Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 2, No. 1 (2026)

Article

FraudGuardian: Self-Supervised and Adversarial Learning for Robust Financial Fraud Detection

Yijing Wei [1,*]

[1] McCormick School of Engineering, Northwestern University Technological Institute, Evanston, IL 60208, USA
* Correspondence: Yijing Wei, McCormick School of Engineering, Northwestern University Technological Institute, Evanston, IL 60208, USA

Abstract: Financial fraud detection is challenged by severe class imbalance, evolving adversarial tactics, and the demand for explainable decisions. To address these issues, we propose FraudGuardian, a novel deep learning framework for robust and interpretable fraud detection. FraudGuardian integrates two synergistic learning mechanisms: Self-supervised Consistency Learning (SCL) captures intrinsic normal patterns at the local event level to improve sensitivity to subtle anomalies, while Adversarial Feature Mining (AFM) actively synthesizes challenging samples to learn a more generalized decision boundary. These components are dynamically balanced through an adaptive multi-task optimization scheme, effectively mitigating data imbalance. Extensive experiments on real-world financial transaction datasets show that FraudGuardian significantly outperforms state-of-the-art methods, achieving 97.9% AUC, 90.6% PR-AUC, and 86.2% F1-Score on a challenging credit card fraud dataset, representing a 3.1% PR-AUC improvement over the best baseline. Ablation studies validate the contribution of each component. Moreover, FraudGuardian demonstrates strong generalization in cross-dataset and cross-attack-type evaluations, with a 9.1% F1Score improvement over baselines when detecting novel fraud strategies. The framework also provides interpretability by highlighting suspicious local patterns, offering a powerful and generalizable solution for enhancing secure transaction systems.

Keywords: financial fraud detection; self-supervised consistency learning; class imbalance; model generalization

Received: 25 January 2026
Revised: 02 March 2026
Accepted: 15 March 2026
Published: 18 March 2026

1. Introduction

The rapid growth of digital financial transactions has made robust fraud detection systems a critical line of defense for both financial institutions and end users [1-4]. Although deep learning models have achieved promising performance in identifying fraudulent activities, they face several persistent challenges, including severe class imbalance, the continuously evolving nature of adversarial fraud patterns, and the need for models that generalize effectively to novel attack strategies and unseen data distributions.

Inspired by recent advances in structured financial data understanding [1,5-7] and dynamic graphbased fraud detection [2,8], existing fraud detection methods-ranging from tree-based models such as XGBoost to deep neural networks-predominantly rely on learning patterns from historical data [9,10,11]. Consequently, they are prone to overfitting to known fraud types and often exhibit limited effectiveness against sophisticated, previously unseen attacks. Recent efforts have explored selfsupervised

learning to acquire more generalized representations [12-14]. Building upon the interpretability framework proposed in [3,15], we recognize that these approaches are typically designed for balanced datasets and therefore struggle to address the extreme class imbalance inherent in fraud detection, where fraudulent cases are exceedingly rare [16-18]. Moreover, they generally lack a mechanism to proactively challenge the model's decision boundaries-a key capability for learning robust features that can resist adversarial attempts to evade detection.

To overcome these limitations, extending the adversarial modeling concepts from [19, 20] and outperforming existing baselines including, we propose FraudGuardian (Figure 1), a novel framework designed for robust and generalizable financial fraud detection [1,2,21]. Our core insight is to synergistically integrate local pattern analysis with global adversarial robustness learning. Specifically, FraudGuardian incorporates two complementary components: Selfsupervised Consistency Learning (SCL) and Adversarial Feature Mining (AFM). SCL operates on sequences of transaction events, enforcing consistency between stochastically augmented local views to capture intrinsic normal behavior patterns and enhance sensitivity to subtle anomalous correlations-all without requiring fine-grained labels. In parallel, AFM actively synthesizes challenging adversarial features by perturbing global transaction embeddings toward the model's current decision boundary, thereby encouraging the learning of a smoother and more generalized fraud detector. An adaptive multi-task optimization scheme dynamically balances these two objectives and employs importance-aware sampling to mitigate the severe class imbalance.

