CovPW-RPC: Certified Relation-Preserving Correlation Control for Paired Time-Series Generation

Anonymous authors
Paper under double-blind review

Abstract

Realistic individual time series can form an unrealistic pair. Correcting their co-movement may also destroy a relation that the sample already captures. We propose CovPW-RPC, a post-sampling controller for paired diffusion generation. It preserves the declared relation and scales only centered common increments. An exact, monotone response maps this scale to log-return correlation. Its inverse selects the smallest feasible edit toward a target interval, subject to action and energy limits. Targets can be training-set constants or predictions from observed context; the action adapts to each generated path. No generator retraining is required. Identical-draw tests reduce rolling-correlation distribution error on all 15 financial pairs and all ten weather-station pairs. A prospective test on 12 asset-disjoint financial pairs reduces this error by 38.52%. Simple training targets retain over 98% of that gain, supporting the value of the path-dependent inverse. A retrospective fixed-scale comparison further isolates this benefit. The unchanged controller also improves a second diffusion construction. On GRU samples, the frozen policy fails, while a distribution-targeted extension improves development and retrospective temporal results. Exact relation preservation coexists with empirical variance, tail and cointegration tradeoffs. The contribution is a tractable control mechanism: specify what a generated path must retain, expose what it can change, and measure the quality of the resulting edit.

1 Introduction

Useful synthetic time series must capture how variables behave together. Two financial paths may each have plausible volatility, yet their co-movement can be wrong. Two sensor traces may each look realistic, yet their difference can drift in ways absent from the data. Joint generation therefore has two obligations: reproduce individual behavior and retain the relations that make the variables a system. Population-aware generation has begun to address this gap (Li et al., 2025b). A further question arises when a generated sample already captures one relation: can we repair another property without undoing it?

We study this question for pairs. The spread describes a relation between levels; return correlation describes how the paths move together. Improving correlation does not automatically preserve the spread. Conversely, keeping the spread unchanged does not fix correlation. This leaves room for a controlled edit. For example, labels in a stored scenario set may depend on relative-price threshold crossings. Preserving the ratio path keeps those labels fixed while common fluctuations change. Such an edit is useful when a generator is already trained, samples are stored, or the reverse sampling loop is unavailable. It should state what it preserves, how far it can move, and when no useful action is supported.

We develop the method for paired diffusion generation. Our approach separates the pair into a protected relation and a common coordinate. For log prices, these are the difference and average of the two channels. CovPW-RPC changes only centered common increments (Figure 1). This simple restriction has a useful consequence: one positive scale gives an exact response curve for log-return correlation. The curve is monotone on nondegenerate windows, so a budget on the edit defines an explicit reachable interval. A context-based policy chooses the smallest action that enters its target

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.

interval. If the target is out of reach, it selects the nearest feasible boundary. Unsupported inputs remain unchanged.

The contribution is an exact, sample-adaptive control rule for this paired setting. Standard null-space geometry identifies which direction preserves the relation. Our scalar response links that direction to attainable correlation and an action budget. Its inverse selects the edit for each realized path. Target calibration is a separate input: even a constant training mean can induce a different action on every sample. Context-based prediction is an optional refinement. Conditional whitening, including CW-Gen (Yang et al., 2026), supplies established generator geometry. A finite-window bound further states when the log-return change direction survives conversion to percentage returns.

The experiments test repair, comparison and quality costs separately. Identical-draw tests isolate the edit across financial and weather pairs. A prospective test confirms a 38.52% Corr reduction on 12/12 new CSI300 pairs. A separate guarded policy leads the frozen five-method comparison. The primary policy transfers across two CSDI constructions but fails on GRU. Later GRU results improve after changing the target and objective, with development and retrospective evidence. This distinction matters: a reusable editing interface still needs a suitable target policy.

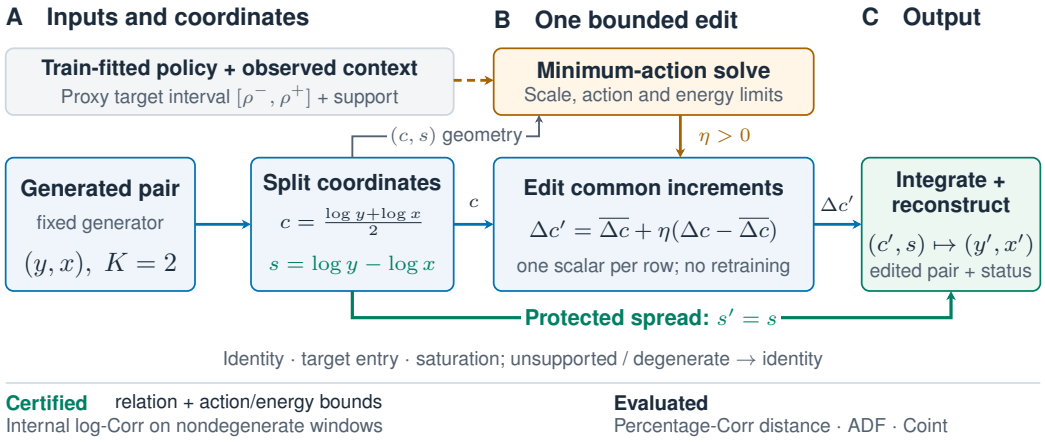

Figure 1: **Control co-movement while retaining the pair relation.** A calibrated target and context-based support check guide the edit. The target may be a training constant or a context prediction. The solver reads both coordinates but edits only common increments. The spread bypasses the edit and is reused in reconstruction. Action and energy limits define the feasible scale; the outcome is identity, target entry or saturation. These are pathwise guarantees. Distributional improvement is measured in the experiments.

2 Related Work

Joint time-series generation. TimeGAN and TimeVAE model temporal structure in multivariate series (Yoon et al., 2019; Desai et al., 2021). Diffusion and score-based modeling (Ho et al., 2020; Song et al., 2021) also support conditional time-series synthesis and imputation (Tashiro et al., 2021). PaD-TS explicitly targets population-level properties, including cross-correlation distributions (Li et al., 2025b). Finance-specific models also address stylized facts and inter-asset dependence (Wiese et al., 2020; Huang et al., 2024; Huh et al., 2026; Cho et al., 2026; Masi et al., 2026). These objectives motivate our evaluation. RPC addresses a different operation: editing a realized pair subject to a fixed path relation.

Whitening and diffusion geometry. Non-isotropic noise and reverse-variance estimation are established ingredients (Voleti et al., 2022; Nichol & Dhariwal, 2021; Bao et al., 2022). CW-Gen jointly learns conditional means and covariances, then uses them in diffusion or flow matching (Yang et al., 2026). We adopt context-covariance whitening as the Base generator. Whitening does not itself specify the relation-preserving edit, its scalar correlation response or its budgeted inverse. Preservation-aware image editing is also prior work (Yu et al., 2023); our analysis concerns a declared time-series relation and the correlation of the resulting path.

Control during and after sampling. PDM and CPS project states or clean estimates during reverse diffusion. SAL/CASAL uses primal–dual Langevin updates, PCD couples pretrained models through a cost and projection, and CCS changes initial noise under diffusion-ODE sampling (Christopher et al., 2024; Narasimhan et al., 2025; Blanke et al., 2026; Luan et al., 2026; Song et al., 2025). Other methods control numerical constraints, attributes, text, regimes or graphs (Coletta et al., 2023; Narasimhan et al., 2024; Jing et al., 2024; Li et al., 2025a; Tanaka et al., 2025; Caulfield & Gleeson, 2025). MCMC correction targets empirical transition statistics, while WaveStitch combines guidance with window stitching (Lin et al., 2026; Shankar et al., 2025). RPC requires only the final sample and context. We evaluate a PDM port and a learned-covariance adaptation on the CSDI backbone. These test mechanisms under shared conditions; they do not reproduce the full published systems. SAL/CASAL, PCD and CCS are related work only. Appendix Table 64 compares the interfaces.

3 Relation-Preserving Correlation Control

3.1 The Protected Relation and the Editable Coordinate

An observed context $\mathbf{X}^{\mathrm{co}} \in \mathbb{R}^{2\times H}$ conditions a generated future $\widetilde{\mathbf{X}}^{\mathrm{ta}} \in \mathbb{R}^{2\times L}$. Our task is conditional trajectory generation: we compare distributions of generated and real paths. For a transformed pair $\mathbf{z}_\ell$, declare the protected relation $r_\ell = \mathbf{a}^\top \mathbf{z}_\ell$ and a common coordinate $c_\ell = \mathbf{b}^\top \mathbf{z}_\ell$, with $\mathbf{T} = [\mathbf{b}^\top; \mathbf{a}^\top]$ invertible. An admissible edit keeps r_ℓ unchanged at every time step.

For finite positive financial paths, set

$$c_{i\ell} = \tfrac{1}{2}(\log y_{i\ell} + \log x_{i\ell}), \qquad s_{i\ell} = \log y_{i\ell} - \log x_{i\ell}. \tag{1}$$

For each generated row i, RPC applies one positive scale η_i:

$$\Delta c'_{i\ell} = \overline{\Delta c}_i + \eta_i(\Delta c_{i\ell} - \overline{\Delta c}_i). \tag{2}$$

Integrating from the original first common coordinate gives c'. Reconstruction uses the original spread:

$$\log y'_{i\ell} = c'_{i\ell} + s_{i\ell}/2, \qquad \log x'_{i\ell} = c'_{i\ell} - s_{i\ell}/2. \tag{3}$$

The edit preserves both endpoints as well as the full spread path, because the centered increments sum to zero. It changes fluctuations around the common trend, leaving the trend itself intact. In particular, $y'_{i\ell}/x'_{i\ell} = y_{i\ell}/x_{i\ell}$. Any predeclared threshold event or first-passage time based only on this ratio is therefore unchanged. This preserves the input scenario's labels; it does not establish that those labels match real data.

Proposition 1 (Exact $K = 2$ admissible-action certificate)**.** *Let $\mathbf{T} = [\mathbf{b}^\top; \mathbf{a}^\top] \in \mathbb{R}^{2\times 2}$ be invertible, with $c_\ell = \mathbf{b}^\top \mathbf{z}_\ell$ and $r_\ell = \mathbf{a}^\top \mathbf{z}_\ell$; write $\mathbf{c} = (c_\ell)_{\ell=1}^L$ and $\mathbf{r} = (r_\ell)_{\ell=1}^L$. A realized output path preserves $\mathbf{a}^\top \mathbf{z}'_\ell = r_\ell$ for every ℓ if and only if there is a unique $\mathbf{u} \in \mathbb{R}^L$ such that, with $\mathbf{e}_1$ the first standard basis vector and $\mathbf{v} = \mathbf{T}^{-1}\mathbf{e}_1$,*

$$\mathbf{z}'_\ell = \mathbf{T}^{-1}[u_\ell, r_\ell]^\top, \qquad \mathbf{z}'_\ell - \mathbf{z}_\ell = (u_\ell - c_\ell)\mathbf{v} \in \ker(\mathbf{a}^\top) = \mathrm{span}\{\mathbf{v}\}.$$

For RPC, $\mathbf{u} = \mathbf{c}'$ is obtained by integrating Equation 2. Together with Equation 3, this preserves log spread and scales centered common-mode increment energy by η_i^2.

3.2 An Exact Correlation Response

Common increments move the two channels together; spread increments move them apart. Scaling their relative energy therefore changes correlation. The next result makes this relation exact on each realized window.

Proposition 2 (Exact scalar response of log-return correlation)**.** *Fix any return window W contained in one controller row, so it shares one η and one row mean. After centering within W, write $u_\ell = \Delta c_\ell - \bar{\Delta} c_W$ and $v_\ell = (\Delta s_\ell - \bar{\Delta} s_W)/2$, and let $A = \mathrm{Var}_W(u)$, $B = \mathrm{Var}_W(v)$, and $C = \mathrm{Cov}_W(u, v)$. For $q = \eta^2$, Pearson correlation of the two reconstructed log-return channels is*

$$\rho_W(q) = \frac{qA - B}{D_W(q)}, \qquad D_W(q) = \sqrt{(qA + B)^2 - 4qC^2}.$$

Where correlation is defined,

$$\frac{d\rho_W}{dq} = \frac{2(qA+B)(AB-C^2)}{D_W(q)^3} \geq 0,$$

with strict inequality when $AB > C^2$. Under this nondegeneracy condition, every positive feasible η-interval maps to an exact reachable interval of realized-path log-return correlation, with a unique interior solution.

Proposition 1 states the admissible geometry; Proposition 2 gives the response of our chosen scalar family. The latter does not cover every relation-preserving path edit. Its value is an exact inverse within a tractable family.

3.3 From a Response Curve to a Bounded Action

On a full row, the average centered energy of the two log-return channels is $E_{\mathrm{pair}}(q) = qA + B$. Starting from $q_0 = 1$, a relative energy budget ϵ_p therefore requires

$$|q-1|A \leq \epsilon_p(A+B). \tag{4}$$

Intersect this condition with a scale box and an action cap to obtain $[\eta_i^-, \eta_i^+]$. The primary policy uses $\eta \in [0.65, 1.35]$, $|\eta - 1| \leq 0.20$ and $\epsilon_p = 0.10$. Monotonicity maps the feasible scales directly to the reachable log-correlation interval.

Target calibration is modular. The inverse below accepts any specified target interval. The original experiments use a context predictor; training-set constants provide simpler alternatives (Table 4). We keep the original predictor fixed for the primary tests.

An affine Ridge model predicts future percentage-return correlation in Fisher-z space. Its five context features describe full and recent correlation, their trend, and common/spread and marginal variance ratios. Fitting, width calibration and support checks use strict-training windows only. Let m_i be the prediction clipped to the training target quantiles $[z_{\mathrm{lo}}, z_{\mathrm{hi}}]$ at $0.005/0.995$, and let σ_z be chronological out-of-fold RMSE. We set

$$z_i^- = \max\{z_{\mathrm{lo}}, m_i - \sigma_z/2\}, \qquad z_i^+ = \min\{z_{\mathrm{hi}}, m_i + \sigma_z/2\}.$$

Applying $\tanh$ and clipping to $[-0.995, 0.995]$ gives $\mathcal{I}_i$. Out-of-fold error sets the tolerance width; the band has no nominal coverage guarantee. It is a numerical proxy for the internal log-return target. We evaluate percentage-return fidelity separately. Appendix B.2 specifies the features, support score and deployment steps.

Minimum-action interval policy. For supported, nondegenerate rows, write ρ_i for the full-row response and solve

$$\rho_i^\star = \mathrm{clip}(\rho_i(1), \mathcal{I}_i), \qquad \eta_i = \operatorname*{arg\,min}_{\eta \in [\eta_i^-, \eta_i^+]} |\rho_i(\eta^2) - \rho_i^\star|, \tag{5}$$

with ties resolved by minimum $|\eta - 1|$. If the baseline already lies inside the target interval, the answer is identity. Otherwise, the controller enters the nearest target boundary or stops at the nearest feasible boundary. Bisection in q uses a 10^{-10} correlation-bracket tolerance and at most 96 steps. Inputs outside hard training support, or with degenerate geometry, return identity. The output includes the scale, relation residual and action status shown in Figure 1.

Target prediction and action adaptation are distinct. For example, if $C_i = 0$, an attainable point target $\rho^\star \in (-1, 1)$ gives $\eta_i^2 = (B_i/A_i)(1+\rho^\star)/(1-\rho^\star)$. Even a shared target can therefore produce path-specific scales. The interval policy also adapts its identity and boundary decisions to each input.

Unless stated otherwise, RPC denotes this *interval policy*. A point-target ablation uses the same solver and support rule with the single target $\tanh(m_i)$. The separate *guarded policy* supplies the frozen system benchmark: it attenuates the action and returns identity for unreachable targets. All share Equations 1–3; their decision rules differ. Appendix Table 5 maps each result to its policy.

3.4 What the Certificate Means for Generation Quality

The exact response uses log returns. The reported Corr metric uses percentage returns and compares rolling-correlation distributions. A finite-window bound connects the two return definitions.

Corollary 1 (Log-to-percentage direction certificate)**.** *For a window W with $n \geq 2$ returns, let $P_W = I_n - \mathbf{1}\mathbf{1}^\top/n$ and $g(z) = e^z - 1$, applied elementwise. For nonconstant vectors, define*

$$R_W(\mathbf{z}) = \frac{2\|P_W(g(\mathbf{z}) - \mathbf{z})\|_2}{\|P_W\mathbf{z}\|_2 + \|P_W g(\mathbf{z})\|_2}, \quad B_W(\mathbf{x}, \mathbf{y}) = \min\{2, R_W(\mathbf{x}) + R_W(\mathbf{y})\}.$$

Then $|\operatorname{corr}_W(g(\mathbf{x}), g(\mathbf{y})) - \operatorname{corr}_W(\mathbf{x}, \mathbf{y})| \leq B_W(\mathbf{x}, \mathbf{y})$. For two admissible actions on the same window, let B_0, B_1 be these bounds and $\Delta_{\log}, \Delta_{\%}$ their correlation changes. Then $|\Delta_{\%} - \Delta_{\log}| \leq B_0 + B_1$. Thus $|\Delta_{\log}| > B_0 + B_1$ certifies agreement in direction.

This certificate concerns a realized change. Whether it moves a generated distribution toward real data also depends on the target policy. Likewise, relation preservation fixes the input spread, including a nonstationary one. A deterministic ADF test on that spread is unchanged; an Engle–Granger test that refits its loading need not be.

There is also a structural reason to evaluate tradeoffs. The two marginal variance changes are

$$\Delta \operatorname{Var}_W(\Delta \log y) = (q-1)A + 2(\eta - 1)C, \qquad \Delta \operatorname{Var}_W(\Delta \log x) = (q-1)A - 2(\eta - 1)C.$$

For $A > 0$, preserving both variances forces $\eta = 1$. A nonidentity edit in this family must therefore change at least one marginal variance. This is not necessarily a quality loss: the new variance may be closer to the real one. It does mean that exact relation preservation and universal marginal preservation are different objectives. Appendix B gives proofs and degenerate cases.

3.5 Generator and Computational Cost

The primary generator is whitened CSDI. It centers each observed context and whitens it using the Cholesky factor of

$$\mathbf{\Sigma}_\lambda^S + 10^{-5}\mathbf{I}, \qquad \mathbf{\Sigma}_\lambda^S = (1-\lambda)\widehat{\mathbf{S}} + \lambda \operatorname{diag}(\widehat{\mathbf{S}}), \quad \lambda = 0.1. \tag{6}$$

It trains and samples with isotropic CSDI, then maps the sample back. *Base* denotes the unedited draw in each panel. A correlation-shaped noise variant with a fixed common-scale edit supplies the remaining comparison branch (Appendix G).

RPC needs no checkpoint update or reverse-loop access. Its path operations are linear in target length, followed by a bounded scalar solve. For the guarded policy's 45 benchmark cells, median generation and controller times are 42.884 and 0.486 seconds. Median cellwise overhead is 1.068%. Appendix E.9 gives the timing scope; this measurement does not describe every controller variant.

4 Experimental Design

Data and comparison roles. The main financial benchmark contains 15 five-minute pairs from cryptocurrency, U.S. equities and CSI300, with horizons $L \in \{288, 390, 240\}$. All splits are chronological. A five-method panel compares complete systems on the same protocol. Separate tests apply multiple controllers to identical Base draws, isolating the effect of the edit. These cover the 15 financial pairs, ten NOAA GHCNh station pairs (Menne et al., 2023), and a prospective set of 12 new, asset-disjoint CSI300 pairs. Each uses three seeds. Only observed context is available to the deployed target policy. Appendix C specifies preprocessing, dates and model fitting.

Metrics follow the task. Our primary control endpoint is Corr, the mean of three distances between real and generated rolling percentage-return correlations: mean error, standard-deviation error and 101-quantile error. The registered window length is 60. Lower values indicate better distributional agreement. Exact preservation of the declared relation is a separate pathwise requirement. ADF and Coint measure financial structure; marginal, tail, temporal and diversity diagnostics measure additional quality costs. We report these costs without requiring one method to win every metric. Their importance depends on the intended use: tail behavior matters directly for risk simulation, and stationarity matters when it is part of the claimed relation. Historical composite ranks retain their original definitions and weights. All original preservation-test decisions are retained (Appendix C.2).

Inference and baseline scope. We average seeds within each pair and use pairs as inferential units. Predeclared contrasts use exact Wilcoxon tests, Holm correction where applicable, and 100,000-resample pair bootstraps. These summaries describe the declared panels. Signed-rank inference assumes independent, symmetric pair differences under the null; pair resampling also assumes independence. Asset-disjoint pairs can still share market shocks, so these intervals do not establish robustness across market periods. Exploratory replays are identified as such. We retain identity, refusal and saturation outcomes. The learned-covariance CSDI baseline adapts conditional whitening. The PDM port projects reverse states on the shared backbone; its retained and changed elements are detailed in Appendix D.5. Full official CW-Gen and PaD-TS are not evaluated here. Ranks therefore describe the named comparison panel, not the full literature. Compute is reported by method; equal training or total inference cost is not assumed. Wider descriptive comparisons and exact decision rules appear in the appendix.

5 RESULTS

5.1 DOES THE EDIT IMPROVE THE SAME GENERATED DRAW?

Table 1 answers this question directly. RPC reduces Corr on all 15 financial pair means, by 15.35% overall. It also improves all ten NOAA station pairs, with a smaller 1.93% reduction. Every arm receives the same generated tensor within its panel. The difference therefore comes from the controller, not a favorable redraw.

Table 1: **RPC improves correlation on identical draws.** Corr errors are pair means after averaging three seeds; lower is better. Each Base/RPC contrast uses the same generated paths. Rows describe different panels and are not pooled into one test. The learned-covariance row is a retrospective adaptation replay, not official CW-Gen.

Panel	Base	RPC	Reduction	Improved pairs
Financial (fresh draws)	.128860	.109075	15.35%	15/15
NOAA GHCNh (new domain)	.144944	.142151	1.93%	10/10
CSI300 (prospective)	.134199	.082505	38.52%	12/12
Learned-cov. replay	.209471	.195491	6.67%	15/15

Prospective confirmation. On the asset-disjoint CSI300 panel, all 36 Base cells and 144 arm endpoints are domain-valid. RPC reduces Corr from 0.134199 to 0.082505, with all 12 pair means improved. The paired difference is -0.051694, the exact one-sided p-value is 0.000244, and the 95% bootstrap upper bound is -0.044362. Figure 2a shows the full pair-level pattern. The point-target ablation reaches a slightly lower mean, 0.080627; the interval policy remains primary because it was fixed before outcomes. Appendix Table 13 reports every arm. A retrospective synchronized calendar-block check retains negative 95% percentile intervals for all three fixed block lengths (Appendix E.1). This probes temporal dependence within the saved panel, conditional on its trained models and draws.

Transfer within diffusion and beyond. The same frozen controller reduces Corr on all 15 archived learned-covariance CSDI pairs, from 0.209471 to 0.195491 (6.67%). A diagnostic excluding windows that cross generated-row joins retains an 8.89% reduction on 14/15 pairs. This supports reuse across two CSDI-based diffusion constructions. In a separate cross-family stress test, the unchanged policy worsens Corr on 11/12 GRU pairs. Its spread and budget certificates still hold; its target policy does not transfer reliably. Appendices D.3 and D.8 give both results. Section 5.3 revisits the GRU limitation with a distribution-targeted variant.

5.2 HOW DOES THE COMPLETE SYSTEM COMPARE?

Guarded RPC achieves the lowest mean Corr error, best mean family rank and best normalized score in the frozen five-method panel (Table 2). Against whitened CSDI, Corr improves on 15/15 pairs; the mean difference is -0.007914 and Holm-adjusted $p = 9.16 \times 10^{-5}$. The strict ADF preservation test fails, and mean Coint distance increases despite passing its historical tolerance. These are separate

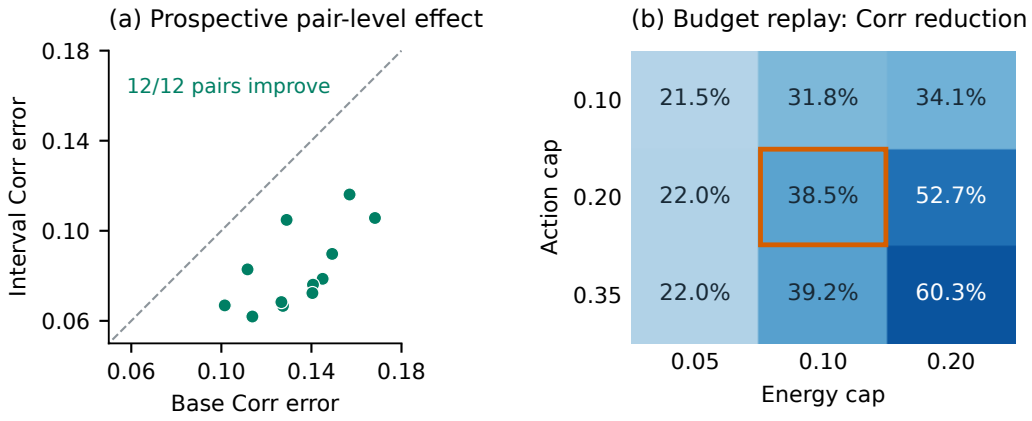

Figure 2: **Consistent pair-level repair and controllable headroom.** (a) Each point is a CSI300 pair, averaged over three seeds. All 12 lie below identity: the RPC interval policy improves Corr on every pair in the prospective test. (b) A later replay varies action and energy caps on the same draws. Cells show percentage reduction in macro Corr error; each improves 12/12 pairs. The outlined cell is the original setting. Panel (b) is post-outcome mechanism evidence, not a new test or a selected optimum.

system runs; the identical-draw tests instead show exact ADF equality. Thus the strongest conclusion is Corr improvement and benchmark-level competitiveness. None of the nine family-by-comparator contrasts involving the learned-covariance adaptation survives Holm correction.

Table 2: **Frozen five-method system comparison.** All methods cover 15 pairs and three seeds. The RPC row uses the guarded policy; learned-covariance CSDI is an adaptation. Corr is primary; ADF and Coint retain their native units. Rank and score retain the original aggregation. Lower is better; bold marks the lowest mean, not significance. Table 19 gives exact contrasts.

Method	Corr	ADF	Coint	Mean rank	Norm. score
Guarded RPC	**0.120009**	17.942	12.685	**2.1222**	**0.417796**
Whitened CSDI	0.127922	17.950	**12.407**	2.5222	0.446799
Corr-shaped + fixed	0.192859	**13.815**	20.338	3.3556	0.510565
Learned-cov. CSDI	0.209471	19.601	16.656	3.4889	0.531562
Isotropic CSDI	0.233791	17.577	15.288	3.5111	0.523551

Family rank retains a unique first place in all 4,944 subsets containing 10–15 pairs. Metric-balanced rank keeps the method first in 4,883 subsets, including two ties; normalized score does so in 4,403. The broader 16-method comparison ranks it first or second under different source-precedence rules (Appendix E.10). These overlapping subsets test sensitivity within the fixed panel; they are not replications.

Comparison with reverse-step projection. The reverse-step projection studies use the same inputs, endpoints and denoiser backbone, but intervene during generation. The earlier study improves Base by 38.75% and does not establish RPC superiority. In a separate 12-pair reserve panel, increasing the energy cap improves both methods on 12/12 pairs (Table 3). RPC at energy cap 0.20 has the lowest observed mean Corr, 0.068134, versus 0.070109 for the PDM port at the same cap. The difference favors RPC on only 8/12 pairs, however, with $p = 0.0881$; the predeclared superiority criterion is not met. The result demonstrates comparable observed errors at these settings, without a statistical equivalence claim. RPC retains the practical advantage of operating on stored draws. Reserve eligibility was amended after initial domain failures; this panel is not an independent confirmation. Appendix D.6 records the amendment, contrasts and timings.

Table 3: Five-arm matched-action/energy follow-up on 12 CSI300 pairs and three seeds. Energy caps are 0.10/0.20, with action cap 0.20. Means average seeds within pair, then pairs equally. Reduction is a ratio of macro means, not an average pair percentage. All 180 endpoints are valid. These arm summaries are descriptive; RPC-over-PDM superiority does not pass.

Arm	Energy cap	Mean Corr distance	Reduction vs Base
Base	—	.137298	—
PDM port	.10	.086639	36.90%
RPC	.10	.086680	36.87%
PDM port	.20	.070109	48.94%
RPC	.20	.068134	50.38%

5.3 Why Does It Work, and What Does It Cost?

Targets and budgets have distinct roles. The fixed CSI300 budget replay yields Corr reductions from 21.52% to 60.31% (Figure 2b). Every tested target-conditioned setting improves all 12 pairs. Yet always choosing the lower feasible boundary worsens all 12; choosing the upper boundary improves only eight. More common-mode energy alone does not explain the result. The target determines the direction, while the budget limits its extent. A strict-training Spline-Ridge alternative also improves all 12 pairs over Base, but beats the original Ridge policy on only five.

