Knowing When Multivariate Forecasts Are Wrong

Binli Luo* **Ning Gui*** **Xianhan Tan**
Central South University
lblhandsome@gmail.com, ninggui@gmail.com, txhts1606@gmail.com

Abstract

At deployment, a multivariate forecaster may expose only a recent input window and a frozen point prediction, with no access to ground truth, calibration residuals, repeated inference, or model internals. We study this deployment-constrained forecast risk localization problem: ranking future timesteps by likely error before they are observed. We introduce SUBDX, a zero-parameter diagnostic that audits the forecast against the local cross-channel subspace of the input history. Each predicted channel vector receives an off-subspace residual score, computed with no labels and no learned parameters; a rank-32 low-rank variant adds about 1.7% parameters on Traffic. Under a linear factor model, we derive closed-form score distributions and a two-population AUROC expression showing how heterogeneous off-subspace error visibility makes the ranking identifiable. On Traffic ($N = 862$), SUBDX reaches AUROC 0.84–0.87 across three backbones and reduces MSE by 65%–69% at 50% coverage; on Solar ($N = 137$), it reaches AUROC 0.78–0.92 with 65%–92% MSE reduction. The same post-hoc score applies to frozen pretrained forecasters on a matched Solar protocol, with Chronos reaching AUROC 0.578 without fine-tuning. Across trained backbones, pretrained models, and synthetic controls, the results show that local cross-channel consistency can provide a practical abstention signal for redundant multivariate systems, while simple controls remain stronger in low-visibility regimes.

1 Introduction

Modern time-series forecasters now range from specialized architectures to frozen foundation models [Lim et al., 2021, Zhou et al., 2021, Wu et al., 2021, Zhou et al., 2022, Nie et al., 2023, Liu et al., 2024, Das et al., 2024, Woo et al., 2024, Ansari et al., 2024, Shi et al., 2025]. Yet deployed systems may expose only point forecasts. Accuracy can be measured after observations arrive; at decision time, the question is which future timesteps to trust.

Existing reliability tools assume information that this setting may not provide. Probabilistic forecasters learn predictive distributions [Salinas et al., 2020, Lim et al., 2021]; ensembles and MC Dropout require repeated inference or architectural access [Lakshminarayanan et al., 2017, Gal and Ghahramani, 2016]; conformal methods need calibration residuals and return intervals rather than per-timestep risk rankings [Romano et al., 2019, Stankeviciute et al., 2021]; learned selectors and test-time revision methods use labeled errors, delayed ground truth, or extra trainable components [Geifman and El-Yaniv, 2017, 2019, Liu et al., 2025, Im and Kwon, 2026]. We study *deployment-constrained forecast risk localization*: given only the recent window $\mathbf{X}$ and a frozen point prediction $\hat{\mathbf{Y}}$, rank future timesteps by risk before observing their outcomes.

Our source of signal is cross-channel consistency. Channel-independent models can be strong predictors [Zeng et al., 2023, Nie et al., 2023, Liu et al., 2024], but redundant multivariate systems still have latent cross-channel geometry. Rather than incorporating this geometry into model training,

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

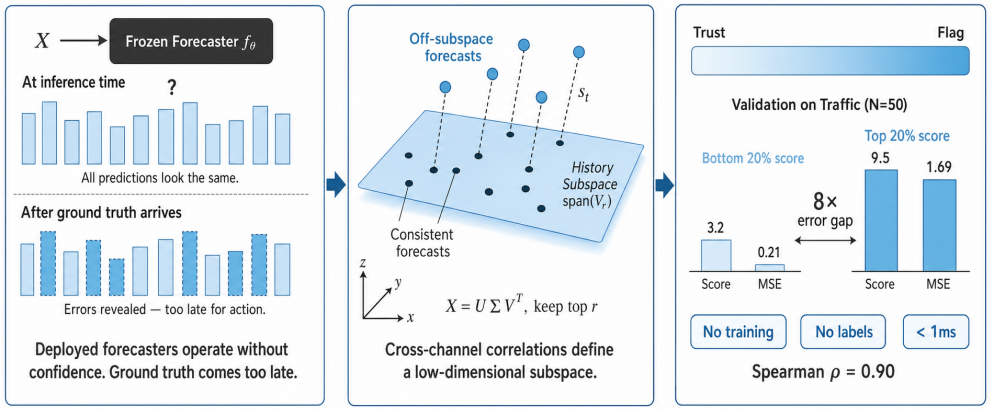

Figure 1: Overview of SUBDX. A frozen forecaster emits point predictions; the recent input window defines a local cross-channel subspace; off-subspace energy gives one label-free risk score per predicted timestep (Traffic example, $N = 50$, Spearman $\rho = 0.90$).

we leverage it post-hoc to audit the predicted trajectories. A forecasted channel vector that departs sharply from the recent low-dimensional channel subspace is structurally implausible under the cross-channel consistency assumption, even before the true future is known.

Figure 1 illustrates the diagnostic. Given $\mathbf{X} \in \mathbb{R}^{L \times N}$, we estimate the dominant channel subspace with truncated SVD, retaining $\mathbf{U} \in \mathbb{R}^{N \times r}$. Each predicted vector $\hat{\mathbf{y}}_t$ receives the off-subspace score

$$s_t = \left\| (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\hat{\mathbf{y}}_t \right\|_2 \tag{1}$$

which requires one SVD and one projection per timestep (a rank-32 low-rank variant adds about 1.7% parameters on Traffic), with no labels, retraining, or access to model internals. Crucially, high ambient dimensionality alone does not guarantee an effective diagnostic. Our linear factor-model analysis explicitly characterizes the operating regime : SUBDX works when errors differ in off-subspace visibility and fails when that visibility is uniform. Empirically, Traffic ($N = 862$) and Solar ($N = 137$) show 65%–69% and 65%–92% MSE reductions at 50% coverage across three backbones, Chronos reaches AUROC 0.578 on a matched Solar protocol without fine-tuning (Section 4.3), and scores correlate across backbones.

Problem contribution. We formulate *deployment-constrained forecast risk localization* as a problem distinct from probabilistic forecasting, conformal calibration, learned selection, and test-time revision. The target is a pre-observation ranking of future timesteps under the constraints that only $\mathbf{X}$ and $\hat{\mathbf{Y}}$ are available: no labels, no model internals, and no repeated inference.

Method and analysis contribution. We introduce a zero-parameter subspace diagnostic that turns redundancy in the input history into an audit signal for the forecast itself. The diagnostic uses cross-channel structure even when the base forecaster is channel-independent, frozen, or externally provided. A factor-model analysis identifies heterogeneous off-subspace error visibility, rather than dimensionality alone, as the success condition and characterizes the corresponding failure mode.

Empirical studies. We evaluate this regime across datasets, backbones, foundation models, channel subsampling, synthetic controls, and correction baselines. The results show what the scalar score solves and what it does not: it supports abstention and monitoring in redundant multivariate systems, while six parameter-free correction mechanisms fail even with oracle selection.

2 Related Work

Multivariate forecasting and cross-channel geometry. Modern forecasters span autoregressive and decomposition Transformers Salinas et al. [2020], Lim et al. [2021], Zhou et al. [2021], Wu et al. [2021], Zhou et al. [2022], linear, patch, and inverted-token architectures Zeng et al. [2023], Nie

et al. [2023], Liu et al. [2024], and frozen foundation models Ansari et al. [2024], Das et al. [2024], Woo et al. [2024], Shi et al. [2025]. Yet deployed pipelines frequently expose only point predictions, providing no pre-observation signal to identify which future timesteps will fail. Moreover, while prior work debates channel-mixing versus channel-independent training Zeng et al. [2023], Nie et al. [2023], Liu et al. [2024], both camps treat inter-channel correlation strictly as a training bias for *generation*, overlooking its utility as a test-time geometric constraint for *auditing* black-box outputs.

Predictive uncertainty and calibration. Existing reliability tools relax at least one constraint of zero-label, black-box deployment. Probabilistic forecasters Salinas et al. [2020], Lim et al. [2021] require specialized training objectives, whereas MC Dropout and deep ensembles Gal and Ghahramani [2016], Lakshminarayanan et al. [2017] need architectural access or repeated forward passes. Conformal methods Romano et al. [2019], Stankeviciute et al. [2021] avoid retraining but require labeled calibration residuals and target marginal interval coverage rather than within-horizon risk ranking, leaving a gap when only a single lookback window and point forecast are available.

Selective prediction and post-hoc revision. Selective classification studies risk-coverage abstention Geifman and El-Yaniv [2017, 2019] using learned confidence heads or supervised calibration. Similarly, time-series refinement and adaptation methods—such as Scaleformer Shabani et al. [2023], PIR Liu et al. [2025], and COSA Im and Kwon [2026]—rely on extra trainable modules, historical error memory, or delayed ground truth. By focusing on trajectory repair, these works also conflate error *localization* with *correction*, overlooking that scalar structural violations can reliably guide abstention even when parameter-free revision fails.

Subspace monitoring versus forecast auditing. Classical statistical monitoring—via Hotelling's T^2 Hotelling [1931], PCA SPE/Q-statistics Jackson and Mudholkar [1979], or Robust PCA Candès et al. [2011]—projects realized observations onto a static subspace learned offline from training data. Such offline designs cannot operate when training corpora are unavailable or non-stationary at test time. Furthermore, classical monitoring lacks a theoretical characterization of forecast-error ranking, failing to explain why subspace auditing succeeds under heterogeneous error footprints but collapses under isotropic, channel-independent errors. Deep anomaly detectors such as Anomaly Transformer Xu et al. [2022], DCdetector Yang et al. [2023], TranAD Tuli et al. [2022], PatchAD Zhong et al. [2025], and GDN Deng and Hooi [2021] learn from normal training data and score observed windows; they differ in scored object (realized observation vs. frozen forecast) and access contract (training data and learned parameters vs. current history and one point forecast). We do not claim a new SPE/Q statistic or residual formula; the contribution is the pre-outcome, structure-dependent visibility law and the deployment contract that makes a frozen forecast auditable without labels or model internals.

3 Method

3.1 Problem Setup

Let a frozen multivariate forecaster f_θ take an input window $\mathbf{X} \in \mathbb{R}^{L \times N}$ and produce a prediction horizon $\hat{\mathbf{Y}} = f_\theta(\mathbf{X}) \in \mathbb{R}^{P \times N}$. At inference time we observe $\mathbf{X}$ and $\hat{\mathbf{Y}}$, but not the future $\mathbf{Y}$. The diagnostic must output one scalar score s_t for each forecasted timestep $t \in \{1, \dots, P\}$, where larger scores indicate higher expected error $e_t = \frac{1}{N}\|\hat{\mathbf{y}}_t - \mathbf{y}_t\|_2^2$. The goal is not to estimate a full predictive distribution; it is to rank future timesteps by risk before labels or residuals exist.

The setting imposes three restrictions. **Label-free** means that neither current errors nor calibration residuals are available when the score is produced. **Post-hoc** means that the forecaster is treated as a black box beyond its point predictions: no retraining, no architectural modification, and no access to hidden states or gradients. **Low overhead** means that the diagnostic should be comparable to a lightweight preprocessing step, not to an ensemble or repeated stochastic inference. These restrictions rule out many standard uncertainty estimators. They still allow one check: whether the prediction itself is geometrically consistent with the recent multivariate history.

3.2 Local Channel Subspace Estimation

SUBDX extracts a local linear representation of cross-channel couplings directly from the lookback window X. To ensure scale invariance across heterogeneous sensors, we evaluate the diagnostic in

the standardized channel space:

$$\tilde{X}_{\ell n} = \frac{X_{\ell n} - \mu_n}{\sigma_n + \epsilon}, \qquad \tilde{\hat{Y}}_{tn} = \frac{\hat{Y}_{tn} - \mu_n}{\sigma_n + \epsilon}, \tag{2}$$

where μ_n and σ_n denote the channel-wise mean and standard deviation (inherited from the forecasting pipeline or computed locally over X), and $\epsilon > 0$ ensures numerical stability; exact train-constant channels use scale one (Appendix D.1). For notational simplicity, we drop the tildes below and write X and $\hat{Y}$ for the standardized matrices.