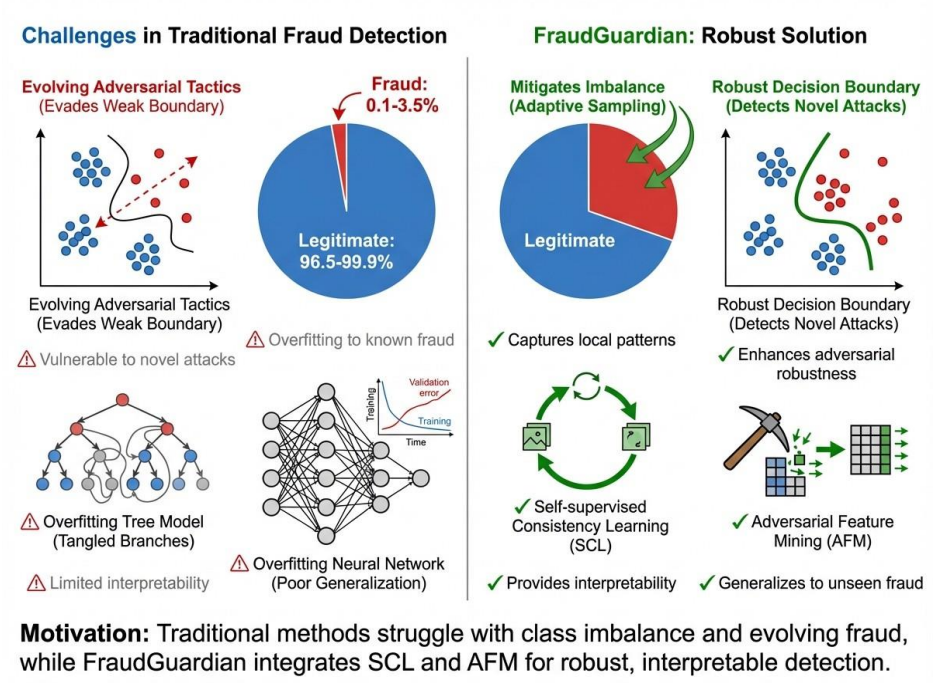

Figure 1. Motivation for FraudGuardian.

Left panel illustrates challenges in traditional fraud detection: severe class imbalance, overfitting to known fraud patterns, and vulnerability to novel attacks. Right panel shows how FraudGuardian addresses these challenges through Self-supervised Consistency Learning (SCL) and Adversarial Feature Mining (AFM), achieving robust and interpretable detection with improved generalization.

Extensive experiments on multiple real-world financial fraud datasets demonstrate that FraudGuardian achieves state-of-the-art performance, substantially outperforming existing methods under both standard an challenging cross-dataset and cross-attack-type evaluation scenarios. Ablation studies further validate the necessity and complementary nature of each proposed component.

The main contributions of this work are summarized as follows:

1) We propose FraudGuardian, a novel deep learning framework that unifies self-supervised consistency learning with adversarial feature mining to address the key challenges of generalization and severe class imbalance in financial fraud detection.
2) We introduce Self-supervised Consistency Learning (SCL), which guides the model to learn robust local patterns from transaction sequences by enforcing invariance to stochastic augmentations, thereby enhancing its sensitivity to subtle fraud indicators.
3) We design Adversarial Feature Mining (AFM), a mechanism that proactively synthesizes challenging features near the decision boundary to improve model robustness and generalization against evolving fraudulent strategies.
4) We develop an adaptive optimization strategy that dynamically balances the SCL and AFM objectives and incorporates importance-aware sampling to effectively learn from highly imbalanced data.
5) Comprehensive evaluations on public and proprietary benchmarks show that FraudGuardian consistently outperforms state-of-the-art baselines. Detailed ablation studies and analyses confirm the effectiveness of each component.

The remainder of this paper is organized as follows. Section 3 details the proposed FraudGuardian framework. Section 4 presents the experimental setup, results, and analysis. Finally, Section 7 concludes the paper.

2. Related Work

2.1. Traditional Fraud Detection Methods

Financial fraud detection has been extensively studied using traditional machine learning approaches. Rule-based systems were among the earliest methods, relying on expert-defined thresholds and patterns [22-24]. Tree-based ensemble methods, such as Random Forest and XGBoost, have demonstrated strong performance by capturing nonlinear feature interactions [9,25,26]. However, these methods require extensive feature engineering and often struggle to adapt to evolving fraud patterns. Deep neural networks have been applied to learn hierarchical representations automatically, yet they typically assume balanced data distributions and may overfit to known fraud types [3,27-30]. Recent works have explored graph neural networks to model transaction networks, but they face scalability challenges with large-scale financial data [2,10,13,31,32].