Simple targets retain sample-adaptive control. We also replace only the target center with observed-context correlation or a constant training mean or median in Fisher-z space (Table 4). The two constant targets recover 98.10% and 98.36% of the original Base-to-RPC improvement. The learned target has a slightly lower macro Corr error, but beats the training mean on only 6/12 pairs. The constants are fitted per pair; the inverse still adapts the action to every generated path. Both constant-target arms use nonconstant scales within all 36 cells. Thus lightweight target calibration suffices for most of the gain. The benefit of context-based target prediction is small here. Context targets trade higher Corr error for lower W1 and Std errors.

A retrospective ablation selects one fixed scale request per pair on 2020 validation data, then projects it into each path's original feasible interval. Its test Corr is 0.132025, versus 0.082505 for RPC: 37.51% lower error for RPC, with 12/12 pair wins. Calendar-block intervals remain negative at all three tested lengths. RPC uses more of the shared energy allowance on average, but has lower W1, Std and tail errors. This supports path-dependent targeting beyond a calibrated fixed request on this panel (Appendix D.11). Separately, purged, nested calibration changes target RMSE from 0.21144 to 0.21266; it does not establish interval coverage (Appendix E.2).

Table 4: **Simple targets work; fixed requests retain less gain.** All edits share draws, support rules and action/energy caps. Target swaps retain the interval width. Fixed requests are validation-selected and projected into each path's feasible interval. Wins count pairs beating primary RPC. W1 and Std errors are multiplied by 10^3; lower is better. Appendices D.10 and D.11 give protocols.

Control / target	Corr	Wins / 12	W1	Std
Base (no edit)	0.134199	0	0.80122	1.39899
Context predictor (primary)	0.082505	–	0.80064	1.39713
Validation-selected request	0.132025	0	0.82933	1.46131
Observed context	0.086924	4	0.79283	1.37789
Training mean	0.083486	6	0.80359	1.40517
Training median	0.083350	3	0.80031	1.39835

Matching the target to rolling dependence. A later variant targets the context's rolling-correlation distribution. On the 2020 development panel, GRU Corr improves by 22.25% with action/energy caps of 0.20/0.10, and by 41.51% with caps of 0.8/0.8. Applying the selected wide-cap policy to all saved 2021–2022 GRU paths, without reselection, reduces mean Corr by 80.41% and improves 12/12 pairs. W1 and Std errors rise by 1.22% and 5.30%. The conservative caps

lower mean Corr by 18.67%, with 9/12 pair wins. The later panel was previously opened; this is retrospective temporal robustness (Appendix F.2), not independent confirmation.

The certificate is informative on realized paths. A replay certifies agreement between log-return and percentage-return change directions on 95.41% of acted CSI300 windows. All 2,485,440 comparable acted windows agree empirically. Three eligible alternative Corr definitions retain 12/12 favorable directions; the shorter $W = 30$ definition is ineligible because of zero-variance reference windows. These checks support the response mechanism and identify its evaluation domain (Appendices E.3 and E.4).

Quality tradeoffs are part of the result. Shared-draw edits preserve spread-based ADF. A non-identity edit must change marginal variance; refitted Coint can also change (Section 3). The fresh financial test fails its prespecified Coint gate: maximum pair degradation is 2.883106, above the 1.340335 margin (Appendix D.1). In the reserve panel, RPC at energy cap 0.20 reduces Corr by 50.38% and improves tail coexceedance on 12/12 pairs. It also worsens kurtosis distance on 11/12, from 6.244903 to 6.270284 in the mean (0.41%), and increases mean Coint distance. The guarded policy's marginal audit improves 24/28 endpoint means and passes all five diversity/copy screens. These results concern that policy; details appear in Appendices E.6 and D.7.

Utility of saved scenarios. On the saved CSI300 draws, RPC also reduces 5% basket-quantile prediction loss by 1.23%, with 12/12 pair wins. A context-only predictor remains stronger, and variance-allocation cost rises by 0.46%. The benefit is improved saved scenarios under a relation constraint; standalone risk forecasting is a different objective (Appendix D.12).

6 Discussion and Limitations

The method adds bounded control to samples that already exist. Its value is an exact account of the permissible edit and its response. The current analysis covers $K = 2$, one fixed relation and a positive scalar common-mode action. Larger baskets and adaptive loadings need additional geometry. The relation must be declared before editing and checked in Base samples: preserving it cannot create stationarity or certify refitted cointegration.

Architecture-independent inputs do not ensure policy transfer. Frozen-policy gains cover two CSDI constructions. The GRU extension supports broader applicability after target adaptation, but confirmation on previously unexposed data remains open. Likewise, simple targets work well on the CSI300 panel; automatic selection among learned, constant and context targets and identity is not validated.

Comparison scope also matters. Full CW-Gen and PaD-TS lack efficacy rows here; the PDM port tests a mechanism, and the CPS-style study tests feasibility. The reserve panel has an eligibility amendment. The reported Corr gains and five-method rank therefore support a focused correlation repair claim. Variance, tail, refitted Coint and downstream costs determine whether an edit is useful for a particular application.

7 Conclusion

CovPW-RPC controls correlation through one bounded common-mode edit while preserving a declared pair relation. Its exact response gives a reachable interval and a minimum-action policy. Shared-draw and prospective tests support correlation repair across financial and non-financial pairs; baseline and quality comparisons identify where it helps and what it costs. This provides a concrete way to add controllability to a trained time-series generator.

AI USE STATEMENT

This revision builds on an existing manuscript with prior author edits. The authors directed the revision. PaperReview.ai provided review feedback. OpenAI Codex assisted with literature searches, mathematical derivations and proof writing, experimental design and implementation, and result interpretation. It also assisted with manuscript drafting and revision, including figures and LaTeX. Checks against source references, saved numerical outputs and executable examples are documented in the supplement. The authors remain responsible for reviewing and approving the manuscript, claims, code and other artifacts.

ETHICS STATEMENT

The experiments use historical time-series records and do not involve human subjects or interventions. Generated financial trajectories are research artifacts, not investment advice, forecasts, or validated risk estimates. They may reproduce biases, regime limitations, and data-quality problems in the underlying records. Any downstream financial use requires independent legal, risk, and domain review; the exact spread invariant does not certify economic validity or safety.

REPRODUCIBILITY STATEMENT

The main paper specifies the estimator boundary, controller, metrics, and inferential units. The appendix provides assumptions and proofs, exact data splits and horizons, metric definitions and units, hyperparameters, frozen ledger provenance, source-precedence sensitivity, negative results, and known artifact gaps. The anonymous supplementary package contains analysis scripts, tables, figures, manifests, frozen summaries, and a NumPy-only controller fit–apply example with its recovered numerical dependencies. Raw data, full generated arrays, and checkpoints are outside the compact package and must be documented or released subject to source licensing before a full from-scratch replication claim is made.

REFERENCES

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: An analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=53L4PzTYB6B.

Matthieu Blanke, Yongquan Qu, Sara Shamekh, and Pierre Gentine. Strictly constrained generative modeling via split augmented Langevin sampling. In *International Conference on Learning Representations*, pp. 145927–145965, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ebd8a977883fd8d5bccf9575e873ce75-Paper-Conference.pdf.

Howard Caulfield and James P. Gleeson. Graph-guided time series generation with applications to financial correlation modeling. In *Learning on Graphs Conference*, 2025. URL https://openreview.net/forum?id=SfUm1571Fn. Extended abstract.

So-Yoon Cho, Jin-Young Kim, Kayoung Ban, Hyeng Keun Koo, and Hyun-Gyoon Kim. Diffolio: A diffusion model for multivariate probabilistic financial time-series forecasting and portfolio construction. *Information Fusion*, 133:104286, 2026. doi: 10.1016/j.inffus.2026.104286. URL https://arxiv.org/abs/2511.07014.

Jacob K. Christopher, Stephen Baek, and Ferdinando Fioretto. Constrained synthesis with projected diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 89307–89333, 2024. doi: 10.52202/079017-2834. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a28af221f2f70be183afc16797a56b91-Paper-Conference.pdf.

Andrea Coletta, Sriram Gopalakrishnan, Daniel Borrajo, and Svitlana Vyetrenko. On the constrained time-series generation problem. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/bfb6a69c0d9e2bc596e1cd31f16fcdde-Abstract-Conference.html.

Abhyuday Desai, Cynthia Freeman, Zuhui Wang, and Ian Beaver. TimeVAE: A variational auto-encoder for multivariate time series generation. *arXiv preprint arXiv:2111.08095*, 2021. doi: 10.48550/arXiv.2111.08095. URL https://arxiv.org/abs/2111.08095.

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://arxiv.org/abs/0912.0902.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Hongbin Huang, Minghua Chen, and Xiao Qiao. Generative learning for financial time series with irregular and scale-invariant patterns. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CdjnzWsQax.

Jeonggyu Huh, Seungwon Jeong, Hyun-Gyoon Kim, Hyeng Keun Koo, and Byung Hwa Lim. MarketGANs: Multivariate financial time-series data augmentation using generative adversarial networks. *arXiv preprint arXiv:2601.17773*, 2026. doi: 10.48550/arXiv.2601.17773. URL https://arxiv.org/abs/2601.17773.

Baoyu Jing, Shuqi Gu, Tianyu Chen, Zhiyu Yang, Dongsheng Li, Jingrui He, and Kan Ren. Towards editing time series. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1186. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/423d0909791493b7c10916fd328c2913-Abstract-Conference.html.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989. doi: 10.1214/aos/1176347265.

Hao Li, Yu-Hao Huang, Chang Xu, Viktor Schlegel, Renhe Jiang, Riza Batista-Navarro, Goran Nenadic, and Jiang Bian. BRIDGE: Bootstrapping text to control time-series generation via multi-agent iterative optimization and diffusion modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34742–34773. PMLR, 2025a. URL https://proceedings.mlr.press/v267/li25ah.html.

Yang Li, Han Meng, Zhenyu Bi, Ingolv T. Urnes, and Haipeng Chen. Population aware diffusion for time series generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18520–18529, 2025b. doi: 10.1609/aaai.v39i17.34038.

Ci Lin, Futong Li, Tet Yeap, and Iluju Kiringa. Preserving temporal dynamics in time series generation. *arXiv preprint arXiv:2604.27182*, 2026. doi: 10.48550/arXiv.2604.27182. URL https://arxiv.org/abs/2604.27182.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9881–9893. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/4054556fcaa934b0bf76da52cf4f92cb-Paper-Conference.pdf.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

Hao Luan, Yi Xian Goh, See-Kiong Ng, and Chun Kai Ling. Projected coupled diffusion for test-time constrained joint generation. In *International Conference on Learning Representations*, pp. 130212–130261, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d3831fa12e608fc9f7690e05b599e69e-Paper-Conference.pdf.

Giuseppe Masi, Andrea Coletta, and Novella Bartolini. High-quality synthetic financial time-series using a GAN-diffusion framework. *arXiv preprint arXiv:2605.27113*, 2026. doi: 10.48550/arXiv.2605.27113. URL https://arxiv.org/abs/2605.27113.

Matthew J. Menne, Simon Noone, Nancy W. Casey, Robert H. Dunn, Shelley McNeill, Diana Kantor, Peter W. Thorne, Karen Orcutt, Sam Cunningham, and Nicholas Risavi. Global historical climatology network-hourly (GHCNh), version 1. NOAA National Centers for Environmental Information, 2023. URL https://doi.org/10.25921/jp3d-3v19.

Sai Shankar Narasimhan, Shubhankar Agarwal, Oguzhan Akcin, Sujay Sanghavi, and Sandeep P. Chinchali. Time Weaver: A conditional time series generation model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37293–37320. PMLR, 2024. URL https://proceedings.mlr.press/v235/narasimhan24a.html.

Sai Shankar Narasimhan, Shubhankar Agarwal, Litu Rout, Sanjay Shakkottai, and Sandeep P. Chinchali. Constrained posterior sampling: Time series generation with hard constraints. In *Advances in Neural Information Processing Systems*, volume 38, pp. 107923–107963, 2025. doi: 10.52202/085713-3599. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/9b01c4a7d3fc49875dad3c13848bcd9e-Paper-Conference.pdf.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

Aditya Shankar, Lydia Y. Chen, Arie van Deursen, and Rihan Hai. WaveStitch: Flexible and fast conditional time series generation with diffusion models. *Proceedings of the ACM on Management of Data*, 3(6), 2025. doi: 10.1145/3769842. URL https://arxiv.org/abs/2503.06231.

Bowen Song, Zecheng Zhang, Zhaoxu Luo, Jason Hu, Wei Yuan, Jing Jia, Zhengxu Tang, Guanyang Wang, and Liyue Shen. CCS: Controllable and constrained sampling with diffusion models via initial noise perturbation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 70516–70552, 2025. doi: 10.52202/085713-2371. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/663ebbb266aec936d726fba950ca693a-Paper-Conference.pdf.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yuki Tanaka, Ryuji Hashimoto, Takehiro Takayanagi, Zhe Piao, Yuri Murayama, and Kiyoshi Izumi. CoFinDiff: Controllable financial diffusion model for time series generation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 9357–9365, 2025. doi: 10.24963/ijcai.2025/1040. URL https://www.ijcai.org/proceedings/2025/1040.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pp. 24804–24816. Curran Associates, Inc., 2021.

Vikram Voleti, Christopher Pal, and Adam Oberman. Score-based denoising diffusion with non-isotropic Gaussian noise models. In *NeurIPS Workshop on Score-Based Methods*, 2022. URL https://openreview.net/forum?id=igC8cJKcb0Q.

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

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=INAeUQ04lT.

Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. Quant GANs: Deep generation of financial time series. *Quantitative Finance*, 20(9):1419–1440, 2020. doi: 10.1080/14697688.2020.1730426.

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

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

Yanfeng Yang, Siwei Chen, Pingping Hu, Zhaotong Shen, Yingjie Zhang, Zhuoran Sun, Shuai Li, Ziqi Chen, and Kenji Fukumizu. Conditionally whitened generative models for probabilistic time series forecasting. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GG01lCopSK.

Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Xi Yu, Xiang Gu, Haozhi Liu, and Jian Sun. Constructing non-isotropic Gaussian diffusion model using isotropic Gaussian diffusion model for image editing. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3357. URL https://papers.nips.cc/paper_files/paper/2023/hash/f25602918e8a0d0c86e3c752ecfbbaa1-Abstract-Conference.html.

Xinyu Yuan and Yan Qiao. Diffusion-TS: Interpretable diffusion for general time series generation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b5b66077d016c037576cc56a82f97f66-Abstract-Conference.html.

A GUIDE TO THE SUPPLEMENT

A short route through the evidence. Five tables cover the empirical argument. Table 1 shows identical-draw repair and prospective confirmation. Table 4 separates target prediction from path-adaptive action. Table 3 compares post-sampling control with reverse-step projection and retains the failed superiority decision. Table 49 places the frozen GRU failure beside the later temporal replay, which is retrospective. Table 32 reports application benefits and costs, including the stronger context-only predictor. The financial Coint gate failure is stated in the main text and detailed in Section D.1.

Section B gives proofs and the complete primary controller. Section C defines data and metrics. Sections D and E provide the full comparison and quality records. Section F contains later candidate studies. The earlier geometry branch in Section G is supporting material, not part of the primary controller. Section H describes the released artifacts and maps functional names to the identifiers in saved records.

Table 5: Controller choices and the evidence they support. RPC without a qualifier denotes the primary interval policy. Results from distinct policies and panels are not pooled.

Policy	Decision rule	Evidence
RPC (primary)	Minimum action toward a calibrated interval; action cap .20, energy cap .10	Identical-draw financial and weather tests; prospective CSI300 test; frozen diffusion and GRU transfer
Guarded RPC	Attenuated action; unreachable targets return identity	Frozen five-method comparison, runtime and marginal-quality audit
RPC at two energy caps	Same interval rule; energy cap .10 or .20	Reserve comparison with the PDM port; quality costs and timings
Distribution-targeted extension	Match rolling-correlation distributions	Development and retrospective GRU temporal replay (Section F)

B CONTROLLER THEORY AND IMPLEMENTATION

The first part proves the realized-path guarantees used by RPC. The background whitening results appear later with the historical geometry studies, so they cannot be mistaken for the controller's novelty.

B.1 RELATION PRESERVATION, SCALAR RESPONSE AND RETURN CONVERSION

We use the coordinates and reconstruction in Equations 1–3. Financial inputs are finite and positive. Guarantees concern the edited input path; improvement in a distributional distance remains empirical.

Proof of Proposition 1. The displayed reconstruction makes the second coordinate of $\mathbf{T}\mathbf{z}'_\ell$ equal to r_ℓ, proving sufficiency. Conversely, for any relation-preserving realized output, set $u_\ell = \mathbf{b}^\top \mathbf{z}'_\ell$. Then $\mathbf{T}\mathbf{z}'_\ell = [u_\ell, r_\ell]^\top$, so invertibility gives the representation and its uniqueness. Subtracting $\mathbf{T}\mathbf{z}_\ell = [c_\ell, r_\ell]^\top$ gives the displacement formula. Finally, with $\mathbf{e}_1, \mathbf{e}_2$ the standard basis, $\mathbf{a}^\top \mathbf{v} = \mathbf{e}_2^\top \mathbf{e}_1 = 0$, and $\mathbf{v} \neq 0$ because $\mathbf{T}$ is invertible. Since $\mathbf{a} \neq 0$ in $\mathbb{R}^2$, its one-dimensional kernel is $\mathrm{span}\{\mathbf{v}\}$. For the specified controller, $\Delta c'_\ell - \overline{\Delta c} = \eta(\Delta c_\ell - \overline{\Delta c})$. Squaring and summing gives the claimed η^2 energy scaling. The centered increments sum to zero, so integration also retains both common-coordinate endpoints. □

What relation preservation leaves free. Under Proposition 2, the centered log-return channels are $\eta u + v$ and $\eta u - v$. Relative to the unedited action $\eta_0 = 1$, direct covariance expansion gives

$$\begin{aligned} \Delta \operatorname{Var}_W(\Delta \log y) &= (\eta^2 - 1)A + 2(\eta - 1)C, \\ \Delta \operatorname{Var}_W(\Delta \log x) &= (\eta^2 - 1)A - 2(\eta - 1)C, \\ \Delta \operatorname{Cov}_W(\Delta \log y, \Delta \log x) &= (\eta^2 - 1)A. \end{aligned}$$

Thus exact spread preservation does not fix either marginal return variance. For $A > 0$, preserving both variances forces $\eta^2 = 1$, hence $\eta = 1$ under the positive-scalar restriction. In this nondegenerate scalar family, a nonidentity correlation repair must change at least one channel variance. The paired-energy budget bounds the aggregate change; it does not guarantee that either variance moves closer to its real-data value. Higher moments and temporal dependence require empirical checks as well (Appendix E.6 and Tables 23–24). By contrast, applying an unchanged deterministic ADF evaluator to the exact same spread gives the same result. This preserves the input diagnostic, including a nonstationary input; it does not create stationarity. A refitted Engle–Granger loading defines a different residual and is not protected. These consequences follow from the existing action algebra and introduce no new empirical endpoint or no-harm guarantee.

The common-increment edit is the positive-valued financial specialization in log-common and log-spread coordinates. The equivalence characterizes exact realized $K = 2$ actions under a fixed linear relation; validation and the declared guards select which bounded common-coordinate map, if any, is useful.

Proof of Proposition 2. Let P_W denote centering over a return window W contained in one controller row. Equation 2 uses one row mean and one positive η, so the constant row-mean term disappears after window centering:

$$P_W \Delta c' = \eta P_W \Delta c = \eta \mathbf{u}.$$

Spread is unchanged, hence $P_W(\Delta s'/2) = P_W(\Delta s/2) = \mathbf{v}$. The two centered reconstructed log-return vectors are therefore $\eta\mathbf{u} + \mathbf{v}$ and $\eta\mathbf{u} - \mathbf{v}$. Using one common normalization for variance and covariance gives

$$\begin{aligned} \mathrm{Cov}_W(\Delta \log y', \Delta \log x') &= qA - B, \\ \mathrm{Var}_W(\Delta \log y') &= qA + B + 2\sqrt{q}\, C, \\ \mathrm{Var}_W(\Delta \log x') &= qA + B - 2\sqrt{q}\, C. \end{aligned}$$

Multiplying the two variances yields $(qA + B)^2 - 4qC^2$, proving the displayed formula for $\rho_W(q)$. Direct differentiation gives

$$\rho'_W(q) = \frac{2(qA + B)(AB - C^2)}{\left((qA + B)^2 - 4qC^2\right)^{3/2}}.$$

Cauchy–Schwarz gives $C^2 \le AB$; the derivative is positive whenever $AB > C^2$ and both channel variances are nonzero. Continuity and strict monotonicity then map every positive feasible q-interval to the exact reachable correlation interval and give a unique solution for any interior target.

The degenerate cases are explicit. If $A = 0 < B$, correlation is -1; if $B = 0 < A$, it is $+1$; and if $A = B = 0$, it is undefined. If $A, B > 0$ and $C^2 = AB$, the centered common and half-spread returns are collinear, so correlation is piecewise -1 or $+1$ and is undefined at the zero-variance boundary $qA = B$. These rows cannot support a unique interior inverse and are handled by the existing sensitivity and feasibility guards. □

Proof of Corollary 1. For a nonzero vector $\mathbf{a}$, write $N(\mathbf{a}) = \mathbf{a}/\|\mathbf{a}\|_2$. If $\mathbf{a}, \mathbf{b} \neq \mathbf{0}$, set $\alpha = \|\mathbf{a}\|_2$, $\beta = \|\mathbf{b}\|_2$, and $c = \mathbf{a}^\top \mathbf{b}/(\alpha\beta)$. Direct expansion gives

$$\left(\frac{2\|\mathbf{a} - \mathbf{b}\|_2}{\alpha + \beta}\right)^2 - \|N(\mathbf{a}) - N(\mathbf{b})\|_2^2 = \frac{2(1 + c)(\alpha - \beta)^2}{(\alpha + \beta)^2} \ge 0.$$

Consequently,

$$\|N(\mathbf{a}) - N(\mathbf{b})\|_2 \le \frac{2\|\mathbf{a} - \mathbf{b}\|_2}{\|\mathbf{a}\|_2 + \|\mathbf{b}\|_2}.$$

Since $P_W \mathbf{x} \neq \mathbf{0}$, the entries of $\mathbf{x}$ are not all equal. The map $g(t) = e^t - 1$ is strictly increasing, so the entries of $g(\mathbf{x})$ are also not all equal and $P_W g(\mathbf{x}) \neq \mathbf{0}$. The same argument applies to $\mathbf{y}$.

Pearson correlation can therefore be written as the inner product of centered unit vectors. Applying the preceding inequality first to $P_W\mathbf{x}, P_W g(\mathbf{x})$ and then to $P_W\mathbf{y}, P_W g(\mathbf{y})$ yields

$$\begin{aligned}&|\mathrm{corr}_W(g(\mathbf{x}), g(\mathbf{y})) - \mathrm{corr}_W(\mathbf{x}, \mathbf{y})| \\ &\quad \le \|N(P_W g(\mathbf{x})) - N(P_W\mathbf{x})\|_2 + \|N(P_W g(\mathbf{y})) - N(P_W\mathbf{y})\|_2 \\ &\quad \le R_W(\mathbf{x}) + R_W(\mathbf{y}).\end{aligned}$$

The left side is at most two because both correlations lie in $[-1, 1]$, which proves the stated bound by B_W.

For comparison with a simpler analytic radius, define

$$\delta_{\mathbf{z}} = \frac{\|P_W(g(\mathbf{z}) - \mathbf{z})\|_2}{\|P_W\mathbf{z}\|_2}.$$

The denominator in $R_W(\mathbf{z})$ is at least $\|P_W\mathbf{z}\|_2$, so $R_W(\mathbf{z}) \le 2\delta_{\mathbf{z}}$. If $\|\mathbf{z}\|_\infty \le M$, Taylor's theorem gives, coordinatewise,

$$|e^{z_j} - 1 - z_j| \le \frac{e^M}{2}|z_j|^2.$$

Because P_W is an orthogonal projection,

$$\delta_{\mathbf{z}} \le \frac{e^M\|\mathbf{z} \odot \mathbf{z}\|_2}{2\|P_W\mathbf{z}\|_2}.$$

This looser expression can be vacuous for nearly constant windows; the data-dependent B_W remains the operative certificate.

For action k, write

$$e_k = \mathrm{corr}_W(g(\mathbf{x}_k), g(\mathbf{y}_k)) - \mathrm{corr}_W(\mathbf{x}_k, \mathbf{y}_k), \qquad |e_k| \le B_k.$$

Then

$$\Delta_{\%} = \Delta_{\log} + e_1 - e_0, \qquad |\Delta_{\%} - \Delta_{\log}| \le B_0 + B_1.$$

Hence $\Delta_{\log} > B_0 + B_1$ implies $\Delta_{\%} > 0$, while $\Delta_{\log} < -(B_0 + B_1)$ implies $\Delta_{\%} < 0$. This proves the two-direction statement. The bound compares two realized correlations only; it contains neither a real-data target nor the mean, standard deviation, and quantile aggregation used by the reported Corr-family distance. It therefore does not imply improvement of that external distance. □

For completeness, take W to be the whole controller row. The centered common-mode energy and average centered energy of the two reconstructed log-return channels are, respectively,

$$E_c(q) = qA, \qquad E_{\mathrm{pair}}(q) = \tfrac{1}{2}\{\mathrm{Var}(\Delta \log y') + \mathrm{Var}(\Delta \log x')\} = qA + B.$$

Thus $|E_c(q) - E_c(q_0)| = |q - q_0|A$. A relative common-mode budget ϵ_c would be

$$|q - q_0|A \le \epsilon_c q_0 A,$$

while a relative paired-energy budget ϵ_p is

$$|q - q_0|A \le \epsilon_p(q_0 A + B).$$

The historical guarded policy and certified CovPW-RPC implementations freeze the second, paired-energy form. Only the common coordinate is edited, but the reference denominator is the total paired centered energy $q_0 A + B$. The legacy code field name `common_mode_energy_budget` describes the edited direction, not that denominator. When $A = 0$, the common channel has no controllable centered energy and the sensitivity guard returns identity.

The proof is deliberately coordinate-specific. Percentage returns apply an exponential map before correlation, and the reported Corr endpoint additionally compares the mean, standard deviation, and 101 quantiles of rolling-correlation distributions. Neither transformation inherits the derivative sign above unconditionally. Corollary 1 supplies a sufficient finite-window margin for the first transformation, but not for distance to the real Corr distribution. Likewise, Engle–Granger refits its loading. These outcomes are tested, not certified by Proposition 2.

B.2 EXACT TARGET AND IMPLEMENTATION SPECIFICATION

The first stored channel is y, and the second is x, as in the main-text reconstruction. For each observed window, define percentage returns $r_y[t] = y_{t+1}/y_t - 1$ and log returns $d_y[t] = \log y_{t+1} - \log y_t$, and define r_x, d_x analogously. Write $z(\rho) = \operatorname{atanh}(\operatorname{clip}(\rho, -0.995, 0.995))$. The five context features, in their stored order, are

$$f_1 = z(\rho_{\text{full}}), \qquad f_2 = z(\rho_{60}), \qquad f_3 = z(\rho_{60}) - z(\rho_{\text{full}}),$$
$$f_4 = \log \frac{\operatorname{Var}((d_y + d_x)/2) + \epsilon}{\operatorname{Var}((d_y - d_x)/2) + \epsilon}, \qquad f_5 = \log \frac{\operatorname{Var}(r_y) + \epsilon}{\operatorname{Var}(r_x) + \epsilon},$$

where $\epsilon = 10^{-12}$. The recent correlation ρ_{60} uses the last $\min(60, T_{\text{context}} - 1)$ percentage returns. All variances use the population convention. An undefined full correlation is replaced by zero; an undefined recent correlation uses the effective full value. The trend feature is a difference of Fisher transforms, not a fitted time slope. The training target is z of the within-future percentage-return correlation. The context–future boundary return is excluded.

Training and width calibration. Training rows are deduplicated by target-start timestamp and sorted chronologically. For N rows, set $s = \max(3, \lfloor N/2 \rfloor)$ and take six equally spaced integer edges from s to N. Each of five folds fits the preceding prefix and scores the next block. Feature means and population standard deviations are fitted on that prefix; scales at or below 10^{-8} are replaced by one. Ridge uses an unpenalized intercept and the grid $\{0, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. We select the largest penalty within one standard error of the lowest mean fold MAE in Fisher-z units, then refit on all training rows. The same expanding folds estimate pooled out-of-fold RMSE σ_z at the selected penalty, with a floor of 10^{-6}. They have no purge gap between adjacent windows. This is a training-only resolution estimate, not independent calibration or a nominal coverage guarantee.

