

Article

Enhancing Contrastive PU Learning for Ad Fraud Detection with Multi-Granularity Diffusion

Lifei Wei ¹ , Weifan Yang ¹ , Yan Meng ² , Xinyu Meng ¹  and Le Yu ^{1*} 

¹ College of Information Engineering, Shanghai Maritime University, 1550 Haigang Avenue, Pudong New Area, Shanghai 201306, China; lfwei@shmtu.edu.cn (L.W.); yangweifan@stu.shmtu.edu.cn (W.Y.); xymeng@shmtu.edu.cn (X.M.)

² School of Computer Science, Shanghai Jiao Tong University, 800 Dongchuan Road, Minhang District, Shanghai 200240, China; yan_meng@sjtu.edu.cn

* Correspondence: yule@shmtu.edu.cn

Abstract

Digital ad fraud continuously evolves through device spoofing, behavior simulation, and coordinated traffic attacks, making the detection of sophisticated fraudulent activities a persistent challenge. In real-world settings, only a small fraction of fraudulent traffic receives reliable labels, while vast amounts of unlabeled data may still contain latent fraud. This severely limits the generalization ability of conventional supervised learning models. To address this issue, we propose DiffVC—a novel contrastive learning framework tailored for ad fraud detection under the Positive-Unlabeled (PU) learning paradigm. DiffVC employs a multi-granularity diffusion augmentation strategy that builds upon a denoising diffusion probabilistic model to generate semantically augmented samples at three granularities: weak, medium, and strong. This strategy expands the latent fraud feature space while preserving diverse semantic information. We further introduce a diffusion-distance-calibrated similarity that dynamically adjusts constraints between augmented samples based on their diffusion distance, thereby improving the discrimination of unlabeled fraud. In addition, we design a dual-gated residual Transformer encoder that adaptively captures high-order feature interactions via gated residual connections and a channel recalibration mechanism. Experimental results on multiple real-world ad fraud detection datasets demonstrate that DiffVC consistently outperforms state-of-the-art methods and achieves stronger generalization. Our results confirm the effectiveness and practical applicability of DiffVC for ad fraud detection under label-scarce conditions.

Keywords: ad fraud; traffic detection; PU learning; diffusion model; contrastive learning



Academic Editor: Guanqiu Qi

Received: 13 July 2026

Revised: 13 August 2026

Accepted: 14 August 2026

Published: 19 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

Digital advertising has become a dominant channel for online monetization through programmatic bidding and real-time ad delivery systems, which process massive volumes of ad requests within milliseconds [1,2]. As Figure 1 illustrates, the digital advertising ecosystem mainly comprises advertisers (e.g., Pepsi), publishers, demand-side platforms (DSPs), supply-side platforms (SSPs, e.g., MoPub), and ad exchanges (e.g., DoubleClick). This system handles the entire pipeline—from request submission and auction to ad rendering—under strict time constraints. Although this architecture enables rapid monetization, it also exposes a broad attack surface. The opacity of the multi-hop supply chain further complicates fraud detection because each intermediary introduces additional opportunities for manipulation.

In recent years, ad fraud has evolved into a sophisticated and organized form of cybercrime, where fraudsters operate large-scale botnets and coordinated click farms [3]. These adversaries inject fraudulent requests into legitimate traffic streams and mimic genuine user behavior to evade conventional filters. Common fraudulent techniques include click fraud [4], impression manipulation [5], and attribution hijacking [6]. Beyond directly draining advertising budgets, these fraudulent activities distort performance analytics, skew training data for algorithmic bidding, and undermine advertisers' trust in programmatic advertising channels. Furthermore, strict system latency requirements constrain the deployment of computationally intensive machine-learning classifiers, creating a trade-off between detection accuracy and response speed—a trade-off that fraudsters continually exploit by developing more sophisticated evasion strategies.

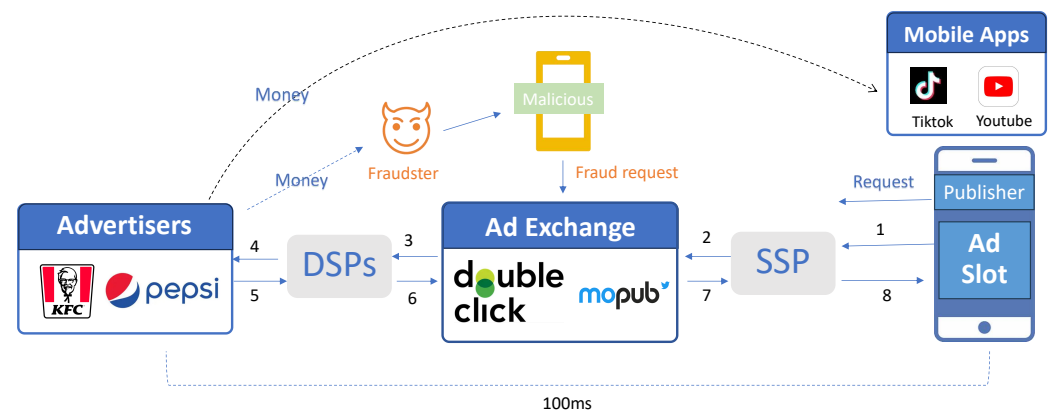


Figure 1. Background of the digital advertising ecosystem and existing ad fraud behaviors.

To combat ad fraud, industrial advertising platforms commonly perform traffic verification by detecting fraudulent patterns in ad transaction logs collected from advertising networks [7]. Traffic that fails verification falls into two categories: general invalid traffic (GIVT) and sophisticated invalid traffic (SIVT). Rule-based methods readily filter GIVT, whereas SIVT presents a substantially greater challenge because attackers deliberately design tactics that imitate legitimate user behavior. Since no single rule reliably identifies SIVT, researchers increasingly employ machine-learning approaches [8,9] to automatically detect these sophisticated fraudulent behaviors.

However, practical ad fraud detection suffers from extreme label scarcity. In real-world advertising systems, only a small fraction of fraudulent traffic receives reliable labels through costly manual inspection, while most traffic remains unlabeled and may contain undiscovered fraudulent instances [10]. Consequently, ad fraud detection naturally fits the Positive-Unlabeled (PU) learning paradigm [11], in which models learn discriminative decision boundaries from a small set of labeled positive samples and a large collection of unlabeled data.

Since fraudulent and legitimate traffic mix heavily within the unlabeled set [12], conventional supervised objectives often fail to learn robust feature representations. This limitation leads to substantial overlap between fraudulent and legitimate representations and weakens generalization to previously unseen fraud patterns [13]. The problem becomes even more pronounced in dynamic advertising environments, where fraudsters continuously evolve their behaviors and gradually deviate from previously observed attack distributions.

To alleviate the supervision shortage in PU learning, a natural approach employs contrastive learning to extract discriminative embeddings from unlabeled data [14]. The ef-

fectiveness of this approach primarily depends on the quality and semantic diversity of data augmentation.

Conventional augmentation methods—such as random perturbation, oversampling, and SMOTE [15]—cannot apply directly to fraud traffic detection, where ad transaction logs are tabular. These methods often ignore complex dependencies between numerical and categorical columns, producing unrealistic samples that violate underlying business logic. Recent advances in denoising diffusion probabilistic models (DDPMs) offer a promising solution for constructing semantically augmented samples from heterogeneous tabular data [16]. However, whether DDPMs serve as an effective augmentation method for fraud traffic detection remains an open question.

Despite these advances, several critical challenges remain unresolved when applying diffusion-based augmentation to PU-based fraud detection. First, existing PU learning approaches for fraud detection [11,17–19] primarily rely on pseudo-labeling or reconstruction-based strategies to leverage unlabeled data. However, these approaches suffer from error propagation and biased representation learning when the unlabeled set contains a large proportion of latent positive samples. Second, although diffusion models demonstrate strong capabilities in generating realistic heterogeneous tabular data [16,20,21], effectively exploiting diffusion processes to construct semantically controllable augmentations for contrastive PU learning remains largely unexplored. In particular, existing methods rarely consider the relationship between diffusion-induced perturbation levels and the semantic consistency of augmented samples. Third, learning effective fraud representations remains challenging due to the complex nonlinear interactions among heterogeneous advertising features [22]. Existing Transformer-based tabular encoders [23,24] typically lack adaptive mechanisms for balancing feature preservation and transformation, which yields suboptimal representations in deep architectures. These challenges motivate us to explore a unified framework that simultaneously addresses supervision scarcity, semantic augmentation, and discriminative representation learning in PU-based ad fraud detection.

In this paper, we propose a novel multi-granularity diffusion method as a data augmentation mechanism for contrastive PU learning in ad fraud detection. We also propose a novel diffusion-distance-calibrated similarity that serves as an adaptive metric among augmented samples, rather than the fixed binary states of similar or dissimilar used in conventional contrastive models. Our design stems from the following observations. First, fraudulent advertising traffic inherently exhibits hierarchical semantic variations across multiple behavioral scales—from temporal fluctuations and lightweight device randomization to larger behavioral shifts [25,26]. This property aligns with our multi-granularity method for constructing hierarchical semantic augmentations. Different diffusion timesteps progressively control the degree of semantic perturbation: smaller timesteps preserve fine-grained instance-level consistency, while larger timesteps encourage broader exploration of the underlying fraud manifold [27]. Second, because semantic consistency varies across diffusion granularities, enforcing identical similarity constraints across all augmented samples becomes suboptimal [28]. This observation further motivates our diffusion-distance-calibrated similarity contrastive learning mechanism, which adaptively models semantic alignment according to diffusion-induced perturbation strengths.

Motivated by these observations, we propose DiffVC, a novel multi-granularity diffusion variable-similarity contrastive learning framework for PU-based ad fraud detection. Specifically, DiffVC first employs a tabular diffusion model to augment samples at weak, medium, and strong semantic levels, progressively expanding the latent fraud distribution. Based on these augmented samples, we further employ a diffusion-distance-calibrated similarity that dynamically adjusts similarity constraints according to semantic distances. To effectively model complex nonlinear interactions in ad traffic, we further introduce a

dual-gated residual Transformer encoder with adaptive feature fusion and channel recalibration. Extensive experiments on multiple real-world ad fraud datasets demonstrate that DiffVC consistently outperforms state-of-the-art methods in AUC, F1-score, and Recall, while exhibiting substantially stronger robustness under temporal distribution shifts.

The main contributions of this work are summarized as follows:

- We propose DiffVC (The source code is available at <https://github.com/yang1230hub/DiffVC.git> (accessed on 27 May 2026)), a unified diffusion-enhanced similarity contrastive learning framework for PU-based ad fraud detection, which jointly integrates semantic augmentation, contrastive representation learning, and unbiased PU classification under extremely limited fraud supervision.
- We introduce a multi-granularity diffusion augmentation strategy for heterogeneous advertising traffic data, where diffusion timesteps naturally induce semantically progressive reconstruction views, expanding the latent fraud distribution while preserving feature dependency structures.
- We propose a diffusion-distance-calibrated similarity contrastive learning mechanism that relaxes the conventional fixed similarity assumption by adaptively modeling semantic consistency across views generated at different diffusion perturbation strengths.
- We design a dual-gated residual Transformer encoder with adaptive residual fusion and channel recalibration to effectively capture high-order nonlinear feature interactions in large-scale advertising traffic.
- Extensive experiments on multiple real-world datasets demonstrate that DiffVC achieves superior performance and robustness under dynamic advertising environments and temporal distribution shifts.

The remainder of this paper proceeds as follows. Section 2 reviews related work. Section 3 introduces the proposed DiffVC framework. Section 4 reports our experimental results and ablation analyses. Finally, Section 5 discusses the implications of our findings, and Section 6 concludes the paper.

2. Related Work

Advertising Fraud Traffic Detection. Machine learning is essential for uncovering behavioral patterns in ad traffic. Mahesh et al. [9] combined XGBoost with spatiotemporal features to separate human and bot clicks. Zhu et al. [29] proposed a locality-sensitive hashing tensor recovery that reconstructs clicks into a high-order tensor, capturing high-dimensional sparse interactions. Li et al. [30] introduced GemmaWithLoRA, which fine-tunes a pre-trained language model via LoRA to detect complex malicious clicks. For mobile fraud, collusive clusters, and covert attacks, Sun et al. [3] detected mobile click farms through device-cluster-level automated analysis, Zhu et al. [31] designed a collaborative mechanism against multi-party collusion, and Cao et al. [32] built AdSherlock, which combines offline pattern generation with an online tree model for efficient client-side detection with reduced false negatives. Targeting search-engine ad fraud, Zhang et al. [33] developed COSEC to identify incoherent search sessions, separating fabricated bot sessions from a macroscopic session-level view.