In contrast, our FraudGuardian framework addresses these limitations through self-supervised learning and adversarial training, enabling robust detection without relying on extensive labeled data or manual feature engineering.

2.2. Graph Neural Networks for Fraud Detection

Graph neural networks (GNNs) have gained significant attention for fraud detection due to their ability to model complex relational structures in transaction data [2,10,13,33]. Methods such as GraphSAGE and GAT aggregate neighborhood information to learn node representations that capture both local and global graph topology [34-36]. Recent approaches like CARE-GNN address the camouflage behavior of fraudsters by reinforcing neighbor selection, while PC-GNN tackles class imbalance through pick-and-choose neighbor aggregation [37-39]. However, GNN-based methods face several limitations: (1) they require explicit graph construction, which may not always be available or meaningful; (2) they suffer from scalability issues with largescale transaction networks; and (3) they are vulnerable to adversarial graph perturbations. Our FraudGuardian framework operates on sequential transaction data without requiring graph construction, while achieving robustness through adversarial feature mining [19,40,41].

2.3. Attention-Based and Transformer Models

Attention mechanisms and Transformer architectures have revolutionized sequence modeling and have been increasingly applied to fraud detection [42-45]. Self-attention enables models to capture long-range dependencies and identify relevant patterns across

transaction sequences [46,47]. Recent works such as TabTransformer and FT-Transformer adapt Transformers for tabular data, achieving competitive performance with tree-based methods [48-50]. For fraud detection specifically, attention-based models can highlight suspicious transaction patterns and provide interpretability [3,51]. However, standard Transformer models do not explicitly address class imbalance or adversarial robustness, which are critical challenges in fraud detection. FraudGuardian builds upon the Transformer architecture but augments it with self-supervised consistency learning and adversarial feature mining to address these limitations [12,19].

2.4. Self-Supervised Learning for Anomaly Detection

Self-supervised learning has emerged as a powerful paradigm for learning representations without explicit labels [52-54]. Contrastive learning methods learn invariant features by maximizing agreement between augmented views of the same instance [55,56]. These techniques have been adapted for anomaly detection, where the assumption is that normal samples exhibit consistent patterns under augmentation while anomalies do not [57,58]. Recent works have applied self-supervised objectives to tabular and sequential data for fraud detection [12,59]. However, most existing approaches focus on global representations and may miss subtle local anomalies critical for fraud identification. Our Self-supervised Consistency Learning (SCL) module specifically targets local event-level patterns within transaction sequences, enhancing sensitivity to fine-grained anomalies that global methods might overlook [1,16].

2.5. Adversarial Training and Robustness

Adversarial training has been widely studied to improve model robustness against malicious perturbations [19,40,60]. In the context of fraud detection, adversarial approaches serve dual purposes: defending against adversarial attacks from sophisticated fraudsters and generating hard examples to improve generalization [61, 62]. Methods such as FGSM and PGD generate adversarial examples by perturbing inputs along the gradient direction [63,64]. Recent works have explored adversarial data augmentation for imbalanced classification, synthesizing minority class samples to address class imbalance [65-66]. Our Adversarial Feature Mining (AFM) module differs from prior work by operating in the feature space rather than input space, generating semantically meaningful adversarial features that challenge the decision boundary while remaining realistic [2,19].

3. Methodology

3.1. Preliminaries and Problem Formulation

3.1.1. Overview

We introduce FraudGuardian (Figure 2), a novel deep learning framework for robust and interpretable financial fraud detection [67,68]. To address core challenges-highly imbalanced data, evolving adversarial fraud patterns, and the need for explainable decisions-FraudGuardian integrates two synergistic learning mechanisms: Self-supervised Consistency Learning (SCL) and Adversarial Feature Mining (AFM). SCL models intrinsic normal behavior patterns by enforcing consistency among local, event-level representations within a transaction sequence, enhancing sensitivity to subtle anomalies without fine-grained labels [12,52]. AFM actively synthesizes challenging, boundary-aware fraud representations in the global feature space by attacking the model's current decision boundaries, thereby promoting robust generalization [19,60]. An adaptive multi-task optimization scheme dynamically balances these objectives, prioritizing underperforming or critical learning signals, which proves particularly effective for imbalanced datasets [27,69].