Support and target. With final training feature means μ and scales σ, define

$$D(f) = \sqrt{\frac{1}{5} \sum_{j=1}^{5} \left(\frac{f_j - \mu_j}{\sigma_j} \right)^2}.$$

The hard radius is the larger of its training 99.5th percentile and its 90th percentile plus 10^{-6}. RPC returns identity above this radius. This is an RMS standardized distance, not a Mahalanobis distance. Let a, b be the 0.5th and 99.5th percentiles of the training targets, using linear quantile interpolation. For the clipped prediction $\bar{m} = \operatorname{clip}(m(f), a, b)$, the target is

$$\mathcal{I}(f) = \left[\tanh(\max(a, \bar{m} - 0.5\sigma_z)) \,,\, \tanh(\min(b, \bar{m} + 0.5\sigma_z)) \right].$$

The solver clips its endpoints to $[-0.995, 0.995]$.

Numerics and replay. Controller arithmetic uses float64. The solver bisects in $q = \eta^2$ for at most 96 steps and stops at correlation-bracket width 10^{-10}. It returns identity when $A \le 10^{-18}$, $B \le 10^{-18}$, or $AB - C^2 \le 10^{-12} \max(AB, 10^{-36})$, or when a baseline or feasible endpoint correlation is undefined. Nonpositive or nonfinite paths are rejected. The prospective CSI300 caches use 240 observed and 240 future points. Their stored target-start timestamp denotes point 240; point zero is 240 trading bars earlier. NOAA uses 288 observed and 288 future hours. Returns are always formed within each window.

The supplementary NumPy-only entry point creates a synthetic training cache, fits the target and calibration, applies the controller, and checks its invariants using preserved historical numerical source files. It tests implementation behavior; it does not reproduce diffusion training or constitute a new benchmark. Section H describes the dependency recovery and the release boundary.

B.3 GENERATOR: OBSERVED-CONTEXT COVARIANCE WHITENING

Whitened CSDI keeps the isotropic denoiser and applies the covariance transform in Equation 6. Coverage is complete for all 45 pair–seed cells. The supporting correlation-shaped generator uses

context correlation in its kernels and a shrunk matrix in its precision loss. These differ from covariance whitening, so the same-metric equivalence does not identify the two empirical branches.

In the earlier 14-method comparison, whitened CSDI ranks first without extra denoiser capacity. Against the correlation-shaped generator with a fixed edit, however, its pairwise result is mixed: 7 wins, 8 losses, one-sided $p = 0.445$. Guarded RPC adds the controller to whitened CSDI for the frozen five-method comparison. Artifact aliases appear in Table 63.

Table 6: Observed-context CovPW conditioning audit over 6,720 unique matrices from 15 seed-deduplicated caches. Values reconstruct the frozen expression in CPU float32; they are not byte-replayed historical runtime matrices or a performance-sensitivity experiment. Here κ_2 is reported only for SPD matrices; all 6,720 are eligible.

Matrix	Median κ_2	q95	q99	Maximum	Chol. failures
Raw $\widehat{\mathbf{S}}$	18.39	19,957.10	125,513.87	461,138.64	0/6,720
Shrunk $\mathbf{\Sigma}^S_{.1}$	10.54	23.53	62.48	1,027.47	0/6,720
$\mathbf{\Sigma}^S_{.1} + 10^{-5}\mathbf{I}$	6.72	17.78	40.70	573.82	0/6,720

We also reconstruct the frozen covariance expression on the 6,720 unique observed test contexts after verifying that the three seed rows share each of 15 caches. Shrinkage reduces the pooled median/q99 spectral condition number from 18.39/125,513.87 to 10.54/62.48; adding the configured 10^{-5} identity jitter reduces them further to 6.72/40.70. Float32 Cholesky has zero failures at all three stages. The diagonal clamp acts in 541/6,720 matrices (8.05%), and 1,606/6,720 shrunk matrices (23.90%) have minimum eigenvalue at or below 10^{-5}, so the regularization is operational rather than cosmetic. This post-hoc audit documents only the bound CPU-float32 reconstruction: the later runtime device is unknown, CPU/CUDA roundoff may change near-threshold categories, and it neither byte-replays historical matrices nor studies sensitivity to λ, jitter, or window length.

B.4 Guarded comparison policy

The guarded controller is fitted only on strict training windows. Its architecture, search domain, uncertainty rule, and fallbacks are frozen before ledger-frozen generation. Table 9 gives the complete operational historical policy specification. The primary RPC policy is specified separately in Section B.2.

Table 7: Frozen guarded RPC controller audit over 45 generated targets (immutable ledger label: CTSDiff-RPC).

Diagnostic	Observed
Controller-input spread certificates	45/45
Regime-aligned rows	0/20,160
Positive-projected rows	0
Median median $\lvert\Delta\eta\rvert$	0.021435
Median q90 $\lvert\Delta\eta\rvert$	0.049447
Maximum $\lvert\Delta\eta\rvert$	0.122159
Maximum relative paired-energy change	0.098469
Median/max hard fallback	19.79%/33.33%

A post-hoc full-benchmark matched-input reanalysis covers all 45 canonical pair–seed cells. It uses the certified inverse of the saved RPC transform, newly evaluates only the recovered base arrays, and reuses SHA-bound frozen reference and candidate metrics. Corr improves on 15/15 pairs (mean -0.007914, Holm $p = 9.155 \times 10^{-5}$, bootstrap upper -0.005040); ADF family distances are exactly equal; and Coint passes the reused ledger-frozen 1.340335 noninferiority margin. These findings isolate the saved controller increment on an opened benchmark and are not a newly registered or independent confirmation.

Guarded RPC supplies the frozen five-method competitiveness result. For row i, $\eta_i^\star$ targets the predicted correlation; $h_i \in [0, 1]$ combines support, signal, sensitivity, and reachability; and $[\eta_i^-, \eta_i^+]$

enforces action and energy. The applied scale is

$$\eta_i = \eta_0 + h_i\big[\mathrm{clip}(\eta_i^\star, \eta_i^-, \eta_i^+) - \eta_0\big]\,. \tag{7}$$

Unsupported or unreachable guarded policy rows set $h_i = 0$. point-target RPC uses the RPC solver and support rule with the single clipped target $\tanh(m_i)$; it is a matched ablation.

C Data and Evaluation Protocol

C.1 Data, chronological splits and computation

The primary financial experiments use 5-minute asset pairs with chronological splits and horizons $L = 288$, 390, and 240 for cryptocurrency, U.S. equities, and CSI300 stocks. Table 8 lists the pairs, boundaries, and cached windows. The paired task fixes $K = 2$ before fitting. Base models and fixed SPRC use train/validation selection; each CovPW-RPC realization fits its ridge, support, and uncertainty policy on strict training windows only. Inference-time control sees observed context and generated samples, never future targets, test metrics, or rankings. The broad and ledger-frozen blocks use the same 15-pair test split and nine diagnostics.

The Electricity and Traffic checks use standard public matrices. Pair selection is frozen from training-split dependence statistics and stratified by relation type. Train-only standardization and positive scaling reuse the pair loader and relation/common evaluator without test leakage. The independent GHCNh T3-v2 panel instead uses 288 context and 288 target hours. Each two-station window is divided channelwise by its own first context value before fitting or generation; the exact divisors and inverse replay are audited.

Table 8: Canonical financial data protocol. Splits are chronological and fixed before comparison. The window counts are train/validation/test; training is capped at 10,000 per pair. All dates are inclusive.

Field	Cryptocurrency	U.S. equity/ETF	CSI300
Pairs	ATA–LIT BTC–ETH FLOW–ZIL SOL–DOGE XRP–BNB	COKE–KO JPM–BAC MA–V MSFT–AAPL SPY–IVV	600048–600837 601288–601988 601318–601628 601398–601939 601857–600028
Horizon L	288	390	240
Training	2020-01-01 to 2023-10-31	2010-01-01 to 2019-12-31	2010-01-01 to 2019-12-31
Validation	2023-11-01 to 2023-12-31	2020-01-01 to 2020-12-31	2020-01-01 to 2020-12-31
Test	2024-01-01 to 2024-12-31	2021-01-01 to 2022-12-31	2021-01-01 to 2022-12-31
Windows	10000/61/366	10000/253/503	10000/243/484

The CTSDiff-Corr family keeps the same CSDI-style denoiser across markets. Its frozen implementation uses sanitized context correlation $\mathbf{R}$ in forward and reverse noise and $\mathbf{R}_{.1} = .9\mathbf{R} + .1\mathbf{I}$ in the full precision loss; the diagonal variant removes off-diagonal precision terms. CTSDiff-SPRC applies validation-selected common-mode scale $\eta = 1.5$ at inference time. The distinct guarded RPC row uses CSDI-CovPW, which whitens with a shrunk context covariance, and adds the frozen guarded controller from the main paper. These branches are not same-metric implementations. Completion and validation used up to nine parallel NVIDIA GeForce RTX 2080 Ti jobs.

C.2 Metric definitions and aggregation

For each evaluation sequence of ADF or Engle–Granger p-values, define $b_j = \mathbb{1}[p_j < 0.05]$. The three reported distances are the absolute real–generated difference in pass proportion, the mean length of maximal consecutive-one runs (defined as zero when no run exists), and the maximum run length. Their units are proportion, windows, and windows. For rolling return correlations, the components are absolute mean difference, absolute standard- deviation difference, and

Table 9: Frozen guarded RPC controller specification (immutable ledger label: CTSDiff-RPC). All fit and selection operations use strict training windows; generation never reads the realized target horizon.

Component	Frozen specification	Safeguard
Target features	Standardized Fisher-z full and recent percentage-return correlation, correlation trend, log common/spread variance ratio, and log marginal variance ratio	Five observed-context features only
Ridge selection	ℓ_2 grid $\{0, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$; five chronological expanding folds; minimum training fraction 0.5; one-standard-error choice	Chronological train-only residuals
Support and uncertainty	Feature-support soft/hard quantiles $0.90/0.995$; target Fisher-z support quantiles $0.005/0.995$; uncertainty multiplier 0.50	Hard out-of-support fallback
Scale solver	$\eta \in [0.65, 1.35]$; 65-point grid followed by 24 deterministic refinements; identity anchor $\eta_0 = 1$	Minimum sensitivity 10^{-4}
Applied policy	Global gain 1.0; maximum deviation $\lvert\eta - \eta_0\rvert \leq 0.20$; relative paired centered log-return energy budget 0.10	Clip to rowwise feasible interval
Fallback and certificate	For guarded policy, return identity on hard OOD, weak sensitivity, or uncertainty-adjusted unreachable targets; transform centered common-mode increments only	Exact finite controller-input log-spread preservation

$101^{-1} \sum_{q \in \{0,.01,\ldots,1\}} |Q_R(q) - Q_G(q)|$, all in correlation units. Thus the family-rank view averages native-unit distances before ranking; the component-balanced view first ranks each component and is less sensitive to the run-length scale.

Table 10: Nine component distances. Each entry is an absolute real–generated difference; lower is better.

Family	Components	Units
ADF	pass ratio; mean/max passing-run length	proportion; windows
Coint	pass ratio; mean/max passing-run length	proportion; windows
Corr	mean; standard deviation; 101-quantile distance	correlation

Table 11: Mean component distances over 15 ledger pairs. Δ is guarded RPC minus CSDI-CovPW; negative is lower. These native units are not combined here.

Family	Component	CovPW	RPC	Δ
ADF	pass ratio	0.128303	0.128304	0.000000
ADF	mean run	4.342473	4.342344	-0.000129
ADF	max run	49.377778	49.355556	-0.022222
Coint	pass ratio	0.119960	0.119682	-0.000278
Coint	mean run	3.833749	3.824827	-0.008922
Coint	max run	33.266667	34.111111	0.844444
Corr	mean	0.079641	0.076184	-0.003457
Corr	standard deviation	0.167116	0.157753	-0.009363
Corr	101-quantile	0.137010	0.126090	-0.010920

All three Corr components are lower for guarded RPC, while the Coint maximum-run component is higher. This explains why the evidence supports Corr superiority and Coint noninferiority under its frozen margin rather than uniform relational improvement.

D Complete Attribution and Comparator Evidence

The primary unit is the asset or station pair after seed averaging. Saved-draw replays change only the controller within a cell. Comparisons between separate generator runs address complete-system performance.

D.1 Fresh financial and non-financial identical-draw tests

The T2 financial protocol creates 45 new base-generation cells and evaluates one Base reference and three frozen controller arms on each shared draw, for 180 arm cells. The RPC-minus-Base Corr difference is negative on all 15 pairs, with mean -0.0197845, exact one-sided $p = 3.05 \times 10^{-5}$, and deterministic 95% pair-bootstrap upper bound -0.0143282. ADF family distances are exactly equal. The prespecified Coint safety gate fails: the maximum pair degradation is 2.883106, above the frozen 1.340335 margin. Because the financial panel was already opened, T2 is a fresh controller-specific execution rather than an independent confirmation.

T3 freezes ten U.S. GHCNh station pairs, three seeds, and 12 target windows per pair before performance outcomes are opened Menne et al. (2023). All 30 generation cells and 1,440 arm–target rows complete. RPC improves Corr on 10/10 pairs: mean -0.00279295, exact one-sided $p = .0009765625$, and 95% bootstrap upper bound -0.00145096. ADF distances are exactly equal. Each 576-hour station path is divided by its own first context value. RPC therefore preserves the log ratio of the two normalized pressure paths, equal to raw log-pressure difference up to a window-fixed constant. The endpoint is the unchanged 60-point rolling correlation of percentage changes, not the internal log-return quantity in Proposition 2. Coint differences are favorable for NYC, SF Bay, Seattle, Atlanta, and Chicago and adverse for Los Angeles, Washington DC, Houston, Twin Cities, and Phoenix. These five favorable and five adverse differences are fully public and are not converted into a safety, noninferiority, ranking, or optimal-setting claim. The primary reduction, complete pair rows, 30-cell audit, and supplemental artifact audit are SHA-bound in the anonymous bundle.

Table 12: Fresh shared-draw tests: RPC minus Base in Corr distance. Tests use pair means after seed averaging. $U_{.95}$ is the one-sided 95% bootstrap upper bound. ADF is equal in both panels. The financial Coint gate fails; GHCNh has five favorable and five adverse Coint changes.

Panel	Base	RPC	Pair W/L/T	Exact p	$U_{.95}$
Financial	.128860	.109075	15/0/0	3.05×10^{-5}	-.014328
GHCNh	.144944	.142151	10/0/0	9.77×10^{-4}	-.001451

D.2 Prospective asset-disjoint CSI300 test

This test was frozen before any independent-test value, generated sample, or endpoint was opened. Twelve CSI300 pairs share no asset with the original 15-pair benchmark. Three seeds and four arms give 36 matched Base cells and 144 arm completions. Every completion binds two reference arrays, two generated arrays, one controller-action array, and one completion record. The two independent structural auditors therefore require an exact set of 864 artifacts before any Corr endpoint is evaluated.

Under the registered $W = 60$ endpoint, the terminal run completes 36/36 Base cells and 144/144 arm evaluations with zero execution failures. The reduction contains 36 `COMPLETE_VALID` Base arms, 108 `COMPLETE_SCIENTIFIC_INFEASIBLE` candidate arm records, and no domain-invalid arm under that definition. This conservative flag means that at least one registered row used a hard identity/refusal or requested a target outside its certified reachable interval. In the latter case, Point and RPC apply a bounded boundary action; they do not refuse the row. The flag therefore does not mean that every row is identity, nor that the generated endpoint is missing or silently filtered. The pre-outcome reduction contract explicitly retains both finite statuses for the Corr endpoint and discloses every candidate arm evaluation.

All 108 candidate-minus-Base seed contrasts are favorable. guarded RPC, point-target RPC, and RPC are each favorable on 36/36 seed cells and 12/12 pair means. Only RPC is inferential: its pair-mean difference is -0.051694, its exact one-sided signed-rank p-value is 0.000244, and its 100,000-resample 95% pair-bootstrap upper bound is -0.044362. The frozen 9/12 win, signed-rank, and bootstrap conditions all pass. Point is lower on average but remains descriptive because the primary arm was fixed before outcomes.

The public reduction SHA-256 is recorded in the anonymous audit manifest. The primary auditor independently replays the endpoint, signed-rank, and bootstrap. A second auditor reimplements all

three without importing the reducer. Both pass. This is an independent asset-disjoint test of Corr repair, not evidence for optimal arm selection, Coint/ADF improvement, universal action coverage, or population-wide financial robustness.

Table 13: Prospective asset-disjoint CSI300 panel-v2: 12 pairs, three seeds, and matched Base contrasts in Corr distance (lower is better). "Reference" marks fields that do not define a candidate contrast.

Arm	Mean distance	Mean diff.	Cell wins	Pair wins	Role
Base	0.134199	0.000000	N/A	N/A	Reference
Guarded RPC	0.111679	-0.022520	36/36	12/12	Descriptive historical
point-target RPC	0.080627	-0.053572	36/36	12/12	Descriptive ablation
RPC	0.082505	-0.051694	36/36	12/12	Frozen primary

D.3 Matched-controller replay on the learned-covariance adaptation

Question and design. An end-to-end comparison can conflate generator quality with the benefit of post-sampling control. We therefore apply the same frozen RPC controller to all 45 archived learned-covariance CSDI cells: 15 financial pairs, three seeds, and 20,160 generated paths. This is a retrospective analysis of an already opened panel. learned-covariance CSDI is the matched-backbone adaptation described in the main paper, not the full official CW-Gen implementation. Base and RPC branch from identical stored arrays within each cell; original strict-training target models, observed contexts, support masks and action budgets are retained. The controller, interval adapter and shared-branch sources match the frozen T2 sources byte for byte. No retraining, regeneration, new target fitting or positive-domain translation is performed. All 45 cells are valid and all pair means are retained.

Correlation effect and endpoint sensitivity. The original Corr-family distance decreases from 0.209471 to 0.195491, a 6.67% reduction with 15/15 favorable pair means. Averaging seeds before inference, the exploratory one-sided exact signed-rank value is 3.05×10^{-5}; the pair-bootstrap 95% interval for the mean difference is $[-0.018896, -0.009731]$. These post-outcome-panel statistics are not new independent confirmation, and their interval is not multiplicity-adjusted across added endpoints. Removing windows that cross generated-row joins changes the descriptive distances to 0.186910 and 0.170293 (8.89% reduction), with 14/15 favorable pairs. SOL_DOGE worsens by 0.001877 under this latter definition. Its real reference remains the original continuous series; this check does not replace the benchmark or retrospectively select its endpoint.

Table 14: Historical same-controller context, 15-pair means after seed averaging. All entries are distances (lower is better). Each Base/RPC pair uses identical draws internally. Rows from different generator runs are *descriptive*: random streams, preprocessing and some floating-point paths differ. This is not an expanded formal ranking or a causal test of whitening.

Run / arm	ADF	Coint	Corr
Learned-cov. Base	19.600562	16.655490	.209471
Learned-cov. + RPC	19.600562	16.709534	.195491
Historical CovPW-T2 Base	16.682694	13.107675	.128860
Historical CovPW-T2 + RPC	16.682694	13.281855	.109075

Historical CovPW-T2+RPC has a lower macro distance in all three families than JMCE+RPC in Table 14; its Corr is lower on 12/15 pairs. BTC_ETH, COKE_KO and SOL_DOGE favor JMCE+RPC. These parallel results do not identify a generator–controller interaction or establish superiority over official CW-Gen. The closer conclusion is that the controller's Corr benefit is not exclusive to the CovPW construction. The failed GRU transfer remains unchanged (Appendix D.8).

Full quality audit. We evaluate 24,879,060 length-60 ADF/Coint windows across BASE and INTERVAL. ADF decisions have zero flips and their family distances are identical. Coint distance

increases by 0.324%, with seven favorable and eight adverse pair differences. The worst degradation is 4.221916 on MA_V, exceeding the borrowed historical tolerance 1.340335. The descriptive guard fails even though its bootstrap mean upper bound (0.778299) is below that tolerance. This borrowed threshold is not a newly justified JMCE noninferiority margin. Neither an unchanged protected spread nor a confidence interval containing zero establishes refitted-Coint equivalence.

Table 15: Learned-covariance marginal-quality costs, evaluated channelwise from float64 within-row log returns. Lower distances are better. Quantile, std and tail distances are multiplied by 10^3; Pearson kurtosis is not excess kurtosis. Improved/adverse counts compare pair means.

Endpoint	Base	+RPC	Change	Improved/adverse
101-quantile	1.038202	1.045491	$+0.70\%$	8/7
Std	.707385	.729330	$+3.10\%$	6/9
Kurtosis	47.706445	47.657366	-0.10%	8/7
1%/99% tail	1.414470	1.454128	$+2.80\%$	6/9
Absolute-return ACF	.084538	.084515	-0.03%	9/6

The added 101-quantile endpoint is not exact Wasserstein distance. Marginal std, kurtosis and tail errors are computed separately for each channel and then averaged; the ACF endpoint uses lags 1–5. These added definitions must not be pooled with differently defined earlier quality tables. The modest ACF/kurtosis improvements do not offset variance/tail deterioration into a joint no-harm claim. The maximum log-spread replay error is 2.22×10^{-16}, supporting the action certificate at numerical precision.

Numerical provenance. A batched evaluator preserves the original AIC lag selection and MacKinnon p-value definitions, with two windows falling back to the scalar implementation. Uniform and near-threshold scalar checks cover 4,320 windows; maximum p-value discrepancy is 8.91×10^{-10} and decisions agree. An independent implementation verifies 2,264 quality checks. An initial marginal audit detected float32 reference arithmetic; every marginal cell was recomputed using float64, retaining the failed audit and prior output. No tolerance was relaxed, and all ADF/Coint p-value hashes stayed unchanged. The old separate-run ADF/Coint components differ from this float64 replay by at most 0.004288, so neither this replay nor Table 14 overwrites the historical main table. Compact aggregate and pair-level results are provided with source bindings; raw financial arrays remain outside the anonymous bundle.

D.4 Frozen benchmark and broader descriptive comparison

The joined view combines the ledger-frozen diffusion neighborhood with the nonoverlapping rows of the canonical broad screen. Both blocks use the same chronological test split, 15 pairs, and nine diagnostic distances. Every cell is measured against held-out real test windows. The join is performed at pair level, not by merging published aggregate ranks. For CSDI/isotropic, CSDI-prewhiten, and CTSDiff-SPRC, the final ledger-frozen three-seed rows replace the earlier broad-screen rows. The resulting matrix contains exactly $15 \times 16 = 240$ unique pair–method cells. Family rank, metric-balanced rank, and normalized score are recomputed over all 16 methods. Only the ledger block is jointly frozen and averaged over three seeds. The broad-screen rows retain their source-block replication and training protocols. The common held-out evaluator therefore aligns test pairs and diagnostics, but does not create matched seed depth or training budget across all 16 methods. The joined rank is descriptive for this reason as well as its source-precedence sensitivity.

Under the declared ledger-frozen-row precedence, guarded RPC (ledger label CTSDiff-RPC) ranks first under all three joined aggregation rules. Its mean family rank is 4.500, metric-balanced rank is 5.796, and normalized score is 0.146. It has the strongest Corr family rank and ties CSDI-CovPW for the strongest Coint rank; original-coordinate CTSDiff-Corr retains the strongest ADF rank. Because these rows use different context estimators and the ledger block uses separate diffusion draws, this decomposition is descriptive rather than a causal attribution of each family difference. The joined result is an actual out-of-sample empirical benchmark. Its global rank was not frozen as one 16-method confirmatory hypothesis because the broad and ledger-frozen blocks overlap. Source precedence is therefore audited explicitly. Under the main policy, the final matched three-seed ledger-frozen rows define the three overlaps. If the earlier broad-screen rows are retained instead,

Table 16: Complete out-of-sample comparison over the same 15 test pairs and nine diagnostics (16 methods; all 240 pair–method cells present). For duplicate methods, the final ledger-frozen three-seed row replaces the earlier broad-screen row; nonoverlapping GAN, VAE, diffusion, and forecasting rows come from the canonical broad screen and retain its replication/training protocols. Ranks are recomputed jointly over the displayed methods, but the join is not a matched-budget comparison. F-rank ranks the mean distance within each family; M-rank first ranks each component and is the unit-robust companion. Lower is better. The daggered guarded RPC row has the historical ledger label CTSDiff-RPC; CSDI-prewhiten is the immutable artifact label for CSDI-CovPW. Predeclared paired inference remains in the ledger-frozen comparison.

Method	Source	F-rank	M-rank	Norm.	ADF	Coint.	Corr.
guarded RPC† (ours)	Frozen	**4.500**	**5.796**	**0.146**	5.033	**4.200**	**4.267**
CSDI-prewhiten	Frozen	4.989	6.233	0.152	5.033	**4.200**	5.733
CTSDiff-SPRC	Frozen	6.644	6.389	0.157	6.067	6.667	7.200
CTSDiff	Broad	6.956	7.681	0.234	**4.867**	4.267	11.733
CSDI-JMCE-CW	Frozen	7.067	7.011	0.184	6.667	6.667	7.867
CSDI-isotropic	Frozen	7.289	6.963	0.190	6.467	6.533	8.867
Non-stationary Transformer	Broad	7.578	7.459	0.218	8.133	7.933	6.667
TimeVAE	Broad	7.889	7.963	0.208	7.333	8.000	8.333
TimeXer	Broad	8.756	9.022	0.234	9.067	9.333	7.867
CTSDiff-diag	Broad	8.756	8.337	0.317	5.133	7.133	14.000
TimeMixer	Broad	8.778	8.496	0.210	9.733	10.400	6.200
Autoformer	Broad	9.111	9.100	0.288	10.667	8.933	7.733
TimeGAN	Broad	10.556	10.504	0.437	9.600	10.400	11.667
iTransformer	Broad	11.022	11.059	0.500	13.800	14.133	5.133
Diffusion-TS	Broad	12.889	11.578	0.474	14.133	12.933	11.600
TimesNet	Broad	13.222	12.407	0.603	14.267	14.267	11.133

CSDI-CovPW ranks first at 4.600 and guarded RPC ranks second at 4.711. This sensitivity limits an unqualified all-method superiority claim; the predeclared paired superiority and noninferiority tests remain in the ledger-frozen five-method experiment.

Before candidate generation and reporting, the source ledger `a5efe4cee11eed58ab4137a34e5f1a57b2109bbf45accfe7f3971b5b1c749af1` froze the five methods, 15 pairs, endpoints, noninferiority margin, one-sided alternatives, Holm family, and PCG64 bootstrap seed. The finalization ledger `de7f0c34b029f9010bb7378b24f95ae58c593260a7f538c495a6e69791a6cccc` performed no statistical recomputation. For Coint, $d_i = D_i^{\mathrm{RPC}} - D_i^{\mathrm{PW}}$. The validation-frozen rule is $\Delta_{\mathrm{NI}} = \max\{0.01, 0.05 D^{\mathrm{PW}}_{\mathrm{Coint,val}}\}$: at most 5% degradation relative to the fixed comparator, with a 0.01 floor. It yields $\Delta_{\mathrm{NI}} = 1.340335$ before test evaluation. The exact one-sided Wilcoxon test is applied to $d_i - \Delta_{\mathrm{NI}}$, and its raw p-value enters the same three-family Holm correction as Corr and ADF. The later 16-method source precedence was not part of this ledger. For backward compatibility, legacy source columns ending in `family_distance` are rank-valued by schema; the join builder validates them against the frozen pair-level ranking table.

The closest learned-covariance comparator is also audited without collapsing its outcomes into a single rank. Table 17 reports all nine common-support hypotheses for CSDI-JMCE-CW. This row adapts the joint mean–covariance mechanism to the matched CSDI backbone; it is not a reproduction of the full CW-Gen wrapper. None of the nine comparisons survives family-wide correction, so it supplies a substantive mechanism baseline rather than evidence of either method uniformly dominating the other.

The sensitivity table is strictly post hoc. It shows that the conclusion does not hold at margins 0.50 or below and therefore must be reported as passing the specific validation-frozen margin, not as generic cointegration preservation. The locally retained bootstrap upper bound is rounded, so the strict decision at the exact rounded boundary 0.518475 is indeterminate and is not tabulated.

Table 17: Strict common-support inference for CSDI-JMCE-CW, an *adapted matched-backbone mechanism baseline*, not a reproduction of the official CW-Gen wrapper. Three seeds are averaged within each of 15 pairs. Differences are CSDI-JMCE-CW minus comparator rank-family distance; negative values favor JMCE. Tests are exact two-sided Wilcoxon tests, with Holm correction over all nine hypotheses and 100,000-replicate pair-bootstrap confidence intervals. No hypothesis survives correction, and every interval includes or touches zero.

