TABULAR ANOMALY DETECTION VIA CAUSAL MATRYOSHKA TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Structural anomalies are samples where individual feature values appear plausible but their combination violates underlying dependencies. Existing anomaly detection (AD) methods often rely on a single fixed latent representation to separate normal and abnormal data, making localized structural violations susceptible to signal dilution in high-dimensional spaces. We propose **NestedAno**, an *unsupervised* tabular AD framework that learns a hierarchy of nested representations, inspired by Matryoshka Representation Learning, and progressively organizes them using topologically aligned causal reconstruction objectives. At inference time, it measures local deviation at each nested scale to construct an *anomaly trajectory* for each sample. The integrated deviation of this trajectory from the normal reference provides the final anomaly score, enabling anomaly evidence to be captured across multiple scales. Extensive experiments on 30 continuous and mixed-type datasets show that NestedAno achieves the highest average AUC-ROC among 17 representative baselines, yielding superior detection accuracy while providing highly interpretable trajectory separation between normal and anomalous data.

1 INTRODUCTION

Anomaly detection (AD) aims to identify rare observations that deviate from normal data and plays a critical role in applications such as fraud detection, healthcare, and cybersecurity. AD often faces severe label scarcity because anomalous samples are rare, diverse, and costly to collect, while available training data are predominantly normal. Consequently, many AD methods adopt an *unsupervised* setting, learning the characteristics of normal data without anomalous examples during training.

A prominent family of tabular AD methods learns compact representations of normal data and detects anomalies through reconstruction or representation-based deviation (Liu et al., 2021; Li et al., 2022; Ye et al., 2025). Such methods are attractive in the unsupervised setting because both representation learning and anomaly scoring can be performed without anomalous training examples. However, they may overlook *structural anomalies*, where individual feature values appear plausible but their combination violates underlying dependencies (e.g., a loan applicant may have a plausible salary but inconsistent with their age and education). CausalAno (Nguyen et al., 2026) addresses this limitation by explicitly incorporating causal dependencies when learning normal tabular structure.

Despite this advantage, CausalAno characterizes normality using a single fixed latent representation and relies on a covariance-based Mahalanobis geometry for anomaly scoring. This design has two limitations. First, a structural violation localized to a small subset of variables may be diluted when all variables are projected into a single high-dimensional representation. Second, the Mahalanobis distance summarizes normality through a single covariance-based geometry, which can be inadequate when the latent distribution is heterogeneous, multimodal, or non-elliptical.

To address these limitations, we propose **NestedAno**, a causally guided nested representation learning framework. Instead of representing each sample with a single latent vector, NestedAno maps it to a hierarchy of nested representations inspired by Matryoshka Representation Learning (Kusupati et al., 2022). Each representation acts as a prefix of the next larger one, providing multiple resolutions of the same sample. It further uses a causal topological ordering to define cumulative variable subsets and impose scale-specific causal reconstruction objectives. This supervision encourages earlier representations to preserve information about upstream variables while larger representations incorporate additional variables. NestedAno employs adversarial learning with a causally structured

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.

generator to provide synthetic samples that encourage a discriminator to distinguish normal observations from structurally inconsistent samples across all representation scales simultaneously. At inference time, it computes multi-scale deviations to form an *anomaly trajectory* that describes how anomaly evidence changes as progressively larger representations are considered. It then integrates the deviations from the normal reference trajectory to obtain the final anomaly score.

To summarize, our main contributions are as follows.

(1) **Novel Framework:** We propose NestedAno, a causally guided nested representation learning framework for tabular AD that constructs a multi-scale anomaly trajectory for each test sample.

(2) **Algorithmic Innovation:** We introduce causal topological supervision to align nested bottlenecks with a causal topological ordering. We additionally introduce a multi-scale anomaly scoring strategy that accumulates structural anomaly evidence across nested representation scales.

(3) **Empirical Validation:** We conduct extensive experiments across 30 tabular datasets. NestedAno achieves an average AUC-ROC of 74.49% on 12 mixed-type datasets, improving upon the strongest baseline CausalAno by ~3%, while maintaining top-tier performance on 18 continuous datasets.

2 Related Works

Existing tabular AD methods characterize normality from different perspectives. Proximity-based approaches such as KNN and IForest (Liu et al., 2008; Ramaswamy et al., 2000) identify isolated samples in the feature space, while distribution-based methods including PCA and ECOD (Shyu et al., 2003; Li et al., 2022) model statistical regularities. Boundary-based methods like DeepSVDD, GOAD, and ICL (Ruff et al., 2018; Bergman and Hoshen, 2020; Shenkar and Wolf, 2022) optimize compact decision boundaries that enclose normal data within a learned latent space. Reconstruction- and generation-based approaches (Pang et al., 2018; Liu et al., 2021; Xu et al., 2023; Qiu et al., 2021; Livernoche et al., 2024; Schlegl et al., 2019; Ye et al., 2025) detect anomalies by measuring deviations from learned data manifolds. Most recently, AnoLLM and LLM-DAS (Tsai et al., 2025; Ye et al., 2026) have adapted Large Language Models for tabular anomaly scoring and detector construction. Most relevant to our work is CausalAno (Nguyen et al., 2026), which integrates causal structure into tabular AD by learning a causal-aware discriminator. However, CausalAno characterizes normality using a single, fixed-capacity latent representation. By compressing complex structural dependencies into one high-dimensional vector, this approach remains susceptible to signal dilution, where localized causal violations are easily obscured by the variance of non-anomalous features. To the best of our knowledge, NestedAno is the first tabular AD framework to exploit Matryoshka-style nested representations. By aligning a causal topological ordering with multi-scale information bottlenecks, it captures anomaly evidence across levels of representational capacity.

3 Framework

3.1 Problem Formulation

Following previous works (Livernoche et al., 2024; Ye et al., 2025; Tsai et al., 2025; Nguyen et al., 2026), we consider *unsupervised* tabular anomaly detection, where only normal samples are available during training. Let $\mathcal{D}_{\text{tr}} = \{\mathbf{x}_i\}_{i=1}^{N}$ denote the training set of N normal records, where each observation $\mathbf{x}_i \in \mathbb{R}^M$ consists of M features. At inference time, we consider the test set $\mathcal{D}_{\text{te}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N'}$, where $y_i \in \{0, 1\}$ indicates whether a test sample is normal ($y_i = 0$) or anomalous ($y_i = 1$). The labels are used only for evaluation and are not available during training.

Our goal is to learn from $\mathcal{D}_{\text{tr}}$ an anomaly detector $\phi : \mathbb{R}^M \to \mathbb{R}$, which models normality across multiple representation granularities and assigns each *unseen* test sample $\mathbf{x} \in \mathcal{D}_{\text{te}}$ an anomaly score $\phi(\mathbf{x})$, where higher values indicate stronger evidence of abnormality.

3.2 Causal Modeling and Anomaly Definitions

Following recent causal tabular frameworks (Nguyen et al., 2026), we assume that normal tabular observations are generated from a Structural Causal Model (SCM) defined over observed variables $\mathbf{X} = \{X_1, \ldots, X_M\}$, exogenous noise terms $\mathbf{E} = \{E_1, \ldots, E_M\}$, and deterministic structural

mechanisms $\mathcal{F} = \{f_1, \ldots, f_M\}$. Each observed variable is generated via $X_i = f_i(\mathrm{PA}_i, E_i)$, where $\mathrm{PA}_i \subseteq \mathbf{X} \setminus \{X_i\}$ denotes the direct causal parents of X_i in an underlying *directed acyclic graph* (DAG) $\mathcal{G}$, and $E_i \sim \mathcal{N}(0, 1)$ represents its exogenous noise.

Definition 1. *Causal Anomaly*: An observation $\mathbf{x} = [x_1, ..., x_M]$ is *causally anomalous* when one or more variables are inconsistent with their normal conditional mechanisms given their causal parents. Formally, under the normal SCM, each variable follows a conditional distribution $p_i(X_i \mid \mathrm{PA}_i)$. A structural anomaly occurs when the observed value x_i has an unusually low likelihood under this mechanism, even if its marginal value remains within the normal range.

For example, given the causal relationship $Age \rightarrow Salary \leftarrow Edu$, a sample is causally anomalous if its observed $Salary$ is highly improbable given its Age and Edu, regardless of whether the $Salary$ value is common in the dataset.

Prior work such as CausalAno (Nguyen et al., 2026) approaches this problem by coupling a causal generator with a single-latent discriminator, scoring anomalies via the Mahalanobis distance in a fixed latent space. However, this method has two key limitations. First, when a structural violation is localized to a small causal subgraph, its signal may be diluted by normal variables projected into the same unified latent vector. Second, Mahalanobis scoring summarizes normality using a single covariance-based quadratic geometry, which can be inadequate when latent representations are multi-modal or non-Gaussian.

To overcome these challenges, **NestedAno** constructs a hierarchy of nested Matryoshka representations $\mathcal{M} = \{m_1, \ldots, m_K\}$ (Kusupati et al., 2022), where $m_1 < \cdots < m_K$ denote progressive latent dimensions. By evaluating local deviations across all scales, NestedAno constructs an **anomaly trajectory** to detect localized structural violations without parametric assumptions.

3.3 Causally Guided Matryoshka Representation Learning

Matryoshka representation: Rather than mapping an input $\mathbf{x}$ to a single latent representation, the discriminator in NestedAno projects $\mathbf{x}$ into a nested sequence of multi-scale embeddings. First, an encoder maps the input to the maximum-dimensional representation $\mathbf{z} \in \mathbb{R}^{m_K}$. For each scale $m_k \in \mathcal{M}$, the corresponding sub-representation is obtained via prefix truncation: $\mathbf{z}^{(m_k)} = \mathbf{z}_{1:m_k} \in \mathbb{R}^{m_k}$. Each successive scale m_k retains the exact features encoded in the preceding bottleneck m_{k-1} while adding $m_k - m_{k-1}$ dimensions of representational capacity. This hierarchy provides multi-resolution views of each sample: localized causal violations that are obscured at full capacity remain prominent within lower-dimensional bottlenecks, while complex multi-variable interactions are resolved at larger capacities.

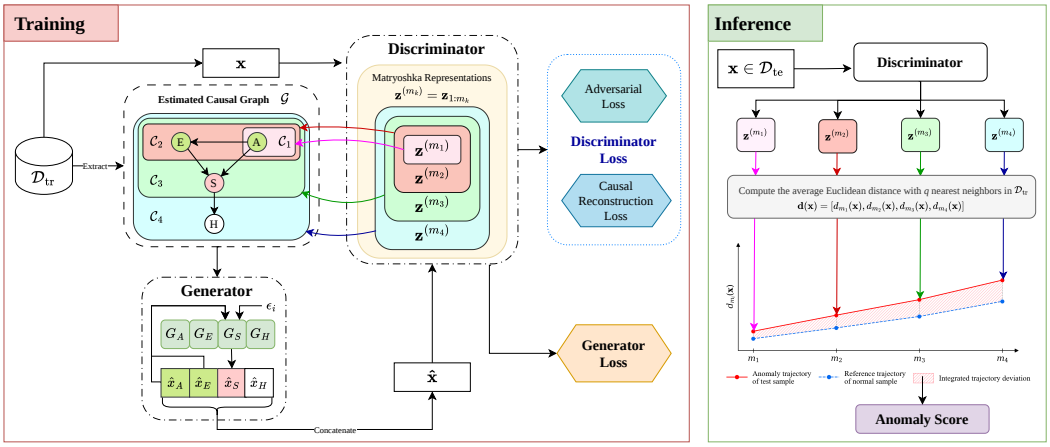


Figure 1: **NestedAno Framework**. During training, the estimated DAG $\mathcal{G}$ guides the sub-generators $\{G_i\}_{i=1}^{M}$ and establishes cumulative variable subsets $\mathcal{C}_k$. The discriminator maps samples into nested prefixes $\mathbf{z}^{(m_k)}$, which are optimized via the multi-scale adversarial and structure-aligned causal reconstruction losses. During inference, q-NN distances form a multi-scale anomaly trajectory, and its integrated deviation from the normal reference trajectory yields the final anomaly score.