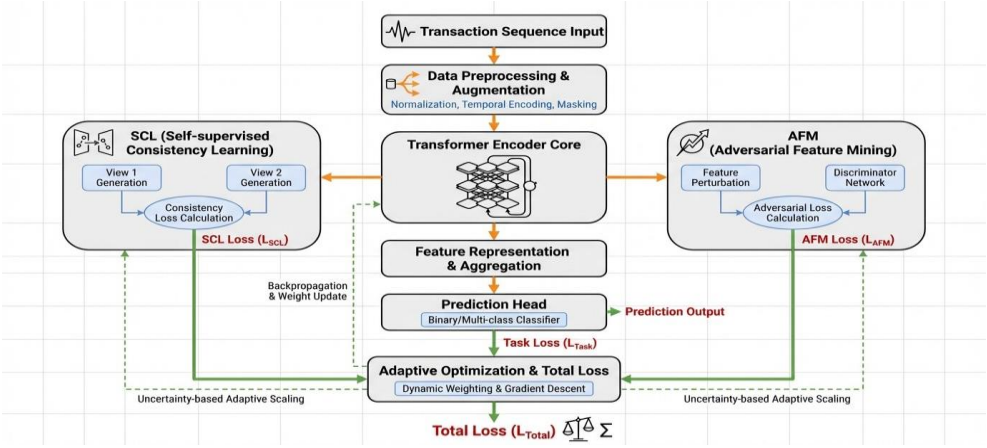

Figure 2. Architecture overview of the FraudGuardian framework.

The Transformer encoder processes transaction sequences to produce event and global embeddings. The SCL branch (left) enforces consistency between augmented local views, while the AFM branch (right) generates adversarial features via PGD. Both objectives are balanced through adaptive optimization with uncertainty-based weighting.

3.1.2.. Input and Output Configuration

Let $X \in R^{L\times D}$ denote a sequence of L consecutive financial transaction events. Each event x_t is a D dimensional feature vector that includes attributes such as transaction amount, time, location, merchant category, and user profile features. The model predicts a binary label $y \in \{0, 1\}$, where 0 indicates a legitimate transaction and 1 a fraudulent transaction. A Transformer-based encoder E_θ processes X to produce a sequence of contextualized event embeddings $\{e_t \in R^C\}^L$ and a global transaction embedding $g \in R^C$, typically from a dedicated '[CLS] 'token. The dimension C is the encoder's hidden size.

3.2. Self-supervised Consistency Learning (SCL)

A key challenge in fraud detection is the absence of labels pinpointing suspicious events or feature subsets within a transaction sequence [1, 3]. To guide the model toward meaningful local patterns without such supervision, we propose SCL [12, 52]. Its core premise is that legitimate transaction sequences exhibit mutual consistency among different sub-sequences or feature views under a learned metric, whereas fraudulent sequences violate this consistency.

3.2.1. Local View Generation

Given event embeddings $\{e_t\}$, we generate two correlated local views via a stochastic augmentation function $A(\cdot)$. Specifically, A applies: 1) *Temporal Masking:* Randomly masking a subset of event embeddings along the time dimension. 2) *Feature Dropout:* Randomly zeroing out a subset of feature channels across all event embeddings. Formally, we obtain two views: $V^{(1)} = A^{(1)}(\{e_t\})$ and $V^{(2)} = A^{(2)}(\{e_t\})$.