Treating $X \in \mathbb{R}^{L \times N}$ as L temporal observations in an N-dimensional channel space, we compute its rank-r truncated singular value decomposition (SVD) and retain the top-r right singular vectors:

$$X = A\Sigma V^\top, \qquad U = V_{:,1:r} \in \mathbb{R}^{N \times r}, \qquad U^\top U = I_r, \tag{3}$$

where the columns of U span the dominant r-dimensional channel subspace of the recent history, yielding the orthogonal projectors

$$P = UU^\top, \qquad P_\perp = I_N - UU^\top. \tag{4}$$

Estimating U per window eliminates any dependency on the original training corpus, naturally tracks gradual regime shifts, and applies seamlessly to third-party foundation models. While offline PCA over the full training set can offer slightly lower estimation variance when historical data are accessible, local SVD is the canonical choice for strict zero-data deployment (see Section 4.5 and Appendix R for empirical comparisons).

3.3 Off-Subspace Diagnostic

For each forecast horizon step $t \in \{1, \ldots, P\}$, SUBDX quantifies structural implausibility via the Euclidean norm of the prediction's orthogonal complement:

$$s_t = \|\mathbf{P}_\perp \hat{\mathbf{y}}_t\|_2 = \left\|(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\hat{\mathbf{y}}_t\right\|_2. \tag{5}$$

In practice, we never materialize the dense $N \times N$ matrix $P_\perp$. Instead, across the entire horizon $\hat{Y} \in \mathbb{R}^{P \times N}$, the scores are computed in batch via low-rank matrix products:

$$Z = \hat{Y}U \in \mathbb{R}^{P \times r}, \qquad R = \hat{Y} - ZU^\top \in \mathbb{R}^{P \times N}, \qquad s_t = \|R_{t,:}\|_2, \tag{6}$$

which can also be verified via the Pythagorean identity $s_t^2 = \|\hat{y}_t\|_2^2 - \|U^\top \hat{y}_t\|_2^2$. Notice that $s_t = 0$ when $\hat{y}_t$ lies entirely within $\mathrm{span}(U)$ and increases as the forecast violates historical inter-channel correlations. We adopt the unweighted residual norm throughout, as our ablation studies (Appendix M.3) confirm that singular-value weighting, Hotelling's T^2 in-subspace statistics, and Robust PCA decompositions yield no systematic gains over Eq. (6). Because s_t measures relative structural violation rather than calibrated coverage, it is designed for selective abstention and shift monitoring rather than point revision.

3.4 Concrete Implementation and Complexity

For each test window, the diagnostic reduces to a compact linear-algebra routine:

$$\begin{aligned} (\mathbf{A}, \mathbf{\Sigma}, \mathbf{V}) &= \mathrm{svd}_r(\tilde{\mathbf{X}}), & \mathbf{U} &= \mathbf{V}_{:,1:r}, \\ \boldsymbol{R} &= \hat{\mathbf{y}} - (\hat{\mathbf{y}}\mathbf{U})\mathbf{U}^\top, & s_t &= \|\boldsymbol{R}_{t,:}\|_2, \quad t = 1, \ldots, P. \end{aligned} \tag{7}$$

The truncation rank is fixed to $r = 8$ in all main experiments and is clipped as $r_{\text{eff}} = \max(0, \min(r, N-1, L-1))$; for $N = 7$, a requested rank of 8 becomes 6, preserving a non-empty residual complement. Because scores are used strictly to rank timesteps within each configuration, no learned calibration or dataset-specific threshold is needed.

Truncated SVD costs $O(LNr)$ and the projections cost $O(PNr)$, with $r \ll N$. On Traffic ($L = 96, N = 862, P = 96$), this adds less than 1% runtime overhead relative to the backbone forward pass, requiring no learned parameters, validation splits, or additional model calls.

3.5 Why the Score Can Rank Errors

The mechanism is clearest under a linear factor model. Suppose the true future satisfies $\mathbf{y}_t = \mathbf{V}\mathbf{z}_t + \boldsymbol{\varepsilon}_t$, where $\mathbf{V} \in \mathbb{R}^{N\times d}$ spans a low-dimensional channel subspace and $\boldsymbol{\varepsilon}_t$ is isotropic observation noise. Writing the forecast error as $\boldsymbol{\delta}_t = \hat{\mathbf{y}}_t - \mathbf{y}_t$, the score is governed by the off-subspace component of the prediction:

$$\mathbf{P}_\perp \hat{\mathbf{y}}_t = \mathbf{P}_\perp \boldsymbol{\varepsilon}_t + \mathbf{P}_\perp \boldsymbol{\delta}_t. \tag{8}$$

Consequently, s_t is informative whenever forecast errors vary in their off-subspace footprint across timesteps. We term this footprint *visibility*: errors aligned with the dominant subspace $\mathrm{span}(\mathbf{V})$ remain largely invisible to SUBDX, whereas errors that inject energy into the $(N-d)$-dimensional orthogonal complement directly inflate s_t. A larger off-subspace dimension $k = N - d$ amplifies this separation, though it cannot generate ranking signal if error visibility is uniform across the horizon.

Under this factor model, the squared score $q_t = s_t^2$ follows a scaled noncentral χ^2 distribution, and the two-population AUROC admits a closed form via the $F_{k,k}$ distribution (proofs in Appendix E). This analysis yields three testable predictions: a larger off-subspace dimension k sharpens discrimination, temporal variance in error visibility is necessary for separation, and observation noise dilutes the signal without altering the underlying geometric mechanism. Experiments in Section 4 evaluate all three predictions.

4 Experiments

We empirically test the three theoretical predictions from Section 3.5: (i) whether a larger off-subspace dimension k sharpens error discrimination across real benchmarks (Section 4.2), channel subsampling (Section 4.6), and synthetic sweeps (Appendix Q); (ii) whether temporal variation in error visibility is necessary, probed via channel-independent DLinear controls (Section 4.2) and model-quality ablations (Section 4.5); and (iii) whether the score captures genuine cross-channel geometry rather than horizon-position or backbone artifacts (Sections 4.4, 4.5).

4.1 Setup

Benchmarks and baselines. We evaluate on six public benchmarks (N from 7 to 862) plus SWaT ($N = 51$) as an industrial probe, using PatchTST, iTransformer, and DLinear trained from scratch and scored post-hoc. Baselines include deep ensembles, MC Dropout, forecast magnitude/change, input volatility, horizon index, and a supervised Ridge oracle. Dataset splits and hyperparameters are detailed in Appendix D, deployment constraint comparisons in Appendix A, and scoring pseudocode in Appendix C (Figure 5).

Evaluation protocol and metrics. We evaluate SUBDX under selective forecasting: at coverage level α, we accept the lowest-scored α-fraction of predicted timesteps and abstain on the rest. We report four metrics: (i) **AUROC** for identifying high-error timesteps (where e_t exceeds the 80th percentile); (ii) Spearman rank correlation $\boldsymbol{\rho}$ between s_t and e_t; (iii) relative **AURC gain** over random abstention based on the Area Under the Risk-Coverage curve; and (iv) **Sel@50** (or Sel@0.5), the percentage MSE reduction achieved at 50% coverage.

4.2 Main Results

Table 1 reports the protocol-locked comparison: every cell uses the producing loader's channel order, training-split normalization, row-wise score, q80 labels, and timestep pooling over the complete frozen test sequence. SUBDX is strongest among same-interface methods in all six Solar/Traffic cells (AUROC 0.779–0.919; Sel@50 65%–92%), where recent history supplies a stable cross-channel constraint. It is close to the best simple control on ETTh2 (0.593–0.617 vs. 0.630–0.632) and slightly above on Electricity/PatchTST (0.598 vs. 0.596), but simple controls are stronger on ETTh1, Weather, Electricity/DLinear, and SWaT. In SWaT, persistence reaches 0.884 while SUBDX reaches 0.604; fixed PCA, an offline-history reference, is also stronger in several cells. The pattern is conditional rather than universal: the score adds a structural signal when errors leave a stable local subspace.

Crucially, large ambient space is insufficient without temporal variation in error visibility ($\tau_1^2 \neq \tau_0^2$), as DLinear on Electricity versus Solar shows. Electricity has $N = 321$, but DLinear's channel-

independent predictions yield nearly isotropic off-subspace errors regardless of magnitude, and AUROC is 0.520. The same backbone reaches 0.779 on Solar ($N = 137$; PatchTST 0.919): Solar's tight factor structure (effective rank $\sim$4) makes large errors violate the manifold while minor errors do not. Thus, k acts only as a statistical amplifier, while heterogeneity in off-subspace error footprints supplies the signal. A 192-cell synthetic controlled factorial isolates this mechanism: with error prevalence and magnitude fixed, the visibility contrast (off-subspace error component 5% vs. 75%) raises AUROC from 0.626 to 0.833 and lowers FN@50 from 73.6% to 31.5%, whereas ambient N from 16 to 128 changes AUROC only from 0.722 to 0.734 (Appendix V).

Because this mechanism is anchored in data geometry rather than training heuristics, it is stable to initialization and rollout length in the original per-backbone protocol (Tables 12, 14): multi-seed AUROC standard deviations stay below 0.007 (Appendix F), and Traffic AUROC remains above 0.85 across horizons $P \in \{24, 96, 192\}$ (Appendix H).

Table 1: Protocol-locked error localization (q80 labels, full frozen test sequence). AUROC is shown for SUBDX and for the strongest of five simple controls (Mag., Last, Var., Hist-vol, Hist-chg). ρ, Sel@50, and FN@50 are for SUBDX. Bold marks cells where SUBDX exceeds the best simple control. SWaT/DLinear is the mean over seeds 42–44; other cells use seed 42. Fixed PCA, an offline-history reference, is reported in Appendix Table 31.

Backbone	Dataset	N	SUBDX	Best control	ρ	Sel@50 (%)	FN@50 (%)
DLinear	ETTh1	7	0.503	0.732 (Mag.)	0.026	0.7	49.6
	Electricity	321	0.520	0.616 (Hist-vol)	0.111	3.8	48.6
	Solar	137	**0.779**	0.766 (Var.)	0.640	71.7	10.5
	Traffic	862	**0.840**	0.804 (Mag.)	0.792	65.3	8.8
	Weather	21	0.543	0.697 (Last)	0.003	6.4	43.6
	SWaT	51	0.604	0.884 (Last)	0.207	11.1	34.8
PatchTST	ETTh2	7	0.593	0.630 (Last)	0.242	17.4	38.4
	Electricity	321	**0.598**	0.596 (Hist-chg)	0.252	11.5	38.3
	Solar	137	**0.919**	0.789 (Last)	0.874	92.3	4.1
	Traffic	862	**0.867**	0.815 (Mag.)	0.809	69.3	7.5
	Weather	21	0.569	0.664 (Last)	0.162	6.3	42.1
iTransformer	ETTh2	7	0.617	0.632 (Last)	0.284	25.0	33.7
	Electricity	321	0.563	0.600 (Hist-vol)	0.202	10.6	42.3
	Solar	137	**0.801**	0.735 (Last)	0.458	65.4	15.1
	Traffic	862	**0.854**	0.817 (Mag.)	0.783	66.4	8.2
	Weather	21	0.561	0.749 (Last)	0.128	8.5	42.8

Comparison against baselines. Appendix Table 31 reports all five simple controls and the offline fixed-PCA reference. On Traffic, SUBDX exceeds the best simple control by 0.036–0.052 across the three backbones; on Weather, forecast-to-last is stronger, reflecting the limited off-subspace capacity at small k. Deep ensembles and supervised Ridge are compared in Appendix I; full signal comparisons are in Appendix P. Unless stated otherwise, the ablation and diagnostic analyses below use the original single-cell protocol; their absolute AUROC values are not directly comparable to the protocol-locked Table 1.

4.3 Foundation Models

Applying SUBDX directly to frozen pretrained forecasters—Chronos-T5-Small [Ansari et al., 2024], TimesFM 2.5 [Das et al., 2024], and Time-MoE [Shi et al., 2025]—under a matched Solar protocol (240 windows, $P = 64$, 10 channels, legal rank 8) shows that zero-shot error geometry varies across models (Table 2). Chronos retains the strongest signal (AUROC 0.578, Sel@50 gain 25.0%), whereas TimesFM 2.5 and Time-MoE rank below random (0.422 and 0.418). In the weak cells, false-negative rows are not more subspace-aligned than caught errors; instead they exhibit stronger perpendicular target-error cancellation (mean cosine -0.995 and -0.975). SUBDX is therefore agnostic to how point predictions are generated, but its diagnostic signal depends on the error geometry of the pretrained model rather than on pretraining lineage.

Model	AUROC	ρ	Sel@50	FN@50
TimesFM 2.5	0.422	0.178	−9.2	62.8
Time-MoE	0.418	0.138	−9.4	64.3
Chronos-T5-S	0.578	0.428	25.0	44.9

Table 2: SUBDX on frozen foundation models (matched Solar protocol, rank 8); Sel/FN are percentages at 50% coverage.

