CLOSURE: CROSS-CHANNEL CLOSURE FOR MULTIVARIATE TIME SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multivariate time series produced by industrial installations and server clusters are under continuous monitoring, and anomaly detection is the first line of defense before a local fault escalates into a system-wide failure. In the absence of labels, the dominant paradigm trains a reconstruction model on normal data and scores the residual to reach a verdict. Recent work has steadily enriched the expressiveness of the recovery map through Transformers, diffusion models, and cross-scale associations. A residual, however, only certifies that an observation can be recovered; it does not certify that the variables still explain one another as they do under normal operation. A coordinated yet unfamiliar operating transition inflates the residual without breaking any relationship, whereas a sensor drifting away from its peers may leave the residual silent. Methods that fold inter-variable dependencies into the reconstruction objective inherit this ambiguity, because their relational evidence is ultimately supervised by the same recovery loss. We therefore propose `CLOSURE`, which evaluates recoverability and cross-channel consistency as two normal-data prediction tasks learned in separation. The recoverability task is solved by a diffusion reconstructor conditioned on a discrete memory of normal temporal bases. The consistency task is solved by a relation-closure operator that predicts every channel exclusively from the others, where the target's own trajectory is withheld and sparse signed gates decide which peers are actually used. The two scores are reconciled only through calibration on normal data. Extensive experiments show that `CLOSURE` consistently outperforms recent reconstruction-, frequency-, and topology-based detectors on public benchmark streams, confirming the value of examining consistency alongside recoverability.

1 INTRODUCTION

Multivariate time series produced by industrial installations and server clusters are under continuous monitoring, and anomaly detection is the first line of defense before a local fault escalates into a system-wide failure (Tuli et al., 2022; Su et al., 2019). Labeled failures are scarce and never enumerate the ways a system can break, so detectors are typically fitted to normal observations alone. Reconstruction supplies the canonical objective for this setting: learn to reproduce normal windows, then read the residual as evidence of departure. Recurrent latent models and adversarial autoencoders established the paradigm (Su et al., 2019; Zhou et al., 2019); Transformers widened the temporal context (Tuli et al., 2022); diffusion models recover observations through denoising and imputation (Zhang et al., 2022; Chen et al., 2023; Xiao et al., 2023; Li et al., 2025b); and cross-scale reconstruction with shared cross-window context further enriches what a model can recover (Li et al., 2025a). These advances strengthen the representation of normal behavior. Whether they strengthen detection depends on a separate question: how faithfully reconstruction difficulty tracks abnormality.

Reconstruction difficulty and abnormality need not coincide. Consider a healthy operating transition in which several sensors change together while preserving their usual coordination. The trajectory may be unfamiliar to the reconstructor and yield a large residual even though no channel relationship has been violated. The converse is equally plausible. A drifting sensor may follow a smooth curve that the model reproduces accurately, while it quietly stops tracking the channels with which it normally varies. Figure 1 contrasts these possibilities. They separate two properties of a single window, *recoverability* and *cross-channel consistency*, and expose an ambiguity in the residual: agreement with a learned reconstructor does not by itself establish agreement with normal relations.

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

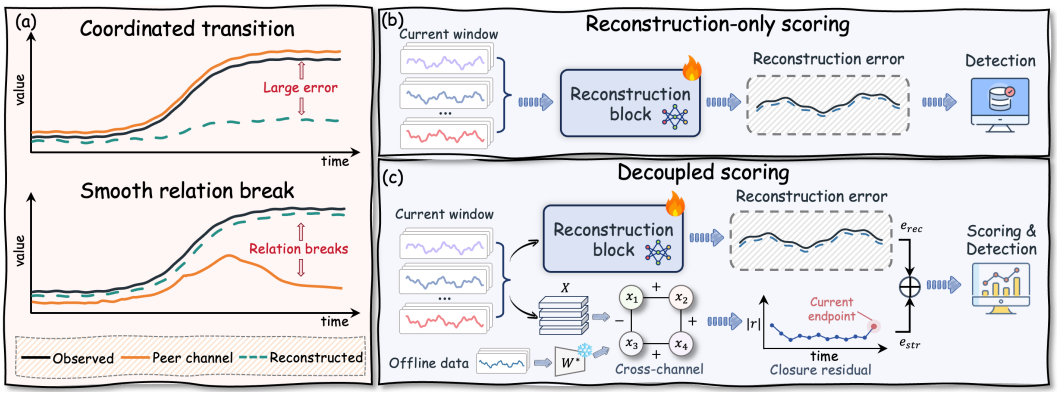

Figure 1: **Reconstructable does not necessarily mean consistent.** (a) A coordinated transition can be hard to reconstruct while every relationship stays intact, whereas a smooth relation break can remain easy to reconstruct. (b) Reconstruction-only scoring compares an observation with its recovery. (c) `CLOSURE` additionally compares the same observation with predictions issued by its peer channels. The depicted $W^\star$ denotes frozen connectivity.

Dependencies already play an important role in anomaly detection. Learned sensor graphs and inter-metric embeddings model interactions among variables (Deng & Hooi, 2021; Li et al., 2021). CATCH learns channel interactions inside frequency patches (Wu et al., 2025), and ScatterAD couples temporal and topological views (Yin et al., 2025). A related line uses temporal association discrepancies as detection evidence (Xu et al., 2021; Yang et al., 2023). These methods establish the value of relational evidence. Our question is therefore not whether dependencies matter, but where their supervision comes from. When inter-variable dependencies are folded into the reconstruction objective, the relational evidence is ultimately shaped by the same recovery loss, and the resulting score inherits the ambiguity above. Such a reference also needs a criterion of its own for deciding which relations enter it. A strong association within one normal interval may reflect a temporary operating condition, and a merely strong correlation does not identify the peers that a nonlinear predictor should actually use. Turning these associations directly into prediction weights therefore grants uncertain relations an influence on detection before their usefulness has been measured. We instead ask how to construct a cross-channel reference whose supervision, and whose selection of relations, are both established by a prediction task separate from reconstruction.

According to these, we introduce `CLOSURE`, a detector that evaluates recoverability and cross-channel consistency as two normal-data prediction tasks learned in separation. The consistency task is carried by a *relation-closure operator*. Each target receives nonlinear messages from its peers through learned signed coefficients and sparse gates, and never from itself; the disagreement between the observed channels and their peer predictions forms a *closure residual* (Batson & Royer, 2019). Because a sustained departure can raise a window's residual background together with its endpoint, we score the endpoint relative to that background, which makes the criterion sensitive to a localized change in coordination. The recoverability task is carried by a diffusion reconstructor conditioned on a discrete memory of normal temporal bases (Li et al., 2025b). The relation operator is neither an input to this reconstructor nor a target of its training loss, and the two paths meet only after their scores are calibrated on normal data. This separation lets us hold the reconstructor fixed and vary the relational criterion alone, so that a change in detection is attributable to relation learning rather than to improved recovery. In summary, we make the following key contributions:

- ❶ **Decoupled scoring.** We formulate a detector that evaluates temporal recoverability alongside cross-channel consistency, keeping the relational reference outside the reconstruction objective and combining the two paths only at the decision stage.
- ❷ **Learned sparse closure.** We instantiate the relational criterion with signed coefficients, per-edge gates, and source-only nonlinear messages, and show that row normalization erases absolute relation strength, motivating a separate gate parameter for selection.
- ❸ **Controlled evaluation.** We prove that a target-only perturbation enters the closure residual unattenuated, and experiments with a frozen reconstructor show that learned relational evidence improves affiliation F1 beyond reconstruction alone.

2 Related Work

Reconstruction-Based Detection. Reconstruction-based detectors learn normal dynamics and score deviations through reconstruction error or likelihood. Recurrent models, adversarial autoencoders, and Transformers provide different temporal representations for this task (Su et al., 2019; Zhou et al., 2019; Tuli et al., 2022). Diffusion approaches formulate recovery as conditional imputation (Chen et al., 2023; Xiao et al., 2023). PhysDiff (Li et al., 2025b) conditions corruption and attention on decomposed temporal bases, while CrossAD (Li et al., 2025a) reconstructs fine-scale observations from coarser views and shares context across windows. These designs enrich what a model can recover the residual. however, they still report agreement with a single recovery map.

