BEYOND FEATURE SUBSPACES: QUERY-CONDITIONED PROJECTION LEARNING FOR OUTLIER INTERPRETATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most outlier interpretation methods answer *why a point is anomalous* by returning a *feature subspace*: an axis-aligned subset S of the d observed coordinates in which the query stands out. We show that this format can be *expressively inadequate*, but in a way that depends on dimensionality and on the search budget in a manner no method can infer a priori. Enumerating every non-empty coordinate subspace of all 47 ADBench Classical benchmarks, we find that 94.9% of outliers admit no subspace that isolates them from every inlier when $d \leq 6$, where only 31–63 candidates exist, yet 0% when $13 \leq d \leq 20$, where exhaustive search succeeds. Beyond $d \approx 20$ exhaustive search is impossible (MNIST: $2^{512}-1$ subsets), and under the cardinality budget $|S| \leq 3$ that any practical search must adopt, 56.0% of outliers remain unseparated for $21 \leq d \leq 40$. Pooled over all 47 benchmarks (33,170 outliers) the rate is 47.7% under the strict criterion and 16.7% under a 99th-percentile reference. Where the search is exhaustive the failures cannot be attributed to pruning or heuristics. We therefore study TMQ (TransforMer-guided Quadruplet network), which explains a query through a *query-conditioned linear projection* whose parameters are re-estimated for that query alone, so the contrast may be a weighted feature combination rather than a coordinate subset. Sampled quadruplets (query, other outlier, near inlier, far inlier) are embedded; a global attention branch and four leave-one-position-out local branches estimate latent contributions; and three margin constraints organize inlier–inlier, query–outlier and query–inlier distances. We characterize the exact zero-loss distance interval and the subtype condition under which its inter-outlier lower bound is useful. Extensive experiments across twelve benchmarks, nine baselines and seven detector references show that TMQ attains the best pooled mean on six of seven metrics. Project page: https://anonymous.4open.science/r/tmq-684E/.

1 INTRODUCTION

Outlier interpretation asks *which measured properties make a detected instance anomalous* (Panjei et al., 2022; Yoon et al., 2026; Yun et al., 2025). Nearly all existing methods answer in the same format: a *feature subspace*. Score-and-search methods enumerate candidate coordinate subsets and keep the one in which the query scores highest (OAMiner Duan et al. (2015), x-PACS Macha & Akoglu (2018), SiNNE Samariya & Ma (2022)); embedding methods learn a representation, rank the observed features by induced importance, and report the top-k set (COIN Liu et al. (2018), ATON Xu et al. (2021), PARs Feng (2024), WANDER Guan et al. (2025)). Different objectives, same answer type, an axis-aligned subset of the observed coordinates.

This paper starts from a measurement that puts that answer type in question. Fix the isolation criterion these methods score by: a subspace S explains query x if, after projecting onto S, the distance from x to its nearest inlier exceeds that from *every* inlier to its own nearest neighbour. We run this test on all 47 ADBench Classical benchmarks, enumerating *every* non-empty subspace whenever $d \leq 20$ and every $|S| \leq 3$ subset whenever that enumeration is affordable. The outcome is not a uniform verdict but two distinct failure regimes, tabulated in Table 1: exhaustive search still fails at $d \leq 6$, where only 31–63 candidate subsets exist; it *succeeds* for all 185 outliers at $13 \leq d \leq 20$; and beyond $d \approx 20$, where exhaustive search is out of reach ($2^{512}-1$ subsets

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.

Table 1: Subspace separability failure over all 47 ADBench Classical benchmarks, grouped by search budget and dimensionality (min-max features, $\leq$ 2,000 sampled inliers and outliers). "Fail" is the share of detected outliers for which *no* subspace in the searched family isolates the query from the inlier reference. Rows marked 2^d-1 enumerate *every* non-empty subspace, so their failures are exact; rows marked $|S|\leq 3$ enumerate every subset of size at most three, which is all a practical search can afford at those dimensions, so their rates are upper bounds. Pooled rates weight each outlier equally. Per-dataset values are in Appendix H.

Search budget	d	**Bench.**	**Outl.**	**Fail, all**	**Fail, q99**
all 2^d-1 subsets	≤ 6	8.0	5204.0	94.9	13.8
all 2^d-1 subsets	7–12	12.0	9624.0	28.5	12.8
all 2^d-1 subsets	13–20	4.0	185.0	0.0	0.0
every $\lvert S\rvert \leq 3$	21–40	14.0	9102.0	56.0	34.7
every $\lvert S\rvert \leq 3$	41–64	3.0	3829.0	50.0	1.9
sampled $\lvert S\rvert \leq 3$	> 64	6.0	5226.0	21.8	6.8
Pooled over all 47		47.0	33170.0	47.7	16.7

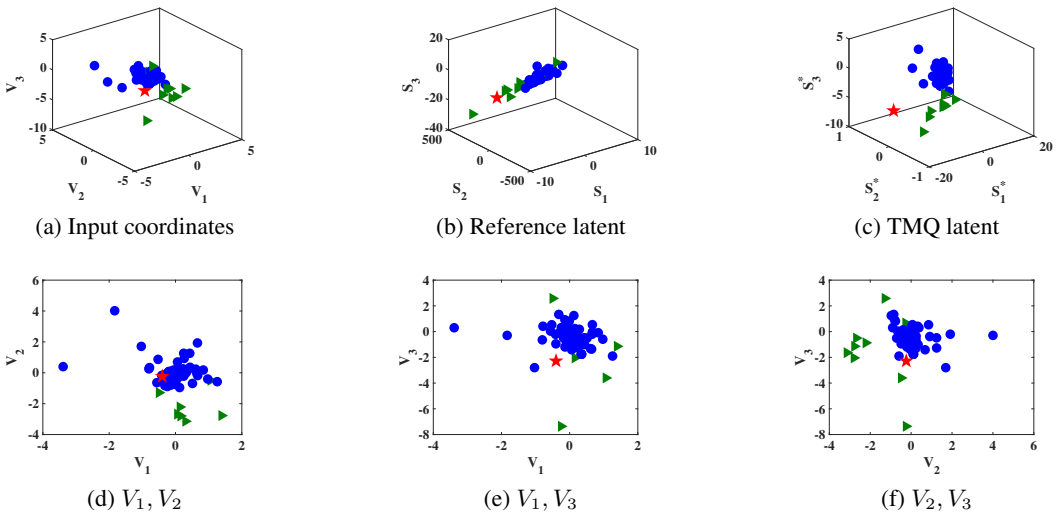
for MNIST), the $|S| \leq 3$ budget that every practical search must adopt leaves 56.0% of outliers unseparated for $21 \leq d \leq 40$. Feasibility is therefore governed by dimensionality and budget jointly (per-dataset rates in Appendix H).

The extreme case is concrete: *vertebral*, a six-feature spinal-disorder benchmark, admits only $2^6-1 = 63$ subspaces, so exhaustive search is trivial, yet all 63 fail for all 30 detected outliers, 73.3% failing even against a 95th-percentile reference; *Wilt* (d=5) fails for all 257. A clinician asking "which of these measurements makes this patient abnormal?" receives no answer in that format, however good the search. The same arises wherever anomalies are defined by *combinations*: a fraudulent transaction can sit inside the normal range on every individual field and differ only in a weighted relation among them. We stress the limit of this evidence: the criterion bounds what the format can *express*, not how useful any method remains; the effect reverses with dimension once search is exhaustive ($\rho = -0.83$, $p < 10^{-6}$, over $d \leq 20$); and 17 of 47 benchmarks fail for under 10% of their outliers. Protocol and sensitivity to the inlier reference are in Appendix H.

TMQ escapes the restriction by changing the answer type. For each query it re-estimates a projection $B_x = \mathrm{diag}(a_x^*)W_x^*$ (Sec. 3), so the contrast is a general direction in $\mathbb{R}^d$, a weighted feature combination, rather than a coordinate selection. Training uses sampled quadruplets (x, x_o, x_i, x_i'), query, nearby other outlier, near inlier, and a non-near inlier, scored by one global and four leave-one-position-out attention branches whose output weights the projection; three margin constraints then order the resulting distances. Figure 2 shows the architecture, Appendix C.1 the full specification.

Figure 1 makes the same point visually on a three-feature fraud case. We are careful about what this measurement does and does not establish. Quadruplet sampling is not itself new: WANDER also samples quadruplets (Guan et al., 2025). We therefore attribute TMQ's behaviour to the global/local attention parameterization of the query-conditioned projection and test that attribution directly through attention ablations (Table 5). The separability measurement likewise does not predict TMQ's margin, over the eleven datasets where both were run, the failure rate is uncorrelated with its F_1 gain over WANDER (Spearman $\rho = +0.22$, $p = 0.52$) and with its absolute F_1 ($\rho = +0.08$, $p = 0.81$), even though TMQ leads WANDER on ten of the eleven. The two are independent pieces of evidence: one characterizes what the answer format can express, the other how well this model performs. Proposition 1 then makes the geometry precise: a zero-loss quadruplet constrains an *interval* of three distances, whereas a triplet objective leaves the query–other-outlier distance unspecified (proof and a non-convexity counterexample in Appendix I), identifying when the extra relation helps (distinct mechanisms) and when it conflicts (same-subtype pairs).

Contributions. **(i)** A measurement rather than an assumption: a separability study over all 47 ADBench Classical benchmarks, exhaustive wherever enumeration is feasible, characterizing *when* the axis-aligned subspace format cannot express the anomaly contrast, low dimensions, where the candidate family is too small, and high dimensions, where no affordable budget suffices, and when it can, with the negative controls reported alongside. **(ii)** TMQ, a query-conditioned projection re-



(a) Input coordinates (b) Reference latent (c) TMQ latent

(d) V_1, V_2 (e) V_1, V_3 (f) V_2, V_3

Figure 1: Three-feature fraud illustration: the query (red star) and another fraud point (green triangle) lie among normal transactions (blue) in the input coordinates, while a learned combination separates them; panels (d)–(f) are two-feature subspaces of the input coordinates. Visualization only: it evidences expressiveness, not manifold preservation or causality.

estimated per query from sampled quadruplets scored by one global and four leave-one-position-out attention branches, returning a weighted feature direction and a ranked readout (Eq. 4) instead of a coordinate subset. **(iii)** An exact account of what the quadruplet objective can and cannot enforce (Proposition 1), including the same-subtype regime where its inter-outlier term conflicts with fidelity. **(iv)** A complete empirical record: pooled and per-dataset results for up to nine baselines under seven detector references, with every incomplete run marked, and an explicit statement of the datasets on which TMQ is not best.

2 Related Work

Score-and-search methods such as OAMiner, x-PACS and SiNNE evaluate candidate subspaces in the observed coordinates (Duan et al., 2015; Macha & Akoglu, 2018; Samariya & Ma, 2022): readable output, but high-dimensional search needs candidate limits or pruning; COIN uses a sparse local classifier; ATON learns a query-conditioned embedding with a triplet deviation objective and outputs feature weights (Liu et al., 2018; Xu et al., 2021); PARs and WANDER supply further subspace comparisons (Feng, 2024; Guan et al., 2025). WANDER is closest: it also samples quadruplets and attaches attention, but estimates the representation with a graph/multi-kernel learner, which TMQ replaces by global and local attention over the four positions. Because "learn an embedding per query, apply attention, read out feature weights" is already the ATON route and the quadruplet already WANDER's, the burden on TMQ is not to show that learning a projection beats enumerating subsets, but to show what this parameterization adds under matched query, quadruplet, objective and readout. SHAP and LIME give local importance scores for a selected predictive model; Anchor returns predictive rules (Lundberg & Lee, 2017; Ribeiro et al., 2016; 2018); converting their outputs to a top-k set permits common overlap scores, but ranking quality and rule coverage remain separate questions. Recent ICLR work attacks the *cost* and *guarantees* of attribution rather than its format (kernel-thinning compression (Baniecki et al., 2025), certified EnsembleSHAP (Wang & Jia, 2026), FAME's minimal abductive explanation (Boumazouza et al., 2026)), but all answer in the observed coordinates, inherit the scoring-function question of Sec. 1, and give guarantees for a classifier's decision rather than an outlier contrast; a certified or rule-based output is *not* refuted by Table 1, since a ranking can be informative even when no subset passes an isolation test. Work on image and telemetry modalities (Xu et al., 2025; Tang et al., 2025; Wang et al., 2025; Zhou & Wu, 2024; Cuellar et al., 2024; Hu et al., 2025) and tabular detection (Ye et al., 2025; Tsai et al., 2025; Yoon et al., 2026; Yun et al., 2025) (the last transforming the features themselves but conditioning

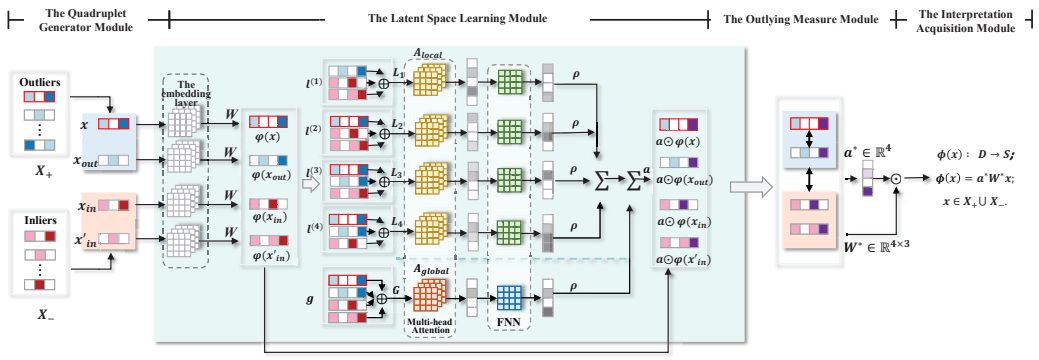


Figure 2: TMQ architecture: quadruplet generation, global/local attention, relational objective, and interpretation acquisition. Each local context withholds one of the four quadruplet positions.

on a user-supplied textual preference rather than the query) emits image-level scores, reconstruction maps, event-level causes or a preference-aligned space, not a per-query contrast against a sampled inlier context. TMQ is post-hoc and complementary: it uses detector-provided outlier and inlier sets and explains contrasts in its learned query projection.

3 METHOD

3.1 PROBLEM AND SCOPE

Let $x \in X_+ \subset \mathbb{R}^d$ be a detected query anomaly and X_- the normal reference set. For an explanation, TMQ estimates a projection into s latent coordinates that exposes how x differs from sampled inliers and, conditionally, other outliers. The output is conditioned on the detector labels, feature preprocessing, sampled neighborhood, and query. The latent contrast is observational; it does not by itself identify a causal failure mechanism. Formal notation and the task definition are given in Appendix 1. This formulation makes the reference choice part of the explanation. If a detector labels a normal instance as an outlier, the sampled outlier context can constrain an irrelevant contrast. If X_- contains contamination, the inlier–inlier scale can expand and change all three margin requirements. For very rare anomalies, k_1 may approach the available outlier count; the resulting comparison is less local than for a dense outlier population. The method is consequently an explanation of a detected population under a specified reference, rather than a universal statement about the semantics of an instance.

3.2 QUADRUPLET SAMPLING AND ATTENTION FLOW

Let $x \in X_+$ be a detected query outlier, with X_+ and X_- the available outliers and inliers. A quadruplet is $q = (x, x_o, x_i, x_i')$: x_o is selected from nearby outliers, x_i from nearby inliers, and x_i' from another sampled inlier pool. With k_1 outlier candidates and k_2 candidates in each inlier pool, the Cartesian construction yields up to $k_1 k_2^2$ quadruplets per query. This growth limits practical batch and memory sizes.

The two inlier positions are deliberately different: one is a near normal comparator for the query, and the other is sampled outside that near-inlier pool. The latter yields a normal variation scale in Eq. 2; the former tests the query against the more difficult normal neighbor. The sampled outlier provides a separate contextual question. An implementation should record the neighbor metric, preprocessing, and random draw because these determine which relations the objective sees. Algorithm 1 summarizes training in the notation of the main paper; Appendix 2 gives the detailed network steps.

Each vector first passes through a learned linear embedding $W \in \mathbb{R}^{s \times d}$. Global attention sees the four embedded positions, while four local branches receive the respective three-position omission contexts. Normalized branch scores are averaged across the sampled quadruplets to produce a latent contribution vector $a \in \mathbb{R}^s$. Omitting one position at a time lets the query and its normal and anomalous references affect attention through distinct relational contexts; the architecture does not

Algorithm 1 Query-conditioned TMQ training pipeline.

Require: query x, detected sets X_+, X_-, k_1, k_2, latent dimension s
collect near outliers, near inliers, and sampled non-near inliers
construct $Q = \{(x, x_o, x_i, x_i')\}$ from their Cartesian product
initialize projection W and global/local attention parameters
repeat
    embed four positions and form one global and four omission contexts
    normalize and average attention scores to obtain latent weights a
    transform the quadruplet positions with $\mathrm{diag}(a)W$
    update parameters against the three-hinge loss in Eq. (2)
until stopping criterion or maximum epoch
return optimized query projection $B_x = \mathrm{diag}(a_x^*)W_x^*$

claim these four contexts are uniquely optimal. Figure 2 shows the four-module architecture, and Appendix C.1 gives the matrix specification and algorithm.

The context structure has a concrete relational interpretation. Omitting x_o leaves query–normal information without another anomaly, while omitting x_i leaves the query, other anomaly, and a sampled non-near normal comparator. Omitting x_i' retains the difficult near inlier, and omitting the query lets the branch encode reference relationships without the fixed anchor. The global branch sees all four jointly. The branches therefore supply distinct reference configurations rather than four arbitrary replicas; subsequent ablation tests whether the extra configurations measurably help. This design rationale does not assert that an attention score alone is a faithful feature attribution.

For a fixed query and optimized parameters, the mapping of a new feature vector z is

$$\phi_x(z) = \mathrm{diag}(a_x^*)W_x^* z. \tag{1}$$

Equation 1 is linear in z. The attention network may depend nonlinearly on the query and its sampled context, but no loss term preserves pairwise distances or a global/local data manifold. TMQ therefore neither claims a nonlinear final projection nor guarantees manifold preservation.

3.3 Relational objective and its scope

Write $\tilde{x} = a \odot Wx$ and analogously transform the other three quadruplet positions. Let $D(u, v) = \|u - v\|_2^2$ and let $m > 0$ be a margin. The per-quadruplet loss is the sum of three hinges:

$$\begin{aligned} L(q) = &[D(\tilde{x}_i, \tilde{x}_i') - D(\tilde{x}, \tilde{x}_o) + m]_+ \\ &+ [D(\tilde{x}_i, \tilde{x}_i') - D(\tilde{x}, \tilde{x}_i) + m]_+ \\ &+ [D(\tilde{x}, \tilde{x}_o) - D(\tilde{x}, \tilde{x}_i) + m]_+. \end{aligned} \tag{2}$$

The first hinge puts the query–outlier distance above the inlier–inlier distance; the second separates the query from the near inlier; the third keeps the other outlier closer than that inlier. The last relation moderates the first, forming a distance hierarchy rather than unlimited repulsion. A triplet objective using only the query–inlier hinge cannot enforce the inter-outlier part. Figure 3 depicts the sampled relationships.

Proposition 1 (Exact feasible interval)**.** *For one quadruplet, write $D_{ii} = D(\tilde{x}_i, \tilde{x}_i')$, $D_{qo} = D(\tilde{x}, \tilde{x}_o)$, and $D_{qi} = D(\tilde{x}, \tilde{x}_i)$. For $m > 0$, $L(q) = 0$ if and only if*

$$D_{qi} \geq D_{ii} + 2m, \qquad D_{qo} \in [D_{ii} + m,\ D_{qi} - m]. \tag{3}$$

For fixed D_{ii}, D_{qi} with $D_{qi} \geq D_{ii} + 2m$, the corresponding triplet hinge can be zero for any nonnegative D_{qo}, including $D_{qo} = 0$.

Proof sketch. Each hinge in Eq. 2 is nonnegative. Their simultaneous vanishing gives $D_{qo} \geq D_{ii} + m$, $D_{qi} \geq D_{ii} + m$, and $D_{qo} \leq D_{qi} - m$. The first and third inequalities imply the stated bound on D_{qi} and already imply the second; the converse follows by substitution. The triplet hinge contains no D_{qo}. □

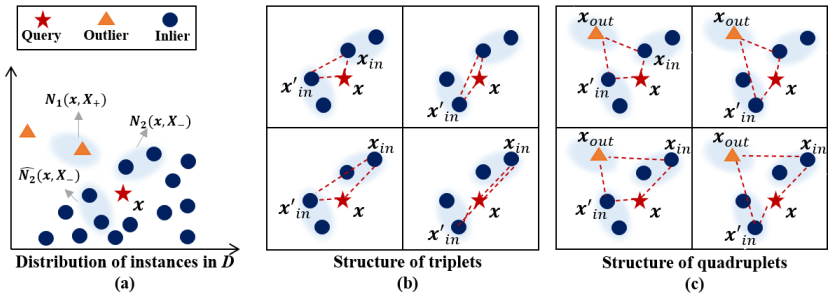


Figure 3: Triplet and quadruplet relational comparisons. The extra other-outlier position constrains a distance that the triplet hinge does not specify.

Proposition 1 makes the benefit conditional. If distinct anomaly subtypes should be separated by at least $D_{ii} + m$, the quadruplet relation supplies an explicit inductive bias absent from the triplet hinge. If x and x_o share a mechanism and fidelity requires $D_{qo} \leq \delta < D_{ii} + m$, the first hinge incurs at least $D_{ii} + m - \delta > 0$. Thus indiscriminately selecting same-subtype outliers creates a conflicting constraint. These statements concern the loss geometry, not an empirical superiority theorem. Appendix I gives the full derivation and a nonconvex counterexample; no unsupported Adam convergence guarantee is used.

