QPATH: JOINT ANOMALY DETECTION FROM MARGINAL QUANTILE FORECASTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series foundation models increasingly use direct quantile forecasting to provide efficient uncertainty estimates without specifying a parametric predictive distribution. Yet marginal quantiles leave dependence across horizon steps unspecified: a path can be ordinary at individual steps while being jointly anomalous. QPath addresses this gap at the detection stage without changing the forecaster or constructing a joint predictive distribution. It combines the mean quantile loss across forecast quantiles with a temporal direction factor estimated from historical median residuals, followed by reference normalization and channel averaging. We evaluate QPath on synthetic anomaly patterns and the univariate and multivariate TSB-AD benchmarks using Chronos-2 and TimesFM 2.5. Across zero-shot and head-only fine-tuned settings, average relative improvements in whole-series VUS-PR over a quantile-loss-only baseline (quantile error) are 22.1% on TSB-AD-U and 4.3% on TSB-AD-M; the univariate absolute gains are 0.069–0.074. Multivariate gains remain inconclusive under source-cluster inference. Removing explicit cross terms eliminates the mean univariate gain in a matched diagonal control; replacing absolute error with quantile loss adds 0.003–0.010 on TSB-AD-U.

1 INTRODUCTION

Early time-series foundation models produced point forecasts, as in the first TimesFM and Timer, or masked reconstructions, as in MOMENT (Das et al., 2024; Liu et al., 2024; Goswami et al., 2024). Probabilistic models represent uncertainty through parametric distributions, mixtures, or categorical distributions over quantized values (Rasul et al., 2024; Woo et al., 2024; Cohen et al., 2024; Ansari et al., 2024). More recently, Chronos-2, TimesFM 2.5, Moirai 2.0, and Toto 2.0 have adopted direct quantile outputs (Ansari et al., 2025; Google Research, 2025; Liu et al., 2025; Khwaja et al., 2026). Quantile regression avoids specifying a parametric predictive distribution (Koenker & Bassett, 1978), while direct multi-step heads avoid sequential trajectory sampling (Amazon Web Services, 2024). Moirai 2.0 and Toto 2.0 explicitly replace mixture-distribution outputs with quantile heads (Liu et al., 2025; Khwaja et al., 2026).

This shift raises an important question for forecasting-based anomaly detection, which scores discrepancies between forecasts and subsequent observations (Malhotra et al., 2015; Hundman et al., 2018; Deng & Hooi, 2021). Recent forecasting reports explicitly motivate quantile outputs for such applications: the Chronos-2 report highlights extreme quantiles for rare-event coverage and anomaly detection (Ansari et al., 2025). But do better-calibrated marginal forecasts supply the information needed to detect a jointly anomalous temporal pattern?

The gap is structural, not merely a matter of forecast accuracy. By Sklar's theorem, even exact continuous marginal distributions can be combined with different copulas to yield different joint laws (Sklar, 1959). Thus per-step quantiles alone do not identify dependence across the forecast horizon. This concerns the output representation, not an absence of temporal modeling inside the forecaster. A path can be ordinary at every step yet violate the temporal relationships expected under normal behavior. For example, residual paths (b, b) and $(b, -b)$ have identical coordinatewise absolute errors and, with forecast quantiles symmetric about the median on a symmetric probability grid, identical mean quantile loss. If same-sign residuals are common and opposite-sign residuals rare, the two paths differ in how typical they are jointly. Better marginal calibration alone cannot resolve this distinction (Figure 1).

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

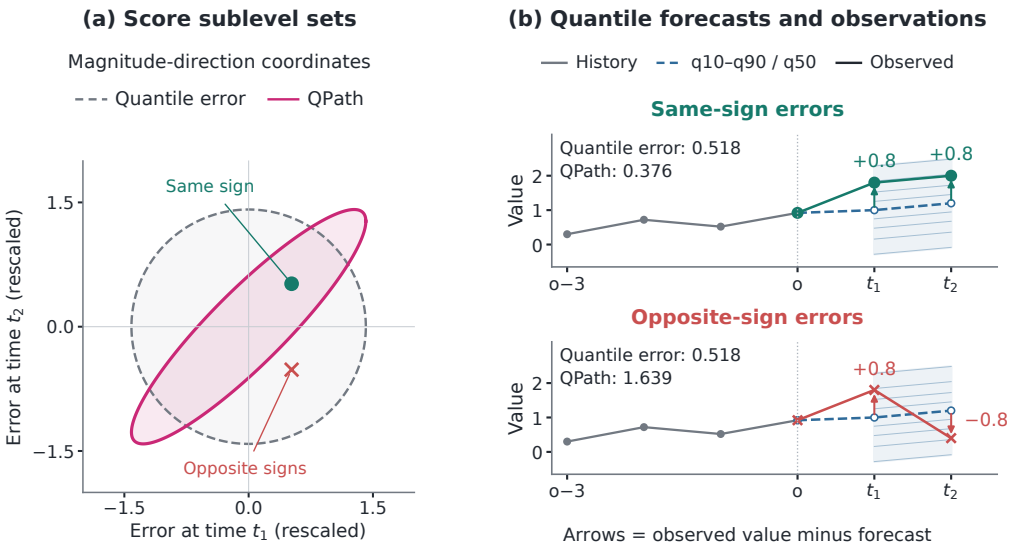


Figure 1: **Equal quantile errors, different temporal configurations.** (a) Quantile error (circle) and QPath (ellipse) in magnitude–direction coordinates $x = \sqrt{2}\, g\, v/\|v\|$ from Equation (5). The axes correspond to $t_1 = o + 1$ and $t_2 = o + 2$; values are rescaled coordinates, not raw median residuals. Here $v = e$ and $P = \left(\begin{smallmatrix} 1 & 0.9 \\ 0.9 & 1 \end{smallmatrix}\right)^{-1}$ is stipulated. Score sublevel sets share an illustrative cutoff, not a matched probability. (b) Constructed observations share history and forecasts $Q_h(\tau) = \widehat{y}_h + \Phi^{-1}(\tau)$, with $\widehat{y} = (1.0, 1.2)$ and $\tau = 0.1, \ldots, 0.9$. Residuals $(0.8, 0.8)$ and $(0.8, -0.8)$ give equal quantile loss but different QPath scores, shown inside the panels. Appendix A.3 derives this analytic illustration.

QPath addresses this missing-dependence problem directly at the detection stage. Mean quantile loss uses the common q10–q90 grid, while a precision-matrix-based direction factor evaluates the signed configuration of the median residual path against a historical reference. Their product combines marginal forecast error with temporal structure. The method changes neither the forecasting model nor its predictions and requires no anomaly-trained head or future-path sampling.

Here, *joint* refers to joint assessment across horizon steps within an observed residual path. Quantile loss supplies magnitude; median residuals supply the direction and historical precision. We do not fit a covariance matrix over all quantile levels or identify the current origin's full conditional joint law. On multivariate data, each channel receives the same temporal scoring rule, followed by reference normalization and averaging. Cross-channel information may enter the forecaster, but the scorer does not fit cross-channel covariance.

Our contributions are a quantile-aware magnitude–direction score, its geometric interpretation, and matched-forecast evaluation across synthetic patterns and both TSB-AD benchmarks. Univariate improvements over quantile error are supported across backbones, forecasting settings, and evaluation periods; multivariate gains are smaller and source-dependent. A matched diagonal control tests explicit temporal cross terms on both benchmarks. A four-way magnitude ablation distinguishes temporal weighting from replacing absolute-error magnitude with quantile loss.

2 Related Work

Marginal forecasts and joint paths. Direct quantile heads describe individual horizon steps without specifying a distribution over future paths. Autoregressive probabilistic models instead generate trajectories through sequential sampling (Rasul et al., 2024; Ansari et al., 2024). Quantile-copula models and correlated-path generation add dependence at forecast time (Wen & Torkkola, 2019; Baron et al., 2025). Energy and variogram scores evaluate multivariate predictive distributions; the latter explicitly compare pairwise increments (Gneiting & Raftery, 2007; Scheuerer & Hamill,

2015). They require a joint predictive law, joint samples, or the relevant joint expectations, rather than only marginal quantiles.

Forecasting-based anomaly detection. Prediction errors underpin recurrent and convolutional detectors (Malhotra et al., 2015; Hundman et al., 2018; Munir et al., 2019), online sequence-memory methods (Ahmad et al., 2017), and graph-based multivariate detectors (Deng & Hooi, 2021; Zhao et al., 2020). TSB-AD compares these approaches with statistical and reconstruction-based alternatives (Liu & Paparrizos, 2024); Blázquez-García et al. (2021) survey this broader literature. In an exploratory six-model comparison, Kim et al. (2026) associate lower symmetric mean absolute percentage error (sMAPE) with higher volume under the precision–recall surface (VUS-PR), within a study primarily focused on reconstruction-based detection.

Scoring and calibrating forecast errors. Telemanom smooths errors and selects nonparametric thresholds (Hundman et al., 2018); hierarchical temporal memory (HTM) compares short- and long-term error statistics (Ahmad et al., 2017); the graph deviation network (GDN) normalizes each sensor's error before taking a sensor maximum (Deng & Hooi, 2021). Malhotra et al. (2015) fit a multivariate Gaussian to errors for the same target predicted from different origins. QPath instead scores the multi-step path from one origin, using precision-matrix weighting for *direction* and mean quantile loss for magnitude. More recently, the 1-Wasserstein adaptive conformal anomaly score ($\mathcal{W}_1$-ACAS) uses pretrained forecasters without additional fine-tuning and adaptively conformalizes horizon-specific errors using weighted empirical quantiles (Martinez Gil et al., 2026). Its emphasis is calibrated tail probabilities and alarm control under stated assumptions; QPath instead ranks completed residual paths through joint temporal geometry.

3 QPATH: QUANTILE ERROR WITH TEMPORAL DIRECTION

3.1 MEAN QUANTILE LOSS

For origin o, channel $c \in \{1, \dots, C\}$, and horizon step $h \in \{1, \dots, H\}$, let $Q_{o,c,h}(\tau)$ denote a forecast quantile and $Y_{o,c,h}$ the subsequently observed value. For $\mathcal{T} = \{0.1, 0.2, \dots, 0.9\}$, define the pinball loss ψ_τ and mean quantile loss g as

$$\psi_\tau(\epsilon) = \epsilon\big(\tau - \mathbf{1}\{\epsilon < 0\}\big), \qquad g_{o,c} = \frac{2}{H|\mathcal{T}|} \sum_{h=1}^{H} \sum_{\tau \in \mathcal{T}} \psi_\tau(Y_{o,c,h} - Q_{o,c,h}(\tau)) \,. \tag{1}$$

The factor two preserves the scale of absolute error when only the median is used. It also gives $g_{o,c} = a_{o,c}$ if all forecast quantiles collapse to the median, where $a_{o,c} = H^{-1} \sum_h |Y_{o,c,h} - Q_{o,c,h}(0.5)|$. Unlike the MAE $a_{o,c}$, $g_{o,c}$ uses deviations from all nine forecast quantiles. It can remain positive when median error is zero. Although quantile loss uses marginal uncertainty, its sum over horizon steps does not itself encode their joint configuration.

This finite-grid mean quantile loss is related to, but not identical to, the continuous ranked probability score (CRPS) (Gneiting & Raftery, 2007; Alexander et al., 2024). Appendix A.4 distinguishes its weights and reference normalization from CRPS and normalized weighted quantile loss.

3.2 HISTORICAL RESIDUAL DIRECTION

The direction factor uses the signed median residuals

$$e_{o,c,h} = Y_{o,c,h} - Q_{o,c,h}(0.5), \qquad v_{o,c,h} = \frac{e_{o,c,h} - \mu_{c,h}}{s_{c,h}}, \qquad v_{o,c} = (v_{o,c,1}, \dots, v_{o,c,H})^\top. \tag{2}$$

The reference $\mathcal{R}$ contains origins whose complete context and target lie in the reference prefix. It is not filtered by anomaly labels and is fitted separately for every series and forecasting setting. Reference medians and median absolute deviations determine $\mu_{c,h}$ and $s_{c,h} > 0$; Appendix B specifies the numerical convention.