Structure-guided progressive supervision: Unconstrained Matryoshka embedding learning does not determine which variables each representation should focus on. Therefore, the nested embeddings may miss necessary dependencies among variables to identify causal anomalies. To align a specific latent dimension with a semantic causal subgraph, we apply the Peter-Clark (PC) algorithm (Spirtes et al., 2000) to $\mathcal{D}_{\text{tr}}$, obtaining an estimated DAG $\mathcal{G}$. Let $\pi = (\pi_1, \ldots, \pi_M)$ denote a topological ordering of variables in $\mathcal{G}$ such that parent nodes precede their descendants. We partition the ordered variables into K cumulative subsets $\mathcal{C}_1 \subset \cdots \subset \mathcal{C}_K = \mathbf{X}$, where each subset is defined by:

$$\mathcal{C}_k = \left\{ X_{\pi_i} \mid 1 \leq i \leq \left\lfloor \frac{k \cdot M}{K} \right\rfloor \right\} \tag{1}$$

Under this formulation, $\mathcal{C}_1$ contains root and early upstream variables, while subsequent subsets progressively incorporate intermediate and downstream descendants. As illustrated in Figure 1, $\mathcal{C}_1$ contains "Age", $\mathcal{C}_2$ contains "Age and Edu", and $\mathcal{C}_4$ contains all variables "Age, Edu, Salary, House".

Once these subsets are defined, we guide each nested representation toward its assigned variables. We use auxiliary reconstruction to encourage each nested representation to retain structural information about its assigned causal subset. To anchor the k-th Matryoshka representation $\mathbf{z}^{(m_k)}$ to its assigned causal subset $\mathcal{C}_k$, we introduce a scale-specific predictor $R_{m_k} : \mathbb{R}^{m_k} \to \mathbb{R}^{|\mathcal{C}_k|}$. This is achieved by minimizing the causal reconstruction error:

$$\mathcal{L}_{\text{CR}}^{(m_k)} = \frac{1}{|\mathcal{C}_k|} \left\| R_{m_k}\left(\mathbf{z}^{(m_k)}\right) - \mathbf{x}_{\mathcal{C}_k} \right\|_2^2, \tag{2}$$

where $\mathbf{x}_{\mathcal{C}_k}$ denotes the ground-truth values of the variables in $\mathcal{C}_k$ corresponding to the input sample $\mathbf{x}$. Because the subsets $\mathcal{C}_k$ are cumulative, minimizing $\mathcal{L}_{\text{CR}}^{(m_k)}$ encourages smaller bottlenecks encode low-depth causal mechanisms, while larger bottlenecks preserve early structural representations and dedicate residual capacity to downstream variables.

3.4 Training with Wasserstein GAN

We adopt the Wasserstein GAN with Gradient Penalty (WGAN-GP) (Arjovsky et al., 2017; Gulrajani et al., 2017), optimizing a structured generator G against a multi-scale discriminator D.

Discriminator optimization: The discriminator evaluates samples across all Matryoshka scales. It consists of an encoder $E : \mathbb{R}^M \to \mathbb{R}^{m_K}$ that maps an input sample $\mathbf{x}$ to a maximum-capacity latent representation, and a shared linear critic head $D_{\text{shared}} : \mathbb{R}^{m_K} \to \mathbb{R}$. To process truncated embeddings of varying dimensions through this head, each sub-representation $\mathbf{z}^{(m_k)}$ is zero-padded to match the maximum capacity m_K. We define the discriminator response for a sample $\mathbf{x}$ at scale m_k as:

$$D^{(m_k)}(\mathbf{x}) = D_{\text{shared}}\left([E(\mathbf{x})_{1:m_k}, \mathbf{0}_{m_K - m_k}]\right) \tag{3}$$

Using this multi-scale formulation, the adversarial objective at each scale m_k is:

$$\mathcal{L}_{\text{adv}}^{(m_k)} = \mathbb{E}_{\hat{\mathbf{x}} \sim p_G}\left[D^{(m_k)}\left(\hat{\mathbf{x}}\right)\right] - \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}}\left[D^{(m_k)}\left(\mathbf{x}\right)\right], \tag{4}$$

where $p_{\mathcal{D}}$ is the distribution of normal training samples and p_G is the generator distribution.

The total discriminator loss aggregates both multi-scale adversarial and causal reconstruction losses, and encourages the 1-Lipschitz constraint:

$$\mathcal{L}_D = \frac{1}{K} \sum_{k=1}^{K} \left(\mathcal{L}_{\text{adv}}^{(m_k)} + \lambda_{\text{CR}} \cdot \mathcal{L}_{\text{CR}}^{(m_k)} \right) + \lambda_{\text{GP}} \cdot \mathbb{E}_{\tilde{\mathbf{x}}}\left[\left(\left\| \nabla_{\tilde{\mathbf{x}}} \bar{D}(\tilde{\mathbf{x}}) \right\|_2 - 1 \right)^2 \right], \tag{5}$$

where λ_{CR} balances the causal reconstruction constraint, $\lambda_{\text{GP}} = 10$ maintains the Lipschitz continuity, and $\bar{D}(\tilde{\mathbf{x}}) = \frac{1}{K} \sum_{k=1}^{K} D^{(m_k)}\left(\tilde{\mathbf{x}}\right)$ represents the ensemble critic output on linear interpolations $\tilde{\mathbf{x}} = \alpha \mathbf{x} + (1 - \alpha)\hat{\mathbf{x}}$ with $\alpha \sim \mathcal{U}(0, 1)$.

Theoretical role of the discriminator: We now analyze the role of the nested discriminator $D^{(m_k)}$ in expanding the information bottleneck within $\mathbf{z}^{(m_k)}$. Let $\mathbf{Z}^{(m_k)}$ be the random variable sampled from the encoder prefix $E(\mathbf{x})_{1:m_k}$ or $E(\hat{\mathbf{x}})_{1:m_k}$ with equal probability. Let $Y \in \{0, 1\}$ be the binary indicator where $Y = 1$ if $\mathbf{Z}^{(m_k)}$ is sampled from the real distribution $E(\mathbf{x})_{1:m_k}$, and $Y = 0$ if sampled from the generated distribution $E(\hat{\mathbf{x}})_{1:m_k}$. We establish the following theorem (proof is provided in *Appendix* H):

Theorem 2. *Given the generator G, we have the following:*

(i) $I(\mathbf{Z}^{(m_k)}; Y) = JSD\left(p(\mathbf{z}^{(m_k)} \mid Y=1) \parallel p(\mathbf{z}^{(m_k)} \mid Y=0)\right)$

(ii) If feature space support $\mathbf{z}^{(m_k)} \subseteq \mathbb{R}^{m_k}$ *is bounded with diameter* $R = \sup_{\mathbf{z},\mathbf{z}'} \|\mathbf{z} - \mathbf{z}'\|_2$*, then:*

$$I(\mathbf{Z}^{(m_k)}; Y) \geq \frac{1}{2R^2}\left(\mathbb{E}_{\mathbf{x}\sim p_{\mathcal{D}}}\left[D^{(m_k)}(\mathbf{x})\right] - \mathbb{E}_{\hat{\mathbf{x}}\sim p_G}\left[D^{(m_k)}(\hat{\mathbf{x}})\right]\right)^2$$

According to this inequality, training the discriminator $D^{(m_k)}$ to maximize the adversarial objective forces the encoder E to produce nested latent representations $\mathbf{z}^{(m_k)}$ that maximize the mutual information $I(\mathbf{Z}^{(m_k)}; Y)$. This maximizes the encoder's capability to distinguish real observations from structurally inconsistent synthetic examples at each causal subset $\mathcal{C}_k$. Conversely, the generator G is updated to minimize $I(\mathbf{Z}^{(m_k)}; Y)$ and, by extension, the Jensen-Shannon Divergence. This minimax dynamic forces the generated representations to match the true data distribution at each bottleneck scale. Consequently, the prefix representations $\mathbf{z}^{(m_k)}$ are effective to capture the localized structural mechanisms assigned to $\mathcal{C}_k$, enhancing their sensitivity to causal anomalies during inference.

Generator optimization: Following CausalAno, the generator G is decomposed into M structural sub-generators $\{G_i\}_{i=1}^M$, where each G_i parameterizes the conditional mechanism of variable X_i:

$$\hat{X}_i = G_i\left(\widehat{\mathrm{PA}}_i, \epsilon_i\right), \quad \epsilon_i \sim \mathcal{N}(0,1) \tag{6}$$

Variables are generated sequentially according to the topological ordering π. Each variable $\hat{X}_i$ receives the synthesized values of its causal parents $\widehat{\mathrm{PA}}_i$ and independent standard normal noise ϵ_i, yielding the assembled sample $\hat{\mathbf{x}} = [\hat{X}_1, ..., \hat{X}_M]$.

The generator is optimized to increase the critic response of generated samples across the entire zero-padded Matryoshka spectrum simultaneously:

$$\mathcal{L}_G = -\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\hat{\mathbf{x}}\sim p_G}\left[D^{(m_k)}\left(\hat{\mathbf{x}}\right)\right] \tag{7}$$

At the ideal optimum, this objective encourages the generated distribution to match the normal data distribution simultaneously across all representation scales. However, as established theoretically in CausalAno (Nguyen et al., 2026), the finite capacity of the structural generators $\{G_i\}_{i=1}^M$ prevents them from exactly reproducing the true structural mechanisms $\mathcal{F}$. Consequently, the generated samples may contain structural discrepancies relative to the normal SCM. These discrepancies provide challenging synthetic examples that encourage the discriminator to learn distinctions between normal observations and structurally inconsistent samples across multiple representation scales.

3.5 Trajectory-based Anomaly Detection

Unlike CausalAno, which scores anomalies in a single high-dimensional latent space under a Gaussian assumption, NestedAno measures local deviation from normal samples at each Matryoshka embedding scale. This avoids imposing a parametric distribution on the latent space while capturing anomaly evidence across multiple representation levels.