Comparator	Family	JMCE W/L/T	Mean diff.	Exact p	Holm p	95% mean CI
CSDI-CovPW	ADF	9/5/1	-0.0889	0.6353	1.0000	[-0.6222, 0.4889]
CTSDiff-SPRC	ADF	7/7/1	0.3444	0.3989	1.0000	[-0.4556, 1.1333]
CSDI-Isotropic	ADF	6/9/0	0.2333	0.4124	1.0000	[-0.2667, 0.7111]
CSDI-CovPW	Coint	5/9/1	0.5889	0.0704	0.6339	[0.0000, 1.1444]
CTSDiff-SPRC	Coint	6/9/0	0.3444	0.6476	1.0000	[-0.6889, 1.3667]
CSDI-Isotropic	Coint	6/9/0	0.3111	0.2816	1.0000	[-0.2222, 0.8444]
CSDI-CovPW	Corr	4/9/2	0.4667	0.2371	1.0000	[-0.0889, 1.0000]
CTSDiff-SPRC	Corr	7/6/2	-0.4667	0.1956	1.0000	[-1.2000, 0.2889]
CSDI-Isotropic	Corr	8/7/0	-0.3111	0.2747	1.0000	[-0.8444, 0.2000]

Table 18: Post-hoc descriptive sensitivity of the Coint noninferiority conclusion to margin Δ. "Pass" requires both Holm-adjusted Wilcoxon $p < 0.05$ and the rounded frozen bootstrap upper bound $0.518475 < \Delta$. Only $\Delta = 1.340335$ was validation-frozen; this table is not a new preregistered test.

Margin Δ	Holm p	Decision
0.00	0.7003	fail
0.25	0.3024	fail
0.50	0.0946	fail
0.60	0.0256	pass
0.75	0.0181	pass
1.00	0.000854	pass
1.340335	0.000183	pass (frozen)

Table 19: Predeclared guarded RPC minus CSDI-CovPW comparison. Negative favors RPC; ADF/Coint tests use prespecified margins.

Family	W/L/T	Mean diff.	Median diff.	Holm p	Boot. upper	Decision
Corr	15/0/0	-0.00791365	-0.00568415	0.00009155	-0.00504202	superiority
ADF	6/6/3	-0.00745039	0	0.151184	9.6984×10^{-6}	10^{-6} gate fails
Coint	9/6/0	+0.278415	-0.00114362	0.0001831	0.518475	passes frozen $\Delta = 1.3403$

Table 20: Outcome-frozen same-endpoint reverse-step comparison on 12 CSI300 pairs and three seeds per pair. Lower Corr-family distance is better. Each difference is candidate minus the named comparator, so negative is favorable; positive relative reduction is favorable. Holm adjusts the two co-primary exact one-sided Wilcoxon tests; $U_{.975}$ is the simultaneous 100,000-resample pair-bootstrap upper bound.

Arm / contrast	Mean D	Δ	Rel. red.	Fav.	p_{H}	$U_{.975}$
Base reference	.134199	—	—	—	—	—
PDM–CSDI vs Base	.082197	-.052002	38.75%	12/12	.000488	-.042528
RPC-Interval vs PDM	.082505	+.000308	-0.37%	5/12	.788086	+.001089

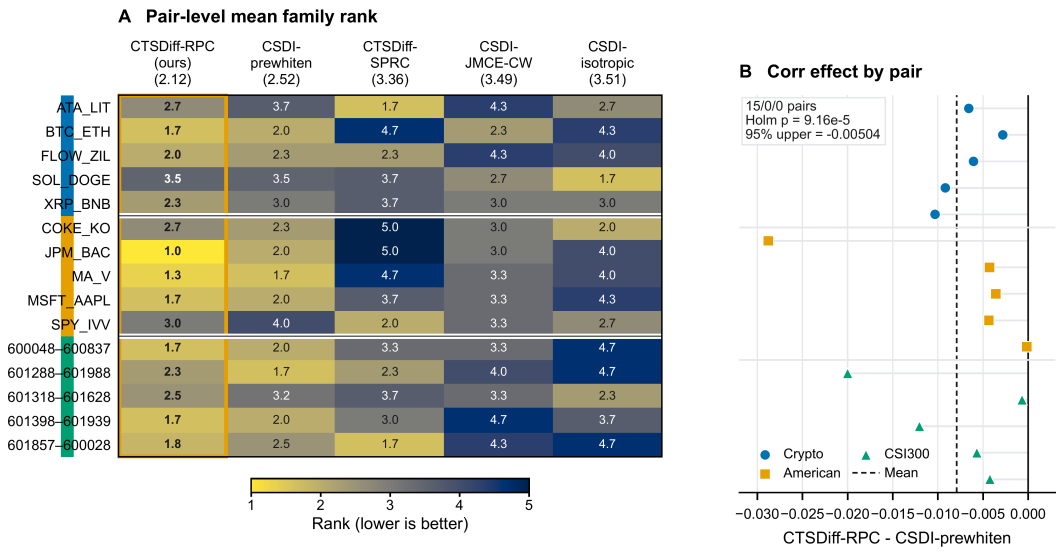

Figure 3: Frozen five-method evidence over 15 pairs. (A) Pair-level mean family ranks; values above the columns are mean ranks across pairs. Lower ranks are better. (B) Pair-level Corr differences between historical guarded RPC (labeled CTSDiff-RPC) and CSDI-CovPW (labeled CSDI-prewhiten). Negative differences favor the controller. The dashed line marks the mean difference.

D.5 Reverse-step projection comparator

This study asks a different question from the CPS qualification gate below: can a reverse-step projector improve the same Corr endpoint when it is actually run on the prospective CSI300 panel? The three arms are authenticated Base, a PDM-semantics-backed CSDI algorithmic port, and the original RPC realization. All use the same 12 pairs and three seeds. Base and PDM share the checkpoint, observed context, initial Gaussian tensor, and every reverse-noise tensor. Each performs 50 denoiser evaluations per row. PDM additionally projects the complete reverse state to the closest point on a Base-reference-conditioned feasible curve after every transition. RPC is the deterministic one-shot transform of the authenticated Base output. Thus denoiser access is matched, but total computation is not.

What the port retains and changes. The retained PDM ingredient is projection after each reverse transition. The task-specific ingredient is the Base-conditioned feasible curve, constructed from the protected relation and edit budgets. This tests intervention location on the present task. It does not reproduce the published method's original task constraints or empirical ranking.

The strict-training target interval, projection curve, action and paired-energy contract, endpoint, contrast direction, and inference rules were frozen before PDM outcomes were opened. The original prospective RPC outcome was already known, so this is outcome-frozen post-outcome evidence rather than a new prospective confirmation. Two co-primary contrasts use candidate-minus- comparator differences, exact one-sided Wilcoxon tests, Holm correction, and simultaneous one-sided 97.5% bounds from 100,000 pair bootstraps. Table 20 reports both contrasts.

All 36 cells and 108 arm endpoints are valid, with no execution failure. PDM–CSDI reduces macro Corr distance from 0.134199 to 0.082197 and passes its Base contrast on all 12 pairs. RPC has macro distance 0.082505. Its frozen contrast against PDM is $+0.000308$, where negative favors RPC; only 5/12 pairs are favorable. The RPC-over-PDM gate and the joint conjunction therefore do not pass. This is not equivalence or noninferiority evidence.

The PDM action audit covers 17,007 row-instances: 13,808 are active and 3,199 remain identity decisions, with no numerical abstention. The maximum curve absolute residual is 2.511×10^{-7}, the maximum curve relative residual is 2.147×10^{-7}, and the maximum log-spread residual is 9.612×10^{-7}. An independently implemented semantic audit reconstructs the effective context, all endpoints, contrasts, tests, bootstraps, action accounting, and artifact inventory, and exactly repro-

duces the primary reduction. The port follows the published PDM intervention locus and projection semantics, but it is not an official-package implementation or checkpoint reproduction.

D.6 Five-Arm Matched-Action/Energy Follow-Up

Question and matching. The earlier PDM comparison and budget-sensitivity replay motivate a direct question: does additional admissible energy improve RPC more than reverse-step projection? The V3.1 follow-up fixes five arms—Base, PDM E10, RPC E10, PDM E20, and RPC E20—on a separately selected 12-pair CSI300 reserve panel with three seeds per pair. Its 24 assets are disjoint from the prospective panel and earlier PDM comparison; the market and 2021–2022 period are shared. The frozen V3.1 selector inherits the superseded V3 reserve pairs, not the prospective panel. New checkpoints use only these pairs' strict-training rows. E10/E20 denote paired-energy caps 0.10/0.20; all controlled arms use action cap 0.20. Checkpoints and target predictors are trained only on the declared strict-training inputs. Each cell samples a single Base; both RPC arms edit that Base, and both PDM arms replay its initial and reverse-noise tape for 50 denoiser evaluations. Context, target interval, support, endpoint and averaging units are matched. PDM adds projection inside the reverse loop, so this is action/energy matching, not equal total computation.

Amendment and retained failures. The completed study is the approved AM03 eligibility amendment, not an unamended independent replication. Its predecessor stopped after 18/36 cells when a second PDM-only context-floor adapter was non-identity. The failure trace did not establish whether the affected final prices were nonpositive. AM03 permits finite positive values below the preparation floor; removing the second translation changes execution eligibility and is disclosed as such. The shared Base retains its original one-time whole-cell floor adaptation. Actual final RPC and PDM outputs must both be finite and strictly positive, without a second translation, clipping or epsilon repair. Their accepted values are retained unchanged. The failed predecessor and an earlier 0/36-cell interface failure are preserved, never resumed or merged. The new run restarts all 36 cells; no successful cells from the failed runs are selected for reuse. Budgets, seeds, targets, metrics, primary contrast and decision thresholds remain fixed.

Complete reporting and acceptance scope. All 36 cells and 180 arm endpoints complete with zero execution failures. Corr and each of the six descriptive quality families have 180/180 valid arm-cell endpoints. No invalid endpoint is dropped or replaced. One frozen terminal auditor verifies the exact source/run inventory, independently recomputes Corr from the actual arrays and authenticated reference cache, and replays both deterministic RPC arms and the four marginal/tail/lag diagnostics. Legacy ADF/Coint p-value vectors are not independently regenerated: their checks cover source and array binding, window accounting, native summaries, and distance arithmetic. Each legacy family records 4,073,076 expected and finite windows per arm/role across all cells, with zero undefined windows. Repeated windows or reference computations are not independent inferential units. The anonymous supplement preserves every pair/seed endpoint and contrast and binds the path-bearing original records by SHA256; raw records and earlier failures remain at their research sources.

Sole primary and serial gate. The sole primary is RPC E20 minus PDM E20, after averaging three seeds within each pair. Success requires complete validity, a negative 12-pair mean, at least nine favorable pairs, exact one-sided Wilcoxon $p \leq 0.05$, and a negative one-sided 95% upper bound from the frozen 100,000 pair bootstraps. The observed mean is -0.001975581, with eight favorable and four adverse pairs; $p = 0.088134765625$, and the upper bound is -0.000166460. The favorable-count and Wilcoxon conditions fail. A negative bootstrap upper bound for the mean does not override those failures: the mean bootstrap and signed-rank test use different statistics and rejection rules. Table 21 retains all primary pair directions.

The gate therefore restricts PDM E20 versus Base to descriptive reporting, despite its 12/12 favorable pairs. Fixed support contrasts show that both methods improve on every pair from E10 to E20 (Table 22). This is evidence of budget headroom in the declared two-cap comparison, not a license to select another budget from outcomes or claim an optimum. The original prospective CSI300 test and earlier PDM result remain distinct studies with their own frozen primaries.

Quality tradeoffs. The four added diagnostics are defined on within-row log returns. Std and raw kurtosis (not excess kurtosis) are averaged over channels and rows; their distances are abso-

Table 21: Complete primary pair differences in the five-arm follow-up. Each entry averages three seeds; negative favors RPC E20 over PDM E20.

Pair	RPC E20 minus PDM E20
sh600779_sz000413	-.007671581
sz002311_sz300002	+.000728705
sh600123_sh600569	-.004918442
sh601918_sz000631	-.000514783
sz000983_sz002128	+.001686653
sh600297_sh600639	-.003144844
sh600169_sh601866	-.004973806
sh600332_sh600848	-.004611447
sh600006_sz000883	+.005077768
sz000536_sz000630	-.006523509
sh600271_sh600703	+.002343962
sh600307_sh600997	-.001185647

Table 22: Fixed key-secondary and support contrasts. Negative differences are favorable; F/E/A are favorable/equal/adverse pair counts out of 12. All entries are descriptive; no secondary test is performed after primary failure.

Contrast	Macro Corr difference	F/E/A
PDM E20 minus Base	-.067188472	12/0/0
RPC E10 minus PDM E10	+.000040723	6/0/6
RPC E20 minus RPC E10	-.018545642	12/0/0
PDM E20 minus PDM E10	-.016529338	12/0/0

lute generated-minus-reference aggregate differences. Tail coexceedance uses strict simultaneous exceedance of the two reference absolute-return 90th percentiles, with linear type-7 quantiles. Lag distance averages absolute differences between row-mean cross-correlations at every lag from -6 to $+6$. Zero-variance undefined quantities invalidate their required endpoint rather than triggering finite-only filtering. Legacy ADF/Coint separately retain their original flattened, potentially cross-row $W = 60$ semantics: ADF on the first/second price ratio and Coint on the ordered second/first level pair. Their pass-ratio and mean/maximum run-length distances are not the boundary-safe primary Corr endpoint.

Tables 23 and 24 retain every fixed quality contrast, including adverse directions and exact ties. These distances have different native scales and are not combined into a score. At E20, RPC has lower macro std, kurtosis, tail-coexceedance and lag-profile distances than PDM. Relative to Base, however, RPC worsens kurtosis on 11/12 pairs and lag distance on 7/12. Coint splits six favorable/six adverse with an adverse mean against both Base and PDM. RPC's exact ADF equality does not certify the refitted Coint endpoint. These are descriptive tradeoffs, not an inferential superiority claim or a reason to suppress an endpoint.

Table 23: All fixed E10 quality contrasts. Negative macro differences are favorable. F/E/A counts are pairs out of 12; all 36 cells per arm are valid. Metrics retain their native units; all results are descriptive.

	PDM E10 minus Base		RPC E10 minus Base		RPC E10 minus PDM E10	
Metric	Mean	F/E/A	Mean	F/E/A	Mean	F/E/A
Std	-2.0501×10^{-7}	5/0/7	-3.0765×10^{-6}	6/0/6	-2.8715×10^{-6}	11/0/1
Kurtosis	+.010338	3/0/9	+.008314	4/0/8	-.002024	12/0/0
Tail	-.003500	12/0/0	-.003554	12/0/0	-.000054	10/0/2
Lag	+.000404	5/0/7	+.000217	5/0/7	-.000186	8/0/4
ADF	-1.0387×10^{-5}	2/7/3	0	0/12/0	$+1.0387\times10^{-5}$	3/7/2
Coint	+.127001	7/0/5	+.155010	7/0/5	+.028009	7/0/5

Table 24: All fixed E20 quality contrasts, with the same definitions as Table 23. RPC improves tail distance against Base but does not uniformly improve kurtosis, lag or Coint.

	PDM E20 minus Base		RPC E20 minus Base		RPC E20 minus PDM E20	
Metric	Mean	F/E/A	Mean	F/E/A	Mean	F/E/A
Std	-1.4161×10^{-5}	6/0/6	-2.4127×10^{-5}	9/0/3	-9.9656×10^{-6}	11/0/1
Kurtosis	+.034778	1/0/11	+.025381	1/0/11	-.009397	11/0/1
Tail	-.005722	12/0/0	-.005958	12/0/0	-.000236	12/0/0
Lag	+.001465	4/0/8	+.000657	5/0/7	-.000808	9/0/3
ADF	$+1.5300\times10^{-5}$	0/8/4	0	0/12/0	-1.5300×10^{-5}	4/8/0
Coint	+.219059	6/0/6	+.284696	6/0/6	+.065637	6/0/6

Table 25: Descriptive stage timing in the matched-budget follow-up. RPC and PDM rows have different scopes; see text.

Recorded stage	Median seconds per cell
Base generation/native batch materialization	45.621
Base whole-cell domain postprocess	4.062
PDM E10 whole-arm materialization	174.913
RPC E10 one-shot postprocess materialization	6.353
PDM E20 whole-arm materialization	240.871
RPC E20 one-shot postprocess materialization	6.524
Shared post-reference reporting and promotion	862.723

Instrumented timing. Table 25 reports median accepted stage times over 36 cells. RPC timing measures incremental postprocessing of the already sampled Base, including materialization/checking; it is not a full generation time. PDM whole-arm time includes reverse sampling, projection and materialization. The shared reporting/promotion stage includes legacy ADF/Coint evaluation. These are implementation-stage measurements, not isolated GPU kernels. Nested stages are not summed as disjoint costs. The historical guarded policy overhead of 1.07% is a different implementation and measurement scope, not the V3.1 incremental ratio. We infer neither equal-compute efficacy nor an end-to-end speedup from the ratio of RPC and PDM stage medians.

D.7 Stored-Sample Quality and Real–Real Temporal References

Status and complete coverage. This is a post-hoc diagnostic of the accepted AM03 arrays, not a new confirmatory test, policy search, or rescue of its failed primary. We fix the added metrics and temporal sampling design before executing this replay, authenticate every read artifact against the original run manifest, and include all 12 pairs, three seeds and five arms. No model is retrained and no generated value is edited. All 180 arm-cell records are valid for the new metrics. An independent centered-window implementation reproduces the accepted Corr distances to a maximum absolute discrepancy of 5.6×10^{-17}. Seeds are averaged within pair and pairs equally; overlapping windows, seeds and temporal phase combinations are not independent inferential units. No additional p-value, noninferiority margin or joint quality score is used.

Channelwise quality definitions. The original AM02 std and kurtosis distances take an absolute difference after averaging channels, which can cancel opposing channel discrepancies. The added version first takes the absolute generated–reference difference in each channel's mean row std or raw kurtosis, then averages the two nonnegative distances. Both definitions are retained. We additionally compare each channel's pooled within-row log returns using one-dimensional Wasserstein-1 and a 99-point quantile distance at $q = 0.01, \ldots, 0.99$. Equal-size empirical Wasserstein-1 is the mean absolute difference between sorted samples. Absolute-return ACF compares row-mean correlations separately at lags $1, 5, 10, 20$, then averages absolute discrepancies over channels and lags. A required undefined row invalidates that component rather than being discarded.

RPC E10 improves Corr on 12/12 pairs but worsens the added quantile distance on 8/12 and channelwise std on 7/12. Its W1 is split 6/6 with a small adverse mean. RPC E20 improves W1 and quantile distances on 8/12, while channelwise std is split 6/6 and kurtosis worsens on 11/12. Absolute-return

Table 26: Post-hoc channelwise quality on all stored AM03 draws. Lower is better. W1, quantile and std distances are multiplied by 10^3. Kurtosis is raw, not excess; ACF is for absolute returns.

Arm	W1	Quantile	Std	Kurtosis	ACF
Base	.965976	.868690	.651676	6.244903	.051390
PDM E10	.969295	.872587	.662405	6.255241	.051139
RPC E10	.967658	.871098	.660281	6.253217	.051180
PDM E20	.961677	.866725	.658265	6.279681	.050771
RPC E20	.955725	.861332	.651027	6.270284	.050927

Table 27: Matched-size temporal reference: absolute Corr-mean, Corr-std, 101-quantile distances and their equally weighted family distance. Model rows use generated versus real late-period blocks; the reference row uses real early versus real late blocks. All entries are descriptive.

Comparison	Mean	Std	Quantile	Family
Real–real	.096666	.030924	.100621	.076071
Base–real	.092002	.186419	.184276	.154233
PDM E10–real	.068037	.123236	.124383	.105219
RPC E10–real	.068296	.123941	.125016	.105751
PDM E20–real	.063882	.095289	.102975	.087382
RPC E20–real	.062073	.096203	.101369	.086548

ACF improves on 11/12 pairs at E10 and 12/12 at E20. Against matched-budget PDM, RPC has lower W1 on 12/12 at both caps but higher absolute-return ACF distance on 12/12 at both caps. Thus neither controller dominates quality; the ACF diagnostic is different from the previously reported cross-series lag profile.

Editing and conditional dispersion. Across 16,980 rows per RPC cap, 3,273 are identity (including 27 unsupported rows). E10 enters the target interval on 3,263 rows and is unreachable on 10,444; E20 enters on 5,934 and is unreachable on 7,773. Identity includes 3,246 already-in-interval rows. Direct log-spread editing error is at most 4.45×10^{-16}. A three-draw diagnostic computes pairwise log-return RMS distance between seeds at the same context, averaging over rows, seed pairs and channels. The mean across asset pairs of the ratio to Base is 0.9991 at E10 and 0.9927 at E20 (ranges 0.9866–1.0100 and 0.9715–1.0087 across pairs). These measurements describe conditional draw dispersion; three draws do not establish full distributional diversity or noninferiority.

Nonoverlapping temporal sampling. Each 240-point CSI300 prediction covers five trading days, while native origins advance 48 five-minute observations. Consequently adjacent reference rows overlap. For each pair with N rows, we use the earliest and latest $\lfloor N/3 \rfloor$ rows as separate candidate periods, leaving the middle third as a temporal gap. Each group contains 30 rows selected every fifth origin, with phase $p \in \{0, \ldots, 4\}$: early indices are $p + 5j$, late indices $N - \lfloor N/3 \rfloor + p + 5j$, $j = 0, \ldots, 29$. The authenticated origin stride verifies no shared future observations within either group or across the two groups. We retain all 25 early/late phase combinations; different combinations overlap and are not independent repetitions. The supplement reports their full values and min/max envelopes, not confidence intervals.

For each combination, real–real compares the early real group with the late real group. Model–real compares the generated late group with exactly that late real group, averaging three seeds. All comparisons use 30 rows, 240 points and the original 60-return rolling-Corr definition, with 5,400 windows per group. Table 27 averages phases within pair and then pairs. This measures temporal variation jointly with finite-sample effects; it is not an irreducible sampling-noise floor.

The ratio of macro family distances to the temporal real–real reference decreases from 2.027 for Base to 1.390 for RPC E10 and 1.138 for RPC E20. However, RPC E20's Corr-std discrepancy remains 0.096203 versus the real–real value 0.030924: a favorable aggregate cannot establish that the remaining error is at a noise floor. These temporal-subset results are distinct from the full-population AM03 endpoint, and their different reference periods prevent interpreting the ratio as

a calibrated measure of distributional fidelity. Larger independently evaluated panels, additional generator families and downstream validation remain necessary.

D.8 Frozen-policy transfer to an independent autoregressive generator

Purpose and scope. We test whether the frozen target policy transfers beyond CSDI, retaining the already-open AM03 panel of 12 asset-disjoint pairs and three seeds. This is a new generator experiment, not an untouched external replication. The custom GRU–Student-t AR is a one-layer, width-32 recurrent network with two standardized log-return inputs and five outputs: two locations, two log scales clipped to $[-7, 3]$, and $0.95\tanh(\cdot)$ correlation. Its bivariate Student-t emission uses five degrees of freedom and one shared chi-square scale across channels. It is neither an official DeepAR reproduction nor a state-of-the-art baseline. Per seed, 2,048 unique chronologically evenly spaced rows are selected by timestamp from the authenticated 10,000-row 2010–2019 training cache. Training uses Adam (10^{-3}), batch size 64, ten epochs, gradient norm cap 5, and the final checkpoint, with training-only normalization and no early stopping. All 36 models run on local MPS; generation uses explicit independent validation/test noise streams and unclipped price reconstruction. This limited training budget prevents equal-compute generator rankings.

Four shared-input arms. BASE is the new generator's unchanged sample. FIXED selects one scalar request per trained model on the complete 2020 validation set, from $\{0.80, 0.82, \ldots, 1.20\}$, by original Corr distance; ties prefer the smallest absolute action, then the smaller request. The request is projected into each row's certified budget interval. POINT and INTERVAL use the unchanged strict-train target predictor, support rule and 0.5 OOF-RMSE Fisher-z interval. All controlled arms share action cap 0.20, energy cap 0.10, and unsupported identity behavior. Test contexts reproduce the original target tape to 1.12×10^{-15}. All four arrays are frozen before their 2021–2022 future references are read. No policy is refitted after these outcomes. Validation NLL beats a training-fitted constant bivariate Student-t on 36/36 cells (mean 2.12334 versus 2.39803); this sanity check does not establish state-of-the-art model quality.

Prespecified narrow transfer test fails. Average the three seeds within each pair and then the 12 pairs equally. The pre-execution gate requires valid complete cells, a negative INTERVAL–BASE Corr difference, at least 9/12 favorable pairs, and one-sided exact signed-rank $p \leq .05$. All 36 cells and 144 arm records are valid, but Corr worsens from 0.054600 to 0.060523, an absolute increase of 0.005923 (10.85%), with only 1/12 favorable pairs and $p = 0.995361$. The gate fails. This is a test for improvement, not a new confirmatory test of harm. The original RPC–PDM superiority failure is unchanged.

Table 28: Custom GRU extension, 12-pair means after seed averaging. All columns are distances to real data (lower is better). W1, quantile and std are multiplied by 10^3; marginal distances are channelwise.

Arm	Corr	W1	Quantile	Std	Kurtosis	Abs. ACF
BASE	.054600	.727825	.653286	.866592	6.095877	.037321
FIXED	.044933	.760320	.683007	.944841	6.098519	.037187
POINT	.065835	.723159	.647077	.859011	6.148269	.037390
INTERVAL	.060523	.725686	.649989	.864413	6.130187	.037373

Secondary comparisons and quality costs. INTERVAL has lower Corr than POINT on 10/12 pairs but lower Corr than FIXED on only 4/12. FIXED lowers macro Corr by 17.71%, yet improves only 6/12 pairs and worsens W1 and quantile distance on 10/12 and std on 9/12. INTERVAL improves W1 on 8/12, quantiles on 9/12 and std on 7/12, but worsens kurtosis on all 12 and absolute-return ACF on 9/12. Its tail-coexceedance distance is 0.008611 versus BASE's 0.008628, with 6/12 favorable pairs; cross-lag distance is 0.013035 versus 0.012777, with 4/12 favorable pairs. These are descriptive comparisons, not additional significance tests or quality noninferiority claims.

Certificate remains valid; distributional improvement does not. Independent array reconstruction checks all 108 controlled arm records; the maximum log-spread discrepancy is 3.34×10^{-16}

and actual-array energy changes obey the 0.10 cap up to floating-point error. INTERVAL leaves 9,975/16,980 rows unchanged (27 unsupported), enters its interval on 4,396 rows, and cannot reach it on 2,609. Mean absolute action is 0.02200 versus POINT's 0.06109. The mean across pairs of cross-seed return-RMS ratios to BASE is 0.99992 for INTERVAL, 1.00157 for POINT and 0.98171 for FIXED. Seeds use separately trained models, so this diagnostic combines training and sampling variation; it does not certify diversity.

A post-outcome decomposition finds Corr-mean/std/quantile distances of $(.058033, .034729, .071037)$ for BASE and $(.058105, .047630, .075834)$ for INTERVAL. About 73% of the aggregate deterioration comes from the std component: mean generated rolling-Corr std falls from .182715 to .169815 against .217444 for the reference. This is consistent with excessive concentration, but does not identify its causal source. A train-fitted target, valid support and smaller edits do not guarantee fidelity of the rolling-Corr distribution for a different generator. Generator-specific validation and an identity option warrant future study on separately reserved outcomes; they are not retrospectively substituted into this failed experiment. Independent recomputation of Corr, W1, std and kurtosis agrees within 1.12×10^{-16} overall.

D.9 Target predictor sensitivity

The prospective controller used a train-only linear Ridge predictor. To test whether the saved-draw CSI300 result is unique to that predictor class, we change one factor to deterministic additive linear-spline Ridge. For each of the same five observed-prefix features, the design adds positive-part hinges at strict-training quantiles .25, .50, and .75. It uses no interactions, new features, random initialization, validation values, or test values. Chronological folds, the regularization grid and one-standard-error rule, model-specific support and out-of-fold uncertainty calibration, target interval, solver, action and energy budgets, saved Base draws, and the $W = 60$ endpoint remain fixed.

Fit occurs from authenticated 2010–2019 strict-training caches. A separate precommit binds the complete fitted model and its inventory before replay opens any saved draw. Replay first reproduces the prospective Ridge bundle and all 36 Base and Ridge endpoints, then applies the spline policy to the same Base arrays. No new generation or GPU work is used.

Table 29: Post-outcome saved-draw predictor-class sensitivity on the 12 CSI300 pairs. The strict-training additive-spline arm changes only the predictor class. Negative differences favor the spline arm. The exact p is two-sided and the reported bound is the one-sided 95% pair-bootstrap upper bound. The Spline–Ridge row is descriptive and has no superiority gate.