Dependency Modeling. Dependencies can guide signal modeling or become detection targets. GDN (Deng & Hooi, 2021) and InterFusion (Li et al., 2021) encode channel interactions through sensor graphs and hierarchical representations, respectively. CATCH (Wu et al., 2025) learns frequency-dependent channel masks, whereas MSCRED (Zhang et al., 2019) reconstructs multiscale inter-sensor signature matrices and scores their residuals. Another line examines temporal associations through association discrepancy and dual-attention contrastive learning (Xu et al., 2021; Audibert et al., 2020; Park et al., 2020; Yang et al., 2023). ScatterAD (Yin et al., 2025) models temporal-node scattering and aligns temporal and topological representations. Across these designs, the relational evidence is produced within the reconstruction or representation objective, so its reference and the recovery map are shaped by the same loss.

Normality Constraints and Scoring. Normal references and constraints on the recovery map address the risk of accurately reconstructing anomalies. MemAE (Gong et al., 2019) retrieves normal latent prototypes for decoding, and vector quantization provides a discrete representation for such prototypes (Van Den Oord et al., 2017). MEMTO (Song et al., 2023) extends memory-guided reconstruction with latent-memory deviations that weight reconstruction errors, demonstrating that a normal reference can also supply an additional scoring criterion. IGAD (Sun et al., 2025) instead imposes idempotency and tightness losses on the reconstruction map. These constraints act on the recovery map or on the representation it decodes from, and their criterion is still read from how well the observation is reproduced.

These methods use learned dynamics, dependencies, or normal prototypes to define detection evidence. `CLOSURE` instead complements reconstruction with a relational reference that is learned separately and never enters the reconstruction objective. Its signed closure operator evaluates whether observed channels remain mutually explainable when each target is withheld from its own prediction, rather than reconstructing a relation matrix or measuring distance to a latent prototype.

3 Methodology

Problem setup. Let $x_t \in \mathbb{R}^D$ contain the retained channels of a multivariate series and $X_t = [x_{t-L+1}, \ldots, x_t]^\top \in \mathbb{R}^{L\times D}$ its length-L window. We seek a score a_t for the *observed endpoint*, using normal data for fitting and calibration. The normal sequence is divided chronologically into training $\mathcal{N}_{\text{tr}}$, validation $\mathcal{N}_{\text{val}}$, and calibration $\mathcal{N}_{\text{cal}}$. Windows stay within their assigned segments, with $L-1$ unused timestamps between segments. All inputs use training-fitted scaling.

Overview. Figure 2 follows one window through two paths that answer different questions. ❶ ***Temporal reconstruction*** measures how accurately a learned model recovers X_t given normal temporal context. ❷ ***Relational closure*** measures whether the channels of X_t remain mutually explainable under a learned normal relation reference. Three components are fitted separately: a discrete basis codec, the reconstructor, and the relation predictor. The relation path receives neither memory codes nor reconstruction outputs, and contributes no gradient to them. Both paths are frozen at detection and meet only through normal-calibrated scoring.

3.1 Memory-Conditioned Temporal Reconstruction

The reconstruction path measures how accurately an entire window can be recovered from normal temporal context, and reports the resulting discrepancy at the observed endpoint. We introduce the basis-conditioned diffusion backbone and represent the normal bases with learned discrete codecs, so that recovery is conditioned on stored normal patterns rather than on the test window itself.

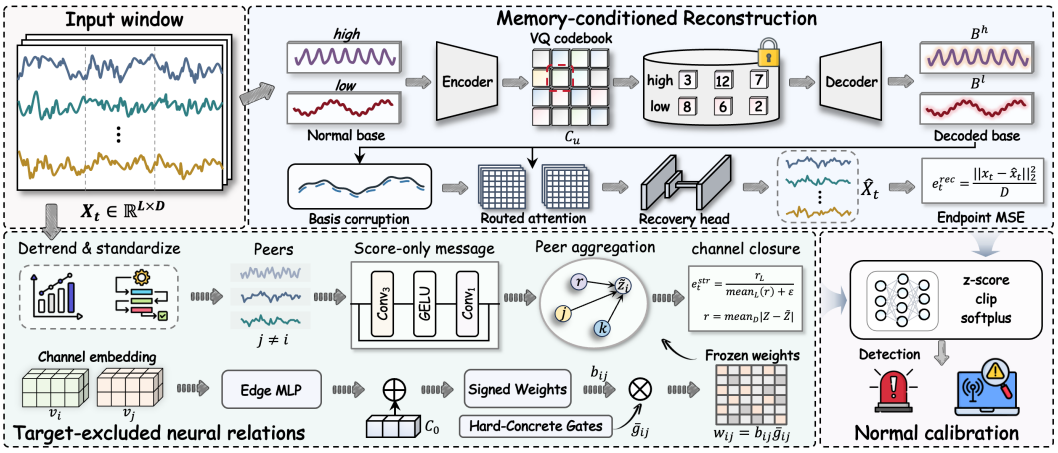

Figure 2: **Overview of CLOSURE.** Paired codebooks store normal temporal bases as code indices; decoded bases condition both corruption and attention. Learned signed coefficients, per-edge gates, and source-only messages predict each target from its peers, never from itself. Independently calibrated scores are combined for detection.

A discrete normal basis memory. High- and low-frequency bases are extracted from normal training windows by adaptive multicomponent decomposition. For band $u \in \{h, l\}$, channel-wise bases are split into length-p patches, normalized by training-fitted scales, and quantized against a codebook $\mathcal{C}_u = \{c_k^u\}_{k=1}^K$ through a straight-through estimator (Van Den Oord et al., 2017):

$$z^u = E_u(b), \quad q_u(b) = \arg\min_{1\le k\le K} \|z^u - c_k^u\|_2^2, \quad \widehat{b} = D_u\big(z^u + \mathrm{sg}[c^u_{q_u(b)} - z^u]\big). \tag{1}$$

Each band owns a codec shared across channels. A memory slot stores the paired code-index sequences of one normal template; reading a slot decodes both bands and restores channel scales to give $\mathcal{B} = (B^h, B^l)$ with $B^u \in \mathbb{R}^{L\times D}$. The shared slot index preserves band pairing. Slots are *sampled* from stored normal context, never retrieved by matching a test query, and no write occurs during detection.

Basis-conditioned recovery. At a fixed diffusion index s with $\rho_s = \sqrt{1-\bar{\alpha}_s}$, the decoded bases enter both corruption and attention. Writing $Q, K, V \in \mathbb{R}^{L\times d}$ for token projections and $U^u \in \mathbb{R}^{L\times d}$ for basis keys,

$$X_t^{(s)} = \sqrt{\bar{\alpha}_s}\,(X_t + \Delta_n) + \rho_s\big[\eta\varepsilon + (1-\eta)(\lambda_s^h B^h + \lambda_s^l B^l)\big],$$
$$A = \mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d}} + \sum_{u\in\{h,l\}}\big[G^u \odot S^u + \widetilde{G}^u \odot \widetilde{S}^u\big]\Big), \quad S^u = \frac{Q(U^u)^\top}{\sqrt{d}}, \quad \widetilde{S}^u = \frac{\widetilde{Q}(\widetilde{U}^u)^\top}{\sqrt{d}}, \tag{2}$$

where $\varepsilon \sim \mathcal{N}(0, I)$, λ_s^u weights each band, and Δ_n is a training-only channel offset repeated across time. Corruption uses the full multichannel bases; attention uses their channel-averaged projections. The sigmoid gates $G^u, \widetilde{G}^u$ modulate basis affinities and are defined in **??**. Since attention aggregates AV, the values still come from the observation. Two attention–feed-forward blocks and a token-wise MLP yield the recovery and its endpoint score:

$$\widehat{X}_t = g_{\mathrm{rec}}(H_t) - \Delta_n, \quad e_t^{\mathrm{rec}} = \frac{1}{D}\,\|x_t - \widehat{x}_t\|_2^2. \tag{3}$$

Recovery is a forward pass at one fixed index, not an iterative reverse trajectory. At detection $\Delta_n = 0$ and the codec is frozen, while basis sampling and Gaussian corruption follow the evaluation seed.

3.2 Target-Excluded Relational Closure

The residual of Eq. 3 leaves undetermined whether the observed channels still explain one another, which is the assessment this path supplies. Let $Z_t \in \mathbb{R}^{L\times D}$ denote $S(X_t)$ standardized channel-wise

across the window, and $z_j = Z_{:,j}$. A predictor able to read its target can copy that trajectory and report agreement without consulting any peer, so every message entering target i comes from sources $j \neq i$ alone. Selection cannot be delegated to the coefficients either.