At test time, for an *unseen* sample $\mathbf{x} \in \mathcal{D}_{\text{te}}$, we compute its latent representation $\mathbf{z}_{\mathbf{x}}$ and extract each prefix $\mathbf{z}_{\mathbf{x}}^{(m_k)}$. At scale m_k, its local deviation from normality is measured by the average Euclidean distance to its q nearest normal training neighbors:

$$d_{m_k}(\mathbf{x}) = \frac{1}{q}\sum_{\mathbf{z}_i^{(m_k)} \in \mathcal{N}_q^{(m_k)}(\mathbf{x})} \left\| \mathbf{z}_{\mathbf{x}}^{(m_k)} - \mathbf{z}_i^{(m_k)} \right\|_2, \tag{8}$$

where $\mathcal{N}_q^{(m_k)}(\mathbf{x})$ denotes the q nearest neighbors of $\mathbf{z}_{\mathbf{x}}^{(m_k)}$ within the training set embeddings $\{\mathbf{z}_i^{(m_k)}\}_{i=1}^N$. Evaluating this distance across all K scales yields the *anomaly trajectory* $\mathbf{d}(\mathbf{x}) = [d_{m_1}(\mathbf{x}), \ldots, d_{m_K}(\mathbf{x})]$. Applying this procedure over all normal training samples produces a *reference trajectory* $\mathbf{d}^{\text{tr}} = [d_{m_1}^{\text{tr}}, ..., d_{m_K}^{\text{tr}}]$, where $d_{m_k}^{\text{tr}} = \frac{1}{N}\sum_{i=1}^N d_{m_k}(\mathbf{x}_i)$. The reference trajectory describes the typical multi-scale behavior of normal data.

Trajectory Integration: The final *anomaly score* $\phi(\mathbf{x})$ quantifies the cumulative deviation of the test sample trajectory above the normal reference trajectory across the discrete scale spectrum $\mathcal{M}$. Formally, we formulate this using the trapezoidal quadrature:

$$\phi(\mathbf{x}) = \sum_{k=1}^{K-1} \frac{m_{k+1} - m_k}{2} \cdot \left[\Delta_k(\mathbf{x}) + \Delta_{k+1}(\mathbf{x})\right], \tag{9}$$

where $\Delta_k(\mathbf{x}) = d_{m_k}(\mathbf{x}) - d^{\mathrm{tr}}_{m_k}$ is the deviation at scale m_k.

This quadrature formulation estimates the total area between the test and reference curves. Anomalies associated with early variables can produce deviations from the initial bottlenecks, while anomalies associated with later variables may exhibit stronger deviations as progressively downstream variables are incorporated. In both cases, $\phi(\mathbf{x})$ integrates sustained divergence across relevant scales, reflecting persistent deviation from normality rather than an isolated deviation at a single representation level. Samples with large scores $\phi(\mathbf{x})$ are identified as anomalies.

4 Experiments

4.1 Experimental Setup

Datasets: We use the same 28 tabular anomaly detection datasets from CausalAno (Nguyen et al., 2026) plus two very large datasets *Fraud* and *Vifd* from AnoLLM (Tsai et al., 2025). In total, we evaluate 30 datasets, including 12 mixed-type datasets with both categorical and continuous features and 18 continuous-only datasets. These datasets cover diverse application domains, including healthcare, finance, cybersecurity, and social sciences. Statistics of 12 mixed-type and 18 continuous datasets are provided in *Appendix A*.

Evaluation metrics: Following established practice in tabular anomaly detection (Ye et al., 2025; Tsai et al., 2025; Nguyen et al., 2026), we randomly assign 50% of the normal samples to training. The test set contains the remaining 50% normal samples together with all anomalies. We use AUC-ROC as the primary metric because it measures the quality of anomaly ranking independently of a fixed decision threshold. F1-score and method rankings are additionally provided in *Appendix* C.

Baseline methods: We compare NestedAno with 17 SOTA detection methods, spanning classical, representation learning, self-supervised, generative, and LLM-based approaches. The baselines are *IForest* (Liu et al., 2008), *KNN* (Ramaswamy et al., 2000), *PCA* (Shyu et al., 2003), *ECOD* (Li et al., 2022), *DeepSVDD* (Ruff et al., 2018), *REPEN* (Pang et al., 2018), *RCA* (Liu et al., 2021), *SLAD* (Xu et al., 2023), *GOAD* (Bergman and Hoshen, 2020), *NeuTraL* (Qiu et al., 2021), *ICL* (Shenkar and Wolf, 2022), *DTE* (Livernoche et al., 2024), *AnoGAN* (Schlegl et al., 2019), *DRL* (Ye et al., 2025), *AnoLLM* (Tsai et al., 2025), *LLM-DAS* (Ye et al., 2026), and *CausalAno* (Nguyen et al., 2026). For reproducibility, we use publicly available implementations of the baseline methods whenever available and evaluate all methods using the same data splits and evaluation protocol.

Implementation details: We retain the structured generator architecture and the common training configuration of CausalAno (Nguyen et al., 2026). Each sub-generators G_i is a 4-layer MLP with type-specific activation function: $\tanh$ for continuous variables and Gumbel-Softmax for categorical variables. We run the PC algorithm with $\alpha = 0.05$ to extract causal graphs, and train the model with 300 epochs, a batch size of 64, and a noise dimension of 16.

To maintain the *unsupervised* setting and avoid dataset-specific hyper-parameter tuning, we use a **fixed configuration** across all datasets. Specifically, we set the Matryoshka embedding scales $\mathcal{M} = \{128, 256, 512, 1024\}$ and the causal reconstruction coefficient $\lambda_{\mathrm{CR}} = 0.1$. We provide sensitivity analyses of $\mathcal{M}$ and λ_{CR} in the ablation studies. To mitigate randomness, each experiment is repeated using three independent runs, and the average performance is reported. Additional implementation details are provided in *Appendix* B.

4.2 Results and Analysis

Results on mixed-type datasets: Table 1 shows that NestedAno achieves the highest overall AUC-ROC (74.49%) on 12 mixed-type datasets, compared with 71.89% achieved by the strongest baseline

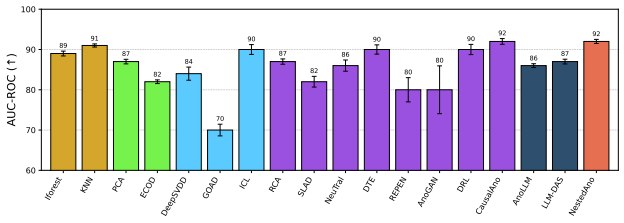
CausalAno. The improvement is ~3%, confirmed statistically significant by a Wilcoxon signed-rank test ($p = 0.03 < 0.05$). NestedAno also achieves a higher AUC-ROC than CausalAno on 10 of the 12 datasets. The improvement is consistent with the use of nested representations, which allow anomaly evidence to be evaluated and aggregated across multiple latent scales rather than relying on a single fixed representation. In addition, the trajectory-based scoring mechanism avoids the covariance-based latent geometry used by CausalAno and does not require a single parametric model of the normal latent distribution.

Table 1: AUC-ROC scores (%) for NestedAno and baselines on 12 mixed-type datasets. **Bold** and underline indicate the best and second-best methods. We report standard deviation in *Appendix* C.

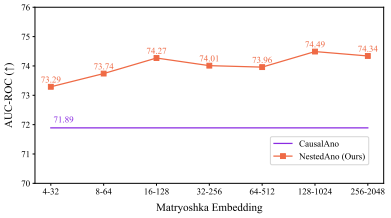
	ACD	Bank	CMC	DAMRE	Fraud	Lymp	NHIS	OS	Seismic	SMD	SPD	Vifd	AVG
Iforest	50.30	51.18	50.61	69.01	48.00	70.50	65.73	77.16	70.58	81.27	48.37	50.67	61.12
PCA	50.59	51.52	51.01	88.57	45.38	85.60	64.98	84.75	70.41	93.40	47.86	49.31	65.28
KNN	49.43	51.24	57.79	92.50	50.05	87.05	67.48	93.30	**74.45**	92.37	49.19	52.27	68.09
ECOD	50.86	51.34	53.94	84.42	46.31	85.45	65.37	76.86	69.80	91.14	49.15	50.28	64.58
DeepSVDD	47.74	49.96	57.15	91.62	50.15	86.85	66.26	79.57	69.39	68.64	48.48	49.93	63.81
RCA	50.08	51.38	50.29	91.46	50.14	92.10	66.13	86.50	73.23	92.60	49.40	52.94	67.19
SLAD	50.10	53.15	49.24	80.39	52.42	93.97	82.52	86.77	71.01	47.32	44.95	56.45	64.02
GOAD	49.69	51.39	48.33	87.26	50.01	87.64	71.51	91.17	71.93	15.45	49.10	49.91	60.28
Neutral	49.25	51.13	54.57	82.71	50.37	77.00	67.33	79.20	67.40	50.75	**53.61**	49.97	61.11
ICL	53.60	53.20	55.75	93.34	48.11	91.55	87.85	**97.07**	71.72	86.70	49.89	56.12	70.41
DTE	51.39	59.04	54.62	91.19	50.00	88.18	50.63	94.52	71.14	82.44	48.60	55.29	66.42
REPEN	50.37	51.27	59.71	78.40	50.22	85.52	66.99	93.17	72.95	78.77	47.88	50.96	65.52
AnoGAN	50.51	53.72	44.58	64.58	45.51	99.45	47.89	73.70	65.31	74.15	51.55	52.60	60.30
AnoLLM	51.00	64.83	53.52	**93.87**	**60.44**	**99.53**	53.48	92.39	73.90	91.79	50.50	55.96	70.10
DRL	49.00	57.61	57.50	89.74	49.25	94.29	89.62	93.39	68.88	82.57	49.19	53.73	69.56
LLM-DAS	50.15	61.72	50.83	90.59	45.38	99.14	80.43	85.97	70.06	92.68	49.99	53.81	69.23
CausalAno	51.64	**67.66**	58.27	92.70	46.09	97.85	80.75	95.02	72.74	**93.73**	51.99	54.22	71.89
NestedAno	**54.03**	66.74	**59.87**	93.00	56.83	99.06	**91.19**	96.94	73.24	92.29	53.13	**57.60**	**74.49**

Results on continuous datasets: Figure 2a reports the AUC-ROC across the 18 continuous datasets. NestedAno achieves the highest average AUC-ROC (92%), performing comparably to CausalAno and exceeding other strong baselines (including KNN, ICL, DTE, and DRL) by 1-2%. These results indicate that the benefit of nested representation learning is not limited to mixed-type data.

In summary, NestedAno performs competitively on continuous datasets while providing clear gains on mixed-type datasets.



(a) Average AUC-ROC scores over 18 continuous datasets for all methods: yellow (proximity-based), green (distribution-based), blue (boundary-based), purple (reconstruction- and generation-based), navy (LLM-based), orange (ours).



(b) Average AUC-ROC scores across the 12 mixed-type datasets for different Matryoshka embedding ranges.

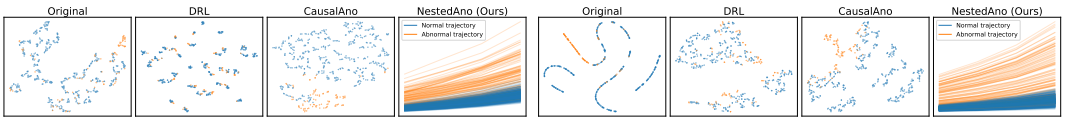
Figure 2: Performance on continuous datasets (a) and Matryoshka embedding analysis (b).

4.3 Ablation Study

Effectiveness of Matryoshka embeddings: Figure 2b evaluates different Matryoshka embedding ranges on 12 mixed-type datasets. Each configuration contains four progressively doubling embedding sizes, e.g., the range 128–1024 corresponds to $\{128, 256, 512, 1024\}$. All evaluated configurations outperform CausalAno, indicating that multi-scale representations provide a benefit over a single fixed latent space. Performance generally improves as the embedding range increases, with 128–1024 achieving the highest average AUC-ROC (74.49%), compared with 71.89% for CausalAno.

