NORMAL SUPPORT IMBALANCE: CORRECTING SUPPORT-DEPENDENT SCORE BIAS IN MULTIVARIATE TIME-SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

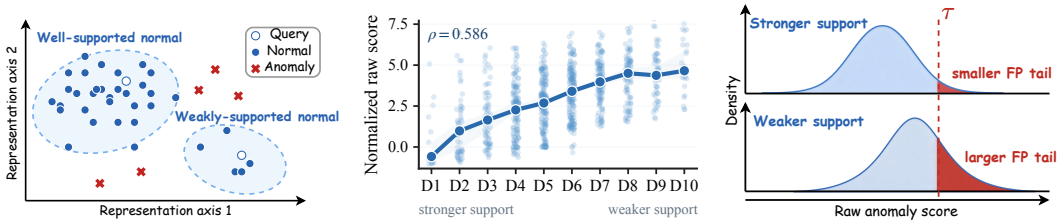
Normal-only multivariate time-series anomaly detectors learn from nominally normal data, yet the resulting models may cover legitimate normal behaviors unevenly. We study this Normal Support Imbalance and find that weakly supported normal observations tend to receive higher raw anomaly scores, creating the potential for false-positive disparities under a shared threshold. We introduce RestoreAD, which uses Row-Uniform regularization to discourage concentrated normal-to-normal relations, retrieves normal references for both anomaly scoring and support estimation, and applies Cross-Support Normality Calibration (CSNC). CSNC aligns training-score medians across support bins while exactly preserving within-bin rankings. On the official TSB-AD-M Eval split, RestoreAD achieves the highest mean on each reported metric among the compared baselines. Diagnostic analyses show that CSNC weakens the dependence of normal anomaly scores on estimated support difficulty and narrows cross-support false-positive-rate gaps. A controlled study further shows that unequal training support can produce score and false-positive disparities even between normal regimes with symmetric representation geometry. These findings highlight normal-reference coverage as an important consideration for reliable anomaly detection. The code is available at `https://anonymous.4open.science/r/RestoreAD/`.

1 INTRODUCTION

Modern industrial and cyber-physical systems generate multivariate sensor streams in which anomalous temporal patterns may signal faults or operational disruptions (Hundman et al., 2018; Su et al., 2019). Because labeled anomalies are scarce and costly to obtain, many multivariate time-series anomaly detection (MTSAD) methods learn normal behavior from training data and assign anomaly scores at test time (Song et al., 2023; Shen, 2025). Depending on the model, these scores may reflect reconstruction or prediction errors (Audibert et al., 2020; Zhao et al., 2020; Tuli et al., 2022), attention-based discrepancies (Xu et al., 2022; Yang et al., 2023), cross-sensor dependencies (Deng & Hooi, 2021; Wu et al., 2025), or comparisons with normal prototypes and stored references (Li et al., 2023; Park & Kang, 2026). Beyond changes in normal behavior over time (Wang et al., 2023; Kim et al., 2024), normality can encompass multiple legitimate operating regimes (Zhou et al., 2026). In multivariate streams, these regimes can differ in their channel-wise temporal patterns, which training data may cover unevenly. For detectors that aggregate channel-wise mismatches, a valid observation with weakly supported patterns can therefore receive a high anomaly score. How does unequal coverage affect scores assigned to normal observations and false-positive rates under a threshold shared across support regimes within the same series?

We refer to this disparity in reference coverage as *Normal Support Imbalance* (NSI): equally legitimate normal behaviors may receive different levels of support from a detector's learned normal reference. Conceptually, support denotes the availability and proximity of relevant normal references in representation space, rather than merely the raw frequency of an operating regime. In multivariate streams, reference coverage can differ across channels of the same normal context. Prior work has described multiple operating modes (Sipple, 2020), represented normality with temporal clusters (Shen et al., 2020), and modeled multiple normal patterns in the frequency domain (Chen et al., 2024a). Capturing these aspects of normality does not by itself ensure comparable ref-

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



(a) Unequal normal support (b) Support–score association (c) False-positive disparity

Figure 1: Normal Support Imbalance: motivation and diagnostic evidence. (a) Schematic of unequal reference coverage for legitimate normal queries. (b) On TSB-AD-M (Liu & Paparrizos, 2024), D1–D10 order support difficulty from stronger to weaker support. For display, raw scores are centered by each series' training median, scaled by its training IQR, and transformed by asinh. Light points show per-series decile summaries after seed aggregation; the dark curve shows cross-series medians with 95% bootstrap confidence intervals over series. The annotated $\rho = 0.586$ is the mean signed per-series Spearman correlation computed from original raw scores of held-out normal observations, rather than the displayed decile summaries. (c) Schematic of how a threshold τ shared across support regimes within a series can yield a larger false-positive tail for a right-shifted, weaker-support normal-score distribution. Panels (a) and (c) are schematic; panel (b) is empirical.

erence coverage across regimes. As illustrated in Figure 1(a), a normal query from a well-supported regime has many nearby references, whereas one from a weakly supported regime has few. Incomplete coverage may make valid queries appear atypical and lead to false alarms (Xing et al., 2026). Weak support alone does not imply an anomaly; related work likewise distinguishes difficult normal samples from anomalous contamination (Zhang et al., 2026). We use *support-dependent score bias* to describe systematic differences in raw-score locations for normal observations across support regimes. Under a shared threshold, such shifts can produce unequal false-positive rates.

We assess whether support difficulty, which increases as normal-reference support weakens, is associated with raw anomaly scores on held-out normal observations. On the TSB-AD-M evaluation split (Liu & Paparrizos, 2024), the mean per-series signed Spearman correlation is 0.586 across 179 valid series, with positive correlations in 174 (Figure 1(b)). This association and the upward score trend across support-difficulty bins are consistent with support-dependent score bias. Figure 1(c) schematically illustrates how a threshold shared across bins within a series can turn a normal-score shift into unequal false-positive rates. In the held-out analysis, setting each series' threshold at the 95th percentile of its training raw scores yields a mean maximum–minimum false-positive-rate (FPR) gap of 33.25 percentage points across support bins.

To address support-dependent score bias, we introduce RestoreAD, a support-aware MTSAD framework designed to improve score comparability across normal-support regimes. It learns channel-resolved representations with Row-Uniform regularization to discourage concentrated normal-to-normal relations, then builds a compact normal reference from selected training windows. For each query, the same channel-wise retrieved references provide a raw anomaly score from query–reference mismatch and a support-difficulty estimate from their relative isolation. *Cross-Support Normality Calibration* (CSNC) aligns training-score medians across support bins by a constant shift per bin, exactly preserving within-bin score differences and rankings.

Our main contributions are summarized as follows:

- **Normal Support Imbalance.** We identify Normal Support Imbalance as a source of support-dependent score bias: uneven reference coverage can raise scores for normal observations and create false-positive disparities across regimes under a shared threshold.
- **Support-aware anomaly detection.** We introduce RestoreAD, combining Row-Uniform regularization, channel-wise reference-based support estimation, and Cross-Support Normality Calibration (CSNC) to improve score comparability across normal-support regimes.
- **Empirical validation.** On the TSB-AD-M evaluation split, RestoreAD achieves the best overall performance across six metrics among baselines. Further analyses support the proposed mechanism and demonstrate reduced cross-support false-positive disparities.

2 PROBLEM FORMULATION

2.1 MULTIVARIATE TIME-SERIES ANOMALY DETECTION

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]^\top \in \mathbb{R}^{T \times C}$ be a multivariate time series of length T with C channels, where $\mathbf{x}_t \in \mathbb{R}^C$ is the observation at time t. For $L \le t \le T$, let $\mathbf{W}_t = [\mathbf{x}_{t-L+1}, \ldots, \mathbf{x}_t]^\top \in \mathbb{R}^{L \times C}$ be the length-L temporal context ending at t. We follow the normal-only MTSAD protocol (Shen, 2025; Park & Kang, 2026): $\mathcal{D}_{\mathrm{tr}} = \{\mathbf{W}_{t_i}\}_{i=1}^N$ comprises windows whose endpoints t_i lie in a predefined training prefix treated as nominally normal. We index these training windows by i below; fitting and calibration use no anomaly labels. In the subsequent segment, $y_t \in \{0, 1\}$ denotes the ground-truth label at t (0: normal; 1: anomalous). A detector trained on $\mathcal{D}_{\mathrm{tr}}$ assigns each time point a raw anomaly score $s_t \in \mathbb{R}$, with larger values indicating greater anomalousness. The raw-score prediction under a threshold τ shared across support regimes within a series is $\hat{y}_t = \mathbb{I}[s_t > \tau]$.

2.2 NORMAL SUPPORT IMBALANCE

For a temporal context $\mathbf{W}_t$, let $\mathbf{Z}_t = f_\theta(\mathbf{W}_t)$ be its learned representation, and let $\mathcal{Z}_{\mathrm{tr}} = \{\mathbf{Z}_i\}_{i=1}^N$ collect the representations of nominal training windows. We use *normal support* to describe the availability and proximity of relevant training representations around $\mathbf{Z}_t$ in representation space. This coverage need not mirror the raw frequency of an operating regime. We later estimate support using the channel-wise relative isolation of retrieved stored references.

We define *Normal Support Imbalance* (NSI) as substantial variation in support across contexts ending at time points labeled normal. Let $H(\mathbf{W}_t)$ denote the conceptual *support difficulty*, with larger values indicating weaker normal support. For $y_a = y_b = 0$, we may have $H(\mathbf{W}_a) \ll H(\mathbf{W}_b)$: both endpoint observations are normal, but $\mathbf{W}_b$ has weaker reference coverage. High support difficulty therefore does not, by itself, establish that the endpoint observation is anomalous.

2.3 SUPPORT-DEPENDENT SCORE BIAS

Normal Support Imbalance can affect detection when raw scores differ systematically across support regimes even among normal observations. For a given series, let b_t index support-difficulty bins, ordered from stronger to weaker support. For analysis, define the normal-score location in bin b as $m_b^{(0)} = \mathrm{median}(s_t \mid y_t = 0, b_t = b)$. We refer to systematic variation in $m_b^{(0)}$ across bins as *support-dependent score bias*, a shift in normal-score baselines.

To illustrate how such variation may arise, consider the working model

$$s_t = A(\mathbf{W}_t) + \Delta(H(\mathbf{W}_t)) + \varepsilon_t, \tag{1}$$

where $A(\mathbf{W}_t)$ represents anomaly-related evidence, $\Delta(H(\mathbf{W}_t))$ a support-dependent score shift, and ε_t residual variation. These components are not separately identified from observed scores. If Δ increases with support difficulty, it can raise the scores of weakly supported normal observations.

For a threshold τ shared across support bins within a series, the bin-wise false-positive rate is $\mathrm{FPR}_b(\tau) = \Pr(s_t > \tau \mid y_t = 0, b_t = b)$. A support-dependent shift can yield unequal false-positive rates, although median differences alone do not determine the score tails. We therefore align score medians estimated from the training prefix across support bins while exactly preserving pairwise score differences and rankings within each bin by construction.

3 METHOD

3.1 FRAMEWORK OVERVIEW

Figure 2 summarizes how RestoreAD addresses the support-dependent score bias described in Section 2.3 through three components: (i) channel-preserving representation learning with Row-Uniform regularization to discourage concentrated normal-to-normal relations; (ii) a compact normal reference whose retrieved entries provide query–reference mismatch for raw anomaly scoring and reference isolation for support-difficulty estimation; and (iii) *Cross-Support Normality Calibration* (CSNC), which aligns training-prefix score medians across support bins while exactly preserving score differences and rankings among time points in the same support bin.