Lemma 1 (*Normalization erases absolute strength*). *Let $\mathcal{W}_\epsilon(C)_{ij} = C_{ij}/(\sum_k |C_{ik}| + \epsilon)$ denote row-wise L_1 normalization. For any $\alpha > 0$ and any matrix E with $s_i(E) = \sum_k |E_{ik}| > 0$,*

$$\|\mathcal{W}_\epsilon(\alpha E)_{i:} - \mathcal{W}_0(E)_{i:}\|_1 = \frac{\epsilon}{\alpha\, s_i(E) + \epsilon} \xrightarrow[\alpha s_i(E)/\epsilon \to \infty]{} 0. \tag{4}$$

A row of uniformly weak associations therefore still receives nearly unit total weight, so we separate *signed contributions*, stating how a peer explains its target, from *gates*, stating whether it participates.

Signed peer contributions. Let $C(\cdot)$ return a temporal Pearson correlation matrix with zero diagonal. Normal training windows $\mathcal{W}_{\mathrm{tr}}$ supply a signed reference over all off-diagonal candidates, which a learned correction then modifies:

$$C_0 = \frac{1}{|\mathcal{W}_{\mathrm{tr}}|} \sum_{X \in \mathcal{W}_{\mathrm{tr}}} C(S(X)), \quad \omega_{ij} = C_{0,ij} + a_\phi([v_i; v_j]), \quad b_{ij} = \frac{\omega_{ij}}{\sum_{k \neq i} |\omega_{ik}| + \epsilon}, \tag{5}$$

where $v_i \in \mathbb{R}^r$ is a channel embedding and a_ϕ an edge MLP. Every pair stays a candidate, including pairs with zero initial correlation, since the additive correction can reverse a sign or introduce an absent connection. Unlike C_0, the learned coefficients are asymmetric, as predicting i from j differs from the reverse.

Per-edge gates. Each candidate carries a trainable logit α_{ij}, relaxed into a hard-concrete gate (Louizos et al., 2017) so that admission is learned by gradient descent while exact zeros stay reachable:

$$g_{ij}(u) = \mathrm{clip}_{[0,1]}\Big(\gamma + (\zeta - \gamma)\, \sigma\big(T_g^{-1}[\log \tfrac{u_{ij}}{1-u_{ij}} + \alpha_{ij}]\big)\Big), \quad w_{ij} = b_{ij}\, g_{ij}(u), \tag{6}$$

with $u_{ij} \sim \mathcal{U}(0,1)$, replaced by its deterministic counterpart at validation and detection. Normalization precedes gating and none follows, since renormalizing would cancel exactly the attenuation the gates express; by Lemma 1 this ordering keeps admission decidable.

Source-only nonlinear messages. Each source trajectory therefore passes through a residual temporal network conditioned on the pair identity:

$$\psi_\phi(z_j; v_i, v_j) = z_j + \mathrm{Conv}_1(\mathrm{GELU}(\mathrm{Conv}_3(z_j) + E_\phi[v_i; v_j])). \tag{7}$$

A single aggregation step forms the prediction, $\big[G_{\phi,\alpha}(Z;u)\big]_{:,i} = \sum_{j \neq i} w_{ij}\, \psi_\phi(z_j; v_i, v_j)$. Subscripts denote kernel widths. The offset $E_\phi[v_i; v_j]$ encodes channel identities, not the target's trajectory, and the single hop admits no all-channel mixer that could route target information back through a peer.

Closure as the relational score. With $G^\star$ the frozen predictor and $\widetilde{Z}_t = G^\star(Z_t)$, the *closure residual* measures how far the observation lies from its peer predictions, aggregated over all retained channels and compared against its own window background:

$$r_{t,\ell} = \frac{1}{D} \sum_{i=1}^{D} \big|Z_t(\ell, i) - \widetilde{Z}_t(\ell, i)\big|, \quad e_t^{\mathrm{str}} = \frac{r_{t,L}}{L^{-1} \sum_{\ell=1}^{L} r_{t,\ell} + \epsilon}. \tag{8}$$

The scored channel set is fixed independently of the gates, so closing a row's connections changes the predictor without removing that target from evaluation. The ratio isolates a *localized change in coordination*; a departure sustained across the window raises numerator and background alike, so this score does not measure anomaly duration.

3.3 Training and Detection

Separate prediction objectives. Normal windows supervise both paths, but the two objectives update disjoint parameters:

$$\begin{aligned} \mathcal{L}_{\mathrm{rec}}(\theta) &= \mathbb{E}_{X,\varepsilon,\mathcal{B},\Delta_n}\left[\frac{\|X - f_\theta(X; \varepsilon, \mathcal{B}, \Delta_n)\|_F^2}{LD}\right], \\ \mathcal{L}_{\mathrm{rel}}(\phi, \alpha) &= \mathbb{E}_{Z,u}\left[\frac{1}{LD} \sum_{\ell=1}^{L} \sum_{i=1}^{D} \mathrm{Huber}_1(Z(\ell,i) - [G_{\phi,\alpha}(Z;u)]_{\ell,i})\right] + \lambda \mathcal{S}(\alpha), \end{aligned} \tag{9}$$

where X is drawn from $\mathcal{N}_{\text{tr}}$, Z is its standardized detrended counterpart, and $\mathcal{S}(\alpha)$ penalizes the mean probability that a gate stays open. The codec is frozen before the reconstructor is trained, and $\mathcal{L}_{\text{rel}}$ sends no gradient to θ or the codec. Every target contributes to the Huber loss even when all of its incoming gates close, so pruning cannot lower the objective by dropping a difficult target. Normal validation selects both checkpoints, by reconstruction error for θ and by deterministic unpenalized Huber loss for (ϕ, α).

Normal-calibrated decision. The two errors share no scale. Each is standardized by its own branch statistics on $\mathcal{N}_{\text{cal}}$ and passed through softplus to give nonnegative evidence, and the fused score is their equal-weight mean:

$$a_t = \tfrac{1}{2}\left(u_t^{\text{rec}} + u_t^{\text{str}}\right), \quad \widehat{y}_t = \mathbb{I}\big[a_t > Q_\delta(\{a_s\}_{s\in\mathcal{I}_{\text{cal}}})\big], \tag{10}$$

where δ is the 99th percentile of a_t over normal calibration endpoints; fusing before thresholding is not equivalent to thresholding each branch and taking a disjunction. Parameters, score transforms, and h stay fixed at test time, and labels enter evaluation only, without point adjustment. This quantile is an operating point, not a guarantee of false-alarm control under distribution shift.

3.4 Theoretical Analysis

Both statements concern the frozen operator $G^\star$ in fixed standardized coordinates, before the endpoint ratio of Eq. 8. The first identifies what the architectural restriction guarantees, the second which perturbations the residual can expose.

Theorem 1 (*Target-excluded closure response*)**.** *Let $E(Z) = Z - G^\star(Z)$ and let e_i be the ith coordinate vector. For any target i and trajectory perturbation $\delta \in \mathbb{R}^L$,*

$$\big[G^\star(Z + \delta e_i^\top)\big]_{:,i} = \big[G^\star(Z)\big]_{:,i}, \quad E(Z + \delta e_i^\top)_{:,i} - E(Z)_{:,i} = \delta. \tag{11}$$

Proof. By Eq. 7, every message entering target i is a function of z_j with $j \neq i$ only, and the frozen coefficients and deterministic gates are input-independent. Perturbing z_i thus leaves $[G^\star(Z)]_{:,i}$ unchanged; subtracting it gives the residual identity. □

No linearity is assumed: ψ_ϕ is nonlinear, but no source of target i is z_i. A target-only change therefore enters the signed residual undiminished, and cannot be absorbed by copying the target. Absolute error need not grow, since a perturbation may cancel an existing residual.

Changes explained by peers. For a general perturbation Δ, let $U(Z, \Delta) = \Delta - [G^\star(Z+\Delta) - G^\star(Z)]$ be the part of the change the predictor fails to reproduce, with mean absolute magnitudes $\mathcal{R}(Z) = \|E(Z)\|_{1,1}/(LD)$ and $\mathcal{D}(Z, \Delta) = \|U(Z, \Delta)\|_{1,1}/(LD)$. Since $E(Z + \Delta) = E(Z) + U(Z, \Delta)$ requires no linearity, the reverse triangle inequality in both orders gives

$$\big|\mathcal{R}(Z + \Delta) - \mathcal{D}(Z, \Delta)\big| \le \mathcal{R}(Z), \quad \big|\mathcal{R}(Z + \Delta) - \mathcal{R}(Z)\big| \le \mathcal{D}(Z, \Delta). \tag{12}$$

When normal closure is small, the first bound places $\mathcal{R}(Z + \Delta)$ near the unexplained part of the change, and the two together give $\mathcal{D}(Z, \Delta) > 2\mathcal{R}(Z) \Rightarrow \mathcal{R}(Z + \Delta) > \mathcal{R}(Z)$. Conversely, a coordinated change the peers reproduce keeps $\mathcal{D}$ small and the residual near its original value, however large $\|\Delta\|$. Neither direction requires the reconstruction error to move, which is what makes the branches non-redundant.