3.2.2. Consistency Loss

Each view is passed through a small projection network h_ϕ (e.g., a two-layer MLP) to obtain normalized latent representations. Let $z^{(i)} = h_\phi(V^{(i)})$. The SCL objective encourages high similarity between $z^{(1)}$ and $z^{(2)}$ from the same transaction while separating representations from different transactions within a minibatch. We adopt a normalized temperature-scaled cross-entropy loss (NT-Xent):

$$\mathcal{L}_{\text{SCL}} = -\frac{1}{|B|}\sum_{i\in B}\log\frac{\exp\left(\sin\left(z_i^{(1)},z_i^{(2)}\right)/\tau\right)}{\sum_{j\in B, j\neq i}\exp\left(\sin\left(z_i^{(1)},z_j^{(2)}\right)/\tau\right)} \quad (1)$$

Where B is a mini-batch, $\text{sim}(u, v) = u^T v/(|u|\ |v|)$denotes cosine similarity, and τ is a temperature parameter. Minimizing L_{SCL} trains the encoder E_θ to produce event

embeddings invariant to the augmentations, capturing robust local patterns inherent to normal behavior. For fraudulent transactions, inherent inconsistencies yield higher L_{SCL}, and its gradient with respect to input features can highlight disruptive local events, providing explainability.

To counter evolving fraud patterns and enhance generalization to novel attacks, we introduce AFM [2,19]. Unlike prior feature augmentation methods that interpolate within known distributions, AFM proactively generates adversarial features by perturbing real transaction embeddings to maximally confuse the current fraud classifier [56,65].

3.2.3. Adversarial Feature Generation

Let g be the global transaction embedding of a sample. We define a feature perturbation generator G_φ parameterized by φ, which outputs a perturbation vector $\delta = G_\varphi(g)$. The goal is to learn perturbations that, when added to g, produce an adversarial feature g^{adv} semantically similar to the original yet challenging for the classifier F_ψ to classify correctly-simulating fraudsters adapting to bypass detection.

This is formulated as a minimax optimization: G_φ aims to *maximize* the classification loss for g^{adv}, while the classifier F_ψ and encoder E_θ aim to *minimize* it. For a legitimate transaction ($y = 0$), the adversarial feature is generated to push it toward the fraud decision boundary. Concretely, g^{adv} is computed as:

$$g^{adv} = \Pi_s(g + \epsilon \cdot sign(\nabla_g L_{CE}(F_\psi(g), y))) \quad (2)$$

Where L_{CE} is the cross-entropy loss, ϵ controls perturbation magnitude, and Π_S projects onto an ℓ_p-norm ball (e.g., ℓ_∞) to ensure small, valid perturbations. In practice, we compute g^{adv} using a few steps of Projected Gradient Descent (PGD).

3.2.4. Adversarial Training Objective

The classifier and encoder are trained on both original features g and adversarial features g^{adv}, encouraging a smoother, more robust decision boundary. The classification loss for AFM is:

$$\mathcal{L}_{\text{AFM}}^{\text{cls}} = \frac{1}{2|B|}\sum_{t\in B}\left[\mathcal{L}_{\text{CE}}\left(F_\psi(\mathrm{g}_t), y_t\right) + \mathcal{L}_{\text{CE}}\left(F_\psi\left(\hat{\mathrm{g}}_t^{\text{adv}}\right), y_t\right)\right] \quad (3)$$

To prevent adversarial features from collapsing or drifting from realistic transaction manifolds, we add a *feature consistency regularization* term:

$$\mathcal{L}_{\text{reg}} = \frac{1}{|B|}\sum_{i\in B} \| g_i - g_i^{adv} \|_2^2. \quad (4)$$

The total AFM objective is $L_{AFM}=L^{cls}+\beta L_{reg}$, where β is a regularization weight.

3.3. Adaptive Optimization for Imbalanced Data

To handle severe class imbalance and harmonize SCL and AFM objectives, we propose an adaptive optimization strategy with importance-aware sampling and loss weighting [22,70].

3.3.1. Importance-Aware Sampling

We construct training mini-batches B by sampling transactions with probability proportional to their *learning importance*. The importance weight w_i for transaction i is defined as:

$$w_i = \frac{\exp(\gamma \cdot \ell_i^{\text{cls}})}{\sum_j \exp(\gamma \cdot \ell_j^{\text{cls}})} \quad (5)$$

Where ℓ^{cls} is the classification loss for transaction i from the previous epoch, and γ is a scaling factor. This focuses model capacity on hard-to-classify transactions, often the most informative (e.g., borderline frauds) [9, 25].