3.4 Mapping the latent projection back to features

After optimization, the mapping for a fixed query is $\phi_x(z) = \mathrm{diag}(a_x^*)W_x^* z$ (Eq. 1). The attention estimator can be nonlinear in its context, but the fixed map in z is linear. A latent contribution entry a_r is not an observed-feature score. To make the explanation operation explicit, form the coefficient matrix $B_x = \mathrm{diag}(a_x^*)W_x^*$ and rank input features by the column magnitudes

$$c_j(x) = \|B_x[:, j]\|_2, \quad j = 1, \ldots, d. \tag{4}$$

One may report the complete ranking or its top k input-feature names. This is a projection-derived readout of the optimized weights, not a causal rule; a value-contrast readout can additionally weight c_j by $|x_j - b_j|$ for an inlier reference b. Which exact readout produced each aggregate score should be confirmed against the released code before those scores are attributed to Eq. 4. Keeping this distinction avoids treating latent indices as if they were input features.

There is a simple mathematical reason to inspect projection columns. For any reference vector b in the same preprocessed coordinates, $\phi_x(x) - \phi_x(b) = \sum_{j=1}^{d}(x_j - b_j)B_x[:, j]$. The triangle inequality gives

$$\|\phi_x(x) - \phi_x(b)\|_2 \leq \sum_{j=1}^{d} |x_j - b_j| \, \|B_x[:, j]\|_2. \tag{5}$$

Thus $|x_j - b_j|c_j$ bounds each coordinate's possible contribution to the projected query–reference contrast. This bound does not uniquely attribute a squared distance: correlated columns can cancel or reinforce one another. For an analyst, reporting the ranked field names with their query values and inlier reference values is more informative than reporting s unnamed latent entries. The same standardized coordinates should be used for training and scoring; otherwise physical units alone could determine the top ranks.

4 Experiments

4.1 Data, baselines, and evaluation

We evaluate on twelve ADBench datasets spanning 3–512 features, nine baselines (ATON, COIN, SHAP, LIME, SiNNE, Anchor, RAE-GA, PARs, WANDER), and seven detector-derived feature-reference proxies (ECOD, iForest, COPOD, HBOS, LUNAR, ROD, MCD). Dataset sizes and anomaly rates are in Table 7; the full table and all detector-level metric values appear in Appendix D.

Table 2: The twelve benchmark datasets; the full table, including anomaly rates, is given in Table 7.

Dataset	n	d	Out.	Dataset	n	d	Out.
wpbc	198	33	47	satimage-2	5,803	36	71
glass	214	7	9	annthyroid	7,200	6	534
vertebral	240	6	30	MNIST	10,000	512	500
yeast	1,484	8	507	smtp	95,156	3	30
fault	1,941	27	673	mammography	11,183	6	260
pageblocks	5,393	10	510	fraud	284,807	29	492

Table 3: Pooled means over all dataset–detector cells; arrows give the preferred direction, **bold** marks the best entry per row. These seven metrics are recorded for five baselines; the remaining baselines are compared on F_1 in Table 4. Full detector-level tables appear in Appendix D.

Metric	**ATON**	**COIN**	**SHAP**	**LIME**	**WANDER**	**TMQ**
AUPR ↑	0.649	0.557	0.557	0.527	0.449	**0.692**
AUROC ↑	0.715	0.612	0.619	0.607	0.424	**0.753**
Precision ↑	0.567	0.483	0.446	0.433	0.379	**0.616**
Jaccard ↑	0.455	0.376	0.360	0.348	0.243	**0.491**
Fidelity ↑	0.637	0.526	0.572	0.503	0.410	**0.649**
Insertion AUPRC ↑	0.884	0.873	0.878	0.866	0.816	**0.900**
Deletion AUPRC ↑	0.808	0.833	0.845	**0.864**	<u>0.845</u>	0.847

The detectors define agreement targets, not independent domain-expert ground truth. Rankings are measured with AUPR/AUROC; equal-cardinality top-k sets are compared by precision, Jaccard, and F_1, while fidelity, insertion, and deletion report additional response-based criteria. Matching k to an annotation's size makes overlap comparable, but does not specify a deployment threshold or measure Anchor rules in their native form.

The annotation description says the reference is the feature subset maximizing a detector score. This cannot be exact exhaustive coordinate search on MNIST's 512 features: there are $2^{512} - 1$ nonempty subsets. ATON's cited protocol reduces datasets with more than 15 features to ten PCA coordinates; whether this study used the same reduction, another candidate family, or an approximation needs confirmation from its annotation scripts. We therefore report every score as agreement under this proxy protocol. This is the principal validation gap rather than an excuse to discard the dataset.

Table 2 lists all twelve datasets. It shows why a single aggregate is incomplete: glass supplies only nine anomalies, smtp thirty, whereas fault and yeast have hundreds and much higher anomaly prevalence. MNIST supplies the high-dimensional case, while smtp, annthyroid, mammography, and vertebral have fewer than ten features. This spread probes differing neighborhood and projection regimes, although it does not substitute for a controlled scaling experiment. The resource and per-detector tables in the appendix allow the size, dimension, and anomaly-rate effects to be inspected rather than hidden behind one mean.

4.2 Main quantitative results

Table 3 reports the across-annotation averages. TMQ has the largest listed mean on six of the seven shown metrics; it is *not* best on deletion AUPRC, where LIME's 0.864 exceeds TMQ's 0.847. Its mean AUPR exceeds ATON by 0.043 and WANDER by 0.243; the latter difference mixes architecture, optimization, and sampling choices. Table 4 gives the per-dataset F_1 for all seven baselines that report it, derived from the detector-level rows in Appendix D. TMQ is best on eight of the twelve datasets, but ATON edges it on wpbc (0.693 vs. 0.692) and satimage-2, SiNNE wins MNIST, and COIN wins smtp. The aggregate thus summarizes a comparison that is not uniformly won; incomplete runs are marked rather than scored as zero.

Detector-level scores show margins that vary widely across datasets and references, so a pooled mean does not imply dominance on any particular one. The full precision/Jaccard, ranking, fidelity/insertion/deletion, resource, and F_1 tables are included as legible multipage tables after the

Table 4: Per-dataset subspace F_1 for the seven baselines that report it, averaged over the detector-derived references for which a run completed; the *Mean* row is pooled over all dataset–detector cells, matching Table 3 and the values quoted in the text. An en dash marks a run that exceeded the two-week compute budget and is excluded from the mean rather than scored as zero (COIN on mammography and fraud; RAE-GA and WANDER on fraud). **Bold** and underline mark the best and second-best entry in each row.

Dataset	ATON	COIN	SiNNE	Anchor	PARs	RAE-GA	WANDER	TMQ
glass	0.461	0.460	0.290	0.430	0.435	0.281	0.371	**0.526**
vertebral	0.600	0.510	0.532	0.477	0.360	0.335	0.390	**0.617**
wpbc	**0.693**	0.455	0.570	0.440	0.252	0.382	0.407	0.692
yeast	0.471	0.406	0.476	0.328	0.474	0.207	0.433	**0.495**
fault	0.532	0.531	0.524	0.513	0.266	0.312	0.526	**0.544**
pageblocks	0.515	0.308	0.411	0.387	0.280	0.168	0.263	**0.550**
MNIST	0.567	0.567	**0.613**	0.540	0.173	0.193	0.512	0.572
annthyroid	0.551	0.571	0.443	0.539	0.390	0.339	0.475	**0.579**
satimage-2	**0.593**	0.504	0.412	0.502	0.438	0.175	0.380	0.585
smtp	0.563	**0.584**	0.521	0.577	0.507	0.366	0.582	0.549
mammography	0.532	–	0.535	0.497	0.382	0.512	0.417	**0.581**
fraud	0.505	–	0.464	0.424	0.399	–	–	**0.555**
Mean	0.550	0.491	0.482	0.473	0.361	0.296	0.433	**0.571**

Table 5: TMQ ablation, averaged over the twelve datasets. The bottom two rows replace a component with a structurally different alternative rather than deleting it.

Variant	AUPR	AUROC	F_1
TMQ	**0.692**	**0.753**	**0.571**
w/o global branch	0.688	0.730	0.561
w/o local branches	0.690	0.732	0.565
MLP for attention	0.628	0.643	0.516
Triplet for quadruplet	0.310	0.391	0.270

bibliography. The t-SNE comparisons, critical-difference panels, training traces, sensitivity plots, and runtimes are also given in the appendix.

The comparison with WANDER is particularly informative because both models receive relational quadruplet context. Against WANDER specifically, TMQ leads by more than 0.20 F_1 on wpbc, vertebral, pageblocks and satimage-2, by only 0.018 on fault and 0.062 on yeast, and smtp reverses the ordering. These gaps vary with dataset and detector proxy, and the per-dataset means in Table 4 are derived summaries of those detector-level rows. A controlled architectural study would keep the four sampled positions, optimizer, projection dimension, and readout fixed, then exchange WANDER's graph/multi-kernel estimator for TMQ's global/local attention. The current comparison is over complete pipelines: it measures effectiveness but does not isolate that one module.

Figure 5 in Appendix gives the corresponding Nemenyi critical-difference diagrams for F_1, precision and Jaccard over average ranks. They summarize the common-format comparison at the stated test level; a connected rank group is not definitively different on the diagram alone, and because the converted top-k sets are the statistical units they do not assess rule coverage for Anchor or the native attribution format of SHAP and LIME.

4.3 Mechanism ablation and training behavior

Swapping quadruplets for triplets produces the largest drop, consistent with the extra D_{qo} relation (though WANDER shares it); replacing attention with an MLP has a smaller but substantial effect, while removing either branch changes AUPR by only 0.004 or 0.002. The split's incremental effect is therefore modest: credit goes to the defined query contexts, and isolating the attention parameterization against WANDER needs a controlled replacement study.

The principal setup uses Adam with initial learning rate 0.1, 5% warm-up, gradient clipping at 1.0, plateau decay, up to 40 epochs, batch size 512, $k_1 = 1$, $k_2 = 50$, and $s = \lfloor 1.5d \rfloor$. A later parameter study explores $k_1 = 3$ and $s = 64$; these are sensitivity settings, not replacements for the primary protocol without run-log confirmation. Training-loss panels in Figure 4 in Appendix show empirical stabilization for four datasets, while the remaining sensitivity and ablation details are given there. The loss is nonconvex, so these curves establish measured optimization behavior rather than global convergence.

Loss trajectories for glass, fault, fraud and MNIST (Figure 4 in Appendix) decline and plateau, which distinguishes an observed stopping behavior from Proposition 1's zero-loss geometry; because competing sampled constraints can leave nonzero hinges, a small average loss neither proves every quadruplet feasible nor implies a convergence theorem.

4.4 Parameter sensitivity and interpretation regimes

AUPR sensitivity panels for k_1, k_2, s and batch size, together with the corresponding AUROC and F_1 plots, are given in Figure 8 in Appendix. These parameters govern different aspects of the same query contrast. Increasing k_1 broadens anomaly comparisons but can bring in unrelated or same-subtype anomalies. Increasing k_2 gives more normal reference pairs but multiplies the candidate count quadratically; at the stated $k_2 = 50$, $k_1 = 1$ already gives up to 2,500 quadruplets per query, and $k_1 = 3$ gives 7,500. With a batch of 512, these cannot all be interpreted as one in-memory attention call without a specified batching rule. The sensitivity study recommends $k_1 = 3$, $k_2 = 50$, $s = 64$, and batches between 128 and 512 for the cases it tests. Those observations are useful starting values, while the main scores use the primary settings stated above.

Increasing s gives the projection more capacity to capture cross-feature combinations, but each additional latent coordinate enlarges W and the attention computation and can distribute the input-feature signal over more columns. Batch size affects the variance of estimated attention scores and resource use; a training plateau at one batch size need not mean the feature ranking is stable across sampled neighborhoods. Interpretation stability should be checked with repeated seeds and reference samples, using rank correlation for a complete ranking or overlap at a prespecified deployment k. The consistency and robustness scores in the full resource table are descriptive evidence; they do not speak to the sensitivity of the projection-column readout of Eq. 4, which should be checked against the accompanying code.

Figure 7 in Appendix shows three latent-visualization regimes: fraud, mammography, and glass. Fraud supplies many anomalies at low prevalence, glass just nine; these examples suggest that neighbor context and anomaly heterogeneity can alter the geometry available to the loss. A t-SNE montage is a qualitative projection of learned embeddings, so it cannot directly establish subtype labels or metric distances in the latent space. Together with Proposition 1, the visualizations motivate a subtype-aware follow-up: compare pairs known to share a mechanism against pairs known to differ, record their D_{qo} and D_{qi} distributions, and test whether the extra lower bound helps or harms feature agreement. We leave such a labeled-subtype study to future work.

5 Computational Cost and Discussion

For one query, candidate quadruplets number at most $q = k_1 k_2^2$, before batching. A dense initial projection incurs $O(qds)$ work; attention costs depend on whether it operates over features, positions, or all candidate columns. The full upper bound includes qs^2 and $q^2 s$ terms, so a blanket linear-in-data-size claim needs a fixed sampling and batch budget. TMQ uses about 1–7 GB over the tested cases, versus tens of MB for a simpler comparator in one setting. Measured runtime trends against dimensionality and data size are plotted in Figure 6 in Appendix. Appendix F gives the complete complexity expression, resource table, and additional parameter plots. TMQ conditions a latent contrast on a query, two inlier references and another anomaly; Proposition 1 states the hierarchy it can impose; and because B_x is re-estimated per query, total cost scales linearly in the detected-outlier count (Appendix F). The first hinge should respect subtype semantics, the fitted map is linear in its input, and neither loss nor visualizations demonstrate manifold preservation; native-format comparison and readout verification remain open checks. Output format, the sensitivity of the readout, and where the extra relation is and is not useful are discussed in Appendix J.

Reproducibility statement

The quadruplet construction, network, objective, algorithm, hyperparameters, dataset inventory, and every reported score are given in the appendix. The annotation search or dimensionality reduction, preprocessing, random seeds, attention tensor orientation, and the feature-readout implementation used to produce each aggregate score are specified in the accompanying code, and the anonymous review archive exposes its dependencies and data-preparation scripts.

AI use statement

Generative AI assisted manuscript editing, and template adaptation. The authors are responsible for checking the final claims and disclosing any further AI use in the submission form; no unperformed experiments are described as new measurements.

Ethics statement

TMQ is a post-hoc feature contrast based on detector labels and a chosen inlier reference. Its output should be reviewed with domain context, especially in high-stakes settings; it is not a causal or automatically actionable explanation.

References

Yahya Almardeny, Noureddine Boujnah, and Frances Cleary. A novel outlier detection method for multivariate data. *IEEE Transactions on Knowledge and Data Engineering*, 34(9):4052–4062, 2020.

Hubert Baniecki, Giuseppe Casalicchio, Bernd Bischl, and Przemyslaw Biecek. Efficient and accurate explanation estimation with distribution compression. In *Proceedings of the 2025 International Conference on Learning Representations, ICLR*, 2025.

Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. In *Proceedings of the 2018 International Conference on Machine Learning, ICML*, pp. 599–608, 2018.

Ryma Boumazouza, Raya Elsaleh, Melanie Ducoffe, Shahaf Bassan, and Guy Katz. FAME: formal abstract minimal explanation for neural networks. In *Proceedings of the 2026 International Conference on Learning Representations, ICLR*, 2026.

Mingjie Cai, Chutian Zhou, Chaoqun Huang, Jiaxin Zhan, and Hamido Fujita. Elastic agents in cooperative feature selection through multi-agent reinforcement learning. *Knowledge Based Systems*, 340:115717, 2026.

Sara Cuellar, Matilde Santos, Fernando Alonso, Ernesto Fábregas, and Gonzalo Farias. Explainable anomaly detection in spacecraft telemetry. *Engineering Applications of Artificial Intelligence*, 133:108083, 2024.

Vincenzo Dentamaro, Paolo Giglio, Donato Impedovo, Giuseppe Pirlo, and Marco Di Ciano. An interpretable adaptive multiscale attention deep neural network for tabular data. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):6995–7009, 2025.

Lei Duan, Guanting Tang, Jian Pei, James Bailey, Akiko Campbell, and Changjie Tang. Mining outlying aspects on numeric data. *Data Mining and Knowledge Discovery*, 29(5):1116–1151, 2015.

Cheng Feng. PARs: predicate-based association rules for efficient and accurate anomaly explanation. In *Proceedings of the 2024 International Conference on Information and Knowledge Management, CIKM*, pp. 612–621, 2024.

Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): a fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track*, 1:59–63, 2012.

Adam Goodge, Bryan Hooi, See-Kiong Ng, and Wee Siong Ng. LUNAR: unifying local outlier detection methods via graph neural networks. In *Proceedings of the 2022 Conference on Artificial Intelligence, AAAI*, pp. 6737–6745, 2022.

Lili Guan, Lei Duan, Xinye Wang, Haiying Wang, and Rui Lin. Weighted-digraph-guided multi-kernelized learning for outlier explanation. *Information Fusion*, pp. 103026, 2025.

Yanbu Guo, Quanming Guo, Shengli Song, Yihan Wang, and Jinde Cao. Heterogeneous graph collaborative representation learning for drug-related microbe prediction with attentive fusion and reciprocal distillation. *Knowledge-Based Systems*, 330:114548, 2025.

Johanna Hardin and David M Rocke. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis*, 44(4): 625–638, 2004.

Zhichao Hu, Likun Liu, Lina Ma, and Xiangzhan Yu. Sinkflow: Fast and traceable root-cause localization for multidimensional anomaly events. *Engineering Applications of Artificial Intelligence*, 139:109582, 2025.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the 2018 Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 7482–7491, 2018.

Katiana Kontolati, Somdatta Goswami, George Em Karniadakis, and Michael D Shields. Learning nonlinear operators in latent spaces for real-time predictions of complex dynamics in physical systems. *Nature Communications*, 15(1):5101, 2024.

Chia-Tung Kuo and Ian Davidson. A framework for outlier description using constraint programming. In *Proceedings of the 2016 Conference on Artificial Intelligence, AAAI*, pp. 1237–1243, 2016.

Zechao Li, Jinhui Tang, and Xiaofei He. Robust structured nonnegative matrix factorization for image representation. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1947–1960, 2018.

Zechao Li, Jinhui Tang, and Tao Mei. Deep collaborative embedding for social image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2070–2083, 2019.

Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. COPOD: copula-based outlier detection. In *Proceedings of the 2020 International Conference on Data Mining, ICDM*, pp. 1118–1123, 2020.

Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George H Chen. ECOD: unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12181–12193, 2022.

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

Ninghao Liu, Donghwa Shin, and Xia Hu. Contextual outlier interpretation. In *Proceedings of the 2018 International Joint Conference on Artificial Intelligence, IJCAI*, pp. 2461–2467, 2018.

Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xie Xing. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 2011 International Conference on Knowledge Discovery and Data Mining, SIGKDD*, pp. 1010–1018, 2011.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 2017 Advances in Neural Information Processing Systems, NeuraIPS*, pp. 4765–4774, 2017.

Meghanath Macha and Leman Akoglu. Explaining anomalies in groups with characterizing subspace rules. *Data Mining and Knowledge Discovery*, 32(5):1444–1480, 2018.

Egawati Panjei, Le Gruenwald, Eleazar Leal, Christopher Nguyen, and Shejuti Silvia. A survey on outlier explanations. *The VLDB Journal*, 31(5):977–1008, 2022.

Elaheh Panjei, Le Gruenwald, Ermeson Leal, Cuong Nguyen, and Sheena Silvia. Outlier explanations for data streams: Outlying attributes and root cause factors of outliers. *Data & Knowledge Engineering*, 154:102349, 2024.

Phu Pham, Quang-Thinh Bui, Ngoc Thanh Nguyen, Robert Kozma, Philip S. Yu, and Bay Vo. Topological data analysis in graph neural networks: surveys and perspectives. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):9758–9776, 2025.

Kangjia Qiao, Jing Liang, Wei-Feng Guo, Yunpeng Wei, Kunjie Yu, and Zhuo Hu. Constrained multi-objective optimization-based temporal network observability for biomarker identification of individual patients. *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2024.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 2016 International Conference on Knowledge Discovery and Data Mining, SIGKDD*, pp. 1135–1144, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: high-precision model-agnostic explanations. In *Proceedings of the 2018 conference on artificial intelligence, AAAI*, volume 32, 2018.

Durgesh Samariya and Jiangang Ma. A new dimensionality-unbiased score for efficient and effective outlying aspect mining. *Data Science and Engineering*, 7(2):120–135, 2022.

Peng Tang, Xiaoxiao Yan, Xiaobin Hu, Kai Wu, Tobias Lasser, and Kuangyu Shi. Anomaly detection in medical images using encoder-attention-2decoders reconstruction. *IEEE Transactions on Medical Imaging*, 44(8):3370–3382, 2025.

Seyed Mohamad Ali Tousi and GN DeSouza. Outlier interpretation using regularized auto encoders and genetic algorithm. In *Proceedings of the 2024 IEEE Congress on Evolutionary Computation, CEC*, pp. 1–8, 2024.

Che-Ping Tsai, Ganyu Teng, Phillip Wallis, and Wei Ding. AnoLLM: large language models for tabular anomaly detection. In *Proceedings of the 2025 International Conference on Learning Representations, ICLR*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 2017 Advances in Neural Information Processing Systems, NeuraIPS*, pp. 5998–6008, 2017.

Nguyen Xuan Vinh, Jeffrey Chan, Simone Romano, James Bailey, Christopher Leckie, Kotagiri Ramamohanarao, and Jian Pei. Discovering outlying aspects in large datasets. *Data Mining and Knowledge Discovery*, 30(6):1520–1555, 2016.

Gang Wang, Yisheng Zou, Songlin He, Yakun Wang, and Ruihong Dai. Anomaly detection and localization via reverse distillation with latent anomaly suppression. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Yanting Wang and Jinyuan Jia. EnsembleSHAP: faithful and certifiably robust attribution for random subspace method. In *Proceedings of the 2026 International Conference on Learning Representations, ICLR*, 2026.

Lei Xing, Wenjun Wang, Guixiang Xue, Hao Yu, Xiaotong Chi, and Weidi Dai. Discovering traffic outlier causal relationship based on anomalous dag. *Advances in Swarm and Computational Intelligence*, pp. 71–80, 2015.