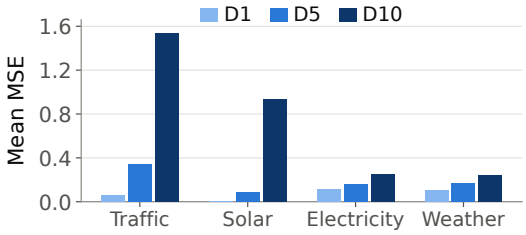

Figure 2: Mean MSE across score deciles; D1 is the lowest-score decile and D10 the highest.

4.4 Selective Forecasting and Diagnostics

Retaining the lowest-scored 10% of predictions on Traffic (PatchTST) slashes MSE by 89%, while retaining 50% reduces error by 62%—substantially outperforming Deep Ensembles, MC Dropout, and the flat horizon-index baseline (Figure 3).

Random rejection and decile monotonicity. Unlike random abstention, which yields zero gain across all coverage levels, SUBDX-guided selection at 50% coverage reduces MSE by 61.5% on Traffic, 16.4% on Electricity, and 15.4% on Weather (Appendix O). Sorting test predictions into score deciles confirms strict monotonicity: realized MSE surges from the lowest- to highest-score decile by 27× on Traffic and 155× on Solar (Figure 2; full decile profiles in Appendix T).

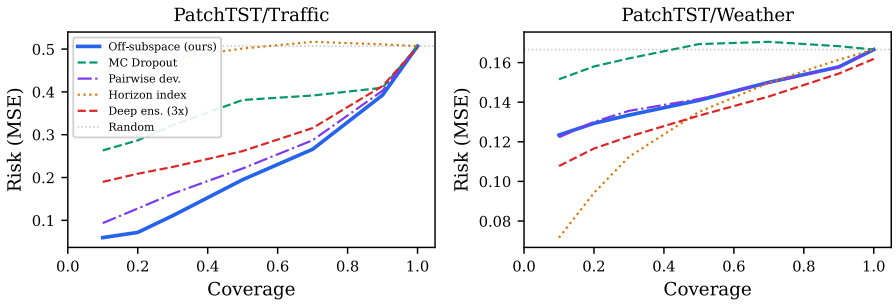

Figure 3: Risk-coverage curves for PatchTST. Traffic benefits sharply from low-score acceptance; Weather is flatter because fewer redundant off-subspace directions are available.

Distribution shift monitoring. Beyond selective forecasting, SUBDX serves as a label-free monitor for cross-channel consistency violations (Table 6). Under additive Gaussian noise ($\sigma = 0.5\times$ channel std) on PatchTST/Traffic, diagnostic scores increase by 44% while MSE rises by only 19%. Conversely, structural input shifts (such as channel permutation or time reversal) inflate error metrics without triggering scores, because the local SVD dynamically adapts to the altered input geometry. Meanwhile, channel dropout mechanically reduces k and suppresses scores, indicating that SUBDX should be paired with basic sensor-availability checks in deployment (Appendix M.2).

Correction failure. Although s_t reliably flags high-risk timesteps, its scalar nature provides no directional gradient in $\mathbb{R}^N$ for point repair. Evaluating six parameter-free mechanisms on the highest-scored 20% of timesteps yields at most a 0.5% MSE reduction—and no more than 1.1% even when using an oracle error selector (Table 18, Appendix M.1). This confirms that scalar subspace residuals are fundamentally tailored for selective abstention rather than parameter-free trajectory correction.

Cross-backbone consistency. Diagnostic scores exhibit cross-architecture alignment, achieving Spearman rank correlations of $\rho = 0.90$ (Traffic) and 0.86 (Solar) between PatchTST and iTransformer, and 0.91 between DLinear and PatchTST on Solar (partial-correlation checks in Appendix N).

4.5 Diagnostic Decomposition

Within-horizon analysis. Computing AUROC independently per step confirms that SUBDX's efficacy stems from genuine error discrimination rather than merely proxying forecast horizon position t (Table 3). Within-horizon performance remains remarkably robust (e.g., 0.900 on Traffic), demonstrating that the diagnostic captures fine-grained, step-wise anomalies across the prediction window.

Table 3: Within-horizon analysis: AUROC computed per-step then averaged eliminates horizon position confound.

Backbone	Dataset	Pooled AUROC	Within-H AUROC	Within-H ρ
PatchTST	Traffic	0.893	0.900	0.785
	Solar	0.850	0.832	0.612
	Electricity	0.691	0.721	0.446
	Weather	0.652	0.586	0.180
	ETTh1	0.569	0.569	0.143
iTransformer	Traffic	0.862	0.870	0.716
	Solar	0.778	0.774	0.561
	Weather	0.618	0.586	0.174

Threshold and bootstrap checks. Error rankings exhibit strong invariance across various risk thresholds and tight statistical bounds (Table 4), confirming that the evaluated abstention signals are stable and statistically reliable rather than artifacts of specific training splits or cutoff choices(extended PR-AUC, Kendall's τ, and bootstrap intervals in Appendices J and S).

Table 4: Threshold and bootstrap checks for PatchTST. AUROC@q uses the qth error percentile as the high-error label; CI is a bootstrap 95% interval for AUROC.

Dataset	AUROC@25	AUROC@50	AUROC@75	Spearman ρ	Bootstrap AUROC CI
Traffic	0.898	0.893	0.862	0.770	[0.853, 0.856]
Electricity	0.682	0.691	0.690	0.378	[0.683, 0.687]
Weather	0.701	0.652	0.633	0.329	[0.627, 0.629]
ETTh1	0.575	0.569	0.567	0.142	[0.567, 0.571]

Model error vs. timestep difficulty. Controlled comparisons across well-trained, undertrained, and structurally corrupted backbones (Table 5) reveal that SUBDX successfully distinguishes model-induced failures from intrinsic data volatility. When cross-channel integrity is artificially destroyed (via channel shuffling), the diagnostic registers massive structural violations while controlling for baseline future geometry, confirming its sensitivity to genuine model degradation.

Table 5: Model quality control: SUBDX across PatchTST variants on the same Traffic test set. Partial ρ controls for the true future's off-subspace distance.

Variant	MSE	AUROC@50	ρ	Partial ρ
Well-trained (20 ep)	0.509	0.893	0.770	0.144
Undertrained (5 ep)	0.515	0.851	0.682	-0.027
Shuffled channels	1.184	0.964	0.894	0.681

Local SVD vs. fixed PCA. While offline global PCA leverages full training corpora to marginally boost performance in stationary high-N regimes, it violates strict zero-data deployment and degrades under non-stationary shifts (Table 7a). Per-window SVD thus provides the necessary adaptive default for black-box environments.

Score-component ablation. Component-wise decomposition demonstrates that the unweighted off-subspace residual norm (SPE) captures virtually all diagnostic power (Table 7b). Elaborate weighting

schemes, in-subspace Hotelling statistics (T^2), and Robust PCA variants yield no systematic gains, confirming that simple geometric projection onto the orthogonal complement is both necessary and sufficient(Appendix M.3).

4.6 Channel Subsampling and Synthetic Validation

Within-dataset channel subsampling directly tests k-amplification. On Traffic, AUROC rises monotonically from 0.617 ($N = 10$) to 0.846 ($N = 400$) as additional channels expand the orthogonal complement (Figure 4). On Solar, AUROC climbs from 0.70 ($N = 10$) to 0.86 ($N = 35$) before plateauing once its compact factor structure ($d \sim 4$) is saturated.

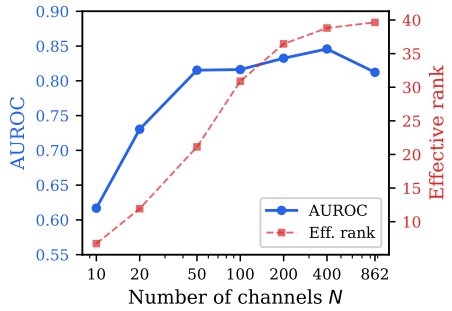

Figure 4: Channel-subsampling sweep on Traffic. AUROC peaks around $N = 400$ as rank grows.

Table 6: Monitoring controlled shifts on PatchTST/Traffic.

Shift	ΔSubDx (%)	ΔMSE (%)	ρ
None	0.0	0.0	0.38
Noise 0.1	+2.4	+0.7	0.38
Noise 0.5	+43.7	+18.5	0.28
Noise 1.0	+110.0	+60.7	0.00
Channel perm.	+8.6	+139.4	0.68
Time reversal	+3.5	+284.7	0.07
Dropout 30%	−18.3	+56.8	0.39
Dropout 70%	−49.3	+132.6	0.34

Table 7: Additional checks. Left: local SVD vs. fixed training-set PCA (AUROC; Δ = SVD minus PCA). Right: signal decomposition (AUROC/AURC gain in %).

(a) Left: local SVD vs. fixed PCA

Dataset	SVD	Fixed PCA	Δ
Traffic	0.856	0.881	−0.026
Solar	0.898	0.907	−0.008
Electricity	0.691	0.778	−0.087
Weather	0.612	0.525	**+0.087**

(b) Right: signal decomposition

Signal	Weather	Traffic
SPE	0.625 / 21.5	0.856 / 59.2
T^2	0.501 / −7.2	0.814 / 42.5
T^2+SPE	0.517 / −2.6	0.815 / 42.9
SPE+RPCA	0.623 / 21.3	0.856 / 59.2

Finally, controlled synthetic experiments (Appendix Q) directly validate Theorem 1: empirical AUROC matches the theoretical $F_{k,k}$ curve across sweeps of k, error visibility ratio τ_1/τ_0, and observation noise σ_y (ranging from 0.57 under low k and narrow visibility to 1.00 under high k and wide spread). Stress tests further confirm that at large k, short lookback windows ($L = 10$), imperfect truncation rank r (consistent with real-data rank sweeps in Appendix G), and heavy-tailed Laplace/Uniform noise have negligible impact on ranking accuracy.

5 Conclusion

SubDx audits frozen multivariate point forecasts against the local cross-channel subspace of the input history, requiring no labels, calibration residuals, or model internals. Across trained backbones, foundation models, and synthetic controls, heterogeneous off-subspace visibility establishes a diagnostic signal; SWaT marks the boundary where persistence dominates. Rather than correcting errors, it provides a lightweight, plug-and-play signal for abstention and real-time distribution monitoring.

Limitations SubDx relies on the input history having a non-trivial cross-channel structure; when this structure is weak or errors are channel-independent, the score carries little information. The analysis assumes a linear factor model and an accurately estimated subspace. The score flags risky timesteps but gives no error direction, supporting abstention and monitoring rather than correction. A seven-cell full-population audit yields 25.3% pooled FN@50, and the 15-cell locked panel yields 28.37% (Appendix U).

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Grant No. 62476289). We would also like to thank the anonymous reviewers for their constructive feedback and suggestions.

References

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.

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*, pages 10148–10167. PMLR, 2024.

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, pages 4027–4035, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in Neural Information Processing Systems*, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, 2019.

Harold Hotelling. The generalization of Student's ratio. *Annals of Mathematical Statistics*, 2(3): 360–378, 1931.

Jeonghwan Im and Hyuk-Yoon Kwon. COSA: Context-aware output-space adapter for test-time adaptation in time series forecasting. In *International Conference on Learning Representations*, 2026.

J. Edward Jackson and Govind S. Mudholkar. Control procedures for residuals associated with principal component analysis. *Technometrics*, 21(3):341–349, 1979.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

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

Zhiding Liu, Mingyue Cheng, Guanhao Zhao, Jiqian Yang, Qi Liu, and Enhong Chen. Improving time series forecasting via instance-aware post-hoc revision. In *Advances in Neural Information Processing Systems*, 2025.

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

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pages 3538–3548, 2019.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020. doi: 10.1016/j.ijforecast.2019.07.001.

Mohammad Amin Shabani, Amir H. Abdi, Lili Meng, and Tristan Sylvain. Scaleformer: Iterative multi-scale refining transformers for time series forecasting. In *International Conference on Learning Representations*, 2023.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations*, 2025.

Kamile Stankeviciute, Ahmed M. Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. In *Advances in Neural Information Processing Systems*, 2021.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 15(6): 1201–1214, 2022.

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*, pages 53140–53164. PMLR, 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022.

Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. DCdetector: Dual attention contrastive representation learning for time series anomaly detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI Conference on Artificial Intelligence*, 2023.