Tabular Data Augmentation and Diffusion Models. To address sample scarcity and long-tail distributions, data augmentation is critical. Gorishniy et al. [34] emphasized that tabular data imposes strict business-logic dependencies, making random interpolation or noise injection prone to generating invalid samples. Denoising diffusion probabilistic models (DDPMs) inherently model the joint distribution: Kotelnikov et al. [16] proposed TabDDPM, which employs Gaussian and multinomial diffusions for continuous and discrete features, enabling end-to-end heterogeneous tabular generation. For improved fidelity

and adaptability, Zhang et al. [20] introduced TabSyn, a latent-space score-matching diffusion that recovers the feature manifold in a disentangled space; Shi et al. [21] built TabDiff, a multimodal diffusion with learnable feature-level weights that enhances diversity while preserving business logic. Liu et al. [35] extended unconditional diffusion to tabular synthesis, and Yang et al. [36] proposed a sensitive-attribute-guided diffusion for fair synthetic data with balanced joint distributions. These works underpin our use of diffusion models for multi-granularity, high-fidelity traffic views.

Contrastive Learning. Self-supervised contrastive learning mines discriminative representations by pulling positives together and pushing negatives apart. SCARF [37] relies on random feature corruption, while SAINT [23] integrates row- and column-wise attention with contrastive pre-training. However, Chen et al. [38] identified a “fixed similarity assumption” that forces rigid semantic alignment and causes semantic drift under strong augmentations such as diffusion-based generation. To address this, Cui et al. [28] proposed CLVS, which dynamically weights inter-view similarity by augmentation strength to learn robust hierarchical features. Tao et al. [39] designed semantic masking to protect core features from random perturbations in anomaly detection. Cui et al. [40] further introduced a dynamic, class-conditioned augmentation strategy that adaptively preserves inter-class core features.

PU Learning. Positive-unlabeled (PU) learning trains classifiers from positive and unlabeled data. Elkan and Noto [11] established the SCAR assumption, proving the learned probability differs from the true probability by a constant factor. Song et al. [17] proposed SatPU with sharpened pseudo-labeling and dynamic re-weighting for industrial anomaly detection. Long and Zhang [41] designed PUNet, combining a variational recurrent autoencoder with CatBoost and a pseudo-labeled candidate set to mitigate prediction drift. Takahashi et al. [18] integrated unbiased PU learning with autoencoders and DeepSVDD for contaminated unlabeled data. An et al. [42] developed PUFA for few-shot visual defect detection via cycle consistency and co-occurrence features. Herurkar et al. [19] proposed Tabular Data Adapters that transform private data to public formats via a shared autoencoder to generate soft labels.

Existing studies provide valuable foundations for advertising fraud detection, tabular diffusion generation, contrastive representation learning, and PU learning; however, several important limitations persist. First, although PU learning methods have received extensive investigation in anomaly detection and industrial applications [17–19,41], they typically rely on pseudo-labeling or reconstruction-based objectives and fail to explicitly exploit the rich semantic information embedded in unlabeled advertising traffic. Meanwhile, existing diffusion-based tabular augmentation methods [16,20,21] mainly focus on synthetic data generation, class balancing, or anomaly detection, while their potential for constructing semantically controllable views in PU contrastive fraud detection remains largely unexplored.

Second, conventional contrastive learning frameworks usually assume fixed similarity relationships between augmented views [23,37], which becomes suboptimal when augmentation strengths vary significantly. Although recent studies explore adaptive similarity modeling [28,38], leveraging diffusion-induced perturbation levels as a continuous semantic distance for contrastive learning has not yet been investigated in tabular fraud detection.

Third, existing Transformer-based tabular encoders [24,43] still face challenges in modeling complex heterogeneous feature interactions. They generally lack adaptive mechanisms to dynamically regulate feature transformation and preservation, which limits representation quality for large-scale advertising traffic.

These limitations motivate us to propose DiffVC, a framework that integrates multi-granularity diffusion augmentation, diffusion-distance-calibrated similarity learning,

and adaptive feature interaction modeling for PU-based advertising fraud detection. By jointly optimizing semantic augmentation and representation learning, DiffVC provides a unified solution that addresses label scarcity and complex behavioral heterogeneity in large-scale advertising traffic.

3. Methodology

This section introduces DiffVC, a multi-granularity diffusion-based similarity contrastive learning framework for advertising fraud detection under the Positive-Unlabeled (PU) learning paradigm. To address the realistic challenge that only a small fraction of fraudulent samples receive labels, DiffVC employs a diffusion model to produce diverse multi-granularity augmented views, integrates similarity contrastive learning with a dual-gated Transformer encoder, and optimizes the entire framework end-to-end via a non-negative PU loss. The following subsections detail the overall framework, problem formulation, diffusion-driven view generation, the dual-gated residual Transformer encoder, diffusion-distance-calibrated similarity contrastive learning, and joint optimization.

3.1. Overall Framework

The overall architecture of DiffVC (Figure 2) comprises three core modules: a diffusion-driven view generator, a dual-gated residual Transformer encoder, and a diffusion-distance-calibrated contrastive learning network. We first transform original tabular data into a unified continuous representation via quantile transformation and categorical embedding. The diffusion model then injects noise across three non-overlapping timestep intervals and performs reverse denoising, producing three augmented views at weak, medium, and strong granularities. A weight-sharing dual-gated Transformer encoder extracts deep features from each view, and an augmentation-aware projection module maps these features into a similarity space to compute a contrastive loss guided by the diffusion distance. Under the PU setting, we use only a small set of positive labels and a large pool of unlabeled samples; multi-view classifications jointly optimize with a weighted PU loss. This loss, together with the diffusion reconstruction loss and the contrastive loss, forms the end-to-end training objective. This design captures discriminative features through diffusion generation and contrastive learning, while the unbiased PU loss drives robust fraud detection even under extreme label scarcity.

3.2. Problem Formulation

Given a tabular dataset for mobile ad fraud traffic detection, $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$ containing N samples, the feature vector of the i -th sample is $x^{(i)} = [x_{\text{num}}^{(i)}, x_{\text{cat}}^{(i)}]$, which consists of continuous numerical features $x_{\text{num}}^{(i)} \in \mathbb{R}^{d_{\text{num}}}$ and discrete categorical features $x_{\text{cat}}^{(i)} \in \mathbb{N}^{d_{\text{cat}}}$, where d_{cat} denotes the number of categorical features. Unlike standard binary classification assumptions, we leverage only the known positive samples for PU learning: we label confirmed positives as 1 and treat the remaining unlabeled data as 0 (i.e., $y^{(i)} = 0$), where normal traffic intermingles with undiscovered fraudulent traffic. Our objective is to learn a mapping function $f(\cdot)$ that, given only partial positive class labels, outputs a fraud probability $\hat{y}^{(i)} \in [0, 1]$ for each sample, thereby effectively distinguishing normal from fraudulent traffic.

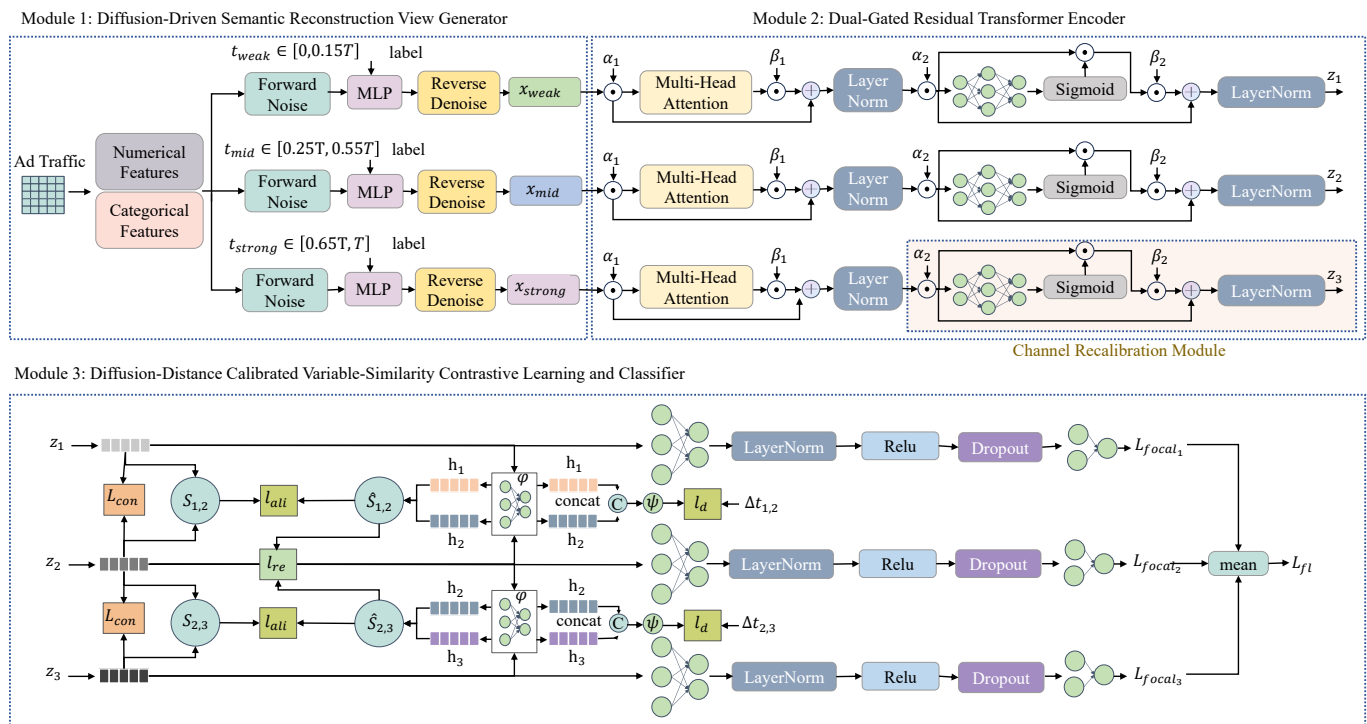


Figure 2. DiffVC model structure.

3.3. Diffusion-Driven Semantic Reconstruction View Generation

To mitigate the severe label scarcity in PU scenarios, we adopt a denoising diffusion probabilistic model that learns the underlying data distribution and generates weak, medium, and strong semantic reconstruction views by controlling the diffusion timestep, thereby substantially enriching sample diversity.

As Figure 3 illustrates, we first transform original input features x_0 via quantile normalization and categorical embedding into a unified continuous representation $x_0 \in \mathbb{R}^d$. The total dimension is $d = d_{\text{num}} + \sum_{m=1}^M K_m$, where d_{num} denotes the number of numerical features, M the number of categorical features, and K_m the cardinality of the m -th categorical feature. The forward diffusion process gradually corrupts x_0 with Gaussian noise according to a predefined variance schedule $\{\beta_t\}_{t=1}^T$, where β_t controls the noise level at step t and T denotes the total number of diffusion steps. For any timestep $t \in [1, T]$, the noisy representation x_t follows the conditional distribution:

$$q(x_t | x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) \mathbf{I}), \quad (1)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ is the cumulative signal retention coefficient. As t increases, $\bar{\alpha}_t \rightarrow 0$ and x_t approaches a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, progressively erasing the original semantics.

To construct hierarchical multi-granularity views, we partition the full timestep range into three non-overlapping intervals and uniformly sample t from each:

$$t_{\text{weak}} \sim \mathcal{U}(0, 0.15T), \quad t_{\text{mid}} \sim \mathcal{U}(0.25T, 0.55T), \quad t_{\text{strong}} \sim \mathcal{U}(0.65T, T). \quad (2)$$

To handle the inherent heterogeneity of advertising traffic—where numerical features (e.g., click-through rates) and categorical features (e.g., device models) follow fundamentally different distributions—our multi-granularity diffusion strategy decouples the forward corruption and reverse reconstruction for each modality.