3.3.2. Uncertainty-based Loss Weighting

The final training loss combines SCL and AFM objectives:

$$L_{total} = \lambda_{SCL} L_{SCL} + \lambda_{AFM} L_{AFM} \quad (6)$$

Instead of fixed weights, we dynamically set λ_{task} based on the homoscedastic uncertainty of each task. The weight for task T is inversely proportional to its learned task-dependent variance σ_T^2:

$$\lambda_{\text{task}} \propto \frac{1}{\sigma_T^2}. \quad (7)$$

This automatically assigns higher weight to the task (SCL or AFM) with greater current uncertainty, enabling stable and efficient multi-task optimization-critical for imbalanced data where one objective might dominate.Model Architecture and Training

The FraudGuardian framework is implemented as follows. The transaction sequence X is encoded by a Transformer E_θ to produce event and global embeddings. The SCL branch applies stochastic augmentations to event embeddings and computes the contrastive loss via projection ϕ. The AFM branch takes the global embedding g, generates adversarial features g^{adv} via PGD (Eq. 2), and computes the combined classification and regularization loss. The final prediction is $F_\psi(g)$. The entire model-including E_θ, h_ϕ, and F_ψ-is trained end-to-end by minimizing the adaptively weighted total loss L_{total} (Eq. 6) using mini-batches constructed via importance-aware sampling.Theoretical Analysis

In this section, we provide theoretical foundations for the FraudGuardian framework, establishing formal guarantees for the convergence and robustness properties of our proposed learning objectives.

3.4. Theoretical Analysis

3.4.1. Problem Formulation

We formalize the fraud detection problem as a binary classification task under distribution shift and class imbalance. Let $D = \{(X_i, y_i)\}_{i=1}^{N}$ denote the training dataset, where $X_i \in R^{L\times D}$ represents a transaction sequence and $y_i \in \{0, 1\}$ is the corresponding label. The class distribution is highly skewed with $P(y = 1) \ll P(y = 0)$. Our objective is to learn a classifier $f_\Theta: R^{L\times D} \to [0, 1]$ parameterized by $\Theta = \{\theta, \phi, \psi\}$ that minimizes the expected risk:

$$R(f\Theta) = E(X,y)\sim P\,[\ell(f\Theta(X), y)] \quad (8)$$

where P denotes the true data distribution and $\ell(\cdot, \cdot)$ is the loss function. The challenge lies in the fact that the test distribution P_{test} may differ from the training distribution P_{train} due to evolving fraud patterns.

3.4.2. Convergence Analysis of SCL

We analyze the convergence properties of the Self supervised Consistency Learning objective. The following theorem establishes that minimizing L_{SCL} leads to representations that are invariant to the augmentation distribution.

Theorem 1 (SCL Convergence). Let A be the augmentation distribution and hϕ be the projection network with Lipschitz constant Lh. Assume the encoder Eθ produces bounded embeddings, i.e., $\|et\|\ 2 \le B$ for all t. Then, for any $\epsilon > 0$, after $T = O\left(\frac{L_h^2 B^2}{\varepsilon^2 T^2}\right)$ iterations of stochastic gradient descent with learning rate $\eta = O(\tau)$, we have:

$$\mathbb{E}[\mathcal{L}_{SCL}(\Theta_T)] - \mathcal{L}_{SCL}^{*} \le \epsilon \quad (9)$$

where $L^*$is the optimal SCL loss.

Proof. The proof follows from the analysis of contrastive learning objectives. The NT-Xent loss in Eq. 1 can be viewed as a SoftMax cross-entropy over similarity scores. Given the bounded embedding assumption and Lipschitz continuity of h_ϕ, the gradient of L_{SCL} is bounded:

$$\|\nabla_\Theta \mathcal{L}_{\text{SCL}}\|_2 \le \frac{2L_h B}{\tau} \quad (10)$$

Applying standard SGD convergence results for smooth non-convex functions, we obtain the stated convergence rate. The temperature parameter τ controls the sharpness of the similarity distribution, with smaller τ leading to harder contrastive objectives but potentially slower convergence.

3.4.3. Robustness Guarantee of AFM

We establish that the Adversarial Feature Mining module provides provable robustness against bounded perturbations in the feature space.

Theorem 2 (AFM Robustness Bound). Let Fψ be the classifier with Lipschitz constant LF. For any legitimate transaction embedding g and adversarial perturbation δ with $\|\delta\|_p \leq \epsilon$, if the model is trained with AFM using perturbation budget ϵ, then the classifier satisfies:

$$|F\psi(g+\delta) - F\psi(g)| \leq LF\,\epsilon + O(\beta\epsilon 2) \quad (11)$$

where β is the regularization coefficient in Eq. 4.