Zhijie Zhong, Zhiwen Yu, Yiyuan Yang, Weizheng Wang, Kaixiang Yang, and C. L. Philip Chen. PatchAD: A lightweight patch-based MLP-mixer for time series anomaly detection. *IEEE Transactions on Big Data*, 11(6):3460–3473, 2025.

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

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27268–27286. PMLR, 2022.

A Deployment Constraints Table

Table 8 summarizes the deployment constraints of the compared approaches.

Table 8: Deployment constraints. Each approach relaxes one constraint but tightens another. SubDx satisfies all three.

Method	Label-free	Post-hoc	Low overhead
Deep ensembles	Yes	No	No
MC Dropout	Yes	Yes	No
Probabilistic forecasting	Yes	No	Yes
Conformal prediction	No	Yes	Yes
Learned revision (PIR, COSA)	No	No	Yes
SubDx (ours)	Yes	Yes	Yes

B Notation Table

Table 9 lists the main symbols; additional quantities are defined with the corresponding theoretical statements.

Table 9: Main notation for the theoretical analysis.

Symbol	Meaning
N	Ambient dimension (number of channels)
d	True rank (intrinsic dimensionality)
r	Truncation rank ($r = 8$ in main experiments)
k	Off-subspace dimension ($N - d$ in theory, $\approx N - r$ in experiments)
$\mathbf{V}$	Factor loading matrix (orthonormal columns)
σ_y^2	Observation noise variance
τ_t^2	Off-subspace forecast error variance at step t
s_t	Off-subspace score (norm)
$\boldsymbol{\delta}_t$	Forecast error $\hat{\mathbf{y}}_t - \mathbf{y}_t$

C Algorithm and Computational Cost

Algorithm 1 gives the scoring procedure, and Figure 5 provides a pipeline overview.

Algorithm 1 SubDx: Post-Hoc Diagnostic Scoring

Require: Input $\mathbf{X} \in \mathbb{R}^{L \times N}$, forecast $\hat{\mathbf{Y}} \in \mathbb{R}^{P \times N}$, rank r, channel statistics μ, σ

$\tilde{\mathbf{X}}_{\ell n} \leftarrow (X_{\ell n} - \mu_n)/(\sigma_n + \epsilon)$; $\tilde{\hat{\mathbf{Y}}}_{tn} \leftarrow (\hat{Y}_{tn} - \mu_n)/(\sigma_n + \epsilon)$

$\mathbf{A}, \mathbf{\Sigma}, \mathbf{V}^\top \leftarrow \mathrm{SVD}_r(\tilde{\mathbf{X}})$ {Keep $\mathbf{U} = \mathbf{V}_{:,1:r}$}

$\mathbf{R} \leftarrow \tilde{\hat{\mathbf{Y}}} - (\tilde{\hat{\mathbf{Y}}}\mathbf{U})\mathbf{U}^\top$

for $t = 1, \ldots, P$ **do**

  $s_t \leftarrow \|\mathbf{R}_{t,:}\|_2$

end for

return $\{s_t\}_{t=1}^P$

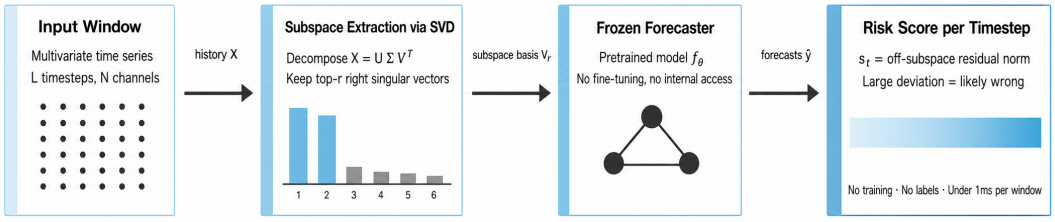

Figure 5: Pipeline overview for SubDx: estimate a local channel subspace from the input history, then score the forecast's off-subspace component. Algorithm 1 specifies normalization and the implicit residual computation without constructing an $N \times N$ projector.

The SVD costs $O(LNr)$ per window; projections cost $O(PNr)$. For PatchTST on Traffic ($L = 96$, $N = 862$, $r = 8$), forward pass takes ∼50ms, SVD adds ∼0.3ms (<1%). On Weather ($N = 21$), SVD is ∼0.05ms. MC Dropout requires 10× forward passes; ensembles require $K\times$ inference cost. On Electricity/PatchTST, scorer-only CPU overhead is 28.11/37.91 ms per window, excluding base-model inference, diagnostics, and CSV replay; this CPU measurement is not directly comparable to the GPU forward-pass numbers above.

D Extended Experiment Setup

Full dataset details, training hyperparameters, and baseline implementations are as follows. All datasets use standard train/val/test splits from their respective repositories. Backbones follow published configurations: PatchTST (patch_len=16, stride=8, d_model=128, 3 layers), iTransformer (d_model=512, 3 layers), DLinear (moving_avg=25). Training: Adam optimizer, MSE loss, 20 epochs, batch_size=32, learning rate 10^{-3} with step decay at epochs 10 and 15. Deep ensembles use 3 independent seeds; MC Dropout uses 10 forward passes with dropout rate 0.1. Unless stated otherwise, the single-cell ablation, rank-sensitivity, and robustness tables in this appendix use the original per-backbone protocol; protocol-locked comparisons are reported in Table 1 and Table 31.

D.1 Epsilon and Constant-Channel Sensitivity

SWaT contains 11 train-constant channels. The pipeline-exact contract used in Table 1 sets their scale to one; the two epsilon-bearing contracts below are diagnostics, and $\epsilon = 0$ is undefined when exact zero variance is present. Table 10 reports the positive-epsilon grid on SWaT/DLinear (8,192 windows; 786,432 timestep rows) under these contracts. The absolute AUROC changes by at most 0.012 across the grid, and the score ranking relative to $\epsilon = 10^{-12}$ stays high (rank correlation ≥ 0.816 for the literal contract and ≥ 0.988 for the pipeline-aware contract). We report these as sensitivity checks and do not select ϵ by test AUROC.

Table 10: Epsilon sensitivity on SWaT/DLinear (8,192 windows; 786,432 timestep rows; 11 train-constant channels). Ranges cover $\epsilon \in \{10^{-12}, 10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}\}$; $\epsilon = 0$ is undefined under exact zero variance.

Normalization contract	AUROC range	FN@50 range (%)	Min rank corr. vs. ϵ=10^{-12}
Current-window literal	0.689–0.701	26.5–27.7	0.816
Train-stats pipeline-aware	0.650–0.653	31.1–31.3	0.988

D.2 Robust Normalization and Contamination

Robust alternatives do not provide a universal default. On ETTh1 (1,024 windows), median/MAD normalization raises DLinear AUROC from 0.665 to 0.695 but lowers iTransformer from 0.709 to 0.505; Huber shows the same pattern (0.665 to 0.691 and 0.716 to 0.495). MCD on 128 windows raises DLinear from 0.185 to 0.888 but carries subset and numerical-warning limitations. Across spike, level-shift, stuck-suffix, and dropout contamination, score rank correlation with the clean ranking ranges 0.257–0.792 and lowest-score top-20 overlap ranges 0.425–0.728. These are scoped ETTh1 diagnostics; the appropriate response is an input-quality flag or abstention, not a universal robustness claim.

Table 11: Clean ETTh1 AUROC under robust normalization (1,024 windows; 40 healthy channels). MCD is reported separately in the text: 0.185 mean/std versus 0.888 MCD on 128 windows for DLinear, with subset and numerical-warning limitations.

Normalization	DLinear	iTransformer
Mean/std SVD	0.665	0.709
Median/MAD (std fallback)	0.695	0.505
Mean/std Huber	0.665	0.716
Median/MAD + Huber	0.691	0.495

E Proofs

Proposition 1 (Score distribution)**.** *Assume observations follow* $\mathbf{y}_t = \mathbf{V}\mathbf{z}_t + \boldsymbol{\varepsilon}_t$ *with* $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma_y^2 \mathbf{I}_N)$*, where* $\mathbf{V} \in \mathbb{R}^{N \times d}$ *has orthonormal columns and* $k = N - d$*. Let the history subspace be estimated perfectly and let the off-subspace forecast error satisfy* $(\mathbf{I} - \mathbf{V}\mathbf{V}^\top)\boldsymbol{\delta}_t \sim \mathcal{N}(\mathbf{0}, \tau_t^2 \mathbf{I}_k)$*. For fixed* $\boldsymbol{\delta}_t$*,* $q_t = s_t^2$ *satisfies*

$$\frac{q_t}{\sigma_y^2} \mid \boldsymbol{\delta}_t \sim \chi_k^2(\lambda_t/\sigma_y^2), \qquad \lambda_t = \|(\mathbf{I} - \mathbf{V}\mathbf{V}^\top)\boldsymbol{\delta}_t\|^2.$$

Marginally, $q_t \sim (\tau_t^2 + \sigma_y^2)\chi_k^2$.

Theorem 1 (Two-population AUROC)**.** *Under the assumptions of Proposition 1, consider two timestep populations with off-subspace error variances* τ_0^2 *and* τ_1^2 *(*$\tau_1^2 > \tau_0^2$*). If* q_i *is independently drawn from population* i*, then*

$$\textit{AUROC} = 1 - F_{F_{k,k}}\left(\frac{\tau_0^2 + \sigma_y^2}{\tau_1^2 + \sigma_y^2}\right),$$

where $F_{F_{k,k}}$ *is the CDF of the* F*-distribution with* (k, k) *degrees of freedom. As* $k \to \infty$*, AUROC* $\to 1$ *whenever* $\tau_1^2 > \tau_0^2$.

Proof of Proposition 1. Under the factor model, $\hat{\mathbf{y}}_t = \mathbf{y}_t + \boldsymbol{\delta}_t = \mathbf{V}\mathbf{z}_t + \boldsymbol{\varepsilon}_t + \boldsymbol{\delta}_t$. The off-subspace projector $\mathbf{P}_\perp = \mathbf{I} - \mathbf{V}\mathbf{V}^\top$ satisfies $\mathbf{P}_\perp \mathbf{V} = \mathbf{0}$, so:

$$\mathbf{P}_\perp \hat{\mathbf{y}}_t = \mathbf{P}_\perp(\mathbf{V}\mathbf{z}_t + \boldsymbol{\varepsilon}_t + \boldsymbol{\delta}_t) = \mathbf{P}_\perp \boldsymbol{\varepsilon}_t + \mathbf{P}_\perp \boldsymbol{\delta}_t.$$

Since $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma_y^2 \mathbf{I}_N)$ and $\mathbf{P}_\perp$ is an orthogonal projector of rank $k = N - d$, we have $\mathbf{P}_\perp \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma_y^2 \mathbf{P}_\perp)$. In an orthonormal basis for the off-subspace, this is $\mathcal{N}(\mathbf{0}, \sigma_y^2 \mathbf{I}_k)$.

For fixed $\boldsymbol{\delta}_t$, the off-subspace component $\mathbf{P}_\perp \boldsymbol{\delta}_t$ is a fixed vector $\boldsymbol{\mu}_t \in \mathbb{R}^k$ with $\|\boldsymbol{\mu}_t\|^2 = \lambda_t$. Therefore:

$$\mathbf{P}_\perp \hat{\mathbf{y}}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \sigma_y^2 \mathbf{I}_k) \qquad \text{(conditional on } \boldsymbol{\delta}_t\text{)}.$$

The squared norm of a k-dimensional Gaussian with mean $\boldsymbol{\mu}_t$ and covariance $\sigma_y^2 \mathbf{I}_k$ is $\sigma_y^2 \chi_k^2(\lambda_t/\sigma_y^2)$ by definition of the noncentral chi-squared distribution. The moments follow from the standard moments of the noncentral χ^2: $\mathbb{E}[\chi_k^2(\lambda)] = k + \lambda$ and $\mathrm{Var}[\chi_k^2(\lambda)] = 2(k + 2\lambda)$.