Visualization: Following CausalAno, we investigate how different methods separate normal and anomalous samples on two datasets OS and DAMRE. We use t-SNE projections (Van der Maaten and Hinton, 2008) to visualize the latent representations learned by DRL and CausalAno, and line plots to visualize anomaly trajectories learned by NestedAno. Visualizations for additional datasets are provided in *Appendix* G.

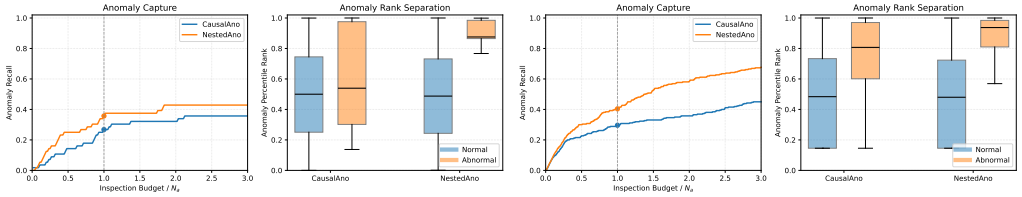
As shown in Figure 3, normal and anomalous samples exhibit substantial overlap in the original observed space, making these datasets challenging for anomaly detection. CausalAno produces more distinct latent-space separation, while DRL exhibits substantial overlap. NestedAno clearly separates normal and anomalous samples from the trajectory perspective, with anomalous samples generally exhibiting higher trajectories than normal samples. This clear separation allows NestedAno to effectively detect anomalies with the use of integrated trajectory deviation for anomaly ranking.



(a) OS (b) DAMRE

Figure 3: Visualization of latent representations learned by DRL and CausalAno, and anomaly trajectories learned by NestedAno on two datasets OS and DAMRE.

Analysis of cases where CausalAno underperforms: We further examine two datasets NHIS and Mammography, where NestedAno substantially improves over CausalAno. Rather than focusing on latent-space separation, we examine the quality of the final anomaly ranking. Figure 4 reports cumulative *anomaly recall* as a function of inspection budget (Andersson et al., 2025) and percentile-rank boxplots (Ramos et al., 2025). The cumulative recall curve measures the fraction of true anomalies recovered as the inspection budget increases, while the percentile-rank boxplots characterize the ranking positions of normal and anomalous samples. On both datasets, NestedAno captures more anomalies under the same budget and places anomalous samples closer to the top of the ranking.



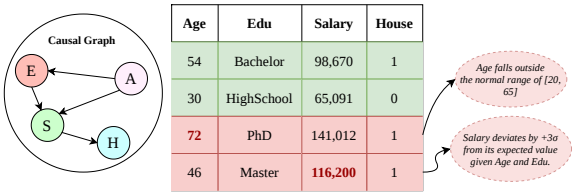
(a) NHIS (b) Mammography

Figure 4: Anomaly-retrieval performance of CausalAno and NestedAno on two datasets NHIS and Mammography. The left panel shows cumulative anomaly recall as the inspection budget increases, normalized by the number of true anomalies N_a in the test set. The right panel shows percentile-rank distributions, with ranks scaled to $[0, 1]$ and larger values indicating positions closer to the top of the anomaly ranking.

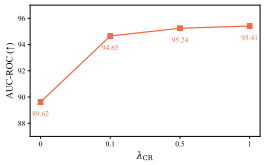
To investigate why CausalAno underperforms on these datasets, we examine whether its covariance-based latent geometry is appropriate for the learned normal representations. NHIS and Mammography exhibit substantial deviations from multivariate normality according to Mardia's multivariate skewness and kurtosis. Specifically, NHIS has a skewness of 41.2 and kurtosis of 5.5, while Mammography has values of 22.9 and 2.8, respectively. For comparison, the corresponding values for CausalAno on OS in Figure 3 are 1.1 and 1.0. In addition, the Henze–Zirkler test rejects multivariate normality on both datasets ($p < 0.01$). These results suggest that the single covariance-based Mahalanobis geometry used by CausalAno may provide a poor approximation to the local structure of the normal latent distribution on these datasets.

Synthetic topological-progression experiment: To examine whether NestedAno captures anomaly evidence differently as progressively later variables are incorporated into the representation, we con-

struct a *Synthetic* dataset using the true DAG shown in Figure 5a. Thus, the topological progression follows $Age \rightarrow Edu \rightarrow Salary \rightarrow House$. We generate $\mathcal{D}_{\text{tr}}$ with 1,000 normal samples and $\mathcal{D}_{\text{te}}$ containing 200 normal samples, 100 Age anomalies, and 100 Salary anomalies. Age anomalies place Age outside its normal range $[20, 65]$, while downstream variables are generated according to their normal mechanisms. Salary anomalies retain normal Age and Edu but perturb $Salary$ by ± 3 standard deviations from its expected value before generating $House$.



(a) True DAG and examples of generated synthetic samples.



(b) AUC-ROC vs. λ_{CR}.

Figure 5: Synthetic topological-progression experiment and sensitivity to λ_{CR}.

In Figure 6, we plot the mean anomaly trajectory for each anomaly type and for normal test samples. The x-axis denotes Matryoshka embedding scales, and the y-axis denotes the anomaly signal. The results show that *causal reconstruction* (CR) supervision influences how anomaly evidence changes across the nested representations. When $\lambda_{\text{CR}} = 0$, the anomaly trajectories do not exhibit a clear distinction between the earliest and later representation scales. As λ_{CR} increases, the trajectories become progressively more consistent with the topological ordering: upstream anomalies are visible at smaller representation scales, whereas anomalies associated with later variables become more pronounced at larger scales. At $\lambda_{\text{CR}} = 1.0$, the Age anomaly is most pronounced at the earliest embedding scale, while the Salary anomaly becomes clearer around the third embedding scale. This behavior is consistent with the progressive incorporation of variables along the topological ordering.

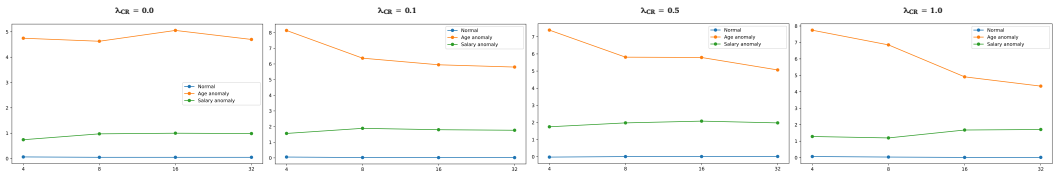


Figure 6: Mean anomaly trajectories under different λ_{CR}. The x-axis denotes the embedding scales. For the 4-dim Synthetic dataset, we use a small range 4–32. The y-axis denotes the anomaly signal.

Effect of λ_{CR} on detection performance: Figure 5b evaluates $\lambda_{\text{CR}} \in \{0, 0.1, 0.5, 1.0\}$. With $\lambda_{\text{CR}} = 0$, the Matryoshka representations are not explicitly constrained by the causal topological organization, yielding the lowest AUC-ROC (89.62%). Increasing λ_{CR} consistently improves performance, reaching 95.41% at $\lambda_{\text{CR}} = 1.0$. This result indicates that causal reconstruction supervision contributes to organizing the nested representations in a way that improves anomaly detection.

Computational efficiency: *Appendix* D reports the runtime comparison. **Robustness to DAG:** *Appendix* E examines the sensitivity to different levels of DAG quality. **Effect of topological ordering:** *Appendix* F investigates the sensitivity to variable-ordering.

5 Conclusion

We address *unsupervised* tabular AD in which structural violations may be obscured within a single latent representation. NestedAno organizes representations into a Matryoshka hierarchy and uses estimated causal structure to progressively guide each scale toward cumulative subsets of variables. Adversarial learning is performed across the nested representations, while inference captures multi-scale anomaly evidence through an anomaly trajectory. Experiments on 30 continuous and mixed-type datasets show that NestedAno achieves the best overall performance, with clear gains on mixed-type data while remaining competitive on continuous datasets. Additional analyses demonstrate the benefits of nested representations, topological ordering, and robustness to imperfect causal graphs.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and manuscript polishing, including improving grammatical correctness, clarity, and readability. They were not used to formulate research hypotheses, design the methodology or experiments, generate or interpret experimental results, or draw scientific conclusions. All AI-assisted edits were reviewed and verified by the authors to ensure that they did not alter the intended technical meaning or introduce unsupported claims. The authors take full responsibility for the accuracy, originality, and final content of this work.

REPRODUCIBILITY STATEMENT

Detailed descriptions of the proposed method, including the model architecture in *Appendix* B, as well as the training setup, data preprocessing, and evaluation protocol, are provided throughout the main text. The full source code will be made publicly available upon acceptance.

REFERENCES

Alex Andersson, Chris Lintott, Rob Fender, Michelle Lochner, Patrick Woudt, Jakob van den Eijnden, Alexander van der Horst, Assaf Horesh, Payaswini Saikia, Gregory Sivakoff, et al. Finding radio transients with anomaly detection and active learning based on volunteer classifications. *Monthly Notices of the Royal Astronomical Society*, 2025.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

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

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. *NeurIPS*, 2017.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *NeurIPS*, 2022.

Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George Chen. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Boyang Liu, Ding Wang, Kaixiang Lin, Pang-Ning Tan, and Jiayu Zhou. RCA: A deep collaborative autoencoder approach for anomaly detection. In *IJCAI*, 2021.

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

Victor Livernoche, Vineet Jain, Yashar Hezaveh, and Siamak Ravanbakhsh. On diffusion modeling for anomaly detection. In *ICLR*, 2024.

Dang Nguyen, Tu Anh Hoang Nguyen, Thuc Duy Le, Svetha Venkatesh, Trung Le, and Sunil Gupta. Causal-aware anomaly detection for tabular data. In *ICML*, 2026.

Guansong Pang, Longbing Cao, Ling Chen, and Huan Liu. Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *KDD*, 2018.

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

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *International Conference on Management of Data*, 2000.

Erikson Carlos Ramos, Leandro Carlos de Souza, and Gustavo Henrique Matos Bezerra Motta. Anomaly Detection Based on Markovian Geometric Diffusion. *Machine Learning and Knowledge Extraction*, 2025.

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

Thomas Schlegl, Philipp Seeböck, Sebastian Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 2019.

Tom Shenkar and Lior Wolf. Anomaly detection for tabular data with internal contrastive learning. In *ICLR*, 2022.

Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. *IEEE Foundations and New Directions of Data Mining*, 2003.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT Press, 2000.

Che-Ping Tsai, Ganyu Teng, Phillip Wallis, and Wei Ding. AnoLLM: Large language models for tabular anomaly detection. In *ICLR*, 2025.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008.

Hongzuo Xu, Yijie Wang, Juhui Wei, Songlei Jian, Yizhou Li, and Ning Liu. Fascinating supervisory signals and where to find them: Deep anomaly detection with scale learning. In *ICML*, 2023.

Hangting Ye, He Zhao, Wei Fan, Mingyuan Zhou, Dan dan Guo, and Yi Chang. DRL: Decomposed Representation Learning for Tabular Anomaly Detection. In *ICLR*, 2025.