4 Experiments

4.1 Experimental Settings

Benchmarks. We evaluate `CLOSURE` on five real-world multivariate time series datasets. (*i*) **MSL** is public data released from NASA (Hundman et al., 2018), which contain the telemetry anomaly data derived from the Incident Surprise Anomaly reports of spacecraft monitoring systems. (*ii*) **SWaT** (Mathur & Tippenhauer, 2016) originates from a critical infrastructure system under 11 days of continuous operations with 51 sensors. (*iii*) **PSM** (Abdulaal et al., 2021) is collected internally from diverse application server nodes from eBay. (*iv*) **SMD** (Su et al., 2019) is a five-week-long dataset released from a large Internet company. We additionally consider two NeurIPS-TS scenarios

Table 1: Anomaly detection performance on five benchmarks. P, R, and F1 denote affiliation metrics (%). The bold text denoting the highest and underlined text denoting the second highest value.

Dataset	SMD			MSL			SMAP			SWaT			PSM		
Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
LOF	57.69	99.10	72.92	49.89	72.18	59.00	47.92	82.86	60.72	53.20	96.73	68.65	53.90	99.91	70.02
iForest	100.00	9.37	17.13	53.87	94.58	68.65	41.12	68.91	51.51	53.03	62.80	62.03	100.00	3.35	6.48
CL-MPPCA	58.29	88.43	60.19	50.10	76.30	60.49	48.50	80.20	60.44	54.32	90.16	67.53	55.27	93.46	69.91
TimesNet	58.50	82.32	55.29	52.10	88.47	55.13	42.40	88.10	50.29	48.37	84.23	51.14	65.16	94.23	66.14
MTAD-GAT	85.90	67.69	75.71	54.96	94.93	69.81	39.05	93.99	55.08	65.90	77.51	71.23	79.90	60.14	68.63
LSTM-VAE	70.26	78.53	74.12	60.30	91.26	72.28	55.10	88.60	67.95	62.44	80.18	70.15	57.23	92.40	71.48
BeatGAN	74.11	81.64	77.69	55.74	98.94	71.30	54.04	98.30	69.74	61.89	83.46	71.08	58.81	99.08	73.81
InterFusion	56.20	90.10	58.11	48.40	92.30	54.79	42.10	93.21	51.67	50.20	88.50	54.14	60.30	95.60	62.87
DeepSVDD	64.98	64.77	64.88	10.53	100.00	19.06	29.73	7.09	11.45	59.11	93.53	72.44	74.05	50.64	60.15
AnoTran	100.00	3.19	6.19	51.04	95.36	66.49	56.91	96.69	71.65	53.63	59.94	57.59	52.01	82.18	64.55
TranAD	59.18	91.26	60.14	53.20	94.10	58.31	45.65	96.20	54.66	52.12	90.30	56.07	63.23	97.10	65.11
NPSR	57.45	89.82	58.57	50.50	92.61	56.13	44.20	95.10	53.36	49.67	88.20	53.68	61.18	96.30	63.63
DCdetector	50.93	95.57	66.45	55.94	95.53	70.56	53.12	98.37	68.99	53.25	98.12	69.03	54.72	86.36	66.99
DiffAD	60.25	92.50	61.34	55.10	95.30	59.91	48.30	97.10	56.69	54.20	91.56	57.70	64.13	98.20	66.15
ImDiffusion	61.36	93.25	62.24	56.22	96.11	60.88	49.17	97.80	57.46	55.45	92.31	58.72	65.22	98.60	66.87
D3R	64.87	97.93	78.04	56.45	95.55	71.81	51.08	94.46	66.30	64.25	77.50	70.25	53.17	100.00	69.43
CLOSURE	81.65	95.73	**82.13**	61.24	96.85	**75.53**	58.62	98.41	**73.68**	63.14	86.71	**73.07**	65.28	98.76	**76.49**

used by PhysDiff (Li et al., 2025b): (*vi*) **GECCO**, named Water in the local release, records nine sensor variables from a water-monitoring system, including pH, temperature, and conductivity. (*vii*) **Creditcard** contains 29 transaction features with rare fraudulent transactions marked as anomalies, extending the benchmark scope to financial fraud detection. For fairness, we split all datasets into training $\mathcal{N}_{\text{tr}}$, validation $\mathcal{N}_{\text{val}}$, and calibration $\mathcal{N}_{\text{cal}}$ in a $70\%/15\%/15\%$ ratio.

Baselines. We conduct a comparative analysis of CLOSURE against 16 baselines for comprehensive evaluations, including (*i*) *the classic methods*: LOF (Breunig et al., 2000), iForest (Liu et al., 2012); (*ii*) *prediction-based methods*: CL-MPPCA (Tariq et al., 2019), TimesNet (Wu et al., 2022), MTAD-GAT (Zhao et al., 2020); (*iii*) *reconstruction-based methods*: LSTM-VAE (Park et al., 2018), BeatGAN (Zhou et al., 2019), InterFusion (Li et al., 2021), OmniAnomaly (Su et al., 2019). Then, from the pespective of model architecture, we compare our model with: (*iv*) *Transformer-based methods*: Anomaly Transformer (AnoTran) (Xu et al., 2021), TranAD (Tuli et al., 2022), NPSR (Lai et al., 2024), DCdetector (Yang et al., 2023); and (*v*) *Diffusion-based methods*: DiffAD (Xiao et al., 2023), ImDiffusion (Chen et al., 2023), D3R (Wang et al., 2023).

Evaluation Metrics. To assess the detection performance of our proposed CLOSURE and the baseline models, we employ three standard metrics: affiliation precision, recall, and F1 (Huet et al., 2022), denoted Aff-P, Aff-R, and Aff-F1. These metrics measure the temporal proximity between predicted anomaly intervals and ground-truth events, with Aff-F1 defined as the harmonic mean of Aff-P and Aff-R. All values are reported as percentages, and no point adjustment is applied. To ensure statistical significance, we report the average results over five independent runs.

4.2 Main Results

4.2.1 Quantitative Results

Performance across benchmarks. Table 1 reports the highest Aff-F1 for CLOSURE among the listed methods on all five datasets. The largest margin is on SMD, where its 82.13 exceeds D3R's 78.04 by 4.09 percentage points. On the spacecraft datasets, CLOSURE reaches 75.53 on MSL and 73.68 on SMAP, exceeding the strongest respective baselines, LSTM-VAE and AnoTran, by 2.29 and 2.03 points. The gain on PSM is 2.68 points over BeatGAN. SWaT presents a smaller gap: 73.07 versus 72.44 for DeepSVDD. The advantage in the reported F1 scores therefore spans the listed application domains, but its magnitude is dataset-dependent.

Precision–recall balance. The individual metrics reveal why high precision or recall alone is insufficient. On SMD, iForest and AnoTran both attain 100.00 precision but have recall below

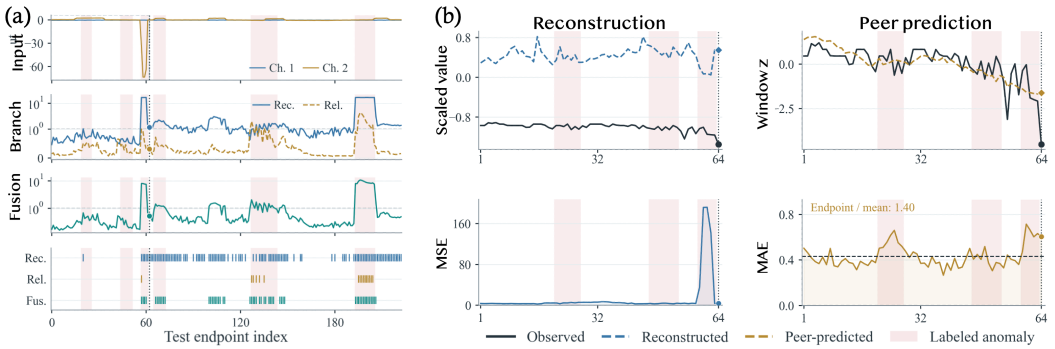

Figure 3: **Suppressing a reconstruction false alarm on SWaT.** (a) Inputs, branch scores, the combined score, and alarm rasters. Scores are divided by their respective frozen thresholds. (b) Reconstruction and peer prediction for one specific channel in the same 64-point window, with residuals averaged over all channels. Reconstruction uses input-scaled coordinates; peer prediction uses window-standardized coordinates. Pink shading marks labeled anomalies.