For the marginal result: since $\mathbf{P}_\perp \boldsymbol{\varepsilon}_t$ and $\mathbf{P}_\perp \boldsymbol{\delta}_t$ are independent (by assumption, Section 3.5), their sum is $\mathbf{P}_\perp \hat{\mathbf{y}}_t \sim \mathcal{N}(\mathbf{0}, (\tau_t^2 + \sigma_y^2)\mathbf{I}_k)$, so $q_t \sim (\tau_t^2 + \sigma_y^2)\chi_k^2$. □

Proof of Theorem 1. Under the factor model with independent draws from two timestep populations $i \in \{0, 1\}$, Proposition 1 gives $q_i \sim (\tau_i^2 + \sigma_y^2)\chi_k^2$ for independent score draws from each population. The AUROC for distinguishing population 1 (higher off-subspace error variance) from population 0 is:

$$\mathrm{AUROC} = \Pr(q_1 > q_0) = \Pr\left(\frac{q_0}{q_1} < 1\right).$$

Define $W_i = q_i/[k(\tau_i^2 + \sigma_y^2)] \sim \chi_k^2/k$, so:

$$\frac{q_0}{q_1} = \frac{(\tau_0^2 + \sigma_y^2)}{(\tau_1^2 + \sigma_y^2)} \cdot \frac{W_0}{W_1}.$$

Since W_0 and W_1 are independent χ_k^2/k random variables, $W_0/W_1 \sim F_{k,k}$ by definition. Therefore:

$$\mathrm{AUROC} = \Pr\left(\frac{W_0}{W_1} < \frac{\tau_1^2 + \sigma_y^2}{\tau_0^2 + \sigma_y^2}\right) = F_{F_{k,k}}\left(\frac{\tau_1^2 + \sigma_y^2}{\tau_0^2 + \sigma_y^2}\right) = 1 - F_{F_{k,k}}\left(\frac{\tau_0^2 + \sigma_y^2}{\tau_1^2 + \sigma_y^2}\right),$$

where the last equality uses the symmetry property $F_{F_{k,k}}(x) = 1 - F_{F_{k,k}}(1/x)$. For $k \to \infty$, the $F_{k,k}$ distribution concentrates at 1, so $\mathrm{AUROC} \to 1$ whenever $\tau_1^2 > \tau_0^2$. □

F Multi-Seed Robustness

Table 12 shows AUROC across 3 seeds for all dataset-backbone pairs where multi-seed was run. Standard deviations are extremely small, especially on high-N data.

G Rank Sensitivity

Figure 6 shows AUROC as a function of the SVD rank r for PatchTST on Weather and Traffic under the original per-backbone protocol. On Traffic ($N = 862$), the score changes little across $r \in \{4, \ldots, 32\}$ (AUROC 0.851 to 0.872). On Weather ($N = 21$), performance peaks at $r = 8$ and is more sensitive to the rank choice.

Table 13 compares label-free rank rules on the SWaT/DLinear cell. All rules use only current-window singular values; none uses forecast outcomes. The precommitted requested-legal rank remains the primary result, while EV90/EV95 and fixed ranks are sensitivity analyses rather than post-hoc selection.

Table 12: Multi-seed AUROC under the original per-backbone protocol (pred_len=96). Results vary little across seeds; absolute values may differ from the protocol-locked Table 1.

Dataset	Backbone	N	Seed 42	Seed 43	Seed 44	Mean	Std
Weather	PatchTST	21	0.626	0.612	0.623	0.620	0.007
	iTransformer	21	0.611	0.608	0.604	0.608	0.004
Electricity	PatchTST	321	0.690	0.691	0.686	0.689	0.003
	iTransformer	321	0.616	0.615	0.614	0.615	0.001
Traffic	PatchTST	862	0.855	0.855	0.854	0.855	0.001
	iTransformer	862	0.810	0.808	0.813	0.810	0.003

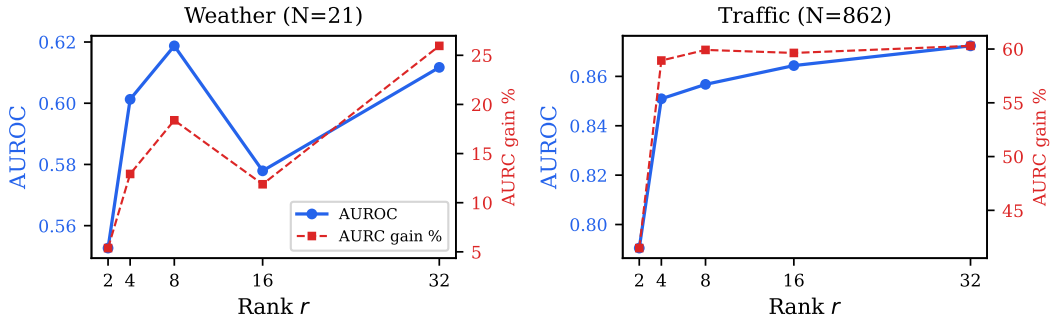

Figure 6: Rank sensitivity for PatchTST. Left: Weather ($N = 21$). Right: Traffic ($N = 862$). The score changes little on high-N data.

Table 13: Label-free rank rules on SWaT/DLinear (seed 42). The primary row uses the precommitted requested-legal rank; the remaining rows are sensitivity analyses.

Rule	AUROC
Requested-legal (primary, $r = 8$)	0.604
Fixed rank 1	0.683
Fixed rank 2	0.706
Fixed rank 4	0.670
Variance 90% (EV90)	0.754
Variance 95% (EV95)	0.753
Effective rank	0.703

H Prediction Horizon Sweep

Table 14 and Figure 7 show results across prediction horizons $P \in \{24, 96, 192\}$. On Traffic, SubDx maintains AUROC > 0.85 for PatchTST across all horizons. On Weather, performance degrades slightly at $P = 192$.

I Deep Ensembles Comparison

Table 15 shows the full deep ensembles comparison across all evaluated configurations. On PatchTST, SubDx beats deep ensembles across all three datasets despite requiring zero additional training.

J Threshold Robustness

Table 16 reports AUROC, PR-AUC, Spearman ρ, and Kendall τ across three error percentile thresholds for all configurations. Rankings are consistent across thresholds: high-N datasets (Traffic, Electricity) show signal regardless of whether the threshold is set at the 25th, 50th, or 75th error percentile. Figure 8 visualizes threshold stability; PR-AUC follows AUROC's high-N/low-N pattern.

Table 14: AUROC across prediction horizons (seed=42).

Backbone	Dataset	$P = 24$	$P = 96$	$P = 192$
PatchTST	Weather	0.602	0.626	0.580
	Traffic	0.851	0.855	0.858
iTransformer	Weather	0.592	0.611	0.586
	Traffic	0.823	0.810	0.785

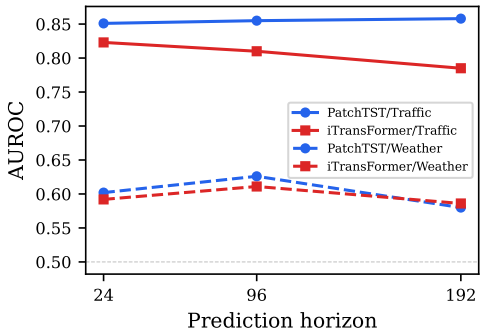

Figure 7: AUROC across prediction horizons. Traffic performance is stable; Weather shows slight degradation at longer horizons.

Table 15: Deep ensembles (3 seeds) vs. SUBDX. Our zero-cost score beats deep ensembles on PatchTST across all datasets.

Backbone	Dataset	SUBDX		Deep Ensembles (3×)	
		AUROC	AURC gain (%)	AUROC	AURC gain (%)
PatchTST	Weather	**0.625**	**21.3**	0.608	18.6
	Electricity	**0.691**	**17.8**	0.671	15.8
	Traffic	**0.856**	**59.7**	0.779	44.0
iTransformer	Weather	0.607	19.7	**0.630**	**21.6**
	Electricity	0.603	10.2	**0.702**	**21.3**

Table 16: Threshold checks: AUROC and PR-AUC at three error percentile thresholds, plus threshold-free rank correlations. Rankings are consistent across thresholds.

Backbone	Dataset	ρ	τ	AUROC			PR-AUC		
				@25	@50	@75	@25	@50	@75
PatchTST	Traffic	0.770	0.576	0.898	0.893	0.862	0.964	0.901	0.711
	Electricity	0.378	0.261	0.682	0.691	0.690	0.865	0.721	0.504
	Weather	0.329	0.224	0.701	0.652	0.633	0.872	0.658	0.367
	ETTh1	0.142	0.095	0.575	0.569	0.567	0.806	0.575	0.313
iTransformer	Traffic	0.702	0.505	0.874	0.862	0.828	0.954	0.862	0.598
	Weather	0.256	0.172	0.657	0.618	0.603	0.853	0.621	0.340

K Qualitative Case Studies

Figure 9 summarizes a Traffic window where the per-timestep score tracks realized MSE and separates high-risk timesteps. Figure 10 shows a representative success case from PatchTST/Traffic: the off-subspace score s_t tracks per-timestep MSE e_t closely (within-window $\rho = 0.94$). Timesteps where the forecast diverges from ground truth correspond to high s_t. Figure 11 shows the complementary failure mode on ETTh1 ($N = 7$): the score is nearly flat across the horizon, providing no discriminative information, consistent with insufficient channel redundancy.

L Implementation Details

All experiments use the following default settings unless otherwise noted:

- SVD rank: $r = 8$
- Prediction horizon: $P = 96$ (main experiments)
- Input length: $L = 512$ (PatchTST, iTransformer), $L = 336$ (DLinear)
- Training: default settings from each backbone's original implementation

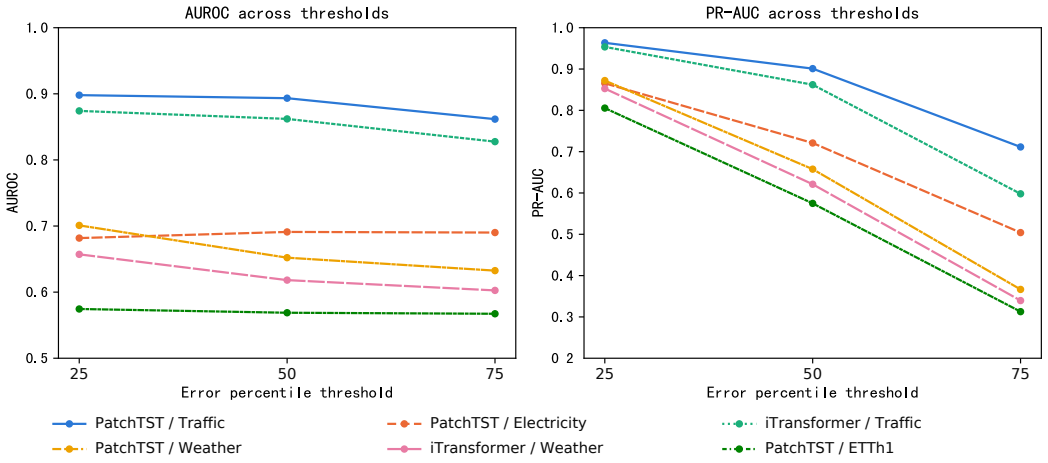

Figure 8: AUROC (left) and PR-AUC (right) at three error percentile thresholds. Rankings among configurations are stable across thresholds.

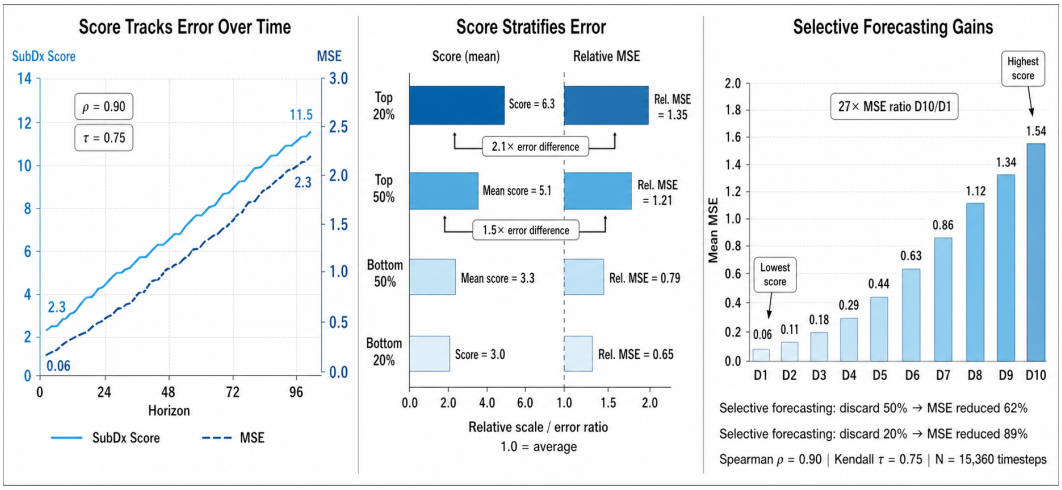

Figure 9: Qualitative Traffic example ($N = 50$, PatchTST). Left: the per-timestep SUBDX score follows the realized per-timestep MSE within the forecast horizon (Spearman $\rho = 0.90$). Right: grouping timesteps by score shows strong risk separation; the top 20% by score has roughly $8\times$ higher MSE than the bottom 20%.