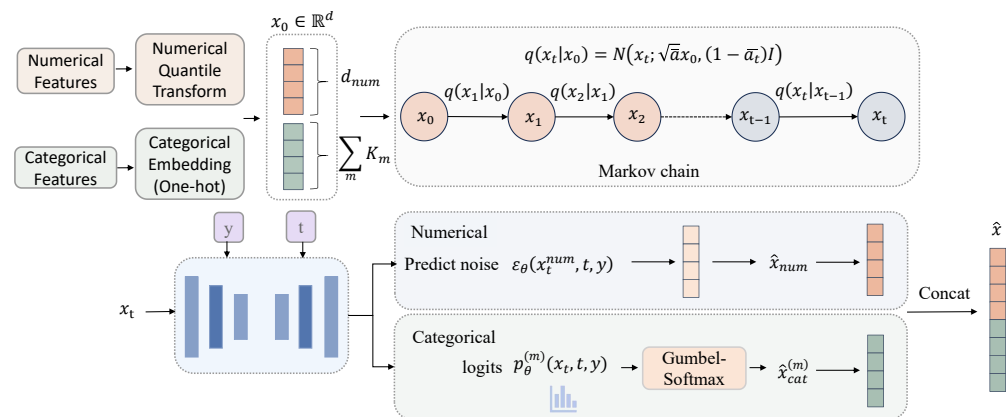


Figure 3. Diffusion-driven semantic reconstruction view generation structure.

For numerical features, the forward process injects Gaussian noise under a variance-preserving schedule, and the reverse network predicts the noise via an L_2 reconstruction loss (Equation (3)). For categorical features, the forward process employs a uniform transition matrix that smooths one-hot vectors toward a uniform prior, and the reverse process recovers discrete tokens by optimizing a multinomial log-likelihood with Gumbel-Softmax reparameterization (Equation (4)). A shared denoising network f_θ jointly optimizes across both modalities through the unified objective in Equation (5), forcing the model to capture cross-modal dependencies and preserve the joint distribution $p(x_{\text{num}}, x_{\text{cat}})$.

The diffusion timestep t further controls the augmentation strength: weak augmentations (t_{weak}) retain the original statistical moments of numerical features and keep categorical features close to one-hot ground truth, preserving high instance-level fidelity; strong augmentations (t_{strong}) drive numerical features toward Gaussian noise and categorical features toward maximum uncertainty, compelling the model to rely on broad structural patterns.

This strategy yields a continuous spectrum from high fidelity to high diversity, spanning both near-in-distribution variations and far-boundary explorations of the fraud manifold. Furthermore, because the reverse process is distribution-driven rather than label-conditioned, the generated views inherently respect business-logic constraints (e.g., incompatible device–OS pairs), which guarantees semantic realism and physical plausibility.

The shared denoising network f_θ parameterized by θ approximates the true reverse distribution $q(x_{t-1} | x_t)$. Taking the noise-corrupted state x_t and timestep t as inputs without any label condition, the generation process under the PU setting remains purely data-driven and robust to label absence and label noise.

To handle the mixed nature of numerical and categorical features in tabular data, the model processes the two feature types separately. Numerical features adopt standard Gaussian denoising:

$$\hat{x}_{\text{num}} = \frac{x_t^{\text{num}} - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t^{\text{num}}, t)}{\sqrt{\bar{\alpha}_t}} \quad (3)$$

where ϵ_θ denotes the injected noise predicted by the network, and a mean squared error loss constrains the prediction accuracy.

Categorical features employ multinomial diffusion to preserve discrete distributions. The network outputs log-probabilities $p_\theta^{(m)}$ for each categorical feature, and the Gumbel-Softmax reparameterization achieves differentiable sampling:

$$\hat{x}_{\text{cat}}^{(m)} = \operatorname{argmax} \left(\operatorname{Softmax} \left(\frac{p_\theta^{(m)}(x_t, t) + g}{\tau_{\text{gumbel}}} \right) \right) \quad (4)$$

where $g \sim \text{Gumbel}(0, 1)$ and τ_{gumbel} is the temperature coefficient that controls the sharpness of the Gumbel-Softmax distribution. During training, we use a smaller temperature coefficient to maintain gradient differentiability; during inference, we apply hard sampling directly to guarantee discrete validity of the predictions.

After obtaining the reconstructed numerical vector $\hat{x}_{\text{num}} \in \mathbb{R}^{d_{\text{num}}}$ from the Gaussian denoising process (Equation (3)) and the reconstructed categorical tokens $\hat{x}_{\text{cat}}^{(m)}$ for each categorical feature from the multinomial diffusion process (Equation (4)), we first concatenate these two components along the feature dimension to form a raw reconstructed sample $\hat{x} = [\hat{x}_{\text{num}}; \hat{x}_{\text{cat}}^{(1)}; \dots; \hat{x}_{\text{cat}}^{(M)}]$, which preserves the original feature ordering. However, directly using this raw concatenation suffers from semantic scale misalignment: numerical features and categorical embeddings reside in different value ranges and representational spaces, which hinders the downstream encoder from effectively learning unified representations.

To address this issue, we introduce a lightweight re-embedding layer that projects the raw concatenated vector into a unified continuous space. Specifically, we apply a linear transformation $\hat{x}_{\text{emb}} = \text{ReLU}(\hat{x}\mathbf{W}_{\text{emb}} + \mathbf{b}_{\text{emb}})$, where $\mathbf{W}_{\text{emb}} \in \mathbb{R}^{(d_{\text{num}} + \sum_m K_m) \times d}$ and $\mathbf{b}_{\text{emb}} \in \mathbb{R}^d$ are learnable parameters, and d is the unified embedding dimension. This transformation aligns the semantic scales of numerical and categorical features, enabling the subsequent dual-gated Transformer encoder to process all features homogeneously. The re-embedding layer shares parameters across all three granularity views ($x_{\text{weak}}, x_{\text{mid}}, x_{\text{strong}}$), ensuring that the augmented samples from different diffusion timesteps map into a consistent representation space. This design also enhances training stability, as the re-embedding layer adaptively adjusts the feature scales during optimization without altering the underlying diffusion generation process. Finally, we concatenate and re-embed the reconstructed numerical and categorical features, yielding the three augmentation views $x_v \in \{x_{\text{weak}}, x_{\text{mid}}, x_{\text{strong}}\}$.

The optimization objective of the diffusion model jointly considers the numerical feature reconstruction MSE and the categorical feature cross-entropy loss, and we define the overall loss function as:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t, x_0, \epsilon} \left[\|\epsilon - \epsilon_{\theta}(x_t^{\text{num}}, t)\|_2^2 + \frac{1}{M} \sum_{m=1}^M \text{CE}(p_{\theta}^{(m)}(x_t, t), x_{\text{cat}, 0}^{(m)}) \right] \quad (5)$$

where the first term minimizes the numerical reconstruction error, and the second term drives the categorical reconstruction to approximate the true distribution. This cross-entropy (CE) term operates exclusively at the feature level for categorical feature reconstruction during the diffusion-based generation process. It ensures that the diffusion model faithfully reconstructs discrete categorical attributes and does not serve the subsequent classification task. This composite design enables joint optimization over numerical and categorical modalities.

This joint constraint ensures that the denoising network accurately captures the joint distribution of the original data, which guarantees that the multi-granularity augmented samples possess both high fidelity and reasonable diversity. Within the PU learning framework, this data augmentation strategy effectively expands the diversity of fraudulent data and enhances the model's detection generalization.

3.4. Dual-Gated Residual Transformer Encoder

Feature interactions in tabular data are global and highly nonlinear, making them difficult for ordinary multilayer perceptrons to capture effectively. Standard Transformers that treat individual features as independent tokens overlook the dependency structure within the feature dimension. To address this issue, we design a dual-gated residual Transformer encoder, whose core comprises a dual-gating mechanism and a channel

recalibration module. This design enables full feature interaction and adaptive selection with a compact structure.

To capture high-order nonlinear feature interactions in large-scale advertising traffic, we design a dual-gated residual Transformer encoder that operates through three complementary mechanisms. First, the multi-head self-attention module explicitly models pairwise feature affinities across all feature dimensions via the dot-product operation $\mathbf{Q}\mathbf{K}^\top$, which effectively captures second-order interactions between every pair of features. By concatenating h heads and applying a linear transformation, the model captures interactions up to h -th order within a single layer. Stacking L such Transformer blocks further compounds this capability, enabling interactions up to $O(h^L)$ -th order. This capability proves particularly critical for advertising traffic data, where fraud patterns often manifest as complex combinatorial relationships across multiple feature fields—for example, the joint occurrence of a specific device type, a particular geographic region, and an anomalous click timing pattern may collectively indicate fraud, whereas any subset of these features alone would be insufficient for reliable detection.

Second, to prevent feature degradation in deep networks and enable adaptive information selection, we introduce dual learnable gating parameters that dynamically balance the contributions of identity mapping, attention-transformed features, and channel-recalibrated features. Third, the channel recalibration module adaptively assigns importance weights to each feature dimension, effectively emphasizing fraud-relevant attributes while suppressing irrelevant or noisy dimensions. Together, these components ensure that the model focuses on genuinely predictive high-order patterns while mitigating overfitting to spurious correlations in large-scale advertising logs.

Given an input feature vector $\mathbf{x} \in \mathbb{R}^d$, we first detail the calculation procedure of the multi-head attention module. The computation of a single attention head is based on scaled dot-product attention:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \quad (6)$$

where $\mathbf{Q} = \mathbf{x}\mathbf{W}^Q$, $\mathbf{K} = \mathbf{x}\mathbf{W}^K$, and $\mathbf{V} = \mathbf{x}\mathbf{W}^V$ are the query, key, and value matrices, $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d \times d_h}$ are projection matrices, $d_h = d/h$ is the dimension per head, and h is the number of heads. Multi-head attention concatenates the outputs of h heads and applies a linear transformation to obtain the final result:

$$\text{MHA}(\mathbf{x}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)\mathbf{W}^O \quad (7)$$

where $\text{head}_i = \text{Attention}(\mathbf{x}\mathbf{W}_i^Q, \mathbf{x}\mathbf{W}_i^K, \mathbf{x}\mathbf{W}_i^V)$, $\mathbf{W}_i^{Q,K,V} \in \mathbb{R}^{d \times d_h}$, and $\mathbf{W}^O \in \mathbb{R}^{d \times d}$.

To mitigate feature degradation in deep networks, we introduce a parameterized gating mechanism after the attention transformation, which adaptively adjusts the fusion ratio between the transformed features and the original features:

$$\mathbf{y} = \alpha_1 \odot \mathbf{x} + \beta_1 \odot \text{MHA}(\mathbf{x}) \quad (8)$$

where $\alpha_1, \beta_1 \in \mathbb{R}^d$ are learnable gating vectors initialized to 1 and 0, respectively. This initialization maintains an approximate identity mapping at the early stage of training to stabilize optimization, and subsequently dynamically adjusts the fusion ratio of the two feature types through gradient backpropagation for adaptive information selection.

Specifically, the gating pair (α_1, β_1) controls the contribution of the multi-head self-attention output relative to the original input features. β_1 determines how much of the attention-transformed features $\text{MHA}(\mathbf{x})$ fuses into the residual stream, while α_1 retains the

direct identity information from $\mathbf{x}$. Their element-wise nature enables dimension-specific control, allowing the model to selectively amplify attention signals for certain feature dimensions while preserving raw features for others.

On this basis, we introduce a channel recalibration module. Drawing inspiration from the squeeze-and-excitation idea [44], this module performs global information compression and channel-wise gating to assign adaptive weights to each dimension:

$$\mathbf{z} = \alpha_2 \odot \mathbf{y} + \beta_2 \odot (\sigma(\mathbf{W}_2 \text{ReLU}(\mathbf{W}_1 \mathbf{y})) \odot \mathbf{y}) \quad (9)$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times (d/4)}$ and $\mathbf{W}_2 \in \mathbb{R}^{(d/4) \times d}$ handle feature compression and recovery, introducing nonlinear transformations through dimensionality reduction followed by expansion; $\sigma(\cdot)$ denotes the Sigmoid activation function, which outputs a channel weight vector with values in $(0, 1)$ and performs adaptive screening of different feature dimensions. The second set of gating parameters α_2, β_2 also initializes to 1 and 0 for adaptively fusing the identity mapping and the recalibrated features.

Here, the second gating pair (α_2, β_2) balances the trade-off between the intermediate representation $\mathbf{y}$ (which already contains fused attention features) and the channel-recalibrated features $\sigma(\mathbf{W}_2 \text{ReLU}(\mathbf{W}_1 \mathbf{y})) \odot \mathbf{y}$. β_2 controls the influence of the squeeze-and-excitation module that adaptively reweights feature channels based on global context, while α_2 preserves the uncalibrated features from $\mathbf{y}$. During training, gradient backpropagation updates both gating pairs, allowing the network to adaptively balance preserving raw features, incorporating attention-derived interactions, and refining representations through channel-wise recalibration.