Hongzuo Xu, Yijie Wang, Songlei Jian, Zhenyu Huang, Yongjun Wang, Ning Liu, and Fei Li. Beyond outlier detection: outlier interpretation by attention-guided triplet deviation network. In *Proceedings of the 2021 Web Conference, WWW*, pp. 1328–1339, 2021.

Shicheng Xu, Wei Li, Zuoyong Li, Tiesong Zhao, and Bob Zhang. Facing differences of similarity: Intra- and inter-correlation unsupervised learning for chest x-ray anomaly detection. *IEEE Transactions on Medical Imaging*, 44(2):801–814, 2025.

Xing Yang, Pan Huang, Le An, Peng Feng, Biao Wei, Peng He, and Kexin Peng. A growing model-based OCSVM for abnormal student activity detection from daily campus consumption. *New Generation Computing*, 40(4):915–933, 2022.

Hangting Ye, He Zhao, Wei Fan, Mingyuan Zhou, Dandan Guo, and Yi Chang. DRL: decomposed representation learning for tabular anomaly detection. In *Proceedings of the 2025 International Conference on Learning Representations, ICLR*, 2025.

Sanghyu Yoon, Dongmin Kim, Suhee Yoon, Ye Seul Sim, Seungdong Yoa, Hye-Seung Cho, Soonyoung Lee, Hankook Lee, and Woohyung Lim. ReTabAD: a benchmark for restoring semantic context in tabular anomaly detection. In *Proceedings of the 2026 International Conference on Learning Representations, ICLR*, 2026.

EungGu Yun, Heonjin Ha, Yeongwoo Nam, and Bryan Dongik Lee. Language-assisted feature transformation for anomaly detection. In *Proceedings of the 2025 International Conference on Learning Representations, ICLR*, 2025.

Jianxiong Zhou and Ying Wu. Outlier-probability-based feature adaptation for robust unsupervised anomaly detection on contaminated training data. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

A Detailed related work

The following method-by-method survey provides additional context for the tabular comparisons.

Score-and-search-based Methods. Score-and-search-based outlier interpretation methods primarily aim to identify feature subspaces that maximize an outlyingness score. Representative methods include: OAMiner Duan et al. (2015), which uses kernel density estimation and an anti-monotonicity-based pruning strategy; x-PACS Macha & Akoglu (2018), which discovers characterizing subspace rules that differentiate anomalous groups; Other approaches employ specific scoring functions, such as density z-score and isolation path length Vinh et al. (2016), or leverage techniques like refining random subspaces based on score discrepancy, employing CP languages Kuo & Davidson (2016), or using nearest neighbor sets with beam search (SiNNE Samariya & Ma (2022)). However, these methods face two key limitations: (i) they search within the input feature space without accounting for inter-dimensional correlations, and (ii) due to the vast search space, they often resort to constraints like maximum subspace length and pruning strategies, which can lead to suboptimal interpretations.

Feature-selection-based Methods. Existing feature-selection-based outlier interpretation methods employ feature selection or transformation techniques to explain outliers. COIN Liu et al. (2018) reformulates interpretation as a classification problem using l_1-norm classifiers to separate outliers from inliers. EAC-FS Cai et al. (2026) is the first unified multi-agent reinforcement learning formulation in which feature grouping, transfer, and selection evolve jointly within a single sequential decision process. ATON Xu et al. (2021) transforms features into an embedding space using triplet deviation loss to emphasize outlier-inlier separation. The Adaptive Multiscale Attention (AMA) mechanism Dentamaro et al. (2025) learns multi-scale compressed feature representations through parallelized excitation layers. WANDER Guan et al. (2025) captures geometric structures via weighted digraphs to learn optimal subspaces. Related interpretable machine learning methods include LIME Ribeiro et al. (2016), which approximates local model behavior; SHAP Lundberg & Lee (2017), which computes Shapley values for feature contributions; and Anchors Ribeiro et al. (2018), which generates rule-based explanations. However, these methods exhibit two key shortcomings: (i) they may overlook complex interdependencies among features, leading to inaccurate interpretations, and (ii) their interpretation instability can limit scalability.

Latent Space Learning Methods. Latent space learning techniques aim to discover meaningful representations that capture underlying data structures. DCE Li et al. (2019) learns a unified latent space to bridge images and tags, while subsequent work Li et al. (2018) employs NMF to ensure

class-consistent embeddings. GLO Bojanowski et al. (2018) trains deep convolutional generators using reconstruction losses. L-DeepONet Kontolati et al. (2024) leverages operator networks in low-dimensional latent spaces for efficient function approximation. This approach has been extended to diverse modalities: topological data analysis integrated with graph neural networks captures multi-scale structural features Pham et al. (2025); heterogeneous graph collaborative representation learning predicts drug-microbe interactions Guo et al. (2025); and constrained multi-objective optimization frameworks encode dynamic patterns for personalized biomarker identification Qiao et al. (2024). However, these methods remain primarily oriented toward image processing, reconstruction, or general prediction tasks. They lack specific architectural components for outlier interpretation, particularly the capability to simultaneously model outlier-inlier discrepancies and inter-outlier distinctions.

Causal-influence-based Explanation Methods. Recent years have witnessed growing interest in causal-influence-based explanation for outlier interpretation. These methods aim to identify causal relationships rather than mere correlations, seeking to uncover the root causes of outliers in dynamic and temporal data. STOTree Liu et al. (2011) focuses on discovering spatio-temporal causal interactions in traffic data streams. Building on these foundations, DAG Xing et al. (2015) develops techniques for discovering traffic outlier causal relationships based on anomalous directed acyclic graphs, and EXOS Panjei et al. (2024) provides comprehensive frameworks for explaining outliers in data streams by identifying both outlying attributes and root cause factors. However, despite their strengths in temporal domains, these causal-influence-based methods face significant challenges when applied to static tabular data scenarios: (i) they depend on strong causal assumptions and domain knowledge that may not be available in many real contexts; and (ii) their computational complexity can be prohibitive for high-dimensional data where temporal dependencies are absent.

Model-based detection methods and their limits for interpretation. A complementary family detects anomalies with an explicit model of normality and then reads off the deviation as an explanation. One-class and growing one-class SVM variants fit a boundary around normal behavior; for instance, a growing model-based OCSVM flags abnormal student activity from daily campus consumption records by updating the normal-behavior boundary as new card data arrive Yang et al. (2022). Their strength is deployment simplicity and a well-defined score, and the incremental update keeps the model current in streaming settings. Their weakness for interpretation is that the score is a scalar distance to a learned boundary: it ranks instances by abnormality but does not by itself select the features responsible, so a post-hoc step (per-feature deviation, gradient, or subset search) is still required to produce a readable explanation. Reconstruction- and density-based detectors share this limitation, which is why TMQ treats detector output as labels and reference sets (Sec. 3.1) and learns the feature-level contrast separately.

B FORMAL PROBLEM AND NOTATION

Let $\mathcal{D}$ be a d-dimensional input space. Given a set of outliers $\mathbf{X}_+$ and a set of inliers $\mathbf{X}_-$ in the full space $\mathcal{D}$. Each instance $\boldsymbol{x} \in \mathbf{X}_+ \cup \mathbf{X}_-$ is represented as a vector in the form $[\boldsymbol{x}(1), \cdots, \boldsymbol{x}(d)]^\top$. Typically, we have $|\mathbf{X}_+| \ll |\mathbf{X}_-|$. Considering relationships among the dimensions within the input space $\mathcal{D}$, we define a mapping function to project all instances from $\mathcal{D}$ into a latent space $\mathcal{S}$. For a query outlier $\boldsymbol{x} \in \mathbf{X}_+$, we use *inlyingness degree* $ID_{\mathcal{S}}(\boldsymbol{x} \mid \mathbf{X}_+ \cup \mathbf{X}_-)$ to measure its outlyingness, compared to other outliers in $\mathbf{X}_+$ and the inliers in $\mathbf{X}_-$ in the latent space $\mathcal{S}$. Without loss of generality, we assume that the lower the value of $ID_{\mathcal{S}}(\boldsymbol{x} \mid \mathbf{X}_+ \cup \mathbf{X}_-)$ is, the more outlying $\boldsymbol{x}$ becomes. We formalize the outlier interpretation task as follows:

Definition 1. *Given $\mathbf{X}_+$ and $\mathbf{X}_-$ in a d-dimensional space $\mathcal{D}$, and a query outlier $\boldsymbol{x} \in \mathbf{X}_+$. The interpretation for $\boldsymbol{x}$ is to learn an optimal latent space $\mathcal{S}$, in which, the inlyingness degree of $\boldsymbol{x}$, $ID_{\mathcal{S}}(\boldsymbol{x} \mid \mathbf{X}_+ \cup \mathbf{X}_-)$, is minimized, i.e.,*

$$\arg\min_{\mathcal{S}} ID_{\mathcal{S}}(\boldsymbol{x} \mid \mathbf{X}_+ \cup \mathbf{X}_-) \tag{6}$$

The main notations used in this paper are listed in Table 6.

Table 6: Summary of main notations.

Notation	Description
$\mathcal{D}$	A d-dimensional input space
$\mathbf{X}_+$	A set of outliers
$\mathbf{X}_-$	A set of inliers
$\mathcal{S}$	A latent space
$\boldsymbol{x}$	A query outlier in $\mathbf{X}_+$
$ID_{\mathcal{S}}(\boldsymbol{x} \mid \mathbf{X}_+ \cup \mathbf{X}_-)$	Inlyingness degree of $\boldsymbol{x}$ in $\mathcal{S}$
$ID^*_{\mathcal{S}}(\boldsymbol{x} \mid \mathbf{X}_+ \cup \mathbf{X}_-)$	The minimised inlyingness degree of $\boldsymbol{x}$ in $\mathcal{S}$
$N_{k_1}(\boldsymbol{x}, \mathbf{X}_+)$	k_1-nearest neighbors of $\boldsymbol{x}$ in $\mathbf{X}_+$
$N_{k_2}(\boldsymbol{x}, \mathbf{X}_-)$	k_2-nearest neighbors of $\boldsymbol{x}$ in $\mathbf{X}_-$
$\widehat{N}_{k_2}(\boldsymbol{x}, \mathbf{X}_-)$	k_2 randomly sampled inliers from $\mathbf{X}_- \setminus N_{k_2}(\boldsymbol{x}, \mathbf{X}_-)$
$\boldsymbol{x}_{out}$	An outlier from $N_{k_1}(\boldsymbol{x}, \mathbf{X}_+)$
$\boldsymbol{x}_{in}$	An inlier from $N_{k_2}(\boldsymbol{x}, \mathbf{X}_-)$
$\boldsymbol{x}'_{in}$	An inlier from $\widehat{N}_{k_2}(\boldsymbol{x}, \mathbf{X}_-)$
$\mathbf{Q}$	A set of quadruplets
$\varphi(\mathbf{Q})$	An embedded quadruplet set
$\widetilde{\mathbf{Q}}$	A transformer-guided quadruplet set

C Full TMQ network specification

C.1 Design of TMQ

The overall network architecture of TMQ is illustrated in Fig. 2, which comprises four modules for learning a query-conditioned projection and relational contrast.

C.1.1 The Quadruplet Generation Module

A set of quadruplets is initially generated as training instances. These quadruplets are utilized to learn the separability of $\boldsymbol{x}$ from other outliers and inliers. Thus, the first quadruplet position is fixed as $\boldsymbol{x}$, and the remaining three positions should well represent other outliers and inliers. We define the set of k_1-nearest neighbors of $\boldsymbol{x}$ in $\mathbf{X}_+$ as $N_{k_1}(\boldsymbol{x}, \mathbf{X}_+)$, and the set of k_2-nearest neighbors of $\boldsymbol{x}$ in $\mathbf{X}_-$ as $N_{k_2}(\boldsymbol{x}, \mathbf{X}_-)$. The set of generated quadruplets $\mathbf{Q}$ is denoted as:

$$\mathbf{Q} = \{(\boldsymbol{x}, \boldsymbol{x}_{out}, \boldsymbol{x}_{in}, \boldsymbol{x}'_{in}) | \boldsymbol{x}_{out} \in N_{k_1}(\boldsymbol{x}, \mathbf{X}_+), \boldsymbol{x}_{in} \in N_{k_2}(\boldsymbol{x}, \mathbf{X}_-), \boldsymbol{x}'_{in} \in \widehat{N}_{k_2}(\boldsymbol{x}, \mathbf{X}_-)\} \tag{7}$$

where $\widehat{N}_{k_2}(\boldsymbol{x}, \mathbf{X}_-)$ denotes a candidate set of inliers randomly sampled from the $\mathbf{X}_-$ without $N_{k_2}(\boldsymbol{x}, \mathbf{X}_-)$ elements. $\boldsymbol{x}_{out}$ is selected from $N_{k_1}(\boldsymbol{x}, \mathbf{X}_+)$ to capture the abnormality of $\boldsymbol{x}$ from other outliers. Similarly, $\boldsymbol{x}_{in}$ is selected from $N_{k_2}(\boldsymbol{x}, \mathbf{X}_-)$ to obtain the abnormality of $\boldsymbol{x}$ from inliers. Furthermore, $\boldsymbol{x}'_{in}$ is randomly sampled from $\widehat{N}_{k_2}(\boldsymbol{x}, \mathbf{X}_-)$ to represent the normality of inliers. Consequently, the size of the generated quadruplet set $\mathbf{Q}$ is $k_1 k_2^2$. Abundant training instances normally bring better performance, but excessively large k_1 and k_2 also result in lower efficiency. We conduct a parameter test in Section G.2 to investigate the effect of k_1 and k_2.

C.1.2 The Latent Space Learning Module

The embedding layer. The embedding layer $\varphi(\cdot)$ utilizes a learnable linear layer for representation, parameterized by the weight matrix $\boldsymbol{W} \in \mathbb{R}^{|\mathcal{S}| \times d}$, to obtain a powerful embedding space with strong representational capacity. This simple transformation can also ensure the scalability of the model.

All instances in the generated quadruplets are embedded into this embedding space to derive an embedded quadruplet set $\varphi(\mathbf{Q})$ as:

$$\varphi(\mathbf{Q}) = \{(\varphi(\boldsymbol{x}), \varphi(\boldsymbol{x}_{out}), \varphi(\boldsymbol{x}_{in}), \varphi(\boldsymbol{x}'_{in})) | \boldsymbol{x}, \boldsymbol{x}_{out}, \boldsymbol{x}_{in}, \boldsymbol{x}'_{in} \in \mathbf{Q}\} \tag{8}$$

The global-local attention mechanism. Attention mechanism Vaswani et al. (2017) has become a de facto standard in many sequence-based tasks, such as language modeling and machine translation, for focusing on the most relevant input parts to make decisions. The primary objective of this mechanism is to measure the contributions of various features for $\boldsymbol{x}$ in the embedding space through feature interaction. The adoption of the global-local attention mechanism enables the model to automatically highlight features in different contexts through dynamic weight allocation, which allows attention parameters to depend on the sampled context.

Within the global-local attention mechanism, the global attention mechanism is composed of a four-head attention mechanism denoted as $A_{global}(\cdot)$, to capture the global dependencies of each element in $\varphi(\mathbf{Q})$. The local attention mechanism consists of four three-head attention mechanisms, each of which is denoted as $A_{local}(\cdot)$, for the purpose of capturing the local dependencies of each element in $\varphi(\mathbf{Q})$. In order to feed into one global attention mechanism and four local attention mechanisms respectively, each quadruplet $(\varphi(\boldsymbol{x}), \varphi(\boldsymbol{x}_{out}), \varphi(\boldsymbol{x}_{in}), \varphi(\boldsymbol{x}'_{in}))$ in $\varphi(\mathbf{Q})$ is first flattened into five vectors as shown below

$$\begin{aligned}
\boldsymbol{g} &= \text{concat}(\varphi(\boldsymbol{x}), \varphi(\boldsymbol{x}_{out}), \varphi(\boldsymbol{x}_{in}), \varphi(\boldsymbol{x}'_{in})) \\
\boldsymbol{l}^{(1)} &= \text{concat}(\varphi(\boldsymbol{x}_{out}), \varphi(\boldsymbol{x}_{in}), \varphi(\boldsymbol{x}'_{in})) \\
\boldsymbol{l}^{(2)} &= \text{concat}(\varphi(\boldsymbol{x}), \varphi(\boldsymbol{x}_{in}), \varphi(\boldsymbol{x}'_{in})) \\
\boldsymbol{l}^{(3)} &= \text{concat}(\varphi(\boldsymbol{x}), \varphi(\boldsymbol{x}_{out}), \varphi(\boldsymbol{x}_{in})) \\
\boldsymbol{l}^{(4)} &= \text{concat}(\varphi(\boldsymbol{x}), \varphi(\boldsymbol{x}_{out}), \varphi(\boldsymbol{x}'_{in}))
\end{aligned} \tag{9}$$

where $\boldsymbol{g} \in \mathbb{R}^{4|\mathcal{S}|\times 1}$, $\boldsymbol{l}^{(1)}, \boldsymbol{l}^{(2)}, \boldsymbol{l}^{(3)}, \boldsymbol{l}^{(4)} \in \mathbb{R}^{3|\mathcal{S}|\times 1}$. We obtain a set of flattened vectors with sizes of $k_1k_2^2$, $k_1k_2^2$, k_2^2, k_1k_2, and k_1k_2, respectively. One global matrix $\boldsymbol{G} = \left[\boldsymbol{g}_1, \cdots, \boldsymbol{g}_{k_1k_2^2}\right]$, and four local matrices $\boldsymbol{L}_1 = \left[\boldsymbol{l}_1^{(1)}, \cdots, \boldsymbol{l}_{k_1k_2^2}^{(1)}\right]$, $\boldsymbol{L}_2 = \left[\boldsymbol{l}_1^{(2)}, \cdots, \boldsymbol{l}_{3|\mathcal{S}|\times k_2^2}^{(2)}\right]$, $\boldsymbol{L}_3 = \left[\boldsymbol{l}_1^{(3)}, \cdots, \boldsymbol{l}_{k_1k_2}^{(3)}\right]$, and $\boldsymbol{L}_4 = \left[\boldsymbol{l}_1^{(4)}, \cdots, \boldsymbol{l}_{k_1k_2}^{(4)}\right]$ are generated by the above five vectors obtained from the quadruplet sets, respectively. We then input $\boldsymbol{G}$ into the global attention mechanism $A_{global}(\cdot)$, resulting in the following global attention matrix:

$$A_{global}(\boldsymbol{G}) = \text{concat}(\text{head}_1, \cdots, \text{head}_4) \cdot \boldsymbol{W}_{\boldsymbol{G}} \tag{10}$$

where $\boldsymbol{W}_{\boldsymbol{G}}$ is the parameter matrix of $A_{global}(\boldsymbol{G})$. The i-th head is denoted by head_i:

$$\text{head}_i = \text{softmax}\left(\frac{(\boldsymbol{G}\boldsymbol{H}_i)(\boldsymbol{G}\boldsymbol{H}_i)^\top}{\sqrt{4|\mathcal{S}|}}\right)\boldsymbol{G}\boldsymbol{H}_i \tag{11}$$

where $\boldsymbol{H}_i$ denotes the parameter matrix of self-attention within the i-th head in $A_{global}(\boldsymbol{G})$. Subsequently, $\boldsymbol{L}_1$, $\boldsymbol{L}_2$, $\boldsymbol{L}_3$, and $\boldsymbol{L}_4$ are fed into the local attention $A_{local}(\cdot)$. Each local attention matrix is denoted as:

$$A_{local}(\boldsymbol{L}_i) = \text{concat}\left(\text{head}_1^{(i)}, \text{head}_2^{(i)}, \text{head}_3^{(i)}\right) \cdot \boldsymbol{W}_{\boldsymbol{L}_i} \tag{12}$$

where $\boldsymbol{W}_{\boldsymbol{L}_i}$ is the parameter matrix of $A_{local}(\boldsymbol{L}_i)$. The j-th head of the i-th local matrix is denoted by $\text{head}_j^{(i)}$:

$$\text{head}_j^{(i)} = \text{softmax}\left(\frac{(\boldsymbol{L}_i\widetilde{\boldsymbol{H}}_j)(\boldsymbol{L}_i\widetilde{\boldsymbol{H}}_j)^\top}{\sqrt{3|\mathcal{S}|}}\right)\boldsymbol{L}_i\widetilde{\boldsymbol{H}}_j \tag{13}$$

where $\widetilde{\boldsymbol{H}}_j$ denotes the parameter matrix of self-attention within the j-th head in $A_{local}(\boldsymbol{L}_i)$.

The feed-forward network. This network contributes to introducing nonlinear transformations and enhancing the modeling capacity of the model towards features. This network consists of two single-hidden-layer feedforward neural networks, $F_{global}(\cdot)$ and $F_{local}(\cdot)$. The global attention mechanism $A_{global}(\boldsymbol{G})$ integrates $F_{global}(\cdot)$ with hidden units, while the local attention mechanism $A_{local}(\boldsymbol{L}_i)$ combines $F_{local}(\cdot)$ with hidden units. Furthermore, this network is connected with the global-local attention mechanism, generating the global attention score:

$$Score_{global} = \delta\big(F_{global}(A_{global}(\boldsymbol{G}))\big) \tag{14}$$

and the local attention score:

$$Score_{local} = \sum_{i=1}^{4} \delta\big(F_{local}(A_{local}(\boldsymbol{L}_i))\big) \tag{15}$$

where $\delta(\cdot)$ is a column-wise min-max normalization function used to ensure that the attention scores are within the range of $[0, 1]$. Based on these scores, given the function with computing the average of the columns $\rho(\cdot)$, we can get the full contribution vector $\boldsymbol{a}$ computed through this network in conjunction with the global-local attention module:

$$\boldsymbol{a} = \rho(Score_{global}) + \rho(Score_{local}) \tag{16}$$

$\boldsymbol{a}$ is directly derived from the quadruplets. The embedding layer, global-local attention mechanism, and feedforward network capture the hierarchical and complex relationships among the features in $\boldsymbol{x}$, other outliers, and the inliers. This is achieved by leveraging the connectivity property and representational capacity of the neural network and attention mechanism. All instances are transformed into a latent space to derive a transformer-guided quadruplet set $\widetilde{\mathbf{Q}}$ as:

$$\begin{aligned}\widetilde{\mathbf{Q}} = \{(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}_{out}, \widetilde{\boldsymbol{x}}_{in}, \widetilde{\boldsymbol{x}}'_{in}) | \widetilde{\boldsymbol{x}} = \boldsymbol{a} \odot \varphi(\boldsymbol{x}), \widetilde{\boldsymbol{x}}_{out} = \boldsymbol{a} \odot \varphi(\boldsymbol{x}_{out}),\\ \widetilde{\boldsymbol{x}}_{in} = \boldsymbol{a} \odot \varphi(\boldsymbol{x}_{in}), \widetilde{\boldsymbol{x}}'_{in} = \boldsymbol{a} \odot \varphi(\boldsymbol{x}'_{in})\}\end{aligned} \tag{17}$$

where $\odot$ denotes element-wise product.