Hangting Ye, Jinmeng Li, He Zhao, Mingchen Zhuge, Dandan Guo, Yi Chang, and Hongyuan Zha. LLM as an Algorithmist: Enhancing Anomaly Detectors via Programmatic Synthesis. In *ICLR*, 2026.

A STATISTICS TABULAR DATASETS

Table 2 summarizes the characteristics of the 30 mixed-type and continuous tabular datasets.

Table 2: Characteristics of 30 benchmark datasets. M_{cat} and M_{con} are the numbers of categorical and continuous features, while N_a is the number of anomalies. *Fraud* and *Vifd* are from AnoLLM (Tsai et al., 2025), while the remaining datasets are from CausalAno (Nguyen et al., 2026).

12 Mixed-type Datasets	N	M	M_{cat}	M_{con}	N_a
ACD	10,000	4	3	1	490 (5%)
Bank	41,188	10	10	0	4,640 (11%)
CMC	1,473	8	8	0	29 (2%)
DAMRE	1,000	5	2	3	100 (10%)
Fraud	151,112	8	6	2	14,159 (9%)
Lymp	148	18	15	3	6 (4%)
NHIS	4,388	4	2	2	56 (1%)
OS	1,000	5	2	3	91 (9%)
Seismic	2,584	18	4	14	170 (7%)
SMD	5,000	7	2	5	250 (5%)
SPD	1,000	12	9	3	764 (76%)
Vifd	15,420	32	24	8	923 (6%)

18 Continuous Datasets	N	M	M_{cat}	M_{con}	N_a
Annthyroid	7,200	6	0	6	534 (7%)
BreastW	699	9	0	9	241 (34%)
Ecoli	336	7	0	7	9 (3%)
Mammography	11,183	6	0	6	260 (2%)
Shuttle	49,097	9	0	9	3,511 (7%)
Thyroid	3,772	6	0	6	93 (2%)
Wine	178	13	0	13	71 (40%)
Waveform	3,443	21	0	21	100 (3%)
WBC	223	9	0	9	10 (4%)
Pendigits	6,870	16	0	16	156 (2%)
Cardio	1,831	21	0	21	176 (10%)
PageBlocks	5,393	10	0	10	510 (9%)
Stamps	340	9	0	9	31 (9%)
Glass	214	7	0	7	9 (4%)
Vowels	1,456	12	0	12	50 (3%)
MagicGamma	19,020	10	0	10	6,688 (35%)
Yeast	1,484	8	0	8	95 (6%)
ImgSeg	2,310	18	0	18	990 (43%)

B NETWORK ARCHITECTURE

B.1 SUB-GENERATOR G_i

Each variable X_i is modeled by a dedicated sub-generator G_i, implemented as a four-layer multilayer perceptron (MLP). Its input consists of the generated parent values $\hat{\mathbf{x}}_{\mathrm{PA}_i}$ and an independent noise variable ϵ_i. The hidden layers expand from 64 to 128 units and use LeakyReLU activations with a negative slope of 0.2. Batch Normalization with a momentum of 0.8 is applied after the second and third layers to improve training stability. The output activation depends on the variable type: continuous variables use a tanh activation, whereas categorical variables use a Gumbel-Softmax layer to produce a differentiable probability vector over the corresponding categories.

B.2 DISCRIMINATOR D

The discriminator uses a three-layer MLP to map each input $\mathbf{x}$ to the maximum-dimensional representation $\mathbf{z} \in \mathbb{R}^{m_K}$. The first two layers contain 256 units with LeakyReLU activations of negative slope 0.2, followed by a linear layer producing $\mathbf{z}$. Each Matryoshka prefix $\mathbf{z}^{(m_k)}$ is then connected to two heads: a shared linear scoring head that outputs a scalar critic response $D^{(m_k)}$, and a scale-specific linear predictor $\mathbb{R}^{m_K}$ that outputs the variables assigned to $\mathcal{C}_k$ for causal reconstruction supervision. Both the generator and the discriminator are optimized using the Adam optimizer with a learning rate of 2×10^{-4}, $\beta_1 = 0.5$, and $\beta_2 = 0.9$.

C FULL PERFORMANCE

C.1 AUC-ROC

AUC-ROC is used as the primary anomaly detection metric. Table 3 provides the results for each dataset.

Table 3: AUC-ROC (standard deviation) for each dataset (*higher is better*).

Dataset	Iforest	KNN	PCA	ECOD	DeepSVDD	GOAD	ICL	RCA	SLAD	Neutral	DTE	REPEN	AnoGAN	AnoLLM	DRL	LLM-DAS	CausalAno	NestedAno
Annthyroid	90.74 (0.01)	80.88 (0.00)	82.98 (0.02)	78.83 (0.00)	73.35 (0.01)	55.21 (0.00)	85.32 (0.01)	71.60 (0.00)	75.38 (0.01)	77.66 (0.02)	97.64 (0.00)	73.16 (0.02)	64.64 (0.14)	91.54 (0.00)	83.20 (0.01)	83.30 (0.02)	**93.72** (0.00)	93.33 (0.01)
BreastW	99.26 (0.00)	99.30 (0.00)	99.12 (0.00)	98.94 (0.00)	97.19 (0.00)	**99.32** (0.00)	99.22 (0.00)	98.89 (0.00)	98.96 (0.00)	98.48 (0.00)	98.86 (0.00)	95.20 (0.02)	99.06 (0.00)	99.13 (0.00)	99.06 (0.00)	98.45 (0.00)	98.88 (0.00)	98.95 (0.00)
Cardio	94.92 (0.00)	91.87 (0.00)	**96.55** (0.00)	93.23 (0.00)	84.87 (0.02)	50.61 (0.02)	86.68 (0.01)	95.17 (0.00)	85.37 (0.02)	84.24 (0.01)	93.56 (0.00)	82.85 (0.04)	93.22 (0.02)	90.98 (0.00)	91.49 (0.03)	94.84 (0.01)	93.83 (0.00)	94.03 (0.00)
Ecoli	81.26 (0.00)	87.85 (0.00)	83.58 (0.01)	75.14 (0.01)	85.25 (0.02)	81.53 (0.01)	**89.75** (0.00)	88.44 (0.03)	77.55 (0.01)	86.45 (0.01)	82.34 (0.01)	84.24 (0.04)	79.03 (0.02)	87.83 (0.01)	86.47 (0.01)	82.41 (0.01)	88.56 (0.00)	87.98 (0.00)
Glass	79.61 (0.00)	84.61 (0.02)	70.84 (0.01)	69.07 (0.02)	83.78 (0.04)	63.36 (0.03)	89.79 (0.04)	70.41 (0.02)	78.28 (0.04)	**90.58** (0.02)	78.18 (0.02)	73.93 (0.03)	51.78 (0.19)	76.16 (0.01)	84.61 (0.03)	69.18 (0.01)	86.77 (0.02)	83.03 (0.02)
ImgSeg	76.95 (0.00)	85.63 (0.01)	71.47 (0.01)	64.42 (0.00)	86.73 (0.01)	80.33 (0.01)	**90.18** (0.00)	73.60 (0.01)	83.21 (0.01)	89.14 (0.01)	80.77 (0.02)	56.17 (0.02)	68.33 (0.09)	68.17 (0.01)	86.88 (0.02)	72.01 (0.01)	88.03 (0.01)	87.53 (0.01)
MagicGamma	77.85 (0.00)	84.56 (0.00)	70.71 (0.00)	64.00 (0.00)	72.30 (0.01)	65.46 (0.01)	78.52 (0.01)	75.76 (0.00)	74.20 (0.00)	81.60 (0.00)	**86.62** (0.00)	76.92 (0.02)	69.38 (0.03)	72.61 (0.00)	79.26 (0.01)	71.27 (0.00)	84.46 (0.01)	83.50 (0.01)
Mammography	87.36 (0.00)	87.04 (0.00)	90.02 (0.00)	**90.65** (0.00)	86.73 (0.01)	75.51 (0.00)	78.94 (0.02)	87.60 (0.00)	73.65 (0.01)	70.02 (0.01)	85.61 (0.03)	85.26 (0.01)	76.19 (0.10)	86.87 (0.00)	84.78 (0.01)	90.62 (0.00)	84.35 (0.03)	87.82 (0.00)