- Seeds: 42, 43, 44 (multi-seed experiments); 42 (single-seed)
- Hardware: NVIDIA RTX 4090D (24GB)

The off-subspace score is computed using `torch.linalg.svd` on the input matrix. For $r = 8$ and $N = 862$ (Traffic), the SVD takes $<$1ms per sample. This cost is negligible relative to backbone inference.

M Within-Horizon Analysis

Later forecast steps are generally harder, so the score might merely rank horizon positions. To test this, we compute AUROC at each step t separately, then average. On Traffic, within-horizon AUROC (0.900) exceeds pooled (0.893), showing that the score ranks errors among same-step predictions. On Weather, within-horizon AUROC drops from 0.652 to 0.586, indicating partial horizon confounding, consistent with the horizon index being competitive there.

The pooled and within-horizon results are reported in Table 3.

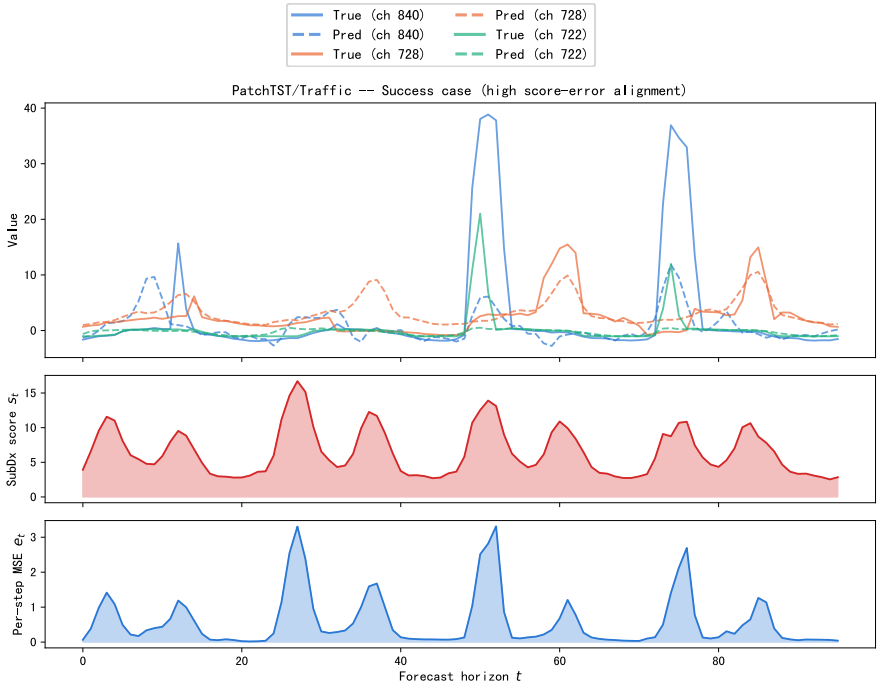

Figure 10: Success case: PatchTST/Traffic. Top: three representative channels (true vs. predicted). Middle: SubDx score s_t. Bottom: per-timestep MSE e_t. High s_t aligns with large forecast errors.

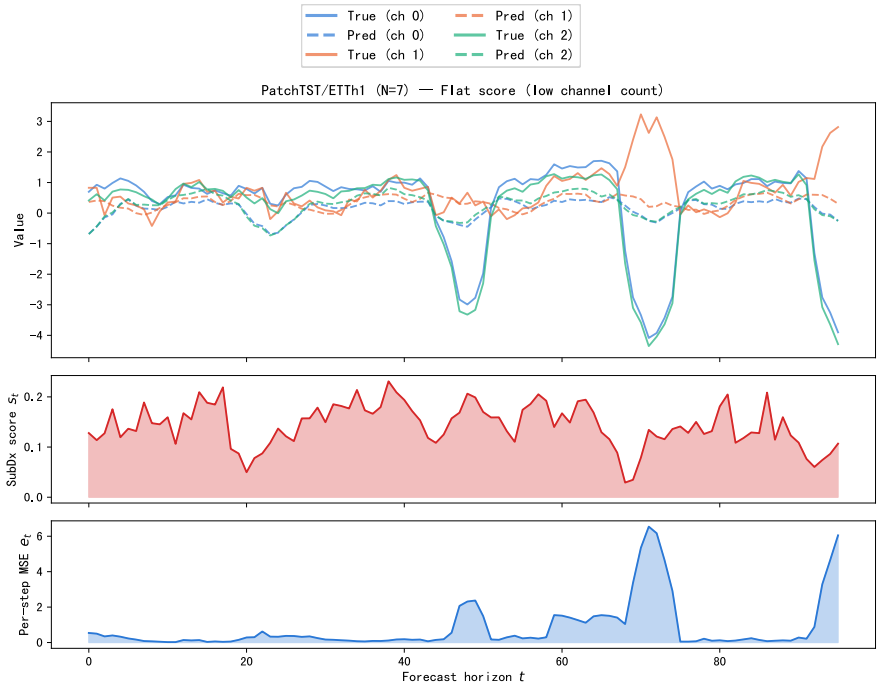

Figure 11: Low-N failure: PatchTST/ETTh1 ($N = 7$). The score is nearly flat despite varying error magnitudes, demonstrating the method's reliance on channel redundancy.

M.1 The Correction Failure

A direct follow-up to error localization is error correction. We tested six parameter-free correction mechanisms: (1) subspace projection ($\hat{\mathbf{y}}_t \leftarrow \mathbf{U}\mathbf{U}^\top \hat{\mathbf{y}}_t$), (2) Ridge-regularized projection, (3) mean

shrinkage toward the training mean, (4) channel-wise mean imputation, (5) norm shrinkage, and (6) last-value repetition. Each was applied to the highest-scored 20% of timesteps.

Table 18 shows that almost all corrections either hurt or negligibly help. The best correction result reduces MSE by 0.5% (Ridge on iTransformer/Electricity; ΔMSE = -0.5%), and even oracle correction, which selects the truly worst timesteps rather than the diagnosed ones, gives at most 1.1% MSE reduction (iTransformer/Electricity, oracle ΔMSE = -1.1%). Most corrections degrade forecasts, especially on Traffic, where all methods increase MSE.

Why do corrections fail when localization works? A scalar score $s_t \in \mathbb{R}$ indicates *that* a timestep is risky but lacks two things needed for correction: (i) a *direction* in $\mathbb{R}^N$, since the full off-subspace residual vector $(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\hat{\mathbf{y}}_t$ spans k dimensions, but collapsing it to the norm $s_t = \|\cdot\|_2$ discards which channels deviate most; and (ii) a *target*, since projection back to $\mathbf{U}\mathbf{U}^\top\hat{\mathbf{y}}_t$ assumes the ground truth lies on the history subspace, which is too strong. Even oracle selection fails under these parameter-free corrections, so the limiting factor is the correction mechanism, not selection accuracy. Future work could condition a learned corrector on the full residual.

Model error vs. timestep difficulty. As noted in Section 3.3, the off-subspace score may partly reflect intrinsic timestep difficulty rather than model-specific failure. To disentangle these effects, we compare three PatchTST variants on the *same* Traffic test set (Table 17): a well-trained model (20 epochs), an undertrained model (5 epochs), and a channel-shuffled variant (well-trained predictions with channel order randomly permuted at test time, destroying cross-channel structure).

Table 17: Model quality control: SubDx score behaviour across PatchTST variants on the same Traffic test set. Partial ρ controls for the true future's off-subspace distance. The shuffled-channel model's partial correlation suggests that the score captures model-specific failure, not just timestep difficulty.

Variant	MSE	AUROC@50	ρ	Partial ρ
Well-trained (20 ep)	0.509	0.893	0.770	0.144
Undertrained (5 ep)	0.515	0.851	0.682	-0.027
Shuffled channels	1.184	0.964	0.894	0.681

The shuffled-channel model has 2.3$\times$ higher MSE, and the score detects this: AUROC rises to 0.964 and partial ρ reaches 0.681, showing that the score captures model-specific structure violations. For well-trained vs. undertrained (MSE 0.509 vs. 0.515), partial ρ is small, indicating that shared signal comes mostly from timestep difficulty when models are similar.

The score is therefore a *risk signal*: it ranks timesteps by predicted error, drawing on both model-specific violations and intrinsic difficulty. For selective forecasting, this distinction matters less than for causal attribution because the goal is risk ranking, not source isolation.

Table 18: Parameter-free correction results: even the best mechanism gives negligible improvement. Oracle correction (selecting truly worst timesteps) also fails for this family of corrections.

Backbone	Dataset	Best Correction	Oracle ΔMSE
PatchTST	Weather	$+1.0\%$	$+0.2\%$
PatchTST	Electricity	-0.1%	-0.5%
PatchTST	Traffic	$+1.7\%$	$+1.3\%$
iTransformer	Weather	$+1.0\%$	$+0.2\%$
iTransformer	Electricity	-0.5%	-1.1%
iTransformer	Traffic	$+0.1\%$	-0.0%

M.2 Distribution Shift Monitoring

One use case for SubDx is model monitoring without ground truth. We evaluate this by applying controlled distribution shifts to the test set of a PatchTST model trained on Traffic, and measuring whether SubDx scores increase with the severity of degradation.

Table 6 in the main text reports the results. Under additive Gaussian noise ($\sigma = 0.5\times$ channel std), SubDx scores increase by 44% while MSE increases by only 19%. In this regime, the diagnostic is *more sensitive* than the error metric itself. At $\sigma = 1.0$, the score doubles (+110%) while MSE rises 61%. With delayed labels, this lets the score move before errors can be measured.

Under structural input shifts (channel permutation, time reversal), MSE increases substantially (+139% and +285%) but SubDx scores barely change (+9% and +4%). This is consistent with the mechanism: these shifts alter the input's cross-channel structure, so the SVD captures a different subspace, and the model's errors respect that shifted subspace rather than violating it. Thus, SubDx responds to errors that violate cross-channel consistency, not to arbitrary increases in error.

Channel dropout (simulating sensor failures) reveals a limitation: dropping 30% of channels reduces k, mechanically lowering SubDx scores (-18%) despite a 57% MSE increase. SubDx monitoring requires the channel structure to remain intact, a condition that should be verified in deployment.

M.3 Ablations and Stability Checks

Score components. Table 19 decomposes the score into its components. The off-subspace norm (Eq. 5) carries essentially all the signal; variance-ratio weighting adds <0.002 AUROC. On Traffic, variance ratio alone is near-random (AUROC = 0.505). Other experiments retain the unweighted norm.

In-subspace vs. off-subspace signal. The SPE/Q-statistic measures off-subspace deviation. A natural complementary signal is Hotelling's T^2 [Hotelling, 1931], which measures whether the forecast lies in an unusual region *within* the history subspace: $T_t^2 = \|\mathbf{\Sigma}_r^{-1}\mathbf{U}^\top \hat{\mathbf{y}}_t\|_2^2$, where $\mathbf{\Sigma}_r$ contains the retained singular values. Under the factor model (Section 3.5), the off-subspace term satisfies $Q_t \sim \chi_k^2$; Gaussian in-subspace coordinates additionally give $T_t^2 \sim \chi_r^2$.

Table 20 reports the comparison. On Traffic ($N = 862$), T^2 alone achieves AUROC = 0.814, well above random, indicating that in-subspace anomaly carries diagnostic information for the worst errors (80th percentile threshold). However, it remains weaker than SPE (0.856), and combining T^2 with SPE yields AUROC = 0.815, near-identical to T^2 alone. The two signals are not complementary: SPE covers the diagnostic information that T^2 provides. On Weather ($N = 21$), T^2 is uninformative (AUROC = 0.501), consistent with the small r ($r = 8$ in $N = 21$ dimensions) leaving limited in-subspace structure to exploit.

Table 19: Score component ablation (PatchTST, pred_len=96). The off-subspace norm is the main source of signal; variance-ratio weighting contributes negligibly.

Component	Weather ($N = 21$)		Traffic ($N = 862$)	
	AUROC	AURC gain (%)	AUROC	AURC gain (%)
Norm only	0.619	18.4	0.855	59.7
Variance ratio only	0.432	−28.1	0.505	0.7
Combined	0.619	18.4	0.857	59.9

Table 20: In-subspace (T^2) vs. off-subspace (SPE) diagnostic signals (PatchTST, pred_len=96). SPE is stronger; combining signals does not improve over SPE alone.