This encoder operates independently on the weak, medium, and strong augmented samples with shared parameters, producing the corresponding output embeddings $\mathbf{z}_{\text{weak}}, \mathbf{z}_{\text{mid}}, \mathbf{z}_{\text{strong}}$. The encoder captures global nonlinear interactions through multi-head attention, ensures flexibility and stability of information flow via dual gating, and achieves adaptive feature selection via channel recalibration. Together, these mechanisms effectively enhance feature extraction from ad fraud traffic.

3.5. Diffusion-Distance-Calibrated Similarity Contrastive Learning

Design rationale. Two considerations motivate our choice of diffusion-distance-calibrated similarity over conventional fixed-similarity contrastive learning. (1) Semantic consistency varies across augmentation granularities: weak views preserve high similarity to the anchor, while strong views diverge to encourage exploration of the fraud manifold. Enforcing uniform similarity across all views would either suppress the diversity benefits of strong augmentations or distort the semantic integrity of weak augmentations. (2) The diffusion timestep provides a natural, continuous measure of perturbation strength that directly links to the amount of semantic information removed during the forward process, making it an ideal calibration signal. The core idea is that views with different augmentation strengths maintain varying degrees of similarity rather than forcing them toward the same anchor point. Our design incorporates four complementary loss terms—InfoNCE for basic contrastive discrimination, distance prediction for timestep awareness, similarity alignment for consistency, and relative similarity for hierarchical structure—which collectively ensure that the learned embedding space respects the inherent semantic hierarchy induced by multi-granularity diffusion.

To achieve this, we build our framework upon a foundation of standard contrastive learning, which typically relies on the InfoNCE loss. InfoNCE loss serves as the de facto standard for contrastive learning; it maximizes the agreement between different augmented views of the same sample (positive pairs) while minimizing the agreement between views of different samples (negative pairs). However, a key limitation of the standard InfoNCE

loss lies in its “fixed similarity assumption,” i.e., it forces all positive pairs to have a uniformly high similarity regardless of their augmentation strength. This rigid constraint proves suboptimal for our multi-granularity diffusion views, where weak and strong augmentations naturally preserve different levels of semantic information.

By leveraging the diffusion timestep as a natural measure of augmentation strength, we construct a similarity contrastive learning framework that comprises an InfoNCE loss, a similarity alignment loss, a distance prediction loss, and a relative similarity constraint. This framework guides the encoder to form a structured embedding topology where a larger diffusion strength corresponds to a lower view similarity.

Encoding of augmentation strength via diffusion timesteps. To operationalize the variable-similarity idea, we associate each augmented view with its specific diffusion timestep t_v . For any pair of views (i, j) , we define the normalized timestep difference $\Delta t_{ij} = |t_j - t_i|/T$ as the augmentation strength gap. We use this gap in three ways. First, the distance prediction loss l_d (Equation (13)) trains an auxiliary network to regress Δt_{ij} from the concatenated embeddings of the two views, which makes the encoder explicitly aware of the relative perturbation strength. Second, the similarity alignment loss l_{ali} (Equation (14)) aligns the cosine similarity S_{ij} from the encoder with the similarity $\hat{S}_{ij}$ computed by the augmentation-aware module. This alignment ensures that the diffusion distance calibrates the actual embedding similarity. Third, the relative similarity constraint l_{re} (Equation (15)) enforces a hierarchical structure $S_{1,2} > S_{2,3}$, which ensures that views with smaller diffusion gaps remain more similar than those with larger gaps. Through these complementary losses, the InfoNCE objective (Equation (11)) no longer treats all positive pairs identically; instead, it adaptively adjusts the contrastive pull based on the diffusion strength. This adaptive mechanism allows weak views to stay close to the anchor while strong views diverge. This encoding mechanism lies at the core of our diffusion-distance-calibrated contrastive learning.

First, we define the cosine similarity between two views $\mathbf{z}_i$ and $\mathbf{z}_j$ in the embedding space as:

$$S_{i,j} = \text{sim}(\mathbf{z}_i, \mathbf{z}_j) = \frac{\mathbf{z}_i \cdot \mathbf{z}_j}{\|\mathbf{z}_i\| \cdot \|\mathbf{z}_j\|} \quad (10)$$

The basic contrastive loss $\mathcal{L}_{\text{con}}$ uses the InfoNCE loss to optimize the distribution of positive and negative sample features:

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau) + \sum_{k=1}^K \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \quad (11)$$

where τ is the temperature coefficient; different granularity views generated from the same sample serve as positive pairs, and views of all other samples within the batch serve as negative pairs. Here, K denotes the number of negative samples within the batch (i.e., views of all other instances in the same batch).

The contrastive learning framework forces the model to identify and preserve shared semantic content across different perturbation levels through InfoNCE-based positive pair alignment. For each anchor sample, we treat its weak, medium, and strong views as positive pairs, while views from different samples serve as negative pairs. The contrastive loss $\mathcal{L}_{\text{con}}$ (Equation (11)) maximizes the agreement between positive pairs in the embedding space while minimizing agreement with negative pairs. Critically, because the positive pairs originate from views generated at substantially different diffusion timesteps—ranging from near-original (t_{weak}) to heavily corrupted (t_{strong})—the only way for the encoder to achieve high positive-pair similarity across all granularities is to extract semantic features that remain invariant across these diverse perturbations. Transient, perturbation-specific artifacts (e.g., noise patterns introduced by the diffusion process at a particular timestep)

do not share across views and thus cannot contribute to consistent positive alignment; the encoder instead focuses on stable, underlying fraud-relevant patterns that persist across the entire perturbation spectrum. This cross-view consistency objective thus biases the learned representations toward capturing intrinsic fraud signatures while discarding superficial variations.

We define the augmentation strength difference as the normalized timestep difference $\Delta t_{i,j} = |t_j - t_i|/T$. An augmentation-aware module $\varphi(\cdot)$ maps the encoder outputs to a new embedding space, yielding $\mathbf{h} = \varphi(\mathbf{z})$, based on which we compute the similarity:

$$\hat{S}_{i,j} = \text{sim}(\varphi(\mathbf{h}_i), \varphi(\mathbf{h}_j)) \quad (12)$$

We concatenate the augmented embedding features of two views to obtain a joint feature $\boldsymbol{\psi}_{i,j} = [\varphi(\mathbf{h}_i) \parallel \varphi(\mathbf{h}_j)]$, and a lightweight predictor $\text{Net}_d(\cdot)$ predicts the corresponding diffusion timestep difference:

$$l_d = \text{MSE}(\text{Net}_d(\boldsymbol{\psi}_{1,2}), \Delta t_{1,2}) + \text{MSE}(\text{Net}_d(\boldsymbol{\psi}_{2,3}), \Delta t_{2,3}) \quad (13)$$

Specifically, we feed the concatenated joint feature $\boldsymbol{\psi}_{ij}$ into a small predictor $\text{Net}_d(\cdot)$, which we implement as a two-layer MLP with ReLU activation, to regress the diffusion timestep difference Δt_{ij} .

This loss drives the augmentation-aware module to accurately capture the diffusion strength differences among views, which provides a reliable basis for calibrating the subsequent similarity computation.

To align the similarity produced by the encoder with that perceived by the augmentation-aware module, we introduce a similarity alignment loss:

$$l_{\text{ali}} = \text{MSE}(\hat{S}_{1,2}, S_{1,2}) + \text{MSE}(\hat{S}_{2,3}, S_{2,3}) \quad (14)$$

Additionally, we impose a relative similarity constraint to ensure that the similarity between the weak and medium augmented samples remains higher than that between the medium and strong augmented samples:

$$l_{\text{re}} = \text{ReLU}(S_{2,3} - S_{1,2}) \quad (15)$$

The above losses together form the similarity contrastive learning system. By leveraging the augmentation strength information encoded in the diffusion timesteps to achieve varying degrees of view contrasting, the model effectively distinguishes different levels of fraudulent features and fraud variants, thereby improving its generalization capability for fraud traffic detection.

3.6. PU Classification and Joint Optimization

In the PU scenario, only positive samples receive labels in the training data, while the remainder remain unlabeled. If we directly adopt a standard cross-entropy loss, the model misclassifies all unlabeled samples as negative, which introduces severe labeling noise. Therefore, we adopt a non-negative PU (nnPU) loss based on the positive class prior π_P for classification.

For a batch of training samples, let P be the set of labeled positive samples and U the set of unlabeled samples. We estimate the prior probability π_P from the proportion of known positive samples in the training set:

$$\pi_P = \frac{|P|}{|P| + |U|_{\text{all}}} \quad (16)$$

where $|U|_{\text{all}}$ denotes the total number of unlabeled samples.

In practice, this estimation relies on the SCAR (Selected Completely At Random) assumption [11], i.e., we assume the labeled positives are randomly sampled from the entire positive population. Although this assumption may not strictly hold in all advertising fraud scenarios, it represents a standard and widely adopted approximation in PU learning. For each dataset, we compute π_P directly from the training split, which reflects the natural positive rate reported in Table 1. We keep this prior fixed throughout training and decouple it from the classifier optimization. While inaccurate prior estimation can bias the risk estimator, the non-negative clipping in Equation (18) mitigates this concern by preventing negative loss values and stabilizing training. To handle the class imbalance between positive and negative samples, we adopt Focal Loss as the base classification loss, which down-weights easy samples and focuses on hard fraudulent examples. We formulate this loss as:

$$\ell(y, p) = -\alpha(1 - p_t)^\gamma \log p_t \quad (17)$$

where $\alpha \in (0, 1)$ is a balancing factor for class imbalance, $\gamma > 0$ is a focusing parameter that down-weights well-classified examples, and $p_t = p$ if $y = 1$ and $p_t = 1 - p$ otherwise.

Based on this base loss, we construct the non-negative PU loss as follows:

$$\mathcal{L}_{\text{PU}} = \pi_P \cdot \frac{1}{|P|} \sum_{i \in P} \ell(1, p_i) + \max\left(0, \frac{1}{|U|} \sum_{j \in U} \ell(0, p_j) - \pi_P \cdot \frac{1}{|P|} \sum_{i \in P} \ell(0, p_i)\right) \quad (18)$$

where $p_i = \text{softmax}(\text{logits})[1]$ denotes the predicted probability of the fraud class for sample i . The first term enforces positive sample classification, driving the model to correctly identify the labeled fraudulent samples. The second term serves as a bias correction for unlabeled samples: it offsets the fitting error caused by fraudulent samples mixed in the unlabeled set by subtracting the expected negative loss on positive samples, while the $\max(0, \cdot)$ function ensures the non-negativity of the loss. This formulation provides unbiased risk estimation for both positive and negative classes, which effectively addresses the label bias problem in PU scenarios.

To exploit the rich information of multi-granularity views, we compute the PU loss for the weak, medium, and strong views separately and fuse them via weighting:

$$\mathcal{L}_{\text{fl}} = w_1 \mathcal{L}_{\text{PU}}(\mathbf{z}_1, y) + w_2 \mathcal{L}_{\text{PU}}(\mathbf{z}_2, y) + w_3 \mathcal{L}_{\text{PU}}(\mathbf{z}_3, y) \quad (19)$$

where w_1, w_2, w_3 are the fusion weights for the weak, medium, and strong views, respectively, satisfying $\sum_{k=1}^3 w_k = 1$.

Finally, DiffVC integrates the above classification loss, all contrastive learning loss terms, and the diffusion reconstruction loss into an end-to-end joint objective:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{fl}} + \lambda_2 \mathcal{L}_{\text{con}} + \lambda_3 l_d + \lambda_4 l_{\text{ali}} + \lambda_5 l_{\text{re}} + \lambda_6 \mathcal{L}_{\text{diff}} \quad (20)$$

the hyperparameters λ_1 to λ_6 control the relative importance of each loss term in the overall objective.