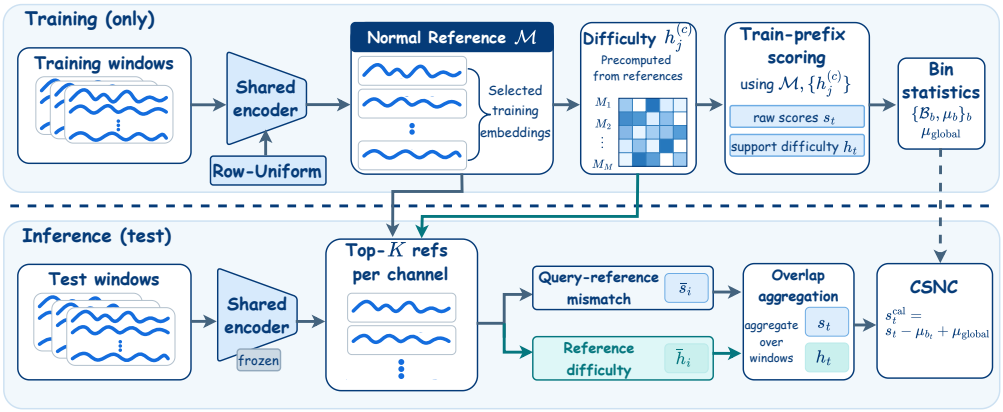


Figure 2: Framework overview of RestoreAD.

We treat the predefined training prefix as nominally normal when learning the encoder, constructing the reference, and estimating calibration statistics, without using anomaly labels. At inference, the encoder and reference yield window-level mismatch and support difficulty, which are averaged over windows covering each time point before CSNC calibrates the resulting raw scores.

3.2 Channel-Preserving Normal Representation

Given a temporal context $\mathbf{W}_t \in \mathbb{R}^{L\times C}$, let $\mathbf{w}_t^{(c)} \in \mathbb{R}^L$ denote its c-th channel, where $c \in \{1,\dots,C\}$. We implement the window-level encoder f_θ using a channel-independent design (Nie et al., 2023), applying a shared temporal encoder g_θ to each channel:

$$\mathbf{z}_t^{(c)} = g_\theta\left(\mathbf{w}_t^{(c)}\right) \in \mathbb{R}^d, \qquad c = 1,\dots,C. \tag{2}$$

The resulting channel-resolved representation is $\mathbf{Z}_t = f_\theta(\mathbf{W}_t) = [\mathbf{z}_t^{(1)},\dots,\mathbf{z}_t^{(C)}]^\top \in \mathbb{R}^{C\times d}$. Parameter sharing provides a common temporal feature extractor across channels; channel-resolved outputs preserve channel identity for relation regularization and reference retrieval.

We train g_θ on windows from the nominally normal training prefix using a VICReg-style objective (Bardes et al., 2022) applied to projected channel representations from two stochastic views, $\mathcal{L}_{\text{rep}} = \lambda_{\text{inv}}\mathcal{L}_{\text{inv}} + \lambda_{\text{var}}\mathcal{L}_{\text{var}} + \lambda_{\text{cov}}\mathcal{L}_{\text{cov}}$. The invariance term encourages agreement between views, the variance term discourages feature collapse, and the covariance term reduces redundancy. We optimize this objective jointly with the Row-Uniform regularizer in Section 3.3.

3.3 Relation-Balanced Representation Learning

The representation objective in Section 3.2 encourages view consistency and discourages feature collapse, but does not control how an anchor distributes its relations over other training windows. A few candidates may therefore receive most of its relation weight. We introduce a *Row-Uniform* regularizer to discourage this concentration without requiring support-regime labels.

At each training step, we sample an anchor set $\mathcal{A}$ and a candidate set $\mathcal{R}$ from $\mathcal{D}_{\text{tr}}$. For each anchor i, we form $\mathcal{C}_i \subseteq \mathcal{R}$ by excluding temporally overlapping windows when non-overlapping candidates are available. For channel c and $j \in \mathcal{C}_i$, let $r_{ij}^{(c)} = \text{sim}(\mathbf{z}_i^{(c)}, \mathbf{z}_j^{(c)})/\tau_{\text{rel}}$, where sim denotes cosine similarity and $\tau_{\text{rel}} > 0$ is a temperature. The relation distribution is $\mathbf{p}_i^{(c)} = \text{softmax}((r_{ij}^{(c)})_{j\in\mathcal{C}_i})$; let $\mathbf{u}_i$ be uniform over $\mathcal{C}_i$. The Row-Uniform objective is defined as

$$\mathcal{L}_{\text{RU}} = \frac{1}{|\mathcal{A}|C}\sum_{i\in\mathcal{A}}\sum_{c=1}^{C} D_{\text{KL}}\left(\mathbf{u}_i \,\|\, \mathbf{p}_i^{(c)}\right). \tag{3}$$

The Row-Uniform term penalizes dispersion in relation logits within each channel, discouraging concentrated weights on a few normal candidates (Appendix B.1).

We jointly optimize representation learning and relation regularization:

$$\mathcal{L} = \mathcal{L}_{\text{rep}} + \lambda_{\text{RU}} \mathcal{L}_{\text{RU}}, \tag{4}$$

where λ_{RU} controls the regularization strength. The learned channel-preserving representations are then used to construct the normal reference and estimate support difficulty in Section 3.4.

3.4 Normal Reference and Support Difficulty

From the learned training-window representations, we construct a compact normal reference $\mathcal{M} = \{\mathbf{M}_j\}_{j=1}^{M}$, where $\mathbf{M}_j = [\mathbf{m}_j^{(1)}, \ldots, \mathbf{m}_j^{(C)}]^\top$ is the channel-resolved representation of a selected training window. We cluster window-level descriptors derived from these representations and select the observed window nearest each cluster center. The reference thus preserves observed training patterns and their channel-specific embeddings for retrieval and support estimation.

We measure how isolated each stored reference is relative to the others. For reference j and channel c, its isolation score and normalized support difficulty are

$$\rho_j^{(c)} = 1 - \max_{q \in \mathcal{V}_j} \operatorname{sim}\left(\mathbf{m}_j^{(c)}, \mathbf{m}_q^{(c)}\right), \qquad h_j^{(c)} = \widehat{F}_c\left(\rho_j^{(c)}\right), \tag{5}$$

where $\mathcal{V}_j$ excludes temporally overlapping references when non-overlapping alternatives are available, and $\widehat{F}_c$ maps isolation values to empirical percentiles among distinct selected training windows in channel c. A larger $h_j^{(c)}$ indicates greater isolation and weaker normal support.

For a query window $\mathbf{W}_i$, let $\mathcal{N}_K^{(c)}(i)$ contain its K most similar references in channel c, and let $d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \operatorname{sim}(\mathbf{a}, \mathbf{b})$. The same neighbors yield both quantities:

$$\bar{s}_i = \frac{1}{CK} \sum_{c=1}^{C} \sum_{j \in \mathcal{N}_K^{(c)}(i)} d_{\cos}(\mathbf{z}_i^{(c)}, \mathbf{m}_j^{(c)}), \qquad \bar{h}_i = \frac{1}{CK} \sum_{c=1}^{C} \sum_{j \in \mathcal{N}_K^{(c)}(i)} h_j^{(c)}. \tag{6}$$

Here, $\bar{s}_i$ measures query–reference mismatch, whereas $\bar{h}_i$ summarizes the relative isolation of the retrieved normal evidence. A large $\bar{h}_i$ does not by itself imply an anomaly; it instead serves as a retrieved-reference proxy for the conceptual support difficulty $H(\mathbf{W}_i)$ defined in Section 2.2.

Let $\mathcal{I}_t$ index scored windows covering t. Averaging their outputs yields point-level quantities:

$$s_t = \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \bar{s}_i, \qquad h_t = \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \bar{h}_i. \tag{7}$$

The resulting h_t conditions the calibration of s_t by CSNC in Section 3.5.

3.5 Cross-Support Normality Calibration

The point-level support difficulty h_t summarizes the support of the normal references used to produce the raw score s_t. CSNC uses this information to estimate support-specific score locations from the nominally normal training prefix and align them across support bins.

Let $\mathcal{T}_{\text{tr}}$ index the scored time points in the training prefix. We partition their support difficulties into at most B quantile bins $\{\mathcal{B}_b\}$, ordered from stronger to weaker support and subject to a minimum bin size. For each bin, we compute $\mu_b = \operatorname{median}\{s_u : u \in \mathcal{T}_{\text{tr}},\, h_u \in \mathcal{B}_b\}$, together with the global median $\mu_{\text{global}} = \operatorname{median}\{s_u : u \in \mathcal{T}_{\text{tr}}\}$. Using the training bin boundaries, we assign each test time point to its support bin b_t and calibrate its raw anomaly score as

$$s_t^{\text{cal}} = s_t - \mu_{b_t} + \mu_{\text{global}}. \tag{8}$$

This correction places the median calibrated score of every training bin at μ_{global}. The shift may be positive or negative, depending on the estimated bin median; weak support does not automatically lower a score. Because the shift is constant within each bin, CSNC exactly preserves within-bin score differences and rankings. Using medians limits sensitivity to extreme training scores. Appendix A presents training and calibration (Algorithm 1) and inference (Algorithm 2).

3.6 Theoretical Properties and Complexity

We summarize the relation-balancing property of Row-Uniform and analyze CSNC under a normal-score location-shift model. Derivations and additional complexity details appear in Appendix B.

Relation balancing. The Row-Uniform penalty equals the log-mean-exp of the relation logits minus their mean. It is nonnegative, vanishes exactly when the logits are equal, and is invariant to common shifts. Near equal logits, its leading term is half their uniform-weighted variance. Together with $\mathcal{L}_{\text{rep}}$, it discourages concentrated relations without fixing absolute similarity levels (Appendix B.1).

Cross-support calibration. CSNC aligns the score medians of nonempty training support bins exactly and preserves all within-bin score differences and rankings (Appendix B.2). For normal observations, consider $S_{\text{raw}} = Z + \Delta(H)$ with $Z \perp H \mid Y = 0$. Let $\mu_b^\star = \text{median}(S_{\text{raw}} \mid Y = 0, H \in \mathcal{B}_b)$ and $\mu_{\text{global}}^\star = \text{median}(S_{\text{raw}} \mid Y = 0)$. The population version of CSNC assigns $S_{\text{cal}}^\star = S_{\text{raw}} - \mu_b^\star + \mu_{\text{global}}^\star$ to observations with $H \in \mathcal{B}_b$.

Proposition 1 (Cross-support FPR disparity bound)**.** *Suppose Δ is L_Δ-Lipschitz on $\mathcal{H}_0 = \text{supp}(H \mid Y = 0)$ and Z has a density bounded above by $\bar{f}$. For each bin $\mathcal{B}_b$ with $\Pr(H \in \mathcal{B}_b \mid Y = 0) > 0$, let $w_b = \sup_{h,h' \in \mathcal{B}_b \cap \mathcal{H}_0} |h - h'|$ and define $\text{FPR}^{\text{cal}}(\tau \mid h) = \Pr(S_{\text{cal}}^\star > \tau \mid Y = 0, H = h)$. For $h \in \mathcal{B}_b \cap \mathcal{H}_0$, $h' \in \mathcal{B}_{b'} \cap \mathcal{H}_0$, and any threshold τ shared across bins,*

$$\left|\text{FPR}^{\text{cal}}(\tau \mid h) - \text{FPR}^{\text{cal}}(\tau \mid h')\right| \le \bar{f} L_\Delta (w_b + w_{b'}). \tag{9}$$

Proposition 1 bounds residual FPR disparity by within-bin variation in the support-dependent shift (Appendix B.3). Averaging over H within each bin yields the same bound for bin-wise FPRs. If Δ is constant within each bin, the calibrated normal-score distributions and FPRs agree across bins (Appendix B.4). Appendix B.5 bounds the effect of median estimation under its stated assumptions.