C.1.3 The Outlying Measure Module

The previous modules are learnable after setting an objective. The objective is to find an optimal latent space that can effectively differentiate among the query outlier, other outliers, and the inliers. Following the idea of Kendall et al. (2018), we formally define the quadruplet-based loss as the inlyingness degree function:

$$ID_{\mathcal{S}}(\boldsymbol{x} \mid \mathbf{X}_+ \cup \mathbf{X}_-) = \sum_{\boldsymbol{q} \in \widetilde{\mathbf{Q}}} (f_1(\boldsymbol{q}) + f_2(\boldsymbol{q})) \tag{18}$$

where

$$f_1(\boldsymbol{q}) = \underbrace{\max\left(\|\widetilde{\boldsymbol{x}}_{in} - \widetilde{\boldsymbol{x}}'_{in}\|_2^2 - \|\widetilde{\boldsymbol{x}}_{out} - \widetilde{\boldsymbol{x}}\|_2^2 + e, 0\right)}_{\textbf{Specificity of the query outlier}} \tag{19}$$

$$\begin{aligned}f_2(\boldsymbol{q}) = &\underbrace{\max\left(\|\widetilde{\boldsymbol{x}}_{in} - \widetilde{\boldsymbol{x}}'_{in}\|_2^2 - \|\widetilde{\boldsymbol{x}}_{in} - \widetilde{\boldsymbol{x}}\|_2^2 + e, 0\right)}_{\textbf{inlier-outlier discrepancy}}\\ &+ \underbrace{\max\left(\|\widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}_{out}\|_2^2 - \|\widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}_{in}\|_2^2 + e, 0\right)}_{\textbf{consistency among outliers}}\end{aligned} \tag{20}$$

, e is the margin. The goal of $f_1(\boldsymbol{q})$ is to ensure that the query outlier $\widetilde{\boldsymbol{x}}$ remains distant from other outliers $\widetilde{\boldsymbol{x}}_{out}$, while maintaining the compactness of the distribution among inliers. $f_1(\boldsymbol{q})$ effectively penalizes the similarity between outliers, thus emphasizing the specificity of the query outliers. The objective of the first term in $f_2(\boldsymbol{q})$ is to push the query outlier $\widetilde{\boldsymbol{x}}$ away from inliers, preventing it from blending into the distribution of inliers. The objective of the second term in $f_2(\boldsymbol{q})$ is to ensure that $\widetilde{\boldsymbol{x}}$ is closer to other outliers than to inliers in the latent space, thereby reinforcing the internal consistency among outliers. As illustrated in Fig. 1(c), $f_1(\boldsymbol{q})$ separates the query outlier ("★") from other outliers ("►"), while $f_2(\boldsymbol{q})$ separates outliers ("★" and "►") from all inliers ("•").

The Adam optimizer is employed to optimize the TMQ by iteratively adjusting the parameters to minimize $ID_{\mathcal{S}}(\boldsymbol{x} \mid \mathbf{X}_+ \cup \mathbf{X}_-)$. The epoch number and batch size are considered as hyper-parameters of TMQ.

C.1.4 The Interpretation Acquisition Module

After the outlying measure module, TMQ obtains the optimized weights $\boldsymbol{W}^*$ and the optimized contribution vector $\boldsymbol{a}^*$ in the latent space learning module. The optimal latent space mapped by $\phi(\cdot)$ is represented as: $\phi(\cdot) : \mathcal{D} \to \mathcal{S}$.

Algorithm 2 *TMQ*

Input: A query outlier: $\boldsymbol{x}$, a set of outliers: $\mathbf{X}_+$,
a set of inliers: $\mathbf{X}_-$
Output: A latent space $\mathcal{S}$

Generate $N_{k_1}(\boldsymbol{x}, \mathbf{X}_+)$, $N_{k_2}(\boldsymbol{x}, \mathbf{X}_-)$, $\widehat{N}_{k_2}(\boldsymbol{x}, \mathbf{X}_-)$;
Generate a set of quadruplets $\mathbf{Q}$ by Equation (7);
Initialize $\boldsymbol{W}$, $\boldsymbol{W_G}$, $\boldsymbol{W}_{\boldsymbol{L}_i}$, $\boldsymbol{H}_i$, $\widetilde{\boldsymbol{H}}_j$;
repeat
    Generate $\varphi(\mathbf{Q})$ by Equation (8);
    Flatten quadruplets to get matrices $\boldsymbol{G}$, $\boldsymbol{L}_1$, $\boldsymbol{L}_2$, $\boldsymbol{L}_3$, and $\boldsymbol{L}_4$;
    Compute $Score_{global}$ by Equation (14);
    Compute $Score_{local}$ by Equation (15);
    Compute $\boldsymbol{a}$ by Equation (16);
    Generate $\widetilde{\mathbf{Q}}$ by Equation (17);
    Optimize network parameters by the quadruplet-based loss $ID_{\mathcal{S}}(\boldsymbol{x} \mid \mathbf{X}_+ \cup \mathbf{X}_-)$ in Equation (18);
until Reach maximum epoch number
Obtain a latent space $\mathcal{S}$;
Return $\mathcal{S}$

C.2 Quadruplet-based Loss vs Triplet Loss

Traditional Triplet Loss Limitations. Existing triplet-based methods such as ATON Xu et al. (2021) employ a triplet loss:

$$L_{\text{triplet}} = \max\left(\|\widetilde{\boldsymbol{x}}_{in} - \widetilde{\boldsymbol{x}}'_{in}\|_2^2 - \|\widetilde{\boldsymbol{x}}_{in} - \widetilde{\boldsymbol{x}}\|_2^2 + e, 0\right) \tag{21}$$

This formulation enforces $\widetilde{\boldsymbol{x}}_{in}$ to be closer to inliers than to the query outlier $\widetilde{\boldsymbol{x}}$, thereby separating outliers from inliers. However, as illustrated in Fig. 3(a), this design critically lacks explicit outlier-outlier separation constraints. In real-world applications, multiple distinct outlier subtypes often coexist within datasets. For instance, in credit card fraud detection, various fraudulent behaviors such as stolen card transactions, account takeovers, and promotional abuse may manifest with different characteristic patterns. The triplet loss cannot effectively distinguish between these different outlier subtypes, potentially grouping disparate anomaly types into the same latent representation. This conflation leads to ambiguous and non-specific interpretations, as the model fails to capture the nuanced differences between outlier categories.

Since the triplet loss function does not incorporate distance constraints between different outliers, the learning process does not actively differentiate the feature characteristics that distinguish various anomaly types. Consequently, during outlier interpretation, the method cannot provide unique, targeted explanations for different outlier categories.

TMQ's Quadruplet-based Loss Advantages. To address these limitations, TMQ introduces a quadruplet-based loss with enhanced relational constraints, extending the traditional triplet structure to (query outlier, other outlier, near inlier, random inlier) as shown in Fig. 3(b). The quadruplet loss incorporates two additional relational components beyond the triplet loss: specifically, the outlier-outlier discrimination term $f_1(\boldsymbol{q})$ explicitly penalizes similarity between the query outlier $\widetilde{\boldsymbol{x}}$ and other outliers $\widetilde{\boldsymbol{x}}_{out}$ to separate distinct anomaly subtypes in the latent space, while the hierarchical relationship term in $f_2(\boldsymbol{q})$ ensures that different outliers remain collectively closer to each other than to inliers even as they are pushed apart, thereby preserving fundamental outlier compactness relative to inliers while enabling effective subtype differentiation.

When the extra relation is appropriate. The inter-outlier term provides a subtype-specific constraint when the sampled outlier represents a meaningfully different mechanism. It is not a universal superiority result: if the pair shares an anomaly subtype, pushing it apart can conflict with a faithful representation. Proposition 1 characterizes the exact zero-loss distance interval and Appendix I derives its limitation. The reported triplet ablation tests the overall objective under the available detector-proxy annotations, rather than proving an interpretation guarantee.

Table 7: Description of the 12 benchmark datasets.

Datasets	**#Instances**	**#Features**	**#Outliers**	**%Rate**
wpbc	198	33	47	23.74
glass	214	7	9	4.21
vertebral	240	6	30	12.50
yeast	1484	8	507	34.16
fault	1941	27	673	34.67
pageblocks	5393	10	510	9.46
satimage-2	5803	36	71	1.22
annthyroid	7200	6	534	7.42
mnist	10000	512	500	5.00
smtp	95156	3	30	0.03
mammography	11183	6	260	2.32
fraud	284807	29	492	0.17

D COMPLETE EVALUATION TABLES

E FULL DETECTOR-ANNOTATION RESULT MATRICES

The numerical entries below are copied from the original manuscript tables. Each row identifies a dataset and annotation detector. $--$ denotes a missing or timed-out result. Tables are split by metric to keep values legible in the ICLR page width.

Table 8: AUPR on all available detector annotations (original reported values).

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
glass	iforest	0.579	0.821	0.829	0.577	0.400	0.801
glass	copod	0.630	0.673	0.656	0.598	0.515	0.812
glass	hbos	0.497	0.751	0.782	0.424	0.435	0.667
glass	ecod	0.703	0.566	0.819	0.664	0.433	0.774
glass	lunar	0.615	0.662	0.759	0.489	0.296	0.722
glass	rod	0.165	0.071	0.238	0.232	0.296	0.279
glass	mcd	0.508	0.607	0.589	0.210	0.444	0.597
wpbc	iforest	0.866	0.746	0.644	0.594	0.539	0.872
wpbc	copod	0.736	0.672	0.604	0.599	0.497	0.760
wpbc	hbos	0.817	0.697	0.639	0.577	0.505	0.826
wpbc	ecod	0.817	0.327	0.520	0.484	0.492	0.832
wpbc	lunar	0.724	0.455	0.654	0.613	0.469	0.742
wpbc	rod	0.818	0.417	0.608	0.607	0.536	0.822
wpbc	mcd	0.812	0.388	0.649	0.640	0.469	0.845
vertebral	iforest	0.716	0.502	0.434	0.452	0.356	0.791
vertebral	copod	0.759	0.491	0.538	0.600	0.384	0.804
vertebral	hbos	0.559	0.357	0.240	0.314	0.455	0.603
vertebral	ecod	0.837	0.415	0.680	0.728	0.451	0.849
vertebral	lunar	0.513	0.283	0.303	0.308	0.529	0.527
vertebral	rod	0.851	0.669	0.859	0.856	0.467	0.829
vertebral	mcd	0.734	0.492	0.390	0.462	0.504	0.787
pageblocks	iforest	0.632	0.334	0.221	0.165	0.194	0.735
pageblocks	copod	0.620	0.351	0.262	0.227	0.236	0.673
pageblocks	hbos	0.576	0.264	0.265	0.211	0.256	0.582
pageblocks	ecod	0.571	0.386	0.450	0.408	0.437	0.588
pageblocks	lunar	0.608	0.449	0.346	0.392	0.434	0.653
pageblocks	rod	0.627	0.483	0.467	0.507	0.414	0.641
pageblocks	mcd	0.483	0.386	0.210	0.125	0.086	0.593
fault	iforest	0.643	0.602	0.558	0.502	0.387	0.697
fault	copod	0.668	0.602	0.561	0.528	0.435	0.706
fault	hbos	0.638	0.589	0.527	0.448	0.345	0.679

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
fault	ecod	0.734	0.443	0.591	0.546	0.459	0.776
fault	lunar	0.618	0.530	0.578	0.534	0.441	0.627
fault	rod	0.681	0.429	0.561	0.537	0.497	0.709
fault	mcd	0.573	0.592	0.596	0.564	0.505	0.596
yeast	iforest	0.655	0.589	0.407	0.252	0.553	0.698
yeast	copod	0.676	0.606	0.372	0.294	0.595	0.695
yeast	hbos	0.573	0.575	0.269	0.159	0.543	0.607
yeast	ecod	0.768	0.418	0.473	0.309	0.584	0.800
yeast	lunar	0.703	0.587	0.514	0.391	0.518	0.714
yeast	rod	0.080	0.074	0.093	0.088	0.281	0.077
MNIST	iforest	0.690	0.647	0.736	0.654	0.450	0.728
MNIST	copod	0.702	0.670	0.745	0.680	0.453	0.732
MNIST	hbos	0.654	0.593	0.762	0.655	0.444	0.684
MNIST	ecod	0.706	0.665	0.734	0.709	0.446	0.737
MNIST	lunar	0.664	0.665	0.676	0.637	0.474	0.678
MNIST	rod	0.707	0.636	0.713	0.681	0.500	0.728
MNIST	mcd	0.547	0.576	0.645	0.633	0.600	0.568
annthyroid	iforest	0.737	0.722	0.613	0.772	0.465	0.842
annthyroid	copod	0.658	0.783	0.604	0.752	0.453	0.772
annthyroid	hbos	0.671	0.686	0.406	0.301	0.468	0.768
annthyroid	ecod	0.498	0.468	0.645	0.768	0.492	0.486
annthyroid	lunar	0.760	0.801	0.736	0.746	0.541	0.790
annthyroid	rod	0.805	0.842	0.812	0.858	0.447	0.846
annthyroid	mcd	0.294	0.277	0.225	0.198	0.492	0.277
satimage-2	iforest	0.770	0.639	0.723	0.609	0.249	0.781
satimage-2	copod	0.704	0.655	0.641	0.576	0.415	0.748
satimage-2	hbos	0.696	0.617	0.798	0.718	0.117	0.733
satimage-2	ecod	0.805	0.774	0.679	0.749	0.391	0.844
satimage-2	lunar	0.538	0.505	0.553	0.608	0.325	0.554
satimage-2	rod	0.789	0.603	0.891	0.542	0.324	0.808
satimage-2	mcd	0.687	0.694	0.667	0.609	0.586	0.711
smtp	iforest	0.806	0.748	0.267	0.406	0.693	0.946
smtp	copod	0.819	0.864	0.647	0.864	0.846	0.909
smtp	hbos	0.806	0.757	0.164	0.513	0.637	0.946
smtp	ecod	0.256	0.436	0.595	0.411	0.250	0.192
smtp	lunar	0.529	0.767	0.403	0.826	0.333	0.458
smtp	rod	0.658	0.186	0.639	0.571	0.585	0.635
smtp	mcd	0.803	0.786	0.196	0.622	0.407	0.763
mammography	iforest	0.674	N/A	0.753	0.589	0.392	0.707
mammography	copod	0.654	N/A	0.739	0.612	0.449	0.706
mammography	hbos	0.573	N/A	0.638	0.448	0.420	0.607
mammography	ecod	0.677	N/A	0.790	0.593	0.507	0.749
mammography	lunar	0.678	N/A	0.712	0.642	0.427	0.728
mammography	rod	0.484	N/A	0.277	0.889	0.957	0.472
fraud	iforest	0.570	N/A	0.500	0.384	N/A	0.667
fraud	copod	0.678	N/A	0.648	0.572	N/A	0.731
fraud	hbos	0.563	N/A	0.517	0.360	N/A	0.630
fraud	ecod	0.693	N/A	0.679	0.598	N/A	0.743
fraud	lunar	0.590	N/A	0.607	0.552	N/A	0.605
fraud	rod	0.608	N/A	0.596	0.562	N/A	0.629

Table 9: AUROC on all available detector annotations (original reported values).

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
glass	iforest	0.658	0.838	0.778	0.647	0.309	0.896
glass	copod	0.649	0.648	0.657	0.639	0.337	0.801

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
glass	hbos	0.644	0.829	0.820	0.644	0.453	0.856
glass	ecod	0.794	0.639	0.851	0.707	0.171	0.840
glass	lunar	0.597	0.604	0.714	0.464	0.358	0.743
glass	rod	0.574	0.365	0.537	0.515	0.104	0.567
glass	mcd	0.667	0.787	0.672	0.363	0.132	0.844
wpbc	iforest	0.898	0.792	0.657	0.624	0.434	0.910
wpbc	copod	0.832	0.741	0.655	0.618	0.468	0.846
wpbc	hbos	0.894	0.783	0.673	0.622	0.409	0.902
wpbc	ecod	0.920	0.483	0.599	0.571	0.374	0.928
wpbc	lunar	0.740	0.498	0.632	0.587	0.427	0.752
wpbc	rod	0.831	0.451	0.603	0.594	0.491	0.843
wpbc	mcd	0.860	0.534	0.648	0.618	0.398	0.875
vertebral	iforest	0.781	0.550	0.644	0.648	0.336	0.808
vertebral	copod	0.860	0.530	0.645	0.668	0.312	0.860
vertebral	hbos	0.717	0.439	0.511	0.518	0.404	0.739
vertebral	ecod	0.880	0.502	0.716	0.755	0.238	0.885
vertebral	lunar	0.573	0.349	0.439	0.390	0.364	0.579
vertebral	rod	0.747	0.510	0.690	0.710	0.632	0.731
vertebral	mcd	0.728	0.499	0.476	0.540	0.445	0.760
pageblocks	iforest	0.804	0.459	0.603	0.535	0.368	0.878
pageblocks	copod	0.734	0.424	0.577	0.529	0.384	0.828
pageblocks	hbos	0.803	0.457	0.612	0.559	0.446	0.709
pageblocks	ecod	0.609	0.477	0.578	0.550	0.526	0.620
pageblocks	lunar	0.682	0.493	0.542	0.567	0.421	0.711
pageblocks	rod	0.631	0.524	0.557	0.552	0.433	0.636
pageblocks	mcd	0.800	0.493	0.576	0.482	0.205	0.878
fault	iforest	0.697	0.674	0.590	0.546	0.424	0.750
fault	copod	0.715	0.662	0.573	0.551	0.457	0.760
fault	hbos	0.743	0.761	0.657	0.601	0.486	0.779
fault	ecod	0.775	0.517	0.593	0.566	0.498	0.822
fault	lunar	0.593	0.527	0.566	0.532	0.495	0.600
fault	rod	0.715	0.483	0.566	0.547	0.465	0.752
fault	mcd	0.517	0.550	0.550	0.548	0.458	0.553
yeast	iforest	0.784	0.776	0.443	0.388	0.386	0.796
yeast	copod	0.728	0.696	0.351	0.340	0.352	0.739
yeast	hbos	0.801	0.840	0.411	0.354	0.428	0.825
yeast	ecod	0.832	0.611	0.507	0.478	0.288	0.850
yeast	lunar	0.706	0.635	0.467	0.359	0.500	0.717
yeast	rod	0.116	0.148	0.276	0.269	0.089	0.095
MNIST	iforest	0.752	0.754	0.755	0.721	0.398	0.803
MNIST	copod	0.724	0.726	0.722	0.684	0.451	0.762
MNIST	hbos	0.809	0.782	0.824	0.795	0.299	0.848
MNIST	ecod	0.729	0.723	0.727	0.728	0.450	0.769
MNIST	lunar	0.666	0.688	0.645	0.621	0.502	0.680
MNIST	rod	0.727	0.686	0.709	0.703	0.473	0.755
MNIST	mcd	0.603	0.680	0.684	0.670	0.655	0.637
annthyroid	iforest	0.741	0.672	0.640	0.814	0.438	0.838
annthyroid	copod	0.782	0.835	0.568	0.733	0.272	0.886
annthyroid	hbos	0.716	0.658	0.562	0.446	0.366	0.791
annthyroid	ecod	0.649	0.610	0.579	0.723	0.679	0.648
annthyroid	lunar	0.628	0.686	0.569	0.617	0.739	0.664
annthyroid	rod	0.651	0.693	0.641	0.720	0.317	0.716
annthyroid	mcd	0.426	0.552	0.397	0.366	0.252	0.435
satimage-2	iforest	0.818	0.675	0.807	0.687	0.337	0.834
satimage-2	copod	0.786	0.702	0.717	0.632	0.490	0.810
satimage-2	hbos	0.813	0.662	0.880	0.813	0.322	0.833
satimage-2	ecod	0.934	0.899	0.721	0.720	0.437	0.942
satimage-2	lunar	0.661	0.581	0.680	0.665	0.383	0.670
satimage-2	rod	0.790	0.607	0.884	0.620	0.310	0.801

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
satimage-2	mcd	0.600	0.578	0.588	0.537	0.478	0.630
smtp	iforest	0.833	0.708	0.444	0.578	0.870	0.961
smtp	copod	0.806	0.889	0.672	0.844	0.618	0.932
smtp	hbos	0.833	0.728	0.089	0.489	0.643	0.961
smtp	ecod	0.356	0.200	0.589	0.578	0.280	0.289
smtp	lunar	0.419	0.867	0.183	0.775	0.359	0.308
smtp	rod	0.656	0.089	0.711	0.397	0.592	0.650
smtp	mcd	0.794	0.733	0.236	0.694	0.713	0.751
mammography	iforest	0.785	N/A	0.773	0.744	0.416	0.803
mammography	copod	0.598	N/A	0.692	0.632	0.344	0.658
mammography	hbos	0.638	N/A	0.681	0.617	0.465	0.661
mammography	ecod	0.677	N/A	0.774	0.731	0.390	0.753
mammography	lunar	0.678	N/A	0.716	0.706	0.489	0.738
mammography	rod	0.751	N/A	0.608	0.898	0.875	0.742
fraud	iforest	0.795	N/A	0.775	0.720	N/A	0.881
fraud	copod	0.771	N/A	0.772	0.743	N/A	0.825
fraud	hbos	0.810	N/A	0.812	0.733	N/A	0.855
fraud	ecod	0.797	N/A	0.806	0.775	N/A	0.849
fraud	lunar	0.637	N/A	0.647	0.620	N/A	0.657
fraud	rod	0.653	N/A	0.631	0.593	N/A	0.678

Table 10: Precision on all available detector annotations (original reported values).