Signal	Weather ($N = 21$)		Traffic ($N = 862$)	
	AUROC	AURC gain (%)	AUROC	AURC gain (%)
SPE (SubDx)	0.625	21.5	0.856	59.2
T^2 (in-subspace)	0.501	−7.2	0.814	42.5
T^2 + SPE	0.517	−2.6	0.815	42.9
SPE + RPCA	0.623	21.3	0.856	59.2

Robust PCA subspace. SVD can be sensitive to outlier channels or timesteps that inflate certain singular directions. We test Robust PCA (RPCA; [Candès et al., 2011]), which decomposes the history into a low-rank component $\mathbf{L}$ and a sparse outlier component $\mathbf{S}$, and extracts the subspace from $\mathbf{L}$. Table 20 shows RPCA + SPE is within 0.001 AUROC of standard SVD + SPE on both datasets. On these clean benchmarks without gross channel-wise outliers, standard SVD is sufficient.

Rank sensitivity. On Traffic, the score changes little across $r \in \{4, 8, 16, 32\}$ (AUROC 0.851 to 0.872). On Weather, $r = 8$ is optimal. Appendix G reports the full sweep; $r = 8$ remains the default.

Prediction horizon. Results are stable across horizons $P \in \{24, 96, 192\}$. PatchTST/Traffic maintains AUROC > 0.85 at all horizons. Full results in Appendix H.

Multi-seed stability. Across 3 seeds, AUROC standard deviation is < 0.003 on Electricity and < 0.001 on Traffic. Bootstrap confidence intervals indicate stable estimates (Appendix S). Full multi-seed results are in Appendix F.

N Cross-Backbone Consistency

Table 21 compares the scores and errors of the two backbones.

Table 21: Cross-backbone SubDx score correlation (PatchTST vs. iTransformer, same test set, seed=42).

Dataset	Score ρ	Partial ρ	Error ρ
Traffic ($N=862$)	0.900	0.833	0.970
Solar ($N=137$)	0.864	0.654	0.878

O Random-Reject Baseline

Table 22 compares random and score-guided rejection at matched coverage.

Table 22: Random-reject vs. SubDx-reject at 50% coverage (PatchTST, pred_len=96, seed=42).

Dataset	Random MSE	SubDx MSE	Gain (%)
Traffic ($N=862$)	0.507	0.195	61.5
Electricity ($N=321$)	0.168	0.141	16.4
Weather ($N=21$)	0.167	0.141	15.4

P Full Signal Comparison

Table 23 reports the complete signal comparison for both backbones.

Table 23: Signal comparison under the original per-backbone protocol (pred_len=96, seed=42). Bold: best non-supervised per config. Absolute values may differ from the protocol-locked Table 1.

Signal	Weather ($N=21$)		Traffic ($N=862$)	
	PatchTST	iTransformer	PatchTST	iTransformer
			AUROC	
SubDx (ours)	0.610	0.598	**0.855**	**0.806**
Pairwise deviation	0.611	0.602	0.825	0.737
Deep ensembles (3×)	0.608	N/A	0.779	N/A
MC Dropout (10×)	0.513	0.595	0.738	0.728
Forecast magnitude	0.474	0.478	0.716	0.723
Forecast change	0.551	0.557	0.711	0.670
Recent volatility	0.573	0.573	0.484	0.480
Horizon index	**0.619**	**0.621**	0.502	0.513
Supervised reference	*0.655*	*0.653*	*0.859*	*0.815*
			AURC gain %	
SubDx (ours)	15.7	16.8	**59.2**	**52.2**
MC Dropout (10×)	1.4	12.6	28.6	26.1
Horizon index	**24.7**	**25.4**	4.2	6.8

Q Synthetic Validation

Figure 12 summarizes the primary synthetic sweeps over off-subspace dimension, visibility spread, observation noise, and their interaction. These controls match the theory: larger k sharpens discrimination, visibility variation enables ranking, and observation noise dilutes the signal.

Table 24 summarizes four additional stress-test dimensions.

Table 24: Synthetic stress-test summary. The signal changes little under estimation perturbations once enough off-subspace dimensions are present, but fails when both k and visibility spread are small.

Stress axis	Sweep	Main observation
Finite history	$L \in [10, 1000]$, $k = 100$	AUROC stays near 0.98 even with $L = 10$.
Rank mismatch	$r \in [2, 150]$, true $d = 100$	Stable until $r > d$ begins absorbing residual directions.
Noise law	Gaussian/Laplace/Uniform	Matched AUROC at medium spread: 0.983/0.980/0.983.
Factorial regime	$k \in \{20, 60, 120\}$, spread/noise grid	AUROC ranges from 0.57 to 1.00; neither large k nor visibility alone is sufficient.

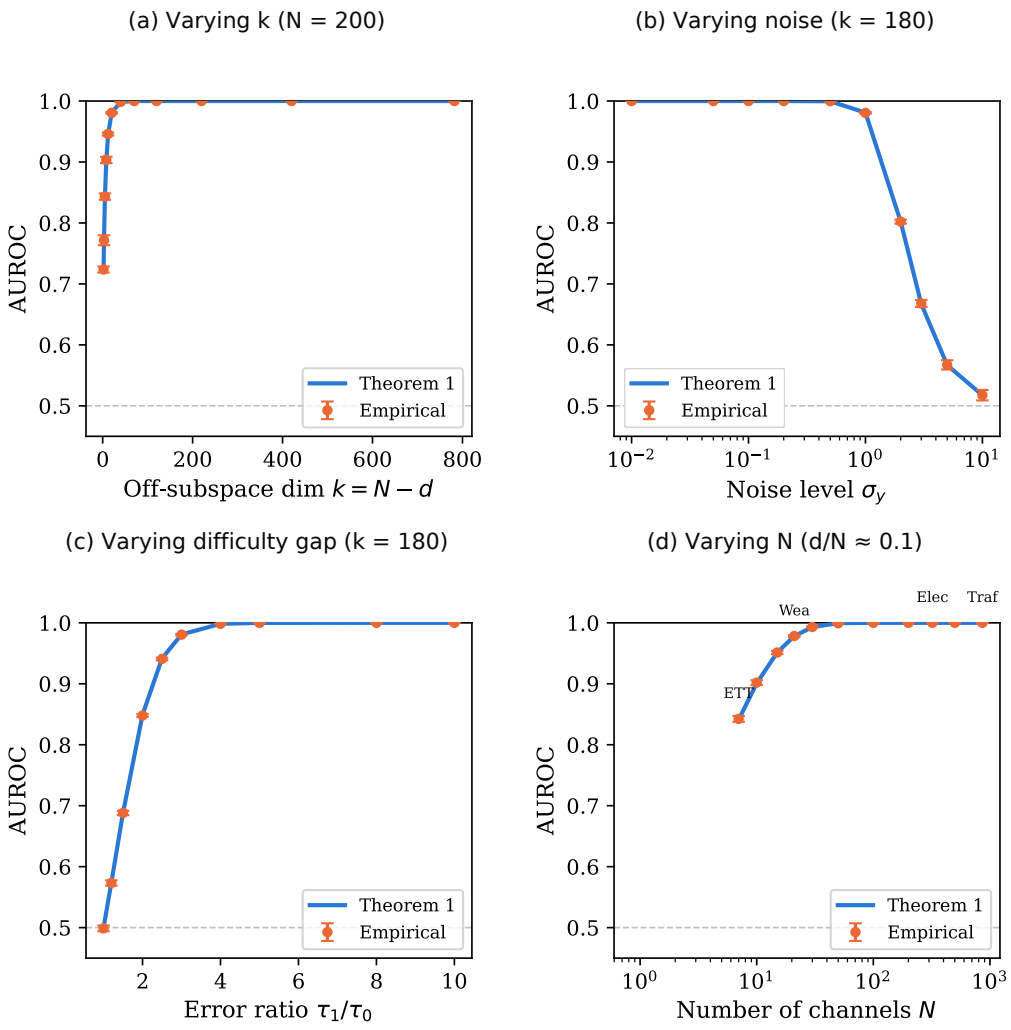

Figure 12: Synthetic validation under a controlled factor model. AUROC rises with off-subspace dimension and visibility spread, falls with observation noise, and is highest when large k and heterogeneous visibility occur together.

D5 (finite history L). Fix $k = 100$, sweep $L \in [10, 1000]$. AUROC is near-constant ($\sim$0.98) even at $L = 10$: when k is large, even a crudely estimated subspace provides sufficient off-subspace projection. This matters for short-window deployments.

D6 (rank mismatch). Fix true $d = 100$ ($k = 100$), sweep algorithmic rank $r \in [2, 150]$. AUROC is stable ($\sim$0.98) across $r \in [2, 128]$, dipping only at $r = 150 > d$. The method does not require r to match d; a small $r \ll N$ suffices, explaining the default $r = 8$ across real datasets.

D7 (noise distribution). Compare Gaussian, Laplace, and Uniform observation noise. At medium spread, AUROC is 0.983/0.980/0.983, respectively, in qualitative agreement with the theory.

D8 (full factorial). A $3 \times 3 \times 2$ sweep ($k \in \{20, 60, 120\}$, spread $\in$ {narrow, medium, wide}, $\sigma_y \in \{0.3, 0.7\}$) shows the interaction: at $k = 20$ with narrow spread and high noise, AUROC drops to 0.57; at $k = 120$ with wide spread and low noise, AUROC reaches 1.00. The two factors interact multiplicatively: neither alone is sufficient. The overall empirical regularity is: AUROC $\approx g(k) \cdot h(\text{std}(\nu)) \cdot q(\sigma_y)$, where $g(k)$ saturates near 1 for $k \gtrsim 40$, $h(\cdot)$ is near-linear in visibility variance, and $q(\cdot)$ captures noise dilution.

R Per-Window vs. Fixed Global Subspace

We also ask whether the per-window recomputed subspace, a defining feature of SubDx, is actually necessary, or whether a fixed subspace estimated once from the training set would suffice. The latter corresponds to classical PCA-based process monitoring (SPE/Q-statistic), where the subspace is learned offline and applied unchanged at test time.

Table 25 compares the two approaches. On large-N datasets (Traffic, Solar, Electricity), the fixed global PCA subspace is slightly more accurate (ΔAUROC $= -0.026$ to -0.087), which is expected: the training-set PCA benefits from seeing more data. On Weather ($N = 21$), however, the per-window SVD is higher (ΔAUROC $= +0.087$), because the smaller channel set makes local adaptation more valuable. Across all datasets, the per-window SVD achieves competitive performance without any access to training data, a prerequisite for the post-hoc, frozen-model setting that SubDx targets.

S Bootstrap Confidence Intervals

To assess the stability of the main results, we compute bootstrap 95% confidence intervals (1,000 resamples) for AUROC, AURC gain, and Sel@0.5 across the primary PatchTST configurations. Table 26 reports the results. The intervals are narrow, indicating that the reported metrics are not artifacts of a particular test split.

Table 25: Per-window SVD (SubDx) vs. fixed global PCA (PatchTST, pred_len=96, seed=42, single training run). Absolute AUROC values may differ from the protocol-locked Table 1 because these are single-cell original-protocol runs; the SVD minus PCA comparison is the relevant signal.

Dataset	SVD (SubDx)	Fixed PCA	ΔAUROC
Traffic ($N=862$)	0.856	0.881	-0.026
Solar ($N=137$)	0.898	0.907	-0.008
Electricity ($N=321$)	0.691	0.778	-0.087
Weather ($N=21$)	0.612	0.525	**+0.087**

Table 26: Bootstrap 95% confidence intervals (1,000 resamples, PatchTST, pred_len=96, single training run). Absolute values may differ from the protocol-locked Table 1 because these are single-cell original-protocol runs; interval widths are the relevant signal.

Dataset	AUROC [95% CI]	AURC gain (%)	Sel@50 (%)
Traffic	0.855 [0.853, 0.856]	59.3	61.5
Solar	0.901 [0.901, 0.902]	75.5	77.1
Electricity	0.685 [0.683, 0.687]	17.3	15.1
Weather	0.628 [0.627, 0.629]	21.1	16.6
ETTh1	0.569 [0.567, 0.571]	5.3	5.3

T Score Decile Monotonicity

A direct way to visualize ranking quality is to sort all test timesteps by their SubDx score and plot the mean MSE in each score decile. Table 27 reports these values. On Traffic and Solar, mean MSE in the highest-score decile exceeds the lowest-score decile by factors of 27$\times$ and 155$\times$, respectively. On Electricity and Weather, the separation is more modest (2 to 3$\times$) but remains monotonic. The ranking signal therefore extends across the score distribution, beyond its extreme quantiles.