Complexity. For A anchors, R candidates, C channels, and embedding dimension d, computing sampled relation logits costs $\mathcal{O}(ARCd)$ per training step; evaluating Row-Uniform once these logits are available costs $\mathcal{O}(ARC)$. Matching a query against M stored references costs $\mathcal{O}(CMd)$ per window. Support estimation reuses retrieved references, while CSNC adds a bin lookup and a constant score shift. Preprocessing, top-K selection, and storage costs appear in Appendix B.6.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate RestoreAD on TSB-AD-M, the multivariate track of the TSB-AD benchmark (Liu & Paparrizos, 2024), using the same dataset partition as PaAno (Park & Kang, 2026). We select hyperparameters on the 20-series *Tuning* split and evaluate the fixed configuration on the 180-series *Eval* split. Within each series, a predefined, nominally normal training prefix is used for fitting and calibration; the subsequent segment is used for testing. Fitting and calibration use no anomaly labels. Dataset statistics and protocol details appear in Appendix C.1.

Baselines. We compare RestoreAD with 22 baselines spanning classical detectors, deep anomaly detectors, and recent time-series models. Representative methods include PCA (Aggarwal, 2017), OmniAnomaly (Su et al., 2019), TranAD (Tuli et al., 2022), Anomaly Transformer (Xu et al., 2022), CATCH (Wu et al., 2025), STREAM-VAE (Özer et al., 2026), and PaAno (Park & Kang, 2026). The main table shows 17 baselines; Appendix C.2 reports full results for all 22.

Implementation details. For baselines covered by TSB-AD (Liu & Paparrizos, 2024), we use its released implementations and multivariate configurations. For additional methods, we use official implementations and select method-specific settings on the *Tuning* split. We evaluate each method with seeds 2027, 2028, and 2029. Using the TSB-AD evaluator without point adjustment, we first average each metric over the 180 *Eval* series and then report the mean and standard deviation across seeds. We implement RestoreAD in PyTorch and use one NVIDIA GeForce RTX 3090 GPU per run. Further configurations and reporting details appear in Appendix C.3.

Metrics. We report AUC-ROC, AUC-PR, and Point-F1 for point-wise detection; Range-F1 (Tatbul et al., 2018) for anomalous ranges; and VUS-ROC and VUS-PR (Paparrizos et al., 2022; Boniol

Table 1: Overall performance on the official TSB-AD-M *Eval* split. Mean ± standard deviation over three seeds. The best mean in each column is boldfaced; the second-best is underlined.

	Method	Range-wise			Point-wise		
		VUS-PR↑	VUS-ROC↑	Range-F1↑	AUC-PR↑	AUC-ROC↑	Point-F1↑
Stat./ML	PCA (2017)	0.2691 ± .0000	0.7058 ± .0000	0.3277 ± .0000	0.2409 ± .0000	0.6759 ± .0000	0.3120 ± .0000
	KMeansAD (2001)	0.2881 ± .0023	0.7298 ± .0004	0.3298 ± .0014	0.2477 ± .0024	0.6918 ± .0001	0.3064 ± .0017
	IForest (ICDM'08)	0.2487 ± .0000	0.7437 ± .0000	0.2434 ± .0000	0.2095 ± .0000	0.7040 ± .0000	0.2798 ± .0000
Deep AD	LSTMAD (ESANN'15)	0.3311 ± .0006	0.7709 ± .0020	0.2938 ± .0017	0.3327 ± .0004	0.7360 ± .0021	0.3953 ± .0023
	OmniAnomaly (KDD'19)	0.3153 ± .0000	0.7434 ± .0005	0.2929 ± .0004	0.3152 ± .0000	0.7056 ± .0009	0.3752 ± .0001
	USAD (KDD'20)	0.4031 ± .0010	0.7674 ± .0020	<u>0.4236</u> ± .0044	0.3621 ± .0004	0.7384 ± .0018	0.4333 ± .0006
	DeepAnT (IEEE Access'19)	0.3555 ± .0007	<u>0.8090</u> ± .0003	0.3016 ± .0012	0.3529 ± .0009	<u>0.7762</u> ± .0014	0.4155 ± .0003
	DAMR (KDD'26)	0.3939 ± .0058	0.7707 ± .0092	0.4103 ± .0074	0.3671 ± .0075	0.7399 ± .0097	0.4159 ± .0068
	STREAM-VAE (IEEE IV'26)	<u>0.4293</u> ± .0029	0.7710 ± .0008	0.4191 ± .0061	<u>0.3882</u> ± .0023	0.7548 ± .0011	<u>0.4547</u> ± .0022
Modern TS	TranAD (PVLDB'22)	0.3178 ± .0004	0.7452 ± .0007	0.2827 ± .0005	0.3179 ± .0004	0.7089 ± .0010	0.3831 ± .0009
	Anomaly Transformer (ICLR'22)	0.1190 ± .0037	0.5136 ± .0032	0.1583 ± .0065	0.0768 ± .0023	0.4756 ± .0047	0.1218 ± .0008
	TimesNet (ICLR'23)	0.3960 ± .0108	0.7091 ± .0010	0.4063 ± .0043	0.3435 ± .0093	0.6680 ± .0014	0.4142 ± .0074
	FITS (ICLR'24)	0.3633 ± .0011	0.7000 ± .0010	0.3248 ± .0040	0.3045 ± .0014	0.6514 ± .0012	0.3689 ± .0005
	PatchTST (ICLR'23)	0.3898 ± .0010	0.7167 ± .0004	0.3816 ± .0027	0.3354 ± .0006	0.6748 ± .0008	0.4066 ± .0007
	CATCH (ICLR'25)	0.3774 ± .0007	0.7320 ± .0004	0.2618 ± .0033	0.3566 ± .0009	0.6937 ± .0005	0.3935 ± .0007
	KAN-AD (ICML'25)	0.2780 ± .0013	0.6842 ± .0009	0.2369 ± .0031	0.2413 ± .0020	0.6392 ± .0009	0.2938 ± .0014
Rep.	PaAno (ICLR'26)	0.4274 ± .0013	0.7947 ± .0007	0.4074 ± .0059	0.3784 ± .0037	0.7633 ± .0009	0.4292 ± .0006
	RestoreAD (Ours)	**0.4662** ± .0047	**0.8167** ± .0050	**0.4627** ± .0068	**0.4260** ± .0024	**0.7897** ± .0047	**0.4655** ± .0028

et al., 2025) across score thresholds and temporal tolerances. Following TSB-AD (Liu & Paparrizos, 2024), VUS-PR is the primary metric. Definitions and evaluation settings appear in Appendix C.4.

4.2 Overall Performance

Table 1 shows that RestoreAD achieves the highest mean on all six metrics on the official TSB-AD-M *Eval* split. On VUS-PR, the primary metric, it outperforms STREAM-VAE and PaAno, the two strongest baselines on that measure. Its lead extends across the remaining range-wise and point-wise metrics, even as the strongest competing method varies between measures. The following ablation and cross-support analyses examine the contributions of relation balancing and CSNC, as well as their effects on support-dependent score disparities.

4.3 Component Ablation

Table 2 assesses the contributions of Row-Uniform (RU) regularization and Cross-Support Normality Calibration (CSNC). Relative to Base, RU improves the mean of all six metrics, with the largest absolute gains in AUC-PR, VUS-PR, and Point-F1. CSNC alone primarily improves AUC-ROC and VUS-ROC. Adding CSNC to RU further improves both ROC metrics, and the full model achieves the highest mean on all six metrics. Because CSNC preserves within-bin rankings, its ROC gains are consistent with changes in score comparisons across support bins. Further design ablations in Appendix D.1 show gains over shuffled support on five metrics, with Range-F1 as an exception, and similar mean performance under quantile and equal-width binning.

4.4 Support-Dependent Score Bias and Cross-Support Comparability

We examine whether weaker normal support is associated with higher raw anomaly scores and whether CSNC improves comparability across support regimes. For each series and seed, we compute the signed Spearman correlation between support difficulty and raw scores over held-out normal observations, then average over valid seeds within each series. Across 179 valid series, the mean signed correlation is 0.586 (95% CI: 0.548–0.621), and 174 of these series have positive correla-

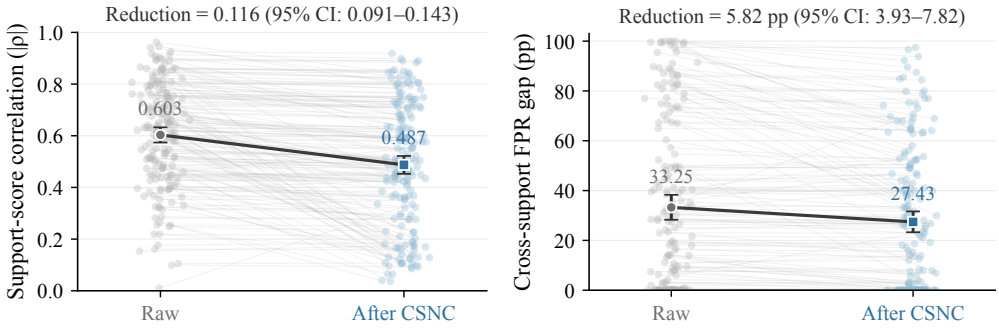
Table 2: Component ablation on the official TSB-AD-M *Eval* split. RU and CSNC denote Row-Uniform regularization and Cross-Support Normality Calibration, respectively. All variants share the encoder architecture, reference-construction procedure, and raw-scoring rule. Results are mean ± standard deviation over seeds 2027, 2028, and 2029.

Variant	RU	CSNC	Range-wise			Point-wise		
			VUS-PR↑	VUS-ROC↑	Range-F1↑	AUC-PR↑	AUC-ROC↑	Point-F1↑
Base			0.4410 ± .0024	0.7874 ± .0010	0.4504 ± .0043	0.3962 ± .0028	0.7602 ± .0009	0.4396 ± .0022
Base + RU	✓		0.4647 ± .0047	0.8009 ± .0023	0.4590 ± .0053	0.4244 ± .0022	0.7729 ± .0030	0.4627 ± .0028
Base + CSNC		✓	0.4414 ± .0022	0.8022 ± .0018	0.4524 ± .0013	0.3970 ± .0029	0.7761 ± .0012	0.4412 ± .0019
RestoreAD	✓	✓	**0.4662** ± .0047	**0.8167** ± .0050	**0.4627** ± .0068	**0.4260** ± .0024	**0.7897** ± .0047	**0.4655** ± .0028

tions. Thus, normal observations with weaker estimated support tend to receive higher raw scores. Further signed and seed-wise statistics appear in Appendix D.2.

Figure 3(a) shows the absolute value of each series' seed-averaged correlation on held-out normal observations, measuring dependence strength regardless of direction. CSNC reduces the cross-series mean from 0.603 to 0.487, a paired reduction of 0.116 (95% CI: 0.091–0.143).

We next assess decision-level disparity under a threshold shared across support bins. For each series and seed, raw and CSNC thresholds are fitted separately at the 95th percentile of their respective training scores and held fixed across that series' support bins. Figure 3(b) shows that CSNC reduces the mean maximum–minimum FPR gap across bins from 33.25 to 27.43 percentage points, a paired reduction of 5.82 points (95% CI: 3.93–7.82). The median within-bin raw-to-calibrated rank correlation is 1.000, while mean anomaly TPR increases by 0.72 percentage points. Together, the reduced score dependence and FPR gap are consistent with attenuated support-dependent score bias. The remaining gap indicates that cross-support disparity persists.

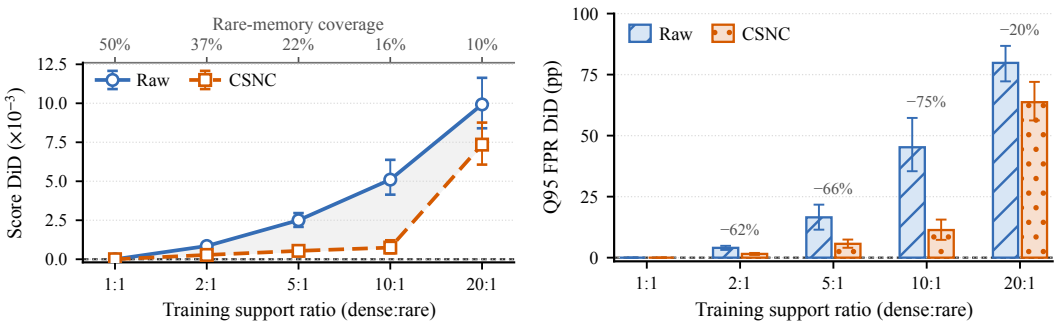


(a) Support–score dependence.