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
glass	iforest	0.657	0.389	0.667	0.537	0.185	0.741
glass	copod	0.611	0.648	0.333	0.333	0.222	0.759
glass	hbos	0.542	0.472	0.380	0.435	0.343	0.583
glass	ecod	0.528	0.722	0.278	0.222	0.111	0.722
glass	lunar	0.570	0.533	0.394	0.446	0.304	0.524
glass	rod	0.111	0.000	0.222	0.111	0.130	0.222
glass	mcd	0.444	0.556	0.111	0.111	0.435	0.667
wpbc	iforest	0.693	0.512	0.465	0.415	0.316	0.783
wpbc	copod	0.734	0.578	0.618	0.596	0.361	0.790
wpbc	hbos	0.651	0.558	0.513	0.487	0.338	0.659
wpbc	ecod	0.733	0.599	0.520	0.513	0.262	0.713
wpbc	lunar	0.618	0.587	0.569	0.543	0.338	0.661
wpbc	rod	0.703	0.587	0.542	0.546	0.412	0.733
wpbc	mcd	0.755	0.562	0.575	0.581	0.289	0.760
vertebral	iforest	0.783	0.533	0.669	0.669	0.258	0.789
vertebral	copod	0.661	0.433	0.467	0.389	0.239	0.717
vertebral	hbos	0.739	0.364	0.489	0.569	0.347	0.714
vertebral	ecod	0.517	0.283	0.033	0.183	0.100	0.600
vertebral	lunar	0.360	0.247	0.228	0.235	0.316	0.410
vertebral	rod	0.733	0.667	0.775	0.792	0.525	0.733
vertebral	mcd	0.714	0.586	0.406	0.406	0.417	0.744
pageblocks	iforest	0.466	0.358	0.315	0.357	0.459	0.475
pageblocks	copod	0.555	0.244	0.128	0.179	0.547	0.634
pageblocks	hbos	0.510	0.290	0.108	0.192	0.507	0.581
pageblocks	ecod	0.471	0.160	0.168	0.144	0.560	0.520
pageblocks	lunar	0.561	0.434	0.407	0.427	0.514	0.587
pageblocks	rod	0.527	0.458	0.381	0.416	0.466	0.555
pageblocks	mcd	0.393	0.097	0.018	0.014	0.384	0.502
fault	iforest	0.645	0.535	0.480	0.463	0.289	0.692
fault	copod	0.551	0.517	0.419	0.388	0.348	0.600
fault	hbos	0.592	0.534	0.494	0.474	0.268	0.629
fault	ecod	0.551	0.518	0.393	0.325	0.383	0.605

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
fault	lunar	0.539	0.567	0.485	0.478	0.376	0.557
fault	rod	0.596	0.510	0.467	0.448	0.442	0.624
fault	mcd	0.520	0.550	0.527	0.522	0.441	0.545
yeast	iforest	0.666	0.512	0.388	0.236	0.233	0.727
yeast	copod	0.551	0.548	0.294	0.145	0.354	0.600
yeast	hbos	0.567	0.557	0.255	0.200	0.183	0.599
yeast	ecod	0.482	0.458	0.199	0.058	0.302	0.528
yeast	lunar	0.607	0.675	0.437	0.367	0.444	0.622
yeast	rod	0.028	0.008	0.006	0.000	0.004	0.008
MNIST	iforest	0.630	0.584	0.634	0.642	0.490	0.669
MNIST	copod	0.609	0.541	0.662	0.631	0.480	0.647
MNIST	hbos	0.612	0.583	0.592	0.588	0.471	0.652
MNIST	ecod	0.558	0.486	0.712	0.607	0.482	0.601
MNIST	lunar	0.576	0.583	0.577	0.568	0.437	0.598
MNIST	rod	0.617	0.554	0.611	0.617	0.472	0.644
MNIST	mcd	0.476	0.493	0.545	0.534	0.355	0.510
annthyroid	iforest	0.572	0.718	0.541	0.712	0.681	0.683
annthyroid	copod	0.587	0.568	0.478	0.592	0.450	0.656
annthyroid	hbos	0.388	0.358	0.293	0.187	0.320	0.384
annthyroid	ecod	0.649	0.585	0.519	0.620	0.456	0.740
annthyroid	lunar	0.685	0.695	0.641	0.641	0.568	0.700
annthyroid	rod	0.731	0.748	0.710	0.743	0.623	0.759
annthyroid	mcd	0.165	0.143	0.109	0.090	0.209	0.321
satimage-2	iforest	0.655	0.560	0.649	0.508	0.133	0.656
satimage-2	copod	0.667	0.543	0.559	0.524	0.399	0.660
satimage-2	hbos	0.754	0.752	0.778	0.678	0.049	0.799
satimage-2	ecod	0.700	0.545	0.587	0.595	0.352	0.700
satimage-2	lunar	0.486	0.433	0.500	0.539	0.244	0.465
satimage-2	rod	0.711	0.514	0.835	0.479	0.197	0.739
satimage-2	mcd	0.650	0.646	0.642	0.554	0.545	0.683
smtp	iforest	0.783	0.608	0.133	0.250	0.648	0.917
smtp	copod	0.800	0.283	0.628	0.817	0.796	0.967
smtp	hbos	0.100	0.025	0.033	0.433	0.057	0.033
smtp	ecod	0.783	0.642	0.589	0.267	0.778	0.950
smtp	lunar	0.467	0.467	0.217	0.778	0.463	0.367
smtp	rod	0.600	0.450	0.567	0.472	0.522	0.567
smtp	mcd	0.308	0.292	0.017	0.372	0.666	0.542
mammography	iforest	0.604	N/A	0.645	0.513	0.471	0.686
mammography	copod	0.584	N/A	0.669	0.493	0.584	0.635
mammography	hbos	0.574	N/A	0.632	0.503	0.115	0.649
mammography	ecod	0.497	N/A	0.509	0.351	0.474	0.549
mammography	lunar	0.599	N/A	0.579	0.585	0.589	0.637
mammography	rod	0.350	N/A	0.077	0.046	0.114	0.342
fraud	iforest	0.598	N/A	0.575	0.509	N/A	0.672
fraud	copod	0.520	N/A	0.411	0.295	N/A	0.585
fraud	hbos	0.587	N/A	0.586	0.532	N/A	0.644
fraud	ecod	0.438	N/A	0.433	0.252	N/A	0.544
fraud	lunar	0.501	N/A	0.504	0.491	N/A	0.529
fraud	rod	0.553	N/A	0.511	0.511	N/A	0.566

Table 11: Jaccard on all available detector annotations (original reported values).

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
glass	iforest	0.561	0.282	0.541	0.437	0.130	0.630
glass	copod	0.565	0.615	0.263	0.263	0.222	0.704
glass	hbos	0.407	0.337	0.253	0.304	0.231	0.482

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
glass	ecod	0.482	0.704	0.222	0.185	0.074	0.667
glass	lunar	0.461	0.398	0.302	0.331	0.238	0.407
glass	rod	0.111	0.000	0.222	0.111	0.000	0.222
glass	mcd	0.444	0.556	0.111	0.111	0.000	0.667
wpbc	iforest	0.595	0.402	0.352	0.322	0.236	0.706
wpbc	copod	0.618	0.452	0.495	0.477	0.253	0.690
wpbc	hbos	0.522	0.432	0.393	0.362	0.218	0.527
wpbc	ecod	0.631	0.510	0.419	0.421	0.224	0.615
wpbc	lunar	0.481	0.449	0.425	0.405	0.241	0.529
wpbc	rod	0.562	0.437	0.397	0.397	0.281	0.600
wpbc	mcd	0.655	0.454	0.459	0.466	0.214	0.656
vertebral	iforest	0.717	0.443	0.592	0.603	0.210	0.722
vertebral	copod	0.611	0.347	0.394	0.329	0.131	0.678
vertebral	hbos	0.679	0.292	0.408	0.500	0.252	0.643
vertebral	ecod	0.500	0.267	0.022	0.178	0.100	0.589
vertebral	lunar	0.308	0.193	0.184	0.188	0.202	0.357
vertebral	rod	0.631	0.562	0.662	0.680	0.433	0.631
vertebral	mcd	0.629	0.479	0.287	0.298	0.290	0.666
pageblocks	iforest	0.339	0.244	0.214	0.243	0.099	0.346
pageblocks	copod	0.508	0.214	0.098	0.138	0.118	0.590
pageblocks	hbos	0.443	0.226	0.072	0.130	0.130	0.510
pageblocks	ecod	0.402	0.121	0.145	0.121	0.250	0.450
pageblocks	lunar	0.428	0.313	0.291	0.309	0.261	0.452
pageblocks	rod	0.391	0.328	0.248	0.279	0.201	0.418
pageblocks	mcd	0.372	0.090	0.013	0.011	0.006	0.483
fault	iforest	0.520	0.405	0.351	0.333	0.198	0.571
fault	copod	0.425	0.393	0.306	0.278	0.235	0.477
fault	hbos	0.457	0.402	0.359	0.342	0.199	0.498
fault	ecod	0.465	0.434	0.313	0.247	0.260	0.520
fault	lunar	0.407	0.430	0.360	0.356	0.255	0.423
fault	rod	0.457	0.369	0.337	0.319	0.320	0.487
fault	mcd	0.382	0.410	0.391	0.389	0.317	0.406
yeast	iforest	0.584	0.425	0.313	0.167	0.237	0.648
yeast	copod	0.492	0.489	0.249	0.116	0.221	0.541
yeast	hbos	0.457	0.444	0.175	0.132	0.282	0.486
yeast	ecod	0.472	0.449	0.189	0.052	0.165	0.519
yeast	lunar	0.472	0.554	0.311	0.257	0.323	0.490
yeast	rod	0.028	0.008	0.006	0.000	0.039	0.008
MNIST	iforest	0.491	0.446	0.493	0.504	0.226	0.535
MNIST	copod	0.481	0.419	0.537	0.510	0.250	0.522
MNIST	hbos	0.473	0.446	0.446	0.446	0.162	0.514
MNIST	ecod	0.485	0.416	0.642	0.533	0.255	0.531
MNIST	lunar	0.435	0.448	0.430	0.420	0.288	0.454
MNIST	rod	0.470	0.409	0.465	0.470	0.296	0.499
MNIST	mcd	0.359	0.371	0.409	0.398	0.371	0.386
annthyroid	iforest	0.498	0.406	0.517	0.680	0.370	0.590
annthyroid	copod	0.355	0.327	0.359	0.462	0.217	0.444
annthyroid	hbos	0.427	0.334	0.267	0.153	0.330	0.522
annthyroid	ecod	0.337	0.294	0.393	0.494	0.457	0.325
annthyroid	lunar	0.504	0.534	0.505	0.518	0.536	0.460
annthyroid	rod	0.546	0.592	0.578	0.619	0.226	0.514
annthyroid	mcd	0.160	0.287	0.142	0.073	0.211	0.181
satimage-2	iforest	0.475	0.389	0.533	0.394	0.093	0.438
satimage-2	copod	0.405	0.282	0.419	0.389	0.278	0.422
satimage-2	hbos	0.447	0.369	0.747	0.652	0.034	0.413
satimage-2	ecod	0.355	0.141	0.459	0.457	0.253	0.389
satimage-2	lunar	0.317	0.325	0.392	0.417	0.163	0.303
satimage-2	rod	0.537	0.387	0.752	0.359	0.125	0.505

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
satimage-2	mcd	0.333	0.532	0.498	0.425	0.410	0.313
smtp	iforest	0.644	0.608	0.111	0.233	0.578	0.619
smtp	copod	0.394	0.283	0.572	0.783	0.461	0.469
smtp	hbos	0.639	0.417	0.033	0.433	0.461	0.590
smtp	ecod	0.183	0.250	0.517	0.267	0.239	0.111
smtp	lunar	0.381	0.267	0.161	0.728	0.273	0.253
smtp	rod	0.317	0.350	0.567	0.417	0.356	0.356
smtp	mcd	0.372	0.292	0.011	0.307	0.367	0.378
mammography	iforest	0.497	N/A	0.519	0.399	0.287	0.586
mammography	copod	0.536	N/A	0.615	0.443	0.217	0.586
mammography	hbos	0.455	N/A	0.498	0.392	0.326	0.534
mammography	ecod	0.423	N/A	0.422	0.263	0.253	0.471
mammography	lunar	0.481	N/A	0.452	0.470	0.311	0.522
mammography	rod	0.350	N/A	0.077	0.046	0.250	0.342
fraud	iforest	0.495	N/A	0.474	0.405	N/A	0.576
fraud	copod	0.482	N/A	0.370	0.257	N/A	0.548
fraud	hbos	0.472	N/A	0.464	0.415	N/A	0.537
fraud	ecod	0.421	N/A	0.420	0.235	N/A	0.530
fraud	lunar	0.375	N/A	0.368	0.361	N/A	0.400
fraud	rod	0.408	N/A	0.364	0.363	N/A	0.421

Table 12: Fidelity on all available detector annotations (original reported values).

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
glass	iforest	0.238	0.345	0.419	0.479	0.272	0.350
glass	copod	0.176	0.356	0.111	0.350	0.390	0.237
glass	hbos	0.152	0.388	0.799	0.138	0.245	0.483
glass	ecod	0.706	0.070	0.047	0.543	0.520	0.333
glass	lunar	0.575	0.490	0.130	0.566	0.146	0.353
glass	rod	0.326	0.103	0.406	0.397	0.442	0.773
glass	mcd	0.561	0.034	0.254	0.458	0.395	0.352
wpbc	iforest	0.884	0.886	0.751	0.532	0.405	0.908
wpbc	copod	0.802	0.787	0.500	0.383	0.446	0.872
wpbc	hbos	0.888	0.781	0.523	0.405	0.447	0.884
wpbc	ecod	0.834	0.775	0.658	0.351	0.392	0.881
wpbc	lunar	0.895	0.000	0.000	0.591	0.381	0.882
wpbc	rod	0.841	0.719	0.600	0.411	0.381	0.890
wpbc	mcd	0.857	0.658	0.505	0.446	0.280	0.881
vertebral	iforest	0.398	0.409	0.220	0.178	0.045	0.324
vertebral	copod	0.145	0.387	0.394	0.249	0.208	0.105
vertebral	hbos	0.470	0.516	0.068	0.069	0.227	0.415
vertebral	ecod	0.498	0.644	0.691	0.660	0.202	0.476
vertebral	lunar	0.154	0.031	0.022	0.118	0.053	0.325
vertebral	rod	0.327	0.506	0.308	0.088	0.266	0.071
vertebral	mcd	0.272	0.002	0.109	0.148	0.477	0.137
pageblocks	iforest	0.812	0.532	0.701	0.694	0.139	0.781
pageblocks	copod	0.819	0.260	0.848	0.833	0.466	0.791
pageblocks	hbos	0.317	0.257	0.086	0.047	0.150	0.361
pageblocks	ecod	0.694	0.252	0.702	0.699	0.472	0.637
pageblocks	lunar	0.982	0.290	0.932	0.453	0.884	0.977
pageblocks	rod	0.564	0.374	0.522	0.667	0.156	0.507
pageblocks	mcd	0.982	0.009	0.998	0.481	0.986	0.982
fault	iforest	0.837	0.785	0.761	0.551	0.537	0.847
fault	copod	0.791	0.739	0.793	0.583	0.575	0.819
fault	hbos	0.811	0.744	0.739	0.666	0.558	0.808
fault	ecod	0.810	0.770	0.784	0.544	0.526	0.827

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
fault	lunar	0.627	0.380	0.475	0.352	0.194	0.305
fault	rod	0.783	0.725	0.684	0.565	0.511	0.778
fault	mcd	0.594	0.570	0.419	0.403	0.553	0.691
yeast	iforest	0.778	0.709	0.696	0.542	0.540	0.788
yeast	copod	0.746	0.680	0.730	0.639	0.508	0.778
yeast	hbos	0.627	0.640	0.694	0.557	0.447	0.603
yeast	ecod	0.733	0.675	0.697	0.588	0.498	0.760
yeast	lunar	0.616	0.810	0.645	0.468	0.369	0.831
yeast	rod	0.739	0.613	0.627	0.497	0.451	0.746
MNIST	iforest	0.627	0.686	0.447	0.508	0.068	0.641
MNIST	copod	0.711	0.668	0.584	0.622	0.012	0.737
MNIST	hbos	0.693	0.676	0.562	0.528	0.064	0.713
MNIST	ecod	0.660	0.686	0.516	0.477	0.044	0.678
MNIST	lunar	0.498	0.382	0.546	0.550	0.083	0.502
MNIST	rod	0.653	0.663	0.445	0.392	0.106	0.636
MNIST	mcd	0.781	0.735	0.768	0.776	0.654	0.779
annthyroid	iforest	0.625	0.582	0.765	0.899	0.411	0.529
annthyroid	copod	0.565	0.411	0.704	0.791	0.405	0.430
annthyroid	hbos	0.735	0.694	0.870	0.885	0.571	0.638
annthyroid	ecod	0.709	0.688	0.775	0.771	0.593	0.644
annthyroid	lunar	0.965	0.986	0.979	0.984	0.977	0.963
annthyroid	rod	0.007	0.000	0.014	0.074	0.042	0.055
annthyroid	mcd	0.992	0.617	0.634	1.000	0.862	0.866
satimage-2	iforest	0.299	0.364	0.190	0.142	0.225	0.555
satimage-2	copod	0.261	0.280	0.287	0.095	0.490	0.354
satimage-2	hbos	0.423	0.570	0.107	0.130	0.143	0.433
satimage-2	ecod	0.196	0.333	0.080	0.166	0.397	0.280
satimage-2	lunar	0.970	0.898	0.914	0.930	0.109	0.966
satimage-2	rod	0.030	0.455	0.294	0.209	0.041	0.048
satimage-2	mcd	0.842	0.884	0.866	0.812	0.843	0.972
smtp	iforest	0.765	0.457	0.588	0.730	0.664	0.616
smtp	copod	0.775	0.547	0.791	0.791	0.307	0.572
smtp	hbos	0.722	0.762	0.270	0.270	0.262	0.962
smtp	ecod	0.905	0.762	0.577	0.577	0.551	0.982
smtp	lunar	0.843	0.884	0.756	0.546	0.728	0.794
smtp	rod	0.398	0.694	0.848	0.778	0.824	0.929
smtp	mcd	0.466	0.216	0.415	0.344	0.435	0.540
mammography	iforest	0.728	N/A	0.830	0.558	0.616	0.761
mammography	copod	0.760	N/A	0.823	0.691	0.798	0.829
mammography	hbos	0.503	N/A	0.515	0.618	0.292	0.512
mammography	ecod	0.766	N/A	0.830	0.684	0.811	0.828
mammography	lunar	0.980	N/A	0.826	0.426	0.841	0.989
mammography	rod	0.626	N/A	0.699	0.358	0.321	0.655
fraud	iforest	0.721	N/A	0.855	0.711	N/A	0.732
fraud	copod	0.743	N/A	0.909	0.718	N/A	0.805
fraud	hbos	0.711	N/A	0.853	0.204	N/A	0.767
fraud	ecod	0.769	N/A	0.893	0.583	N/A	0.833
fraud	lunar	0.776	N/A	0.765	0.617	N/A	0.734
fraud	rod	0.716	N/A	0.860	0.400	N/A	0.771

Table 13: Insertion AUPRC on all available detector annotations (original reported values).