PageBlocks	92.72 (0.00)	94.78 (0.00)	93.18 (0.00)	91.31 (0.00)	91.08 (0.00)	86.86 (0.00)	94.38 (0.00)	93.91 (0.00)	86.33 (0.02)	91.20 (0.00)	95.09 (0.01)	88.67 (0.00)	85.22 (0.01)	88.96 (0.00)	96.80 (0.00)	93.02 (0.00)	**97.41** (0.00)	96.09 (0.00)
Pendigits	96.69 (0.01)	**99.86** (0.00)	94.20 (0.00)	92.73 (0.00)	80.46 (0.06)	27.11 (0.03)	98.77 (0.01)	97.04 (0.00)	93.38 (0.01)	96.33 (0.01)	97.95 (0.00)	92.60 (0.02)	94.37 (0.03)	95.15 (0.00)	98.47 (0.01)	94.72 (0.00)	99.57 (0.00)	**99.86** (0.00)
Shuttle	99.62 (0.00)	99.89 (0.00)	99.36 (0.00)	99.30 (0.00)	99.49 (0.00)	98.83 (0.00)	99.97 (0.00)	99.54 (0.00)	99.81 (0.00)	99.77 (0.00)	99.76 (0.00)	99.38 (0.00)	99.06 (0.00)	99.84 (0.00)	99.93 (0.00)	99.52 (0.00)	**99.98** (0.00)	99.85 (0.00)
Stamps	94.01 (0.00)	92.49 (0.01)	93.22 (0.01)	88.43 (0.02)	85.40 (0.04)	54.32 (0.06)	90.43 (0.02)	92.57 (0.01)	66.94 (0.05)	83.02 (0.04)	89.91 (0.02)	85.87 (0.05)	63.66 (0.24)	93.68 (0.01)	93.42 (0.01)	91.81 (0.01)	**94.96** (0.01)	93.59 (0.02)
Thyroid	99.03 (0.00)	97.61 (0.00)	98.39 (0.00)	97.87 (0.00)	85.01 (0.01)	68.44 (0.02)	98.60 (0.00)	97.12 (0.00)	94.85 (0.01)	84.59 (0.02)	**99.09** (0.00)	92.95 (0.02)	96.06 (0.02)	98.91 (0.00)	98.15 (0.00)	97.76 (0.00)	98.49 (0.00)	98.73 (0.00)
Vowels	78.32 (0.02)	97.43 (0.00)	63.98 (0.01)	59.83 (0.00)	87.99 (0.00)	90.86 (0.02)	98.69 (0.00)	89.45 (0.00)	97.01 (0.00)	**99.21** (0.00)	97.80 (0.00)	71.96 (0.07)	52.31 (0.03)	67.63 (0.02)	98.92 (0.01)	67.32 (0.01)	98.45 (0.00)	98.41 (0.00)
Waveform	71.60 (0.01)	**76.06** (0.01)	65.46 (0.00)	60.56 (0.00)	58.62 (0.02)	48.08 (0.01)	74.44 (0.03)	65.66 (0.01)	47.91 (0.00)	73.57 (0.01)	63.49 (0.02)	75.63 (0.06)	58.53 (0.10)	65.20 (0.00)	66.08 (0.05)	69.02 (0.00)	74.85 (0.01)	71.39 (0.00)
WBC	99.18 (0.00)	99.28 (0.00)	99.35 (0.00)	99.28 (0.00)	91.59 (0.01)	96.39 (0.02)	98.32 (0.01)	98.41 (0.00)	99.22 (0.00)	87.41 (0.04)	94.67 (0.03)	98.66 (0.01)	99.44 (0.00)	**99.59** (0.00)	97.70 (0.01)	98.63 (0.00)	97.69 (0.00)	98.72 (0.00)
Wine	93.55 (0.01)	99.31 (0.00)	99.72 (0.00)	74.37 (0.01)	96.65 (0.02)	99.55 (0.00)	97.79 (0.03)	95.51 (0.01)	99.17 (0.01)	99.37 (0.00)	**99.78** (0.00)	30.94 (0.07)	98.92 (0.01)	99.74 (0.00)	99.47 (0.00)	99.25 (0.00)	99.41 (0.01)	99.58 (0.00)
Yeast	**86.34** (0.01)	80.32 (0.00)	85.82 (0.01)	79.13 (0.01)	63.35 (0.03)	17.50 (0.01)	77.65 (0.03)	81.69 (0.01)	38.33 (0.04)	53.32 (0.04)	79.35 (0.04)	76.47 (0.04)	**87.22** (0.03)	75.31 (0.01)	76.10 (0.02)	85.39 (0.01)	78.89 (0.01)	81.25 (0.00)
ACD	50.30 (0.02)	49.43 (0.00)	50.59 (0.02)	50.86 (0.02)	47.74 (0.00)	49.69 (0.02)	53.60 (0.01)	50.08 (0.01)	50.10 (0.01)	49..25 (0.02)	51.39 (0.00)	50.37 (0.01)	50.51 (0.02)	51.00 (0.01)	49.00 (0.01)	50.15 (0.01)	51.64 (0.01)	**54.03** (0.01)
Bank	51.18 (0.00)	51.24 (0.00)	51.52 (0.00)	51.34 (0.00)	49.96 (0.01)	51.39 (0.00)	53.20 (0.01)	51.38 (0.00)	53.15 (0.00)	51.13 (0.00)	59.04 (0.01)	51.27 (0.00)	53.72 (0.01)	64.83 (0.00)	57.61 (0.01)	61.72 (0.00)	**67.66** (0.01)	66.74 (0.01)
CMC	50.61 (0.01)	57.79 (0.01)	51.01 (0.00)	53.94 (0.01)	57.15 (0.02)	48.33 (0.01)	55.75 (0.01)	50.29 (0.01)	49.24 (0.00)	54.57 (0.01)	54.62 (0.04)	59.71 (0.04)	44.58 (0.02)	53.52 (0.01)	57.50 (0.00)	50.83 (0.00)	58.27 (0.01)	**59.87** (0.02)
DAMRE	69.01 (0.02)	92.50 (0.00)	88.57 (0.01)	84.42 (0.01)	91.62 (0.00)	87.26 (0.05)	93.34 (0.00)	91.46 (0.00)	80.39 (0.05)	82.71 (0.02)	91.19 (0.01)	78.40 (0.06)	64.58 (0.02)	**93.87** (0.01)	89.74 (0.01)	90.59 (0.00)	92.70 (0.00)	93.00 (0.00)
Fraud	48.00 (0.01)	50.05 (0.00)	45.38 (0.00)	46.31 (0.00)	50.15 (0.00)	50.01 (0.00)	48.11 (0.01)	50.14 (0.00)	52.42 (0.01)	50.37 (0.00)	50.00 (0.00)	50.22 (0.00)	45.51 (0.01)	**60.44** (0.00)	49.25 (0.00)	45.38 (0.00)	46.09 (0.01)	56.83 (0.01)
Lymp	70.50 (0.08)	87.05 (0.04)	85.60 (0.03)	85.45 (0.03)	86.85 (0.03)	87.64 (0.02)	91.55 (0.04)	92.10 (0.03)	93.97 (0.02)	77.00 (0.05)	88.18 (0.07)	85.52 (0.05)	99.45 (0.00)	**99.53** (0.00)	94.29 (0.05)	99.14 (0.01)	97.85 (0.01)	99.06 (0.00)
NHIS	65.73 (0.04)	67.48 (0.00)	64.98 (0.04)	65.37 (0.03)	66.26 (0.02)	71.51 (0.01)	87.85 (0.02)	66.13 (0.04)	82.52 (0.03)	67.33 (0.03)	50.63 (0.03)	66.99 (0.03)	47.89 (0.01)	53.48 (0.01)	89.62 (0.01)	80.43 (0.03)	80.75 (0.03)	**91.19** (0.00)
OS	77.16 (0.01)	93.30 (0.01)	84.75 (0.01)	76.86 (0.00)	79.57 (0.03)	91.17 (0.01)	**97.07** (0.00)	86.50 (0.00)	86.77 (0.02)	79.20 (0.03)	94.52 (0.01)	93.17 (0.00)	73.70 (0.07)	92.39 (0.01)	93.39 (0.01)	85.97 (0.01)	95.02 (0.00)	96.94 (0.01)
Seismic	70.58 (0.01)	**74.45** (0.01)	70.41 (0.01)	69.80 (0.01)	69.39 (0.01)	71.93 (0.01)	71.72 (0.00)	73.23 (0.00)	71.01 (0.00)	67.40 (0.00)	71.14 (0.01)	72.95 (0.01)	65.31 (0.11)	73.90 (0.03)	68.88 (0.02)	70.06 (0.00)	72.74 (0.00)	73.24 (0.01)
SMD	81.27 (0.01)	92.37 (0.00)	93.40 (0.00)	91.14 (0.00)	68.64 (0.02)	15.45 (0.01)	86.70 (0.01)	92.60 (0.00)	47.32 (0.04)	50.75 (0.02)	82.44 (0.02)	78.77 (0.01)	74.15 (0.07)	91.79 (0.00)	82.57 (0.05)	92.68 (0.00)	**93.73** (0.00)	92.29 (0.01)
SPD	48.37 (0.01)	49.19 (0.01)	47.86 (0.01)	49.15 (0.01)	48.48 (0.01)	49.10 (0.04)	49.89 (0.01)	49.40 (0.01)	44.95 (0.02)	**53.61** (0.02)	48.60 (0.02)	47.88 (0.01)	51.55 (0.03)	50.50 (0.01)	49.19 (0.01)	49.99 (0.02)	51.99 (0.01)	53.13 (0.02)
Vifd	50.67 (0.01)	52.27 (0.01)	49.31 (0.01)	50.28 (0.01)	49.93 (0.01)	49.91 (0.01)	56.12 (0.01)	52.94 (0.01)	56.45 (0.00)	49.97 (0.00)	55.29 (0.02)	50.96 (0.01)	52.60 (0.04)	55.96 (0.01)	53.73 (0.00)	53.81 (0.00)	54.22 (0.01)	**57.60** (0.01)