(b) Cross-support FPR gap.

Figure 3: **Support-dependent score bias and cross-support FPR disparity on TSB-AD-M *Eval*.** (a) Absolute Spearman correlation between support difficulty and anomaly scores on held-out normal observations, taken after averaging signed correlations over valid seeds within each series. (b) Maximum–minimum FPR gap across support bins. For each series and seed, raw and CSNC thresholds are fitted separately at the 95th percentile of their respective training scores and held fixed across bins. Light pairs represent series after seed averaging; large markers and error bars show cross-series means and 95% bootstrap confidence intervals. Top annotations give paired mean reductions and their 95% paired-bootstrap confidence intervals. Support bins and thresholds use training data only; evaluation labels are used only for these diagnostics.

Controlled support-imbalance study. To separate training support from intrinsic pattern complexity, we construct two normal regimes with mirror-symmetric representation geometry and vary their relative training sample counts from 1:1 to 20:1. We hold the total training size, test set, representation mapping, and memory budget fixed, and evaluate both assignments of the rare regime. For each score version, a Q95 threshold is fitted using the balanced-condition training scores and held fixed across ratios. We measure difference-in-differences (DiD): the change from 1:1 in the



(a) Score DiD.

(b) Q95 FPR DiD.

Figure 4: **Controlled support-imbalance study.** DiD is the change from 1:1 in the rare regime minus that in the dense regime. (a) Median normal-score DiD; top labels show the rare regime's share of memory centers. (b) Q95 FPR DiD; percentages show the relative change from Raw to CSNC (negative values indicate reductions). Raw and CSNC thresholds are fitted separately in the 1:1 condition and held fixed across ratios. Error bars show 95% bootstrap CIs over 10 seeds.

rare regime minus the corresponding change in the dense regime. The construction and evaluation protocol for this study are detailed in Appendix C.5.

Figure 4 shows that the rare regime's share of memory centers falls from 50% to 10% as the training ratio increases. Raw score DiD rises across the five conditions (mean within-seed Spearman $\rho = 0.99$). At 10:1, CSNC reduces score DiD from 0.0051 to 0.00075 and Q95 FPR DiD from 45.2 to 11.3 percentage points; the latter is a 75% relative reduction. At 20:1, the rare regime receives only 10% of memory centers, and the FPR DiD reduction falls to 20%. The controlled comparison shows that unequal training support can produce score and decision disparities even between geometrically symmetric normal regimes. CSNC reduces these disparities, although its correction is incomplete.

A hyperparameter sensitivity analysis finds modest variation in the six detection metrics over the tested bin counts and minimum bin sizes (Appendix D.3). Three case studies visualize localized rescue, regime-level realignment, and rescue near a true anomaly (Appendix D.4; Fig. 5).

5 Related Work

Multivariate time-series anomaly detection. Deep MTSAD methods model normality through reconstruction, prediction, dependencies, or representation learning. Reconstruction and prediction methods detect deviations from temporal regularities (Su et al., 2019; Audibert et al., 2020; Li et al., 2021; Tuli et al., 2022), while graph and attention methods capture temporal and cross-variable dependencies (Zhao et al., 2020; Deng & Hooi, 2021; Xu et al., 2022; Yang et al., 2023). Recent work explores memory and prototype modeling (Song et al., 2023; Shen, 2025), decomposition and diffusion (Wang et al., 2023), causal contrastive learning (Kim et al., 2025), smooth local approximation (Zhou et al., 2025), idempotent reconstruction (Sun et al., 2025), and low-rank attention (Shimillas et al., 2026). PaAno (Park & Kang, 2026) uses lightweight patch representations to score queries against normal training patches. How normal heterogeneity shapes anomaly-score distributions remains unclear. Heterogeneous normality and score calibration are discussed in Appendix E.

6 Conclusion

We study Normal Support Imbalance, where normal patterns receive unequal reference coverage and may acquire different anomaly-score baselines. RestoreAD combines relation-regularized representation learning with Cross-Support Normality Calibration to improve score comparability across support regimes while preserving within-bin rankings. On TSB-AD-M, RestoreAD leads the compared methods on all reported metrics, while diagnostic and controlled studies show reduced support-dependent score and false-positive disparities. These findings highlight the importance of normal-reference coverage for reliable anomaly detection across heterogeneous operating regimes.

AI USE STATEMENT

We used generative AI to improve the clarity and readability of the manuscript.

REFERENCES

Charu C. Aggarwal. *Outlier Analysis*. Springer International Publishing, 2 edition, 2017.

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

Sarah Alnegheimish, Zelin He, Matthew Reimherr, Akash Chandrayan, Abhinav Pradhan, and Luca D'Angelo. M^2ad: Multi-sensor multi-system anomaly detection through global scoring and calibrated thresholding. In *AISTATS*, 2025.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In *SIGKDD*, 2020.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *ICLR*, 2020.

Paul Boniol, Ashwin K. Krishna, Marine Bruel, Qinghua Liu, Mingyi Huang, Themis Palpanas, Ruey S. Tsay, Aaron Elmore, Michael J. Franklin, and John Paparrizos. Vus: effective and efficient accuracy measures for time-series anomaly detection. *The VLDB Journal*, 2025.

Evgeny Burnaev and Vladislav Ishimtsev. Conformalized density-and distance-based anomaly detection in time-series data. *arXiv preprint arXiv:1608.04585*, 2016.

Chris U. Carmona, François-Xavier Aubet, Valentin Flunkert, and Jan Gasthaus. Neural contextual anomaly detection for time series. In *IJCAI*, 2022.

Feiyi Chen, Yingying Zhang, Zhen Qin, Lunting Fan, Renhe Jiang, Yuxuan Liang, Qingsong Wen, and Shuiguang Deng. Learning multi-pattern normalities in the frequency domain for efficient time series anomaly detection. In *ICDE*, 2024a.

Zejian Chen, Zuoyong Li, Xinwei Chen, Xiaobo Chen, Haoyi Fan, and Rong Hu. Rectifying inaccurate unsupervised learning for robust time series anomaly detection. *Information Sciences*, 2024b.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *AAAI*, 2021.

Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track*, 2012.

Thi Kieu Khanh Ho and Narges Armanfard. Contaminated multivariate time-series anomaly detection with spatio-temporal graph conditional diffusion models. In *UAI*, 2025.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *SIGKDD*, 2018.

Dongmin Kim, Sunghyun Park, and Jaegul Choo. When model meets new normals: Test-time adaptation for unsupervised time-series anomaly detection. In *AAAI*, 2024.

HyunGi Kim, Jisoo Mok, Dongjun Lee, Jaihyun Lew, Sungjae Kim, and Sungroh Yoon. Causality-aware contrastive learning for robust multivariate time-series anomaly detection. In *ICML*, 2025.

HyunGi Kim, Jisoo Mok, Hyungyu Lee, Juhyeon Shin, and Sungroh Yoon. Candi: Curated test-time adaptation for multivariate time-series anomaly detection under distribution shift. In *AAAI*, 2026.

Tian Lan, Hao Duong Le, Jinbo Li, Wenjun He, Meng Wang, Chenghao Liu, and Chen Zhang. Towards foundation models for zero-shot time series anomaly detection: Leveraging synthetic data and relative context discrepancy. In *ICML*, 2026.

Yuxin Li, Wenchao Chen, Bo Chen, Dongsheng Wang, Long Tian, and Mingyuan Zhou. Prototype-oriented unsupervised anomaly detection for multivariate time series. In *ICML*, 2023.

Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. COPOD: copula-based outlier detection. In *ICDM*, 2020.

Zhihan Li, Youjian Zhao, Jiaqi Han, Ya Su, Rui Jiao, Xidao Wen, and Dan Pei. Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. In *SIGKDD*, 2021.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *ICDM*, 2008.

Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In *NeurIPS*, 2024.

Jianming Lv, Yaquan Wang, and Shengjing Chen. Adaptive multivariate time-series anomaly detection. *Information Processing & Management*, 2023.

Pankaj Malhotra, Lovekesh Vig, Gautam M. Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In *ESANN*, 2015.

Natalia Martinez Gil, Fearghal O'Donncha, Wesley M. Gifford, Nianjun Zhou, Dhaval C. Patel, and Roman Vaculin. Adaptive conformal anomaly detection with time series foundation models for signal monitoring. In *ICLR*, 2026.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, 1990.

Xudong Mou, Rui Wang, Bo Li, Tianyu Wo, Jie Sun, Hui Wang, and Xudong Liu. Roca: Robust contrastive one-class time series anomaly detection with contaminated data. *IEEE Transactions on Neural Networks and Learning Systems*, 2026.

Mohsin Munir, Shoaib Ahmed Siddiqui, Andreas Dengel, and Sheraz Ahmed. Deepant: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*, 2019.

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

Kadir-Kaan Özer, René Ebeling, and Markus Enzweiler. Stream-vae: Dual-path routing for slow and fast dynamics in vehicle telemetry anomaly detection. In *IEEE IV*, 2026.

John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S Tsay, Aaron J Elmore, and Michael J Franklin. Volume under the surface: A new accuracy evaluation measure for time-series anomaly detection. *Proc. VLDB Endow.*, 2022.

Jinju Park and Seokho Kang. Paano: Patch-based representation learning for time-series anomaly detection. In *ICLR*, 2026.

Chen Qiu, Timo Pfrommer, Marius Kloft, Stephan Mandt, and Maja Rudolph. Neural transformation learning for deep anomaly detection beyond images. In *ICML*, 2021.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *ICML*, 2018.

Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, 2014.

Ke-Yuan Shen. Learn hybrid prototypes for multivariate time series anomaly detection. In *ICLR*, 2025.

Lifeng Shen, Zhuocong Li, and James T. Kwok. Timeseries anomaly detection using temporal hierarchical one-class network. In *NeurIPS*, 2020.

Charalampos Shimillas, Kleanthis Malialis, Konstantinos Fokianos, and Marios Polycarpou. Low rank transformer for multivariate time series anomaly detection and localization. In *ICLR*, 2026.

Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouet. Anomaly detection in streams with extreme value theory. In *SIGKDD*, 2017.

John Sipple. Interpretable, multidimensional, multimodal anomaly detection with negative sampling for detection of device failure. In *ICML*, 2020.

Junho Song, Keonwoo Kim, Jeonglyul Oh, and Sungzoon Cho. Memto: memory-guided transformer for multivariate time series anomaly detection. In *NeurIPS*, 2023.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *SIGKDD*, 2019.

Xin Sun, Heng Zhou, and Chao Li. Multivariate time series anomaly detection with idempotent reconstruction. In *NeurIPS*, 2025.

Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. Precision and recall for time series. In *NeurIPS*, 2018.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. Tranad: deep transformer networks for anomaly detection in multivariate time series data. *Proc. VLDB Endow.*, 2022.

Chengsen Wang, Zirui Zhuang, Qi Qi, Jingyu Wang, Xingyu Wang, Haifeng Sun, and Jianxin Liao. Drift doesn't matter: Dynamic decomposition with diffusion reconstruction for unstable multivariate time series anomaly detection. In *NeurIPS*, 2023.

Rui Wang, Chongwei Liu, Xudong Mou, Kai Gao, Xiaohui Guo, Pin Liu, Tianyu Wo, and Xudong Liu. Deep contrastive one-class time series anomaly detection. *arXiv preprint arXiv:2207.01472*, 2022.

Yining Wang, Fujun Han, Ke Li, Shuhan Liu, Tao Chen, and Zhaohui Peng. Damr: Dual adaptive multi-head representation learning for multivariate time series anomaly detection. In *SIGKDD*, 2026.