The joint optimization strategy leverages the multi-granularity diffusion views through the coordinated action of all loss terms on the shared encoder. The classification loss $\mathcal{L}_{\text{fl}}$ applies the PU objective independently to each view, which ensures that the encoder produces fraud-discriminative representations across all perturbation levels. The contrastive loss $\mathcal{L}_{\text{con}}$ treats views from the same sample as positive pairs and views from different samples as negative pairs; by contrasting weak-vs-strong and medium-vs-strong pairs, it enforces the model to identify semantic content that persists across perturbations while suppressing instance-specific noise. The auxiliary losses ($l_d, l_{\text{ali}}, l_{\text{re}}$) im-

pose a structured embedding topology where similarity decreases with increasing diffusion distance. This topology encourages the encoder to organize features along a semantically meaningful axis that separates perturbation-sensitive patterns from more invariant fraud signatures. Finally, the diffusion reconstruction loss $\mathcal{L}_{\text{diff}}$ ensures that all generated views remain on the real data manifold and prevents the exploitation of unrealistic artifacts. Through this multi-task pressure, the optimization compels the shared encoder to capture a broad spectrum of fraud-relevant features across all granularities.

Table 1. The report summarizes four datasets.

Class	CTR	CPA	M	Martech
Positive	4917	1241	21,664	242,240
Negative	5083	3759	41,347	257,760
SUM	10,000	5000	63,011	500,000

During training, we jointly optimize the diffusion generation module, dual-gated Transformer encoder, projection network, and classifier from random initialization, which achieves deep fusion of self-supervised data augmentation, contrastive representation learning, and unbiased PU classification. Under the extreme label scarcity and complex sample distributions typical of real-world advertising scenarios, this joint optimization strategy leverages multi-granularity diffusion views to mine diverse fraudulent features, employs similarity contrastive learning to learn highly discriminative representations, and ultimately achieves accurate and robust ad fraud detection via the unbiased PU loss.

4. Experiments

This section provides a detailed description of our experimental design and results analysis. We conduct comparative experiments, parameter sensitivity analysis, ablation studies, generalization tests, and a distributional comparison between the diffusion-generated views and the original data.

4.1. Dataset Descriptions

To comprehensively evaluate the feasibility and applicability of our proposed model, we conduct experiments on four datasets: one proprietary dataset from a Chinese technology company and three public benchmark datasets (CPA Fraud Detection, CTR Prediction, and Martech Fraud Detection).

- **CPA Fraud Detection Dataset.** This dataset originates from a Cost-Per-Action online advertising scenario and specifically targets click fraud research. It contains 5000 click records with features covering user behavior, device information, and interaction patterns. We refer to it as the CPA dataset.
- **CTR Prediction Dataset.** This dataset originally targets click-through rate estimation tasks, where the goal is to predict the probability of a user clicking on an advertisement. It contains 10,000 advertising click logs, covering user attributes, ad content, and contextual information. Because of the high overlap in feature space between fraud detection and CTR estimation, we use it as an evaluation dataset. We refer to it as the CTR dataset.
- **Martech Fraud Detection Dataset.** This dataset is a large-scale public dataset for predicting ad click fraud. It consists of 500,000 advertising traffic records, where each sample includes 19 attribute fields and one fraud label. We refer to it as the Martech dataset.
- **Real-world Ad Log Dataset.** A Chinese technology company provides this dataset, which collects from real-world advertising traffic. It records key features such as

user behavior, geographic information, and device attributes. We utilize approximately 63,000 advertising traffic logs from 6 to 15 May 2020, consisting of roughly 41,000 normal samples and 22,000 fraudulent samples. We refer to it as the M dataset.

We present the statistical summaries of the four datasets in Table 1.

4.2. Data Preprocessing

To ensure a fair and reproducible evaluation, all datasets undergo the following preprocessing pipeline.

Missing value handling. We impute numerical features with missing values using the median of the corresponding feature column, and categorical features with the mode. We exclude features with an excessively high proportion of missing values from the analysis.

Numerical feature normalization. We transform all numerical features via quantile normalization to follow a Gaussian distribution, which mitigates the impact of skewed distributions and outliers. We fit this transformation solely on the training set and then apply the learned quantile mapping to the validation and test sets to prevent data leakage.

Categorical feature encoding. We first hash high-cardinality categorical features (e.g., device IDs, IP prefixes) to a fixed vocabulary size of 1000 to control embedding dimensionality, following the approach of TabTransformer [24]. We label-encode low-cardinality categorical features. We initialize all categorical embeddings as 32-dimensional vectors and train them jointly with the rest of the network.

Positive-Unlabeled split. Following the PU learning paradigm, we treat all labeled fraudulent samples as positives (P) and treat the remaining samples (including both legitimate traffic and undetected fraud) as unlabeled (U). We guarantee that the labeled positives form a random subset of all fraudulent samples in each dataset, with the proportion of labeled positives set to the natural label rate of that dataset (as reported in Table 1). We do not use negative labels during training.

Train/validation/test split. For each dataset, we randomly split the data into training (70%), validation (15%), and test (15%) sets, ensuring that the Positive-Unlabeled structure remains preserved in the training split. We use the validation set for early stopping and hyperparameter tuning. For the generalization test (Section 4.8), the test set consists entirely of temporally separated logs (16–18 May 2020) that we do not use during training or validation.

4.3. Experimental Setup

Parameter Configuration. We set the total number of diffusion timesteps to $T = 100$ and linearly increase the forward variance from 0.0001 to 0.02. We normalize numerical features via quantile transformation and embed categorical features into 32-dimensional continuous vectors. Based on the diffusion timestep, we define three views with distinct augmentation strengths: weak ($t = 0.15T$), medium ($t = 0.55T$), and strong ($t = T$). The dual-gated residual Transformer encoder comprises 4 stacked blocks with a hidden dimension of 64 and 4 attention heads. We train the model for 100 epochs using the AdamW optimizer with a learning rate of 0.001 and a batch size of 256. We set the loss weights as follows: $\lambda_1 = \lambda_3 = \lambda_4 = \lambda_5 = \lambda_6 = 1.0$, while $\lambda_2 = 0.5$ to prevent the contrastive loss from dominating the overall gradient.

Justification of hyperparameter choices. The diffusion timestep budget $T = 100$ follows the standard practice in tabular diffusion models [16,20], balancing reconstruction fidelity and computational efficiency. We empirically determine the timestep intervals for weak ($0.15T$), medium ($0.55T$), and strong (T) views to ensure (a) sufficient separation between granularities (non-overlapping intervals), and (b) coverage of the full noise spectrum, from minimal to near-total semantic corruption. We choose the Transformer encoder's

hidden dimension (64) and number of heads (4) to match the feature dimensionality after embedding ($d \approx 64$), which ensures that each attention head operates on a $d_h = 16$ dimensional subspace—a common practice for attention-based tabular models [24]. We select the learning rate 10^{-3} and batch size 256 based on the parameter sensitivity analysis (Section 4.5), where these values achieve the best trade-off between convergence stability and gradient signal strength across all datasets. By default, we set the loss weights λ_1 through λ_6 to 1.0 to assign equal importance to all objectives, except for $\lambda_2 = 0.5$, which reduces the contrastive loss weight to prevent it from dominating the overall gradient, as recommended by prior contrastive learning studies [28].

PU learning setup. In all experiments, we follow the standard PU learning paradigm where we label only positive samples. For each dataset, we use the entire training set with its natural positive rate, treating all labeled fraud samples as positives (P) and all remaining samples as unlabeled (U). The unlabeled set contains a mixture of legitimate traffic and undetected fraud, consistent with real-world deployment scenarios. We emphasize that we do not use explicit negative labels during training; the classifier learns to distinguish fraud from non-fraud solely through the nnPU loss (Equation (18)). We estimate the positive class prior π_P as described in Section 3.6. We adapt all baseline methods to this PU setting following the same protocol: for methods that require negative labels, we treat unlabeled samples as negative for training, which serves as a strong but practical baseline. This setup ensures a fair comparison under realistic label-scarcity conditions.

Experimental environment. All experiments are conducted on a machine running Windows 11, equipped with an NVIDIA GeForce RTX 4060 Laptop GPU, an Intel(R) Core(TM) i7-14650HX CPU, and 32 GB of RAM.

To mitigate random fluctuations, all experiments are repeated independently five times with different random initializations. We report the mean and standard deviation for all metrics. All baseline methods follow the same experimental protocol.

4.4. Comparative Experiments

To verify the effectiveness of DiffVC, we compare it against five representative methods: XGBoost, Multi-Layer Perceptron (MLP), TabTransformer (which specifically designs for tabular data) [24], TrafficFormer (which tailors for fraud traffic detection) [43], and DeepFM (a deep learning approach for ad fraud detection) [45]. We tune all baseline methods according to their original papers or publicly available implementations to ensure optimal performance. Table 2 reports the mean and standard deviation of AUC, F1-score, and Recall over five independent runs.

Specifically, on the CTR dataset, DiffVC achieves an AUC of 0.9227 (ranking second to TabTransformer's 0.9427) but attains the highest Recall of 0.9029, which outperforms the best baseline by 4.1% (from 0.8672). On CPA, it attains a near-perfect AUC of 0.9941, surpassing TrafficFormer (0.9691) by 2.5%, and achieves an F1-score of 0.9214, which exceeds TrafficFormer's 0.8522 by 7.8%. On M, DiffVC secures the highest AUC of 0.9425, ahead of TabTransformer (0.9387) and TrafficFormer (0.8959) by 4.7%; its F1-score of 0.8298 is competitive and surpasses TabTransformer (0.8180) and DeepFM (0.7979). On MarTech, it dominates all metrics with an AUC of 0.9167, an F1-score of 0.8444, and a Recall of 0.8720, surpassing the second-best method (TrafficFormer) by 1.1%, 0.9%, and 0.5%, respectively. Overall, DiffVC achieves the highest AUC on CPA, M, and MarTech, the highest F1-score on CPA and MarTech, and the best Recall on both CTR and CPA, consistently outperforming the state-of-the-art baselines across diverse real-world tasks.

Table 2. Comparative experimental results between DiffVC and the baseline models.

Dataset	Metric	XGBoost	MLP	TabTransformer	TrafficFormer	DeepFM	DiffVC (Our)
CTR	AUC	0.8016 ± 0.017	0.6282 ± 0.014	0.9427 ± 0.002	0.8378 ± 0.010	0.9052 ± 0.007	0.9227 ± 0.004
	F1-Score	0.6428 ± 0.020	0.5788 ± 0.022	0.8684 ± 0.009	0.7511 ± 0.015	0.8330 ± 0.009	0.8490 ± 0.005
	Recall	0.5192 ± 0.022	0.5751 ± 0.053	0.8672 ± 0.025	0.7352 ± 0.031	0.8294 ± 0.018	0.9029 ± 0.025
CPA	AUC	0.5203 ± 0.005	0.8614 ± 0.041	0.9520 ± 0.004	0.9691 ± 0.021	0.9066 ± 0.014	0.9941 ± 0.002
	F1-Score	0.3831 ± 0.105	0.5600 ± 0.090	0.7834 ± 0.020	0.8522 ± 0.063	0.7207 ± 0.030	0.9214 ± 0.013
	Recall	0.5725 ± 0.395	0.9108 ± 0.099	0.8505 ± 0.101	0.9226 ± 0.092	0.7882 ± 0.037	0.9258 ± 0.023
M	AUC	0.7568 ± 0.020	0.8866 ± 0.010	0.9387 ± 0.006	0.8959 ± 0.002	0.8361 ± 0.015	0.9425 ± 0.006
	F1-Score	0.5562 ± 0.030	0.8733 ± 0.006	0.8180 ± 0.008	0.8893 ± 0.002	0.7979 ± 0.006	0.8298 ± 0.011
	Recall	0.4197 ± 0.022	0.9315 ± 0.012	0.8782 ± 0.014	0.9407 ± 0.006	0.7356 ± 0.012	0.8310 ± 0.019
MarTech	AUC	0.8652 ± 0.002	0.9065 ± 0.002	0.9135 ± 0.002	0.9066 ± 0.001	0.8886 ± 0.006	0.9167 ± 0.005
	F1-Score	0.7510 ± 0.011	0.8278 ± 0.003	0.8368 ± 0.003	0.8371 ± 0.001	0.8217 ± 0.006	0.8444 ± 0.006
	Recall	0.6953 ± 0.027	0.8006 ± 0.005	0.8106 ± 0.010	0.8681 ± 0.002	0.8142 ± 0.011	0.8720 ± 0.006

Bold values indicate the best performance.