10.00. `CLOSURE` instead combines 81.65 precision with 95.73 recall. On SWaT, its recall is lower than DeepSVDD's (86.71 versus 93.53), while precision is higher (63.14 versus 59.11), yielding the higher reported F1. These comparisons show different operating points rather than uniform improvement in every metric.

NeurIPS-TS results. Table 2 extends the comparison to transaction data and water monitoring. On Creditcard, `CLOSURE` records an Aff-F1 of 68.54, compared with 62.62 for D3R, the strongest listed baseline. On GECCO, its 65.08 exceeds BeatGAN's 57.25. The respective margins are 5.92 and 7.83 percentage points. In both cases, `CLOSURE` also has higher precision and recall than the baseline with the next-highest F1. On Creditcard, the improvements over D3R are relatively balanced: 7.37 points in precision and 6.00 points in recall. On GECCO, the gain is more recall-oriented, with a 12.38-point increase over BeatGAN alongside a 6.46-point improvement in precision. The reported advantage on these two scenarios thus combines event retrieval with more precise alarm placement, rather than recall alone.

Table 2: Performance on the Creditcard and GECCO scenarios of NeurIPS-TS. P, R, and F1 denote affiliation metrics (%). The bold denotes the highest.

	Creditcard			GECCO		
	P	R	F1	P	R	F1
iForest	1.57	88.16	2.93	2.55	97.35	5.08
BeatGAN	51.76	60.48	62.39	62.71	58.94	<u>57.25</u>
AnoTran	55.86	57.55	56.69	54.77	57.21	55.96
DCdetector	49.86	67.67	57.41	38.30	59.70	46.60
D3R	56.90	69.61	<u>62.62</u>	59.00	36.61	45.18
`CLOSURE`	64.27	75.61	**68.54**	69.17	71.32	**65.08**

4.2.2 Qualitative Analysis

Suppressing a reconstruction false alarm. Figure 3 examines a SWaT case where the two paths yield different decisions. Panel (a) shows the first run of at least three normal endpoints where reconstruction alarms but the combined detector does not. At endpoint 62, reconstruction reaches 1.18 times its threshold, while the relational and combined scores reach 0.31 and 0.75 of their respective thresholds. The alarm raster confirms that the combined rule removes this false alarm. Scores are normalized by their own thresholds, not by fusion weights. Labels are used only for post-evaluation case selection and shading.

Two measurements of the same window. Panel (b) expands the same 64-point window. The specific channel has the largest absolute endpoint closure residual; both residual curves still average all 25 channels. Reconstruction is visibly displaced from the observation. Peer prediction uses detrended, window-standardized coordinates, so its raw error is not directly comparable to reconstruction error. The endpoint closure residual is 1.40 times the window mean, yet its calibrated score remains below threshold. Earlier positions contain labeled anomalies, but the selected endpoint is normal. The case illustrates how relational consistency provides a separate criterion for suppressing a reconstruction alarm without changing the reconstructed signal.

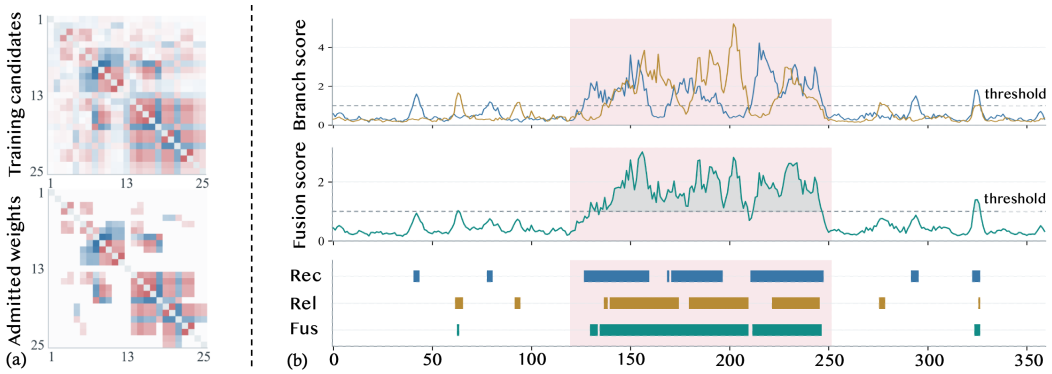

Figure 4: **Selective relations and complementary scoring.** (a) Archived SWaT correlations and admitted weights before row normalization, with red and blue marking positive and negative entries. (b) Hand-constructed scores, where blue, gold, and teal mark reconstruction, relation, and their mean.

Selective relations and complementary scoring. Figure 4 separates two decisions using archived and hand-constructed material rather than one model run. In panel (a), diffuse candidate correlations give way to a sparser reference while groups of positive and negative relations remain: whether a candidate exists and whether it contributes to prediction are separate decisions, carried here by signed coefficients and per-edge gates (Eq. 5 and 6). In panel (b), the equal-weight average detects 114 points of the designed 132-point interval against 97 and 91 individually, while normal-interval false alarms fall from 14 and 11 to 4. Averaging retains a strong response from either branch and suppresses isolated excursions, yet discards six points found by reconstruction alone: fusion is neither an alarm union nor a guarantee of improved coverage.

4.3 Ablation Study

Table 3 examines the two scoring paths and the learned basis memory. The single-path variants keep one path and omit the other; the raw-basis variant replaces the VQ codec with the original basis memory. Adding relational scoring to reconstruction raises Aff-F1 from 69.99 to 73.07 on SWaT and from 78.10 to 82.13 on SMD, exceeding relation-only scoring by 4.99 and 6.34 points. The relational path is the weaker standalone detector here, yet removing it lowers both precision and recall, consistent with peer consistency acting as a complement rather than a replacement. Replacing VQ memory with raw bases lowers Aff-F1 by 1.50 and 1.60 points. Since this variant still supplies normal basis templates and retains relational scoring, the gap indicates a benefit from learning the basis representation rather than merely providing normal temporal context; it evaluates encoder, codebook, and decoder jointly and does not isolate quantization.

Table 3: **Component ablations.** P, R, and F1 are affiliation metrics. The bold denotes the highest.

	SWaT			SMD		
Variant	P	R	F1	P	R	F1
Rec. only	60.42	83.15	69.99	72.60	84.50	78.10
Rel. only	58.76	80.93	68.08	70.60	81.80	75.79
Raw bases	61.85	84.92	71.57	74.80	87.20	80.53
Full model	63.14	86.71	**73.07**	81.65	95.73	**82.13**

5 Conclusion

We revisited a premise of reconstruction-based anomaly detection, that difficulty of recovery tracks abnormality. A residual certifies that a window can be reproduced, and leaves undetermined whether its channels still explain one another. `CLOSURE` therefore evaluates recoverability and cross-channel consistency as two normal-data prediction tasks learned in separation: a diffusion reconstructor conditioned on a discrete memory of normal temporal bases answers the first, and a relation-closure operator answers the second by predicting every channel from the others, with the target's trajectory withheld and per-edge gates deciding which peers participate. Comprehensive experiments demonstrate the effectiveness of our methods.

AI USE STATEMENT

We used generative AI tools to help run the experiments, translate and polish the text, and search the literature; every cited paper was checked against its source by the authors. We carefully checked all code to ensure that it is correct. We take responsibility for the final content of this work, including text, claims and artifacts.

ETHICS STATEMENT

This work does not raise ethical concerns.

REPRODUCIBILITY STATEMENT

The main text specifies the model components, separate training objectives, anomaly scores, and evaluation metrics. The implementation separates fitting and calibration on normal data from test-time scoring. Reproducing the numerical results additionally requires the dataset versions and splits, preprocessing settings, training configurations, and scoring scripts associated with each run. Details are illustrated in appendix.

REFERENCES

Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2485–2494, 2021.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3395–3404, 2020.

Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International conference on machine learning*, pp. 524–533. PMLR, 2019.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

Yuhang Chen, Chaoyun Zhang, Minghua Ma, Yudong Liu, Ruomeng Ding, Bowen Li, Shilin He, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. Imdiffusion: Imputed diffusion models for multivariate time series anomaly detection. *arXiv preprint arXiv:2307.00754*, 2023.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 4027–4035, 2021.

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1705–1714, 2019. URL https://doi.org/10.1109/ICCV.2019.00179.

Alexis Huet, Jose Manuel Navarro, and Dario Rossi. Local evaluation of time series anomaly detection algorithms. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 635–645, 2022.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 387–395, 2018.