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

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. CATCH: Channel-aware multivariate time series anomaly detection via frequency patching. In *ICLR*, 2025.

Shiwang Xing, Jianwei Niu, and Tao Ren. Doknowad: Calibrating normal representations with refined domain knowledge to enhance time series anomaly detection. In *AAAI*, 2026.

Hongzuo Xu, Yijie Wang, Songlei Jian, Qing Liao, Yongjun Wang, and Guansong Pang. Calibrated one-class classification for unsupervised time series anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 2024a.

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

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with $10k$ parameters. In *ICLR*, 2024b.

Takehisa Yairi, Yoshikiyo Kato, and Koichi Hori. Fault detection by mining association rules from house-keeping data. In *Proceedings of the 6th International Symposium on Artificial Intelligence, Robotics and Automation in Space*, 2001.

Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. Dcdetector: Dual attention contrastive representation learning for time series anomaly detection. In *SIGKDD*, 2023.

Ruyi Zhang, Hongzuo Xu, Songlei Jian, Yusong Tan, Haifang Zhou, and Rulin Xu. Angel or devil: Discriminating hard samples and anomaly contaminations for unsupervised time series anomaly detection. *Neural Networks*, 2026.

Shuai Zhang, Chuan Zhou, Yang Liu, Peng Zhang, Xixun Lin, and Shirui Pan. Conformal anomaly detection in event sequences. In *ICML*, 2025.

Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Multivariate time-series anomaly detection via graph attention network. In *ICDM*, 2020.

Guoxiang Zhong, Fagui Liu, Jun Jiang, Bin Wang, and C.L. Philip Chen. Refining one-class representation: A unified transformer for unsupervised time-series anomaly detection. *Information Sciences*, 2024.

Quan Zhou, Changhua Pei, Fei Sun, Han Jing, Zhengwei Gao, Haiming Zhang, Gaogang Xie, Dan Pei, and Jianhui Li. KAN-AD: Time series anomaly detection with kolmogorov–arnold networks. In *ICML*, 2025.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One Fits All: Power general time series analysis by pretrained LM. In *NeurIPS*, 2023.

Xiaohui Zhou, Yijie Wang, Hongzuo Xu, and Yizhou Li. Modeling heterogeneous normality in time series anomaly detection. *Information Processing & Management*, 2026.

A Training and Inference Procedures

Algorithm 1 summarizes the training and calibration procedure on the nominally normal training prefix. Algorithm 2 computes calibrated anomaly scores for each test time point.

Algorithm 1: Training and Calibration of RestoreAD

Input: Training windows $\mathcal{D}_{\mathrm{tr}} = \{\mathbf{W}_i\}_{i=1}^{N}$ and scored times $\mathcal{T}_{\mathrm{tr}}$; reference size M, retrieval size K, maximum bin count B, and minimum bin size $m_{\min}$.

Output: Encoder f_θ; reference $\mathcal{M}$ and difficulties $\{h_j^{(c)}\}_{j,c}$; bins $\{\mathcal{B}_b\}_b$ and medians $\{\mu_b\}_b$, μ_{global}.

Initialize the shared channel-wise encoder g_θ
for *each training iteration* **do**
  Sample stochastic views and anchor/candidate windows from $\mathcal{D}_{\mathrm{tr}}$
  Compute $\mathcal{L}_{\mathrm{rep}}$ and $\mathcal{L}_{\mathrm{RU}}$ using valid relations
  Update θ to minimize $\mathcal{L}_{\mathrm{rep}} + \lambda_{\mathrm{RU}}\mathcal{L}_{\mathrm{RU}}$
end
Encode training windows: $\mathbf{Z}_i \leftarrow f_\theta(\mathbf{W}_i)$ for $i = 1, \ldots, N$
Construct $\mathcal{M}$ by clustering window descriptors and selecting the nearest training window per center
Compute $\{h_j^{(c)}\}_{j,c}$ using Eq. 5
for *each* $\mathbf{W}_i \in \mathcal{D}_{\mathrm{tr}}$ **do**
  Obtain $(\bar{s}_i, \bar{h}_i)$ from the same channel-wise top-K references using Eq. 6
end
for *each* $t \in \mathcal{T}_{\mathrm{tr}}$ **do**
  $(s_t, h_t) \leftarrow |\mathcal{I}_t|^{-1} \sum_{i \in \mathcal{I}_t} (\bar{s}_i, \bar{h}_i)$
end
Form up to B quantile bins $\{\mathcal{B}_b\}_b$ from $\{h_t\}_{t \in \mathcal{T}_{\mathrm{tr}}}$, with at least $m_{\min}$ points per bin
for *each support bin* $\mathcal{B}_b$ **do**
  $\mu_b \leftarrow \mathrm{median}\{s_t : t \in \mathcal{T}_{\mathrm{tr}},\ h_t \in \mathcal{B}_b\}$
end
$\mu_{\mathrm{global}} \leftarrow \mathrm{median}\{s_t : t \in \mathcal{T}_{\mathrm{tr}}\}$
return f_θ, $\mathcal{M}$, $\{h_j^{(c)}\}_{j,c}$, $\{(\mathcal{B}_b, \mu_b)\}_b$, μ_{global}

Algorithm 2: Inference with RestoreAD

Input: Scoring windows $\mathcal{D}_{\mathrm{te}}$ covering test times $\mathcal{T}_{\mathrm{te}}$; retrieval size K; outputs of Algorithm 1.

Output: Calibrated scores $\{s_t^{\mathrm{cal}}\}_{t \in \mathcal{T}_{\mathrm{te}}}$.

for *each* $\mathbf{W}_i \in \mathcal{D}_{\mathrm{te}}$ **do**
  $\mathbf{Z}_i \leftarrow f_\theta(\mathbf{W}_i)$
  Obtain $(\bar{s}_i, \bar{h}_i)$ from the same channel-wise top-K references using Eq. 6
end
for *each* $t \in \mathcal{T}_{\mathrm{te}}$ **do**
  $(s_t, h_t) \leftarrow |\mathcal{I}_t|^{-1} \sum_{i \in \mathcal{I}_t} (\bar{s}_i, \bar{h}_i)$
  Find b_t such that $h_t \in \mathcal{B}_{b_t}$
  $s_t^{\mathrm{cal}} \leftarrow s_t - \mu_{b_t} + \mu_{\mathrm{global}}$
end
return $\{s_t^{\mathrm{cal}}\}_{t \in \mathcal{T}_{\mathrm{te}}}$

B Proofs and Additional Theoretical Analysis

This appendix derives the properties in Section 3.6 and analyzes finite-sample behavior and complexity. We mark population calibration quantities with $\star$ and empirical estimates with hats.

B.1 Row-Uniform Identity and Local Expansion

For a fixed anchor i and channel c, let $m = |\mathcal{C}_i|$ and let $\mathbf{r} = (r_1, \dots, r_m)$ denote its relation logits. Set $\mathbf{p} = \mathrm{softmax}(\mathbf{r})$, $\mathbf{u} = (1/m, \dots, 1/m)$, and $\bar{r} = m^{-1} \sum_{j=1}^{m} r_j$. Then

$$\begin{aligned} D_{\mathrm{KL}}(\mathbf{u} \| \mathbf{p}) &= \frac{1}{m} \sum_{j=1}^{m} \log \frac{u_j}{p_j} \\ &= \log \left(\frac{1}{m} \sum_{j=1}^{m} e^{r_j} \right) - \bar{r} \geq 0. \end{aligned} \tag{10}$$

The identity follows by substituting $p_j = e^{r_j} / \sum_k e^{r_k}$ into the KL expression. Jensen's inequality gives nonnegativity, with equality if and only if $r_1 = \cdots = r_m$.

Shift invariance and local quadratic approximation. A common shift $r_j \mapsto r_j + a$ adds a to both the log-mean-exp term and $\bar{r}$ in Eq. 10, leaving their difference unchanged. Let $\widetilde{r}_j = r_j - \bar{r}$ and $\|\widetilde{\mathbf{r}}\|_\infty = \max_j |\widetilde{r}_j|$. Since $\sum_j \widetilde{r}_j = 0$, Taylor expansion as $\|\widetilde{\mathbf{r}}\|_\infty \to 0$ yields

$$\begin{aligned} D_{\mathrm{KL}}(\mathbf{u} \| \mathbf{p}) &= \log \left(\frac{1}{m} \sum_{j=1}^{m} e^{\widetilde{r}_j} \right) \\ &= \frac{1}{2m} \sum_{j=1}^{m} \widetilde{r}_j^2 + \mathcal{O}\big(\|\widetilde{\mathbf{r}}\|_\infty^3\big) . \end{aligned} \tag{11}$$

Thus, the leading-order term is half the uniform-weighted variance of the relation logits.

B.2 Exact Finite-Sample Identities of CSNC

Let $\mathcal{T}_{\mathrm{tr}}$ index scored time points in the training prefix. For a nonempty support bin $\mathcal{B}_b$, define $\mathcal{T}_b = \{t \in \mathcal{T}_{\mathrm{tr}} : h_t \in \mathcal{B}_b\}$ and $\mu_b = \mathrm{median}\{s_t : t \in \mathcal{T}_b\}$. CSNC applies the same shift $c_b = \mu_{\mathrm{global}} - \mu_b$ to every score in this bin. The resulting scores satisfy

$$\begin{aligned} \mathrm{median}\{s_t^{\mathrm{cal}} : t \in \mathcal{T}_b\} &= \mu_b + c_b = \mu_{\mathrm{global}}, \\ s_t^{\mathrm{cal}} - s_{t'}^{\mathrm{cal}} &= s_t - s_{t'}, \qquad t, t' \in \mathcal{T}_b. \end{aligned} \tag{12}$$

The first identity follows from translation equivariance of the median and establishes exact alignment of training-bin score locations. The second follows because the shared shift cancels; it also holds for any two test observations assigned to the same bin. Thus, CSNC preserves all within-bin pairwise score differences and rankings at both training and inference.

B.3 Proof of Proposition 1

Proof. Let $m_Z = \mathrm{median}(Z \mid Y = 0)$, and let F_Z denote the corresponding CDF. For a nonempty support bin $\mathcal{B}_b$, set $\mathcal{H}_b = \mathcal{B}_b \cap \mathcal{H}_0$ and define $\Delta_b^- = \inf_{h \in \mathcal{H}_b} \Delta(h)$ and $\Delta_b^+ = \sup_{h \in \mathcal{H}_b} \Delta(h)$. The within-bin oscillation $\omega_b = \Delta_b^+ - \Delta_b^-$ satisfies $\omega_b \leq L_\Delta w_b$.

Conditional on $Y = 0$ and $H \in \mathcal{B}_b$, we have $Z + \Delta_b^- \leq S_{\mathrm{raw}} \leq Z + \Delta_b^+$ almost surely. Since $Z \perp H \mid Y = 0$, monotonicity and translation equivariance of the median give

$$m_Z + \Delta_b^- \leq \mu_b^\star \leq m_Z + \Delta_b^+.$$

Define the residual shift $\eta_b(h) = \Delta(h) + m_Z - \mu_b^\star$. For every $h \in \mathcal{H}_b$, the median bound yields

$$|\eta_b(h)| \leq \omega_b \leq L_\Delta w_b. \tag{13}$$

For $h \in \mathcal{H}_b$, population CSNC gives $S_{\mathrm{cal}}^\star = Z - m_Z + \mu_{\mathrm{global}}^\star + \eta_b(h)$. Setting $a = \tau - \mu_{\mathrm{global}}^\star + m_Z$, we obtain $\mathrm{FPR}^{\mathrm{cal}}(\tau \mid h) = 1 - F_Z(a - \eta_b(h))$. Because $\|f_Z\|_\infty \leq \bar{f}$, the CDF F_Z is $\bar{f}$-Lipschitz.