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
glass	iforest	0.750	1.000	0.736	0.771	0.950	0.893
glass	copod	0.854	0.861	0.705	0.761	0.635	0.861
glass	hbos	0.830	1.000	0.813	0.817	1.000	1.000

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
glass	ecod	0.718	0.635	0.732	0.718	0.372	0.705
glass	lunar	0.705	0.761	0.622	0.692	0.888	0.893
glass	rod	1.000	1.000	1.000	1.000	0.950	1.000
glass	mcd	0.385	0.678	0.402	0.461	0.653	0.611
wpbc	iforest	0.990	0.975	0.986	0.972	0.975	0.993
wpbc	copod	0.984	0.976	0.963	0.962	0.973	0.986
wpbc	hbos	1.000	0.979	0.974	0.973	0.970	1.000
wpbc	ecod	0.993	0.972	0.919	0.950	0.975	0.989
wpbc	lunar	0.988	0.980	0.958	0.963	0.963	0.995
wpbc	rod	0.901	0.791	0.734	0.787	0.688	0.906
wpbc	mcd	0.925	0.917	0.967	0.933	0.985	1.000
vertebral	iforest	0.928	0.754	1.000	0.870	0.841	0.872
vertebral	copod	0.880	0.729	0.801	0.869	0.799	0.883
vertebral	hbos	0.904	0.852	0.971	0.926	0.842	0.878
vertebral	ecod	0.685	0.719	0.800	0.825	0.584	0.636
vertebral	lunar	0.481	0.490	0.531	0.537	0.587	0.597
vertebral	rod	0.467	0.568	0.568	0.566	0.511	0.527
vertebral	mcd	0.554	0.473	0.654	0.661	0.566	0.523
pageblocks	iforest	0.975	0.986	0.974	0.986	0.954	0.983
pageblocks	copod	0.904	0.921	0.878	0.885	0.874	0.898
pageblocks	hbos	0.839	0.888	0.925	0.816	0.797	0.891
pageblocks	ecod	0.825	0.794	0.740	0.749	0.775	0.814
pageblocks	lunar	0.954	0.850	0.965	0.948	0.922	0.969
pageblocks	rod	0.917	0.931	0.958	0.957	0.941	0.932
pageblocks	mcd	0.991	0.994	0.993	0.993	0.956	0.993
fault	iforest	0.987	0.982	0.981	0.972	0.967	0.992
fault	copod	0.981	0.971	0.964	0.962	0.968	0.982
fault	hbos	0.984	0.987	0.973	0.962	0.966	0.987
fault	ecod	0.988	0.975	0.972	0.964	0.959	0.986
fault	lunar	0.943	0.901	0.933	0.905	0.882	0.947
fault	rod	0.944	0.916	0.905	0.906	0.864	0.942
fault	mcd	0.977	0.968	0.957	0.960	0.951	0.981
yeast	iforest	0.995	0.993	0.984	0.963	0.974	0.995
yeast	copod	0.987	0.986	0.978	0.951	0.949	0.991
yeast	hbos	0.975	0.979	0.980	0.959	0.955	0.980
yeast	ecod	0.987	0.991	0.982	0.942	0.949	0.992
yeast	lunar	0.984	0.979	0.965	0.939	0.957	0.960
yeast	rod	0.957	0.951	0.936	0.873	0.877	0.961
MNIST	iforest	0.969	0.953	0.954	0.968	0.738	0.972
MNIST	copod	0.971	0.962	0.949	0.952	0.765	0.973
MNIST	hbos	0.981	0.967	0.967	0.973	0.854	0.984
MNIST	ecod	0.967	0.953	0.948	0.955	0.816	0.966
MNIST	lunar	0.911	0.925	0.934	0.920	0.762	0.921
MNIST	rod	0.824	0.763	0.709	0.734	0.604	0.819
MNIST	mcd	0.980	0.975	0.982	0.983	0.911	0.983
annthyroid	iforest	0.976	0.967	0.993	0.987	0.731	0.958
annthyroid	copod	0.771	0.731	0.770	0.826	0.614	0.749
annthyroid	hbos	0.979	0.984	0.993	0.993	0.850	0.974
annthyroid	ecod	0.719	0.698	0.721	0.741	0.853	0.710
annthyroid	lunar	0.879	0.859	0.937	0.856	0.641	0.822
annthyroid	rod	0.793	0.748	0.826	0.836	0.794	0.756
annthyroid	mcd	0.721	0.423	0.456	0.678	0.742	0.655
satimage-2	iforest	0.999	0.949	0.998	0.944	0.969	0.999
satimage-2	copod	0.836	0.836	0.815	0.890	0.779	0.824
satimage-2	hbos	0.989	0.961	0.958	0.956	0.973	0.973
satimage-2	ecod	0.958	0.908	0.858	0.884	0.917	0.956
satimage-2	lunar	0.998	0.992	0.991	0.987	0.903	0.996
satimage-2	rod	0.804	0.799	0.611	0.714	0.814	0.797

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
satimage-2	mcd	0.994	0.941	0.971	0.971	0.827	0.960
smtp	iforest	0.875	0.949	0.970	0.891	0.773	0.814
smtp	copod	0.720	0.886	0.694	0.422	0.641	0.698
smtp	hbos	0.996	0.972	1.000	0.930	0.769	0.983
smtp	ecod	0.939	0.942	0.786	0.746	0.485	0.939
smtp	lunar	0.953	0.967	0.987	0.935	0.564	0.971
smtp	rod	0.431	0.567	0.468	0.670	0.509	0.667
smtp	mcd	0.440	0.629	0.488	0.652	0.725	0.668
mammography	iforest	0.973	N/A	0.996	0.982	0.639	0.979
mammography	copod	0.963	N/A	0.988	0.979	0.756	0.989
mammography	hbos	0.915	N/A	0.910	0.948	0.756	0.919
mammography	ecod	0.961	N/A	0.987	0.970	0.917	0.981
mammography	lunar	0.950	N/A	0.953	0.949	0.806	0.967
mammography	rod	0.938	N/A	0.960	0.920	0.867	0.953
fraud	iforest	0.918	N/A	0.991	0.883	N/A	0.969
fraud	copod	0.972	N/A	0.998	0.821	N/A	0.987
fraud	hbos	0.964	N/A	0.994	0.840	N/A	0.984
fraud	ecod	0.977	N/A	0.997	0.683	N/A	0.991
fraud	lunar	0.948	N/A	0.975	0.848	N/A	0.972
fraud	rod	0.736	N/A	0.883	0.725	N/A	0.810

Table 14: Deletion AUPRC on all available detector annotations (original reported values).

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
glass	iforest	0.950	0.888	1.000	0.917	0.888	1.000
glass	copod	0.950	1.000	1.000	1.000	0.950	0.917
glass	hbos	0.950	0.888	0.950	1.000	0.950	1.000
glass	ecod	0.888	0.888	0.950	0.950	0.917	1.000
glass	lunar	0.830	0.893	0.950	0.950	0.771	0.771
glass	rod	1.000	1.000	1.000	1.000	1.000	1.000
glass	mcd	0.813	0.430	0.692	0.793	0.441	0.650
wpbc	iforest	0.825	0.933	0.893	0.963	0.887	0.921
wpbc	copod	0.917	0.949	0.847	0.868	0.818	0.864
wpbc	hbos	0.855	0.913	0.963	0.978	0.898	0.882
wpbc	ecod	0.850	0.912	0.966	0.969	0.831	0.876
wpbc	lunar	0.916	0.958	0.952	0.966	0.883	0.921
wpbc	rod	0.907	0.912	0.962	0.983	0.869	0.918
wpbc	mcd	0.856	0.911	0.915	0.908	0.944	0.936
vertebral	iforest	0.750	0.763	0.729	0.710	0.817	0.675
vertebral	copod	0.828	0.860	0.766	0.889	0.775	0.784
vertebral	hbos	0.875	0.918	0.978	0.868	0.974	0.894
vertebral	ecod	0.870	0.717	0.854	0.818	0.970	0.894
vertebral	lunar	0.646	0.575	0.664	0.673	0.586	0.603
vertebral	rod	0.856	0.895	0.898	0.869	0.973	0.821
vertebral	mcd	0.583	0.473	0.654	0.616	0.566	0.523
pageblocks	iforest	0.949	0.895	0.986	0.975	0.980	0.966
pageblocks	copod	0.942	0.888	0.955	0.957	0.937	0.977
pageblocks	hbos	0.657	0.713	0.675	0.632	0.674	0.678
pageblocks	ecod	0.950	0.904	0.854	0.880	0.904	0.969
pageblocks	lunar	0.893	0.807	0.891	0.960	0.881	0.909
pageblocks	rod	0.829	0.851	0.840	0.901	0.915	0.853
pageblocks	mcd	0.942	0.961	0.989	0.982	0.926	0.944
fault	iforest	0.867	0.910	0.894	0.893	0.925	0.886
fault	copod	0.900	0.931	0.877	0.917	0.920	0.906
fault	hbos	0.839	0.915	0.887	0.914	0.930	0.855
fault	ecod	0.867	0.918	0.915	0.931	0.937	0.903

Dataset	Detector	ATON	COIN	SHAP	LIME	WANDER	TMQ
fault	lunar	0.573	0.774	0.690	0.746	0.823	0.574
fault	rod	0.817	0.906	0.871	0.910	0.910	0.860
fault	mcd	0.847	0.887	0.910	0.829	0.886	0.813
yeast	iforest	0.866	0.897	0.891	0.937	0.930	0.893
yeast	copod	0.538	0.926	0.835	0.831	0.866	0.799
yeast	hbos	0.856	0.862	0.787	0.881	0.906	0.866
yeast	ecod	0.556	0.909	0.861	0.875	0.876	0.836
yeast	lunar	0.787	0.835	0.808	0.807	0.881	0.749
yeast	rod	0.849	0.856	0.863	0.927	0.898	0.901
MNIST	iforest	0.781	0.782	0.835	0.852	0.943	0.854
MNIST	copod	0.809	0.819	0.839	0.859	0.966	0.816
MNIST	hbos	0.834	0.810	0.899	0.897	0.967	0.866
MNIST	ecod	0.821	0.801	0.870	0.876	0.952	0.844
MNIST	lunar	0.595	0.722	0.683	0.579	0.800	0.594
MNIST	rod	0.847	0.791	0.900	0.887	0.952	0.878
MNIST	mcd	0.791	0.782	0.772	0.749	0.777	0.801
annthyroid	iforest	0.708	0.763	0.776	0.933	0.645	0.764
annthyroid	copod	0.584	0.855	0.717	0.921	0.819	0.812
annthyroid	hbos	0.755	0.782	0.851	0.898	0.789	0.860
annthyroid	ecod	0.839	0.915	0.913	0.979	0.930	0.931
annthyroid	lunar	0.451	0.561	0.525	0.651	0.561	0.605
annthyroid	rod	0.829	0.916	0.943	0.989	0.925	0.916
annthyroid	mcd	0.396	0.583	0.570	0.427	0.504	0.603
satimage-2	iforest	0.955	0.952	0.968	0.958	0.930	0.970
satimage-2	copod	0.876	0.898	0.873	0.762	0.749	0.897
satimage-2	hbos	0.825	0.939	0.978	0.961	0.931	0.839
satimage-2	ecod	0.954	0.940	0.983	0.898	0.898	0.964
satimage-2	lunar	0.548	0.746	0.483	0.444	0.898	0.497
satimage-2	rod	0.907	0.931	0.964	0.855	0.878	0.851
satimage-2	mcd	0.539	0.521	0.584	0.454	0.853	0.821
smtp	iforest	0.896	0.866	0.861	1.000	0.635	0.950
smtp	copod	0.761	0.656	0.988	0.996	0.683	0.890
smtp	hbos	0.948	0.969	0.933	0.984	0.739	0.969
smtp	ecod	0.982	0.839	1.000	1.000	0.903	0.942
smtp	lunar	0.946	0.658	0.975	0.996	0.456	0.970
smtp	rod	0.586	0.690	0.708	0.700	0.903	0.884
smtp	mcd	0.692	0.796	0.848	0.694	0.604	0.794
mammography	iforest	0.778	N/A	0.555	0.936	0.867	0.800
mammography	copod	0.742	N/A	0.640	0.840	0.917	0.841
mammography	hbos	0.579	N/A	0.692	0.896	0.833	0.665
mammography	ecod	0.766	N/A	0.653	0.861	0.867	0.867
mammography	lunar	0.518	N/A	0.571	0.885	0.806	0.608
mammography	rod	0.671	N/A	0.590	0.905	0.917	0.705
fraud	iforest	0.959	N/A	0.954	0.634	N/A	0.966
fraud	copod	0.961	N/A	0.963	0.959	N/A	0.969
fraud	hbos	0.971	N/A	0.986	0.951	N/A	0.983
fraud	ecod	0.971	N/A	0.986	0.996	N/A	0.985
fraud	lunar	0.883	N/A	0.855	0.784	N/A	0.875
fraud	rod	0.935	N/A	0.933	0.765	N/A	0.949

Table 15: F_1 values (methods ATON–Anchor).

Dataset	Detector	ATON	COIN	SiNNE	Anchor
glass	iforest	0.443	0.718	0.252	0.449
glass	copod	0.608	0.482	0.236	0.609
glass	hbos	0.374	0.617	0.201	0.326

Dataset	Detector	ATON	COIN	SiNNE	Anchor
glass	ecod	0.650	0.485	0.420	0.509
glass	lunar	0.572	0.421	0.460	0.596
glass	rod	0.271	0.000	0.363	0.263
glass	mcd	0.307	0.494	0.101	0.259
wpbc	iforest	0.758	0.644	0.616	0.452
wpbc	copod	0.640	0.571	0.522	0.422
wpbc	hbos	0.691	0.567	0.550	0.394
wpbc	ecod	0.738	0.290	0.620	0.363
wpbc	lunar	0.621	0.390	0.528	0.532
wpbc	rod	0.721	0.336	0.614	0.493
wpbc	mcd	0.683	0.388	0.538	0.426
vertebral	iforest	0.614	0.470	0.575	0.489
vertebral	copod	0.693	0.511	0.566	0.461
vertebral	hbos	0.403	0.320	0.363	0.265
vertebral	ecod	0.731	0.494	0.614	0.471
vertebral	lunar	0.405	0.463	0.379	0.394
vertebral	rod	0.706	0.725	0.640	0.695
vertebral	mcd	0.649	0.584	0.589	0.565
pageblocks	iforest	0.584	0.180	0.431	0.292
pageblocks	copod	0.543	0.227	0.408	0.325
pageblocks	hbos	0.532	0.228	0.399	0.338
pageblocks	ecod	0.468	0.460	0.447	0.466
pageblocks	lunar	0.532	0.398	0.449	0.485
pageblocks	rod	0.499	0.526	0.467	0.576
pageblocks	mcd	0.449	0.139	0.275	0.228
fault	iforest	0.536	0.555	0.535	0.493
fault	copod	0.568	0.586	0.535	0.520
fault	hbos	0.519	0.511	0.519	0.391
fault	ecod	0.621	0.471	0.580	0.497
fault	lunar	0.481	0.535	0.476	0.563
fault	rod	0.582	0.471	0.578	0.524
fault	mcd	0.419	0.589	0.443	0.601
yeast	iforest	0.558	0.443	0.566	0.305
yeast	copod	0.597	0.530	0.564	0.393
yeast	hbos	0.457	0.347	0.388	0.198
yeast	ecod	0.649	0.471	0.586	0.370
yeast	lunar	0.555	0.640	0.523	0.556
yeast	rod	0.008	0.006	0.228	0.144
MNIST	iforest	0.602	0.541	0.678	0.511
MNIST	copod	0.597	0.614	0.638	0.567
MNIST	hbos	0.558	0.428	0.610	0.421
MNIST	ecod	0.607	0.603	0.580	0.574
MNIST	lunar	0.537	0.626	0.578	0.591
MNIST	rod	0.599	0.591	0.659	0.564
MNIST	mcd	0.466	0.567	0.546	0.549
annthyroid	iforest	0.654	0.590	0.606	0.597
annthyroid	copod	0.523	0.515	0.675	0.390
annthyroid	hbos	0.582	0.510	0.647	0.519
annthyroid	ecod	0.484	0.476	0.173	0.410
annthyroid	lunar	0.659	0.705	0.450	0.743
annthyroid	rod	0.697	0.742	0.475	0.782
annthyroid	mcd	0.256	0.458	0.074	0.333
satimage-2	iforest	0.675	0.560	0.452	0.551
satimage-2	copod	0.601	0.440	0.444	0.439
satimage-2	hbos	0.635	0.539	0.410	0.535
satimage-2	ecod	0.531	0.247	0.446	0.254
satimage-2	lunar	0.491	0.491	0.354	0.491
satimage-2	rod	0.696	0.558	0.422	0.563

Dataset	Detector	ATON	COIN	SiNNE	Anchor
satimage-2	mcd	0.525	0.695	0.353	0.684
smtp	iforest	0.789	0.756	0.817	0.741
smtp	copod	0.547	0.442	0.737	0.430
smtp	hbos	0.786	0.782	0.802	0.769
smtp	ecod	0.263	0.400	0.022	0.401
smtp	lunar	0.541	0.636	0.239	0.632
smtp	rod	0.480	0.621	0.518	0.607
smtp	mcd	0.532	0.452	0.513	0.456
mammography	iforest	0.560	N/A	0.625	0.419
mammography	copod	0.543	N/A	0.585	0.637
mammography	hbos	0.477	N/A	0.560	0.434
mammography	ecod	0.582	N/A	0.570	0.588
mammography	lunar	0.570	N/A	0.577	0.611
mammography	rod	0.463	N/A	0.292	0.292
fraud	iforest	0.425	N/A	0.425	0.283
fraud	copod	0.590	N/A	0.502	0.469
fraud	hbos	0.403	N/A	0.409	0.230
fraud	ecod	0.601	N/A	0.530	0.437
fraud	lunar	0.483	N/A	0.458	0.557
fraud	rod	0.530	N/A	0.462	0.567

Table 16: F_1 values (methods PARs–TMQ).

Dataset	Detector	PARs	RAE-GA	WANDER	TMQ
glass	iforest	0.407	0.184	0.440	0.593
glass	copod	0.479	0.409	0.412	0.691
glass	hbos	0.377	0.164	0.586	0.488
glass	ecod	0.514	0.311	0.378	0.645
glass	lunar	0.493	0.356	0.423	0.605
glass	rod	0.387	0.500	0.089	0.256
glass	mcd	0.387	0.044	0.267	0.406
wpbc	iforest	0.323	0.399	0.439	0.752
wpbc	copod	0.293	0.392	0.426	0.656
wpbc	hbos	0.278	0.364	0.354	0.706
wpbc	ecod	0.244	0.438	0.375	0.742
wpbc	lunar	0.167	0.400	0.426	0.606
wpbc	rod	0.267	0.340	0.472	0.679
wpbc	mcd	0.189	0.339	0.357	0.703
vertebral	iforest	0.333	0.266	0.324	0.661
vertebral	copod	0.286	0.310	0.346	0.710
vertebral	hbos	0.333	0.154	0.328	0.443
vertebral	ecod	0.333	0.645	0.242	0.720
vertebral	lunar	0.333	0.405	0.442	0.437
vertebral	rod	0.333	0.276	0.542	0.664
vertebral	mcd	0.571	0.290	0.509	0.684
pageblocks	iforest	0.235	0.118	0.418	0.657
pageblocks	copod	0.258	0.193	0.161	0.616
pageblocks	hbos	0.250	0.140	0.231	0.529
pageblocks	ecod	0.323	0.172	0.227	0.490
pageblocks	lunar	0.339	0.251	0.392	0.544
pageblocks	rod	0.367	0.270	0.355	0.481
pageblocks	mcd	0.187	0.034	0.057	0.532
fault	iforest	0.245	0.211	0.519	0.559
fault	copod	0.267	0.320	0.555	0.581
fault	hbos	0.199	0.169	0.416	0.529
fault	ecod	0.280	0.358	0.563	0.647

Dataset	Detector	PARs	RAE-GA	WANDER	TMQ
fault	lunar	0.299	0.322	0.647	0.472
fault	rod	0.274	0.391	0.542	0.590
fault	mcd	0.299	0.411	0.437	0.427
yeast	iforest	0.492	0.165	0.496	0.607
yeast	copod	0.465	0.284	0.553	0.611
yeast	hbos	0.468	0.069	0.398	0.509
yeast	ecod	0.458	0.342	0.567	0.691
yeast	lunar	0.486	0.170	0.562	0.546
yeast	rod	0.478	0.209	0.024	0.007
MNIST	iforest	0.174	0.116	0.521	0.618
MNIST	copod	0.194	0.143	0.579	0.603
MNIST	hbos	0.158	0.059	0.390	0.573
MNIST	ecod	0.167	0.203	0.573	0.612
MNIST	lunar	0.177	0.149	0.543	0.532
MNIST	rod	0.237	0.262	0.579	0.599
MNIST	mcd	0.101	0.421	0.400	0.470
annthyroid	iforest	0.622	0.144	0.519	0.733
annthyroid	copod	0.602	0.176	0.342	0.626
annthyroid	hbos	0.114	0.346	0.477	0.665
annthyroid	ecod	0.538	0.586	0.355	0.466
annthyroid	lunar	0.378	0.290	0.615	0.621
annthyroid	rod	0.434	0.632	0.680	0.671
annthyroid	mcd	0.045	0.198	0.335	0.268
satimage-2	iforest	0.424	0.182	0.424	0.646
satimage-2	copod	0.432	0.192	0.233	0.619
satimage-2	hbos	0.355	0.039	0.365	0.607
satimage-2	ecod	0.495	0.123	0.251	0.565
satimage-2	lunar	0.406	0.199	0.491	0.477
satimage-2	rod	0.565	0.213	0.674	0.679
satimage-2	mcd	0.392	0.280	0.220	0.500
smtp	iforest	0.595	0.233	0.667	0.808
smtp	copod	0.762	0.944	0.614	0.639
smtp	hbos	0.146	0.000	0.690	0.794
smtp	ecod	0.771	0.267	0.603	0.164
smtp	lunar	0.375	0.422	0.413	0.385
smtp	rod	0.429	0.317	0.586	0.521
smtp	mcd	0.468	0.381	0.502	0.535
mammography	iforest	0.416	0.627	0.427	0.636
mammography	copod	0.437	0.420	0.342	0.609
mammography	hbos	0.373	0.532	0.484	0.540
mammography	ecod	0.437	0.583	0.387	0.664
mammography	lunar	0.400	0.596	0.461	0.615
mammography	rod	0.229	0.312	0.400	0.420
fraud	iforest	0.327	N/A	N/A	0.552
fraud	copod	0.451	N/A	N/A	0.630
fraud	hbos	0.309	N/A	N/A	0.488
fraud	ecod	0.447	N/A	N/A	0.658
fraud	lunar	0.429	N/A	N/A	0.484
fraud	rod	0.430	N/A	N/A	0.518

Table 17: Consistency across all original datasets and available methods.

Dataset	Detector	TMQ	ATON	LIME	SHAP	COIN	WANDER
glass		1.00	0.98	0.99	0.96	1.00	0.95
wpbc		1.00	0.98	0.99	1.00	0.95	0.97

Dataset	Detector	TMQ	ATON	LIME	SHAP	COIN	WANDER
vertebral		1.00	0.97	1.00	1.00	0.98	0.97
pageblocks		1.00	0.96	0.99	1.00	0.95	0.94
fault		1.00	0.98	0.99	1.00	1.00	0.97
yeast		1.00	0.96	0.99	1.00	0.99	0.97
MNIST		0.97	0.98	0.99	1.00	1.00	0.96
annthyroid		0.96	0.97	1.00	1.00	1.00	0.92
satimage-2		0.95	0.98	0.94	1.00	0.96	0.94
smtp		0.97	0.98	0.98	0.87	0.97	0.97
mammography		0.97	0.97	0.99	1.00	N/A	0.94
fraud		0.97	0.98	0.96	1.00	N/A	N/A

Table 18: Robustness across all original datasets and available methods.