For $n = |\mathcal{R}|$ reference paths, form

$$\widehat{V}_c = \frac{1}{n-1} \sum_{o \in \mathcal{R}} (v_{o,c} - \bar{v}_c)(v_{o,c} - \bar{v}_c)^\top, \qquad A_c = 0.9\widehat{V}_c + 0.1 I_H, \qquad P_c = A_c^{-1}. \tag{3}$$

The sample mean $\bar{v}_c$ centers only the covariance; it is not additionally subtracted from the transformed query. Each channel retains an $H \times H$ temporal matrix. This historical reference is not a prediction of the current origin's conditional covariance, nor is it refitted using a vector of quantile losses.

3.3 Joint scoring and normalization

For nonzero v, QPath combines magnitude and direction as

$$\rho_c(v) = \sqrt{\frac{v^\top P_c v}{v^\top v}}, \qquad \boxed{m_{o,c} = g_{o,c}\, \rho_c(v_{o,c}).} \tag{4}$$

At $v = 0$ the neutral rule is $\rho_c(0) = 1$, so $m = g$, including when $g > 0$. The factor can be below or above one, scaling the quantile loss according to temporal direction.

For a fixed channel, suppress the channel and origin indices and define magnitude–direction coordinates for $v \neq 0$

$$x = \sqrt{H}\, g\, \frac{v}{\|v\|}, \qquad g = \frac{\|x\|}{\sqrt{H}}, \qquad m(g, v) = \sqrt{\frac{x^\top P x}{H}}. \tag{5}$$

In these coordinates, scalar score cutoffs define a sphere for quantile error and an ellipsoid for QPath (Figure 1). This algebraic representation does not imply a Gaussian likelihood or calibrated probability. Appendix A gives the score properties and explains the neutral zero-direction case.

Both QPath and the quantile error baseline use a floored channel reference mean as the reference denominator:

$$\kappa_c(s) = \max\left\{ \frac{1}{n} \sum_{o' \in \mathcal{R}} s_{o',c}, 10^{-6} \right\}, \qquad S_o(s) = \frac{1}{C} \sum_{c=1}^{C} \frac{s_{o,c}}{\kappa_c(s)}. \tag{6}$$

QPath sets $s = m$; the *quantile error* baseline sets $s = g$. Each denominator is fitted from the same checkpoint's reference, not its test scores. At $C = 1$, normalization preserves the raw-score ranking. At $C > 1$, each method's denominator also determines channel weights, so their difference is not a direction-only causal effect.

A path score is assigned to its last observed target, $o + H$. Between sampled test endpoints, the last completed score is held forward; Appendix B gives the offline warm-up convention. Accumulating the temporal covariance and factorizing it cost $O(nCH^2 + CH^3)$; scoring each path costs $O(CH^2 + CH|\mathcal{T}|)$. Quadratic forms are computed with triangular solves rather than an explicit matrix inverse.

4 Experimental setup

4.1 Data and forecasting settings

We evaluate the official TSB-AD evaluation subsets: TSB-AD-U (U) contains 350 univariate series from 23 sources and TSB-AD-M (M) contains 180 multivariate series from 17 sources (Liu & Paparrizos, 2024). The benchmark-provided training segment is the reference prefix. *Full* evaluates the complete series, including the reference prefix; *Test* evaluates only the suffix. Full is primary to match the whole-series scope of the TSB-AD benchmark; Test assesses performance outside the reference prefix. One U test suffix contains no positive anomaly label, leaving 349 evaluable Test series. M has 180 in both periods.

Chronos-2 and TimesFM 2.5 use context length 256 and forecast horizon 64; TimesFM in tables denotes version 2.5 unless stated otherwise. Real-data experiments use zero-shot (ZS) and head-only fine-tuned (FT) forecasts, labeled Zero-shot and Fine-tuned in tables. FT averages the metrics of seeds 42, 43, and 44 *within each series*; it is not a prediction ensemble or an enlarged paired sample. Checkpoints are selected by prefix median-forecast MAE when validation is available, otherwise by a fixed update budget. No anomaly label or QPath score selects a checkpoint. On M, Chronos-2 conditions jointly on all channels and TimesFM 2.5 forecasts channels separately. These modes are not a controlled test of conditioning. Within each forecasting setting, every scorer receives identical forecasts and the same reference origins. Scoring uses q10–q90 because this is the common output grid of both models; Chronos-2's additional extreme quantiles are outside this comparison.

The synthetic study uses Gaussian AR(1) noise (coefficient 0.8) and five anomaly types, with 40 series per type: spikes, level shifts, trend changes, seasonal shifts, and variance bursts, totaling 200 series. A further 160 series under the same noise process vary the strength of seasonal shifts and trend changes to 0.5 or 2.0. Synthetic forecasts are ZS only. The generators, seeds, and anomaly definitions are in Appendix D.

4.2 Evaluation and statistical reporting

We report VUS-PR with automatic window selection (AUTO VUS-PR) (Paparrizos et al., 2022), using a fixed evaluator revision and no point adjustment. Reference origins use stride one; test scoring uses stride one on U and synthetic data, and stride 16 with forward hold on M. Only the published-detector comparison uses the published-window rule to align evaluation with the published values (Appendix F). Appendix B specifies time mapping and software revisions.

We use two-sided series-paired t tests and report per-comparison p values and nominal 95% intervals. Real-data comparisons additionally use 10,000-replicate source-cluster bootstrap intervals, targeting an equal-weight mean over series. Synthetic tests use $N = 40$ within each type–strength–backbone setting. The abstract averages the four setting-level relative gains in Full VUS-PR across two backbones and ZS/FT, not per-series percentage changes. These analyses describe benchmark-informed method comparisons; neither FT nor Test is an independent-data confirmation.

5 Results

5.1 Synthetic anomaly shapes under Gaussian AR(1) noise

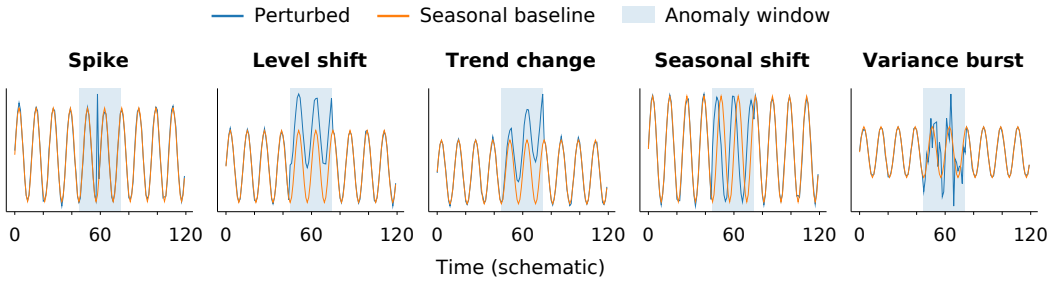


Figure 2: **Synthetic anomaly patterns (schematic).** A seasonal baseline and perturbed observations illustrate the five types; shading marks the anomaly window. These explanatory curves are not model forecasts or experimental trajectories. Generation details are in Appendix D.

Table 1: **Synthetic anomaly detection under Gaussian AR(1) noise.** Zero-shot Full AUTO VUS-PR; AR coefficient 0.8, 40 series per type.

Anomaly type	Chronos-2		TimesFM	
	Quantile error	QPath	Quantile error	QPath
Spike	0.061	**0.264**$^{\uparrow}$	0.038	**0.300**$^{\uparrow}$
Level shift	0.412	**0.507**$^{\uparrow}$	0.453	**0.529**$^{\uparrow}$
Trend change	**0.334**	0.318	**0.347**	0.341
Seasonal shift	0.422	**0.476**$^{\uparrow}$	0.469	**0.513**$^{\uparrow}$
Variance burst	0.806	**0.931**$^{\uparrow}$	0.838	**0.929**$^{\uparrow}$

$^{\uparrow}$/$^{\downarrow}$ indicate a significant gain/loss, respectively, under two-sided paired tests with nominal $p < 0.05$ within each type and backbone. Test and Gaussian-noise strength-sweep results are in Appendix D.

The synthetic study tests whether temporal weighting improves anomaly discrimination under identical forecasts, rather than comparing forecasting accuracy. Figure 2 illustrates the five anomaly types, and Table 1 compares the resulting detection performance. Under Gaussian AR(1) noise,

QPath improves on spikes, level shifts, seasonal shifts, and variance bursts in Full evaluation for both backbones, with nominal $p < 0.05$. Trend-change means are lower, with $p \geq 0.05$.

The Gaussian gains also occur in Test evaluation. In the strength sweep, half-strength seasonal shifts improve with both backbones at nominal $p < 0.05$. Chronos-2 has negative double-strength trend-change effects in both periods and a negative half-strength trend-change effect in Test at the same criterion. Using mean quantile loss as the magnitude does not prevent shape-dependent failures of temporal weighting. Appendix D gives every type–strength contrast.

5.2 UNIVARIATE RESULTS

Table 2: **Univariate results on TSB-AD-U.** Mean AUTO VUS-PR under identical forecasts; higher is better.

Backbone	Forecast setting	Full series		Test suffix	
		Quantile error	QPath	Quantile error	QPath
Chronos-2	Zero-shot	0.318	**0.389**	0.332	**0.402**
	Fine-tuned	0.326	**0.395**	0.340	**0.408**
TimesFM	Zero-shot	0.322	**0.394**	0.337	**0.407**
	Fine-tuned	0.326	**0.400**	0.340	**0.413**

Full/Test: 350/349 series. FT averages three seeds within each series. Table 13 gives six-decimal means. Throughout, bold marks the best displayed mean per method comparison (ties included), not significance.

Table 2 compares QPath with quantile error under identical forecasts. On U, QPath improves Full VUS-PR in all four backbone–forecast settings, by 0.069–0.074. The average setting-level relative gain is 22.1%. With fine-tuned forecasts, Full means are 0.395 for Chronos-2 and 0.400 for TimesFM 2.5, versus 0.326 for quantile error with either backbone. Test means reach 0.408 and 0.413, with gains of 0.068–0.073 across all four settings.

All eight U contrasts across Full and Test have positive paired mean differences supported by both inference procedures (Appendix C). Thus the observed gains are not confined to one backbone, forecasting setting, or prefix-including evaluation. This supports the combined quantile-loss and temporal-direction score.

5.3 MULTIVARIATE RESULTS AND SOURCE VARIATION

Table 3: **Multivariate results and paired differences.** AUTO VUS-PR on all 180 M series.

Backbone	Forecast setting	Evaluation period	Quantile error	QPath	Δ	p
Chronos-2	Zero-shot	Full	0.348	**0.367**	+0.019	0.0057
		Test	0.360	**0.374**	+0.014	0.0315
	Fine-tuned	Full	0.351	**0.368**	+0.017	0.0151
		Test	0.363	**0.375**	+0.012	0.0663
TimesFM	Zero-shot	Full	0.344	**0.353**	+0.008	0.2650
		Test	0.356	**0.360**	+0.004	0.5043
	Fine-tuned	Full	0.347	**0.362**	+0.016	0.0477
		Test	0.358	**0.370**	+0.011	0.1127

Δ is QPath minus quantile error. Tests are two-sided and nominal; all eight source-cluster intervals include zero (Appendix C). Means and differences are rounded separately.

On M, the average Full relative gain is 4.3%, and all eight Full/Test means favor QPath. However, all source-cluster intervals include zero. For ZS, Chronos-2 has $p < 0.05$ in both periods, whereas TimesFM 2.5 does not. Under FT, both Full paired tests have $p < 0.05$, whereas neither Test contrast does (Table 3). We report the positive mean differences without claiming source-wide superiority or equivalence.

Source diagnostics show why a single mean is insufficient. Under FT, 11 of 17 M sources have positive mean differences for Chronos-2 and nine for TimesFM 2.5. Exathlon, SMAP, and CATSv2

improve, while several ECG-related sources decline. Appendix G reports every source, seed, and per-series win/loss count, together with auxiliary detection metrics.