Thus, for $h \in \mathcal{H}_b$ and $h' \in \mathcal{H}_{b'}$,

$$\begin{aligned}\left|\mathrm{FPR}^{\mathrm{cal}}(\tau \mid h) - \mathrm{FPR}^{\mathrm{cal}}(\tau \mid h')\right| &\le \bar{f}\,|\eta_b(h) - \eta_{b'}(h')| \\ &\le \bar{f}\,(\omega_b + \omega_{b'}) \\ &\le \bar{f} L_\Delta (w_b + w_{b'}).\end{aligned} \tag{14}$$

□

B.4 Exact Comparability under Binwise Location Shifts

Corollary 1 (Exact comparability under binwise location shifts)**.** *Under the normal-score model $S_{\mathrm{raw}} = Z + \Delta(H)$ with $Z \perp H \mid Y = 0$, suppose $\Delta(h) = \delta_b$ for all $h \in \mathcal{B}_b \cap \mathcal{H}_0$ in every bin with positive normal probability. Then population CSNC yields the same calibrated normal-score distribution across bins, giving equal FPRs at any shared threshold.*

Proof. Fix a nonempty support bin $\mathcal{B}_b$. Since $\Delta(h) = \delta_b$ on $\mathcal{B}_b \cap \mathcal{H}_0$, we have $S_{\mathrm{raw}} = Z + \delta_b$ conditional on $Y = 0$ and $H \in \mathcal{B}_b$. Let $m_Z = \mathrm{median}(Z \mid Y = 0)$. Translation equivariance gives $\mu_b^\star = m_Z + \delta_b$, so population CSNC cancels the bin-specific shift:

$$\begin{aligned}S_{\mathrm{cal}}^\star &= S_{\mathrm{raw}} - \mu_b^\star + \mu_{\mathrm{global}}^\star \\ &= Z - m_Z + \mu_{\mathrm{global}}^\star.\end{aligned}$$

Because $Z \perp H \mid Y = 0$, this score has the same conditional distribution in every support bin. Hence, $\Pr(S_{\mathrm{cal}}^\star > \tau \mid Y = 0, H \in \mathcal{B}_b)$ is independent of b for any shared threshold τ. □

B.5 Finite-Sample Analysis of CSNC

Proposition 1 uses population bin medians. To isolate median-estimation error, we consider a fixed support partition with at most B nonempty bins and independent normal calibration scores within each bin. Let $\widehat{\mu}_b$ be the empirical median in bin $\mathcal{B}_b$ and set $\varepsilon_b = \widehat{\mu}_b - \mu_b^\star$. For a fresh normal test observation, let $\widehat{\mathrm{FPR}}^{\mathrm{cal}}(\tau \mid h)$ denote its false-positive probability conditional on the fitted calibration statistics. For $h \in \mathcal{B}_b \cap \mathcal{H}_0$ and $h' \in \mathcal{B}_{b'} \cap \mathcal{H}_0$, the argument of Proposition 1 gives

$$\begin{aligned}\left|\widehat{\mathrm{FPR}}^{\mathrm{cal}}(\tau \mid h) - \widehat{\mathrm{FPR}}^{\mathrm{cal}}(\tau \mid h')\right| &\le \bar{f} L_\Delta (w_b + w_{b'}) \\ &\quad + \bar{f}\big(|\varepsilon_b| + |\varepsilon_{b'}|\big).\end{aligned} \tag{15}$$

The empirical global median adds a common shift that cancels in this bound.

Theorem 1 (Finite-sample cross-support FPR bound)**.** *Assume the conditions of Proposition 1 and fix the support partition. For each nonempty bin $\mathcal{B}_b$, let F_b and f_b be the CDF and density of $S_{\mathrm{raw}} \mid (Y = 0, H \in \mathcal{B}_b)$. Suppose the bin contains n_b independent calibration scores drawn from F_b, and the test score is independent of the calibration data. Assume $f_b(s) \ge \underline{f} > 0$ whenever $|s - \mu_b^\star| \le r$. Let $n_{\mathrm{obs}} = \min_{b:n_b>0} n_b$ and $\epsilon_\alpha = \sqrt{\log(2B/\alpha)/(2n_{\mathrm{obs}})}$ for $0 < \alpha < 1$. If $\epsilon_\alpha \le \underline{f} r$, then, with probability at least $1 - \alpha$ over the calibration data, the following holds for every pair of nonempty bins, all $h \in \mathcal{B}_b \cap \mathcal{H}_0$ and $h' \in \mathcal{B}_{b'} \cap \mathcal{H}_0$, and every threshold τ:*

$$\left|\widehat{\mathrm{FPR}}^{\mathrm{cal}}(\tau \mid h) - \widehat{\mathrm{FPR}}^{\mathrm{cal}}(\tau \mid h')\right| \le \bar{f} L_\Delta (w_b + w_{b'}) + \frac{2\bar{f}}{\underline{f}}\epsilon_\alpha. \tag{16}$$

Proof. The Dvoretzky–Kiefer–Wolfowitz inequality (Massart, 1990) and a union bound over at most B bins imply that, with probability at least $1 - \alpha$, all empirical bin CDFs satisfy

$$\max_{b:n_b>0} \sup_s |\widehat{F}_b(s) - F_b(s)| \le \epsilon_\alpha.$$

On this event, the local density lower bound and $\epsilon_\alpha \le \underline{f} r$ give $|\widehat{\mu}_b - \mu_b^\star| \le \epsilon_\alpha / \underline{f}$ simultaneously for every nonempty bin. Inserting this estimate into Eq. 15 proves the theorem. □

Approximation–estimation tradeoff. The first term in Eq. 16 reflects residual variation of $\Delta(H)$ within the support bins; the second arises from estimating their medians. Finer bins can reduce the former but leave fewer calibration scores in each bin. For approximately balanced bins with $n_{\text{obs}} \approx N_{\text{cal}}/B$, the finite-sample estimation term scales with the bin count as

$$\mathcal{O}\left(\frac{\bar{f}}{\underline{f}}\sqrt{\frac{B\log(2B/\alpha)}{N_{\text{cal}}}}\right),$$

where N_{cal} is the total number of normal calibration scores. This tradeoff motivates a moderate number of bins and a minimum-bin-size constraint to stabilize empirical median estimates.

B.6 Additional Complexity Analysis

We detail training, one-time preparation, inference, and storage costs beyond the encoder. Let N be the number of training windows, M the number of stored references, A the number of sampled anchors, and R the number of candidates per anchor.

Training relations. For C channels and embedding dimension d, computing sampled relation logits costs $\mathcal{O}(ARCd)$ per step. Once computed, Row-Uniform adds $\mathcal{O}(ARC)$ work.

Offline preparation. Summarizing the channel-resolved training representations as $2d$-dimensional window descriptors costs $\mathcal{O}(NCd)$. Clustering these descriptors into M groups costs $\mathcal{O}(INMd)$ for I full assignment passes; with mini-batches, the cost instead scales with the total number of descriptors processed. Selecting the observed training window nearest each center costs $\mathcal{O}(NMd)$, and computing channel-wise reference isolation costs $\mathcal{O}(M^2Cd)$. Training-window scoring uses the same retrieval procedure as inference. Given N_{cal} point-level support–score pairs, fitting quantile bins and their score medians costs at most $\mathcal{O}(N_{\text{cal}}\log N_{\text{cal}})$.

Inference and storage. Computing similarities to all references costs $\mathcal{O}(CMd)$ per query window, followed by top-K selection (e.g., $\mathcal{O}(CM\log K)$ with a heap). The entries yield both the raw score and support difficulty; computing the latter adds $\mathcal{O}(CK)$ aggregation. CSNC requires $\mathcal{O}(\log B)$ bin lookup and $\mathcal{O}(1)$ correction per time point. Excluding encoder parameters, the reference embeddings, isolation values, and calibration statistics occupy $\mathcal{O}(MCd + MC + B)$ space.

C Experimental Details

C.1 Dataset Details

TSB-AD-M. The multivariate track of TSB-AD (Liu & Paparrizos, 2024) comprises 200 series from 17 source datasets. The benchmark assigns 20 series to *Tuning* and 180 to *Eval* and specifies a train–test boundary for each series. Table 3 reports split sizes, average lengths, and anomaly ratios.

Table 3: Statistics of TSB-AD-M.

Split	# Series	Avg. Length	Anom. Ratio
Tuning	20	98,164.1	5.7%
Eval	180	108,826.7	5.0%
All	200	107,760.4	5.1%

We select hyperparameters on *Tuning* and evaluate the fixed configuration on *Eval*. Model fitting and score calibration use only the preceding segment of each series, treated as nominally normal; the subsequent segment is used for testing. Anomaly labels are excluded from representation learning, reference construction, support estimation, and CSNC calibration. On *Eval*, labels are used for performance metrics and post hoc analyses of held-out normal observations.

C.2 Baseline Details

Compared methods. We compare RestoreAD with 22 baselines spanning statistical and machine-learning methods, deep anomaly detectors, and recent time-series models. Statistical and machine-learning methods include PCA (Aggarwal, 2017), KMeansAD (Yairi et al., 2001), IForest (Liu

Table 4: Full performance comparison on the TSB-AD-M *Eval* split. Results are mean ± standard deviation over three runs; best and second-best means are shown in **bold** and underline, respectively.

	Method	Range-wise			Point-wise		
		VUS-PR↑	VUS-ROC↑	Range-F1↑	AUC-PR↑	AUC-ROC↑	Point-F1↑
Stat/ML	PCA (2017)	0.2691 ± .0000	0.7058 ± .0000	0.3277 ± .0000	0.2409 ± .0000	0.6759 ± .0000	0.3120 ± .0000
	KMeansAD (2001)	0.2881 ± .0023	0.7298 ± .0004	0.3298 ± .0014	0.2477 ± .0024	0.6918 ± .0001	0.3064 ± .0017
	IForest (ICDM'08)	0.2487 ± .0000	0.7437 ± .0000	0.2434 ± .0000	0.2095 ± .0000	0.7040 ± .0000	0.2798 ± .0000
	COPOD (ICDM'20)	0.2015 ± .0000	0.6820 ± .0000	0.2422 ± .0000	0.2046 ± .0000	0.6518 ± .0000	0.2688 ± .0000
	HBOS (KI'12)	0.1856 ± .0000	0.6691 ± .0000	0.2438 ± .0000	0.1610 ± .0000	0.6331 ± .0000	0.2439 ± .0000
Deep AD	AutoEncoder (MLSDA'14)	0.3056 ± .0016	0.7161 ± .0019	0.3359 ± .0040	0.2823 ± .0024	0.6712 ± .0027	0.3414 ± .0028
	LSTMAD (ESANN'15)	0.3311 ± .0006	0.7709 ± .0020	0.2938 ± .0017	0.3327 ± .0004	0.7360 ± .0021	0.3953 ± .0023
	OmniAnomaly (KDD'19)	0.3153 ± .0000	0.7434 ± .0005	0.2929 ± .0004	0.3152 ± .0000	0.7056 ± .0009	0.3752 ± .0001
	USAD (KDD'20)	0.4031 ± .0010	0.7674 ± .0020	0.4236 ± .0044	0.3621 ± .0004	0.7384 ± .0018	0.4333 ± .0006
	DeepAnT (IEEE Access'19)	0.3555 ± .0007	0.8090 ± .0003	0.3016 ± .0012	0.3529 ± .0009	0.7762 ± .0014	0.4155 ± .0003
	DAMR (KDD'26)	0.3939 ± .0058	0.7707 ± .0092	0.4103 ± .0074	0.3671 ± .0075	0.7399 ± .0097	0.4159 ± .0068
	STREAM-VAE (IEEE IV'26)	0.4293 ± .0029	0.7710 ± .0008	0.4191 ± .0061	0.3882 ± .0023	0.7548 ± .0011	0.4547 ± .0022
Modern TS	TranAD (PVLDB'22)	0.3178 ± .0004	0.7452 ± .0007	0.2827 ± .0005	0.3179 ± .0004	0.7089 ± .0010	0.3831 ± .0009
	Anomaly Transformer (ICLR'22)	0.1190 ± .0037	0.5136 ± .0032	0.1583 ± .0065	0.0768 ± .0023	0.4756 ± .0047	0.1218 ± .0008
	TimesNet (ICLR'23)	0.3960 ± .0108	0.7091 ± .0010	0.4063 ± .0043	0.3435 ± .0093	0.6680 ± .0014	0.4142 ± .0074
	FITS (ICLR'24)	0.3633 ± .0011	0.7000 ± .0010	0.3248 ± .0040	0.3045 ± .0014	0.6514 ± .0012	0.3689 ± .0005
	PatchTST (ICLR'23)	0.3898 ± .0010	0.7167 ± .0004	0.3816 ± .0027	0.3354 ± .0006	0.6748 ± .0008	0.4066 ± .0007
	OFA (NeurIPS'23)	0.3524 ± .0032	0.6859 ± .0011	0.3588 ± .0068	0.2990 ± .0030	0.6370 ± .0017	0.3669 ± .0023
	CATCH (ICLR'25)	0.3774 ± .0007	0.7320 ± .0004	0.2618 ± .0033	0.3566 ± .0009	0.6937 ± .0005	0.3935 ± .0007
	KAN-AD (ICML'25)	0.2780 ± .0013	0.6842 ± .0009	0.2369 ± .0031	0.2413 ± .0020	0.6392 ± .0009	0.2938 ± .0014
	Time-RCD (ICML'26)	0.2063 ± .0000	0.6640 ± .0000	0.2023 ± .0000	0.1616 ± .0000	0.6171 ± .0000	0.2270 ± .0000
Rep.	PaAno (ICLR'26)	0.4274 ± .0013	0.7947 ± .0007	0.4074 ± .0059	0.3784 ± .0037	0.7633 ± .0009	0.4292 ± .0006
	RestoreAD (Ours)	**0.4662** ± .0047	**0.8167** ± .0050	**0.4627** ± .0068	**0.4260** ± .0024	**0.7897** ± .0047	**0.4655** ± .0028