Chih-Yu Andrew Lai, Fan-Keng Sun, Zhengqi Gao, Jeffrey H Lang, and Duane Boning. Nominality score conditioned time series anomaly detection by point/sequential reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024.

Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Ben Hu. Revisiting time series outlier detection: Definitions and benchmarks. *NeurIPS Datasets and Benchmarks*, 20: 21, 2021.

Beibu Li, Qichao Shentu, Yang Shu, Hui Zhang, Ming Li, Ning Jin, Bin Yang, and Chenjuan Guo. CrossAD: Time series anomaly detection with cross-scale associations and cross-window modeling. In *Advances in Neural Information Processing Systems*, 2025a.

Long Li, Wencheng Zhang, Shi Yuan, Hongle Guo, and Wanghu Chen. PhysDiff: A physically-guided diffusion model for multivariate time series anomaly detection. In *Advances in Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=ElTbpJp7b9.

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 *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 3220–3230, 2021.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):1–39, 2012.

Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*, 2017.

Aditya P Mathur and Nils Ole Tippenhauer. Swat: A water treatment testbed for research and training on ics security. In *2016 international workshop on cyber-physical systems for smart water networks (CySWater)*, pp. 31–36. IEEE, 2016.

Daehyung Park, Yuuna Hoshi, and Charles C Kemp. A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. *IEEE Robotics and Automation Letters*, 3(3):1544–1551, 2018.

Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14360–14369. IEEE, 2020.

Junho Song, Keonwoo Kim, Jeonglyul Oh, and Sungzoon Cho. Memto: Memory-guided transformer for multivariate time series anomaly detection. *Advances in Neural Information Processing Systems*, 36:57947–57963, 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 *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2828–2837, 2019.

Xin Sun, Heng Zhou, and Chao Li. Multivariate time series anomaly detection with idempotent reconstruction. In *Advances in Neural Information Processing Systems*, 2025.

Shahroz Tariq, Sangyup Lee, Youjin Shin, Myeong Shin Lee, Okchul Jung, Daewon Chung, and Simon S Woo. Detecting anomalies in space using multivariate convolutional lstm with mixtures of probabilistic pca. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2123–2133, 2019.

Shreshth Tuli, Giuliano Casale, and Nicholas R Jennings. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *arXiv preprint arXiv:2201.07284*, 2022.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

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. *Advances in neural information processing systems*, 36:10758–10774, 2023.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

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 *International Conference on Learning Representations*, 2025.

Chunjing Xiao, Zehua Gou, Wenxin Tai, Kunpeng Zhang, and Fan Zhou. Imputation-based time-series anomaly detection with conditional weight-incremental diffusion models. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2742–2751, 2023.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642*, 2021.

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*, pp. 3033–3045, 2023.

Tao Yin, Shaochen Fu, Zhibin Zhang, Li Huang, Xiaohong Zhang, Yiyuan Yang, Kaixiang Yang, and Meng Yan. ScatterAD: Temporal-topological scattering mechanism for time series anomaly detection. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=UB5u3KvOS1.

Chuxu Zhang, Dongjin Song, Yuncong Chen, Xinyang Feng, Cristian Lumezanu, Wei Cheng, Jingchao Ni, Bo Zong, Haifeng Chen, and Nitesh V. Chawla. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1409–1416, 2019. doi: 10.1609/aaai.v33i01.33011409. URL https://doi.org/10.1609/aaai.v33i01.33011409.

Weiqi Zhang, Chen Zhang, and Fugee Tsung. Grelen: Multivariate time series anomaly detection from the perspective of graph relational learning. In *IJCAI*, volume 734, pp. 2390–2397, 2022.

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 *2020 IEEE international conference on data mining (ICDM)*, pp. 841–850. IEEE, 2020.

Bin Zhou, Shenghua Liu, Bryan Hooi, Xueqi Cheng, and Jing Ye. Beatgan: Anomalous rhythm detection using adversarially generated time series. In *IJCAI*, volume 2019, pp. 4433–4439, 2019.

A BASELINE DESCRIPTIONS

We summarize the methods appearing in the main comparison tables and experimental setup; the descriptions concern their modeling principles rather than a claim that their original evaluation protocols are identical.

- **LOF** (Breunig et al., 2000) compares an observation's local density with those of its nearest neighbors and assigns a larger outlier score to observations in relatively sparse neighborhoods.
- **iForest** (Liu et al., 2012) isolates observations through random recursive partitions and scores points that require shorter isolation paths as more anomalous.
- **CL-MPPCA** (Tariq et al., 2019) combines convolutional LSTM modeling with mixtures of probabilistic principal component analyzers to capture temporal behavior and multiple normal operating patterns.
- **TimesNet** (Wu et al., 2022) reshapes a sequence into period-dependent two-dimensional tensors and uses convolutional blocks to model within-period and across-period variation.
- **MTAD-GAT** (Zhao et al., 2020) combines feature-wise and temporal graph attention with forecasting and reconstruction objectives to model dependencies across variables and time.
- **LSTM-VAE** (Park et al., 2018) models normal sequences with recurrent latent-variable encoding and decoding, using the resulting probabilistic reconstruction to assess anomalous observations.
- **BeatGAN** (Zhou et al., 2019) regularizes time-series reconstruction through adversarial training and identifies anomalous regions from discrepancies between observations and generated reconstructions.
- **InterFusion** (Li et al., 2021) uses a hierarchical variational model to capture inter-metric and temporal dependencies and evaluates observations through their reconstruction probabilities.
- **DeepSVDD (?)** learns a representation concentrated around a normal-data center and detects anomalies from their distance to that center rather than reconstruction error.
- **Anomaly Transformer (AnoTran)** (Xu et al., 2021) contrasts a temporal prior association with learned series associations and combines their discrepancy with reconstruction evidence for detection.
- **TranAD** (Tuli et al., 2022) uses attention-based reconstruction with self-conditioning and adversarial training to model multivariate windows and score their reconstruction deviations.
- **NPSR** (Lai et al., 2024) combines point-wise and sequence-wise reconstruction to estimate nominality and condition anomaly scores on the distinction between point and contextual deviations.
- **DCdetector** (Yang et al., 2023) learns multiscale representations with an asymmetric dual-attention contrastive objective and derives detection evidence without a reconstruction objective.
- **DiffAD** (Xiao et al., 2023) detects anomalies through conditional diffusion-based imputation, progressively adjusting conditioning weights to preserve information from observed values.
- **ImDiffusion** (Chen et al., 2023) learns diffusion-based time-series imputers and uses their denoised estimates to measure departures from temporal and cross-variable context.
- **D3R** (Wang et al., 2023) combines dynamic decomposition with diffusion reconstruction to separate nonstationary drift from the deviations used for anomaly detection.
- **OmniAnomaly** (Su et al., 2019), also mentioned in the experimental setup, uses a stochastic recurrent model with a flexible variational posterior and reconstruction probabilities to detect multivariate anomalies.

B Datasets and Data Handling

B.1 Benchmark Descriptions

Server Machine Dataset (SMD). SMD contains five weeks of monitoring data from 28 machines, with 38 recorded metrics per machine (Su et al., 2019). Its source release supplies separate training and test sequences and test anomaly annotations for each machine. Machine identities define separate sequences, not adjacent portions of one continuous trajectory.

Mars Science Laboratory (MSL). MSL contains spacecraft telemetry from the Curiosity rover and expert-annotated anomalous intervals (Hundman et al., 2018). The source collection contains 27 telemetry streams. A stream's model inputs include a monitored telemetry variable and encoded command context; these input columns should not all be interpreted as distinct physical sensors.

Soil Moisture Active Passive (SMAP). SMAP is the companion NASA spacecraft collection with 55 telemetry streams and annotated anomalous intervals (Hundman et al., 2018). As in MSL, the monitored signal is accompanied by command-related inputs. The number of streams in the collection is distinct from the feature dimension of an individual input window.

Secure Water Treatment (SWaT). SWaT records the operation of a six-stage water-treatment testbed through 51 sensor and actuator variables (Mathur & Tippenhauer, 2016). The commonly used release contains normal-operation data and a separate attack period. Measurements describe interacting processes such as flow, tank level, and valve operation, making cross-variable consistency relevant to monitoring.

Pooled Server Metrics (PSM). PSM contains multivariate monitoring signals collected from eBay application servers (Abdulaal et al., 2021). The released training and test files describe server behavior through anonymized metrics, with anomaly labels supplied separately for testing. It complements SMD with a different operational environment and dependency structure.

GECCO. The GECCO scenario in the NeurIPS time-series benchmark originates from drinking-water quality monitoring (Lai et al., 2021). It is named `Water` in the local release and contains nine measurement variables, including pH, temperature, and conductivity. Its annotations concern departures in a continuously monitored water-quality process.