Statistical significance analysis. To rigorously assess whether the performance improvements of DiffVC over the baseline methods are statistically significant, we perform a linear mixed model (LMM) analysis across all six methods (XGBoost, MLP, TabTransformer, TrafficFormer, DeepFM, and DiffVC) on the four datasets. For each evaluation metric (AUC, F1-score, and Recall), we fit a separate LMM with method as a fixed effect and dataset as a random intercept to account for dataset-specific variability. Table 3 summarizes the likelihood ratio (LR) tests for the fixed effect of method. The results yield highly significant p -values for all three metrics (AUC: $\chi^2 = 52.11$, $df = 5$, $p = 5.12 \times 10^{-10}$; F1-score: $\chi^2 = 70.07$, $df = 5$, $p = 9.90 \times 10^{-14}$; Recall: $\chi^2 = 54.21$, $df = 5$, $p = 1.90 \times 10^{-10}$), which indicates that the observed differences among methods are extremely unlikely to result from random chance. Post hoc pairwise comparisons using Bonferroni correction further confirm that DiffVC significantly outperforms each baseline method across all datasets ($p < 0.05$ for all pairwise comparisons). These results provide strong statistical evidence that the performance gains of DiffVC are genuine and not attributable to random fluctuations.

Table 3. Likelihood ratio test results from the linear mixed model analysis comparing all six methods across four datasets.

Metric	Likelihood Ratio (χ^2)	Degrees of Freedom (df)	p -Value
AUC	52.11	5	5.12×10^{-10}
F1-Score	70.07	5	9.90×10^{-14}
Recall	54.21	5	1.90×10^{-10}

Computational cost analysis. To evaluate the practical deployability of DiffVC, we measure the inference latency of all methods on the CTR dataset (10,000 samples) using a single NVIDIA RTX 4060 GPU with a batch size of 256 samples. Table 4 shows that DiffVC achieves an inference time of 1.59 s per 256 samples, which outperforms TrafficFormer (3.37 s) but falls moderately behind TabTransformer (0.68 s), MLP (0.42 s), DeepFM (0.33 s), and XGBoost (0.23 s). This aligns with our expectations, as DiffVC employs a more expressive dual-gated Transformer encoder that captures higher-order feature interactions at the cost of moderate computational overhead. Crucially, the inference latency of 1.59 s per 256 samples comfortably falls within the acceptable range for most real-time ad fraud detection pipelines, where batch processing is common. Moreover, we require the diffusion module only during training and entirely disable it during inference, meaning that the inference cost of DiffVC depends solely on the lightweight Transformer encoder forward

pass. For streaming scenarios that demand ultra-low latency, we can further optimize the model via quantization or pruning without significant performance degradation, as prior work on efficient Transformer deployment demonstrates.

Table 4. Inference latency comparison of all methods on the CTR dataset (batch size = 256).

Method	Inference Time(s) per 256 Samples
XGBoost	0.23
MLP	0.42
TabTransformer	0.68
TrafficFormer	3.37
DeepFM	0.33
DiffVC (ours)	1.59

4.5. Parameter Sensitivity Analysis

To investigate the sensitivity of DiffVC to key hyperparameters, we conduct independent experiments on the CTR dataset by varying the number of training epochs, learning rate, and batch size. In each experiment, we change only the target parameter while keeping all others at their default settings. Figure 4a–c visualize the results.

The results show that the model’s AUC and F1-score fluctuate minimally as the number of training epochs varies from 80 to 300, which indicates that the model converges early and remains relatively insensitive to training duration. The learning rate has the most pronounced impact: performance stays stable as the learning rate decreases from a higher value, but degrades sharply once it falls below a critical threshold. This finding suggests that the joint optimization of diffusion and contrastive learning requires a sufficiently high learning rate to sustain effective gradient updates.

This sensitivity arises from the distinct gradient characteristics of the two learning objectives. The diffusion reconstruction loss $\mathcal{L}_{\text{diff}}$ yields relatively smooth and stable gradients, as it operates as a regression objective over the denoising network. In contrast, the InfoNCE-based contrastive loss $\mathcal{L}_{\text{con}}$ produces sparser and more competitive gradients, because they rely on pairwise similarities across the entire batch. The combined gradients from both objectives update the shared encoder parameters. In the early training stages, contrastive gradients tend to have smaller magnitudes than diffusion gradients due to the stochasticity of negative sampling and the high-dimensional embedding space. A sufficiently high learning rate is therefore necessary to amplify these contrastive signals, enabling them to compete with and balance the diffusion reconstruction gradients. Below a critical threshold, the contrastive objective effectively suppresses, leading to the sharp performance drop that Figure 4c observes.

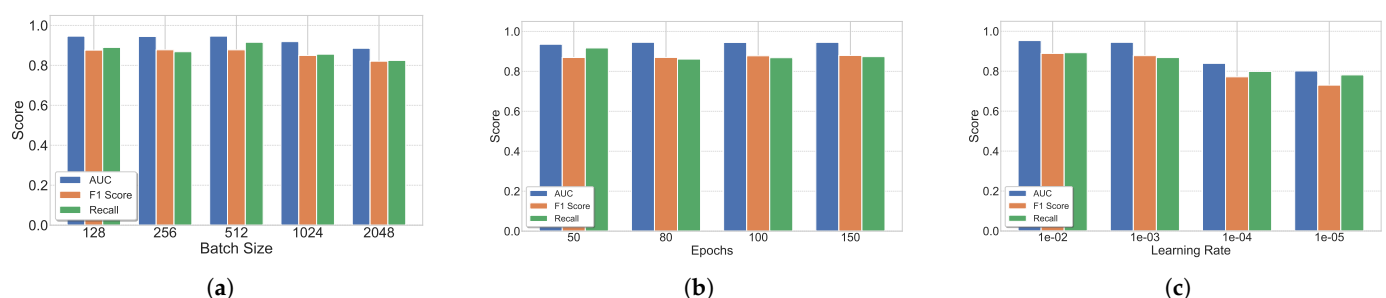


Figure 4. Parameter sensitivity experiment results using different batch sizes, training epochs, and learning rates. (a) Experimental results with different batch sizes. (b) Experimental results with different training epochs. (c) Experimental results with different learning rates.

The batch size exhibits a peak-shaped trend, where a moderate batch size achieves the best balance between precision and recall. Excessively small batches lead to unstable gradient estimation, while excessively large batches weaken the discriminative signal in contrastive learning.

In summary, DiffVC performs robustly under default settings and remains largely insensitive to variations in the number of epochs and batch size within a reasonable range; performance degrades only at extremely low learning rates.

4.6. Ablation Study

To quantify the contributions of the three core modules, we sequentially remove the diffusion-driven multi-view generation module (DDPM), replace the dual-gated residual Transformer encoder with a standard MLP, and remove the similarity contrastive learning component (Vari-Con) while retaining only the basic InfoNCE loss.

Specifically, the full similarity contrastive learning system consists of four components: the InfoNCE loss $\mathcal{L}_{\text{con}}$ (Equation (11)) that pulls positive pairs closer and pushes negative pairs apart, the distance prediction loss l_d (Equation (13)) that captures diffusion strength differences, the similarity alignment loss l_{ali} (Equation (14)) that aligns encoder and augmentation-aware similarities, and the relative similarity constraint l_{re} (Equation (15)) that enforces hierarchical view ordering. In the ablation setting, we remove the three diffusion-calibrated components (l_d , l_{ali} , and l_{re}) while retaining only the standard InfoNCE loss $\mathcal{L}_{\text{con}}$. This variant treats all positive pairs identically regardless of their augmentation strengths, reverting to the fixed similarity assumption of conventional contrastive learning. The performance gap between this ablated variant and the full model quantifies the contribution of the diffusion-distance-calibrated similarity to learning robust hierarchical representations.

Table 5 presents the ablation results on the four datasets: CTR, CPA, M, and Martech.

Table 5. Ablation experiment results of removing the diffusion model (DDPM), the Transformer encoder, and the similarity contrastive learning (Simi-Con) module, respectively.

DDPM	Transformer	Simi-Con	CTR	CPA	M	MarTech
✓			0.7971 ± 0.019	0.9166 ± 0.014	0.8226 ± 0.011	0.8259 ± 0.007
✓	✓		0.8594 ± 0.005	0.9585 ± 0.031	0.9234 ± 0.004	0.8769 ± 0.005
✓		✓	0.8155 ± 0.014	0.9510 ± 0.012	0.8842 ± 0.033	0.8433 ± 0.005
	✓	✓	0.9048 ± 0.005	0.9468 ± 0.010	0.9020 ± 0.008	0.8698 ± 0.003
✓	✓	✓	0.9227 ± 0.004	0.9941 ± 0.002	0.9425 ± 0.006	0.9167 ± 0.005

Bold values indicate the best performance. ✓ indicates that the corresponding module is adopted.

The full model consistently achieves the best AUC on all datasets, which confirms the effectiveness of the synergy among the three modules. When training an MLP solely on diffusion-augmented data, performance remains the lowest, indicating that data augmentation alone cannot replace an effective representation learning framework. Among the three modules, removing the Transformer encoder causes the most severe performance degradation; across multiple datasets, the AUC drop exceeds 0.05, which confirms its central role in modeling complex feature interactions in tabular data. Removing the similarity contrastive learning component also leads to an overall performance decline, demonstrating that contrastive learning is crucial for learning discriminative feature representations. Removing the diffusion module also results in a consistent and significant performance loss, which shows that the diffusion-augmented samples provide indispensable multi-granularity training signals for the encoder. In summary, the synergistic design of the three modules forms the foundation that enables DiffVC to achieve optimal performance; removing any single

module leads to a non-negligible performance drop, which validates the rationality and necessity of the overall architecture.

4.7. Distribution Comparison Between Data Generated by the Diffusion Model and Original Data

To verify the quality of the augmented samples, we visualize the distribution of numerical and categorical features for the weak, medium, and strong views. As Figures 5 and 6 show, the three augmented views align closely with the original data in value ranges and category frequencies. No anomalous samples appear outside the original distribution range, which indicates that the diffusion reconstruction process strictly respects the joint distribution of the data.

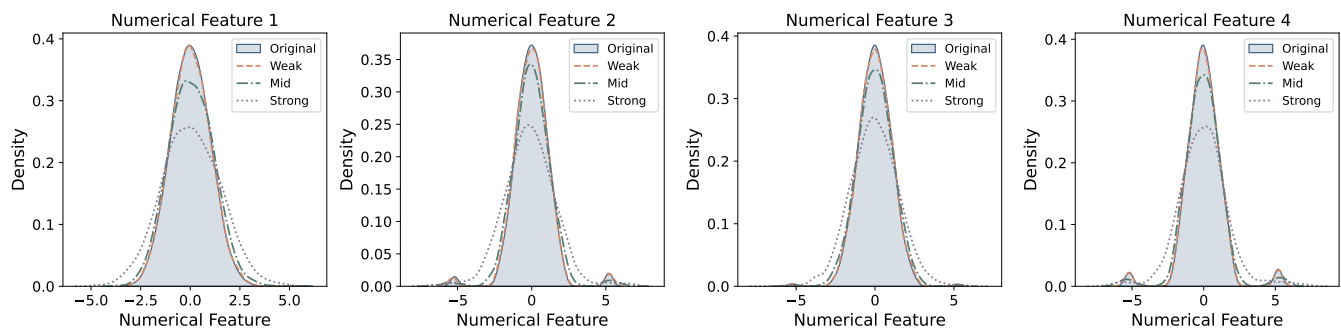


Figure 5. Distribution comparison result plot of numerical features for the three views generated by the diffusion model and the original data.

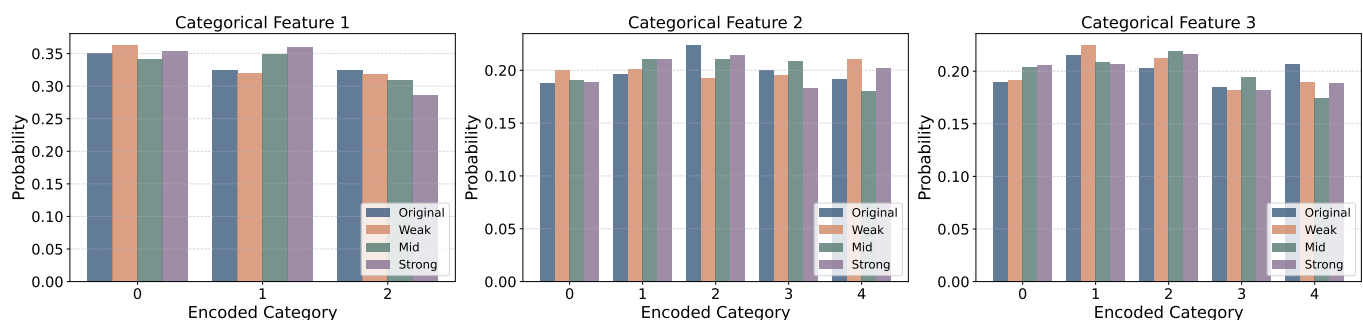


Figure 6. Distribution comparison result plot of categorical features for the three views generated by the diffusion model and the original data.