et al., 2008), COPOD (Li et al., 2020), and HBOS (Goldstein & Dengel, 2012). Deep anomaly detectors include AutoEncoder (Sakurada & Yairi, 2014), LSTMAD (Malhotra et al., 2015), OmniAnomaly (Su et al., 2019), USAD (Audibert et al., 2020), DeepAnT (Munir et al., 2019), DAMR (Wang et al., 2026), and STREAM-VAE (Özer et al., 2026). Recent time-series models include TranAD (Tuli et al., 2022), Anomaly Transformer (Xu et al., 2022), TimesNet (Wu et al., 2023), FITS (Xu et al., 2024b), PatchTST (Nie et al., 2023), OFA (Zhou et al., 2023), CATCH (Wu et al., 2025), KAN-AD (Zhou et al., 2025), and Time-RCD (Lan et al., 2026). We also include the representation-based PaAno (Park & Kang, 2026). Table 1 presents 17 baselines; Table 4 reports the complete comparison across all six evaluation metrics.

C.3 Implementation and Evaluation Details

RestoreAD configuration. We use length-96 windows with stride one and standardize each channel within each window. A shared temporal encoder with a convolutional stem and three gated dilated blocks produces $d = 64$-dimensional channel embeddings. Training uses two weakly augmented views and combines a VICReg-style loss, with invariance, variance, and covariance weights of 25, 25, and 1, respectively, with Row-Uniform regularization of weight 1. For Row-Uniform, we sample up to 64 anchors and 128 candidates, apply the temporal-overlap mask in Section 3.3, and use a logit temperature of 0.1. We train for 100 iterations with AdamW at an initial learning rate of 3×10^{-4}.

To construct the normal reference, we concatenate the mean and element-wise maximum across each training window's channel embeddings into a $2d$-dimensional descriptor. We cluster these descriptors and select the observed training window whose descriptor is nearest to each center. For N training windows, the reference budget is $M = \min\{N - 1, \max\{500, \mathrm{round}(0.1N)\}\}$ slots; different centers may select the same window. Scoring retrieves $K = 3$ references per channel. We average window-level scores and support difficulties over windows covering each time point before point-level calibration. CSNC uses up to $B = 10$ quantile bins of at least $m_{\min} = 64$ training points.

Table 5: Baseline configurations on TSB-AD-M. Listed values are those used in our evaluation; unlisted parameters follow the corresponding released defaults.

Method	Input setting	Key hyperparameters
Statistical & Machine Learning		
PCA	$w = 100$	$n_{\text{comp}} = 0.25$
KMeansAD	$w = 40$	$n_{\text{clusters}} = 10$
IForest	$w = 100$	$n_{\text{est}} = 25$, max features $= 0.8$
COPOD	–	Default
HBOS	$w = 1$	$n_{\text{bins}} = 30$, tolerance $= 0.5$
Deep Anomaly Detection		
AutoEncoder	$w = 100$	hidden dims $= [128, 64]$
DeepAnT	$w = 50$	channels $= [32, 32, 40]$
LSTMAD	$w = 150$	$lr = 8 \times 10^{-4}$
OmniAnomaly	$w = 100$	$lr = 2 \times 10^{-3}$
USAD	$w = 100$	$lr = 10^{-3}$
STREAM-VAE	$w = 100$	latent dim $= 64$, $lr = 10^{-3}$, BS $= 128$, Ep $= 50$
DAMR	$w = 60$	hidden dim $= 16$, conv. hidden dim $= 8$, heads $= 3$, $lr = 10^{-6}$, BS $= 512$, Ep $= 40$, wd $= 10^{-3}$
Modern Time-Series Models		
TranAD	$w = 10$	$lr = 10^{-3}$
AnomalyTransformer	$w = 50$	$lr = 10^{-3}$
TimesNet	$w = 96$	$lr = 10^{-4}$
FITS	$w = 100$	$lr = 10^{-3}$
PatchTST	$w = 100$	$lr = 10^{-4}$, BS $= 128$
OFA	$w = 50$	Default
CATCH	$L = 96$, p/s $= 16/8$	$d = 128$, $h = 2$, $lr = 5 \times 10^{-4}$, BS $= 32$, Ep $= 5$
KAN-AD	$w = 96$	$lr = 10^{-3}$, BS $= 1024$, Ep $= 100$
Time-RCD	$w = 512$	BS $= 1$
Representation-Based		
PaAno	patch $= 96$	mem $= 0.1$, top-$K = 3$, $lr = 10^{-4}$, BS $= 512$, It $= 100$

w: window size; L: sequence length; p/s: patch size/stride; d: model dimension; h: number of heads; lr: learning rate; BS: batch size; Ep: epochs; It: iterations; wd: weight decay; mem: memory ratio.

Implementations and configurations. For baselines included in TSB-AD (Liu & Paparrizos, 2024), we use its released implementations and multivariate settings; for the remaining methods, we use the authors' implementations. The DeepAnT implementation is labeled `CNN` in TSB-AD. Any method-specific configuration selection uses only the official *Tuning* split (Appendix C.1). Table 5 lists the key settings used for evaluation; unlisted settings follow the released defaults.

Evaluation and reporting. We evaluate all baselines on the 180-series TSB-AD-M *Eval* split using the TSB-AD evaluator without point adjustment. For each seed, we average each of the six metrics across series, then report the mean and standard deviation over seeds 2027, 2028, and 2029. Baselines with identical outputs across seeds have zero reported standard deviation.

C.4 Evaluation Metrics

For each evaluation series, we compare point-level anomaly scores with ground-truth labels using the TSB-AD evaluator (Liu & Paparrizos, 2024), without point adjustment. As described in Appendix C.3, we compute each of the six metrics per series before averaging across series and runs.

Point-wise measures. AUC-ROC summarizes the receiver operating characteristic curve across score thresholds. The measure reported as AUC-PR follows the TSB-AD implementation and uses non-interpolated average precision, rather than trapezoidal integration of the precision–recall curve. Point-F1 is the maximum point-wise F1 over the evaluator's precision–recall thresholds.

Range-wise measures. Range-F1 uses range-based precision and recall (Tatbul et al., 2018), accounting for anomaly-range coverage and prediction fragmentation. We report its maximum over 100 evenly spaced score thresholds. VUS-ROC and VUS-PR (Paparrizos et al., 2022; Boniol et al., 2025) evaluate performance across score thresholds and temporal tolerances. Under the TSB-AD autocorrelation rule, all methods share a maximum tolerance determined from the first channel of each series when computing both VUS-ROC and VUS-PR.

The reported Point-F1 and Range-F1 use evaluation labels to select their threshold maxima. These thresholds are used only for evaluation, not for model fitting or score calibration.

C.5 Controlled Support-Imbalance Protocol

This study isolates the effect of training support on reference-based scoring and calibration. For each of ten seeds, we generate six-channel normal windows from fixed channel-specific sinusoidal templates with bounded phase perturbations and stochastic amplitude, drift, and noise variation. Each sample in regime A has an exact sign-reflected counterpart in regime B, so the two regimes have symmetric within-regime geometry. A normalizer fitted to the paired training pool and a fixed 64-dimensional orthonormal projection are shared across support ratios within each seed. The projection and normalization preserve the sign reflection. We do not retrain the RestoreAD encoder; this experiment probes the reference, scoring, and calibration mechanisms rather than Row-Uniform.

We hold the total training size at $2{,}000$ windows and the balanced test set at 600 windows per regime. From a fixed paired training pool, we vary the dense-to-rare training ratio over $\{1{:}1, 2{:}1, 5{:}1, 10{:}1, 20{:}1\}$ and evaluate both assignments of the rare regime. For each ratio and assignment, we rebuild a 48-center reference, recompute prototype-isolation-based support difficulty, and refit the training-score calibration using up to ten support bins. Raw scores use the three nearest centers. Raw and CSNC Q95 thresholds are fitted separately from the balanced-condition training scores and then held fixed across ratios. We report the rare-minus-dense change relative to the corresponding balanced condition. The two rare-regime assignments are averaged within each seed; 95% confidence intervals use $5{,}000$ bootstrap resamples of the ten seed-level values.

D Additional Experimental Results

D.1 CSNC Design Ablations

We examine whether CSNC benefits from the estimated support assignments and whether its performance depends on the binning rule. All variants use the same Base+RU raw scores and differ only in calibration (Table 6). Shuffling the support assignments yields scores close to those without calibration on five metrics, although it increases Range-F1. Relative to shuffled support, CSNC improves the mean of the other five metrics, and their 95% paired-bootstrap confidence intervals lie above zero (Table 7). Range-F1 instead favors the shuffled control. Quantile and equal-width bins yield similar mean performance across the six metrics. These results support the role of informative support assignments while showing little difference between the two binning rules tested here.

Statistical comparison. We compute 95% paired-bootstrap confidence intervals using $5{,}000$ resamples. Each resample draws from the 180 evaluation series with replacement, preserving the pairing between CSNC and shuffled support for each selected series. The confidence intervals describe

Table 6: CSNC design ablation on the official TSB-AD-M *Eval* split. All variants use the same Base+RU raw scores. Results are mean ± standard deviation over seeds 2027, 2028, and 2029; the best mean in each column is boldfaced.

Design	VUS-PR↑	VUS-ROC↑	Range-F1↑	AUC-PR↑	AUC-ROC↑	Point-F1↑
No calibration	0.4647 ± .0047	0.8009 ± .0023	0.4590 ± .0053	0.4244 ± .0022	0.7729 ± .0030	0.4627 ± .0028
Shuffled support	0.4642 ± .0045	0.8007 ± .0026	**0.4738** ± .0043	0.4238 ± .0022	0.7722 ± .0031	0.4623 ± .0028
Equal-width bins	0.4640 ± .0046	0.8149 ± .0008	0.4628 ± .0065	0.4238 ± .0027	0.7877 ± .0017	0.4643 ± .0038
CSNC (quantile bins)	**0.4662** ± .0047	**0.8167** ± .0050	0.4627 ± .0068	**0.4260** ± .0024	**0.7897** ± .0047	**0.4655** ± .0028

Table 7: Paired differences between CSNC and shuffled support on the official TSB-AD-M *Eval* split. Differences are CSNC minus shuffled support.