C.2 F1-SCORE

F1-score is reported as a complementary metric. Table 4 provides the F1-score and standard deviation for each dataset.

C.3 RANK

We additionally compare methods using their average AUC-ROC ranks across datasets. On each dataset, methods are ranked from 1 (**best**) to 18 (**worst**), and Figure 7 reports the resulting average ranks.

D COMPUTATIONAL EFFICIENCY

Figure 8 compares the average wall-clock runtime (in minutes) of all methods across 30 datasets. All runtime results are measured on a computer with GPU: NVIDIA GeForce RTX 4070 Super, CPU: 20 cores, Memory: 16GB.

Table 4: F1-score (standard deviation) for each dataset (*higher is better*).

Dataset	Iforest	KNN	PCA	ECOD	DeepSVDD	GOAD	ICL	RCA	SLAD	Neutral	DTE	REPEN	AnoGAN	AnoLLM	DRL	LLM-DAS	CausalAno	NestedAno
Annthyroid	55.12 (0.02)	43.94 (0.00)	47.94 (0.02)	38.82 (0.00)	42.88 (0.03)	23.03 (0.00)	53.12 (0.01)	37.33 (0.01)	40.51 (0.00)	42.44 (0.01)	**78.72** (0.01)	33.77 (0.01)	29.46 (0.11)	60.36 (0.01)	51.12 (0.01)	46.44 (0.02)	62.80 (0.02)	63.30 (0.02)
BreastW	96.27 (0.01)	**96.55** (0.01)	95.99 (0.01)	94.61 (0.01)	93.36 (0.01)	96.41 (0.01)	96.13 (0.01)	95.99 (0.01)	95.71 (0.01)	96.13 (0.01)	96.27 (0.01)	91.70 (0.04)	96.40 (0.00)	96.13 (0.00)	96.13 (0.01)	95.57 (0.01)	95.44 (0.00)	95.58 (0.01)
Cardio	68.94 (0.01)	66.86 (0.01)	**81.06** (0.01)	65.91 (0.02)	58.33 (0.05)	27.46 (0.03)	67.43 (0.01)	76.51 (0.01)	60.60 (0.01)	53.79 (0.02)	68.94 (0.02)	56.25 (0.01)	69.51 (0.07)	62.31 (0.00)	69.70 (0.04)	70.27 (0.04)	71.59 (0.03)	72.92 (0.02)
Ecoli	70.37 (0.05)	74.08 (0.05)	**77.78** (0.00)	37.03 (0.05)	**77.78** (0.00)	55.56 (0.09)	74.08 (0.05)	59.26 (0.05)	44.44 (0.09)	66.67 (0.09)	**77.78** (0.00)	59.26 (0.26)	25.92 (0.14)	**77.78** (0.00)	66.67 (0.09)	48.15 (0.21)	**77.78** (0.00)	**77.78** (0.00)
Glass	11.11 (0.00)	18.52 (0.05)	11.11 (0.00)	14.81 (0.05)	18.52 (0.05)	14.81 (0.05)	**29.63** (0.19)	18.52 (0.10)	11.11 (0.00)	25.92 (0.14)	18.52 (0.05)	03.70 (0.05)	11.11 (0.00)	11.11 (0.00)	18.52 (0.05)	11.11 (0.00)	18.52 (0.10)	11.11 (0.00)
ImgSeg	75.49 (0.00)	85.32 (0.01)	72.39 (0.01)	65.86 (0.00)	85.22 (0.01)	77.14 (0.00)	**86.97** (0.01)	75.15 (0.01)	83.37 (0.01)	85.93 (0.00)	82.66 (0.01)	63.40 (0.01)	68.08 (0.06)	71.21 (0.00)	83.81 (0.02)	73.07 (0.01)	86.57 (0.01)	85.96 (0.01)
MagicGamma	70.51 (0.00)	77.02 (0.00)	65.40 (0.00)	60.32 (0.00)	66.73 (0.00)	60.88 (0.01)	70.76 (0.01)	69.35 (0.00)	67.80 (0.00)	74.86 (0.00)	**79.91** (0.00)	71.13 (0.01)	64.14 (0.01)	65.31 (0.00)	72.91 (0.01)	65.57 (0.00)	77.43 (0.01)	76.23 (0.01)
Mammography	40.25 (0.01)	40.77 (0.02)	47.05 (0.01)	**53.59** (0.00)	48.97 (0.03)	28.97 (0.01)	26.67 (0.00)	34.87 (0.01)	13.72 (0.01)	13.08 (0.01)	34.87 (0.04)	29.23 (0.02)	28.72 (0.09)	35.90 (0.01)	33.21 (0.01)	49.23 (0.01)	37.43 (0.04)	42.18 (0.01)
PageBlocks	63.07 (0.01)	75.10 (0.01)	64.44 (0.00)	57.19 (0.00)	67.91 (0.01)	63.20 (0.01)	74.84 (0.00)	67.45 (0.00)	57.52 (0.04)	64.25 (0.02)	76.01 (0.02)	63.53 (0.01)	43.47 (0.02)	58.36 (0.01)	80.72 (0.01)	63.73 (0.00)	**82.88** (0.00)	78.43 (0.01)
Pendigits	56.20 (0.01)	91.45 (0.02)	44.02 (0.01)	42.74 (0.01)	36.75 (0.07)	00.00 (0.00)	77.35 (0.14)	56.84 (0.03)	38.03 (0.05)	46.58 (0.04)	57.69 (0.02)	30.98 (0.03)	41.88 (0.20)	47.65 (0.02)	67.52 (0.10)	54.49 (0.01)	83.98 (0.04)	**91.88** (0.01)
Shuttle	97.07 (0.00)	97.66 (0.00)	95.72 (0.00)	91.69 (0.00)	97.86 (0.00)	96.25 (0.00)	97.93 (0.00)	96.94 (0.00)	97.47 (0.00)	98.40 (0.00)	97.36 (0.00)	96.77 (0.00)	95.46 (0.00)	98.09 (0.00)	98.13 (0.00)	97.77 (0.00)	**98.59** (0.00)	97.80 (0.00)
Stamps	63.44 (0.02)	58.06 (0.05)	59.14 (0.04)	47.31 (0.11)	51.62 (0.10)	17.20 (0.04)	61.29 (010)	56.99 (0.03)	23.66 (0.06)	48.39 (0.07)	51.61 (0.03)	36.56 (0.09)	32.26 (0.19)	66.67 (0.03)	64.51 (0.05)	55.91 (0.04)	**73.12** (0.04)	66.67 (0.05)
Thyroid	**81.00** (0.02)	63.80 (0.01)	73.83 (0.01)	63.44 (0.00)	49.82 (0.07)	41.58 (0.02)	74.91 (0.03)	61.29 (0.01)	64.16 (0.05)	27.96 (0.05)	**81.00** (0.01)	37.64 (0.04)	59.50 (0.11)	79.21 (0.01)	71.68 (0.05)	62.37 (0.02)	70.97 (0.02)	73.84 (0.01)
Vowels	30.00 (0.07)	70.67 (0.03)	18.67 (0.01)	22.00 (0.00)	54.00 (0.07)	42.00 (0.04)	76.67 (0.06)	42.67 (0.02)	70.00 (0.02)	82.00 (0.02)	79.33 (0.01)	26.00 (0.02)	05.33 (0.02)	21.33 (0.01)	**82.67** (0.05)	22.67 (0.05)	81.33 (0.01)	77.33 (0.05)
Waveform	11.67 (0.02)	28.67 (0.00)	08.67 (0.00)	07.67 (0.00)	08.67 (0.03)	04.00 (0.02)	38.00 (0.08)	14.00 (0.01)	03.33 (0.01)	**44.67** (0.03)	12.67 (0.04)	31.67 (0.04)	06.67 (0.03)	09.33 (0.00)	20.00 (0.07)	12.33 (0.03)	31.33 (0.01)	29.67 (0.05)
WBC	86.67 (0.05)	**90.00** (0.00)	83.33 (0.05)	83.33 (0.05)	43.33 (0.05)	73.33 (0.05)	73.33 (0.12)	80.00 (0.08)	86.67 (0.05)	30.00 (0.08)	56.67 (0.17)	73.33 (0.05)	**90.00** (0.00)	86.67 (0.05)	76.67 (0.05)	76.67 (0.05)	70.00 (0.08)	80.00 (0.00)
Wine	91.55 (0.02)	97.65 (0.01)	98.12 (0.01)	81.69 (0.02)	93.42 (0.03)	97.18 (0.00)	96.24 (0.02)	95.30 (0.01)	98.12 (0.02)	97.18 (0.00)	98.12 (0.01)	62.91 (0.03)	96.71 (0.01)	97.65 (0.01)	98.12 (0.01)	95.77 (0.01)	**98.59** (0.00)	98.12 (0.01)
Yeast	48.77 (0.02)	35.09 (0.01)	44.56 (0.01)	33.68 (0.00)	22.81 (0.01)	01.05 (0.00)	35.44 (0.01)	37.89 (0.01)	07.02 (0.02)	09.12 (0.05)	35.44 (0.09)	25.27 (0.04)	**53.33** (0.05)	31.23 (0.01)	32.28 (0.04)	42.46 (0.01)	37.19 (0.00)	37.54 (0.03)
ACD	08.09 (0.01)	08.64 (0.00)	08.64 (0.01)	10.61 (0.01)	07.96 (0.02)	08.91 (0.01)	10.20 (0.00)	08.57 (0.01)	08.98 (0.01)	09.46 (0.02)	11.63 (0.00)	08.64 (0.01)	09.73 (0.01)	10.75 (0.01)	08.57 (0.00)	10.27 (0.01)	10.14 (0.00)	**12.79** (0.01)
Bank	21.62 (0.00)	21.18 (0.00)	21.96 (0.00)	21.96 (0.00)	20.18 (0.01)	21.42 (0.00)	24.86 (0.01)	21.33 (0.01)	22.89 (0.00)	21.03 (0.00)	34.40 (0.01)	22.08 (0.00)	24.96 (0.02)	39.53 (0.00)	28.07 (0.01)	35.38 (0.00)	**42.27** (0.01)	39.83 (0.01)
CMC	13.79 (0.00)	10.34 (0.00)	06.90 (0.00)	06.90 (0.00)	11.49 (0.03)	09.19 (0.02)	10.34 (0.00)	12.64 (0.02)	12.64 (0.02)	**16.09** (0.03)	10.34 (0.03)	08.05 (0.02)	03.45 (0.03)	08.05 (0.02)	10.34 (0.03)	06.90 (0.00)	12.64 (0.02)	**16.09** (0.02)
DAMRE	37.00 (0.08)	81.67 (0.01)	74.00 (0.02)	65.00 (0.01)	77.67 (0.03)	63.33 (0.02)	82.33 (0.02)	81.67 (0.00)	44.33 (0.12)	51.00 (0.04)	75.00 (0.02)	50.00 (0.09)	29.67 (0.02)	**84.33** (0.01)	69.33 (0.02)	84.00 (0.02)	84.00 (0.01)	83.33 (0.00)
Fraud	15.87 (0.01)	17.63 (0.00)	14.76 (0.00)	15.46 (0.00)	17.01 (0.00)	16.94 (0.00)	15.34 (0.01)	17.59 (0.00)	13.75 (0.02)	17.34 (0.00)	17.07 (0.00)	17.42 (0.00)	16.88 (0.00)	**25.22** (0.00)	16.25 (0.00)	14.74 (0.00)	14.70 (0.00)	19.04 (0.01)
Lymp	27.78 (0.08)	66.67 (0.00)	66.67 (0.00)	38.89 (0.08)	55.56 (0.08)	61.11 (0.08)	66.67 (0.00)	72.22 (0.08)	66.67 (0.00)	55.56 (0.16)	72.22 (0.08)	61.11 (0.08)	83.33 (0.14)	**88.89** (0.08)	**88.89** (0.08)	77.78 (0.08)	66.67 (0.00)	83.33 (0.00)
NHIS	21.43 (0.04)	26.79 (0.01)	09.52 (0.01)	14.29 (0.00)	26.19 (0.03)	26.79 (0.00)	33.93 (0.03)	27.38 (0.01)	08.34 (0.04)	08.33 (0.01)	02.38 (0.01)	27.38 (0.01)	01.79 (0.00)	00.00 (0.00)	11.91 (0.02)	29.17 (0.01)	27.98 (0.02)	**34.52** (0.02)
OS	50.55 (0.02)	68.50 (0.02)	64.47 (0.02)	51.65 (0.02)	50.55 (0.05)	62.64 (0.03)	**87.54** (0.01)	65.20 (0.03)	49.82 (0.01)	41.76 (0.07)	79.49 (0.03)	64.47 (0.01)	38.83 (0.07)	69.60 (0.02)	69.23 (0.03)	65.93 (0.01)	86.81 (0.01)	86.45 (0.04)
Seismic	26.47 (0.03)	**32.35** (0.01)	28.63 (0.02)	28.24 (0.00)	23.33 (0.02)	28.82 (0.01)	30.98 (0.01)	30.59 (0.00)	30.39 (0.01)	21.96 (0.01)	24.12 (0.03)	30.19 (0.01)	29.41 (0.05)	32.16 (0.01)	24.12 (0.02)	28.82 (0.02)	28.82 (0.00)	30.79 (0.03)
SMD	34.40 (0.02)	52.93 (0.01)	**57.47** (0.00)	51.07 (0.01)	16.67 (0.02)	0.00 (0.00)	46.53 (0.05)	53.33 (0.03)	04.80 (0.01)	09.20 (0.01)	42.53 (0.04)	29.07 (0.01)	30.00 (0.05)	51.20 (0.02)	38.40 (0.09)	54.80 (0.02)	54.67 (0.02)	54.53 (0.02)
SPD	86.65 (0.00)	86.65 (0.00)	86.21 (0.00)	86.74 (0.00)	86.26 (0.00)	87.09 (0.01)	86.52 (0.00)	86.56 (0.00)	85.99 (0.00)	**87.21** (0.00)	86.61 (0.00)	86.39 (0.00)	86.78 (0.00)	86.74 (0.00)	86.43 (0.00)	86.08 (0.00)	86.65 (0.00)	86.95 (0.00)
Vifd	11.77 (0.01)	14.23 (0.00)	10.87 (0.01)	11.41 (0.01)	11.45 (0.00)	11.77 (0.01)	17.26 (0.00)	15.06 (0.01)	16.65 (0.01)	11.74 (0.00)	12.64 (0.01)	13.07 (0.01)	12.78 (0.03)	**18.78** (0.00)	16.51 (0.01)	13.51 (0.01)	15.49 (0.00)	16.03 (0.00)

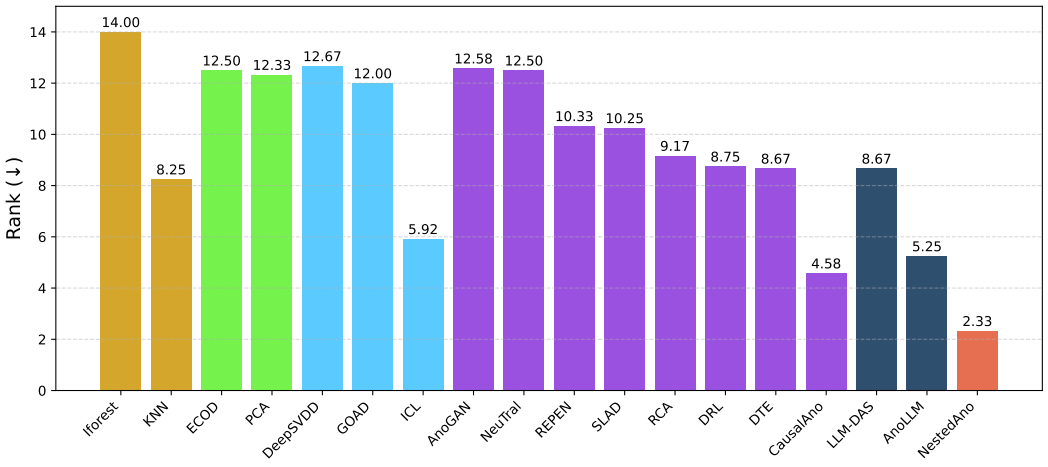


Figure 7: Average ranks across all 30 datasets (*lower is better*).