In particular, the medium and strong views exhibit slight distributional shifts in localized regions: the peak densities of some numerical features decrease slightly, the proportion of samples in edge intervals increases marginally, and certain categorical probabilities show minor fluctuations. This behavior aligns with the increased noise intensity at higher diffusion timesteps. This “overall fidelity, local augmentation” characteristic ensures that the augmented samples remain faithful to the real data manifold while providing a reasonable expansion of the data distribution for training. Such smooth expansion helps the model learn boundary cases, which effectively enhances its generalization ability.

4.8. Generalization Verification

To further evaluate the generalization capability of DiffVC in real-world deployment, we construct a test set using three consecutive days of new advertising traffic logs (16–18 May 2020) from a Chinese technology company. We completely withhold this dataset from the training and validation stages; it contains approximately 23,000 samples and exhibits a natural temporal shift relative to the training data. Table 6 summarizes the performance of all models.

Table 6. Generalization test experiment results using real-world ad log data.

Model	AUC	F1 Score	Recall
XGBoost	0.7193	0.5292	0.4028
MLP	0.7414	0.5177	0.3554
TabTransformer	0.8668	0.7046	0.642
TrafficFormer	0.8595	0.7084	0.7799
DeepFM	0.7819	0.6518	0.7054
DiffVC (our)	0.931	0.8071	0.836

Bold values indicate the best performance.

The experimental results demonstrate that DiffVC attains an AUC of 0.931, an F1-score of 0.8071, and a Recall of 0.836, striking a superior balance across all metrics. In contrast, while TabTransformer and TrafficFormer maintain AUCs around 0.86, TabTransformer yields a low Recall of 0.64 (which poses a high risk of missed detections), and TrafficFormer achieves a Recall of 0.78, still falling short of our method. DeepFM exhibits insufficient discriminative power (AUC 0.78), XGBoost records the lowest AUC, and the MLP's Recall plummets to a mere 0.36, rendering it unable to capture fraudulent samples effectively. The robust generalization of DiffVC across time periods stems from two key factors: the diffusion-driven multi-granularity augmentation suppresses temporal overfitting, while the similarity contrastive learning guides the encoder to extract time-invariant fraud patterns.

Specifically, the similarity contrastive learning mechanism achieves this through two main aspects. First, the InfoNCE-based contrastive loss (Equation (11)) encourages the encoder to maximize agreement between different augmented views of the same sample. Since these views are generated with varying diffusion timesteps, the model identifies and preserves shared semantic content that persists across different perturbation levels while discarding superficial variations such as timestamp-specific artifacts or transient changes in request patterns. This cross-view consistency objective biases the encoder toward learning stable, time-invariant representations that are robust to temporal distribution shifts.

Second, the diffusion-distance-calibrated similarity mechanism (Equations (13) and (14)) imposes a structured topology on the embedding space, where views with smaller diffusion gaps (e.g., weak vs. medium) pull closer together than those with larger gaps (e.g., weak vs. strong). This hierarchical constraint prevents the model from over-relying on any single augmentation level and encourages the encoder to capture features that remain consistent across the entire granularity spectrum. Such features—including anomalous device fingerprint patterns, abnormal click frequency distributions, or inconsistent geolocation sequences—are intrinsic to fraud behaviors rather than tied to specific time periods. Consequently, the learned representations generalize effectively to unseen traffic logs collected days after training, as Table 6 confirms.

5. Discussion

In this study, we propose DiffVC, a novel framework that integrates diffusion-based multi-granularity data augmentation, diffusion-distance-calibrated similarity contrastive learning, and a dual-gated residual Transformer encoder for advertising fraud detection under the Positive-Unlabeled (PU) learning paradigm. Our experimental results consistently demonstrate that DiffVC outperforms strong baselines across four real-world and public datasets. In this section, we interpret these findings in the context of prior research, discuss their theoretical and practical implications, and highlight the limitations of our approach.

Comparison with existing fraud detection methods. Traditional fraud detection models, such as XGBoost and MLP, rely heavily on hand-crafted features and often struggle to capture high-order nonlinear interactions among heterogeneous tabular attributes. More recent deep learning approaches, including TabTransformer and TrafficFormer, improve

representation learning through self-attention or pre-training strategies. Nevertheless, they remain vulnerable to semantic drift when encountering out-of-distribution fraudulent samples, especially under the extreme label scarcity inherent in real-world advertising systems. DiffVC bridges this gap by leveraging a denoising diffusion probabilistic model (DDPM) to generate semantically meaningful augmented views at multiple granularities. Unlike conventional data augmentation methods (e.g., SMOTE or random noise injection), the diffusion-based generation preserves the joint distribution of categorical and numerical features through decoupled Gaussian and multinomial diffusion processes, which ensures that augmented samples stay on the real data manifold. This characteristic aligns with recent advances in tabular diffusion models [16,20], which emphasize the importance of preserving feature dependencies and business-logic constraints when generating heterogeneous tabular data.

The role of diffusion-distance-calibrated similarity contrastive learning. A key innovation of our work is the diffusion-distance-calibrated contrastive loss, which relaxes the fixed-similarity assumption inherent in standard InfoNCE. As our ablation study (Table 5) shows, removing the variable similarity module causes a substantial performance drop, especially on the large-scale Martech dataset (a decrease of over 7 percentage points in AUC). This result supports the view that forcing identical semantic alignment between weak and strong augmented views leads to representation collapse or overfitting to spurious correlations [28,46]. In contrast, our dynamic similarity weighting—which leverages the diffusion timestep as a natural measure of augmentation strength—enables the model to learn hierarchical representations: it preserves fine-grained instance-level consistency from weak views while encouraging broader exploration of the underlying fraud manifold through strong views. The joint optimization of the similarity alignment loss, distance prediction loss, and relative similarity constraint ensures that the encoder forms a structured embedding topology where larger diffusion gaps correspond to lower view similarity. This mechanism proves particularly beneficial for PU-based fraud detection, where fraudulent patterns evolve rapidly and models must generalize across temporal shifts while learning from only a limited set of labeled positives.

Effectiveness of the dual-gated residual Transformer encoder. Replacing the dual-gated residual Transformer with a standard MLP causes the most severe performance degradation across all datasets, which confirms that tabular feature interactions in advertising traffic are inherently nonlinear and global. The proposed encoder, equipped with parameterized dual-gating mechanisms and squeeze-and-excitation blocks [44], adaptively fuses original and transformed features through residual connections while mitigating the degradation problem in deep networks. The first gating parameter controls the contribution of the multi-head self-attention output, which ensures flexible information flow; the second gating parameter balances identity mapping and channel-recalibrated features, which prevents the loss of low-level feature information during forward propagation. This design choice aligns with recent observations that simple MLPs often underfit or overfit tabular data [34], whereas attention-based models with adaptive gating capture cross-feature dependencies more effectively. The channel recalibration mechanism, which employs feature compression and expansion through dimensionality reduction, performs adaptive screening of different feature dimensions—a critical capability when dealing with sparse categorical features such as device IDs, geographic attributes, and ad slot IDs in advertising logs.

Practical implications for the advertising ecosystem. From a practical standpoint, DiffVC offers several deployable advantages. First, the cross-period generalization test (Section 4.7) demonstrates robust detection performance on unseen traffic logs collected three days after the training period, which indicates that the model adapts to evolving fraud tactics without frequent retraining. The diffusion-generated multi-granularity views

effectively suppress temporal overfitting, while the similarity contrastive learning guides the encoder to capture time-invariant fraud patterns. Second, the multi-granularity views generated by the diffusion model serve as a standalone data augmentation tool for other fraud detection systems, providing a plug-and-play solution for imbalanced tabular data under PU scenarios. Third, the non-negative PU loss with Focal Loss as the base classification loss effectively addresses the labeling bias problem and enables unbiased risk estimation for positive and negative class classification. Fourth, the variable similarity contrastive learning framework extends beyond advertising fraud; we can readily extend it to other tabular anomaly detection domains, such as financial transaction fraud, network intrusion detection, or healthcare claim auditing, where labeled anomalies are scarce and unlabeled data contain latent fraud instances.

It should be noted, however, that these conclusions are drawn from experiments on four datasets, with a temporal gap of only three days in the cross-period generalization test. While the results are encouraging, they represent an initial step toward understanding the model's generalization capabilities. Further validation on larger-scale deployments and longer temporal horizons would be necessary to fully establish the practical applicability of DiffVC in diverse advertising environments.

Limitations. Despite its superior performance, DiffVC has several limitations that also point to future directions. First, the sequential denoising over 100 timesteps and the iterative generation of three views per sample introduce considerable training overhead, which may limit rapid iteration and real-time deployment. Second, the loss weights (λ_1 to λ_6) and timestep intervals require careful tuning, as suboptimal settings can affect the fidelity-diversity balance. Third, the current design uses only three discrete granularities; an adaptive sampling strategy based on sample difficulty can further improve efficiency. Fourth, while the binary PU loss effectively handles Positive-Unlabeled scenarios, multi-class fraud settings (e.g., distinguishing click farms, attribution hijacking, and malicious script injection) remain unexplored. Addressing these limitations—through lightweight diffusion architectures, learnable schedulers, multi-class extension, and robust prior estimation—constitutes our primary future work.

6. Conclusions

This paper presents DiffVC, a multi-granularity diffusion-based similarity contrastive learning framework for advertising fraud detection under the Positive-Unlabeled (PU) paradigm. The framework integrates three core innovations to enable robust fraud detection under extreme label scarcity: (1) a diffusion-driven multi-granularity augmentation strategy that generates weak, medium, and strong views while preserving the original data distribution; (2) a dual-gated residual Transformer encoder for capturing high-order nonlinear feature interactions; and (3) a diffusion-distance-calibrated similarity contrastive loss that dynamically adjusts similarity constraints according to perturbation strength.

Experimental results on four real-world datasets confirm that DiffVC meets this objective. DiffVC consistently achieves the highest AUC, F1-score, and Recall among all competing methods, with up to 7.8% improvement in F1-score over the best baselines. Ablation studies verify the indispensable contribution of each module. Crucially, cross-period generalization tests on unseen traffic logs (collected three days after training) demonstrate robust performance (AUC 0.931), suggesting its potential for practical deployability in dynamic advertising environments, though we require further validation on larger-scale and longer-term deployments. Future work focuses on lightweight diffusion architectures, adaptive granularity scheduling, and multi-class fraud identification.

Author Contributions: Conceptualization, L.W., W.Y. and L.Y.; methodology, L.W., W.Y. and L.Y.; software, W.Y. and L.Y.; validation, Y.M., X.M. and L.Y.; formal analysis, W.Y. and L.Y.; investigation,

W.Y.; resources, L.Y. and Y.M.; data curation, L.W. and W.Y.; writing—original draft preparation, L.W., W.Y. and L.Y.; writing—review and editing, L.W., Y.M. and L.Y.; visualization, W.Y. and X.M.; supervision, L.W., Y.M., X.M. and L.Y.; project administration, L.W. and L.Y.; funding acquisition, L.W., X.M. and L.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant No. 61972241, 62402309, 62562012, the Natural Science Foundation of Shanghai under Grant No. 22ZR1427100, the Soft Science Project of Shanghai under Grant No. 25692107400.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The CPA Fraud Detection Dataset is available on Kaggle platform <https://www.kaggle.com/datasets/ziya07/fraud-detection-dataset> (accessed on 16 August 2025); the CTR Prediction Dataset is available on GitHub <https://github.com/thaychansy/click-through-rate-prediction/tree/main> (accessed on 11 November 2024); the Martech Fraud Detection Dataset is available on GitHub <https://github.com/wxydaydayup/-MarTech-Challenge-10-> (accessed on 2 November 2022). The advertising log dataset provided by a Chinese tech company is private data and can be requested from the author if needed.