5.4 MATCHED DIAGONAL CONTROL

To isolate the explicit query cross terms, we hold mean quantile loss g, transformed residuals v, the full precision matrix P, and the QPath reference denominators fixed, and score

$$m^{\mathrm{diag}}_{o,c} = g_{o,c}\sqrt{\frac{\sum_h P_{c,hh} v^2_{o,c,h}}{v^\top_{o,c} v_{o,c}}}, \qquad S^{\mathrm{diag}}_o = \frac{1}{C}\sum_c \frac{m^{\mathrm{diag}}_{o,c}}{\kappa_c(m)}. \tag{7}$$

The zero-direction rule is unchanged. Diagonal entries come from the same full precision matrix P, not a separately fitted diagonal covariance. This removes explicit query cross terms without changing channel weights; it does not remove every dependence-related feature of the fitted reference.

Table 4: **Matched diagonal control.** Full AUTO VUS-PR.

Data	Backbone	Forecast setting	Quantile error	Diagonal	QPath	Δ	p
U	Chronos-2	Zero-shot	0.318	0.316	**0.389**	$+0.073$	$< 10^{-15}$
		Fine-tuned	0.326	0.324	**0.395**	$+0.070$	$< 10^{-15}$
	TimesFM	Zero-shot	0.322	0.319	**0.394**	$+0.075$	$< 10^{-15}$
		Fine-tuned	0.326	0.324	**0.400**	$+0.076$	$< 10^{-15}$
M	Chronos-2	Zero-shot	0.348	0.350	**0.367**	$+0.017$	0.0002
		Fine-tuned	0.351	0.353	**0.368**	$+0.015$	0.0012
	TimesFM	Zero-shot	0.344	0.347	**0.353**	$+0.005$	0.3021
		Fine-tuned	0.347	0.349	**0.362**	$+0.014$	0.0215

Δ is QPath minus Diagonal; p is the nominal paired value. Diagonal and QPath share the same mean quantile loss, full precision matrix P, and QPath reference denominator. U: $N = 350$; M: $N = 180$. Full/Test paired and source-cluster intervals are in Appendix C.3. Means and differences are rounded separately.

In U, QPath exceeds this control in all eight backbone–forecast–period settings, by 0.069–0.076, with all nominal $p < 10^{-15}$ and positive source-cluster intervals (Table 4; Appendix C.3). In M, Chronos-2 has nominal $p < 0.004$ in all four settings; its FT Test source interval includes zero. For TimesFM 2.5, only FT Full has $p < 0.05$ ($p = 0.0215$), and all four source intervals include zero. Explicit cross terms therefore have strong support in U, while their multivariate benefit depends on the forecasting setting and source composition.

Figure 3 shows the two main scores under identical forecasts. QPath remains elevated across the labeled interval, whereas quantile error has a brief peak. The full post-anomaly response is also shown; Appendix C.4 gives the selection rule and numerical details.

5.5 CONTEXT AMONG PUBLISHED FORECASTING-BASED DETECTORS

Table 5: **Forecasting-based detector context.** Published-window Full AUTO VUS-PR.

Detector	Forecaster	TSB-AD-U	TSB-AD-M
LSTMAD	trained LSTM	0.325	0.307
CNN	trained CNN	0.343	0.313
Lag-Llama	Lag-Llama (Zero-shot)	0.273	–
Chronos	Chronos (Zero-shot)	0.272	–
TimesFM	TimesFM 1.0 (Zero-shot)	0.300	–
Quantile error*	Chronos-2, TimesFM 2.5	0.333	0.357
QPath*	Chronos-2, TimesFM 2.5	**0.407**	**0.371**

*Mean across the four backbone–forecast settings, not one model run. Published rows are from the fixed leaderboard snapshot. Our rows use the published-window rule; they are not interchangeable with Table 2. A dash means not reported. Appendix F records the comparison assumptions.

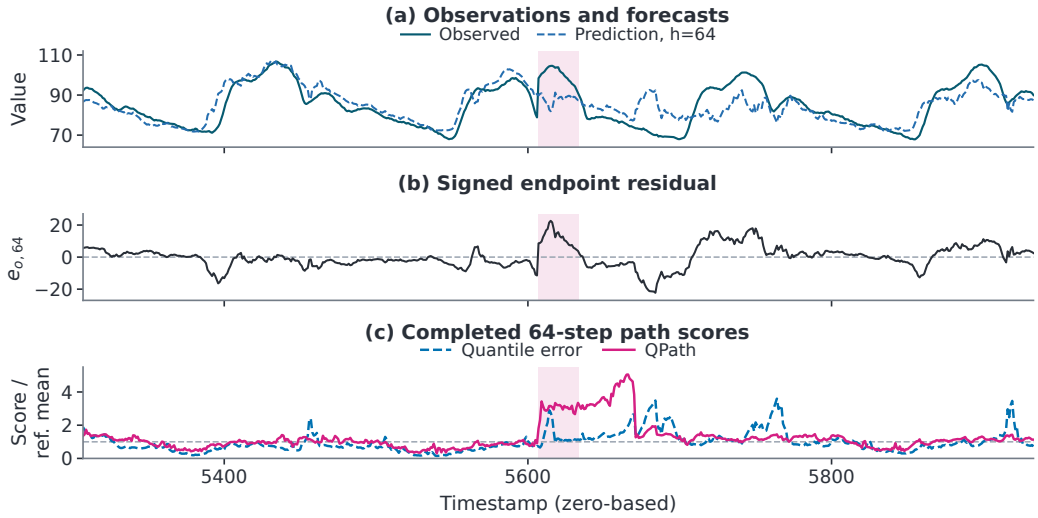


Figure 3: **Quantile error and QPath on a selected real series.** Chronos-2 zero-shot on TSB-AD-U, UCR 396. (a) Observations and rolling 64-step-ahead median forecasts. (b) Endpoint residual $e_{o,64} = Y_{o,64} - Q_{o,64}(0.5)$; the sole channel index is omitted throughout this figure. (c) Completed 64-step path scores normalized by each method's reference mean; the line at one is a reference guide, not an alarm threshold. These scores use all nine quantiles and are not sliding-window scores of (b). Shading marks the labeled anomaly; responses can persist afterward because completed paths still contain anomalous observations. The case was selected by the largest Full MAE-QPath–MAE gain on U and kept without reselection on quantile scores.

Table 5 provides descriptive context from a fixed TSB-AD leaderboard snapshot (TheDatumOrg, 2026), not a controlled comparison of models or protocols. Our rows average both backbones and ZS/FT under the published-window rule; Appendix F documents its assumptions. QPath exceeds these forecasting-based means, but not the best results from other detector families.

5.6 Magnitude ablation: absolute error versus quantile loss

To separate adding direction from changing magnitude, we compare four scores: MAE a, MAE-QPath $a\rho(v)$, quantile error g, and QPath $g\rho(v)$. Each uses its own reference denominator under Equation (6). No model is refitted for this ablation.

Table 6: **Magnitude ablation: MAE versus quantile loss.** Full AUTO VUS-PR.

Data	Backbone	Forecast setting	MAE	MAE-QPath	Quantile error	QPath
U	Chronos-2	Zero-shot	0.314	0.386	0.318	**0.389**
		Fine-tuned	0.320	0.391	0.326	**0.395**
	TimesFM	Zero-shot	0.317	0.389	0.322	**0.394**
		Fine-tuned	0.318	0.390	0.326	**0.400**
M	Chronos-2	Zero-shot	0.348	0.366	0.348	**0.367**
		Fine-tuned	0.349	0.366	0.351	**0.368**
	TimesFM	Zero-shot	0.345	0.350	0.344	**0.353**
		Fine-tuned	0.343	0.358	0.347	**0.362**

QPath uses mean quantile loss as its magnitude. All four methods share forecasts; the two QPath variants share the median-residual direction factor. Each method uses its own reference denominator. Six-decimal values are in Table 13.

QPath has a higher Full mean than MAE-QPath in all eight settings (Table 6), but the added gains are smaller than those over quantile error. On U they range from 0.003 to 0.010, with TimesFM 2.5 showing positive paired and source-cluster intervals in ZS/FT and Full/Test. The corresponding

source-cluster intervals include zero for Chronos-2 on U and for both backbones on M. Mean quantile loss therefore aligns the magnitude with the forecasting interface, while its additional benefit varies across forecasting settings and sources. Appendix C gives all paired magnitude comparisons.

Appendix E provides the complementary MAE-based controls and fixed-seed synthetic traces.

5.7 Comparison with a Mahalanobis-type score

A direct alternative to QPath is the standardized quadratic score $q = v^\top Pv/H$ (hereafter Quad), whose square root is a shrinkage-regularized Mahalanobis norm about the reference median ($v = 0$). QPath uses the same precision but replaces the standardized length $\sqrt{v^\top v/H}$ by quantile loss. On U, the strictly increasing square root and positive scalar reference normalization preserve the ranking of q. Table 7 compares their VUS-PR with QPath using the same forecasting settings and series.

Table 7: **QPath versus a Mahalanobis-type score on U.** AUTO VUS-PR.

Backbone	Forecast setting	Full series			Test suffix		
		Quad	QPath	Δ	Quad	QPath	Δ
Chronos-2	Zero-shot	0.377	**0.389**	+0.012	0.389	**0.402**	+0.013
	Fine-tuned	0.382	**0.395**	+0.013	0.394	**0.408**	+0.014
TimesFM	Zero-shot	0.376	**0.394**	+0.018	0.388	**0.407**	+0.019
	Fine-tuned	0.376	**0.400**	+0.023	0.388	**0.413**	+0.025

Δ is QPath minus Quad. Quad uses $v^\top Pv/H$; on U, taking its strictly increasing square root preserves rankings and, under the fixed rank-threshold evaluator, VUS-PR. The table compares descriptive means on the same series. No paired significance test for QPath–Quad is reported. Means and differences are rounded separately.

QPath has higher means in all eight settings; the comparison is descriptive, with no paired uncertainty estimate for QPath–Quad. Appendix E separately analyzes the MAE-radius variant.

6 Discussion

Quantile loss measures marginal forecast error, while the median-residual reference assesses temporal configuration. The matched diagonal control supports an explicit cross-term contribution in U, larger than the gain from replacing MAE magnitude alone. Channel averaging extends this temporal score without estimating cross-channel dependence.

The precision matrix describes second-order reference structure. Under Gaussian AR(1) noise with coefficient 0.8, temporal weighting improves spikes, level shifts, seasonal shifts, and variance bursts in both evaluation periods. Trend changes do not improve at base strength and can deteriorate in the strength sweep. Multivariate performance also depends on channel relevance, score normalization, and source composition.

Full includes the prefix used to fit the reference; on M, the FT paired-test conclusions differ between Full and Test. Fine-tuning and reference fitting also share prefix data, which can make reference residuals optimistic. Test excludes that prefix from evaluation, not from fitting. Overlapping reference paths reduce effective sample size, and the fixed shrinkage choice need not be optimal across sources. Cross-fitted references and regularization sensitivity remain useful extensions.

7 Conclusion

QPath combines quantile loss and historical median-residual direction for joint temporal anomaly assessment. Under matched forecasts, QPath shows consistent U gains and smaller, source-dependent M gains; the matched diagonal control directly supports an explicit cross-term contribution, especially in U. Synthetic failures and magnitude ablations clarify the scope of this detection-stage use of dependence.

AI USE STATEMENT

The authors established the overall mathematical framework and the main approach to the mathematical arguments. Generative AI assisted with detailed derivations, checking mathematical steps, and drafting mathematical exposition. The authors designed the synthetic-data generation procedures, made the experimental decisions, and conducted the training and evaluation experiments. AI assistance with the synthetic-data generators in Appendix D was limited to code implementation. Generative AI also assisted with research-idea exploration, conceptual framing and hypothesis refinement, experiment planning, method implementation and evaluation-code preparation, interpretation of experimental summaries, translation into English, and manuscript drafting, editing, and formatting. During manuscript preparation, scorer unit tests and table-regeneration checks were run, and newly added literature claims were checked against primary sources. The synthetic data were produced by executing the scripted generators; reported empirical values come from recorded experimental outputs. The authors retain responsibility for reviewing all AI-assisted content and for the final claims, results, and artifacts.

ETHICS STATEMENT

The study evaluates anomaly scores on existing benchmark data. Scores are not calibrated probabilities or validated operational alarms. Deployment would require application-specific assessment of false positives, detection delays, and data handling.