Dataset	Detector	TMQ	ATON	LIME	SHAP	COIN	WANDER
glass		1.00	0.96	0.98	0.98	0.92	0.97
wpbc		1.00	0.97	0.97	1.00	0.86	0.97
vertebral		1.00	0.96	0.98	0.99	0.96	0.98
pageblocks		1.00	0.96	0.93	0.98	0.57	0.95
fault		1.00	0.98	0.98	0.99	0.94	0.97
yeast		1.00	0.96	0.79	0.99	0.92	0.96
MNIST		0.97	0.98	0.98	1.00	0.97	0.94
annthyroid		0.96	0.97	1.00	0.99	0.98	0.93
satimage-2		0.95	0.98	0.93	1.00	0.96	0.96
smtp		0.98	0.98	0.97	0.66	0.97	0.96
mammography		0.97	0.97	0.88	1.00	N/A	0.95
fraud		0.97	0.98	0.96	1.00	N/A	N/A

Table 19: Sparsity across all original datasets and available methods.

Dataset	Detector	TMQ	ATON	LIME	SHAP	COIN	WANDER
glass		5.00	4.00	3.00	2.00	2.00	6.00
wpbc		7.00	8.00	3.00	3.00	7.00	6.00
vertebral		4.00	4.00	2.00	2.00	2.00	3.00
pageblocks		7.00	7.00	4.00	9.00	3.00	3.00
fault		9.00	8.00	6.00	9.00	7.00	9.00
yeast		5.00	5.00	7.00	5.00	6.00	7.00
MNIST		9.00	8.00	9.00	6.00	9.00	9.00
annthyroid		5.00	4.00	5.00	5.00	4.00	5.00
satimage-2		6.00	8.00	5.00	9.00	9.00	9.00
smtp		1.00	3.00	2.00	1.00	3.00	2.00
mammography		3.00	3.00	3.00	5.00	N/A	3.00

Dataset	Detector	TMQ	ATON	LIME	SHAP	COIN	WANDER
fraud		6.00	6.00	6.00	7.00	N/A	8.00

Table 20: Stability across all original datasets and available methods.

Dataset	Detector	TMQ	ATON	LIME	SHAP	COIN	WANDER
glass		0.57	0.20	0.42	0.38	0.39	0.46
wpbc		0.57	0.16	0.38	0.33	0.20	0.54
vertebral		0.78	0.29	0.62	0.90	0.50	0.72
pageblocks		0.73	0.26	0.71	0.50	0.89	0.47
fault		0.47	0.18	0.37	0.25	0.18	0.49
yeast		0.49	0.17	0.44	0.23	0.17	0.48
MNIST		0.41	0.20	0.45	0.31	0.16	0.17
annthyroid		0.35	0.30	1.00	0.79	0.42	0.36
satimage-2		0.60	0.60	0.79	0.79	0.37	0.55
smtp		0.62	0.58	1.00	1.00	0.68	0.64
mammography		0.49	0.24	0.76	0.42	N/A	0.23
fraud		0.21	0.17	0.24	0.22	N/A	N/A

Table 21: Time across all original datasets and available methods.

Dataset	Detector	TMQ	ATON	LIME	SHAP	COIN	WANDER
glass		61.35	4.87	111.43	2.15	9.72	7.55
wpbc		577.35	299.12	854.71	149.75	61.63	37.20
vertebral		536.09	54.49	331.18	4.76	43.52	23.96
pageblocks		12487.90	452.92	3264.01	1994.02	9806.67	92794.42
fault		6744.84	343.81	7709.64	3117.58	1707.34	4961.54
yeast		8828.62	491.52	2621.21	577.45	1168.35	2861.52
MNIST		10314.61	356.25	10397.50	2271.71	20290.65	223941.96
annthyroid		15670.73	631.49	6632.85	476.74	17950.78	143636.98
satimage-2		1501.20	95.33	1334.03	136.85	3046.23	8686.76
smtp		543.30	74.50	218.97	18.09	557.31	7383.20
mammography		6255.80	241.79	773.69	96.04	N/A	336039.73
fraud		8856.73	609.01	347.00	3953.62	N/A	N/A

Table 22: Mem across all original datasets and available methods.

Dataset	Detector	TMQ	ATON	LIME	SHAP	COIN	WANDER
glass		2582.91	541.84	1655.73	60.14	251.47	130.96
wpbc		248.50	128.98	1926.59	877.42	194.11	312.71
vertebral		2188.11	267.45	87.80	209.36	2.45	3294.02
pageblocks		7175.92	182.46	203.91	655.50	1564.86	2991.70

Dataset	Detector	TMQ	ATON	LIME	SHAP	COIN	WANDER
fault		7343.32	1532.54	9391.69	2047.99	46.90	732.77
yeast		973.60	354.03	1525.11	413.49	16.61	57.81
MNIST		759.33	5.79	3120.93	1956.94	2309.45	3812.70
annthyroid		2841.97	1681.79	206.04	3535.11	6517.00	4400.00
satimage-2		286.62	60.91	1423.56	256.17	138.76	8.82
smtp		1122.52	4540.08	2658.51	829.45	1777.20	1951.56
mammography		230.55	362.20	168.80	450.53	N/A	171.60
fraud		1112.81	3220.19	1141.00	1179.75	N/A	N/A

F Reported complexity expression

Algorithm 2 presents the procedure of TMQ. In Steps (1-2), TMQ samples the data into three candidate sets and forms a set of quadruplets, resulting in a complexity of $O(|\mathbf{X}_+ \cup \mathbf{X}_-|dk_1k_2^2)$. In Steps (3-11), TMQ captures the relationships among the features, yielding a complexity of $O(d|\mathcal{S}|k_1k_2^2 + k_1k_2^2|\mathcal{S}|^2 + k_1^2k_2^4|\mathcal{S}|)$. In Steps (11-13), the complexity is $O(k_1k_2^2|\mathcal{S}|)$. For simplicity, we do not explicitly present batch training process. In practical usage, TMQ can be trained by setting proper epoch number and batch size. The overall time complexity of TMQ is represented as $O(|\mathbf{X}_-|dk_1k_2^2 + d|\mathcal{S}|k_1k_2^2 + k_1k_2^2|\mathcal{S}|^2 + k_1^2k_2^4|\mathcal{S}|)$. Hyper-parameters k_1 and k_2 are normally set below 60 and $|\mathcal{S}|$ is 64 or $\lfloor 1.5d \rfloor$, and $|\mathbf{X}_+| \ll |\mathbf{X}_-|$ as well. Thus, TMQ has quadratic time complexity w.r.t. data dimensionality and approximate linear time complexity w.r.t. data size.

G Full experimental record and additional figures

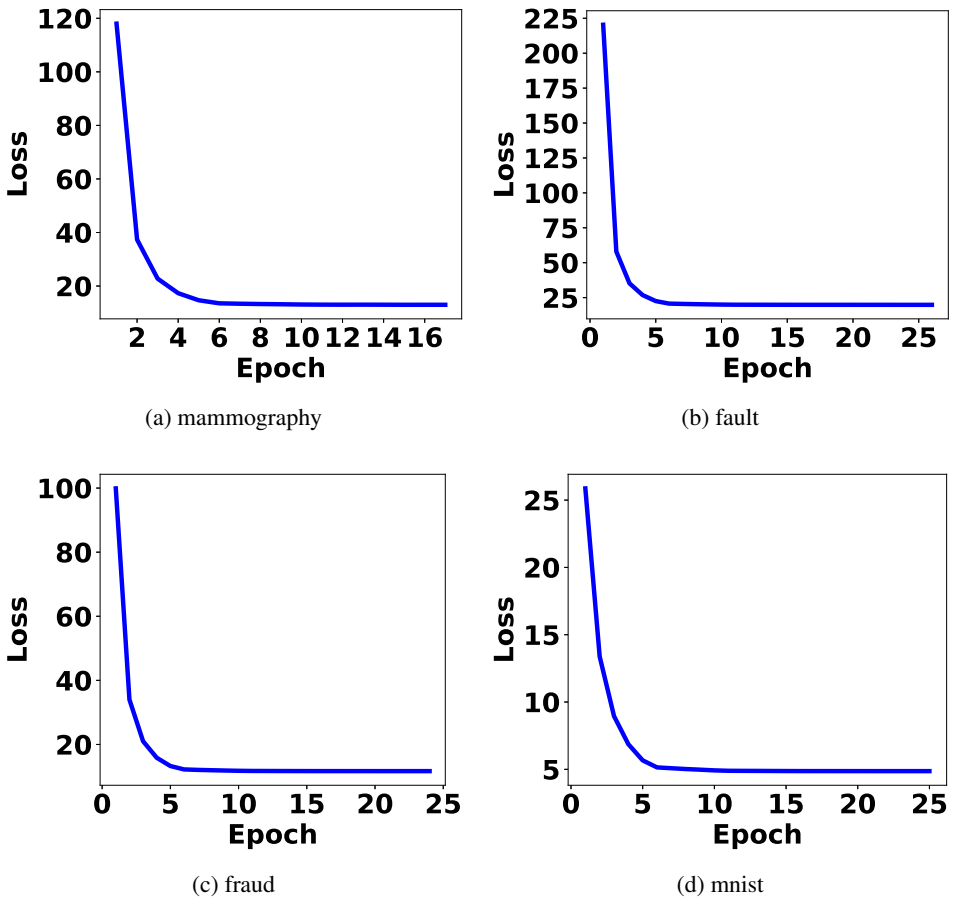
G.1 Experimental Settings

Datasets. We evaluate the proposed method on twelve benchmark datasets, which are sourced from the ADBench benchmark[1]. The selection covers multiple observed ranges of: (i) data dimensionality (3 to 512), (ii) outlier ratio (0.03% to 34.67%), and (iii) diverse application domains, including finance, healthcare, industrial monitoring, remote sensing, and e-commerce. Although conducting experiments on all 57 ADBench datasets is computationally prohibitive due to the iterative Transformer-guided optimization in our approach, the selected datasets provide diverse regimes for the reported comparison. Table 7 summarizes the detailed statistics of all datasets used in the evaluation.

Detector-derived annotation proxies. We derive feature-subspace annotation proxies from ECOD Li et al. (2022), iForest Liu et al. (2012), COPOD Li et al. (2020), HBOS Goldstein & Dengel (2012), LUNAR Goodge et al. (2022), ROD Almardeny et al. (2020), and MCD Hardin & Rocke (2004). The maximum-score-subset description does not specify a feasible candidate family or dimensionality-reduction procedure for 512-dimensional MNIST. Exact search over all feature subsets would require $2^{512} - 1$ evaluations. These are detector-preference proxies; the complete candidate-generation and preprocessing protocol should be checked against the released scripts before treating the high-dimensional results as exact ground-truth explanations.

Baselines. To ensure a comprehensive and up-to-date comparison, we evaluate TMQ against a diverse set of outlier interpretation methods from two complementary perspectives: classical baselines and recently proposed approaches. The rationale behind this selection is to capture both well-established techniques that have been widely validated in the literature and emerging methods that reflect recent advances in the field, thereby enabling a thorough assessment of TMQ's relative strengths and generalizability. Specifically, we compare with nine representative baselines: ATON Xu et al. (2021), COIN Liu et al. (2018), SHAP Lundberg & Lee (2017), LIME Ribeiro et al. (2016), SiNNE Samariya & Ma (2022), Anchor Ribeiro et al. (2018), RAE-GA Tousi & DeSouza (2024), PARs Feng (2024), and WANDER Guan et al. (2025). These methods span a broad spectrum of

[1] https://github.com/Minqi824/ADBench



(a) mammography

(b) fault

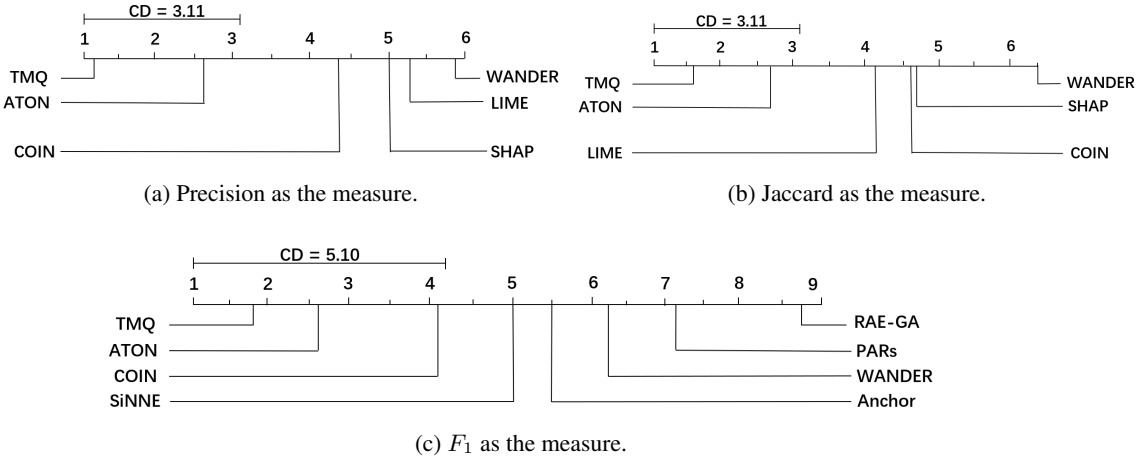
(c) fraud

(d) mnist

Figure 4: Loss curves for four real-world datasets.

interpretation paradigms, including feature weighting, subspace mining, rule-based explanations, and latent space learning. This diverse set of competitors not only covers a wide range of technical approaches, but also includes state-of-the-art (SOTA) methods in outlier interpretation, providing a reliable and rigorous benchmark for evaluating the effectiveness and validity of TMQ.

Parameter Settings & Implementations. For TMQ, we use the Adam optimizer (initial learning rate 0.1) stabilized by a 5% linear warm-up, gradient clipping (max norm 1.0), and learning rate decay on plateau, alongside early stopping (patience 5, max 40 epochs, typically converging in 15–25) and a batch size of 512, with hyper-parameters set to $k_1 = 1$, $k_2 = 50$, latent dimension $|\mathcal{S}| = \lfloor 1.5d \rfloor$, and Transformer hidden layers of $2|\mathcal{S}|$ and $\lfloor 1.5|\mathcal{S}| \rfloor$. To ensure fair comparison, we configure the baselines as follows: ATON Xu et al. (2021) uses $r = 30$, $\alpha = 0.8$, $d = 64$, and 10 epochs; Anchors Ribeiro et al. (2018) sets beam width 10 and coverage threshold 0.95; COIN Liu et al. (2018) tunes the ℓ_1-SVM regularization $C \in \{0.1, 1, 10, 100\}$ via 5-fold cross-validation; LIME Ribeiro et al. (2016) uses an RBF-SVM surrogate with 5000 perturbed samples and explanation features matching the ground-truth size; SHAP Lundberg & Lee (2017) employs TreeSHAP/KernelSHAP with 5000 background samples; SiNNE Samariya & Ma (2022) sets ensemble 100, sampling 8, and beam width 10; PARs Feng (2024) uses $\theta = \max(10/|\mathcal{D}|, 0.01)$, $\gamma = 0.9$, and $\lambda = 5$; RAE-GA Tousi & DeSouza (2024) configures 100 generations, population 20, uniform crossover, mutation probability 0.1, tournament $k = 5$, regularization $\alpha = 0.02$, and $k_i = 3$; and WANDER Guan et al. (2025) uses Adam (learning rate 0.1) with early stopping, $k_1 = 1$, $k_2 = 50$, and bandwidth 0.68. All experiments are executed on an Ubuntu 20.04 system equipped with four GeForce RTX 2080 Ti GPUs and 512-GB memory.



(a) Precision as the measure.

(b) Jaccard as the measure.

(c) F_1 as the measure.

Figure 5: Critical difference diagrams of Nemenyi test. The average ranks for each method on 12 datasets are shown.

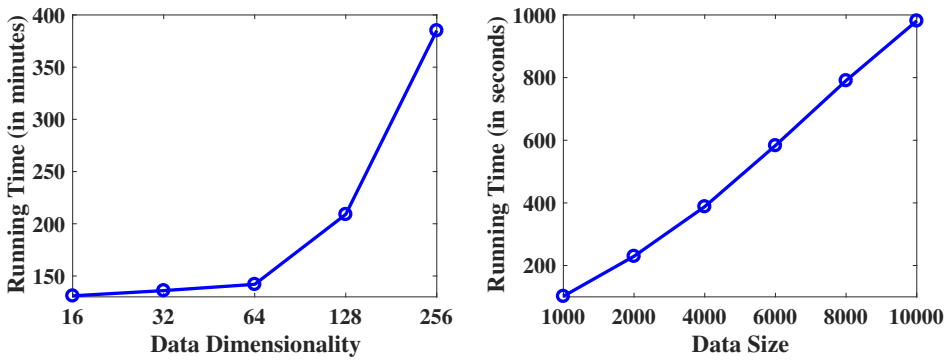


Figure 6: Running time w.r.t. data dimensionality and data size.

Evaluation Metrics. For a shared subspace comparison we convert every method's output to a feature subspace. This measures overlap under one interface, while rankings and rules require separate native-format assessment. For methods directly producing subspaces, we evaluate interpretability using the F_1 score based on ground-truth $\mathcal{G}$ and predicted $\mathcal{P}$, where precision $= |\mathcal{G} \cap \mathcal{P}|/|\mathcal{P}|$ and recall $= |\mathcal{G} \cap \mathcal{P}|/|\mathcal{G}|$. For methods outputting feature weight vectors, we convert them to subspaces by selecting the top-k features (with $k = |\mathcal{G}|$), following Xu et al. (2021); Guan et al. (2025). Subspace quality is then assessed via Precision (which equals recall since $|\mathcal{P}| = |\mathcal{G}| = k$) and the Jaccard Index ($|\mathcal{G} \cap \mathcal{P}|/|\mathcal{G} \cup \mathcal{P}|$). Additionally, for weight-based methods, we evaluate ranking quality independently of subspace size using AUPR and AUROC. Finally, we conduct comprehensive fidelity tests for weight-based methods–including insertion/deletion curves, stability (bootstrap Jaccard), sparsity profiling, and robustness across detector annotations–since their continuous importance scores specifically require verification of sensitivity, consistency, conciseness, and reliability, concerns that are less applicable to discrete subspace methods.

G.2 Effectiveness Evaluation

Experimental Analysis. We report ten runs per dataset and mark computations exceeding two weeks as N/A. Tables 8–16 give the full detector-level outcomes. TMQ has the strongest averages on several metrics, while deletion AUPRC and some individual cases favor comparators. These scores measure agreement with detector-derived annotation proxies and response tests under the

Table 23: Average performance of ablation study and improvement rates of TMQ over its ablated versions.

Method	AUPR	AUROC	F_1
TMQ	**0.692**	**0.753**	**0.571**
TMQ-global	0.688 (0.6%)	0.730 (3.2%)	0.561 (1.8%)
TMQ-local	0.690 (0.3%)	0.732 (2.9%)	0.565 (1.2%)
TMQ-trans	0.628 (9.6%)	0.643 (13.5%)	0.516 (10.6%)
TMQ-quad	0.310 (122.6%)	0.391 (87.2%)	0.270 (109.3%)

conversion protocol above. Figure 4 displays four training-loss trajectories; their stabilization is an empirical observation rather than a global convergence result.

Analysis of Quadruplet Loss Regimes and Case Study. To empirically validate our theoretical analysis, we conduct a 3D t-SNE latent space visualization (Fig. 7) across three representative datasets covering key regimes: fraud (heterogeneous outliers), mammography (conditional drift), and glass (extreme sparsity), excluding COIN, RAE-GA, and WANDER due to scalability constraints. The results demonstrate that for fraud, TMQ effectively separates the query outlier from inliers while keeping other anomalies proximate to the normal distribution, enabling differentiated explanations. In mammography, TMQ preserves outlier cohesion and ensures clear inlier separation, supporting context-aware interpretations. Conversely, in the extremely sparse glass dataset, all methods exhibit limited differentiation, with TMQ only marginally outperforming ATON. These patterns are consistent with explicit separation among outliers being useful for heterogeneous or conditionally varying anomalies but yields diminishing returns under extreme sparsity, where insufficient data fundamentally constrains the learning of meaningful separations.

Statistical Analysis. We convert all methods' outputs into feature subspaces and display Nemenyi critical-difference diagrams for F_1, Precision, and Jaccard (Fig. 5). These ranks summarize the common-format comparisons at the stated test level. Connected ranks within the critical difference should not be called definitively distinct from the diagram alone. The exact paired units, ties, missing values, and native-format quality need separate verification; the diagram is a statistical summary, not a per-method verdict.

Running Time and Memory Analysis. Table 17 lists the resource, consistency, robustness, stability, and sparsity entries. TMQ uses about 1–7 GB over the evaluated settings, versus 46.9 MB for COIN in one case; WANDER reaches much higher memory in some cases. Figure 6 reports measured time trends over the plotted range. The complexity expression includes both $q|\mathcal{S}|^2$ and $q^2|\mathcal{S}|$ terms for $q = k_1 k_2^2$, so approximate linear scaling with dataset size is conditional on a fixed sampling/batching budget and implementation orientation. No million-instance run is included.

Ablation Study. To validate the significance of TMQ's core components, we evaluate four ablated variants: *TMQ-global* (removes global attention, retaining local), *TMQ-local* (removes local attention, retaining global), *TMQ-trans* (replaces the global-local Transformer with a parameter-matched three-layer MLP), and *TMQ-quad* (replaces the quadruplet-based loss with a triplet deviation-based loss Xu et al. (2021)). As detailed in Table 23, TMQ consistently and substantially outperforms all four variants across every dataset and evaluation metric. The large triplet ablation effect and smaller global/local branch effects should be considered separately when interpreting the architectural contribution.

Parameter Test. The sensitivity study examines k_1, k_2, latent dimension, and batch size on five selected datasets, with the full AUPR/AUROC/F_1 panels in Fig. 8. It discusses $k_1 = 3$, $k_2 = 50$, $|\mathcal{S}| = 64$, and batches between 128 and 512 as tested choices. These are empirical settings within the plotted study, not a proof of an optimum, manifold recovery, freedom from overfitting, or general convergence. The primary benchmark configuration above states $k_1 = 1$ and $|\mathcal{S}| = \lfloor 1.5d \rfloor$; the settings for each reported run should be checked against run logs.