Creditcard. The Creditcard scenario contains transaction records with fraud annotations and is included in the same benchmark (Lai et al., 2021). The local release uses 29 transaction features. Unlike regularly sampled sensor measurements, consecutive observations represent transaction events; window position therefore describes event order rather than a fixed physical time interval.

C Algorithmic Details and Pseudocode

We detail the independent fitting of the basis codec, reconstructor, and sparse relation operator in Section 3, followed by calibration of their equal-weight score. The relational formulation assumes $D \geq 2$ retained channels. We write $\mathcal{W}_q$ for the windows from segment $q \in \{\mathrm{tr}, \mathrm{val}, \mathrm{cal}, \mathrm{test}\}$ and store each input window together with its aligned detrended copy.

C.1 Fitting the Three Components

Basis codec. For a normalized basis patch b, let $z = E_u(b)$, $c = c^u_{q_u(b)}$, and $\widehat{b} = D_u(z + \mathrm{sg}[c - z])$, as in Eq. 1. Each band minimizes the usual straight-through quantization objective (Van Den Oord et al., 2017):

$$\mathcal{L}^u_{\mathrm{VQ}} = \mathbb{E}_b\Big[\mathrm{MSE}(b, \widehat{b}) + \mathrm{MSE}(\mathrm{sg}[z], c) + \beta_{\mathrm{vq}}\,\mathrm{MSE}(z, \mathrm{sg}[c])\Big]. \tag{13}$$

Here MSE averages squared coordinate differences, and $\beta_{\mathrm{vq}} > 0$ weights commitment to the selected code. The codebook loss updates the selected code vector, while the straight-through reconstruction gradient and commitment loss update the encoder. The fitted codecs and template indices are

subsequently frozen. Each stored template retains paired high- and low-frequency code sequences; the two bands are read with the same template index.

Reconstruction and relation fitting. The reconstructor minimizes $\mathcal{L}_{\text{rec}}$ with the frozen memory. Independently, the relation predictor minimizes $\mathcal{L}_{\text{rel}}$ over all retained targets. Its candidates comprise all $D(D-1)$ directed off-diagonal pairs, not just the nonzero entries of C_0. One hard-concrete gate vector is drawn per training minibatch; validation uses deterministic gates. The scored target set is not changed by pruning.

Algorithm 1 Normal-data fitting and calibration of `CLOSURE`

Require: Normal sequence, window length L, fixed preprocessing and optimization settings
Ensure: Frozen detector state Ω
Split normal data into $\mathcal{N}_{\text{tr}}, \mathcal{N}_{\text{val}}, \mathcal{N}_{\text{cal}}$ with purges
Fit the input scaler on $\mathcal{N}_{\text{tr}}$; transform each segment separately
Detrend each transformed segment channel-wise; construct aligned windows
Extract paired high/low bases from normal training windows
Fit basis normalization and band codecs with Eq. 13
Encode paired templates into memory $\mathcal{M}$; freeze codecs and memory
for each reconstruction update **do**
    Sample $X \in \mathcal{W}_{\text{tr}}$, paired memory slots, noise, and training offsets
    Update only θ using $\mathcal{L}_{\text{rec}}$ in Eq. 9
end for
Select $\theta^\star$ by normal-validation reconstruction loss; freeze it
Compute C_0 from detrended training windows using Eq. 5
Initialize channel embeddings, source-message network, edge MLP, and gate logits
for each relation update **do**
    Sample detrended training windows and standardize them to obtain Z
    Draw one gate vector; form signed weights by normalizing *before* gating
    Predict each target using only other channels, with no post-gate normalization
    Update only (ϕ, α) using all-target $\mathcal{L}_{\text{rel}}$
end for
Select $(\phi^\star, \alpha^\star)$ by deterministic validation Huber loss; freeze $G^\star$
Compute both raw scores on every complete calibration window
Fit (μ_b, σ_b) and transforms $\mathcal{T}_b$, $b \in \{\text{rec}, \text{str}\}$
Set $h = Q_{0.99}(\{[\mathcal{T}_{\text{rec}}(e_t^{\text{rec}}) + \mathcal{T}_{\text{str}}(e_t^{\text{str}})]/2\}_{t \in \mathcal{I}_{\text{cal}}})$
return $\Omega = (\text{preprocessing}, \mathcal{M}, \text{codecs}, \theta^\star, G^\star, \mathcal{T}_{\text{rec}}, \mathcal{T}_{\text{str}}, h)$

Algorithm 2 Frozen endpoint detection with `CLOSURE`

Require: Frozen state Ω, test sequence, and fixed evaluation random seed
Ensure: Endpoint scores a_t and alarms $\widehat{y}_t$ for every complete window
Apply the training-fitted scaler and prescribed offline detrending to the test segment
Construct aligned input and detrended windows; retain the final minibatch
for each endpoint $t = L, \ldots, n_{\text{test}}$ **do**
    Read paired high/low basis templates from frozen memory using one shared slot index
    Decode bases; sample the prescribed corruption noise; set $\Delta_n = 0$
    Recover $\widehat{X}_t$ with one fixed-index forward pass using Eq. 2
    $e_t^{\text{rec}} \leftarrow D^{-1} \left\| X_t(L,:) - \widehat{X}_t(L,:) \right\|_2^2$
    Standardize the aligned detrended window to obtain Z_t
    $\widetilde{Z}_t \leftarrow G^\star(Z_t)$ using frozen coefficients and deterministic gates
    $r_{t,\ell} \leftarrow D^{-1} \sum_{i=1}^{D} |Z_t(\ell, i) - \widetilde{Z}_t(\ell, i)|$, $\ell = 1, \ldots, L$
    $e_t^{\text{str}} \leftarrow r_{t,L} / (L^{-1} \sum_{\ell=1}^{L} r_{t,\ell} + \epsilon)$
    $a_t \leftarrow [\mathcal{T}_{\text{rec}}(e_t^{\text{rec}}) + \mathcal{T}_{\text{str}}(e_t^{\text{str}})]/2$
    $\widehat{y}_t \leftarrow \mathbb{I}[a_t > h]$
end for
return scores and alarms; access test labels only for subsequent evaluation

D Detailed Theoretical Analysis

We prove Lemma 1, Theorem 1, and Eq. 12, and derive the gate penalty. The results concern a frozen operator in fixed standardized coordinates. They do not imply that every anomaly violates its relations, that the branches are statistically independent, or that fusion improves detection. Nor does an empirical normal-data quantile guarantee a distribution-free false-alarm rate under dependence or distribution shift. Detection accuracy and event coverage require empirical evaluation.

D.1 Proof of the Normalization Lemma

Fix a row i, let $s = \sum_j |E_{ij}| > 0$, and take $\alpha > 0$ and $\epsilon > 0$. Positive scaling preserves signs, so

$$\begin{aligned} \mathcal{W}_\epsilon(\alpha E)_{ij} - \mathcal{W}_0(E)_{ij} &= \frac{\alpha E_{ij}}{\alpha s + \epsilon} - \frac{E_{ij}}{s} \\ &= -\frac{\epsilon E_{ij}}{s(\alpha s + \epsilon)}. \end{aligned} \tag{14}$$

Taking absolute values and summing over j proves the exact identity

$$\sum_j |\mathcal{W}_\epsilon(\alpha E)_{ij} - \mathcal{W}_0(E)_{ij}| = \frac{\epsilon \sum_j |E_{ij}|}{s(\alpha s + \epsilon)} = \frac{\epsilon}{\alpha s + \epsilon}. \tag{15}$$

The right-hand side tends to zero when $\alpha s/\epsilon \to \infty$, establishing Lemma 1. If $\epsilon = 0$, every positive scaling gives exactly the same normalized row. The case $s = 0$ is excluded from the lemma because $\mathcal{W}_0(E)_{i:}$ is undefined; the implemented ϵ-regularized map instead returns a zero row.

Why the order of gating matters. Let $s_i = \sum_{j \neq i} |\omega_{ij}|$ and let m_i be the absolute row mass after deterministic gating. Since $0 \leq \bar{g}_{ij} \leq 1$,

$$m_i = \sum_{j \neq i} |b_{ij}| \bar{g}_{ij} = \frac{\sum_{j \neq i} |\omega_{ij}| \bar{g}_{ij}}{s_i + \epsilon} \leq \frac{s_i}{s_i + \epsilon} \leq 1. \tag{16}$$

If all gates in a row equal $c \in [0, 1]$, its mass is $cs_i/(s_i + \epsilon)$: the gates retain control over attenuation after normalization. A second exact normalization would remove this control because, for $c > 0$ and a nonzero row $b_{i:}$,

$$\frac{cb_{ij}}{\sum_k |cb_{ik}|} = \frac{b_{ij}}{\sum_k |b_{ik}|}. \tag{17}$$

With a positive numerical stabilizer, cancellation is approximate when the row mass dominates that stabilizer, as quantified by the lemma. Normalization thus preserves relative contributions but not their absolute scale. This motivates separate attenuation parameters; it does not prove that learned sparsity necessarily improves detection.