Table 27: Mean per-timestep MSE by SubDx score decile (PatchTST, pred_len=96, single training run). Decile 1 = lowest scores (most confident), Decile 10 = highest scores (least confident). Monotonicity is the relevant signal, not absolute MSE values.

Dataset	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Traffic	0.057	0.084	0.192	0.297	0.346	0.384	0.510	0.716	0.964	1.537
Solar	0.006	0.014	0.022	0.039	0.086	0.113	0.174	0.258	0.447	0.932
Electricity	0.113	0.144	0.152	0.157	0.161	0.172	0.184	0.189	0.189	0.253
Weather	0.101	0.120	0.131	0.145	0.167	0.171	0.178	0.183	0.198	0.244

U Why SubDx Works Better on Some Datasets

The effectiveness of SubDx depends on the concentration of singular values in the input history, which we call *spectral concentration*. Figure 13 contrasts two regimes. High spectral concentration (left) occurs when a few dominant singular values capture most of the variance, indicating strong cross-channel redundancy; the off-subspace dimension is large and errors are easily detected. Low spectral concentration (right) occurs when singular values decline slowly, indicating weak cross-channel structure; here the subspace spans nearly all directions and the diagnostic signal is diluted.

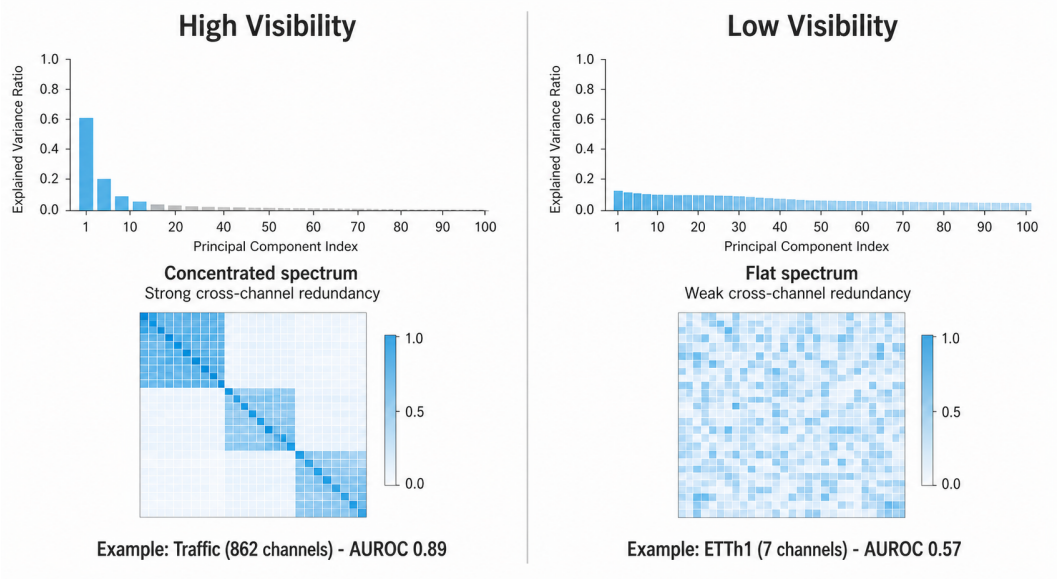

Figure 13: Spectral concentration explains SubDx performance across datasets. Left: concentrated singular values (high spectral concentration) give strong cross-channel structure, enabling clear error detection (e.g., Traffic, AUROC 0.89). Right: flat singular value spectra (low spectral concentration) provide weak structure and dilute the diagnostic signal (e.g., ETTh1, AUROC 0.57).

Table 28 reports the full-population real-error audit over 15,231,744 forecast timesteps from one frozen cell per dataset. High-error visibility is below all-error visibility in all seven cells, so aligned errors are an intrinsic false-negative mode; the per-cell variation also shows that alignment alone does not explain every miss.

Table 28: Full-population false negatives and visibility (one frozen cell per dataset). FN@50 uses the q80 high-error threshold and accepts the lowest-scored 50%; visibility is $\|\mathbf{P}_\perp(\hat{y}-y)\|_2/(\|\hat{y}-y\|_2+10^{-12})$.

Cell	FN@50 (%)	Mean visibility		
		All	q80	FN@50
ETTh1 / DLinear	49.6	0.031	0.013	0.012
ETTh2 / iTransformer	33.7	0.050	0.027	0.028
Electricity / PatchTST	38.3	0.905	0.816	0.805
Solar / PatchTST	4.1	0.588	0.482	0.267
Traffic / PatchTST	7.5	0.875	0.776	0.706
Weather / PatchTST	42.1	0.185	0.131	0.121
SWaT / DLinear	24.9	0.900	0.832	0.824
Pooled (7 cells)	25.3	–	–	–

On the 15-cell protocol-locked panel (the non-SWaT cells of Table 31; 9,280,512 timesteps; same q80 threshold and lowest-score 50% acceptance rule), pooled FN@50 is 28.37%. The difference from the 7-cell audit reflects cell composition and population size, not a change in the scoring rule: Table 28 isolates one frozen cell per dataset for the full-population visibility profile, whereas the 15-cell panel is the cross-method evaluation set.

V Controlled Geometry Factorial

To isolate visibility from channel count, redundancy, and error magnitude, we run a synthetic factorial with 192 cells (four seeds; 128 windows; history length 96; horizon 24; 20% high-error labels; high/low error norms 1.00/0.25). Table 29 reports marginal AUROC and FN@50 for each factor. Ambient dimension $N \in \{16, 64, 128\}$ changes AUROC only from 0.722 to 0.734. In contrast, the visibility contrast from $\alpha = 0.05$ to $\alpha = 0.75$ raises AUROC from 0.626 to 0.833 and lowers FN@50 from 73.6% to 31.5%, with error prevalence and magnitude fixed across cells. This is mechanism evidence on synthetic data, not a replacement for the real-data panels.

Table 29: Controlled synthetic factorial (192 cells, four seeds). Each block reports marginal means for one factor; error prevalence and magnitude are fixed across cells.

Factor	Level	AUROC	FN@50 (%)
Ambient N	16	0.722	53.5
	64	0.734	51.6
	128	0.732	52.5
Idiosyncratic noise	0.05	0.846	30.3
	0.30	0.613	74.8
Off-subspace error α	0.05	0.626	73.6
	0.75	0.833	31.5

W Foundation Models

We evaluate SubDx on three frozen pretrained forecasters under one matched Solar protocol: Chronos-T5-Small [Ansari et al., 2024], TimesFM 2.5 [Das et al., 2024], and Time-MoE [Shi et al., 2025]. All checkpoints are pinned and used without fine-tuning; SubDx scores their raw point predictions. Table 30 reports AUROC, rank correlation, Sel@50, FN@50, and the visibility/cancellation diagnostics. TimesFM 2.5 and Time-MoE rank below random; in both cells, false-negative rows show stronger perpendicular cancellation (mean cosine -0.995 and -0.975) rather than higher subspace alignment. All reported numbers use pinned checkpoints; runs whose script did not load the labelled checkpoint are excluded.

Table 30: SubDx on frozen foundation models under the matched Solar protocol (240 windows, $P = 64$, 10 channels, legal rank 8). No fine-tuning.

Model	AUROC	ρ	Sel@50 (%)	FN@50 (%)	Vis. all	Vis. q80	Canc. cos
TimesFM 2.5	0.422	0.178	-9.2	62.8	0.211	0.122	-0.995
Time-MoE	0.418	0.138	-9.4	64.3	0.123	0.098	-0.975
Chronos-T5-Small	0.578	0.428	25.0	44.9	0.276	0.188	-0.929

X Protocol-Locked Re-evaluation

Table 31 reports the full loader-aligned comparison; Table 1 summarizes the same panel for SubDx. Every row uses the producing loader's channel order, training-split normalization, the row-wise SubDx score, q80 timestep-MSE labels, and timestep-level pooling over the complete frozen test sequence. Fixed PCA is an offline-observation reference whose global basis uses the complete training history; the remaining columns share the same deployment interface (current history and one frozen point forecast). For SWaT, the channel order matches the loader (OT last), and target-reconstruction parity passes at zero maximum absolute difference for all three seeds. The locked panel supersedes the legacy per-backbone rows for cross-method comparison; the legacy rows are retained only in the original single-cell analyses below.

Table 31: Protocol-locked AUROC under the shared same-interface contract (q80 labels, full frozen test sequence). SubDx uses the precommitted requested-legal rank. Fixed PCA* is an offline-history reference and is not same-interface. SWaT/DLinear is the mean over seeds 42–44.

Cell	SubDx	Magn.	Last	Var.	Hist-vol	Hist-chg	Fixed PCA*
ETTh1 / DLinear	0.503	0.732	0.634	0.553	0.485	0.504	0.518
ETTh2 / PatchTST	0.593	0.490	0.630	0.555	0.592	0.574	0.578
ETTh2 / iTransformer	0.617	0.532	0.632	0.559	0.623	0.607	0.500
Electricity / DLinear	0.520	0.561	0.505	0.404	0.616	0.607	0.755
Electricity / PatchTST	0.598	0.577	0.508	0.420	0.593	0.596	0.770
Electricity / iTransformer	0.563	0.590	0.506	0.417	0.600	0.597	0.764
Solar / DLinear	0.779	0.626	0.755	0.766	0.371	0.362	0.901
Solar / PatchTST	0.919	0.661	0.789	0.715	0.466	0.419	0.937
Solar / iTransformer	0.801	0.546	0.735	0.694	0.418	0.384	0.910
Traffic / DLinear	0.840	0.804	0.669	0.767	0.512	0.524	0.885
Traffic / PatchTST	0.867	0.815	0.689	0.772	0.528	0.553	0.885
Traffic / iTransformer	0.854	0.817	0.701	0.758	0.534	0.560	0.891
Weather / DLinear	0.543	0.482	0.697	0.515	0.671	0.640	0.512
Weather / PatchTST	0.569	0.453	0.664	0.523	0.606	0.606	0.484
Weather / iTransformer	0.561	0.401	0.749	0.586	0.700	0.630	0.482
SWaT / DLinear	0.604	0.750	0.884	0.672	0.633	0.662	0.692

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (Section 1) state the three contributions—a post-hoc diagnostic, empirical validation, and a negative correction result—each matched by a corresponding section, and they state the scope boundaries explicitly.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated Limitations section after the conclusion discusses the need for non-trivial cross-channel structure, the channel-independent error failure mode, the linear factor-model and subspace-estimation assumptions, and the fact that the score identifies risky timesteps but not the error direction.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions (linear factor model, isotropic noise, and known subspace) are stated in Proposition 1 and Theorem 1; proof sketches are in Section 3.5, and full proofs are in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4.1 names the benchmarks and describes the evaluation protocol; Appendix D specifies the backbones, metrics, baselines, optimizer, and training hyperparameters; Algorithm 1 (Appendix C) specifies the scoring procedure. The rank $r = 8$ is fixed throughout without tuning.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code is not publicly released. The method is fully specified in Algorithm 1 (Appendix C) and Section 4.1; the datasets are established benchmarks (SWaT is available from iTrust under its access terms).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 describes the evaluation protocol and benchmarks; Appendix D specifies the backbones, optimizer, and training hyperparameters, and Section 4.3 covers the foundation models.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Table 1 reports means over 3 seeds where available; Appendix F reports multi-seed means and standard deviations, and Appendix S reports bootstrap confidence intervals. Single-seed results are noted. Spearman ρ and Kendall τ provide threshold-free ranking metrics.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix L specifies the GPU (NVIDIA RTX 4090D, 24GB), default settings, and seeds; Appendix C reports forward-pass and SVD times. These details are sufficient to reproduce the runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Foundational ML research on time-series diagnostics. No human subjects, no sensitive data. The datasets are established benchmarks used under their original terms.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper does not separately discuss societal impacts; it is foundational method work on forecast-risk diagnostics, and it does not release new data or models with high misuse risk.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new models or datasets posing misuse risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: Backbone and foundation-model papers are cited (PatchTST, iTransformer, DLinear, Chronos, TimesFM, Time-MoE). The datasets are established benchmarks used under their original terms, but the paper does not include explicit license statements or redistribute the data.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new datasets or pretrained models are introduced; SubDx is fully specified in Algorithm 1 (Appendix C).

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human-subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human-subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not an important, original, or non-standard component of the core method; SubDx uses SVD and linear algebra.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.