Contrast	Macro Δ	Fav.	p_{2s}	$U_{.95}$
Spline-Interval vs Base	-.051295	12/12	.000488	-.043431
Spline-Interval vs Ridge-Interval	+.000399	5/12	—	—

All 36 spline endpoints are valid. Spline-Interval has macro Corr distance 0.082904 and improves Base by 0.051295, or 38.22%, on all 12 pairs. The predeclared corroboration gate passes. The descriptive Spline-minus-Ridge macro difference is $+0.000399$: 5/12 pairs favor Spline and 7/12 favor Ridge. An independent, failure-aware audit reimplements the feature pipeline, controller, endpoints, exact signed-rank test, bootstrap, contrast direction, and frozen gate; it reproduces the reported reduction. The result excludes the narrow explanation that the saved-draw gain occurs only with the linear Ridge under this one controlled nonlinear alternative. It does not establish predictor superiority, universal predictor-class robustness, regime robustness, or independent confirmation. The same five non-purged chronological folds serve selection and out-of-fold calibration. Section E.2 adds a separate training-only purged and nested sensitivity check; the deployed original policies are unchanged.

D.10 Simple target substitutions on the original draws

The main CSI300 result establishes the effect of the complete RPC edit. To isolate the role of its target predictor, we replace only its center with three alternatives: observed-context correlation, the strict-training future Fisher-z mean, and its median. The latter two are constant within each pair. All interval substitutions reuse the original fitted support rule, half-RMSE width, training quantile

Table 30: Retrospective simple-target sensitivity on the original CSI300 saved Base. All 12 pairs, three seeds, and all saved rows are retained. Lower distances are better. RPC substitutions keep the original half-RMSE width, support rule, inverse solver, and action/energy budgets; point substitutions additionally set the width to zero. W1 and Std refer to percentage returns and are averaged over the two channels. Changed is the percentage of rows receiving a nonidentity edit. These are descriptive comparisons on an already examined test set.

Target / rule	Corr	W1 ($\times 10^3$)	Std ($\times 10^3$)	Changed (%)
Base	0.134199	0.80122	1.39899	0.00
Learned / interval (original)	0.082505	0.80064	1.39713	81.19
Learned / point (original)	0.080627	0.80050	1.39638	100.00
Observed context / interval	0.086924	0.79283	1.37789	80.92
Training mean / interval	0.083486	0.80359	1.40517	81.95
Training median / interval	0.083350	0.80031	1.39835	81.80
Observed context / point	0.084996	0.79239	1.37694	100.00
Training mean / point	0.082257	0.80399	1.40537	100.00
Training median / point	0.081849	0.80062	1.39830	100.00

limits, inverse solver, scale box, action cap 0.20 and paired-energy cap 0.10. We also retain the corresponding point-target variants, which additionally set the width to zero. The protocol fixes all nine arms before their new endpoints are computed, after the original test outcome was known. It is a retrospective diagnostic, not a new prospective test or a selection of the primary policy.

Training constants use the original 2010–2019 windows, deduplicated by target-start timestamp across the three seed caches. Test futures are used only to evaluate output quality. The replay first reconstructs all 36 original Base and RPC endpoints, with maximum Corr discrepancy 4.16×10^{-16} and maximum scale discrepancy 4.44×10^{-16}. It then evaluates every specified arm on all saved rows. All 324 pair–seed–arm endpoints are valid. No generation or generator training is repeated.

The two constant intervals recover 98.10% and 98.36% of the original Base-to-Interval improvement. The learned interval has a lower mean error but wins on only 6/12 pairs against the training mean and 9/12 against the training median. Both constant point targets have slightly lower mean error than the original RPC, while the original learned Point has the lowest observed mean among these arms. The constants retain the controller's sample-specific inverse and its identity and boundary rules. The archived action extrema confirm nonconstant scales within every one of the 36 cells for each constant-interval arm. Training-mean control leaves 3,069 of 17,007 rows unchanged; training-median control leaves 3,095 unchanged. Thus constant targets do not implement a fixed scalar edit. They support effective control with lightweight target calibration, while providing little evidence for a large extra benefit from the context-based predictor. They do not select a new primary policy or validate automatic selection among target families.

Marginal diagnostics use percentage returns within each row, without cross-row joins. W1 is the exact empirical Wasserstein-1 distance after pooling those returns within a channel. Std is the absolute difference of their pooled standard deviations. Both are averaged over channels, then over seeds within a pair and over pairs. The context target has higher Corr error than the learned target but lower W1 and Std errors. The original RPC has small average improvements in both relative to Base; channel-specific values and the complete tradeoffs are released. Across all edited arms, maximum log-spread and endpoint errors are 4.44×10^{-16} and 8.88×10^{-16}. The maximum relative paired-energy change is 0.10 up to floating-point tolerance.

This target-swap diagnostic supplies no new significance test or superiority gate. Section D.11 adds a separate validation-selected fixed-request comparison using recovered original model weights. Its scale selection uses validation outcomes only. The anonymous supplement retains all target-swap arm, pair and seed reductions and a replay script; source financial arrays remain outside the compact release.

D.11 Validation-selected fixed requests on the original CovPW panel

This retrospective ablation isolates the effect of path-dependent targeting. It reuses the original 12 CSI300 pairs, three model seeds and saved 2021–2022 Base draws. The original test outcomes were

already known. We specified the new comparator before computing its test scores and retained the original RPC policy.

Validation and selection. All 36 original epoch-30 checkpoints were recovered and matched their historical hashes. We generated new validation samples locally with these weights; no model was retrained. For each pair, we greedily selected nonoverlapping 480-point windows from its strict-start 2020 cache. This retained 23–24 windows per pair, 285 unique windows and 855 seed–window instances. Row selection used indices, not numeric values. Each window contains 240 context points and 240 future points. Generation receives the context only. It uses the original normalization, regime alignment, 50-step diffusion equations and positive-domain projection. A one-window implementation check matches the native CPU sampler exactly; its CPU–MPS output difference is below 6×10^{-8}. The original test draws are unchanged.

For each pair, we evaluate the 41 requests $\{0.80, 0.81, \ldots, 1.20\}$ and minimize validation Corr averaged over the three seeds. Exact ties favor the request nearest one, then the smaller value. One request is frozen per pair and shared across seeds. At application, it is clipped into each path's original feasible interval. Both arms use the same 0.20 action cap, 0.10 energy cap, support refusal and degenerate-path identity rule. Budget projection can still make the applied scale vary by path. The comparator omits path-specific target matching, rather than all path-dependent feasibility handling.

Results and cost. The fixed request reduces mean Corr from 0.134199 to 0.132025, or 1.62%. RPC reaches 0.082505: 37.51% lower than the fixed request, with 12/12 pair wins. Its mean paired difference is -0.049520, with a 100,000-resample pair-bootstrap 95% interval $[-0.062295, -0.036530]$. The two-sided sign and signed-rank tests both give $p = 0.000488$. Pair tests do not remove shared market shocks. Synchronized calendar-block resampling addresses this sensitivity using the original trading calendar and 2,000 replicates per block length. The 95% intervals for 5, 10 and 20 trading days are respectively $[-0.056176, -0.042644]$, $[-0.055821, -0.041909]$ and $[-0.056391, -0.041520]$. These are conditional checks on the observed panel, frozen models and draws, not an independent prospective trial.

Table 31: Fixed-request attribution on the original CovPW inputs. Errors are lower-is-better. W1 and Std are scaled by 10^3. Energy is the mean absolute relative change in centered paired energy. Tail error uses real-data 90th-percentile absolute log-return thresholds.

Policy	Corr	W1	Std	Tail	Mean $\lvert\eta - 1\rvert$	Energy
Base	0.134199	0.80122	1.39899	0.021950	0	0
Fixed request	0.132025	0.82933	1.46131	0.023579	0.056846	0.060455
RPC	0.082505	0.80064	1.39713	0.018119	0.074447	0.070668

RPC has lower measured errors but uses more realized editing energy than the fixed request (Table 31). The comparison holds the allowed constraints fixed; it does not establish dominance at equal realized edit size. Together with the constant-target results, it supports simple calibrated targets followed by path-dependent inversion. It does not establish a large benefit from the learned target predictor.

All 36 original Base/RPC cells reproduce the archived endpoints before the new test arm is evaluated. A second numerical formulation checks all 1,476 validation-grid scores and 108 test endpoints. It uses cumulative return moments for rolling correlation and a detrended-log-path expression for reconstruction. Its largest Corr discrepancy is below 10^{-10}. The evidence release retains the complete validation grid, pair/seed results, selected requests and every bootstrap replicate.

D.12 A Direct Use of Saved Scenarios

Correlation repair is useful only if its gains matter for the intended task. We therefore add a retrospective risk-prediction check using the original 12-pair, three-seed CSI300 panel. The protocol fixes the task, arms and endpoints before these utility scores are computed. The underlying data were already exposed. No generator is retrained, no trajectory is resampled, and no policy is selected using utility.

Task and primary endpoint. For each stored 240-point trajectory, we form 239 within-row percentage returns. The basket return is $y_t = (r_{1,t} + r_{2,t})/2$. Its lower 5% quantile q is estimated from the generated basket using linear quantile interpolation. The context-only baseline uses the same calculation on observed returns. We score each prediction on actual future basket returns with the quantile-consistent loss (Gneiting, 2011):

$$L_\tau(q, y) = (\tau - \mathbf{1}\{y < q\})(y - q), \qquad \tau = 0.05.$$

Prediction receives no actual future. This task estimates a quantile of the next window's return mixture; it does not estimate terminal wealth or a separate quantile at every forecast horizon. Each saved model provides one trajectory per context. Positive multiplicative price normalization cancels in percentage returns. Returns never cross row boundaries. Scores average over returns, rows, seeds within a pair, and then pairs equally.

Table 32: Saved-scenario utility. Quantile loss is multiplied by 10^4; allocation variance by 10^6. Lower is better. Changes and pair wins refer to each panel's Base. Context-only estimates a risk quantile directly; it does not generate relation-preserving scenarios. The two development panels share data but differ from the original panel.

Strategy	5% loss	Change	Wins / 12	Alloc. var.
Original CovPW: retrospective test reuse				
Base	3.5504	–	–	9.0263
Fixed request	3.5577	+0.20%	5	9.0573
RPC	3.5066	-1.23%	12	9.0680
Context-only	3.0419	-14.32%	12	7.4860
CovPW: exposed development windows				
Base	4.0439	–	–	11.4096
Context target, small caps	3.9549	-2.20%	12	11.4296
Context target, wide caps	3.8648	-4.43%	11	11.2399
Context-only	3.3875	-16.23%	12	9.0684
GRU: exposed development windows				
Base	3.7239	–	–	10.8020
Context target, small caps	3.7136	-0.28%	7	10.7568
Context target, wide caps	3.7348	+0.29%	4	10.7243
Context-only	3.3875	-9.03%	12	9.0684

Original-panel result. RPC lowers the primary loss by 1.23% versus Base and 1.44% versus the validation-selected fixed request. It improves all 12 pairs against Base. The mean paired difference is -4.38025×10^{-6}, with a pair-bootstrap 95% interval of $[-5.02317, -3.70355] \times 10^{-6}$ from 100,000 replicates. Synchronized circular blocks of 5, 10 and 20 trading dates give intervals $[-5.16232, -3.58983]$, $[-5.28335, -3.50005]$ and $[-5.25586, -3.51801] \times 10^{-6}$, respectively, from 2,000 replicates each. These intervals describe sensitivity conditional on the saved panel and fitted models. They do not represent new-data confirmation. The predeclared 1% and 10% quantile losses also fall, by 2.74% and 0.82%, with 12/12 pair wins for each.

The actual 5% breach rate moves from 6.03% to 5.49%. Coverage alone is insufficient: the fixed request moves it to 5.74% while increasing quantile loss. The context-only predictor has lower loss than all three synthetic arms; RPC remains 15.28% above it. The result supports a useful repair of saved scenarios, rather than leadership in standalone risk prediction.

A secondary allocation diagnostic. We also choose a constant long-only weight using only the scenario's sample covariance. The minimum-variance solution is $w = \mathrm{clip}((v_2 - c)/(v_1 + v_2 - 2c), 0, 1)$; a degenerate denominator uses $w = 1/2$. Its realized future variance is reported in Table 32. RPC increases this cost by 0.46% and worsens 11/12 pairs. The equal-weight reference variance is 8.44653×10^{-6}, below all three synthetic-arm allocations. Thus the quantile benefit does not imply improved allocation or net trading returns. Execution costs and leverage are outside this diagnostic.

Development extensions. We apply the same evaluator to the previously declared later 12 development windows for CovPW and GRU, keeping both the small-budget context policy and the

previously selected wide-budget policy fixed. No utility-driven reselection occurs. CovPW's quantile loss decreases under both policies. GRU's small-budget policy lowers it by 0.28%, whereas the wide-budget policy raises it by 0.29% despite its 41.51% Corr reduction. The latter lowers allocation variance by 0.72%. These small utility changes quantify task-dependent costs; they do not erase the substantial development Corr gains or establish new-data generalization. Context-only prediction remains stronger in both panels.

Numerical verification. All 108 model cells and 432 arm cells are complete. A separate row-wise loss and covariance formulation reproduces every endpoint. Synthetic checks cover scale invariance, degenerate covariance and the allocation solution against a dense weight grid. The complete scoring run takes 11.67 seconds on the local CPU, including verification and resampling; this is not a generator-training speed claim. Every row-level prediction, pair reduction and adverse result is retained. The replay is a second implementation by the same coordinator, not external peer review.

D.13 CPS-STYLE QUALIFICATION STUDY

We test whether an iterative clean-estimate projection can enter the same paired-generation protocol before comparing downstream quality. The CPS-style adaptation acts at the CSDI DDPM clean-estimate hook. It reuses the same checkpoint, context, reverse-noise tape, physical endpoints, strict-train target, and Base-derived matched-budget contract. It is not an official CPS implementation reproduction. The 15 pairs and three hash-derived infrastructure replicates give 45 Base cells. These complete before the train-only (q=.995) action cap is calibrated. Three frozen schedules then produce 135 candidate trials. Candidate selection receives no validation data, test outcome, ground truth, or earlier T4 outcome.

Table 33: Outcome-blind CPS-style schedule qualification. Counts are over 45 strict-training cells per schedule and do not use downstream performance.

Schedule	Active steps (weights)	Pass	Budget infeasible	Scientific infeasible	Exec. fail	All 45 eligible
C0 final only	0 (1.00)	2	16	27	0	No
C1 late 10%	0, 4 (.75, .25)	1	1	43	0	No
C2 late 20%	0, 4, 9 (.60, .25, .15)	0	0	45	0	No

The terminal categories in Table 33 are mutually exclusive. C0 has 2 passes, 16 matched-budget infeasibilities, and 27 other scientific infeasibilities; C1 has 1, 1, and 43; C2 has 0, 0, and 45. Across schedules, the failed certificates include action, paired energy, target-interval entry, final-enabled-row behavior, and, for C1/C2, active-step reachability. Thus the no-go is not explained by one tunable scalar. Loosening these checks could create an exploratory arm, but it would no longer be the frozen matched-budget comparator evaluated here. The frozen all-cell rule therefore marks every schedule ineligible. The reducer returns `NO_GO_CPS_SCHEDULE`; it selects no schedule and opens no efficacy metric. These terminal aggregates, rather than unreleased candidate-level subcounts, define the anonymous evidence boundary.

The first sealed execution is retained as a fail-closed provenance record. It stopped before a scientific reduction when candidate physical endpoints were not bitwise equal to their Base anchors after reconstruction. A fresh run used a mechanical endpoint-anchoring repair plus precommit equality checks. The scientific freeze, candidate registry, thresholds, and metrics were unchanged, and the complete 45-to-135 design was rerun. The fresh queue has zero execution failures. Its primary reduction and independent worker replay are byte-identical; a third read-only replay validates all 135 candidate artifacts and the exact private artifact set. This is negative feasibility evidence. It is not a CPS efficacy result, noninferiority test, ranking, optimal-schedule claim, or statement about constrained samplers in general.

E MECHANISM, QUALITY AND ROBUSTNESS

These checks explain the effect and its limits. Post-outcome analyses retain their original status; they do not create additional independent tests or change the predeclared decisions.

E.1 SYNCHRONIZED CALENDAR-BLOCK SENSITIVITY

Asset-disjoint pairs can still share market shocks. We therefore supplement the original pair-level inference with a retrospective temporal sensitivity check on the fixed CSI300 panel. We use circular moving blocks of trading dates, following the block-resampling principle for dependent observations (Künsch, 1989). A common date-multiplicity vector is applied to every pair, seed, arm and real reference. Each pair retains its original missing-origin pattern; absent trajectories are not imputed. This preserves synchronized shocks and local temporal dependence at the chosen block length. It conditions on the 12 pairs, their fitted models and their saved draws.

The original caches contain 5,669 unique prediction origins and 17,007 seed-row instances over 475 trading dates, from 11 January 2021 to 23 December 2022. Source-bound native indices and timestamps verify their alignment. With missing paired observations and shifted intraday origins, the actual 240-point prediction span reaches eight trading days. We fix lengths of 5, 10 and 20 trading days before computing these resampling results, after the original outcome was known. Five days is the nominal span; 10 and 20 exceed every observed span. For each length, 2,000 replicates sample circular blocks uniformly, concatenate them, and truncate to 475 dates.

Every replicate uses the original within-row percentage-return $W = 60$ correlations, their mean/std/101-quantile distance, three-seed averaging and equal pair weights. Exact integer-weighted quantiles match literal repetition within 4.0×10^{-15}; no histogram approximation is used. Resampled block joins create no new return window.

Table 34: Retrospective synchronized calendar-block sensitivity of RPC minus Base Corr on the frozen 12-pair panel. The original point contrast is -0.051694. Each fixed length uses 2,000 circular moving-block replicates, exact integer-weighted empirical quantiles, three-seed averaging within pair and equal pair weights. Five days is nominal; the verified maximum prediction span is eight trading days. These descriptive intervals do not replace the registered primary analysis.

Block days	Median	95% percentile interval	95th percentile
5	-0.051749	[-0.056994, -0.046346]	-0.047076
10	-0.051724	[-0.057795, -0.046123]	-0.046970
20	-0.051379	[-0.057426, -0.046176]	-0.046858

All three 95% percentile intervals lie below zero (Table 34). The improvement is therefore stable in this fixed-panel temporal diagnostic. Circular resampling still assumes sufficient temporal regularity, and the finite block choices do not establish robustness to arbitrary long-range dependence or regime changes. These intervals are neither a new independent confirmation nor a replacement for the original pair-level test.

E.2 PURGED AND NESTED TRAINING CALIBRATION

The original five expanding folds reused their selection data for out-of-fold width calibration and did not remove overlapping windows. A separate retrospective diagnostic first reproduces all 12 fitted models, penalty-selection scores and original RMSEs from the authenticated 2010–2019 caches. The 311,526 unique training origins map exactly to the source-bound native time sequences. All 60 original fold boundaries contain overlap between training and validation context/future support.

We then retain a training row only when its final future point precedes the first validation context point. For native prediction-origin indices o_i, the condition is

$$o_i + 239 < o_{\text{first validation}} - 240.$$

This removes 96–155 training rows at each outer boundary. Actual timestamps verify the resulting separation. One construction keeps the original selected penalty. The other selects the penalty within each purged outer-training prefix, using five purged inner folds, the original grid, MAE criterion and one-standard-error rule. Feature standardization is fitted within each training subset. Only the nested construction separates penalty selection from the outer validation outcomes; the fixed-penalty construction isolates purging while retaining the original selection.

Equal-pair mean Fisher-z RMSE changes from 0.211444 to 0.211738 under fixed-penalty purging, and to 0.212660 under nested purged selection (Table 35). The nested/original RMSE ratio ranges

Table 35: Retrospective strict-training calibration sensitivity. Entries are equal-pair means over 12 pairs and the same 155,766 outer-validation origins (these overlapping observations are not treated as independent). Purging removes native context/future overlap; nested selection uses only purged outer training data. The last column is descriptive and has no nominal coverage interpretation. The original test policy is unchanged.

Construction	RMSE_z	MAE_z	Within $\pm 0.5\sigma_{\mathrm{original}}$
Original non-purged	0.211444	0.154433	44.60%
Fixed-lambda purged	0.211738	0.154670	44.55%
Nested purged	0.212660	0.155188	44.54%

from 0.9713 to 1.0344 across pairs. This finite check finds a small change in the aggregate training error scale. The equal-pair mean fraction of unclipped Fisher-z residuals within half each pair's original OOF RMSE is 44.54% for nested calibration. That is a descriptive residual threshold, not the coverage of the clipped deployment interval or a nominal prediction guarantee. The original test policy, fitted model and width remain unchanged. The release retains all fold counts, selected penalties, actual boundary timestamps and pair-level results.

E.3 Log-to-percentage direction coverage

After the prospective outcome was known, a specification-frozen audit applied Corollary 1 to every length-60 return window in the 36 saved Base/RPC cells. It rehashed all 288 input artifacts and retained 3,061,260 total windows. A row is acted when $|\eta - 1| > 10^{-12}$. Of 2,485,440 windows in acted rows, 2,371,323 satisfy the strict sufficient margin, giving 95.4086% certificate coverage. No certified window contradicts the predicted direction or violates the bound. Empirically, all 2,485,440 comparable acted windows have the same log-return and percentage-return change sign; agreement in the remaining 4.59% is descriptive, not certified. These overlapping windows form a census of the saved paths, not 2.49 million independent inferential units. This is a post-outcome audit of realized-path bridge coverage. It changes no endpoint, arm, threshold, or inference and does not certify distance-to-real improvement, select a setting, or provide independent confirmation. The output-only repair history and complete result inventory remain in the anonymous audit bundle.

E.4 Sensitivity to the Corr definition

After the prospective $W = 60$ result was known, we froze four alternate Corr definitions before opening their values: Rolling $W = 30$, Rolling $W = 120$, and exponentially weighted $W = 60$ with half-lives 30 and 60. The replay uses the same percentage returns, four arms, three seeds, 12 pairs, aggregation direction, and exact 864-file source inventory. It performs no training, generation, controller rerun, epsilon regularization, window deletion, imputation, p-value, confidence interval, rank, or setting selection. The published $W = 60$ endpoint is replayed as a source-closure check: all 144 arm distances agree within 8.33×10^{-17}, below the frozen 10^{-10} tolerance. The first evaluator stopped before writing scientific outputs when $W = 30$ exposed missing zero-denominator handling. The fresh failure-aware run was preserved; result04 only hardened numerical and source-closure checks, and its three scientific tables are byte-identical to result03.

Rolling $W = 120$ and both exponentially weighted definitions have 144/144 valid arm–seed endpoints. Each candidate arm is favorable against Base on all 12 pair means in each definition, matching the $W = 60$ direction. Rolling $W = 30$ instead has 132 domain-invalid arm–seed endpoints. Every invalidity comes from the shared reference side: 11 pairs across all three seeds contain at least one exact zero-variance reference window, so all four corresponding arm endpoints are invalid. Generated-side arrays have no matching invalidity. We therefore retain all 720 cell rows, propagate the 11 unavailable pair contrasts per candidate, and fail the frozen four-alternate conjunction. The evidence supports direction consistency across the three full-coverage alternates, not universal window robustness. This is a post-outcome specification replay, not an independent test.

Table 36: Complete post-outcome, specification-frozen Corr-definition replay on the prospective CSI300 artifacts. Differences are candidate minus matched Base; negative is favorable. "Cells" gives valid candidate arm–seed endpoints (Base has the same coverage), while "Pairs" gives valid candidate-minus-Base contrasts. Favorable lower bounds retain 12 as the denominator; invalid pairs contribute zero to that bound but are neither ties nor adverse. "Reduction" is the eligible-block macro reduction. Rolling $W = 30$ is domain-ineligible because 11 reference pairs contain an exact zero-variance window; its single eligible pair is not promoted to a panel mean. guarded RPC and point-target RPC remain descriptive. Execution PASS is not an across-definition PASS, and no new p-value, confidence interval, rank, or optimal-definition claim is added.

Arm	Definition	Cells	Pairs	Fav. LB/12	Mean diff.	Reduction
Guarded RPC	Rolling $W = 30$	3/36	1/12	1/12	domain-ineligible	
Guarded RPC	Rolling $W = 60$	36/36	12/12	12/12	-0.022520	16.78%
Guarded RPC	Rolling $W = 120$	36/36	12/12	12/12	-0.024617	16.40%
Guarded RPC	EW $W = 60, h = 30$	36/36	12/12	12/12	-0.022216	16.67%
Guarded RPC	EW $W = 60, h = 60$	36/36	12/12	12/12	-0.022465	16.76%
point-target RPC	Rolling $W = 30$	3/36	1/12	1/12	domain-ineligible	
point-target RPC	Rolling $W = 60$	36/36	12/12	12/12	-0.053572	39.92%
point-target RPC	Rolling $W = 120$	36/36	12/12	12/12	-0.057513	38.33%
point-target RPC	EW $W = 60, h = 30$	36/36	12/12	12/12	-0.053009	39.78%
point-target RPC	EW $W = 60, h = 60$	36/36	12/12	12/12	-0.053460	39.89%
RPC	Rolling $W = 30$	3/36	1/12	1/12	domain-ineligible	
RPC	Rolling $W = 60$	36/36	12/12	12/12	-0.051694	38.52%
RPC	Rolling $W = 120$	36/36	12/12	12/12	-0.055405	36.92%
RPC	EW $W = 60, h = 30$	36/36	12/12	12/12	-0.051170	38.40%
RPC	EW $W = 60, h = 60$	36/36	12/12	12/12	-0.051593	38.49%

E.5 Action, energy and target-direction sensitivity

After the prospective result was known, we froze a mechanism-stress grid before opening any value from this replay. It reuses the 36 authenticated Base cells, references, and observed-prefix features. The nine target-conditioned policies cross $\delta \in \{.10, .20, .35\}$ with $\epsilon_p \in \{.05, .10, .20\}$; two target-free controls take the lower or upper feasible boundary under the deployed $(.20, .10)$ contract. No training, generation, target-model refit, endpoint change, or GPU work occurs. All policies and both controls are retained. The formal replay verifies the exact 864-file source inventory and reports 432/432 valid cell endpoints, 396 non-Base action audits, 144 pair endpoints, and 132 contrasts.

Table 37: Post-outcome CSI300 action/energy stress on the frozen 36 Base cells. All grid points and both direction controls are reported. Differences are candidate minus Base Corr-family distance (lower is better); F/T/A denotes favorable/tied/adverse pair means. The controls are matched-budget mechanism controls, not external baselines. No setting is selected.

Policy	δ	ϵ_p	Eligible	F/T/A	Mean Δ	Reduction
Target-conditioned	.10	.05	12/12	12/0/0	-0.028883	21.52%
Target-conditioned	.10	.10	12/12	12/0/0	-0.042658	31.79%
Target-conditioned	.10	.20	12/12	12/0/0	-0.045717	34.07%
Target-conditioned	.20	.05	12/12	12/0/0	-0.029528	22.00%
Targeted (deployed)	.20	.10	12/12	12/0/0	-0.051694	38.52%
Target-conditioned	.20	.20	12/12	12/0/0	-0.070690	52.68%
Target-conditioned	.35	.05	12/12	12/0/0	-0.029570	22.03%
Target-conditioned	.35	.10	12/12	12/0/0	-0.052650	39.23%
Target-conditioned	.35	.20	12/12	12/0/0	-0.080939	60.31%
Lower-bound control	.20	.10	12/12	0/0/12	+0.076441	-56.96%
Upper-bound control	.20	.10	12/12	8/0/4	-0.012327	9.19%

Every grid contract is fully eligible and favorable on all 12 pair means. The reductions show robustness to both tighter budgets and wider descriptive headroom. They do not identify an optimal setting: the deployed center remains the registered policy and exactly reconstructs all 36 frozen eta vectors and generated arrays. Its maximum endpoint, eta, and log-array replay errors are 4.44×10^{-16}, 4.44×10^{-16}, and 8.88×10^{-16}, respectively. The opposite boundary is adverse on 12/12 pairs;

the upper boundary is weaker and mixed. Thus feasibility and common-mode energy alone do not explain the gain. The train-only target direction is empirically consequential.

Across the center policy's 17,007 row decisions, 3,199 (18.81%) are minimum-action identities, 3,320 (19.52%) enter the target interval, and 10,488 (61.67%) saturate at a certified boundary; none is an out-of-support fallback. The center policy's maximum action and paired-energy change are 0.20 and 0.10 up to floating-point tolerance. Across all stress policies, the corresponding maxima are 0.35 and 0.20; maximum spread and endpoint errors are below 9×10^{-16}. These are post-outcome sensitivity and mechanism results. They add no p-value, confidence interval, rank, policy-selection claim, independent confirmation, or external guided-method comparison.

E.6 Marginal quality and diversity

Because every saved rowwise scale is positive, we algebraically invert the 45 saved guarded RPC candidates and replay RPC, obtaining maximum log/relative replay errors of 7.77×10^{-16}/7.81×10^{-16} and maximum base–candidate log-spread discrepancy 4.44×10^{-16}. This SHA-bound, post-hoc matched-input reanalysis is conditional on saved candidates and scales; it is neither a registered CRN experiment nor a byte-level regeneration of the original sampling orchestration. Observed context enters only the unchanged novelty/copy/diversity screen; future targets are opened only after generation for 28 marginal endpoints using within-future log returns. Seeds are averaged within pair, leaving 15 inferential units; exact two-sided Wilcoxon tests are Holm-adjusted separately within three endpoint families.