E Robustness to Causal Graph

We evaluate the sensitivity of NestedAno to the quality of the estimated causal graph. Starting from the ground-truth graph of the Synthetic dataset, we progressively perturb its structure to construct six graph-quality levels, as illustrated in Figure 9. We report AUC-ROC under each setting to examine how structural misspecification affects anomaly detection.

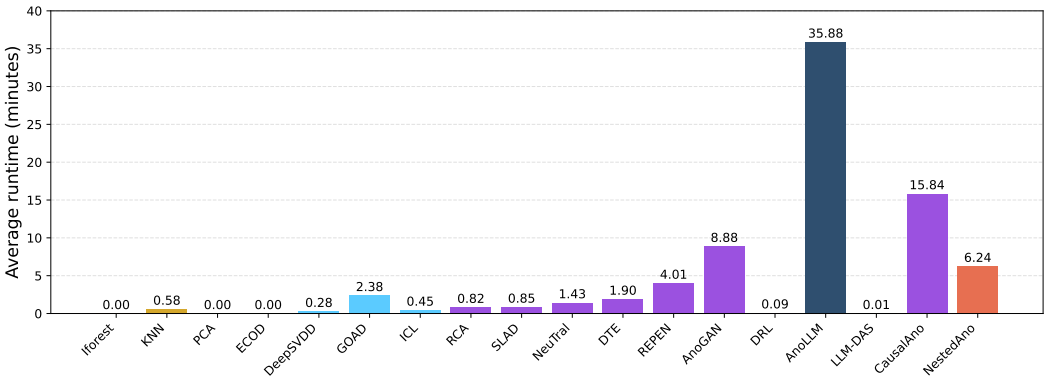


Figure 8: Average runtime (in *minutes*) of top methods over all 30 datasets.

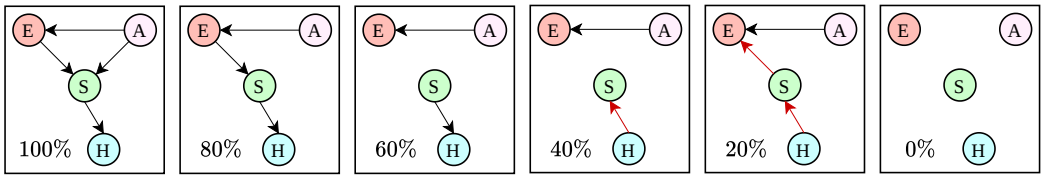


Figure 9: Causal graph configurations used for robustness evaluation on Synthetic dataset. The graph-quality levels represent: 100%—true causal mechanism; 80%—missed one parent of Salary; 60%—both parents of Salary missed; 40%—reversed causal direction; 20%—effects treated as causes; and 0%—no causal structure.

As shown in Figure 10, NestedAno performs best when the causal graph closely matches the ground-truth structure, while moderate graph misspecification leads to only limited degradation. Performance decreases more noticeably when causal dependencies are severely distorted or entirely removed. Overall, these results indicate that informative causal structure benefits anomaly detection, while NestedAno remains relatively robust to imperfect estimated graphs.

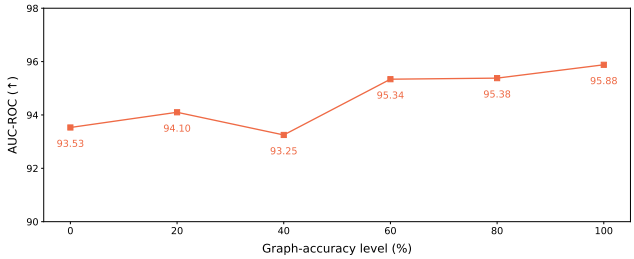


Figure 10: AUC-ROC of NestedAno under different causal graph-quality levels on the Synthetic dataset.

F Effect of Variable Ordering

The structure-guided supervision in NestedAno relies on the ordering of variables assigned to the nested representations. To examine whether this ordering is important, we compare three strategies on the Synthetic dataset: *Original Order*, which follows the input column order; *Random Order*, which randomly permutes the variables; and *Causal Order*, which follows the parent-to-child structure of the causal graph. Figure 11 illustrates the three ordering schemes used in this ablation.

As shown in Figure 12, Causal Order yields the highest performance, outperforming both Original Order and Random Order. Random ordering performs slightly better than the original column

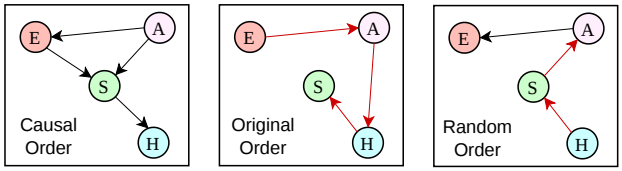


Figure 11: Variable-ordering strategies for assigning variables to the nested representations.

ordering, suggesting that an arbitrary permutation may occasionally align more favorably with the underlying dependency structure.

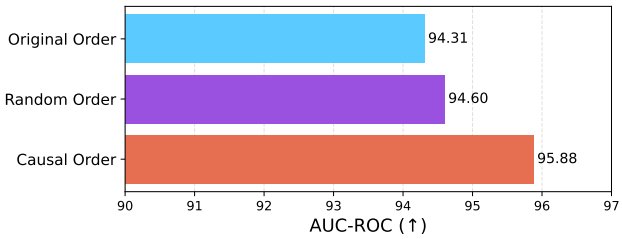
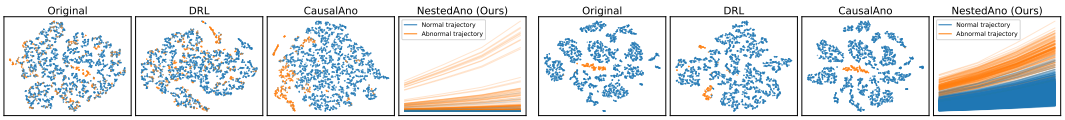


Figure 12: Effect of variable ordering on NestedAno performance on the Synthetic dataset.

G Additional Visualization Results

Figure 13 provides the additional visualization results for two datasets as Annthyroid and Pendigits, complementing the OS and DAMRE results presented in the main paper. Similar behavior is observed on both datasets, where anomalous samples generally exhibit higher trajectories than normal samples.



(a) Annthyroid (b) Pendigits

Figure 13: Additional visualization results on Annthyroid and Pendigits.

H Proof of Theorem 2

(i) By definition, the mutual information between the continuous variable $\mathbf{Z}^{(m_k)}$ and the discrete variable Y is the Kullback-Leibler divergence between their joint distribution and the product of their marginals:

$$I(\mathbf{Z}^{(m_k)}; Y) = D_{\mathrm{KL}}\left(p(\mathbf{z}^{(m_k)}, y) \parallel p(\mathbf{z}^{(m_k)})P(y)\right)$$

Expanding this expectation across the discrete values of $Y \in \{0, 1\}$, and letting $p_y(\mathbf{z}^{(m_k)}) = p(\mathbf{z}^{(m_k)} \mid Y = y)$ with $P(Y = 1) = P(Y = 0) = \frac{1}{2}$:

$$\begin{aligned} I(\mathbf{Z}^{(m_k)}; Y) &= \sum_{y\in\{0,1\}} P(y) \int p(\mathbf{z}^{(m_k)} \mid y) \ln \frac{p(\mathbf{z}^{(m_k)} \mid y)P(y)}{p(\mathbf{z}^{(m_k)})P(y)} d\mathbf{z}^{(m_k)} \\ &= P(Y = 1) \int p_1(\mathbf{z}^{(m_k)}) \ln \frac{p_1(\mathbf{z}^{(m_k)})}{p(\mathbf{z}^{(m_k)})} d\mathbf{z}^{(m_k)} + P(Y = 0) \int p_0(\mathbf{z}^{(m_k)}) \ln \frac{p_0(\mathbf{z}^{(m_k)})}{p(\mathbf{z}^{(m_k)})} d\mathbf{z}^{(m_k)} \\ &= \frac{1}{2} D_{\mathrm{KL}}\left(p_1(\mathbf{z}^{(m_k)}) \parallel p(\mathbf{z}^{(m_k)})\right) + \frac{1}{2} D_{\mathrm{KL}}\left(p_0(\mathbf{z}^{(m_k)}) \parallel p(\mathbf{z}^{(m_k)})\right) \end{aligned}$$

Since the marginal $p(\mathbf{z}^{(m_k)}) = \frac{1}{2}p_1(\mathbf{z}^{(m_k)}) + \frac{1}{2}p_0(\mathbf{z}^{(m_k)})$, the RHS is precisely the definition of the Jensen-Shannon Divergence:

$$I(\mathbf{Z}^{(m_k)}; Y) = \text{JSD}\left(p(\mathbf{z}^{(m_k)} \mid Y = 1) \parallel p(\mathbf{z}^{(m_k)} \mid Y = 0)\right)$$

(ii) By Pinsker's inequality, the Jensen-Shannon Divergence is bounded below by the squared Total Variation (TV) distance:

$$\text{JSD}(p_1 \parallel p_0) \geq \frac{1}{2}\|p_1 - p_0\|_{\text{TV}}^2$$

If the feature space support $\mathbf{z}^{(m_k)} \subseteq \mathbb{R}^{m_k}$ is bounded with diameter $R = \sup_{\mathbf{z},\mathbf{z}'} \|\mathbf{z} - \mathbf{z}'\|_2$, the TV distance bounds the 1-Wasserstein distance W_1:

$$\|p_1 - p_0\|_{\text{TV}} \geq \frac{1}{R}W_1(p_1, p_0)$$

Recalling that the adversarial objective at scale m_k in Equation (4) acts as the Kantorovich-Rubinstein dual formulation bounding the Wasserstein distance:

$$W_1(p_1, p_0) \geq \mathbb{E}_{\mathbf{x}\sim p_{\mathcal{D}}}\left[D^{(m_k)}(\mathbf{x})\right] - \mathbb{E}_{\hat{\mathbf{x}}\sim p_G}\left[D^{(m_k)}(\hat{\mathbf{x}})\right] = -\mathcal{L}_{\text{adv}}^{(m_k)}$$

Substituting this into the TV bound yields the target mutual information bound:

$$I(\mathbf{Z}^{(m_k)}; Y) \geq \frac{1}{2R^2}\left(\mathbb{E}_{\mathbf{x}\sim p_{\mathcal{D}}}\left[D^{(m_k)}(\mathbf{x})\right] - \mathbb{E}_{\hat{\mathbf{x}}\sim p_G}\left[D^{(m_k)}(\hat{\mathbf{x}})\right]\right)^2 = \frac{1}{2R^2}\left(-\mathcal{L}_{\text{adv}}^{(m_k)}\right)^2$$