Metric	Mean difference	95% CI
VUS-PR	$+0.00203$	$[+0.00072,\ +0.00358]$
VUS-ROC	$+0.01592$	$[+0.00989,\ +0.02272]$
Range-F1	-0.01116	$[-0.01763,\ -0.00557]$
AUC-PR	$+0.00217$	$[+0.00085,\ +0.00365]$
AUC-ROC	$+0.01750$	$[+0.01104,\ +0.02458]$
Point-F1	$+0.00320$	$[+0.00147,\ +0.00524]$

uncertainty in the mean paired difference across series; the mean $\pm$ standard deviation in Table 6 separately describes variation across the three seeds.

D.2 Cross-Support Diagnostic Details

Figure 3 (a) reports the absolute Spearman correlation between support difficulty and anomaly scores on held-out normal observations. Here, we retain its sign to examine the direction of the association.

For each evaluation series and seed, we compute the signed correlation separately for raw and CSNC scores. A series–seed pair is valid only if both correlations are finite. We average the signed correlations across valid seeds within each series and then compute cross-series statistics. The absolute analysis uses the same validity mask and takes the absolute value *after* this seed averaging. One of the 180 evaluation series has no valid correlations under any seed, leaving 179 valid series; 178 have valid results under all three seeds. Confidence intervals are obtained from $5{,}000$ bootstrap resamples of evaluation series, with raw and CSNC results kept paired.

Table 8 shows a predominantly positive association before calibration. The mean signed correlation is 0.5858 (95% CI: 0.5477–0.6209), and 174 of 179 valid series (97.2%) have a positive seed-averaged correlation. Among the 178 series with three valid seeds, 173 (97.2%) are positive in every seed. Since larger support difficulty indicates weaker estimated normal support, these results show that weakly supported normal observations tend to receive higher raw anomaly scores.

After CSNC, the mean signed correlation is 0.4690 (95% CI: 0.4291–0.5102). The seed-averaged correlation remains positive in 174 of 179 series, but its mean magnitude is lower. Positive correlations persist across all three seeds in 162 of the 178 complete series. Taking absolute values of the per-series seed-averaged correlations reproduces the means of 0.603 and 0.487 shown in Figure 3 (a). Thus, the signed and absolute analyses describe the same series under the same validity rule.

For the FPR diagnostic in Figure 3 (b), we compute the gap separately for each series and seed before averaging it across valid seeds within a series. Raw and CSNC scores have separately fitted thresholds: each is set at the 95th percentile of its corresponding training-score distribution and applied unchanged across all support bins of that series. Cross-series confidence intervals resample the resulting per-series values while preserving the raw–CSNC pairing.

Table 8: Signed Spearman correlation between support difficulty and anomaly scores on the TSB-AD-M *Eval* split. Statistics summarize per-series coefficients averaged over valid seeds. Confidence intervals use $5{,}000$ resamples of evaluation series. Final row: 178 series with three valid seeds.

Statistic	Raw	After CSNC
Mean signed ρ	0.5858	0.4690
95% CI	$[0.5477,\ 0.6209]$	$[0.4291,\ 0.5102]$
Median signed ρ	0.6206	0.5018
Positive seed-averaged ρ	$174/179$ (97.2%)	$174/179$ (97.2%)
Positive in all three seeds	$173/178$ (97.2%)	$162/178$ (91.0%)

D.3 CSNC Hyperparameter Sensitivity

We vary the number of support bins B and the minimum bin size $m_{\min}$ one at a time around the frozen configuration $(B, m_{\min}) = (10, 64)$. Specifically, we test $B \in \{5, 10, 20\}$ with $m_{\min} = 64$ and $m_{\min} \in \{32, 64, 128\}$ with $B = 10$. All configurations use the same Base+RU detector and raw scores and are evaluated over three seeds on the TSB-AD-M *Eval* split.

Table 9 shows limited variation within this tested grid. The range of displayed means across the five settings is at most 0.0012 for each metric other than Range-F1, whose range is 0.0033. Relative to the frozen configuration, no setting differs by more than 0.0009 on the other five metrics or 0.0017 on Range-F1. Across all six reported metrics, these results indicate that performance is not highly sensitive to either hyperparameter within the examined range.

Table 9: CSNC sensitivity on the official TSB-AD-M *Eval* split. Results are mean ± standard deviation over seeds 2027, 2028, and 2029. The dagger marks the frozen default.

B	$m_{\min}$	**VUS-PR↑**	**VUS-ROC↑**	**Range-F1↑**	**AUC-PR↑**	**AUC-ROC↑**	**Point-F1↑**
5	64	0.4655 ± .0045	0.8158 ± .0045	0.4610 ± .0064	0.4254 ± .0024	0.7889 ± .0041	0.4649 ± .0028
10†	64	0.4662 ± .0047	0.8167 ± .0050	0.4627 ± .0068	0.4260 ± .0024	0.7897 ± .0047	0.4655 ± .0028
20	64	0.4667 ± .0046	0.8169 ± .0053	0.4640 ± .0047	0.4264 ± .0022	0.7899 ± .0051	0.4659 ± .0029
10	32	0.4661 ± .0043	0.8167 ± .0049	0.4643 ± .0085	0.4259 ± .0022	0.7898 ± .0045	0.4653 ± .0026
10	128	0.4663 ± .0042	0.8166 ± .0046	0.4611 ± .0032	0.4262 ± .0022	0.7897 ± .0044	0.4656 ± .0026

D.4 Case Studies

Figure 5 shows three evaluation segments illustrating how CSNC changes decisions across support regimes. In panel (a), normal observations in a localized high-difficulty region exceed the raw threshold but fall below the calibrated threshold. The high-difficulty FPR, computed over the corresponding full series, decreases by 7.15 percentage points. Panel (b) shows broader score realignment across a weakly supported normal region, with a full-series high-difficulty FPR reduction of 3.85 points. In panel (c), high-difficulty normal observations are rescued near a true anomaly: the corresponding FPR decreases by 3.46 points, while full-series TPR changes by $+0.30$ points. These selected cases illustrate the decision-level effects underlying the aggregate analyses; the displayed segments alone are not used to calculate the reported changes.

E Additional Related Work

Normality modeling and difficult normal samples. Beyond homogeneous normality assumptions, memory-, prototype-, and reference-based methods preserve representative normal patterns (Song et al., 2023; Shen, 2025; Park & Kang, 2026), while one-class and self-supervised methods learn compact or transformation-sensitive representations (Ruff et al., 2018; Bergman & Hoshen, 2020; Qiu et al., 2021; Carmona et al., 2022; Wang et al., 2022). Robust approaches further address contaminated training data (Xu et al., 2024a; Chen et al., 2024b; Zhong et al., 2024; Ho & Armanfard, 2025; Mou et al., 2026). HNAD (Zhou et al., 2026) decomposes heterogeneous normality into latent subclasses, whereas PLDA (Zhang et al., 2026) distinguishes informative hard normals from anomaly contamination. DoKnowAD (Xing et al., 2026) further expands normality coverage using auxiliary domain knowledge. Yet score shifts from unequal normal coverage remain underexplored.

Score calibration and anomaly decisions. A complementary line of work converts continuous anomaly scores into reliable decisions. Statistical and streaming methods estimate adaptive thresholds from score distributions or tail behavior (Ahmad et al., 2017; Siffer et al., 2017; Hundman et al., 2018). Score normalization and calibrated thresholding address temporal or system heterogeneity (Lv et al., 2023; Alnegheimish et al., 2025), while test-time adaptation tracks evolving normality under distribution shift (Kim et al., 2024; 2026). Conformal methods instead map scores to statistically calibrated decisions (Burnaev & Ishimtsev, 2016; Zhang et al., 2025; Martinez Gil et al., 2026). These studies mainly target global calibration, distribution shift, or marginal error control; score comparability across heterogeneous normal-support regimes remains less studied.

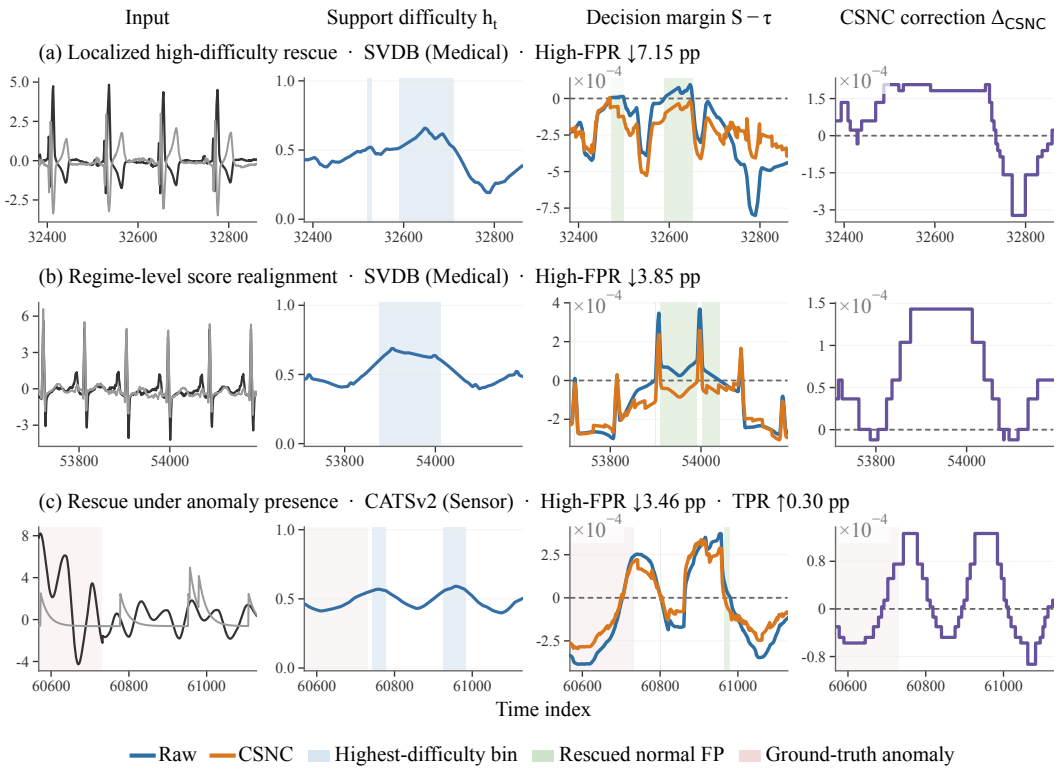


Figure 5: **Case studies of support-conditioned calibration.** Each row shows a test segment from one evaluation series under seed 2029. Columns display input signals, point-level support difficulty h_t, variant-specific decision margins $S^{\text{raw}} - \tau^{\text{raw}}$ and $S^{\text{CSNC}} - \tau^{\text{CSNC}}$, and the score correction $\Delta_{\text{CSNC}} = S^{\text{raw}} - S^{\text{CSNC}}$. Blue, green, and pink shading mark the highest-difficulty bin, rescued normal false positives, and ground-truth anomalies, respectively. Raw and CSNC thresholds are fitted separately at the 95th percentile of their corresponding training-score distributions and shared across support bins within each series. High-FPR denotes the false-positive rate within the highest-difficulty bin. Changes in the row headings are computed over each full evaluation series and averaged over three seeds; the displayed segments are illustrative.

F Limitations and Future Work

Our empirical evaluation follows the TSB-AD-M protocol with a nominally normal training prefix. Future work could examine broader datasets and settings in which normal operating regimes evolve over time. Because CSNC calibrates score locations using an estimated normal reference, its effectiveness also depends on how well that reference covers different normal behaviors. Jointly improving reference construction and support-conditioned calibration is a natural next step.

Our theoretical analysis isolates support-dependent score shifts under simplified distributional and sampling assumptions. Extending it to temporally dependent observations and adaptive support partitions would provide a fuller account of calibration in practical time-series settings.