REPRODUCIBILITY STATEMENT

Appendix B specifies the experimental protocol. The anonymous real-data supplement provides cache-based scoring for quantile error, QPath, and the matched diagonal control, optional MAE ablations, timeline construction, a pinned-evaluator adapter, and paired statistical analysis. Cohort lists, configurations, input schemas, tests, and validation records document this workflow. The manuscript source includes numerical results and table/figure scripts. Replay requires matching forecast/factor caches or channel-score arrays, raw TSB-AD series, and the specified evaluator checkout. Backbone training/inference code, checkpoints, full caches, and the synthetic benchmark pipeline are not redistributed in the real-data supplement.

REFERENCES

Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134–147, 2017.

Carol Alexander, Michael Coulon, Yang Han, and Xiaochun Meng. Evaluating the discrimination ability of proper multi-variate scoring rules. *Annals of Operations Research*, 334(1):857–883, 2024. doi: 10.1007/s10479-022-04611-9. URL https://doi.org/10.1007/s10479-022-04611-9.

Amazon Web Services. Fast and accurate zero-shot forecasting with Chronos-Bolt and AutoGluon. AWS Machine Learning Blog, December 2024. URL https://aws.amazon.com/blogs/machine-learning/fast-and-accurate-zero-shot-forecasting-with-chronos-bolt-and-autogluon/.

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

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and

Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. URL https://arxiv.org/abs/2510.15821.

Ethan Baron, Boris N. Oreshkin, Ruijun Ma, Hanyu Zhang, Kari Torkkola, Michael W. Mahoney, Andrew Gordon Wilson, and Tatiana Konstantinova. Efficiently generating correlated sample paths from multi-step time series foundation models. In *NeurIPS Workshop on Recent Advances in Time Series Foundation Models: Have We Reached the BERT Moment?*, 2025. URL https://arxiv.org/abs/2510.02224.

Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A. Lozano. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys*, 54(3):1–33, 2021. doi: 10.1145/3444690.

Ben Cohen, Emaad Khwaja, Kan Wang, Clément Masson, Elise Ramé, Youssef Doubli, and Othmane Abou-Amal. Toto: Time series optimized transformer for observability. *arXiv preprint arXiv:2407.07874*, 2024. URL https://arxiv.org/abs/2407.07874.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4027–4035, 2021. doi: 10.1609/aaai.v35i5.16523.

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

Google Research. TimesFM-2.5-200M: Official model card. Hugging Face model card, 2025. URL https://huggingface.co/google/timesfm-2.5-200m-pytorch.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 387–395, 2018. doi: 10.1145/3219819.3219845.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukova, Othmane Abou-Amal, Chenghao Liu, Ameet Talwalkar, and David Asker. Toto 2.0: Time series forecasting enters the scaling era. *arXiv preprint arXiv:2605.20119*, 2026. URL https://arxiv.org/abs/2605.20119.

Jongwon Kim, Samuel Yoon, Young Myoung Ko, Yerin Kim, Sung Il Kim, and Jaeung Tae. Complexity- and statistics-guided anomaly detection in time series foundation models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d06768735ce7d11ade5baef7099b4e05-Paper-Conference.pdf.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi: 10.2307/1913643.

Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Hu. Revisiting time series outlier detection: Definitions and benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025. URL https://arxiv.org/abs/2511.11698.

Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/c3f3c690b7a99fba16d0efd35cb83b2c-Abstract-Datasets_and_Benchmarks_Track.html.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32369–32399, 2024. URL https://proceedings.mlr.press/v235/liu24cb.html.

Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pp. 89–94, 2015. URL https://www.esann.org/sites/default/files/proceedings/legacy/es2015-56.pdf.

Natalia Martinez Gil, Fearghal O'Donncha, Wesley M. Gifford, Nianjun Zhou, Dhaval C. Patel, and Roman Vaculin. Adaptive conformal anomaly detection with time series foundation models for signal monitoring. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=7uFbs68MSI.

Mohsin Munir, Shoaib Ahmed Siddiqui, Andreas Dengel, and Sheraz Ahmed. DeepAnT: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*, 7:1991–2005, 2019. doi: 10.1109/ACCESS.2018.2886457.

John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S. Tsay, Aaron Elmore, and Michael J. Franklin. Volume under the surface: A new accuracy evaluation measure for time-series anomaly detection. *Proceedings of the VLDB Endowment*, 15(11):2774–2787, 2022. doi: 10.14778/3551793.3551830.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. Lag-Llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2024. URL https://arxiv.org/abs/2310.08278.

Michael Scheuerer and Thomas M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015. doi: 10.1175/MWR-D-14-00269.1.

Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

TheDatumOrg. TSB-AD leaderboard and per-series evaluation results. https://thedatumorg.github.io/TSB-AD/, 2026. Per-series results from https://github.com/TheDatumOrg/TSB-AD, leaderboard snapshot 6beac72; accessed 24 September 2026.

Ruofeng Wen and Kari Torkkola. Deep generative quantile-copula models for probabilistic forecasting. In *ICML Time Series Workshop*, 2019. URL https://arxiv.org/abs/1907.10697.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Multivariate time-series anomaly detection via graph attention network. In *IEEE International Conference on Data Mining (ICDM)*, 2020. URL https://arxiv.org/abs/2009.02040.

A ALGEBRAIC PROPERTIES AND QUANTILE-SCORE GEOMETRY

A.1 GENERIC MAGNITUDE–DIRECTION DECOMPOSITION

For a nonnegative radius r, define $f(r, v) = r\rho(v)$, where $\rho(v) = \sqrt{v^\top P v/(v^\top v)}$ for $v \neq 0$ and $\rho(0) = 1$. The main score uses $r = g$; the MAE ablation uses $r = a$.

Proposition 1 (Direction invariance and bounds). *For $P \succ 0$ and $v \neq 0$, $\rho(tv) = \rho(v)$ for $t \neq 0$ and*

$$r\sqrt{\lambda_{\min}(P)} \leq f(r, v) \leq r\sqrt{\lambda_{\max}(P)}.$$

For $P = I$, $f(r, v) = r$, including at $v = 0$ under the neutral convention.

Proof. Write $u = v/\|v\|$. The Rayleigh quotient $u^\top P u$ is bounded by the extreme eigenvalues. Taking square roots and multiplying by $r \geq 0$ gives the result. Any nonzero scaling cancels between the two quadratic forms. The identity case is immediate. Moreover, $\widehat{V} \succeq 0$ and $A = 0.9\widehat{V} + 0.1I$ imply $\lambda_{\min}(A) \geq 0.1$, so $\rho(v) \leq \sqrt{10}$ for $v \neq 0$; the zero-direction convention also satisfies this bound. This invariance concerns the factor and a fixed transformed direction, not an arbitrary change to observations or quantile predictions that can also change g. □

Proposition 2 (Signed temporal cross terms). *Let D be diagonal with entries in $\{-1, 1\}$, $v \neq 0$, and hold r fixed. Then*

$$f(r, Dv)^2 - f(r, v)^2 = \frac{r^2}{v^\top v} v^\top (DPD - P)v.$$

The difference vanishes for diagonal P. Nonzero off-diagonal entries distinguish suitable sign configurations when $r > 0$, but need not increase every score.

Proof. Since $D^\top D = I$, the denominator is unchanged; expanding and subtracting the numerators proves the identity. If P is diagonal, $DPD = P$. If $P_{hk} \neq 0$, choosing a vector supported on h, k and flipping one sign gives $v^\top (DPD - P)v = -4P_{hk}v_h v_k$. The sign depends on the path. For $D = -I$, the difference is zero for every P: the direction factor responds to relative sign patterns, not a simultaneous sign reversal of the whole transformed path. Holding r fixed is essential: for asymmetric forecast quantiles, a sign flip of raw median residuals can also change g. These are algebraic score properties. □

A.2 WHY ALL-QUANTILE COLLAPSE RECOVERS MAE

The grid $\mathcal{T} = \{0.1, \ldots, 0.9\}$ has mean $1/2$. If $Q_h(\tau) = Q_h(0.5)$ for every τ, then

$$\frac{2}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \psi_\tau(e_h) = |e_h|.$$

For $e_h \geq 0$, use $2e_h \operatorname{mean}(\tau) = e_h$; for $e_h < 0$, use $2e_h(\operatorname{mean}(\tau) - 1) = -e_h$. Averaging over h gives $g = a$. This is a limiting consistency check, not an assertion that actual model quantiles coincide. With noncollapsed quantiles, a zero median residual $e = 0$ can coexist with positive mean quantile loss. Separately, the neutral rule $\rho(0) = 1$ applies only when the transformed vector $v = 0$, equivalently $e = \mu$ coordinatewise. If $e = 0$ but $\mu \neq 0$, the direction factor is computed normally from the nonzero transformed vector.

A.3 CIRCLE AND ELLIPSE IN FIGURE 1

The example stipulates $H = 2$, $v = e$, $A = \left(\begin{smallmatrix} 1 & 0.9 \\ 0.9 & 1 \end{smallmatrix}\right)$ and $P = A^{-1}$. It does not estimate this reference from the displayed paths. The quantiles are $Q_h(\tau) = \widehat{y}_h + \Phi^{-1}(\tau)$ with $\widehat{y} = (1.0, 1.2)$ and the nine-level grid. Observations are $Y^N = (1.8, 2.0)$ and $Y^A = (1.8, 0.4)$, where N and A label the same-sign and opposite-sign examples, respectively, giving residuals $(0.8, 0.8)$ and $(0.8, -0.8)$. The symmetric probability grid and forecast quantiles symmetric about each median make both magnitudes $g \simeq 0.51834$.

For $v \neq 0$, $x = \sqrt{2}\, g\, v/\|v\|$ gives

$$g = \sqrt{x^\top x/2}, \qquad m = \sqrt{x^\top P x/2}.$$

The illustrative unnormalized cutoff $\eta = 1$ therefore gives the circle $x^\top x \le 2$ and ellipse $x^\top P x \le 2$. The plotted points are (g, g) and $(g, -g)$, not the raw residual points. QPath scores are $g/\sqrt{1.9} \simeq 0.37604$ and $g/\sqrt{0.1} \simeq 1.63913$. These are magnitude–direction coordinate sublevel sets rather than probability contours; attainable points depend on the fixed quantile forecasts.

At $v = 0$ the nonzero-direction coordinate map is not used: the neutral rule is $m = g$, including $g > 0$. Reference normalization, omitted from the illustration, rescales the thresholds separately. The circle expresses mean quantile loss in rescaled coordinates; it does not introduce an RMSE detector.

A.4 Relation to quantile scores and CRPS

For a distribution F with a finite first moment, the CRPS loss has the quantile representation

$$\mathrm{CRPS}(F, y) = 2 \int_0^1 \psi_\tau\big(y - F^{-1}(\tau)\big)\, d\tau$$

(Gneiting & Raftery, 2007; Alexander et al., 2024). QPath uses a finite average over nine interior quantiles at each horizon step, followed by a horizon average. These uniform discrete weights and the omission of the tails do not equal the continuous integral. We therefore call g the mean quantile loss, rather than an exact CRPS. It is also distinct from normalized weighted quantile loss used for forecast evaluation: here normalization is by each channel's floored reference score mean. The common q10–q90 grid permits matched scoring across both forecasters; additional extreme quantiles are not evaluated.

B Implementation and evaluation details

B.1 Reference transform and numerical evaluation

For each channel and horizon, the reference location and scale are

$$\mu_{c,h} = \mathrm{median}_{o \in \mathcal{R}}\, e_{o,c,h}, \qquad s_{c,h} = \max\big\{\gamma\ \mathrm{median}_{o \in \mathcal{R}}\, |e_{o,c,h} - \mu_{c,h}|, 10^{-6}\big\},$$

where $\gamma = 1.482602218505602$ is the normal-consistency constant of the median absolute deviation. The fitted transform, consisting of μ_c, s_c, and the Cholesky factor of A_c, is reused for QPath, MAE-QPath, and the univariate quadratic comparison. The sample covariance in Equation (3) has denominator $n-1$ and centers the reference vectors by their sample mean; that mean is not subtracted from queries.