Table 38: Post-hoc 15-pair secondary matched-input audit. Here Δ =guarded RPC minus its reconstructed controller input, so lower is better; seeds are averaged before pair-level analysis. Novelty is a frozen point-estimate screen; marginal intervals and adjusted p-values are descriptive and introduce no new no-harm margin.

Audit/endpoint	Direction summary	Effect summary	p_{Holm}/boundary
Novelty/copy/diversity	five gates pass	copy change ≈ 0	frozen screen
Marginal endpoints	24/28 favorable	10 favorable, 0 adverse	descriptive
Return W1, channel 1	15/0/0	mean -3.95×10^{-6}	.00098
Return W1, channel 2	12/3/0	mean -2.84×10^{-6}	.2888
Abs.-return ACF, ch. 1	14/1/0 at lags 1/5/20	all favorable	.00073 each
Abs.-return ACF, ch. 2	13/2/0 at lags 1/5/20	all favorable	.0251 each
Raw-return ACF	mixed	no endpoint survives	$\geq .723$

All five frozen novelty gates pass. Of 28 marginal endpoints, 24 have favorable mean differences and ten favorable effects survive Holm correction; none of the adverse effects does. The four positive means comprise two numerical-zero mean errors and two nonsignificant raw-return-ACF effects. The marginal intervals are descriptive and do not constitute a new noninferiority or formal no-harm claim.

E.7 Earlier asset-disjoint audit

The registered audit contains five CSI300 pairs that share no asset with the ledger-frozen 15-pair benchmark and three independently trained seeds per pair. For every pair–seed cell, CSDI-CovPW generates one stochastic base sample and the base and guarded RPC branches receive the same post-validity-projection input. A shared positive projection affects 3/7,146 rows before the split, so the comparison estimates the controller increment conditional on that map rather than an end-to-end effect on raw sampler output.

Registered Corr differences are favorable on all five pairs: mean -0.025316, median -0.022929, exact one-sided $p = .03125$, and 95% bootstrap upper bound -0.020432. ADF differences are exactly zero. The Coint differences are $-0.554010, 0.530896, -0.008629, -0.121616, -0.009574$: one exceeds the registered 0.153875 margin, and the bootstrap upper bound is 0.206803, so the Coint safety gate fails. The maximum base–candidate log- spread error is 4.44×10^{-16}.

After this outcome was opened, a structural-coverage rule retained only pairs whose two assets have full interval coverage. Although independent of candidate metrics, the rule's timing makes the

Table 39: Registered five-pair asset-disjoint common-random-number audit. Each pair averages three independently trained seeds; both branches receive the same stochastic base sample and positive-domain projection. Entries are guarded RPC minus CSDI-CovPW family distances, so negative values favor RPC. The registered five-pair result remains primary: Corr and ADF pass, whereas Coint safety fails because one pair exceeds the prespecified margin $\Delta_{\mathrm{NI}} = 0.153875$. The later four-pair structural-coverage subset is post-outcome and is not shown as a replacement result.

Pair	ΔCorr	ΔADF	ΔCoint	Registered pair-level status
sh601991_sz002194	-0.02406170	0	-0.55401033	Corr favorable; ADF equal; Coint safe
sh601158_sz000661	-0.04023145	0	0.53089643	Corr favorable; ADF equal; **Coint unsafe**
sh600050_sh600060	-0.01694457	0	-0.00862858	Corr favorable; ADF equal; Coint safe
sh600236_sz002244	-0.02292886	0	-0.12161612	Corr favorable; ADF equal; Coint safe
sh600282_sh600361	-0.02241411	0	-0.00957394	Corr favorable; ADF equal; Coint safe
Five-pair mean	-0.02531614	0	-0.03258651	5/5 Corr favorable; 5/5 ADF equal; 1/5 Coint unsafe
Five-pair median	-0.02292886	0	-0.00957394	N/A
Registered aggregate	**Pass**	**Pass**	**Fail**	Corr: exact one-sided $p = .03125$, 95% boot. upper -0.02043219

Coint's maximum pair difference is 0.53089643 and its 95% bootstrap upper bound is 0.20680328, both above the registered margin. Maximum base–candidate log-spread error is 4.44×10^{-16}.

resulting four-pair analysis post-outcome. On this subset, Corr and Coint differences are negative on all four retained pairs and ADF remains zero, but this sensitivity cannot replace the registered five-pair result.

All four sensitivity Corr and Coint effects have the favorable sign. Four nonzero pairs nevertheless have minimum attainable one-sided exact Wilcoxon $p = 1/16 = 0.0625$. The percentile bootstrap is descriptive and conditional on these four pairs. It does not establish population-level superiority, and the exact-test floor makes the small pair-level sample size explicit.

E.8 Conditional guard-removal effects

Post-hoc guard-removal audit. On the already opened full-15 benchmark, we replayed the frozen controller on its certified same-input reconstructed bases and removed one declared component at a time. Pair-level effects use arm-minus-full-RPC distances after averaging three seeds, with five-removal Holm correction applied separately within ADF, Coint, and Corr. Because support, uncertainty attenuation, sensitivity, reachability, and energy constraints interact and their row-level activations can overlap, these are conditional, nonexclusive, and nonadditive component-removal associations. They do not establish guard necessity, isolated causal contributions, independent confirmation, universal safety, or a new superiority/noninferiority gate.

E.9 Runtime, action coverage and remaining sensitivity checks

The 45 guarded RPC ledger cells retain stage timing. Median generation time is 42.884 s (range 42.606–103.348 s), median controller time is 0.486 s (0.378–1.217 s), and the median cellwise controller/generation ratio is 1.068% (0.751–1.394%); total controller time divided by total generation time is 1.037%. The process-level peak allocated memory has median 4.745 GB, but measurement begins before generation and ends after control. It therefore does not estimate controller-incremental memory or compare prewhitening with original coordinates.

Across 20,160 guarded RPC controller rows, nonexclusive trigger counts are 189 support (0.938%), zero sensitivity, and 3,674 reachability (18.224%); their union produces 3,833 hard fallbacks (19.013%). Because triggers can overlap, these frequencies describe deployed behavior rather than mutually exclusive or causal attribution. A post-hoc 45-cell same-input leave-one-out audit instead finds exact ADF ties, no Holm-significant Coint change, and lower Corr distance after every removal (Appendix E.8); with overlapping activations, these are conditional associations rather than necessity or isolated-causality tests.

For guarded policy, pairwise hard-fallback counts range from 102/1,344 (7.59%) to 406/1,344 (30.21%); the full 15-pair vector remains in the audit CSV. T2 and T3 use mutually exclusive RPC statuses. "Boundary saturation" means a safe action at the nearest feasible bound, not refusal. In the CSI300 compact record, all 108 candidate arm records—nine per pair—carry the conservative

Table 40: Post-hoc full-15 same-input controller-component removal audit. Differences are ablated arm minus full RPC, so positive values indicate a larger (worse) distance after removal. Seeds are averaged before pair-level analysis. W/L/T counts pairs where removal is lower/higher/equal; Holm correction covers the five removals separately within each family. Guard activations can overlap; their interactions make these conditional, nonexclusive, nonadditive contrasts, not necessity or isolated causal effects.

Removed	Family	Mean	Median	W/L/T	Raw p	Holm p
Support	ADF	0	0	0/0/15	1	1
Support	Coint	$+6.241\times10^{-5}$	$+3.152\times10^{-6}$	7/8/0	0.8469	1
Support	Corr	-4.404×10^{-4}	-2.805×10^{-4}	13/2/0	0.00116	0.003479
Uncertainty	ADF	0	0	0/0/15	1	1
Uncertainty	Coint	-0.06636	-2.102×10^{-4}	8/7/0	0.9341	1
Uncertainty	Corr	-9.017×10^{-4}	-0.001166	14/1/0	0.006714	0.01343
Sensitivity	ADF	0	0	0/0/15	1	1
Sensitivity	Coint	-0.02957	$+4.026\times10^{-4}$	7/8/0	0.8904	1
Sensitivity	Corr	-5.560×10^{-4}	-2.826×10^{-4}	11/4/0	0.03015	0.03015
Reachability	ADF	0	0	0/0/15	1	1
Reachability	Coint	-0.03125	-0.002074	9/6/0	0.06372	0.3186
Reachability	Corr	-0.008469	-0.009121	15/0/0	6.104×10^{-5}	3.052×10^{-4}
Energy bound	ADF	0	0	0/0/15	1	1
Energy bound	Coint	+0.07206	-0.001117	8/7/0	0.9341	1
Energy bound	Corr	-0.007314	-0.005976	14/1/0	1.831×10^{-4}	7.324×10^{-4}

Mean-difference confidence intervals (95%):

Removed	Family	Confidence interval
Support	ADF	[0, 0]
Support	Coint	$[-2.206\times10^{-4}, +3.897\times10^{-4}]$
Support	Corr	$[-7.465\times10^{-4}, -1.903\times10^{-4}]$
Uncertainty	ADF	[0, 0]
Uncertainty	Coint	[-0.2223, +0.02281]
Uncertainty	Corr	[-0.001269, -4.517×10^{-4}]
Sensitivity	ADF	[0, 0]
Sensitivity	Coint	[-0.08902, $+8.715\times10^{-4}$]
Sensitivity	Corr	$[-9.754\times10^{-4}, -1.506\times10^{-4}]$
Reachability	ADF	[0, 0]
Reachability	Coint	[-0.0681, -0.002197]
Reachability	Corr	[-0.01105, -0.005908]
Energy bound	ADF	[0, 0]
Energy bound	Coint	[-0.08249, +0.257]
Energy bound	Corr	[-0.01091, -0.004447]

Table 41: Sign-reversed consistency check against the earlier post-hoc same-input core audit. No new hypothesis test or gate is introduced.

Family	Prior Full−Id.	Implied Full−Id.	|difference|
ADF	0	0	0
Coint	+0.2784	+0.2784	0
Corr	-0.007914	-0.007914	3.764×10^{-16}

Table 42: Descriptive controller actions in the same post-hoc audit. Counts are pooled over 45 cells and are nonexclusive. The identity raw-clamped count describes unused prepared geometry.

Quantity	Identity	Full RPC
Cell-mean mean $\|\eta-1\|$	0	0.02462
Maximum $\|\eta-1\|$	0	0.1222
η range	[1, 1]	[0.9083, 1.122]
Maximum relative energy change	0	0.09847
Fallbacks Supp./Sens./Reach./union	0/0/0/0	189/0/3674/3833
Prepared raw-clamped rows	6229	6229
Certified arm-cells	315/315	

Table 43: Observed controller decision coverage. The no-action column is a hard fallback for guarded policy and a minimum-action identity for supported RPC rows; it does not mean universal refusal. Bounded saturation is reported separately. Percentages use all registered controller rows; "not recorded" marks categories not exposed as mutually exclusive by that frozen policy.

Panel	Policy	Rows	Identity/fallback	Bounded saturation	Entered interval
Historical 15-pair	guarded RPC	20,160	3,833 (19.01%)	not recorded	not recorded
Fresh financial T2	RPC	20,160	5,207 (25.83%)	11,832 (58.69%)	3,121 (15.48%)
Frozen GHCNh T3	RPC	360	12 (3.33%)	345 (95.83%)	3 (0.83%)
Post-hoc CSI300 replay	RPC	17,007	3,199 (18.81%)	10,488 (61.67%)	3,320 (19.52%)

arm-level flag because at least one row is identity/refusal or saturated. That record does not expose an exact row-frequency table, so we do not infer one from the prospective reduction. The later authenticated post-outcome reconstruction does expose the RPC row decisions; its separate counts are shown in Table 43.

Post-hoc validation-only shrinkage sensitivity. The failure-aware aggregation reports all 135 cells from 15 historical pairs: 86 are domain-valid and 49 are domain-invalid, with no execution failures. Both .05 and .20 have nine complete three-seed candidate/reference pair blocks out of 15, so neither alternative is performance-eligible across the panel. Among those nine eligible pairs, .05 has favorable/adverse directions of 7/2 for ADF, 8/1 for Coint, and 3/6 for Corr; the corresponding counts for .20 are 4/5, 3/6, and 6/3. Mean candidate-minus-.10 effects are $(-0.321, -0.122, +0.0054)$ for .05 and $(+1.355, +0.508, -0.0115)$ for .20 in (ADF, Coint, Corr) order. Table 44 discloses every pair position, including ineligible blocks, and the public CSV retains all unrounded values. These are descriptive validation effects, not p-values, confidence intervals, ranks, test robustness, noninferiority, or evidence for an optimal λ.

Post-hoc structural dose controls. The full-panel RPC dose audit contains all 15 pairs, 45 cells, and 225 cell–dose rows. It verifies identity at $\alpha = 0$, recovery of the deployed candidate at $\alpha = 1$, exact common-channel shifts, linear common-shift scaling, quadratic energy scaling, and preservation of the controller-input log spread. This establishes only a structural common-mode strength knob; no real future target or downstream quality endpoint is opened.

The generated-only Coint dose gate is post-outcome and is not an independent confirmation of the registered five-pair audit. It selects retained doses $(0, .5, 1, 1, 1)$ for anonymous pairs A–E, with mean retained dose 0.7 and maximum generated-only Coint-family displacement $0.125354 < 0.153875$. The triangle-inequality certificate therefore passes under the same metric, but pair A falls

Table 44: Post-hoc validation-only CovPW shrinkage diagnostic on all 15 historical benchmark pairs. "Valid" gives domain-valid seeds at $(\lambda = .05, .10, .20)$ out of three. Each effect is the candidate mean minus the $\lambda = .10$ mean in (ADF, Coint, Corr) family order; lower is favorable. Effects are reported only when both candidate and reference have all three valid seeds. The complete unrounded rows and all invalid cells are retained in the public audit artifacts.

Anonymous pair	Valid $(.05/.10/.20)$	$\Delta_{.05}$ (ADF, Coint, Corr)	$\Delta_{.20}$ (ADF, Coint, Corr)
01	$3/3/3$	$(-0.894, -1.631, +0.0162)$	$(+1.038, -0.098, -0.0317)$
02	$3/3/3$	$(-0.034, -0.265, +0.0090)$	$(+0.379, +0.376, -0.0263)$
03	$0/0/0$	ineligible	ineligible
04	$3/3/3$	$(-0.510, -1.029, +0.0090)$	$(-1.456, -1.977, -0.0205)$
05	$3/3/3$	$(-0.902, -0.005, +0.0114)$	$(-0.542, +0.448, -0.0256)$
06	$3/3/3$	$(+3.668, -1.493, +0.0008)$	$(+15.115, +6.395, -0.0005)$
07	$3/3/3$	$(-3.130, -0.116, +0.0150)$	$(-2.884, +0.684, -0.0249)$
08	$3/3/3$	$(-0.762, +5.113, -0.0014)$	$(-0.208, -3.137, +0.0055)$
09	$3/3/3$	$(-1.447, -1.448, -0.0046)$	$(+0.106, +1.771, +0.0104)$
10	$3/3/3$	$(+1.126, -0.225, -0.0069)$	$(+0.650, +0.113, +0.0104)$
11	$1/1/1$	ineligible	ineligible
12	$0/0/0$	ineligible	ineligible
13	$0/1/1$	ineligible	ineligible
14	$0/0/0$	ineligible	ineligible
15	$0/0/0$	ineligible	ineligible

back to identity and supplies no efficacy evidence. The original registered Coint-safety failure is retained unchanged.

Still open are prospective context-window, λ/floor, and adaptive-precision selection, branchwise memory, official-package PDM and CPS reproductions, matched SAL/PCD/CCS rows, a CPS efficacy comparison after an eligible schedule exists, purged or nested nonlinear-predictor calibration, correlation-prewhitening, and independent tests across broader variables and regions. The completed T2 closes fresh full-benchmark shared-draw attribution but remains on an opened panel; T3 supplies independent non-financial evidence, and the CSI300 panel supplies an asset-disjoint prospective financial test. None establishes population-wide robustness. Anonymous summaries bind timing and guard aggregates; raw financial run-level JSON remains withheld pending anonymization.

E.10 Exhaustive sensitivity of existing benchmark leadership

This post-outcome analysis reuses the frozen pair-level benchmark; it introduces no new training, generated samples, or independent test data. We distinguish the jointly frozen five-method, three-seed block from the 16-method descriptive join, whose source blocks retain different training and replication protocols. All original endpoints and paired inferential decisions remain unchanged.

Pair and metric sensitivity. Guarded RPC remains uniquely first by mean family rank in every one of the $4{,}944$ subsets retaining 10–15 of the original 15 pairs. The result is exhaustive over these subsets, rather than a sampled bootstrap. It also remains first under all three original aggregation rules after deleting any single pair, any single one of the nine raw diagnostic metrics, or any single competing method. For metric deletion, the remaining metrics are reaveraged within their original family; for comparator deletion, within-pair ranks and normalization are recomputed. These results support stability to the specified benchmark perturbations, not population-wide or metric-independent superiority.

Aggregation limits. On the full five-method panel, our mean ranks for (ADF, Coint, Corr) are

$$(2.4333, 2.2000, 1.7333),$$

componentwise no greater than those of any comparator. Thus no nonnegative weighting of these three *rank summaries* can strictly demote our aggregate position; using only ADF produces a tie with CSDI-CovPW. This statement does not apply to raw distances or every other aggregation rule. For metric-balanced rank and normalized score, another method can tie us after moving respectively 0.214850 and 0.143243 of total weight from the equal-weight vector, where movement is measured by $\frac{1}{2}\|w - (1/3, 1/3, 1/3)\|_1$. Removing the entire Corr family gives positions 1, 2, and

Table 45: Frozen five-method panel: first-place counts, including ties, over every subset of each size. Overlapping subsets are not independent trials; the fractions are descriptive sensitivity summaries. Metric-balanced rank has two tied-first cases among the 2,948 first-place 10-pair subsets; all other reported first-place cases are unique.

Retained pairs	Subsets	Family rank	Metric rank	Normalized score
10	3,003	3,003	2,948	2,601
11	1,365	1,365	1,359	1,249
12	455	455	455	434
13	105	105	105	103
14	15	15	15	15
15	1	1	1	1

4 under family rank, metric-balanced rank, and normalized score, respectively. On individual markets, normalized-score positions are 1 (American), 2 (CSI300), and 4 (Crypto); omitting American or CSI300 gives position 2. We therefore do not claim universal aggregation or subgroup robustness.

Exhaustive source-precedence sensitivity. We enumerate all $2^3 = 8$ global source choices for the three overlapping methods: CSDI-isotropic, CSDI-CovPW, and CTSDiff-SPRC. Each choice applies to all 15 pairs; no source is selected separately by pair or outcome. All three aggregate rules place us first in the four combinations using the registered CSDI-CovPW rows, and second in the four using its earlier broad-screen rows. Within this finite factorial audit, the binary first/second switch is determined by that comparator's source choice, although the other source choices also change numerical scores. This localizes rather than eliminates the previously disclosed source sensitivity. The matched frozen block is the basis for the core ranking claim; the 16-method join remains descriptive breadth evidence.

Reproducibility. Input hashes are checked against the original benchmark builder. The audit reproduces 1,191 archived score/rank comparisons with maximum absolute error 3.56×10^{-15}, then retains all deletion, source, subset, and weight cases. An independent implementation uses pair-count average ranks and direct subset aggregation; continuous weight radii are checked through geometric vertex enumeration. The analysis includes all 231 family-weight grid points at spacing 0.05, without selecting a preferred weighting. No sensitivity fraction is interpreted as a p-value, probability of SOTA, or replacement for the original confirmatory tests.

E.11 Target information and rejected refinements

Why diagnose targets separately? The invariant is a realized-path property; distance to a real distribution also depends on the target supplied to the controller. We investigate this separation on the already-open GRU development data. These experiments use a research prototype with a rolling-distribution objective, not a newly selected version of the main RPC policy. No result below changes the registered test, the GRU transfer failure or the ranking table.

Privileged-target diagnostic. We reuse 12 pairs, three seeds and 24 rows per pair from the 2020 development panel. Each context plus future spans 480 native points; retained intervals have no shared points. Three chronological blocks contain eight rows each. The prototype uses one scalar coefficient per generated row to edit centered common log increments, 80 Adam steps at learning rate .03, identity initialization, the best iterate including identity, coefficient cap .20, conservative relative energy cap .10 and normalized edit penalty .001. Targets summarize length-60 rolling percentage-return correlations. Optimization uses float32 MPS; final projection, reconstruction and evaluation use float64. This is a batch-specific optimization, not a fitted forecasting policy.

ORACLE_SELF supplies the actual future of the same eight-row block and therefore directly fits the scoring outcomes. ORACLE_OTHER supplies the other two blocks (16 rows), excluding every current-block native point. It may nevertheless use later observations and is noncausal. SELF has eight target rows, whereas OTHER and historical comparators use 16; we do not hide this difference

by duplicating rows. Both arms have privileged future information, remain temporally related to the evaluation data, and are neither deployable predictors nor mathematical performance bounds.

Table 46: Target-information diagnostic on opened GRU development data. All seven arms are reported; the final two are intentionally nondeployable. Corr is a boundary-safe family distance; Std and W1 changes refer to channelwise distance to real data, not the edit budget. A positive quality change is adverse. Pair counts follow three-seed averaging.

Arm	Corr	Corr wins/12	Std change	W1 change
BASE	.061399	–	–	–
Historical FIXED	.063887	4	-5.68%	-1.81%
Historical CONTEXT	.063809	5	-4.16%	-2.32%
FIXED + gate	.057838	3	-0.52%	-0.13%
CONTEXT + gate	.062356	1	-0.54%	-0.38%
ORACLE_SELF	.028772	12	$+4.65\%$	$+1.18\%$
ORACLE_OTHER	.028371	12	$+4.23\%$	$+0.94\%$

Historical FIXED selects 16 evenly spaced 2010–2017 future rows. Historical CONTEXT retrieves two training examples per query context by log common-increment variance. Its name denotes retrieval, not the direct context-persistence baseline below. The gates combine training feature support with the maximum of historical 2018–2019 block discrepancy scores. FIXED+gate has three improved and nine identical pairs; CONTEXT+gate has one improved, ten identical and one adverse pair. All arms share the same Base tensors. These historical alternatives had already failed their separate resource-screening conditions; their apparent benefits on selected summaries do not retrospectively change those decisions.

SELF and OTHER lower pair-pooled macro Corr by 53.14% and 53.79%, respectively, on 12/12 pairs. At block level SELF improves all 108 comparisons, whereas OTHER improves 65 and worsens 43. OTHER's slightly lower pooled score is therefore not evidence of better per-block forecasting or optimality. Both worsen marginal variance, quantile and W1 distances; full higher-moment, ACF, tail and cross-lag results accompany the compact evidence. The 216 new optimizations demonstrate observed fitting capacity with privileged targets, not prospective efficacy, and cannot isolate target information as the sole cause. All 252 arm records are recomputed by an independent implementation; 6,397 checks pass, including temporal membership, reconstruction and budgets.

Historical target-summary forecasting screen. A separate, subsequently frozen screen uses only 2010–2019 history and does not reopen 2020 price arrays. FIT uses 151 eight-row blocks with pre-2016 origins, SELECT 55 blocks from 2016–2017, and AUDIT 59 blocks from 2018–2019. All 12 pairs are retained; insufficient-history blocks and incomplete tails are disclosed. Each historical reference must end before the current block's earliest context begins. AUDIT may use previously completed historical windows, but fitted coefficients and selection remain frozen. These years have already participated in development, so AUDIT is not pristine holdout confirmation. Contexts are available at their respective row origins; we make no claim that all eight are known at the earliest origin of a batch.

EXPANDING16 uses 16 evenly spaced available past rows, RECENT16 the latest 16, and CONTEXT the current observed contexts. A shared ridge model predicts the 103-component future Corr summary minus RECENT16, from eight features: mean, std, q10 and q90 for context and RECENT16. Training uses pair-balanced weights, training-only standardization, an unpenalized intercept and $\lambda \in \{.1, 1, 10\}$. Predictions are clipped to valid scalar ranges and quantiles sorted; these operations do not guarantee a coherent probability distribution. The simple comparator and ridge strength are selected globally on SELECT, then ridge is refitted on FIT+SELECT before AUDIT outcomes are opened for this study. There are no artificial seed replicates of real targets.

The selected ridge ($\lambda = .1$) lowers AUDIT error by 11.26% relative to selected CONTEXT persistence, with 9/12 favorable pairs and all three macro components lower. It is nevertheless 13.47% worse on SELECT. The frozen expansion rule requires improvement on SELECT as well as at least a 10% AUDIT reduction, eight favorable pairs and no worse macro component. Thus the conjunction fails, and the candidate is not integrated into RPC. The positive later-period result motivates further target research, not a retroactive success claim. Five numerical prechecks and 6,728 independent

Table 47: Historical target-summary prediction, not generated-path quality. Distances use the original equally weighted Corr mean/std/quantile family; blocks are averaged within pair, then pairs equally. All frozen candidates are retained. No candidate is selected on AUDIT.

Predictor	SELECT (2016–2017)	AUDIT (2018–2019)
RECENT16	.092090	.054165
EXPANDING16	.091843	.061434
CONTEXT persistence	.045283	.046759
Ridge $\lambda = .1$	.051384	.041493
Ridge $\lambda = 1$	.071051	.042698
Ridge $\lambda = 10$	.090737	.055244

implementation checks pass, including augmented-system least-squares verification; no generator or controller is run in this screen.

A rejected variance-budget refinement. A separate analytic cap shrinks scalar edits toward identity to bound both channelwise log-return variance changes relative to Base. Four budgets (1%, 2.5%, 5%, 10%) and equal-average-action shrinkage controls are screened on two earlier disclosed development periods: 12 pairs, three seeds, ten arms, 720 records. All four global selectors choose BASE; every nonzero arm has a worse macro Corr distance on the later period. The resource gate fails and no cap is added to the submitted method. This negative result is consistent with the identity $\mathrm{Var}(x - y) = \mathrm{Var}(x) + \mathrm{Var}(y) - 2\,\mathrm{Cov}(x, y)$: if the difference and both marginal variances are fixed exactly, correlation cannot change. Relaxed variance caps permit edits but do not guarantee better distance to real data, Coint or tail behavior.

F Development Candidates and Task-Aligned Tradeoffs

These studies ask how much useful performance remains beyond the frozen RPC policy. Candidate selection uses the exposed 2020 development panel. A later replay applies its selected GRU policy to saved 2021–2022 paths without reselection. Those years were also previously evaluated, so neither study supplies pristine held-out confirmation. We retain all candidates and the historical GRU failure. These extensions do not replace the primary method or the five-method benchmark.

F.1 Context-Targeted Scalar Search

The new solver keeps one scalar per row but optimizes a rolling percentage-return distribution objective. It uses cumulative sums to compute rolling moments and a vectorized coarse grid with local refinement. The exact energy interval follows from Equation 4. The search evaluates 163 candidates rather than taking 80 Adam steps. This is a finite search, not a certificate of continuous global optimality. The original log-spread and both path endpoints remain fixed.

The panel has 12 pairs, three seeds and 24 nonoverlapping windows per pair. Each window contains 240 context and 240 future points. The first 12 windows select one global configuration from 13 candidates; the last 12 evaluate it without reselection. Both parts belong to data used in earlier development. The selected configuration uses each row's observed context as its target, action and paired-energy caps of 0.8, and an edit penalty of 0.001. It is denoted `CONTEXT_EXACT80` in the artifacts. The larger caps, target and objective differ from the main RPC policy.

For CovPW, later-window Corr falls from 0.154655 to 0.041621 (73.09%), with 12/12 pairs improved. W1 improves by 4.35%, while standard-deviation error rises by 2.11%. For GRU, Corr falls from 0.054506 to 0.031879 (41.51%), with 9/12 pairs improved; W1 and standard-deviation errors rise by 2.13% and 5.03%. The controlled GRU remains better on seven of eight reported quality endpoints than controlled CovPW, so the larger relative CovPW gain is not an end-to-end leadership result.

A matched-target comparison also helps locate the improvement. With the conservative action cap 0.20 and energy cap 0.10 fixed, the same grid solver gives GRU Corr errors of 0.064548 for the historical target and 0.042380 for the observed-context target, against 0.054506 for Base. The latter

Table 48: Development-only context-targeted scalar search. Each row averages three seeds within each of 12 pairs, then averages pairs. The table uses the later 12 windows per pair from the exposed 2020 development data. W1 and Std are multiplied by 10^3; all columns are errors, lower is better. The search is different from the primary RPC policy.