D.2 Expected Gate Count and Deterministic Evaluation

For the hard-concrete gate in Eq. 6, assume $T_g > 0$ and $\gamma < 0 < 1 < \zeta$. The event $g_{ij} > 0$ is unaffected by clipping at the upper endpoint. Writing $u = u_{ij}$ and $\alpha = \alpha_{ij}$ gives

$$\begin{aligned} g_{ij} > 0 &\iff \sigma\left(\frac{\log \frac{u}{1-u} + \alpha}{T_g}\right) > \frac{-\gamma}{\zeta - \gamma} \\ &\iff \log \frac{u}{1-u} > T_g \log \frac{-\gamma}{\zeta} - \alpha. \end{aligned} \tag{18}$$

Because $u \sim \mathcal{U}(0, 1)$, $\Pr(\log \frac{u}{1-u} \leq x) = \sigma(x)$, and therefore

$$\pi_{ij} := \Pr(g_{ij} > 0) = \sigma\left(\alpha_{ij} - T_g \log \frac{-\gamma}{\zeta}\right), \qquad \mathcal{S}(\alpha) = \frac{1}{D(D-1)} \sum_{i \neq j} \pi_{ij}. \tag{19}$$

This is the expected fraction of open candidate gates, following the hard-concrete relaxation of Louizos et al. (2017); it is not a confidence probability that an edge is physically correct. An open

gate can still multiply a zero coefficient, so gate counts and nonzero effective-coefficient counts need not agree.

Validation and detection replace sampling by the deterministic surrogate

$$\bar{g}_{ij} = \mathrm{clip}_{[0,1]}\left(\gamma + (\zeta - \gamma)\sigma(\alpha_{ij})\right). \tag{20}$$

The reference implementation uses $T_g = 2/3$, $\gamma = -0.1$, and $\zeta = 1.1$. In general, $\bar{g}_{ij} \neq \mathbb{E}[g_{ij}]$. Closing every incoming gate of target i sets its prediction to zero, but its Huber loss remains in the objective; pruning cannot remove the target from supervision or scoring.

D.3 Proof of Target-Excluded Closure Response

Fix the fitted parameters, channel embeddings, and deterministic gates. Then target i has the functional form

$$G_i^\star(Z) = \sum_{j \neq i} w_{ij}^\star \, \psi_{\phi^\star}(z_j; v_i^\star, v_j^\star), \qquad z_j = Z_{:,j}. \tag{21}$$

The weights and identity embeddings do not depend on the current observation. Let $Z' = Z + \delta e_i^\top$. For every $j \neq i$, $Z'_{:,j} = z_j$, hence every summand in Eq. 21 is unchanged and $G_i^\star(Z') = G_i^\star(Z)$. Subtracting predictions from observations yields

$$\begin{aligned} E_i(Z') - E_i(Z) &= (z_i + \delta - G_i^\star(Z')) - (z_i - G_i^\star(Z)) \\ &= \delta. \end{aligned} \tag{22}$$

This proves both identities in Theorem 1. The argument is exact for nonlinear source messages and arbitrary perturbation magnitude. At differentiable points, the equivalent Jacobian statement is

$$\frac{\partial G_i^\star}{\partial z_i} = 0_{L \times L}, \qquad \frac{\partial E_i}{\partial z_i} = I_L. \tag{23}$$

The result requires exclusion from the entire forward path, not just a zero diagonal in one matrix. A target-dependent gate, a shared all-channel feature mixer, or multihop feedback through another channel could invalidate the argument. None is present in the specified one-hop operator. A change to channel i can still alter predictions for other targets, since it is allowed to serve as their source.

Signed response versus error magnitude. The identity does not imply $\|E_i(Z')\|_1 > \|E_i(Z)\|_1$. For example, choosing $\delta = -E_i(Z)$ makes the new target residual zero. What always holds is

$$\|\delta\|_1 - \|E_i(Z)\|_1 \leq \|E_i(Z')\|_1 \leq \|\delta\|_1 + \|E_i(Z)\|_1. \tag{24}$$

Thus $\|\delta\|_1 > 2\,\|E_i(Z)\|_1$ suffices to increase that target's absolute residual, but a conclusion about all-channel aggregation additionally requires accounting for the other targets.

D.4 Proof of the General Perturbation Bounds

For an arbitrary $\Delta \in \mathbb{R}^{L \times D}$, define

$$U = \Delta - \big(G^\star(Z + \Delta) - G^\star(Z)\big), \qquad E' = E(Z + \Delta), \quad E = E(Z). \tag{25}$$

Direct substitution gives the exact identity $E' = E + U$. Let $\|A\|_{1,1} = \sum_{\ell,i} |A_{\ell i}|$. Applying the reverse triangle inequality first to E' and U, then to E' and E, gives

$$\begin{aligned} \Big|\|E'\|_{1,1} - \|U\|_{1,1}\Big| &\leq \|E' - U\|_{1,1} = \|E\|_{1,1}, \\ \Big|\|E'\|_{1,1} - \|E\|_{1,1}\Big| &\leq \|E' - E\|_{1,1} = \|U\|_{1,1}. \end{aligned} \tag{26}$$

Division by LD proves both inequalities in Eq. 12. In particular,

$$\mathcal{D}(Z, \Delta) - \mathcal{R}(Z) \leq \mathcal{R}(Z + \Delta) \leq \mathcal{D}(Z, \Delta) + \mathcal{R}(Z). \tag{27}$$

If $\mathcal{R}(Z) \leq \varepsilon_0$, the perturbed residual lies within ε_0 of the unexplained perturbation magnitude. If $\mathcal{D}(Z, \Delta) > 2\mathcal{R}(Z)$, the lower bound is strictly larger than $\mathcal{R}(Z)$. Conversely, if $\mathcal{D}(Z, \Delta) \leq \eta_0$, the second inequality in Eq. 12 limits the change in residual magnitude to η_0, regardless of the size of Δ itself.

Linear special case. For $G^\star(Z) = ZW^\top$, the unexplained change becomes

$$U = \Delta(I - W^\top). \tag{28}$$

Thus a perturbation satisfying $\Delta = \Delta W^\top$ leaves the residual unchanged. In the nonlinear case, the corresponding condition is the finite-increment equality $G^\star(Z + \Delta) - G^\star(Z) = \Delta$, not membership in a fixed linear subspace. These statements describe predictability under the learned operator, not a recovered causal graph or a universal definition of anomalous behavior.

D.5 Coordinates, Endpoint Ratios, and Decision Scope

Fixed standardized coordinates. The perturbations above act on Z, after preprocessing. If a raw window is perturbed and then standardized again, the relevant change is the *induced* difference between the two standardized windows, not the raw additive perturbation. Channel-wise preprocessing preserves target exclusion from peer trajectories, but centering can remove an additive shift and rescaling can change its magnitude. Offline detrending also limits any claim about streaming detection delay.

The endpoint ratio is a different quantity. Suppress the window index and set $b = L^{-1}\sum_{\ell=1}^{L} r_\ell$. For a second nonnegative residual profile r', write $d_L = r'_L - r_L$ and $d_b = L^{-1}\sum_\ell (r'_\ell - r_\ell)$. The exact score difference is

$$\frac{r'_L}{b + d_b + \epsilon} - \frac{r_L}{b + \epsilon} = \frac{d_L(b + \epsilon) - r_L d_b}{(b + \epsilon)(b + d_b + \epsilon)}. \tag{29}$$

Both denominators are positive. Therefore an endpoint-error increase raises the ratio only when $d_L(b + \epsilon) > r_L d_b$; a simultaneous increase in the background can offset it. Since $b \geq r_L/L$,

$$0 \leq \frac{r_L}{b + \epsilon} < L, \qquad r_1 = \cdots = r_L = c > 0 \implies \frac{r_L}{b + \epsilon} = \frac{c}{c + \epsilon}. \tag{30}$$

A uniformly large residual therefore produces a ratio near one, not a score proportional to its absolute magnitude. This readout emphasizes local departures within a window and does not measure how long an anomaly persists.