For a saved factor $A_c = L_c L_c^\top$, compute $w_c = L_c^{-1} v_c$, so $v_c^\top P_c v_c = \|w_c\|^2$. The direction factor is evaluated from this numerator and $v_c^\top v_c$, without clipping the query residuals. If the FP64 Cholesky factorization of A_c fails, A_c is recomputed from the same standardized reference in 80-decimal-digit arbitrary-precision arithmetic and factorized there, without eigenvalue clipping, jitter, or ridge reselection; scores then use that stored factor. For U, when converting a normalized QPath score sequence to FP64 would change its order or ties, the evaluator receives the dense ranks of the arbitrary-precision scores instead. Different forecasting settings do not share fitted transforms or denominator values.

The score-normalization mean uses every reference origin in $\mathcal{R}$, not the durations of the test-period score holds. If the reference mean is zero, the same positive floor in Equation (6) is used; the channel remains in the average. A zero reference mean does not imply that future scores are zero.

B.2 Head-only fine-tuning

Each series and seed starts from the original pretrained forecast head. Chronos-2 updates the terminal weight and bias of its existing output-patch embedding (1,032,528 parameters). TimesFM 2.5 updates the terminal weights of its existing point and continuous-quantile output projections (14,745,600 parameters). Unused output rows need not receive nonzero gradients. All other model parameters and native normalization paths remain fixed. Chronos-2 uses its official inverse transformation, including the required $\sinh$ operation; TimesFM 2.5 uses its native output composition and flip-invariance.

A sample contains 256 context and 64 target timestamps. With prefix length L, the split point is $\lfloor 4L/5 \rfloor$. Training and validation footprints must lie completely on their respective sides. If fewer

than 16 validation origins remain, the full-prefix fixed-budget rule is used. Applying this eligibility rule to the benchmark prefix lengths gives 230 validation-eligible and 120 fixed-budget series in U, and 127 and 53, respectively, in M. For three constant-prefix TimesFM 2.5 series, head parameters do not change during fine-tuning; these series remain in the evaluation.

Training origins are drawn into a fixed seed-specific bank by time bins, with at most 2,048 origins; validation uses at most 256. Let N_i be the eligible training-origin count before sampling. The update budget is

$$K_i = \min\left(200, 10 \left\lceil \frac{N_i}{32} \right\rceil\right).$$

The effective batch contains up to 32 origins. Multivariate batches include all channels of those origins; microbatching preserves the same averaged objective. AdamW uses learning rate 10^{-4}, zero weight decay, cosine decay over K_i, no warmup, and gradient clipping at one. Model execution is FP32 with AMP and TF32 disabled. Frozen encoder features are reused during head updates; hidden features are not inputs to QPath.

Chronos-2 minimizes its native pinball loss over 21 quantiles; TimesFM 2.5 minimizes its native objective, a squared-error loss on the point output plus pinball losses on nine quantiles. Candidate checkpoints are the distinct elements of $\{\min(K_i, s) : s \in \{50, 100, 200\}\}$. Validation median MAE selects a candidate, with exact ties retaining the earlier one. Without validation, the last checkpoint is selected. The entire budget runs, and step zero is not a selected FT candidate. Seeds are 42, 43, and 44. Anomaly labels do not select a head checkpoint. Reference origins span the complete prefix after checkpoint selection, including timestamps used for head training. This is training-reference overlap rather than test-label leakage; it may nevertheless make the reference error distribution optimistic. With a contiguous prefix of length L, the stride-one reference count is $n = \max(0, L - 319)$. At least two complete reference paths are required. In both U and M, the shortest prefix has length 500, giving minimum $n = 181$ under this rule; synthetic prefixes give $n = 1{,}681$. The reference implementation rejects $n < 2$. The overlapping 320-timestamp footprints are not independent samples.

B.3 Time mapping and fixed software inputs

For a zero-based origin o, context is $[o - 255, o]$ and the target is $[o + 1, o + 64]$. The path score is assigned to endpoint $o + 64$. The first valid endpoint is 319. For a full-length offline evaluation, the preceding 319 timestamps repeat that first score; this padding is not a causal alarm before the first valid endpoint. U uses stride-one reference and evaluation origins. For M, the prefix is scored at stride-one reference endpoints and the suffix at stride-16 endpoints with forward hold. Full includes the reference prefix; Test restricts the evaluated timestamps to the suffix. AUTO window selection uses the evaluator's first raw feature. The settings `opt` and `thre=250` belong to the metric implementation, not a selected anomaly operating threshold. The fixed evaluator selects score thresholds by positions in the sorted score sequence; order- and tie-preserving transformations therefore preserve the evaluated label partitions.

Model and evaluator revisions

Chronos-2: `29ec3766d36d6f73f0696f85560a422f50e8498c`

TimesFM 2.5: `1d952420fba87f3c6dee4f240de0f1a0fbc790e3`

TSB-AD: `e0975a5f7d3e65ab77e9fab24d1b5b51acda8f48`

C Complete real-data comparisons

QPath uses mean quantile loss multiplied by the temporal direction factor; quantile error uses magnitude alone. Both use their own reference denominators. The two contrasts below distinguish adding direction from replacing the MAE radius. All results in this section use the same evaluator revision and the Full and Test periods defined in Section 4.1.

C.1 Uncertainty and paired sampling

For each series, compare methods within the same forecast and evaluation period. FT first averages seed 42, 43, and 44 metrics within the series; it does not average predictions or weights and does not treat seeds as extra datasets. Real-data tests are two-sided paired t with $N - 1$ degrees of freedom. Their p values and 95% intervals are nominal. The source-cluster bootstrap has 10,000 draws over 23 U or 17 M sources while targeting an equal-weight mean over series; source resampling keeps within-source series together. Neither procedure corrects for all benchmark-informed method development.

Tables 8 and 9 give QPath minus quantile error. Both U backbones have positive paired and source-cluster intervals under ZS and FT in both periods. Every M source-cluster interval includes zero. For ZS on M, both Chronos-2 contrasts have $p < 0.05$, whereas neither TimesFM 2.5 contrast does. For FT on M, the paired p values are below 0.05 in Full but not Test. These results do not establish source-wide superiority or noninferiority.

Table 8: **QPath minus quantile error: Full.** AUTO VUS-PR differences.

Data	Backbone	Forecast setting	Δ	95% paired t CI	p	95% source CI
U	Chronos-2	Zero-shot	$+0.0710$	$[+0.0568,\ +0.0852]$	2.33×10^{-20}	$[+0.0340,\ +0.0998]$
		Fine-tuned	$+0.0687$	$[+0.0542,\ +0.0832]$	1.27×10^{-18}	$[+0.0310,\ +0.0981]$
	TimesFM	Zero-shot	$+0.0719$	$[+0.0562,\ +0.0876]$	1.43×10^{-17}	$[+0.0283,\ +0.1050]$
		Fine-tuned	$+0.0735$	$[+0.0579,\ +0.0892]$	2.69×10^{-18}	$[+0.0297,\ +0.1056]$
M	Chronos-2	Zero-shot	$+0.0195$	$[+0.0057,\ +0.0332]$	0.0057	$[-0.0036,\ +0.0453]$
		Fine-tuned	$+0.0171$	$[+0.0034,\ +0.0309]$	0.0151	$[-0.0064,\ +0.0436]$
	TimesFM	Zero-shot	$+0.0082$	$[-0.0063,\ +0.0227]$	0.2650	$[-0.0194,\ +0.0352]$
		Fine-tuned	$+0.0155$	$[+0.0002,\ +0.0309]$	0.0477	$[-0.0184,\ +0.0507]$

Uncertainty and FT aggregation follow Appendix C.1. U: $N = 350$, 23 sources; M: $N = 180$, 17 sources.

Table 9: **QPath minus quantile error: Test.** AUTO VUS-PR differences.

Data	Backbone	Forecast setting	Δ	95% paired t CI	p	95% source CI
U	Chronos-2	Zero-shot	$+0.0704$	$[+0.0557,\ +0.0850]$	4.57×10^{-19}	$[+0.0309,\ +0.1009]$
		Fine-tuned	$+0.0678$	$[+0.0529,\ +0.0826]$	1.68×10^{-17}	$[+0.0281,\ +0.0985]$
	TimesFM	Zero-shot	$+0.0694$	$[+0.0532,\ +0.0857]$	1.29×10^{-15}	$[+0.0235,\ +0.1050]$
		Fine-tuned	$+0.0731$	$[+0.0568,\ +0.0894]$	5.24×10^{-17}	$[+0.0274,\ +0.1072]$
M	Chronos-2	Zero-shot	$+0.0141$	$[+0.0013,\ +0.0269]$	0.0315	$[-0.0044,\ +0.0385]$
		Fine-tuned	$+0.0119$	$[-0.0008,\ +0.0247]$	0.0663	$[-0.0072,\ +0.0368]$
	TimesFM	Zero-shot	$+0.0044$	$[-0.0087,\ +0.0176]$	0.5043	$[-0.0199,\ +0.0287]$
		Fine-tuned	$+0.0112$	$[-0.0027,\ +0.0250]$	0.1127	$[-0.0192,\ +0.0444]$

Uncertainty and FT aggregation follow Appendix C.1. U: $N = 349$, 23 sources; M: $N = 180$, 17 sources.

Tables 10 and 11 give QPath minus MAE-QPath. TimesFM 2.5 U supports the additional gain under both procedures in ZS/FT and Full/Test. The corresponding source-cluster intervals include zero for Chronos-2 on U and both backbones on M. Table 12 complements the Full ablation with absolute Test means for all four methods. Positive mean differences do not imply that every series benefits.

C.2 Definition of the headline relative improvement

For each benchmark $\mathcal{B}$, the abstract reports

$$G_{\mathcal{B}} = \frac{1}{4} \sum_{b \in \{\text{Chronos-2,TimesFM 2.5}\}} \sum_{f \in \{\text{ZS,FT}\}} 100 \frac{\overline{\text{VUSPR}}^{\text{QPath}}_{\mathcal{B},b,f} - \overline{\text{VUSPR}}^{\text{QE}}_{\mathcal{B},b,f}}{\overline{\text{VUSPR}}^{\text{QE}}_{\mathcal{B},b,f}},$$

where QE denotes quantile error, using Full AUTO VUS-PR only. The unrounded means give 22.0668% for U and 4.3376% for M, rounded to 22.1% and 4.3%; Table 13 displays six-decimal means. These are descriptive setting-level relative changes, not percentage-point differences or

Table 10: **QPath minus MAE-QPath: Full.** AUTO VUS-PR differences.

Data	Backbone	Forecast setting	Δ	95% paired t CI	p	95% source CI
U	Chronos-2	Zero-shot	$+0.0028$	$[-0.0001,\ +0.0058]$	0.0577	$[-0.0010,\ +0.0072]$
		Fine-tuned	$+0.0036$	$[+0.0006,\ +0.0066]$	0.0194	$[-0.0013,\ +0.0093]$
	TimesFM	Zero-shot	$+0.0044$	$[+0.0015,\ +0.0073]$	0.0030	$[+0.0011,\ +0.0079]$
		Fine-tuned	$+0.0096$	$[+0.0056,\ +0.0136]$	2.86×10^{-6}	$[+0.0037,\ +0.0154]$
M	Chronos-2	Zero-shot	$+0.0016$	$[-0.0026,\ +0.0059]$	0.4500	$[-0.0051,\ +0.0076]$
		Fine-tuned	$+0.0025$	$[-0.0013,\ +0.0063]$	0.1978	$[-0.0046,\ +0.0096]$
	TimesFM	Zero-shot	$+0.0031$	$[+0.0004,\ +0.0058]$	0.0250	$[-0.0028,\ +0.0109]$
		Fine-tuned	$+0.0044$	$[+0.0015,\ +0.0073]$	0.0034	$[-0.0008,\ +0.0122]$

Uncertainty and FT aggregation follow Appendix C.1. U: $N = 350$, 23 sources; M: $N = 180$, 17 sources.

Table 11: **QPath minus MAE-QPath: Test.** AUTO VUS-PR differences.