Generator	Control	Corr	W1	Std	Tail
CovPW	None	0.154655	1.049664	1.021241	0.022993
CovPW	Context search	0.041621	1.004024	1.042799	0.011884
GRU	None	0.054506	0.906297	0.972667	0.011952
GRU	Context search	0.031879	0.925603	1.021601	0.010228

improves 10/12 pairs and reduces macro Corr by 22.25%, with W1 and marginal Std error increases of 0.92% and 2.44%; tail-coexceedance error falls by 5.92%. Thus positive development results do not require the larger 0.8 budgets. This contrast changes only target source within the new rolling-distribution solver. It does not isolate the cause of the earlier whole-row RPC failure, whose objective and target construction differ. All 13 candidates remain in the released results, and the original selected configuration is unchanged.

The full 24-window ADF replay has zero decision flips. Coint p-value distribution error increases from 0.164409 to 0.166452 for CovPW and from 0.058296 to 0.060677 for GRU. These diagnostics use within-row 60-point log-price windows and AIC lag selection. They differ from the historical pass-rate/run-length families and are not merged with them.

On local MPS, isolated eight-row microbenchmarks compare the grid and Adam solvers on the same conservative feasible interval. Median times are 0.007560 versus 0.409079 seconds for CovPW, and 0.025355 versus 1.036584 seconds for GRU. These correspond to 54.11-fold and 40.88-fold solver speedups. They do not measure training or end-to-end generation speed. The numerical reductions and source hashes are retained in the development evidence package.

F.2 Frozen GRU Policy Across Time

We next apply the 2020-selected context policy to every saved 2021–2022 GRU trajectory. The panel contains 12 pairs, three seeds and 5,660 unique pair–origin rows, or 16,980 seed–window instances. No generator is retrained and no path is resampled. The target, grid search, penalty and caps stay fixed. We record the protocol and input hashes before computing these new policy outcomes. Control receives only a generated path and its observed context. A declared identity fallback handles undefined input correlations; it is used on zero rows. Every row passes the spread, endpoint and edit-budget checks.

Table 49: Retrospective GRU temporal replay on saved 2021–2022 paths. All changes and Corr pair wins refer to Base. W1 and Std columns are error changes; positive values are costs. The last two policies were fixed in the 2020 study, and are applied here without reselection.

Policy	Corr	Change	Wins / 12	W1 Δ	Std Δ
Base	0.054600	–	–	–	–
Original fixed request	0.044933	-17.71%	6	+4.46%	+9.03%
Original RPC	0.060523	+10.85%	1	-0.29%	-0.25%
Context target, small caps	0.044405	-18.67%	9	+1.23%	+3.33%
Context target, wide caps	0.010695	-80.41%	12	+1.22%	+5.30%

Correlation repair persists beyond development windows. The wide-cap policy reduces Corr from 0.054600 to 0.010695, an 80.41% reduction with 12/12 pair wins. Its pair-bootstrap 95% interval for the mean difference is $[-0.056378, -0.032778]$ from 100,000 replicates. This is conditional descriptive sensitivity to pairs, not a new confirmatory test or a temporal bootstrap. The conservative policy also improves Corr by 18.67%, with 9/12 pair wins. The original RPC still worsens Corr by 10.85%; its result belongs to a different target policy. Thus a failed frozen policy does not imply that the shared relation-preserving action cannot help this generator.

Quality and utility costs. Under wide caps, W1 and Std errors increase by 1.22% and 5.30%. Tail-coexceedance and cross-lag errors decrease by 6.45% and 29.15%. The 5% basket-quantile loss increases by 1.06%, while allocation variance decreases by 0.67% on 11/12 pairs. Under conservative caps, these utility changes are $+0.008\%$ and -0.37%, respectively. The context-only quantile predictor remains stronger. All eight quality endpoints and both utility endpoints appear in the numerical supplement; the policy was not reselected using them. Correlation repair is therefore substantial, but its application value still depends on the task.

What the replay establishes. The strategy was selected on 2020 outcomes and is unchanged here. However, earlier evaluation on the 2021–2022 panel revealed the original GRU failure and motivated later development. Reusing this panel expands temporal coverage without restoring data independence. The result supports retrospective temporal robustness and cross-family applicability; confirmation on previously unexposed data remains a separate question.

All 108 historical Base, Fixed and RPC arm–cell endpoints reproduce. A second cumulative-moment implementation also agrees with the new Corr scores. The two new policies take 19.06 seconds on local MPS; full scoring and numerical verification take 171.65 seconds on CPU. These times cover saved-path control and evaluation only. The executable package runs both the original RPC reference and the new distribution-targeted policy from a clean directory. Its synthetic examples need no financial data or private project path. This verifies implementation portability; it does not add statistical independence or reproduce full generator training.

F.3 Return Representation and Volatility Calibration

The next study trains two pooled CSDI generators on 19,155 full windows ending before 2018. One models log returns; the other models common log-return increments and spread levels. They share initialization, training rows, noise and a fixed 2,048-update schedule. Each batch has 24 rows, two per pair. Both use observed-context whitening and 50-step DDPM sampling. A separate internal panel contains 24 nonoverlapping 2018–2019 windows per pair. No full-training-period refit or multi-seed confirmation has been completed.

A channel-symmetric Ridge model predicts log volatility from six observed features: log standard deviations over the last 239, 120, 60 and 30 returns, and log mean absolute returns over the last 239 and 60 returns. It fits only pre-2018 data. Calibration rescales centered generated returns by the predicted-to-generated standard-deviation ratio, clipped to $[0.5, 2]$. This preserves the first and last future price but generally changes the spread. When RPC follows this step, its invariant is the *calibrated input spread*, not the earlier generator output.

The original internal screen requires improvement in a composite quality score together with separate Corr, variance, temporal/tail and stationarity conditions. No new candidate meets every condition. Return+CAL+RPC improves the composite by 3.99% but narrowly worsens Corr and increases the temporal/tail family. We retain that failed decision. A later, task-oriented reassessment keeps the candidate for further study because uniform nondegradation is stronger than the requirements of the stated control task. That reassessment is a research decision, not a retrospective validation success.

Table 50: Return-space development candidate: all 24 windows per pair in 2020, 12 pairs, one seed. W1 and Std are multiplied by 10^3. ADF is absolute rejection-rate error, not the historical ADF family. All columns are errors. External generators use their existing training configurations; the new return models use pre-2018 training only.

Arm	Corr	W1	Std	Abs-ACF	Tail	ADF
Original CovPW + RPC	0.026295	0.773664	0.679648	0.053772	0.003719	0.139503
GRU + RPC	0.026733	0.867717	0.960665	0.037350	0.011855	0.063363
Return + RPC	0.026423	0.743617	0.965929	0.049706	0.011579	0.160892
Return + CAL + RPC	0.026188	0.555180	0.421218	0.049794	0.016489	0.164825
Hybrid + RPC	0.116526	1.503045	1.278056	0.027103	0.072205	0.128415
Hybrid + CAL + RPC	0.109423	0.994644	0.578406	0.027533	0.046127	0.127053

On all 24 development windows, Return+CAL+RPC has 36.02% lower W1 error than GRU+RPC and wins that metric on 12/12 pairs. Its Corr error is 2.04% lower, with only 7/12 pair wins. The

small Corr margin does not establish significant superiority. Absolute-return ACF error increases by 33.32%. Tail-coexceedance error increases from 0.011855 to 0.016489, an absolute change of 0.004634. ADF rejection-rate error rises from 0.063363 to 0.164825. These costs matter for temporal and financial structure even when marginal distribution fidelity is the main goal.

Both hybrid variants have much worse Corr than the return variants. A subsequent calibration ablation also finds that the learned volatility target is not consistently better than simply using observed-context volatility. Before RPC, learned calibration has 1.55% worse W1 and 15.19% worse tail error on 2020, despite 11.24% better standard-deviation error. Thus the evidence supports investigating representation and calibration jointly; it does not isolate learned volatility prediction as the source of the gain.

F.4 How These Results Inform the Main Claim

The useful pattern is a tradeoff, not universal dominance. Correlation repair and a preserved relation define the present contribution. W1, variance, temporal dependence and tails describe the quality of the resulting generator. Correlated marginal diagnostics are not independent votes for a winner. ADF and Coint become primary application requirements when stationarity or cointegration is claimed, and tail error becomes central for risk simulation.

The anonymous numerical supplement keeps all 13 scalar-search candidates, all eight new representation/calibration arms, the six external development comparators, and every pair-level metric. It also retains the weaker hybrid and prior return variants. None of these development comparisons constitutes full same-protocol superiority over official CW-Gen or PaD-TS. The retained candidates require independently defined confirmation with task priorities fixed in advance.

G Historical Geometry and Supporting Experiments

This section retains the earlier CTSDiff-Corr branch and related screens. They explain the path to the present controller and document alternatives that were not adopted. Their kernels and objectives differ from the CSDI-CovPW generator used in the primary RPC experiments.

G.1 Whitening and matched-SPD reference theory

We first analyze an idealized matched-SPD reference process in which the same fixed $\mathbf{\Sigma}$ shapes the forward kernel, reverse covariance, and quadratic objective. This makes the coordinate and KL relationships explicit, but is not an exact characterization of frozen CTSDiff-Corr, whose kernels use $\mathbf{R}$ and whose precision loss uses $\mathbf{R}_{\cdot 1}$. We state the results for one target time column $\mathbf{x}_t \in \mathbb{R}^K$; the full sequence case follows by summing over target columns. Let $0 < \beta_t = 1 - \alpha_t < 1$, $\bar{\alpha}_0 = 1$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and assume $\mathbf{\Sigma} \succ 0$ is fixed after conditioning on the observed context.

Proposition 3 (Same-metric coordinate equivalence)**.** *Let $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top \succ 0$ be fixed after conditioning. If*

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\mathbf{L}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

then $\mathbf{y}_t = \mathbf{L}^{-1}\mathbf{x}_t$ satisfies the isotropic forward marginal

$$\mathbf{y}_t = \sqrt{\bar{\alpha}_t}\mathbf{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}.$$

Conversely, mapping this isotropic kernel back by $\mathbf{L}$ gives the colored kernel with covariance $(1-\bar{\alpha}_t)\mathbf{\Sigma}$.

Proof. Left-multiply the colored equation by $\mathbf{L}^{-1}$; the converse follows by left-multiplying the isotropic equation by $\mathbf{L}$. □

The proposition requires the identical $\mathbf{\Sigma}$ on both sides and is a statement about the Gaussian forward kernel. It neither identifies finite learned reverse functions, proves that frozen CTSDiff-Corr optimizes the matched KL derived below, nor equates the correlation-based CTSDiff-Corr and covariance-based CSDI-CovPW experiments.

Lemma 1 (Covariance-preserving forward marginal)**.** *Consider the forward kernel*

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{\Sigma}) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_{t-1}, \beta_t \mathbf{\Sigma}).$$

Then

$$q(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{\Sigma}) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{\Sigma}).$$

Proof. For $t = 1$, the statement is the definition of the kernel. Assume it holds at $t-1$. Then

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\mathbf{z}_t, \qquad \mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}),$$

and by the induction hypothesis $\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\mathbf{u}_{t-1}$, with $\mathbf{u}_{t-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. The independent noise terms give covariance $\alpha_t(1-\bar{\alpha}_{t-1})\mathbf{\Sigma} + \beta_t\mathbf{\Sigma} = (1-\bar{\alpha}_t)\mathbf{\Sigma}$, and the mean is $\sqrt{\alpha_t\bar{\alpha}_{t-1}}\mathbf{x}_0 = \sqrt{\bar{\alpha}_t}\mathbf{x}_0$. □

Proposition 4 (Exact pairwise kernel control)**.** *Let*

$$\mathbf{\Sigma}_\gamma = \begin{bmatrix} 1 & \gamma\rho \\ \gamma\rho & 1 \end{bmatrix}, \qquad |\gamma\rho| < 1.$$

Then $\mathbf{\Sigma}_\gamma \succ 0$, its eigenvalues are $1 \pm \gamma\rho$, and a Gaussian vector $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\gamma)$ has $\mathrm{Corr}(\epsilon_1, \epsilon_2) = \gamma\rho$. Consequently, the standardized residual of the conditional forward marginal has correlation $\gamma\rho$ at every diffusion step.

Proof. The eigenvectors $(1,1)^\top$ and $(1,-1)^\top$ have eigenvalues $1+\gamma\rho$ and $1-\gamma\rho$, respectively. The condition $|\gamma\rho| < 1$ makes both eigenvalues positive. Because the diagonal entries are one, covariance and correlation coincide for the two Gaussian coordinates. By the forward marginal,

$$\frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0}{\sqrt{1-\bar{\alpha}_t}} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\gamma),$$

which gives the same correlation at every t. □

This is a kernel guarantee, not an output-fidelity theorem. Covariance clipping can saturate the affine law, and the learned reverse process can produce non-monotone downstream diagnostics. The result is restricted to the declared $K = 2$ paired-generation task; basket generation requires a different relation definition and is not analyzed here.

Lemma 2 (Closed-form covariance-matched posterior)**.** *Under the same assumptions,*

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0, \mathbf{\Sigma}) = \mathcal{N}(\tilde{\boldsymbol{\mu}}_t, \tilde{\beta}_t\mathbf{\Sigma}),$$

where

$$\tilde{\boldsymbol{\mu}}_t = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t,$$
$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t.$$

Proof. By the forward marginal, the pair $(\mathbf{x}_{t-1}, \mathbf{x}_t)$ conditioned on $\mathbf{x}_0$ is jointly Gaussian. Its covariance blocks are scalar multiples of $\mathbf{\Sigma}$. Applying the conditional Gaussian formula gives the stated mean. The conditional covariance simplifies to

$$(1-\bar{\alpha}_{t-1})\mathbf{\Sigma} - \frac{\alpha_t(1-\bar{\alpha}_{t-1})^2}{1-\bar{\alpha}_t}\mathbf{\Sigma} = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\mathbf{\Sigma}.$$

□

Theorem 1 (Mahalanobis denoising objective)**.** *Fix $t \in \{2, \ldots, T\}$, so $\tilde{\beta}_t > 0$. Let the model reverse kernel use covariance $\tilde{\beta}_t\mathbf{\Sigma}$ and the standard noise-prediction mean*

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right).$$

If $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, then

$$\mathrm{KL}(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0, \mathbf{\Sigma}) \| p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{\Sigma})) = c_t(\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta)^\top \mathbf{\Sigma}^{-1}(\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta),$$

where $c_t = \beta_t^2/[2\tilde{\beta}_t\alpha_t(1-\bar{\alpha}_t)]$ is independent of θ and $\mathbf{\Sigma}$.

At $t = 1$, the posterior is a point mass and $\tilde{\beta}_1 = 0$. The reconstruction term is separate from this Gaussian KL identity.

Proof. The two Gaussian kernels have the same covariance $\tilde{\beta}_t\mathbf{\Sigma}$. Their KL divergence is therefore a quadratic form in the mean difference with precision $(\tilde{\beta}_t\mathbf{\Sigma})^{-1}$. Substituting $\mathbf{x}_0 = (\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon})/\sqrt{\bar{\alpha}_t}$ into the exact posterior mean gives the standard noise form

$$\tilde{\boldsymbol{\mu}}_t = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right).$$

Thus

$$\tilde{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_\theta = \frac{\beta_t}{\sqrt{\alpha_t(1-\bar{\alpha}_t)}}(\boldsymbol{\epsilon}_\theta - \boldsymbol{\epsilon}).$$

Substituting this expression into the KL formula gives the result. □

Proposition 5 (Stability to covariance perturbation)**.** *Let* $\mathbf{\Sigma}, \widetilde{\mathbf{\Sigma}} \succ 0$ *satisfy* $\lambda_{\min}(\mathbf{\Sigma}), \lambda_{\min}(\widetilde{\mathbf{\Sigma}}) \geq m > 0$*, and let* $d_t = 1 - \bar{\alpha}_t$*. For a shared forward mean* $\boldsymbol{\mu}_t$*,*

$$W_2\Big(\mathcal{N}(\boldsymbol{\mu}_t, d_t\mathbf{\Sigma}), \mathcal{N}(\boldsymbol{\mu}_t, d_t\widetilde{\mathbf{\Sigma}})\Big) \leq \frac{\sqrt{d_t}}{2\sqrt{m}}\|\mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}}\|_F.$$

For any residual $\mathbf{r}$*, the corresponding quadratic losses obey*

$$\left|\mathbf{r}^\top\mathbf{\Sigma}^{-1}\mathbf{r} - \mathbf{r}^\top\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{r}\right| \leq \frac{\|\mathbf{r}\|_2^2}{m^2}\|\mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}}\|_2.$$

Proof. Write $\mathbf{S} = \mathbf{\Sigma}^{1/2}$ and $\widetilde{\mathbf{S}} = \widetilde{\mathbf{\Sigma}}^{1/2}$. Coupling the two Gaussians with the same $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ gives

$$W_2^2 \leq d_t\,\mathbb{E}\|(\mathbf{S} - \widetilde{\mathbf{S}})\mathbf{z}\|_2^2 = d_t\|\mathbf{S} - \widetilde{\mathbf{S}}\|_F^2.$$

Let $\mathbf{X} = \mathbf{S} - \widetilde{\mathbf{S}}$. The identity

$$\mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}} = \mathbf{S}\mathbf{X} + \mathbf{X}\widetilde{\mathbf{S}}$$

and the Frobenius inner product imply

$$2\sqrt{m}\|\mathbf{X}\|_F^2 \leq \langle \mathbf{X}, \mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}}\rangle_F \leq \|\mathbf{X}\|_F\|\mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}}\|_F.$$

Therefore $\|\mathbf{X}\|_F \leq \|\mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}}\|_F/(2\sqrt{m})$, which proves the first bound. For the second, use the resolvent identity

$$\mathbf{\Sigma}^{-1} - \widetilde{\mathbf{\Sigma}}^{-1} = \mathbf{\Sigma}^{-1}(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma})\widetilde{\mathbf{\Sigma}}^{-1}$$

and $\|\mathbf{\Sigma}^{-1}\|_2, \|\widetilde{\mathbf{\Sigma}}^{-1}\|_2 \leq 1/m$, followed by $|\mathbf{r}^\top\mathbf{A}\mathbf{r}| \leq \|\mathbf{A}\|_2\|\mathbf{r}\|_2^2$. □

This proposition guarantees stability, not sample quality. For the matched-SPD reference, any positive eigenvalue floor gives controlled changes in both the forward Gaussian and matched quadratic loss. In frozen CTSDiff-Corr these roles are stabilized separately: $|\rho| \leq .999$ plus 10^{-6} sampler jitter for the kernels, and $\lambda = .1$ shrinkage plus 10^{-5} precision floor for the loss. Whether either design improves relational diagnostics remains an empirical question.

G.2 Validation-gated relational calibration

This subsection documents the fixed-SPRC and GammaSPRC validation gate, not the adaptive CovPW-RPC policy. The calibration view is not finance-specific. Let r_ℓ and m_ℓ denote relation and common coordinates; for positive financial paths they are log spread and log common. A relation-preserving family $\{\mathcal{T}_h : h \in \mathcal{H}\}$ keeps r_ℓ fixed and changes only centered increments of m_ℓ. The candidate set includes a true identity map h_{id} that leaves base samples unchanged.

The implemented validation gate is:

1. train the relevant base generator and generate validation samples;
2. apply every candidate $\mathcal{T}_h$ to the validation samples;

3. compute the predeclared validation score $J_{\mathrm{val}}(h)$, using the same relation/common metric families that will be reported for the dataset;
4. select $h^\star \in \arg\min_{h\in\mathcal{H}} J_{\mathrm{val}}(h)$;
5. apply only $h^\star$ to the test samples and record a pseudo-run manifest.

The gate selects the nontrivial CTSDiff-SPRC scale on the canonical and robustness financial benchmarks. The separate GammaSPRC gate covers Electricity and Traffic. Its cross-domain conclusion is conservative but useful: common-mode calibration can steer relation statistics, yet a frozen policy does not improve every domain, so base CTSDiff remains the external-validity row.

Because identity is a candidate, minimization gives $J_{\mathrm{val}}(h^\star) \le J_{\mathrm{val}}(h_{\mathrm{id}})$. The safeguard concerns validation only. It neither proves test improvement nor removes the need to report uncalibrated CTSDiff. Its role is practical: calibration may repair a relation/common trade-off, while the identity candidate lets the procedure decline unsupported intervention.

G.3 Precision-Loss Shrinkage Evidence and Sensitivity Boundaries

Table 51: Legacy exploratory precision-loss shrinkage sweep on BTC–ETH, SPY–IVV, and sh601857–sh600028 (lower is better). $\lambda = .05$ has six runs per pair versus three for other values, and SPY–IVV uses an obsolete 234-step horizon rather than the final 390; this is not a confirmatory sensitivity experiment.

λ	Mean native-unit distance	Normalized score
0	27.895	0.338
0.03	14.724	0.415
0.05	10.823	0.365
0.10	**8.001**	0.355
0.20	8.417	**0.318**

The values above are retained for transparency but do not answer the requested full-protocol sensitivity question. In particular, there is no controlled context-window sweep, no adaptive precision estimator comparison, and no final 15-pair matched-seed sweep over loss shrinkage λ or the precision floor. A later post-outcome CSI300 replay covers Rolling $W = 30, 60, 120$ and two exponentially weighted $W = 60$ definitions. Its full-coverage alternatives retain the primary direction, while $W = 30$ exposes a reference-side zero-variance domain boundary. A separate post-outcome CSI300 replay stresses a fixed 3×3 action/energy grid and two matched-budget direction controls without refitting the target model. Neither analysis varies the internal controller-window construction, ridge feature set, or nonlinear policy class.

G.4 Original Broad and Frozen Robustness Screens

Table 52 restores the complete canonical 13-method broad screen. Table 53 reports the separate frozen 15-pair robustness set under the same broad protocol. These tables retain the original-coordinate mechanism evidence and show the GAN/VAE, diffusion, and forecasting-adapted rows that remain available in the broader descriptive comparison.

Table 54 reports raw distance to real data on the canonical 15-pair benchmark. This effect-size view preserves the original scales of ADF, cointegration, and rolling-correlation diagnostics. It complements family-balanced rank and is the view in which CTSDiff-Corr's price-level structural advantage is clearest.

G.5 P0-Light Correlation-Control Screen

This screen compares context correlation, fixed train-set correlation, batch-permuted context correlation, and a random valid pair correlation inside CTSDiff-Corr. It uses BTC–ETH, SPY–IVV, and sh601857–sh600028 with the same seed and pair/horizon protocol. The screen is smaller than the main benchmark and serves attribution rather than ranking.

Table 52: Canonical 15-pair all-model diagnostic benchmark. Lower is better. Family rank gives equal weight to ADF stationarity, cointegration, and rolling-correlation families within each pair. Raw preserves original diagnostic scale; Norm. is reported as a sensitivity view. CTSDiff-SPRC is the validation-selected spread-preserving calibration of CTSDiff.

Model	Family rank	Raw	Norm.	ADF	Coint.	Corr.
CTSDiff-SPRC	**4.704**	7.223	**0.152**	10.682	10.800	0.187
CSDI	5.656	10.436	0.180	18.052	13.048	0.208
CTSDiff	5.674	**7.082**	0.231	**10.682**	**10.233**	0.332
Non-stationary Transformer	5.759	20.571	0.216	33.554	27.967	0.192
TimeVAE	6.056	17.572	0.206	21.962	30.539	0.216
TimeMixer	6.430	25.173	0.209	34.313	41.058	0.148
CTSDiff-diag	6.433	11.638	0.315	13.959	20.451	0.504
TimeXer	6.811	20.578	0.233	29.955	31.579	0.198
Autoformer	7.033	33.750	0.287	58.407	42.634	0.209
TimeGAN	8.326	36.343	0.435	56.813	51.798	0.418
iTransformer	8.904	97.843	0.499	142.688	150.703	**0.138**
Diffusion-TS	9.207	72.289	0.473	122.798	93.672	0.396
TimesNet	10.007	140.327	0.603	262.789	157.840	0.352

Table 53: Frozen 15-pair all-model robustness check. The pair set is disjoint from the canonical benchmark and balanced across cryptocurrency, U.S. equity/ETF, and CSI300 pairs. It is reported as robustness evidence, not as the final unopened confirmatory test. CTSDiff-family rows average three seeds per pair; external baseline rows follow the canonical all-model protocol. Lower is better.

Model	Family rank	Raw	Norm.	ADF	Coint.	Corr.
CTSDiff-SPRC	**4.467**	**5.252**	**0.140**	6.505	**9.111**	0.140
CSDI	5.219	6.659	0.156	8.441	11.382	0.154
CTSDiff	5.878	5.444	0.243	**6.505**	9.512	0.316
TimesNet	5.985	39.852	0.251	79.302	40.073	0.182
TimeXer	6.044	13.364	0.191	14.386	25.577	**0.129**
Non-stationary Transformer	6.098	11.080	0.185	14.423	17.728	0.188
Autoformer	6.704	12.419	0.264	21.615	15.396	0.248
CTSDiff-diag	6.748	9.441	0.286	12.215	15.710	0.398
TimeVAE	7.141	26.908	0.301	26.384	54.182	0.157
TimeMixer	7.144	19.136	0.225	26.515	30.734	0.160
TimeGAN	8.919	23.945	0.443	41.047	30.381	0.407
iTransformer	9.019	46.571	0.455	51.571	87.973	0.168
Diffusion-TS	11.231	67.128	0.781	88.230	112.281	0.404

Table 54: Raw distance-to-real aggregate on the canonical 15-pair benchmark. Lower is better. This effect-size view preserves the original scale of ADF, cointegration, and rolling-correlation diagnostics.

Model	Raw distance	Normalized score	Pairs
CTSDiff	**7.082**	0.231	15
CTSDiff-SPRC	<u>7.223</u>	**0.152**	15
CSDI	10.436	<u>0.180</u>	15
CTSDiff-diag	11.638	0.315	15
TimeVAE	17.572	0.206	15
Non-stationary Transformer	20.571	0.216	15
TimeXer	20.578	0.233	15
TimeMixer	25.173	0.209	15
Autoformer	33.750	0.287	15
TimeGAN	36.343	0.435	15
Diffusion-TS	72.289	0.473	15
iTransformer	97.843	0.499	15
TimesNet	140.327	0.603	15

The fixed global metric has the lowest family-balanced rank, while the context metric gives the lowest raw distance and the weakest rolling-correlation-family rank. The random metric is weakest overall. The result supports relational geometry but rejects a stronger necessity claim: sample-specific adaptation does not uniformly dominate every plausible control.

G.6 Dependence-Scaling Validation Boundary

CTSDiff-Corr's clearest weakness is rolling correlation, so we test gamma scaling and Dual-Geometry variants on three representative validation pairs. Stronger return-level geometry lowers rolling-correlation distance but raises raw, ADF, and cointegration distances relative to $\gamma = 1.0$. We therefore stop before a 15-pair expansion. This negative result motivates SPRC directly: adjust common-mode returns while holding the finite positive log-spread coordinate fixed, rather than replacing the price-level geometry.

Table 55: Validation-only dependence-scaling ablation on three representative pairs. Lower is better. Dual-Geometry variants substantially reduce rolling-correlation distance but fail the validation gate because they increase raw, ADF, and cointegration distances relative to the main CTSDiff setting.

Model	Family rank	Raw	ADF	Coint.	Corr.	Raw/main	ADF/main	Coint./main	Corr./main
CTSDiff, $\gamma = 1.0$ (main)	5.981	8.838	6.789	19.265	0.460	1.000	1.000	1.000	1.000
CTSDiff, $\gamma = 1.2$ sensitivity	5.019	12.271	9.960	26.404	0.447	1.388	1.467	1.371	0.972
Dual-Geometry, lowest rank	5.130	24.530	37.821	35.566	0.202	2.776	5.571	1.846	0.439
Dual-Geometry, lowest Corr.	5.537	27.830	37.858	45.455	0.177	3.149	5.576	2.359	0.385

G.7 Selected Canonical15 Component Screen

The component screen asks where relational covariance geometry is most useful inside CTSDiff. Five selected rows cover the same 15 canonical pairs: CSDI-Isotropic, CTSDiff-LossOnly, CTSDiff-SamplerOnly, CTSDiff-KernelOnly, and CTSDiff-FullComponent. ForwardOnly remains a three-pair negative control and is not expanded. The full 75-target matrix is audited for configuration identity, checkpoints, generated batches, paired fixed-slice outputs, and launch-log errors.

All 75 pair–component targets are complete, with zero launch-stamp log-error rows. The two diagnostic views measure different units. The lightweight screen summarizes whole generated slices,

Table 56: Selected-component screen on 15 canonical pairs. Lower rank is better. Wins count pairs with the lowest family rank in this five-row comparison. All 75 targets are complete, with no launch-stamp log errors. LossOnly is strongest on the lightweight diagnostic; FullComponent is strongest on the path diagnostic. Neither wins uniformly.