Acknowledgments: The author would like to express gratitude to the anonymous reviewers for their valuable comments and suggestions. This work was supported by the National Natural Science Foundation of China under Grant No. 61972241, 62402309, 62562012, the Natural Science Foundation of Shanghai under Grant No. 22ZR1427100, the Soft Science Project of Shanghai under Grant No. 25692107400.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Quantifying the Cost of Ad Fraud: 2023–2028. Technical Report, Juniper Research. 2023. Available online: <https://fraudblocker.com/data/ad-fraud-data-facts> (accessed on 26 October 2023).
2. Choi, H.; Mela, C.F.; Balseiro, S.R.; Leary, A. Online display advertising markets: A literature review and future directions. *Inf. Syst. Res.* **2020**, *31*, 556–575. [CrossRef]
3. Sun, S.; Yu, L.; Zhang, X.; Xue, M.; Zhou, R.; Zhu, H.; Hao, S.; Lin, X. Understanding and detecting mobile ad fraud through the lens of invalid traffic. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, 15–19 November 2021; pp. 287–303. [CrossRef]
4. Oentaryo, R.; Lim, E.P.; Finegold, M.; Lo, D.; Zhu, F.; Phua, C.; Cheu, E.Y.; Yap, G.E.; Sim, K.; Nguyen, M.N.; et al. Detecting click fraud in online advertising: A data mining approach. *J. Mach. Learn. Res.* **2014**, *15*, 99–140.
5. Springborn, K.; Barford, P. Impression fraud in on-line advertising via {Pay-Per-View} networks. In Proceedings of the 22nd USENIX Security Symposium (USENIX Security 13), Washington, DC, USA, 14–16 August 2013; pp. 211–226.
6. Zhu, T.; Shou, C.; Huang, Z.; Chen, G.; Zhang, X.; Meng, Y.; Hao, S.; Zhu, H. Unveiling collusion-based ad attribution laundering fraud: Detection, analysis, and security implications. In Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security, Salt Lake City, UT, USA, 14–18 October 2024; pp. 2963–2977. [CrossRef]
7. Nagaraja, S.; Shah, R. Clicktok: Click fraud detection using traffic analysis. In Proceedings of the 12th Conference on Security and Privacy in Wireless and Mobile Networks, Miami, FL, USA, 15–17 May 2019; pp. 105–116. [CrossRef]
8. Kirkwood, B.; Vanamala, M.; Seliya, N. Click fraud detection of online advertising using machine learning algorithms. In *Proceedings of the 2024 IEEE International Conference on Electro Information Technology (eIT)*; IEEE: Piscataway, NJ, USA, 2024; pp. 586–590. Available online: <https://ieeexplore.ieee.org/abstract/document/10609899> (accessed on 31 July 2024).
9. Mahesh, V.B.; Chandra, K.V.S.; Babu, L.S.P.; Sowjanya, V.A.; Mohammed, M. Clicking fraud detection for online advertising using machine learning. In *Proceedings of the 2024 4th International Conference on Intelligent Technologies (CONIT)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6. Available online: <https://ieeexplore.ieee.org/abstract/document/10627189> (accessed on 15 August 2024).
10. Cho, G.; Cho, J.; Song, Y.; Kim, H. An empirical study of click fraud in mobile advertising networks. In *Proceedings of the 2015 10th International Conference on Availability, Reliability and Security*; IEEE: Piscataway, NJ, USA, 2015; pp. 382–388. Available online: <https://ieeexplore.ieee.org/document/7299940> (accessed on 19 October 2015).

11. Elkan, C.; Noto, K. Learning classifiers from only positive and unlabeled data. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NE, USA, 24–27 August 2008; pp. 213–220. [CrossRef]
12. Vasquez, C.O.; De Weerd, J.; vanden Broucke, S. The hidden cost of fraud: An instance-dependent cost-sensitive approach for positive and unlabeled learning. In *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications*; PMLR: Cambridge, MA, USA, 2022; pp. 53–67. Available online: <https://proceedings.mlr.press/v183/vasquez22a.html> (accessed on 23 September 2022).
13. Gong, C.; Wang, Q.; Liu, T.; Han, B.; You, J.; Yang, J.; Tao, D. Instance-Dependent Positive and Unlabeled Learning with Labeling Bias Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 4163–4177. [CrossRef] [PubMed]
14. Acharya, A.; Sanghavi, S.; Jing, L.; Bhushanam, B.; Choudhary, D.; Rabbat, M.; Dhillon, I. Positive unlabeled contrastive learning. *arXiv* **2022**, arXiv:2206.01206. [CrossRef]
15. Sisodia, D.; Sisodia, D.S. A hybrid data-level sampling approach in learning from skewed user-click data for click fraud detection in online advertising. *Expert Syst.* **2023**, *40*, e13147. [CrossRef]
16. Kotelnikov, A.; Baranchuk, D.; Rubachev, I.; Babenko, A. Tabddpm: Modelling tabular data with diffusion models. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2023; pp. 17564–17579. [CrossRef]
17. Song, K.; Liu, C.; Jiang, D. A positive-unlabeled learning approach for industrial anomaly detection based on self-adaptive training. *Neurocomputing* **2025**, *647*, 130488. [CrossRef]
18. Takahashi, H.; Iwata, T.; Kumagai, A.; Yamanaka, Y. Deep positive-unlabeled anomaly detection for contaminated unlabeled data. *arXiv* **2024**, arXiv:2405.18929. [CrossRef]
19. Herurkar, D.; Hees, J.; Tzvetkov, V.; Dengel, A. Tabular Data Adapters: Improving Outlier Detection for Unlabeled Private Data. *arXiv* **2025**, arXiv:2504.20862.
20. Zhang, H.; Zhang, J.; Srinivasan, B.; Shen, Z.; Qin, X.; Faloutsos, C.; Rangwala, H.; Karypis, G. Mixed-type tabular data synthesis with score-based diffusion in latent space. *arXiv* **2023**, arXiv:2310.09656. [CrossRef]
21. Shi, J.; Xu, M.; Hua, H.; Zhang, H.; Ermon, S.; Leskovec, J. TabDiff: A unified diffusion model for multi-modal tabular data generation. In Proceedings of the NeurIPS 2024 Third Table Representation Learning Workshop, Vancouver, BC, Canada, 14 December 2024. Available online: <https://neurips.cc/virtual/2024/103158> (accessed on 10 October 2024).
22. Wu, Y.; Xu, Y.; Li, J. Fraudulent traffic detection in online advertising with bipartite graph propagation algorithm. *Expert Syst. Appl.* **2021**, *185*, 115573. [CrossRef]
23. Somepalli, G.; Goldblum, M.; Schwarzschild, A.; Bruss, C.B.; Goldstein, T. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv* **2021**, arXiv:2106.01342. [CrossRef]
24. Huang, X.; Khetan, A.; Cvitkovic, M.; Karnin, Z. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv* **2020**, arXiv:2012.06678. [CrossRef]
25. Li, W.; Zhong, Q.; Zhao, Q.; Zhang, H.; Meng, X. Multimodal and contrastive learning for click fraud detection. *arXiv* **2021**, arXiv:2105.03567. [CrossRef]
26. Zhu, T.; Meng, Y.; Hu, H.; Zhang, X.; Xue, M.; Zhu, H. Dissecting click fraud autonomy in the wild. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*; ACM: New York, NY, USA, 2021; pp. 271–286. [CrossRef]
27. Yue, Z.; Wang, J.; Sun, Q.; Ji, L.; Chang, E.; Zhang, H. Exploring diffusion time-steps for unsupervised representation learning. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024; Volume 2024, pp. 8209–8219. [CrossRef]
28. Cui, H.; Chen, S.; Li, J.; Yang, J. Enhancing Contrastive Learning with Variable Similarity. In Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems, San Diego, CA, USA, 2–7 December 2025. Available online: <https://neurips.cc/virtual/2025/loc/san-diego/poster/116034> (accessed on 19 September 2025).
29. Zhu, F.; Zhang, C.; Zheng, Z.; Al Otaibi, S. Click fraud detection of online advertising—LSH based tensor recovery mechanism. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 9747–9754. [CrossRef]
30. Dai, B.; Wang, L.; Zhao, X.; Guo, Y.; Obaidat, M.S. GemmaWithLoRA: A New Approach to Click Fraud Detection. In *Proceedings of the 2024 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–8. Available online: <https://ieeexplore.ieee.org/document/10736481> (accessed on 5 November 2024).
31. Zhu, T.; Huang, Z.; Zhou, L.; Chen, G.; Meng, Y.; Zhu, H. Collaborative ad fraud detection in ad networks. *IEEE Trans. Netw.* **2025**, *33*, 1779–1794. [CrossRef]
32. Cao, C.; Gao, Y.; Luo, Y.; Xia, M.; Dong, W.; Chen, C.; Liu, X. AdSherlock: Efficient and deployable click fraud detection for mobile applications. *IEEE Trans. Mob. Comput.* **2020**, *20*, 1285–1297. [CrossRef]
33. Zhang, R.; Sridhar, R.P.; Yao, M.; Yang, Z.; Oygenblik, D.; Xu, H.; Dave, V.; Herley, C.; England, P.; Saltaformaggio, B. Identifying incoherent search sessions: Search click fraud remediation under real-world constraints. In *Proceedings of the 2025 IEEE Symposium on Security and Privacy (SP)*; IEEE: Piscataway, NJ, USA, 2025; pp. 93–111. Available online: <https://ieeexplore.ieee.org/document/11023363> (accessed on 16 June 2025).

34. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting deep learning models for tabular data. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 18932–18943. [[CrossRef](#)]
35. Liu, T.; Fan, J.; Tang, N.; Li, G.; Du, X. Controllable tabular data synthesis using diffusion models. *Proc. ACM Manag. Data* **2024**, *2*, 1–29. [[CrossRef](#)]
36. Yang, Z.; Yu, H.; Guo, P.; Zanna, K.; Yang, X.; Sano, A. Balanced mixed-type tabular data synthesis with diffusion models. *Trans. Mach. Learn. Res.* **2025**, *2025*, 3537. [[CrossRef](#)]
37. Bahri, D.; Jiang, H.; Tay, Y.; Metzler, D. Scarf: Self-supervised contrastive learning using random feature corruption. *arXiv* **2021**, arXiv:2106.15147. [[CrossRef](#)]
38. Chen, Q.; Lacomis, J.; Schwartz, E.J.; Neubig, G.; Vasilescu, B.; Le Goues, C. Varclr: Variable semantic representation pre-training via contrastive learning. In Proceedings of the 44th International Conference on Software Engineering, Pittsburgh, PA, USA, 21–29 May 2022; pp. 2327–2339. [[CrossRef](#)]
39. Tao, S.; Zhu, T.; Wang, H.; Meng, X. SemanticMask: A Contrastive View Design for Anomaly Detection in Tabular Data. In Proceedings of the IJCAI, Jeju Island, Republic of Korea, 3–9 August 2024; pp. 2370–2378. [[CrossRef](#)] [[PubMed](#)]
40. Cui, W.; Hosseinzadeh, R.; Ma, J.; Wu, T.; Sui, Y.; Golestan, K. Tabular data contrastive learning via class-conditioned and feature-correlation based augmentation. *arXiv* **2024**, arXiv:2404.17489. [[CrossRef](#)]
41. Long, G.; Zhang, Z. PUNet: A semi-supervised anomaly detection model for network anomaly detection based on positive unlabeled data. *Comput. Mater. Contin.* **2024**, *81*, 327. [[CrossRef](#)]
42. An, S.; Kim, J.; Kim, S.; Chikontwe, P.; Jung, J.; Jeon, H.; Park, S.H. Few-shot anomaly detection using positive unlabeled learning with cycle consistency and co-occurrence features. *Expert Syst. Appl.* **2024**, *256*, 124890. [[CrossRef](#)]
43. Zhou, G.; Guo, X.; Liu, Z.; Li, T.; Li, Q.; Xu, K. Trafficformer: An efficient pre-trained model for traffic data. In *Proceedings of the 2025 IEEE Symposium on Security and Privacy (SP)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1844–1860. Available online: <https://ieeexplore.ieee.org/document/11023272> (accessed on 16 June 2025).
44. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141. [[CrossRef](#)]
45. Ma, D.; Wan, F. Research on Intelligent Recognition of Ad Click Fraud Based on Deep FM Heterogeneous Integration Model. *Int. J. High Speed Electron. Syst.* **2025**, *35*, 2540726. [[CrossRef](#)]
46. Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; Isola, P. What makes for good views for contrastive learning? *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6827–6839. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.