Data	Backbone	Forecast setting	Δ	95% paired t CI	p	95% source CI
U	Chronos-2	Zero-shot	$+0.0044$	$[+0.0012,\ +0.0075]$	0.0075	$[-0.0003,\ +0.0095]$
		Fine-tuned	$+0.0034$	$[+0.0002,\ +0.0065]$	0.0349	$[-0.0016,\ +0.0092]$
	TimesFM	Zero-shot	$+0.0050$	$[+0.0022,\ +0.0077]$	0.0004	$[+0.0014,\ +0.0083]$
		Fine-tuned	$+0.0102$	$[+0.0062,\ +0.0141]$	6.55×10^{-7}	$[+0.0043,\ +0.0168]$
M	Chronos-2	Zero-shot	$+0.0024$	$[-0.0017,\ +0.0065]$	0.2457	$[-0.0049,\ +0.0085]$
		Fine-tuned	$+0.0032$	$[-0.0002,\ +0.0067]$	0.0639	$[-0.0038,\ +0.0104]$
	TimesFM	Zero-shot	$+0.0037$	$[+0.0013,\ +0.0062]$	0.0032	$[-0.0017,\ +0.0110]$
		Fine-tuned	$+0.0044$	$[+0.0018,\ +0.0069]$	0.0008	$[-0.0007,\ +0.0117]$

Uncertainty and FT aggregation follow Appendix C.1. U: $N = 349$, 23 sources; M: $N = 180$, 17 sources.

Table 12: **Magnitude ablation: MAE versus quantile loss.** Test AUTO VUS-PR.

Data	Backbone	Forecast setting	MAE	MAE-QPath	Quantile error	QPath
U	Chronos-2	Zero-shot	0.328	0.398	0.332	**0.402**
		Fine-tuned	0.334	0.404	0.340	**0.408**
	TimesFM	Zero-shot	0.332	0.402	0.337	**0.407**
		Fine-tuned	0.332	0.403	0.340	**0.413**
M	Chronos-2	Zero-shot	0.359	0.371	0.360	**0.374**
		Fine-tuned	0.361	0.372	0.363	**0.375**
	TimesFM	Zero-shot	0.356	0.357	0.356	**0.360**
		Fine-tuned	0.355	0.365	0.358	**0.370**

QPath uses mean quantile loss as its magnitude. All four methods share forecasts; the two QPath variants share the median-residual direction factor. Each method uses its own reference denominator. Six-decimal values are in Table 13.

Table 13: **Six-decimal real-data mean AUTO VUS-PR.** Full and Test evaluation.

Data	Backbone	Forecast setting	Evaluation period	MAE	MAE-QPath	Quantile error	QPath
U	Chronos-2	Zero-shot	Full	0.314101	0.386237	0.318060	**0.389078**
U	Chronos-2	Zero-shot	Test	0.328111	0.398102	0.332093	**0.402462**
U	Chronos-2	Fine-tuned	Full	0.319541	0.391175	0.326048	**0.394772**
U	Chronos-2	Fine-tuned	Test	0.333833	0.404351	0.339980	**0.407736**
U	TimesFM	Zero-shot	Full	0.316931	0.389370	0.321911	**0.393779**
U	TimesFM	Zero-shot	Test	0.331658	0.401893	0.337406	**0.406853**
U	TimesFM	Fine-tuned	Full	0.317894	0.390285	0.326333	**0.399875**
U	TimesFM	Fine-tuned	Test	0.332353	0.403054	0.340104	**0.413224**
M	Chronos-2	Zero-shot	Full	0.348155	0.365529	0.347681	**0.367165**
M	Chronos-2	Zero-shot	Test	0.358990	0.371365	0.359705	**0.373770**
M	Chronos-2	Fine-tuned	Full	0.349167	0.365501	0.350844	**0.367984**
M	Chronos-2	Fine-tuned	Test	0.360872	0.371544	0.362846	**0.374778**
M	TimesFM	Zero-shot	Full	0.344736	0.349622	0.344483	**0.352709**
M	TimesFM	Zero-shot	Test	0.355952	0.356663	0.355959	**0.360404**
M	TimesFM	Fine-tuned	Full	0.343440	0.357845	0.346742	**0.362253**
M	TimesFM	Fine-tuned	Test	0.354990	0.365142	0.358353	**0.369512**

These numerical means supply the main tables and the setting-level relative improvements. Paired differences are rounded separately in the inferential tables.

averages of per-series relative changes. Full and Test are overlapping evaluations, not independent replications.

C.3 MATCHED DIAGONAL CONTROL

Equation (7) changes only the explicit off-diagonal query terms. Quantile loss, median-residual transformation, fitted precision matrix P, reference origins, forecasts, and QPath reference denominators are identical. In particular, the diagonal entries are those of $P = A^{-1}$ rather than reciprocals of the diagonal of A. The control is not normalized by its own reference mean. At $v = 0$, its direction factor is one and the mean quantile loss is preserved.

The two backbones and ZS plus seeds 42, 43, and 44 cover 4,240 series-level forecasting runs. The diagonal control has 8,480 evaluation records across Full and Test, of which eight are the U suffix with no positive label; the other 8,472 are defined. FT is derived by averaging the three seed metrics within each series. Neither forecasts nor reference statistics are refitted for this comparison.

Tables 14 and 15 give all paired results with per-comparison two-sided p values and both uncertainty summaries. In U, all eight source-cluster intervals have positive endpoints. In M, Chronos-2's ZS intervals and FT Full interval are positive, though the latter lower endpoint is only 0.00011; its FT Test interval spans zero. All four TimesFM 2.5 intervals span zero. Nominal paired $p = 0.0215$ for TimesFM 2.5 FT Full is therefore evidence under the series-paired model, not a source-wide guarantee.

Table 14: **Matched diagonal control on U.** QPath minus the shared-denominator diagonal control.

Backbone	Forecast setting	Evaluation period	Diagonal	QPath	Δ	p
Chronos-2	Zero-shot	Full	0.31571	**0.38908**	$+0.07336$	1.57×10^{-20}
		Test	0.33024	**0.40246**	$+0.07223$	2.52×10^{-19}
	Fine-tuned	Full	0.32428	**0.39477**	$+0.07049$	8.17×10^{-19}
		Test	0.33851	**0.40774**	$+0.06923$	5.84×10^{-18}
TimesFM	Zero-shot	Full	0.31867	**0.39378**	$+0.07511$	3.76×10^{-18}
		Test	0.33504	**0.40685**	$+0.07182$	3.55×10^{-16}
	Fine-tuned	Full	0.32427	**0.39987**	$+0.07560$	1.64×10^{-18}
		Test	0.33875	**0.41322**	$+0.07447$	1.83×10^{-17}
Paired difference uncertainty				95% paired t CI		95% source CI
Chronos-2	Zero-shot	Full		$[+0.05879,\ +0.08794]$		$[+0.03560,\ +0.10254]$
		Test		$[+0.05734,\ +0.08711]$		$[+0.03272,\ +0.10217]$
	Fine-tuned	Full		$[+0.05572,\ +0.08526]$		$[+0.03226,\ +0.10041]$
		Test		$[+0.05431,\ +0.08414]$		$[+0.02960,\ +0.09999]$
TimesFM	Zero-shot	Full		$[+0.05902,\ +0.09119]$		$[+0.03073,\ +0.10850]$
		Test		$[+0.05532,\ +0.08831]$		$[+0.02580,\ +0.10682]$
	Fine-tuned	Full		$[+0.05961,\ +0.09160]$		$[+0.03143,\ +0.10763]$
		Test		$[+0.05815,\ +0.09080]$		$[+0.02876,\ +0.10790]$

Mean quantile loss, the full precision matrix P, and QPath reference denominators are fixed. Intervals and tests are nominal; source intervals use 10,000 source-cluster replicates. U: 350 Full and 349 Test series from 23 sources.

Table 16 summarizes Full win/tie/loss counts, positive-source counts, and leave-one-source-out mean ranges. These diagnostics do not change the benchmark composition or define a new test. All conditions, individual seeds, auxiliary metrics, and per-source comparisons are provided in the supplementary CSV files. Synthetic quantile-diagonal traces and experiments are not part of this control.

Table 15: **Matched diagonal control on M.** QPath minus the shared-denominator diagonal control.

Backbone	Forecast setting	Evaluation period	Diagonal	QPath	Δ	p
Chronos-2	Zero-shot	Full	0.35040	**0.36717**	+0.01676	0.0002
		Test	0.35947	**0.37377**	+0.01430	0.0006
	Fine-tuned	Full	0.35286	**0.36798**	+0.01513	0.0012
		Test	0.36228	**0.37478**	+0.01250	0.0033
TimesFM	Zero-shot	Full	0.34738	**0.35271**	+0.00533	0.3021
		Test	0.35803	**0.36040**	+0.00237	0.5911
	Fine-tuned	Full	0.34864	**0.36225**	+0.01362	0.0215
		Test	0.35990	**0.36951**	+0.00961	0.0508

Paired difference uncertainty			95% paired t CI	95% source CI
Chronos-2	Zero-shot	Full	[+0.00798, +0.02554]	[+0.00319, +0.03275]
		Test	[+0.00626, +0.02235]	[+0.00207, +0.02982]
	Fine-tuned	Full	[+0.00604, +0.02421]	[+0.00011, +0.03214]
		Test	[+0.00422, +0.02079]	[−0.00053, +0.02879]
TimesFM	Zero-shot	Full	[−0.00484, +0.01550]	[−0.01521, +0.02612]
		Test	[−0.00632, +0.01106]	[−0.01536, +0.02232]
	Fine-tuned	Full	[+0.00203, +0.02520]	[−0.01170, +0.03926]
		Test	[−0.00003, +0.01926]	[−0.01336, +0.03477]

Mean quantile loss, the full precision matrix P, and QPath reference denominators are fixed. Intervals and tests are nominal; source intervals use 10,000 source-cluster replicates. M: 180 series from 17 sources in each period.

Table 16: **Cross-term source and downside diagnostics.** QPath minus Diagonal, Full AUTO VUS-PR.

Data	Backbone	Forecast setting	Win/tie/loss	Large drops	Positive sources	LOSO mean range
U	Chronos-2	Zero-shot	255/5/90	14/6	17/23	[+0.0629, +0.0801]
U	Chronos-2	Fine-tuned	252/3/95	17/6	17/23	[+0.0597, +0.0780]
U	TimesFM	Zero-shot	239/3/108	22/9	16/23	[+0.0636, +0.0829]
U	TimesFM	Fine-tuned	242/3/105	24/8	15/23	[+0.0641, +0.0834]
M	Chronos-2	Zero-shot	122/6/52	7/3	11/17	[+0.0125, +0.0212]
M	Chronos-2	Fine-tuned	118/5/57	8/3	13/17	[+0.0108, +0.0209]
M	TimesFM	Zero-shot	86/8/86	12/7	8/17	[+0.0009, +0.0150]
M	TimesFM	Fine-tuned	88/5/87	12/7	8/17	[+0.0057, +0.0245]

All series remain in the primary comparison. Ties use 10^{-12}. Large drops count $\Delta < -0.05$ / $\Delta < -0.10$. LOSO removes one source at a time; its range is not a confidence interval. Source and seed details are supplied in CSV form.

C.4 SELECTED REAL-DATA ILLUSTRATION

Figure 3 uses Chronos-2 ZS on UCR series `396_UCR_id_94_Medical_tr_1875_1st_5607.csv`. The selection criterion is the largest Full MAE-QPath minus MAE VUS-PR gain across U; the same case is used for the quantile scores without further selection. It is an outcome-selected illustration, not a representative or independent sample. The displayed range is 5307–5933 and the labeled anomaly is 5607–5633 (27 timestamps), all zero-based. Both methods use the same forecasts and their original reference denominators, with no smoothing or time shift. Their Full VUS-PR values are 0.02014 (quantile error) and 0.74554 (QPath); Test values are 0.02649 and 0.71980. The line at score one marks reference-relative scale rather than a fitted alarm threshold. Rolling $h = 64$ predictions supply one endpoint residual per origin; the plotted scores are computed from each origin's complete 64-step, nine-quantile forecast path.