Component	Light rank	Light wins	Path rank	Path wins
CSDI-Isotropic	2.902	2	2.906	1
CTSDiff-LossOnly	**2.665**	**6**	3.350	1
CTSDiff-SamplerOnly	2.896	5	3.236	4
CTSDiff-KernelOnly	3.091	1	2.922	2
CTSDiff-FullComponent	3.446	1	**2.586**	**7**

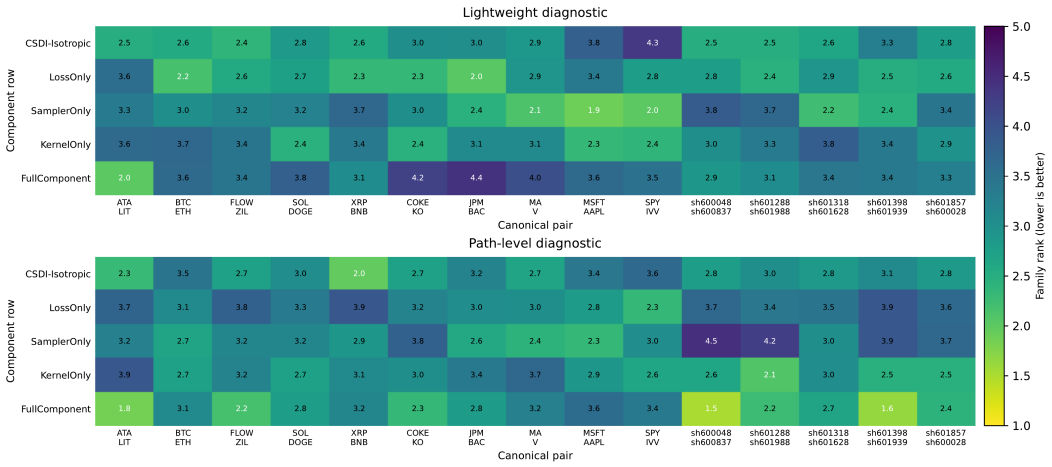

Figure 4: Canonical15 component family-rank heatmap for the selected five-row matrix. Lower ranks are better. The lightweight diagnostic favors LossOnly and SamplerOnly on more pairs, whereas the path-level diagnostic favors FullComponent and keeps the coordinate/control rows competitive on several pairs.

while the path-level screen compares each scenario path with non-overlapping real windows of equal horizon. This distinction changes the ordering: lightweight statistics favor LossOnly and SamplerOnly, whereas path-level diagnostics favor FullComponent and keep CSDI-Isotropic and KernelOnly close.

Figure 4 exposes the pair-level structure behind the aggregates. Every CTSDiff component is best on at least one pair, and FullComponent is the strongest path-level row. Yet LossOnly and SamplerOnly lead the lightweight screen, while the isotropic coordinate control remains competitive pathwise. The insertion point of geometry is therefore diagnostic-dependent. That is a mechanism result, not a winner-take-all result.

The screen is exploratory and paired by construction. We report family ranks and pair wins rather than treating 75 rows as independent samples. It supports relational preconditioning and metric-family sensitivity, but not joint necessity of every component or uniform dominance by FullComponent. A stronger claim would require a prespecified component endpoint, paired confidence intervals, and multiplicity control.

G.8 Controlled Synthetic Mechanism Diagnostic

Three predefined synthetic regimes separate a stationary OU relation, within-path regime switching, and correlated returns without mean reversion. Historical fixed SPRC leaves its CTSDiff source-path relation distances unchanged to numerical precision while reducing return-correlation distance in all nine matched seed–regime cells. CSDI nevertheless has the lowest unstandardized overall mean on

two of the three regimes, so this is bounded construct-validity evidence rather than a dominance or adaptive-RPC result.

Table 57: Controlled synthetic mechanism diagnostic over three training and generation seeds (lower is better). Relation, return-correlation, and marginal columns average four, three, and three component distances; Overall averages all ten unstandardized components. Fixed SPRC ($\eta = 1.5$) is a deterministic transform of the corresponding CTSDiff source paths and preserves their log spread up to 3.89×10^{-16}; CSDI uses separate trained/generated runs. This historical diagnostic does not evaluate guarded adaptive CovPW-RPC or establish general dominance.

Regime	Method	Relation	Return-corr.	Marginal	Overall
OU-coint.	CSDI	0.0059	0.0211	0.1294	**0.0476**
OU-coint.	CTSDiff	0.0059	0.2391	0.1711	0.1254
OU-coint.	CTSDiff-SPRC	0.0059	0.2022	0.1435	0.1061
Regime-switch	CSDI	0.0295	0.0672	0.0359	0.0427
Regime-switch	CTSDiff	0.0509	0.0819	0.0253	0.0525
Regime-switch	CTSDiff-SPRC	0.0509	**0.0322**	**0.0154**	**0.0346**
Corr-only	CSDI	0.0084	**0.0748**	0.6810	**0.2301**
Corr-only	CTSDiff	0.0104	0.2448	0.6845	0.2829
Corr-only	CTSDiff-SPRC	0.0104	0.1013	**0.6696**	0.2354

G.9 Secondary Six-Dimensional RSF Diagnostics

RSF-Eval augments ADF, cointegration, and rolling correlation with raw-path spread, tail coexceedance, and risk diagnostics. Explicit manifests provide complete coverage for the original twelve-model artifact set. Because the aggregate mixes hypothesis-test, distributional, and portfolio-risk proxy distances, it remains a secondary sanity check. CSDI ranks first, TimeVAE second, and CTSDiff third. CTSDiff has the lowest risk distance and remains strong on cointegration and spread, while its pre-SPRC rolling-correlation weakness persists.

Table 58: Secondary six-dimensional RSF diagnostics with complete raw-path coverage. Lower is better. The aggregate averages ADF, cointegration, rolling return-correlation, spread, tail coexceedance, and risk dimensions. Lowest and second-lowest values are bolded and underlined, respectively.

Model	ADF	Coint.	Corr.	Spread	Tail	Risk	Mean
CSDI	0.216	0.225	0.288	**0.197**	0.061	0.035	**0.171**
TimeVAE	**0.125**	0.233	0.301	0.318	**0.031**	0.084	<u>0.182</u>
CTSDiff	0.177	<u>0.186</u>	0.463	<u>0.208</u>	0.048	**0.017**	0.183
TimeMixer	0.215	0.273	<u>0.205</u>	0.276	0.061	0.091	0.187
Non-stationary Transformer	0.238	0.230	0.260	0.243	0.049	0.116	0.189
TimeXer	0.230	0.243	0.300	0.263	<u>0.036</u>	0.094	0.195
CTSDiff-diag	<u>0.140</u>	**0.165**	0.706	0.283	0.045	<u>0.025</u>	0.227
Autoformer	0.414	0.345	0.287	0.387	0.174	0.123	0.288
iTransformer	0.539	0.754	**0.195**	0.387	0.121	0.109	0.351
TimesNet	0.680	0.666	0.528	0.456	0.094	0.055	0.413
TimeGAN	0.448	0.449	0.602	0.455	0.715	0.282	0.492
Diffusion-TS	0.538	0.410	0.566	0.967	0.734	0.766	0.664

G.10 Downstream Spread-Reversal Sanity Check

The spread-reversal check asks whether generated paths preserve enough pair structure for a simple mean-reversion diagnostic. For each pair, we estimate the hedge coefficient on the real test path, form the log spread, and compare a one-step reversal signal between real and generated paths. Distances cover PnL distribution, signal frequency, moments, hit rate, Sharpe ratio, tail loss, and threshold-transfer regret. This is a structure check, not a deployable trading claim.

The result reproduces the main trade-off. TimeVAE and CTSDiff-diag lead this downstream aggregate. CSDI and CTSDiff are nearly tied at 0.164 and 0.165. CTSDiff has lower hit-rate distance, while CSDI better matches signal frequency. CTSDiff therefore preserves several spread-related structures without leading every downstream metric.

Table 59: Downstream spread-reversal sanity check with complete raw-path coverage. Lower is better. The score averages pair-normalized distances for reversal PnL distribution, signal frequency, PnL moments, hit rate, Sharpe, tail loss, and synthetic-to-real threshold-transfer regret. Lowest and second-lowest values are bolded and underlined, respectively.

Model	Score	PnL-W1	Signal	Hit	Sharpe	Pairs
TimeVAE	**0.104**	**0.001**	**0.061**	0.083	0.159	15
CTSDiff-diag	<u>0.117</u>	<u>0.001</u>	0.081	<u>0.059</u>	**0.141**	15
CSDI	0.164	0.001	0.093	0.081	<u>0.144</u>	15
CTSDiff	0.165	0.001	0.116	**0.043**	0.158	15
TimeXer	0.179	0.002	<u>0.077</u>	0.153	0.365	15
TimeMixer	0.182	0.001	0.091	0.151	0.399	15
Non-stationary Transformer	0.241	0.001	0.104	0.071	0.206	15
Autoformer	0.289	0.100	0.158	0.116	0.237	15
TimeGAN	0.292	0.008	0.082	0.160	0.369	15
TimesNet	0.318	0.008	0.137	0.280	0.740	15
iTransformer	0.329	0.004	0.142	0.313	0.803	15
Diffusion-TS	0.698	3.503	0.149	0.276	0.686	15

G.11 Electricity and Traffic comparisons

These supporting tables include forecasting-adapted comparators. They are not an evaluation of all listed architectures as native generative models or a modern foundation-model benchmark.

The non-financial audit covers 12 frozen Electricity pairs and 12 frozen Traffic pairs. Unlike the market benchmark, pairs are chosen from training data by generic dependence criteria. ADF and Engle–Granger statistics are not primary because the target relation is load sharing, congestion sharing, or lagged consumption rather than a financial spread equilibrium. We instead evaluate relation/common coordinates, return correlation, marginal tails, and lag profiles. The four generative rows, including CSDI, are CSDI, TimeGAN Yoon et al. (2019), TimeVAE Desai et al. (2021), and Diffusion-TS Yuan & Qiao (2024). The six forecasting-adapted rows are Autoformer, Non-stationary Transformer, TimesNet, TimeXer, TimeMixer, and iTransformer Wu et al. (2021); Liu et al. (2022); Wu et al. (2023); Wang et al. (2024b;a); Liu et al. (2024). For these rows, the forecasted target path is evaluated as a context-conditioned trajectory under the same diagnostics. Their predictive objectives differ from distributional generators, so they are descriptive sequence alternatives rather than generative equivalents.

The expanded audit adds three transparent classical controls and has exact 14-method, 12-pair, three-seed coverage per domain: 504 rows for Electricity and 504 for Traffic, with no duplicate keys or non-finite entries. Block bootstrap resamples paired temporal blocks; paired-return bootstrap resamples synchronized return pairs; common/spread AR(1) fits separate first-order autoregressions in the transformed coordinates. CTSDiff has the lowest family-balanced rank on both datasets. Relative to CSDI, mean distance is 14.1% lower with 10/12 pair wins on Electricity; on Traffic, it is 8.1% lower with 6/12 wins. Block bootstrap is competitive and has the lower Electricity mean-

distance aggregate. These checks support external validity of the base pair geometry, not universal dominance or the CTSDiff-v2 population objective.

Table 60: External-validity checks on non-financial relational pairs. We freeze 12 Electricity pairs and 12 Traffic pairs using train-only dependence statistics, run three seeds, and compare CTSDiff with CSDI, three other generative methods, and six forecasting-adapted Time-Series-Library methods. The displayed family ranks are computed within this 11-method view; the 14-method ranks with classical controls are reported separately in Table 61. Relation, return-correlation, marginal, and lag columns are family averages; lower is better.

Dataset	Method	Family rank	Mean dist.	Relation	Return corr.	Marginal	Lag
Electricity	CTSDiff	**2.29**	**0.097**	**0.010**	**0.043**	**0.285**	**0.045**
Electricity	CSDI	2.71	0.113	0.010	0.076	0.308	0.053
Electricity	Autoformer	4.75	0.149	0.030	0.113	0.365	0.088
Electricity	TimesNet	5.61	0.185	0.033	0.122	0.474	0.115
Electricity	TimeXer	5.76	0.190	0.033	0.108	0.513	0.094
Electricity	TimeVAE	5.90	0.195	0.044	0.121	0.504	0.097
Electricity	TimeMixer	6.35	0.315	0.035	0.104	0.962	0.127
Electricity	Non-stationary Transformer	6.76	0.177	0.040	0.168	0.376	0.148
Electricity	iTransformer	7.26	0.322	0.038	0.146	0.936	0.147
Electricity	Diffusion-TS	8.88	1.749	3.075	0.187	2.076	0.154
Electricity	TimeGAN	9.74	1.032	0.065	0.262	3.362	0.212
Traffic	CTSDiff	**2.78**	**0.124**	0.046	**0.052**	**0.329**	**0.038**
Traffic	CSDI	3.35	0.135	**0.031**	0.070	0.368	0.042
Traffic	Autoformer	5.10	0.300	0.049	0.102	0.908	0.079
Traffic	TimesNet	5.42	0.285	0.052	0.094	0.848	0.098
Traffic	TimeXer	5.65	0.436	0.051	0.077	1.420	0.103
Traffic	TimeMixer	5.66	0.659	0.047	0.105	2.214	0.098
Traffic	Non-stationary Transformer	5.88	0.292	0.053	0.111	0.855	0.100
Traffic	TimeVAE	6.70	0.452	0.070	0.134	1.399	0.091
Traffic	iTransformer	6.84	0.657	0.047	0.150	2.154	0.132
Traffic	Diffusion-TS	9.03	1.931	3.317	0.274	2.328	0.172
Traffic	TimeGAN	9.59	1.014	0.073	0.259	3.286	0.229

Table 60 reports the 11-method neural/generative/adapted view. The expanded table below adds the classical controls and recomputes every rank within the 14-method union. CTSDiff remains first by family-balanced rank on both datasets. Block bootstrap ranks third on Electricity and second on Traffic and has the lowest Electricity mean-distance aggregate. This disagreement is why family-balanced rank and family behavior, rather than one distance, form the primary summary.

Table 61: Classical-baseline robustness check for non-financial relational benchmarks. We add three transparent CPU baselines to the Electricity and Traffic protocols: block bootstrap, paired-return bootstrap, and common/spread AR(1). The table shows CTSDiff, CSDI, and the classical rows. Family rank is computed within the full 14-method comparison including the eleven neural/adapted rows in Table 60; lower is better.

Dataset	Method	Family rank	Mean dist.	Relation	Return corr.	Marginal	Lag
Electricity	CTSDiff	**3.06**	0.097	**0.010**	**0.043**	0.285	**0.045**
Electricity	CSDI	3.65	0.113	**0.010**	0.076	0.308	0.053
Electricity	Block bootstrap	5.12	**0.087**	0.027	0.084	**0.180**	0.054
Electricity	Common/spread AR(1)	7.74	0.337	0.031	0.078	1.061	0.161
Electricity	Paired-return bootstrap	8.23	0.346	0.099	0.076	1.006	0.160
Traffic	CTSDiff	**3.74**	**0.124**	0.046	**0.052**	**0.329**	0.038
Traffic	CSDI	4.45	0.135	0.031	0.070	0.368	0.042
Traffic	Block bootstrap	4.21	0.144	**0.021**	0.058	0.430	**0.037**
Traffic	Common/spread AR(1)	7.82	0.682	0.032	0.071	2.331	0.160
Traffic	Paired-return bootstrap	7.91	0.279	0.143	0.068	0.710	0.161

G.12 Validation-Only Pairwise Relation-Control Response

The control-response audit uses the same 12 frozen pairs per domain, five predeclared values $\gamma \in \{0.5, 0.75, 1.0, 1.25, 1.5\}$, and three seeds. It contains exactly $12 \times 5 \times 3 = 180$ rows per domain,

with no duplicate keys or non-finite responses. These inspected validation outputs test whether the implemented relation axis responds. They do not test whether a selected setting improves an unopened endpoint.

For each pair, we average generated return correlation over seeds and compute Spearman correlation across the five γ values. A positive rank response means a coefficient above 0.5, which is weaker than requiring every adjacent step to increase. Electricity responds on 11/12 pairs, with mean coefficient 0.908, 95% pair-bootstrap interval $[0.750, 1.000]$, and strict increase on 9/12 pairs. Traffic responds on 12/12 pairs, with mean coefficient 0.950, interval $[0.900, 0.992]$, and strict increase on 8/12 pairs. Mean endpoint ranges are 0.491 and 0.414, with intervals $[0.217, 0.787]$ and $[0.204, 0.640]$. The evidence establishes a useful cross-domain response coordinate while rejecting universal strict monotonicity.

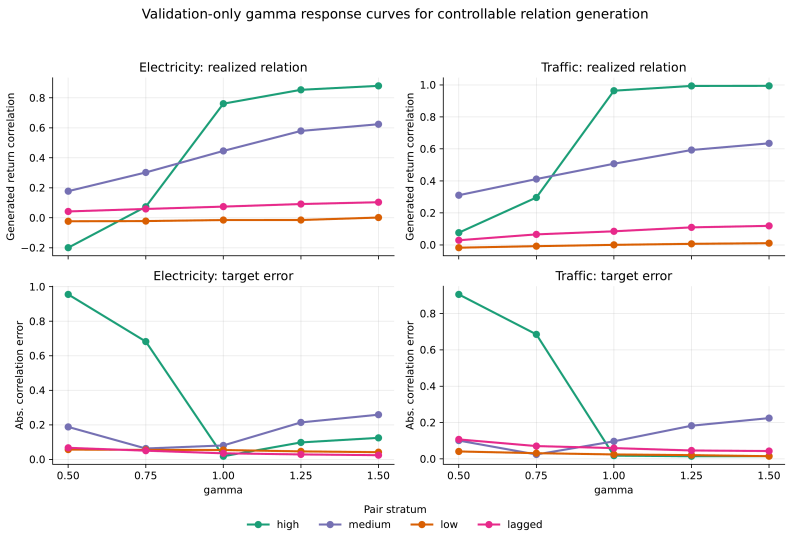

Figure 5: Validation-only response of generated pair dependence to the frozen kernel scale γ on Electricity and Traffic. Curves establish response behavior of the relation control; they are not a test-set model-selection result and do not imply that downstream fidelity improves monotonically.

G.13 Non-Financial GammaSPRC Calibration Gate

The GammaSPRC gate asks whether combined gamma control and SPRC should replace base CTSDiff as the cross-domain row. Candidate policies are selected on validation data and applied unchanged to test data. Electricity is favorable: the category policy lowers mean distance from 0.097 to 0.090 and improves family rank. Traffic is stricter. Category calibration over-corrects, while a correlation-aware policy slightly lowers mean distance and correlation error but remains behind base CTSDiff in family-balanced rank with a weak paired test. GammaSPRC therefore demonstrates controllability and dataset-specific calibration, while base CTSDiff remains the cross-domain model.

G.14 Preregistered Population-Objective Screen

We separately evaluated a population-aware CTSDiff-v2 objective on a frozen validation matrix of 15 pairs, three seeds, and matched RelPop, NoPop, and PaD arms. The analysis contained 2,970 metric cells and used pairs as inferential units. The preregistered evaluator returned `STOP_AND_CONTRACT`: family no-harm failed for cointegration and spread against NoPop and for spread against PaD. Aggregate effects and pair wins could not override this safeguard. We stopped before routing or non-financial expansion and did not open the independent test. CTSDiff-v2 remains a negative validation ablation, not an active contribution.

Table 62: Frozen non-financial GammaSPRC policy gate. Candidate policies are selected on validation categories and then applied to the test split. Gate rank is computed within the displayed CTSDiff/CSDI/GammaSPRC comparison set. The test gate improves Electricity but does not consistently replace base CTSDiff on Traffic; lower is better.

Dataset	Model	Mean dist.	Gate rank	Rank score	Corr. mean	Wins vs CTSDiff	p_{less}
Electricity	CTSDiff	0.097	4	4.00	**0.038**	N/A	N/A
Electricity	CSDI	0.113	5	4.81	0.094	2/12	0.990
Electricity	GammaSPRC-Category	**0.090**	**1**	**3.49**	0.042	7/12	0.133
Electricity	GammaSPRC-CorrCategory	0.093	3	3.80	0.040	6/12	0.485
Electricity	GammaSPRC-FamilyCategory	**0.090**	2	**3.49**	0.042	7/12	0.133
Traffic	CTSDiff	0.124	**1**	**4.13**	0.053	N/A	N/A
Traffic	CSDI	0.135	3	4.87	0.080	6/12	0.661
Traffic	GammaSPRC-Category	0.140	5	5.62	0.139	4/12	0.924
Traffic	GammaSPRC-CorrCategory	**0.124**	2	4.35	**0.044**	7/12	0.455
Traffic	GammaSPRC-FamilyCategory	0.134	4	4.92	0.080	4/12	0.883

G.15 Preregistered Projected-Safeguard Screen

We also evaluated a separately named projected-safeguard successor, CTSDiff-Guard, on six new asset-disjoint validation pairs, three seeds, and matched Base, NoGuard, and Full arms. All 54 targets completed under the frozen identity. An outcome-blind structural audit verified exactly $54 \times 38 = 2{,}052$ metric cells, equal 18-run arm coverage, finite nonnegative distances, and paired coverage before the unchanged evaluator was opened. The pair remained the inferential unit, with 10,000 pair-cluster bootstrap resamples and the preregistered seed.

The evaluator returned `STOP_AND_RETAIN_BASE_PAPER`. For the useful-control comparison, Full versus Base, the Corr-family mean log ratio was -0.0313, but its one-sided 95% upper bound was 0.3356; the 4/6 pair-win condition alone could not override this uncertainty failure. For Full versus NoGuard, the Coint safeguard passed. Spread failed because one pair's mean log ratio 0.2554 exceeded the frozen $\log(1.25) = 0.2231$ maximum-harm bound. Common had mean log ratio 0.0393, one-sided upper bound 0.1653, 2/6 pair wins, and maximum pair ratio 0.3095, so it failed all three frozen safeguard conditions.

These results show a coordinate-specific and pair-heterogeneous validation failure, not an independent-test result or an equivalence finding. We stopped before expansion, routing, post-result tuning, or any test split. CTSDiff-Guard is therefore negative validation evidence about the difficulty of simultaneous relation-family control and is not a method, performance, or novelty contribution of the active paper.

G.16 Qualitative Diffusion Trajectory

Figure 6 visualizes one BTC–ETH validation window. The forward row follows the exact CTSDiff noising equation with a fixed noise draw and the trained schedule. The reverse row records sampler states from the denoising chain rather than reconstructing them from final outputs. The figure therefore shows the process the implementation actually runs, while remaining a qualitative sanity check.

A companion dense process explorer records 50 forward snapshots and 50 reverse sampler states for the same validation example. It overlays the matched target and visualizes covariance ellipses, spread and first-difference histograms, and a PCA process map. These panels are qualitative; no main benchmark number is computed from them.

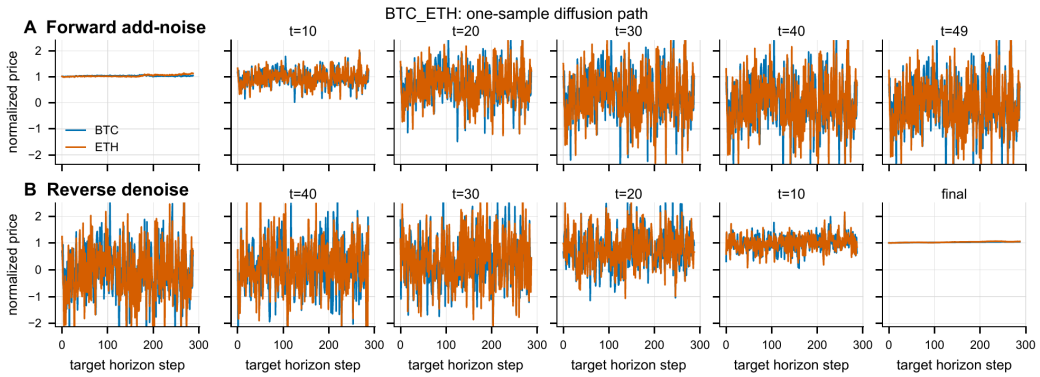

Figure 6: Forward add-noise and reverse denoise trajectories for one BTC–ETH sample. Each panel shows the generated target horizon for both assets at a selected diffusion step. The final reverse panel is the sampled target path produced by the model.

H Reproducibility Artifacts

Code and numerical evidence. The compilation archive supplies the manuscript, tables, figures and bibliography. Separate anonymous bundles supply controller code, evaluation scripts, pair-level reductions, bootstrap replicates and SHA-256 manifests. The primary fit–apply reference runs with NumPy; the distribution-targeted extension uses PyTorch and supports CPU and MPS. Both synthetic examples run from an isolated directory without the original project on the import path. These examples check implementation, not empirical efficacy.

The weather release supplies its frozen 37-file code root and a fetcher that verifies the 80 annual NOAA input hashes. For the financial controller, the executable reference restores two fitting dependencies omitted from the earlier compact weather package. Their bytes match earlier source manifests, but the original weather inventory did not bind these external files; the provenance record preserves that gap. A saved-cell replay reproduces all 475 actions within 2.23×10^{-16} and output paths within 8.89×10^{-16}.

The evidence bundles retain complete arm summaries, failures and selection rules for the tests in Sections D–F. These include all 324 simple-target endpoints, 432 utility reductions and 180 GRU temporal arm–cell summaries. Baseline ports, amended eligibility and retrospective studies retain their original scope. Artifact manifests carry the detailed execution history; the tables report the scientific comparisons.

What is not included. Compact public bundles omit raw financial arrays, generated trajectories and generator checkpoints. The 36 original financial checkpoints have been recovered and matched to historical hashes locally. Full public training still needs documented data access and a portable entry point; the historical launcher binds a particular CUDA runtime and launch ledger. The supplied code reproduces the core controller on user-supplied samples and the aggregate reductions, not full financial training from scratch.

Names in saved records. We use functional names in the main text. Table 63 maps them to immutable artifact labels. Earlier figures retain their original labels; these are aliases, not additional methods. Table 64 separately distinguishes evaluated adaptations from related methods that were not evaluated.

Table 63: Functional names and immutable artifact aliases. The aliases identify saved results; they do not define additional contributions.

Name used in the paper	Saved aliases	Role
RPC (interval policy)	CovPW-RPC-Interval; CRPC-Interval; `CRPC_INTERVAL`	Primary bounded inverse
Guarded RPC	CovPW-RPC-v10; CTSDiff-RPC	Frozen system comparison
Point-target RPC	CovPW-RPC-Point	Primary-policy ablation
Whitened CSDI	CSDI-CovPW; CSDI-prewhiten	Primary base generator
Learned-covariance CSDI	CSDI-JMCE-CW	Matched-backbone adaptation
Corr-shaped + fixed	CTSDiff-SPRC	Supporting comparison branch
Correlation-shaped generator	CTSDiff-Corr; CTSDiff	Earlier geometry studies
Common-increment edit	SPRC	Earlier name for the shared transform

Table 64: Related interfaces and implementation status. This is not a performance table. The rows are not efficacy-matched: they differ in intervention locus, required model access, and controlled or protected object. "Not evaluated" means that no method-native result for the cited method is reported in this paper.

Family / examples	Intervention locus	Model access	Controlled or protected object	Evidence boundary in this paper
Reverse-step projection (PDM; CPS) Christopher et al. (2024); Narasimhan et al. (2025)	Projection within reverse denoising; CPS acts on clean estimates	Reverse loop, denoiser outputs, and a constraint/projection operator	Constraint-set feasibility; CPS targets hard time-series constraints	A PDM-semantics-backed PDM–CSDI port is evaluated on the same Corr endpoint; it is not an official-code reproduction. The CPS-style adaptation remains strict-training qualification only.
Split/primal–dual constrained sampling (SAL/CASAL) Blanke et al. (2026)	Auxiliary-variable primal–dual updates within Langevin sampling	Score or energy gradients and explicit constraint functions	Strict mathematical or physical feasibility	Related work only; no method-native implementation or result is reported.
Multi-model compositional projection (PCD) Luan et al. (2026)	Coupled guidance and projection at reverse steps	Two separately pretrained diffusion models and a differentiable joint cost	A joint cost or constraint between separately generated outputs	Related work only; its two-model assumption differs from the joint CSDI backbone.
Initial-noise / global-statistic control (CCS) Song et al. (2025)	Initial-noise perturbation before diffusion-ODE decoding	ODE sampler and a noise-to-output controller	Individual-sample proximity or a sample-mean target	Related work only; no time-series adaptation or relation-path certificate is evaluated.
Post-sampling relation-preserving control (CovPW-RPC)	One bounded terminal edit after joint sampling	A generated joint sample and a train-fitted policy; no denoising-loop access	Corr response through the common coordinate; declared log-spread path is protected	Evaluated here. Exact certificates cover the declared relation and action; downstream Corr, ADF, and Coint remain empirical.
Learned dependence modeling (CW-Gen / CSDI-JMCE-CW) Yang et al. (2026)	Conditional mean/covariance learning in the generator	A trainable backbone, or the matched-backbone adaptation used here	Joint conditional means and sliding-window covariances; no explicit protected action coordinate	CSDI-JMCE-CW is an adapted mechanism baseline, not an official CW-Gen reproduction or a control contract.