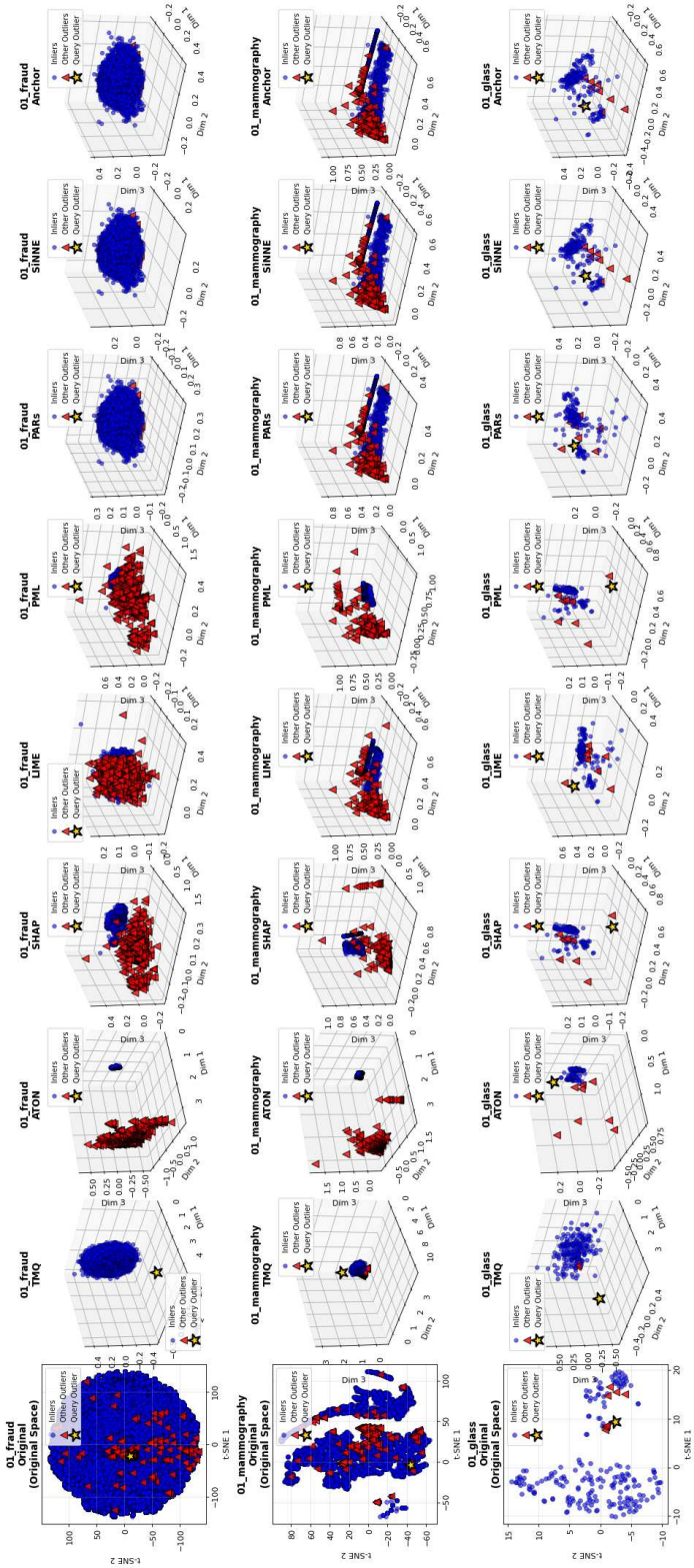


Figure 7: 3D t-SNE visualizations of latent representations learned by different interpretation methods across three outlier regimes. Each row corresponds to a dataset (fraud, mammography, glass) and each column to a method (Original, TMQ, ATON, SHAP, LIME, PARs, SiNNE, Anchor). "★" is a query outlier, "▲" is another outlier, and "●" is an inlier.

Table 24: Per-dataset subspace separability failure over all 47 ADBench Classical benchmarks. Features are min-max normalized; up to 2,000 inliers and 2,000 outliers are drawn with seed 0 where a dataset exceeds those caps, and the counts used are reported. "Bench. n" is the total number of instances, "Outl." the number of outliers evaluated. "Budget" gives the searched family: *all* 2^d-1 enumerates every non-empty subspace, so the failure rate is exact and no search could do better; *every* $|S| \leq 3$ enumerates every subset of size at most three; *sampled* $|S| \leq 3$ draws 30,000 such subsets uniformly for dimensions where even that enumeration is unaffordable, so those rates are upper bounds on the true failure rate. Reference levels: *all* requires the query to be more isolated than every inlier, *q99* and *q95* than the 99th and 95th percentile of inlier isolation.

Dataset	n	d	**Outl.**	**Budget**	**Fail all**	**Fail q99**	**Fail q95**
http	3	567498	2000	all 2^d-1	98.5	0.0	0.0
skin	3	245057	2000	all 2^d-1	95.8	9.5	0.8
smtp	3	95156	30	all 2^d-1	33.3	26.7	23.3
Wilt	5	4819	257	all 2^d-1	100.0	98.8	47.5
annthyroid	6	7200	534	all 2^d-1	88.8	29.8	0.2
mammography	6	11183	260	all 2^d-1	78.8	28.1	11.2
thyroid	6	3772	93	all 2^d-1	81.7	4.3	0.0
vertebral	6	240	30	all 2^d-1	100.0	96.7	73.3
glass	7	214	9	all 2^d-1	77.8	66.7	0.0
Pima	8	768	268	all 2^d-1	89.2	56.0	22.0
yeast	8	1484	507	all 2^d-1	96.6	84.2	64.5
Stamps	9	340	31	all 2^d-1	83.9	0.0	0.0
WBC	9	223	10	all 2^d-1	10.0	0.0	0.0
breastw	9	683	239	all 2^d-1	11.3	1.7	0.0
shuttle	9	49097	2000	all 2^d-1	3.5	0.0	0.0
PageBlocks	10	5393	510	all 2^d-1	68.0	14.1	2.0
cover	10	286048	2000	all 2^d-1	4.1	0.1	0.0
donors	10	619326	2000	all 2^d-1	5.8	0.0	0.0
magic.gamma	10	19020	2000	all 2^d-1	66.8	28.7	4.9
vowels	12	1456	50	all 2^d-1	2.0	0.0	0.0
wine	13	129	10	all 2^d-1	0.0	0.0	0.0
pendigits	16	6870	156	all 2^d-1	0.0	0.0	0.0
Lymphography	18	148	6	all 2^d-1	0.0	0.0	0.0
Hepatitis	19	80	13	all 2^d-1	0.0	0.0	0.0
Cardiotocography	21	2114	466	every $\lvert S\rvert \leq 3$	40.1	8.6	0.9
Waveform	21	3443	100	every $\lvert S\rvert \leq 3$	71.0	7.0	0.0
cardio	21	1831	176	every $\lvert S\rvert \leq 3$	2.8	0.0	0.0
ALOI	27	49534	1508	every $\lvert S\rvert \leq 3$	90.5	59.9	24.7
fault	27	1941	673	every $\lvert S\rvert \leq 3$	70.7	13.1	1.0
fraud	29	284807	492	every $\lvert S\rvert \leq 3$	15.0	5.1	0.0
WDBC	30	367	10	every $\lvert S\rvert \leq 3$	0.0	0.0	0.0
Ionosphere	32	351	126	every $\lvert S\rvert \leq 3$	1.6	0.8	0.0
letter	32	1600	100	every $\lvert S\rvert \leq 3$	6.0	0.0	0.0
WPBC	33	198	47	every $\lvert S\rvert \leq 3$	83.0	57.4	6.4
landsat	36	6435	1333	every $\lvert S\rvert \leq 3$	54.2	28.8	1.1
satellite	36	6435	2000	every $\lvert S\rvert \leq 3$	32.4	10.8	0.1
satimage-2	36	5803	71	every $\lvert S\rvert \leq 3$	1.4	0.0	0.0
celeba	39	202599	2000	every $\lvert S\rvert \leq 3$	74.9	73.4	73.4
SpamBase	57	4207	1679	every $\lvert S\rvert \leq 3$	65.2	2.6	0.7
campaign	62	41188	2000	every $\lvert S\rvert \leq 3$	40.9	1.4	0.0
optdigits	64	5216	150	every $\lvert S\rvert \leq 3$	0.0	0.0	0.0
musk	166	3062	97	sampled $\lvert S\rvert \leq 3$	0.0	0.0	0.0
backdoor	196	95329	2000	sampled $\lvert S\rvert \leq 3$	10.9	4.1	0.0
speech	400	3686	61	sampled $\lvert S\rvert \leq 3$	42.6	3.3	0.0
census	500	299285	2000	sampled $\lvert S\rvert \leq 3$	39.3	11.0	8.6
mnist	512	10000	500	sampled $\lvert S\rvert \leq 3$	6.3	0.0	0.0
InternetAds	1555	1966	368	sampled $\lvert S\rvert \leq 3$	16.8	13.6	13.6

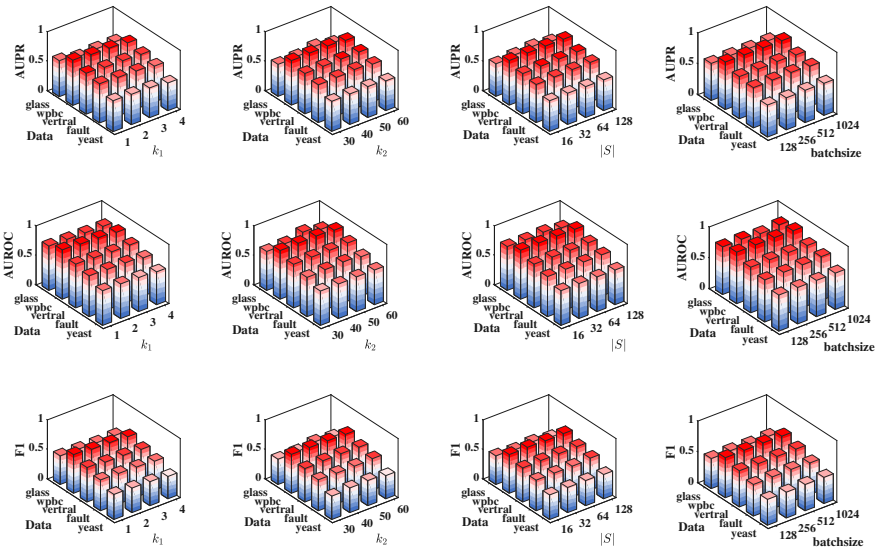


Figure 8: Parameter test results w.r.t. k_1, k_2, $|\mathcal{S}|$, and batch size.

H SUBSPACE SEPARABILITY MEASUREMENT

This appendix records the measurement summarized in Table 1 of the main text, with the complete per-dataset results in Table 24, so that the claim about the axis-aligned subspace format can be checked or contested.

Setup. All 47 datasets of the ADBench Classical split are used, downloaded from the benchmark's public release and stored as `.npz` arrays; every dataset in the main comparison is included among them. Features were min-max normalized per column using the minimum and maximum over all rows, matching the preprocessing used by the released TMQ implementation. Where a dataset has more than 2000 inliers or more than 2000 labelled outliers, 2000 of each are drawn uniformly at random with seed 0; the counts actually used appear in Table 24 so no rate silently rests on a subsample the reader cannot see. The inlier cap matters little because nearest-neighbour distances in low-dimensional subspaces are insensitive to subsample size; the outlier cap affects only the 11 largest sets (skin, donors, census, magic.gamma, campaign, celeba, shuttle, cover, backdoor, http, satellite), for which the reported rate covers the sampled 2000 rather than all $146,016$ labelled outliers. The same draw is used for all subspaces of a dataset. Distances are Euclidean in the projected subspace, computed in exact float64 arithmetic: a float32 gemm shortcut was tested and rejected because its cancellation error flips borderline comparisons and failed to reproduce published rates.

Criterion. Let $S \subseteq \{1, \dots, d\}$, $S \neq \emptyset$, be a coordinate subset, X_- the inliers and x the query outlier. Write $d_S(u, V)$ for the Euclidean distance between the projections of u and its nearest point in V onto S. For an inlier $p \in X_-$ define its isolation scale $\iota_S(p) = d_S(p, X_- \setminus \{p\})$, the distance to its nearest *other* inlier. We say S *separates* x at reference level q if

$$d_S(x, X_-) \;>\; \mathrm{Quantile}_q\big(\{\iota_S(p) : p \in X_-\}\big), \tag{22}$$

with $q = 100$ (the maximum) giving the strict reading "more isolated than every inlier" used for the main-text headline figure, and $q \in \{99, 95\}$ giving permissive readings. An outlier *fails* when no non-empty S satisfies Eq. (22).

Search budget, and why it differs by dimension. Three budgets are used, and each rate is only as strong as its budget allows:

- *All* $2^d - 1$ *subsets* for every $d \leq 20$ (24 datasets). No pruning, beam, cardinality cap or scoring heuristic is involved, so a failure is a statement about the axis-aligned family itself:

no better search algorithm can succeed on such a query, because nothing in that family succeeds. These are the only exactly interpretable rows.

- *Every* $|S| \leq 3$ *subset* for $21 \leq d \leq 64$ where that enumeration is affordable (17 datasets, up to 30,913 subsets). Failure rates here are *upper bounds*: searching more coordinates could only separate more queries. The bound is nonetheless the relevant one for practice, since any deployed search must cap $|S|$ or prune.
- 30,000 *uniformly sampled* $|S| \leq 3$ *subsets* for $d > 64$ (6 datasets, up to InternetAds at d=1555), also upper bounds and the loosest of the three.

Because a smaller searched family can only find fewer separating subspaces, the direction of every approximation is known: sampled and capped budgets overstate failure, never understate it. No headline figure in the main text is taken from a sampled row.

Results. Pooled over all 47 benchmarks and 33,170 evaluated outliers, 47.7% fail under the strict criterion, 16.7% at $q = 99$ and 8.4% at $q = 95$. Restricting to the 24 benchmarks where every non-empty subset was enumerated (15,013 outliers), the rates are 51.2%, 13.0% and 4.6% respectively. The distribution is strongly bimodal rather than uniformly high: 17 benchmarks fail for fewer than 10% of their outliers and 2 fail for all of them, with a per-dataset median of 39.3%.

The dimension dependence runs in the opposite direction from what a "high-dimensional search is hopeless" reading would predict, and only within a fixed budget. Among the 24 exhaustively searched benchmarks, failure falls as dimension rises (Spearman $\rho = -0.83$, $p = 7e-07$): at $d \leq 6$ there are just 31–63 candidate subsets and 94.9% of outliers fail, at 7–12 there are up to 4095 and 28.5% fail, and at 13–20 exhaustive search separates every outlier, 0.0% failing over all 185 of them. Among the 17 benchmarks sharing the $|S| \leq 3$ budget the association is absent ($\rho = -0.12$, $p = 0.64$), because that budget is fixed while d varies from 21 to 64. The correct reading is therefore that feasibility, not dimensionality alone, governs the outcome: exhaustive search is cheap enough to succeed in a middle band, too small a candidate family dooms the low band, and no affordable budget suffices in the high band.

Two benchmarks fail completely under exhaustive search: *vertebral* (d=6, 63 subsets, all 30 outliers, 73.3% failing even at $q = 95$) and *Wilt* (d=5, 31 subsets, all 257 outliers, 47.5% at $q = 95$). At the other end, *wine*, *pendigits*, *Hepatitis*, *Lymphography*, *WDBC*, *musk* and *optdigits* fail for none of their outliers, which is why the main text reports this as a characterization of regimes rather than a universal impossibility.

Failure is not an artifact of how many inliers were sampled: the strict rate is uncorrelated with n_{in} used (Spearman $\rho = +0.14$, $p = 0.37$), and both extremes occur at the 2000-inlier cap (Wilt 100%, pendigits 0%) as well as at small n_{in} (WPBC 83.0% with 151 inliers, glass 77.8% with 205).

Relation to the empirical comparison, and a negative control. The measurement bounds what the subspace *format* can express; it does not predict by how much any given method improves. Pairing per-dataset strict failure rates with per-dataset mean F_1 from Table 4 of the main text gives Spearman $\rho = +0.22$ ($p = 0.52$, $n = 11$) against TMQ's gain over WANDER, and $\rho = +0.08$ ($p = 0.81$, $n = 12$) against TMQ's absolute F_1. Neither association is significant, so we report the two findings separately and do not claim that TMQ's advantage is largest where separability fails; TMQ nonetheless leads WANDER on 10 of the 11 datasets where both were run (fraud has no WANDER result). What the two measurements jointly support is weaker and safer: the subspace format fails for a large and dimension-dependent share of outliers, and TMQ outperforms the subspace-output baselines on the shared protocol, with the mechanism of that lead established by the ablation in Sec. G.2 rather than by the failure rates. Note that separability is a property of the *data* and so is available for all 47 ADBench Classical benchmarks, whereas the F_1 comparison requires model runs and is available only for the twelve evaluated datasets; the two cover different dataset sets and are never pooled into a single statistic.

Scope. Eq. (22) formalizes an isolation-style reading of "distinguishes the query from the normals", which is the criterion score-and-search methods optimize. Methods whose explanation target differs, rules with coverage guarantees (Anchors), attribution rankings evaluated by insertion/deletion curves (SHAP, LIME), are not refuted by this measurement, since a ranking can be

informative even when no subset passes an isolation test. The measurement is also specific to axis-aligned subsets; a criterion admitting oblique directions would not fail in the same way, which is precisely the expressiveness gap that motivates learning a projection. Two further limits are worth stating. First, the reference is detector-independent: it uses the labels the benchmark ships, so a failure does not imply that any particular detector was wrong to flag the query. Second, the caps on inliers and outliers mean the pooled rates are weighted by the number of outliers actually evaluated (33,170 of the 146,016 labelled outliers across the suite), so they describe the evaluated sample rather than every ADBench instance; Table 24 reports both totals and evaluated counts so the weighting can be inspected.

I PROOFS AND SCOPE OF THE RELATIONAL OBJECTIVE

Full proof of Proposition 1. Let $u = D_{ii}$, $v = D_{qo}$, and $w = D_{qi}$. The three summands of Eq. 2 are $[u - v + m]_+$, $[u - w + m]_+$, and $[v - w + m]_+$. Because all three are nonnegative, their sum is zero exactly when $v \geq u + m$, $w \geq u + m$, and $v \leq w - m$. Combining the first and third gives $w \geq u + 2m$, which implies the second. Thus v must lie in $[u + m, w - m]$ and the interval is nonempty precisely when $w \geq u + 2m$. Conversely, every v in that interval makes all three hinges zero. The triplet objective $[u - w + m]_+$ does not contain v; once $w \geq u + m$, it can vanish for any $v \geq 0$. □

Same-subtype conflict. Suppose the desired same-subtype geometry has $D_{qo} \leq \delta$ for a pair whose inlier scale is D_{ii}, with $\delta < D_{ii} + m$. The first hinge satisfies $[D_{ii} - D_{qo} + m]_+ \geq D_{ii} + m - \delta > 0$. No parameter choice can attain zero loss for that quadruplet while respecting this upper bound. This gives a precise reason to avoid unconditional inter-outlier repulsion when the sampled anomalies share a mechanism. We do not implement such a subtype-aware sampler in this work.

Nonconvexity counterexample. Nonconvexity holds even before attention is added. Consider the one-dimensional projection $\phi(z) = wz$, margin $m = 1$, and fixed positions $x = 0$, $x_o = 1/2$, $x_i = 1$, $x_i' = 1$. Then $D_{ii} = 0$, $D_{qo} = w^2/4$, and $D_{qi} = w^2$. For sufficiently small $|w|$, all three hinges are active and $L(w) = 3 - 2w^2$. For $\epsilon > 0$ small enough, $L(0) = 3 > \{L(\epsilon) + L(-\epsilon)\}/2 = 3 - 2\epsilon^2$, violating midpoint convexity. Adding learned attention does not restore a general convexity guarantee. Empirical training traces therefore should not be conflated with proof of global convergence.

J READING EXPLANATIONS IN THEIR OUTPUT FORMAT

An explanation is finally consumed in a concrete output format, and the choice of format determines which questions it can answer. TMQ's latent weights describe how a projection was estimated; an analyst inspecting one instance needs input feature names and values. A top-k set answers a constrained review question, which k recorded fields most distinguish this record, while a full ranking retains information when the inspection budget has not yet been set, at the cost of requiring the reader to choose a cut point. Rule-based or native-format outputs (Anchor rules, SHAP rankings) answer related but distinct questions about predictive behavior rather than detector-referenced contrast, which is why the shared-subspace comparison in the main text is complemented, not replaced, by format-specific caveats.

Eq. 5 of the main text gives one defensible bridge between the learned projection and input fields: each coordinate's possible contribution to a projected query–reference contrast is bounded by $|x_j - b_j| \, \|B_x[:, j]\|_2$. This bound does not uniquely attribute squared distance, because correlated columns may cancel or reinforce one another; it should therefore be read as a screening quantity for ordering candidate fields, not as a decomposition of the detector's score. Before any aggregate score is attributed to this readout, the released implementation should be checked: which ranking (column magnitude, value-weighted magnitude, or a hybrid), which preprocessing (standardization, clipping), and which reference vector b were actually used. The detector proxies support comparison under common feature targets; domain-expert or subtype annotations would further test whether the detected contrasts correspond to actionable anomaly mechanisms. Keeping the numerical evidence,

the algebraic geometry, and the semantic reading explicitly separated prevents the projection indices from being over-read as causal or detector-internal attributions.

Where the extra relation is and is not useful. The empirical breadth and the relational analysis point to different next decisions for an analyst. On a dataset with many detected anomaly types, the other-outlier position can help distinguish a query-specific contrast only if that sampled outlier is semantically different from the query. On a sparse dataset such as glass or smtp, there may be too few anomalies to form a stable local comparison, even if normal references are plentiful. Conversely, MNIST's high feature count tests representation capacity but makes the reference-annotation search protocol and memory budget especially consequential. These are not reasons to omit those datasets: the complete dataset and detector-level tables make the differing regimes inspectable. They identify where a query-conditioned projection is a useful hypothesis and where the extra relation requires subtype evidence.