D GAUSSIAN-NOISE SYNTHETIC ANOMALY PATTERNS AND QUANTILE RESULTS

D.1 DATA AND FORECASTING PROTOCOL

The study asks how scoring responds to anomaly shape when the injected deviation is controlled. Using the point-wise/pattern-wise distinction of Lai et al. (2021), we construct five patterns: spike, level shift, trend change, seasonal shift, and variance burst. Each of the 200 base series has 4,000 timestamps, with a 2,000-step reference prefix and 2,000-step test suffix. The base is the sum of one or two sinusoids with periods from $\{24, 48, 96, 168\}$ and amplitudes summing to one, a slow linear trend, and Gaussian AR(1) noise with coefficient 0.8 and stationary standard deviation 0.2.

A pattern anomaly occupies one 100–200-step segment at least 300 steps from either test boundary. Level shifts add a constant, trend changes add a ramp, and variance bursts amplify the noise. Seasonal shifts mix two fixed constructions: a period multiplier of 1.5 with continuous phase for 25 series, and a half-cycle phase shift for 15. The four pattern types are rescaled to mean absolute injected deviation one. A spike series instead contains three isolated spikes of height 1.2 at least 300 steps apart. Each type has 40 fixed seeds. Figure 2 illustrates the shapes schematically.

The strength sweep reuses the Gaussian-noise seasonal-shift and trend-change seeds at strengths 0.5 and 2.0, yielding 160 series. Strength one is the original Gaussian condition, not another independent sample. Thus the reported synthetic evaluation contains 360 series variants, or 720 series–backbone runs across the two models. All synthetic forecasts are ZS with context 256 and horizon 64. Reference and test origins have stride one, and score placement and evaluation follow the U protocol.

The controlled quantity is the injected deviation, not the actual forecast residual or quantile loss. These are end-to-end forecasting experiments, including model reactions after anomalous observations enter the context. They are not an exact equal-loss causal experiment. Inference retains $N = 40$ within each type–strength–backbone condition rather than pooling repeated seeds across strengths.

D.2 MEANS AND TYPE-SPECIFIC PAIRED COMPARISONS

Tables 1 and 17 report the detection results and descriptive cohort means. Tables 18 and 19 report all Gaussian-noise type–strength paired differences and nominal two-sided p values. The main contrast is QPath minus quantile error. Replacing the MAE radius is a separate comparison from adding direction; the supplementary numerical data distinguish both contrasts.

Table 17: **Gaussian-noise synthetic cohort means.** Zero-shot AUTO VUS-PR, descriptive only.

Study	Backbone	Evaluation period	N	Quantile error	QPath
Base strength	Chronos-2	Full	200	0.4069	**0.4993**
Base strength	TimesFM	Full	200	0.4289	**0.5224**
Strength sweep	Chronos-2	Full	160	0.3940	**0.3970**
Strength sweep	TimesFM	Full	160	0.4225	**0.4278**
Base strength	Chronos-2	Test	200	0.4422	**0.5182**
Base strength	TimesFM	Test	200	0.4536	**0.5352**
Strength sweep	Chronos-2	Test	160	**0.4259**	0.4237
Strength sweep	TimesFM	Test	160	0.4482	**0.4499**

All rows use Gaussian noise. Base strength: five types with 40 series each. Strength sweep: two types and two strengths; repeated seeds across strengths are not independent observations. Inference uses type–strength-specific pairs rather than pooling these means.

Table 18: **Gaussian: QPath minus quantile error.** Zero-shot AUTO VUS-PR.

Anomaly type	Backbone	Full Δ	Full p	Test Δ	Test p
Spike	Chronos-2	$+0.2033$	4.84×10^{-14}	$+0.1580$	3.44×10^{-12}
	TimesFM	$+0.2615$	2.67×10^{-14}	$+0.2278$	1.91×10^{-12}
Level shift	Chronos-2	$+0.0948$	7.33×10^{-5}	$+0.0843$	0.0002
	TimesFM	$+0.0763$	0.0002	$+0.0734$	0.0002
Trend change	Chronos-2	-0.0161	0.2851	-0.0233	0.0894
	TimesFM	-0.0065	0.6179	-0.0142	0.2585
Seasonal shift	Chronos-2	$+0.0541$	0.0039	$+0.0491$	0.0072
	TimesFM	$+0.0443$	0.0135	$+0.0380$	0.0317
Variance burst	Chronos-2	$+0.1257$	4.18×10^{-7}	$+0.1118$	5.85×10^{-7}
	TimesFM	$+0.0918$	1.22×10^{-7}	$+0.0830$	1.44×10^{-7}

Each comparison uses $N = 40$ and a two-sided paired t test. The reported p values are per-comparison. A negative difference favors the quantile-error baseline.

Table 19: **Gaussian-noise strength sweep: QPath minus quantile error.** Zero-shot AUTO VUS-PR.

Anomaly type (strength)	Backbone	Full Δ	Full p	Test Δ	Test p
Seasonal shift (0.5)	Chronos-2	$+0.0435$	0.0265	$+0.0423$	0.0282
	TimesFM	$+0.0492$	0.0107	$+0.0467$	0.0193
Seasonal shift (2.0)	Chronos-2	$+0.0178$	0.1503	$+0.0091$	0.4425
	TimesFM	-0.0068	0.5406	-0.0110	0.3359
Trend change (0.5)	Chronos-2	-0.0216	0.1024	-0.0281	0.0427
	TimesFM	-0.0152	0.3043	-0.0193	0.2036
Trend change (2.0)	Chronos-2	-0.0278	0.0144	-0.0323	0.0017
	TimesFM	-0.0060	0.4458	-0.0095	0.2068

Each comparison uses $N = 40$ and a two-sided paired t test. The reported p values are per-comparison. A negative difference favors the quantile-error baseline.

Under Gaussian noise, spikes, level shifts, seasonal shifts, and variance bursts improve with nominal $p < 0.05$ for both backbones and periods, while trend-change contrasts have $p \geq 0.05$. The TimesFM 2.5 seasonal-shift Test gain has $p = 0.0317$. In the strength sweep, half-strength seasonal shifts improve in both periods for both backbones. Chronos-2 double-strength trend changes decline in Full ($p = 0.0144$) and Test ($p = 0.0017$); its half-strength Test trend also declines ($p = 0.0427$). In the strength sweep, three Full and four Test contrasts have $p < 0.05$: two gains and one loss in Full, and two gains and two losses in Test. The tables report both positive and negative differences.

E MAE-MAGNITUDE VARIANTS

E.1 MAGNITUDE ALTERNATIVES

The four-way ablation uses a, $a\rho(v)$, g, and $g\rho(v)$ with separate reference normalization. MAE-only contrasts use the same per-series sampling and nominal two-sided paired t reporting convention. They address the MAE-radius variant, separately from the mean-quantile-loss comparisons in Appendix C.

E.2 MATCHED DIAGONAL CONTROL WITH MAE MAGNITUDE

For any channel and forecasting setting, the MAE-based matched diagonal control is

$$m^{\mathrm{MAE,diag}} = a\sqrt{\frac{\sum_h P_{hh} v_h^2}{v^\top v}},$$

with the neutral rule at zero and the same MAE-QPath reference denominator. It removes explicit off-diagonal query terms while fixing the precision matrix and channel weights. Table 20 reports the available aggregate MAE comparison for Chronos-2 ZS on M, with MAE-QPath gains of 0.01547 on Full and 0.01309 on Test. It complements the all-setting quantile-based control in Appendix C.3; MAE-diagonal trajectories are also available for the illustrative synthetic series. The diagonal entries are taken from the full precision matrix P, not from a separately fitted diagonal covariance.

Table 20: **Matched diagonal control with MAE magnitude.** Chronos-2 Zero-shot on M ($N = 180$).

Evaluation period	MAE	MAE-diagonal	MAE-QPath	Δ	p
Full	0.34815	0.35006	**0.36553**	+0.01547	0.0010
Test	0.35899	0.35827	**0.37137**	+0.01309	0.0026

Only the explicit off-diagonal query terms are removed; the full-score reference denominator is fixed. Δ is MAE-QPath minus MAE-diagonal. p is the nominal two-sided paired t value.

E.3 MAHALANOBIS-TYPE SCORE COMPARISON

For $q = v^\top Pv/H = d\rho(v)^2$, where $d = v^\top v/H$, the square root is strictly increasing on nonnegative scores, and positive scalar single-channel normalization preserves the ranking equivalence of q and $\sqrt{q}$. The main comparison with QPath appears in Table 7. Table 21 reports the complementary MAE-QPath contrast in both evaluation periods. Five of the eight MAE-QPath contrasts have nominal $p < 0.05$. These tests compare MAE-QPath with Quad, not the quantile-based QPath with Quad.

E.4 MAE-RADIUS SYNTHETIC EXAMPLES

Figure 4 gives the five fixed-seed Gaussian AR(1) examples for the MAE-radius ablation. The general diagonal-control definition above also applies to these traces. The quantile-based real-data illustration in Figure 3 is described separately in Appendix C.4.

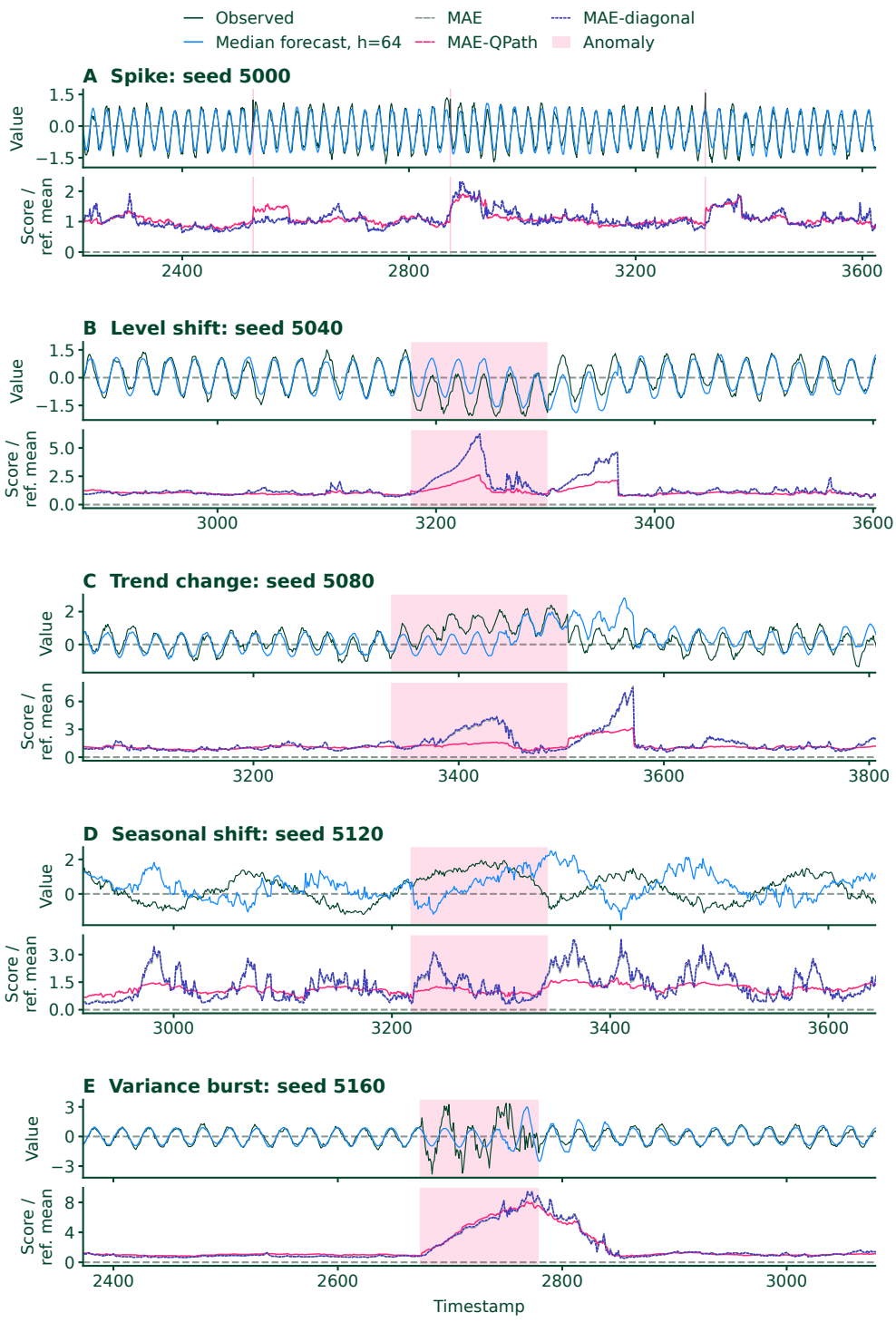


Figure 4: **MAE-radius ablation: synthetic scoring examples.** Panels A–E show Gaussian AR(1)-noise series with seeds 5000, 5040, 5080, 5120, and 5160, respectively, using Chronos-2 ZS. Each pairs observations and 64-step-ahead median predictions with reference-normalized MAE (dashed), MAE-QPath (solid), and MAE-diagonal (dotted) scores. Shading marks anomalous timestamps.

Table 21: **MAE-QPath versus the quadratic score on U.** AUTO VUS-PR.

Backbone	Forecast setting	Evaluation period	Quad	MAE-QPath	Δ	p
Chronos-2	Zero-shot	Full	0.37698	**0.38624**	+0.00926	0.0614
Chronos-2	Zero-shot	Test	0.38901	**0.39810**	+0.00909	0.0699
Chronos-2	Fine-tuned	Full	0.38188	**0.39118**	+0.00929	0.0613
Chronos-2	Fine-tuned	Test	0.39401	**0.40435**	+0.01034	0.0392
TimesFM	Zero-shot	Full	0.37614	**0.38937**	+0.01323	0.0171
TimesFM	Zero-shot	Test	0.38829	**0.40189**	+0.01360	0.0157
TimesFM	Fine-tuned	Full	0.37639	**0.39028**	+0.01389	0.0126
TimesFM	Fine-tuned	Test	0.38841	**0.40305**	+0.01465	0.0086

Δ is MAE-QPath minus Quad. Both use the same transformed median residuals and precision. p is the nominal two-sided paired t value. Five of the eight MAE-QPath–Quad contrasts have $p < 0.05$.

F AUTOMATIC-WINDOW RULES FOR PUBLISHED COMPARISONS

The published comparison uses leaderboard snapshot `6beac72`; our score evaluation uses revision `e0975a5`, whose full hash is given in Appendix B.3. These identifiers fix the published result files and the evaluation code, respectively. The evaluator estimates the VUS window from the first raw feature. A 2026 TSB-AD revision changed the rule for estimated periods below 6 or between 301 and 303; the evaluator revision used here includes that change, while the published values predate it. The two rules differ on 27 of 350 U series and 20 of 180 M series. A deterministic spectral-residual (SR) check on the 27 affected U series matches the published-window values for 25 and both rules for two; none matches only the current rule. This directly checks SR on those U series, not every detector or any M result.

Table 22 reports quantile-error and QPath VUS-PR under both window rules. Only published-window values enter the descriptive published-detector table. The other published U and M results are assumed to follow that rule; this is not direct verification. Forecasts, scores, labels, and series membership do not change, and equal-window cases reuse their existing metric values. The Table 5 stars denote arithmetic averages over the four backbone–forecast settings, not a single model execution.

Table 22: **Current versus published-window evaluation.** Full AUTO VUS-PR.

Data	Backbone	Forecast setting	Quantile error		QPath	
			Current	Published-window	Current	Published-window
U	Chronos-2	Zero-shot	0.31806	0.32815	**0.38908**	**0.40215**
U	Chronos-2	Fine-tuned	0.32605	0.33609	**0.39477**	**0.40790**
U	TimesFM	Zero-shot	0.32191	0.33182	**0.39378**	**0.40673**
U	TimesFM	Fine-tuned	0.32633	0.33639	**0.39987**	**0.41280**
M	Chronos-2	Zero-shot	0.34768	0.35685	**0.36717**	**0.37578**
M	Chronos-2	Fine-tuned	0.35084	0.36006	**0.36798**	**0.37661**
M	TimesFM	Zero-shot	0.34448	0.35356	**0.35271**	**0.36129**
M	TimesFM	Fine-tuned	0.34674	0.35574	**0.36225**	**0.37094**

Only the published-window columns enter Table 5. All other primary and ablation results use the current evaluator revision. Only the automatic-window rule differs between these columns.

The comparison includes LSTMAD, CNN, Lag-Llama, the original Chronos, and TimesFM 1.0. These versions and their one-step error scoring differ from the present backbones and horizon scoring. Other detector families in the reported snapshot score higher, including Time-RCD+MAFT on U (0.586) and StreamVAE on M (0.435). Thus the descriptive comparison does not claim state-of-the-art performance or attribute all differences to QPath.

G SOURCE AND FINE-TUNING-SEED DIAGNOSTICS

Tables 23 and 24 report every U and M source for FT Full evaluation. The headline score is an equal-weight mean over series; source means diagnose heterogeneity without changing the benchmark

composition. Counts differ substantially, including single-series sources. The diagnostics explain variation without changing the paired inference target.

Table 23: **All 23 U sources: FT Full differences.** AUTO VUS-PR.

Source	N	Chronos-2		TimesFM	
		QPath minus Quantile error	QPath minus MAE-QPath	QPath minus Quantile error	QPath minus MAE-QPath
CATSv2	1	+0.1312	+0.0075	+0.1332	+0.0107
Daphnet	1	+0.0336	+0.0014	+0.0415	+0.0018
Exathlon	30	+0.0488	−0.0037	+0.0438	−0.0036
IOPS	15	+0.1661	+0.0151	+0.1599	+0.0113
LTDB	8	−0.0586	+0.0175	−0.0811	+0.0058
MGAB	8	+0.0005	+0.0003	+0.0007	+0.0005
MITDB	7	−0.0133	+0.0018	−0.0244	+0.0038
MSL	7	+0.0384	+0.0081	+0.0232	−0.0031
NAB	23	+0.0370	+0.0192	+0.0339	+0.0287
NEK	8	+0.0047	+0.0088	−0.0104	−0.0112
OPPORTUNITY	27	−0.0169	+0.0042	−0.0127	+0.0071
Power	1	−0.0102	+0.0034	−0.0089	+0.0054
SED	2	+0.0038	+0.0014	+0.0081	+0.0028
SMAP	17	+0.0188	−0.0023	+0.0585	+0.0190
SMD	33	+0.1579	+0.0093	+0.1463	−0.0018
Stock	8	+0.0041	+0.0036	+0.0050	+0.0040
SVDB	18	−0.0054	+0.0156	−0.0463	+0.0233
SWaT	1	−0.0039	+0.0010	−0.0035	+0.0003
TAO	2	−0.0055	−0.0007	−0.0051	+0.0002
TODS	13	+0.0384	+0.0069	+0.0453	+0.0094
UCR	70	+0.1020	−0.0075	+0.1214	+0.0112
WSD	20	+0.1913	−0.0039	+0.2068	−0.0010
YAHOO	30	+0.0959	+0.0076	+0.1301	+0.0284

Source means are descriptive. The headline mean weights series equally across the complete benchmark.

Table 24: **All 17 M sources: FT Full differences.** AUTO VUS-PR.

Source	N	Chronos-2		TimesFM	
		QPath minus Quantile error	QPath minus MAE-QPath	QPath minus Quantile error	QPath minus MAE-QPath
CATSv2	5	+0.1249	+0.0026	+0.1392	+0.0222
CreditCard	1	+0.0004	+0.0018	−0.0002	+0.0018
Daphnet	1	+0.0806	−0.0097	+0.0807	−0.0075
Exathlon	25	+0.0802	+0.0104	+0.0980	+0.0019
GECCO	1	+0.2330	+0.0008	+0.2136	+0.0197
Genesis	1	+0.0021	+0.0004	−0.0002	+0.0007
GHL	23	+0.0002	+0.0001	−0.0004	+0.0001
LTDB	4	−0.0428	+0.0121	−0.0639	+0.0044
MITDB	11	−0.0050	+0.0000	−0.0225	+0.0003
MSL	14	−0.0073	−0.0153	−0.0197	−0.0038
OPPORTUNITY	7	+0.0014	−0.0003	+0.0009	−0.0023
PSM	1	+0.0027	+0.0009	+0.0044	+0.0038
SMAP	25	+0.0491	−0.0092	+0.0631	+0.0002
SMD	20	−0.0041	+0.0000	+0.0015	−0.0032
SVDB	28	−0.0253	+0.0205	−0.0538	+0.0258
SWaT	2	−0.0092	+0.0006	−0.0067	−0.0010
TAO	11	+0.0029	−0.0001	+0.0038	−0.0001

Source means are descriptive. The headline mean weights series equally across the complete benchmark.

Table 25 gives all three fine-tuning seeds. No seed is chosen for superior anomaly detection, and FT does not increase the number of independent series. Small seed variation measures robustness to this fine-tuning procedure, not independent confirmation on unseen sources. Table 26 reports wins,

ties, losses, and large-drop counts. Table 27 gives average precision (AP), area under the receiver operating characteristic curve (AUROC), and volume under the ROC surface (VUS-ROC) for FT; VUS-PR remains the primary metric.

Table 25: **Individual fine-tuning seeds.** Full mean AUTO VUS-PR and QPath–quantile error differences.

Data	Backbone	QPath			Δ versus Quantile error		
		42	43	44	42	43	44
U	Chronos-2	0.39481	0.39492	0.39458	+0.06877	+0.06896	+0.06844
U	TimesFM	0.39998	0.39992	0.39972	+0.07358	+0.07354	+0.07351
M	Chronos-2	0.36790	0.36795	0.36810	+0.01700	+0.01730	+0.01711
M	TimesFM	0.36172	0.36235	0.36269	+0.01491	+0.01564	+0.01597

All three seeds contribute to FT within each series. These are descriptive seed-specific means, not independent datasets or a seed-selection procedure.

Table 26: **Per-series gains and losses: QPath minus quantile error.** Full AUTO VUS-PR.

Data	Backbone	Forecast setting	Median Δ	Wins	Ties	Losses	$\Delta < -0.05$	$\Delta < -0.10$
U	Chronos-2	Zero-shot	+0.0167	256	4	90	10	6
U	Chronos-2	Fine-tuned	+0.0135	249	3	98	17	7
U	TimesFM	Zero-shot	+0.0105	241	3	106	27	8
U	TimesFM	Fine-tuned	+0.0133	251	3	96	26	8
M	Chronos-2	Zero-shot	+0.0017	124	4	52	12	8
M	Chronos-2	Fine-tuned	+0.0013	119	4	57	12	7
M	TimesFM	Zero-shot	+0.0003	98	5	77	20	11
M	TimesFM	Fine-tuned	+0.0002	96	3	81	22	12

Ties use tolerance 10^{-12}. Large-drop columns are nested counts, not additional disjoint categories. Every series contributes to the benchmark mean.

Table 27: **Auxiliary detection metrics.** FT, Full evaluation.

Metric	Data	Backbone	Quantile error	QPath	Δ	p
AP	U	Chronos-2	0.26181	**0.30309**	+0.04128	7.21×10^{-15}
AP	U	TimesFM	0.26334	**0.30636**	+0.04302	5.72×10^{-13}
AP	M	Chronos-2	0.29389	**0.30680**	+0.01291	0.0116
AP	M	TimesFM	0.29046	**0.30024**	+0.00978	0.0983
AUROC	U	Chronos-2	0.72258	**0.75759**	+0.03502	1.13×10^{-10}
AUROC	U	TimesFM	0.71674	**0.75173**	+0.03499	3.69×10^{-10}
AUROC	M	Chronos-2	0.66675	**0.67401**	+0.00727	0.0826
AUROC	M	TimesFM	0.66378	**0.67234**	+0.00856	0.0350
VUS-ROC	U	Chronos-2	0.78567	**0.81489**	+0.02922	1.06×10^{-10}
VUS-ROC	U	TimesFM	0.78046	**0.81225**	+0.03179	6.94×10^{-11}
VUS-ROC	M	Chronos-2	0.71762	**0.72522**	+0.00760	0.0804
VUS-ROC	M	TimesFM	0.71282	**0.72303**	+0.01021	0.0185

Δ is QPath minus quantile error. AP is average precision, not trapezoidal PR-AUC. VUS-ROC uses automatic window selection. Paired p values are nominal. U uses 350 series and M uses 180.