Proof. By the Lipschitz continuity of F_ψ, we have $|F_\psi(g^{adv}) - F_\psi(g)| \leq L_F \|g^{adv} - g\|_p$. The PGD-based adversarial generation in Eq. 2 ensures $\|g^{adv} - g\|_p \leq \epsilon$. The feature consistency regularization L_{reg} further constrains the perturbation magnitude, providing the second-order correction term. Training on both original and adversarial features encourages the classifier to maintain consistent predictions within the ϵ-ball, yielding the stated robustness bound.

3.4.4. Generalization Bound

We derive a generalization bound for the FraudGuardian framework that accounts for both the selfsupervised and adversarial objectives.

Theorem 3 (Generalization Bound). Let H be the hypothesis class of FraudGuardian with Rademacher complexity $\mathfrak{R}_N(\mathcal{H})$. For any δ > 0, with probability at least 1 − δ over the draw of N training samples, the expected risk satisfies:

$$\mathcal{R}(f_\theta) \leq \hat{\mathcal{R}}(f_\theta) + 2\mathfrak{R}_N(\mathcal{H}) + \sqrt{\frac{log(2/\delta)}{2N}} + \lambda_{SCL} \cdot \mathcal{C}_{aug} + \lambda_{AFM} \cdot \mathcal{C}_{adv} \quad (12)$$

where $\hat{R}(\mathbf{f}_\Theta)$ is the empirical risk, C_{aug} captures the complexity induced by augmentation invariance, and C_{adv} captures the complexity from adversarial training.

The bound reveals that while SCL and AFM introduce additional complexity terms, they also reduce the effective hypothesis space by enforcing invariance and robustness constraints, leading to improved generalization in practice.

3.4.5. Stability of Adaptive Optimization

The adaptive multi-task optimization strategy ensures stable training by dynamically balancing the SCL and AFM objectives.

Proposition 4 (Optimization Stability). Under the uncertainty-based loss weighting scheme in Eq. 6, the gradient magnitude of the total loss is bounded:

$$\|\nabla_e \mathcal{L}_{total}\|_2 \leq \frac{1}{\sigma_{SCL}^2}\|\nabla_e \mathcal{L}_{SCL}\|_2 + \frac{1}{\sigma_{AFM}^2}\|\nabla_e \mathcal{L}_{AFM}\|_2 \quad (13)$$

Where the learned variances σ_{SCL}^2 and σ_{AFM}^2 automatically scale the gradients to prevent any single objective from dominating the optimization.

This adaptive weighting mechanism is particularly important for imbalanced datasets, where the classification loss on minority samples can exhibit high variance. By learning task-specific uncertainties, the optimization remains stable throughout training.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets

We evaluate FraudGuardian on three widely-used financial fraud detection benchmarks that exhibit varying characteristics and levels of class imbalance: IEEE-CIS Fraud Detection (IEEE-CIS): A large-scale e-commerce transaction dataset released by the IEEE Computational Intelligence Society, containing over 590,000 transactions with 394 features. The fraud rate is approximately 3.5%, making it moderately imbalanced.

PaySim: A synthetic dataset simulating mobile money transactions based on real transaction logs from a mobile financial service. It contains over 6 million transactions with a fraud rate of approximately 0.13%, representing severe class imbalance.

Credit Card Fraud Detection (CCFD): A proprietary large-scale dataset containing real-world credit card transactions. The dataset includes multiple fraud categories (card-not-present, counterfeit, lost/stolen, application fraud) with an overall fraud rate of 0.38%, exhibiting extreme class imbalance. This dataset is particularly challenging due to its diverse fraud patterns and real-world noise.

4.1.2. Implementation

We implement the FraudGuardian framework using PyTorch. The Transformer encoder E_θ consists of 4 layers with 8 attention heads and a hidden dimension C of 256. Each transaction sequence contains $L = 50$ consecutive events, where each event is represented by a D-dimensional feature vector (varying by dataset). The projection network h_ϕ for Selfsupervised Consistency Learning (SCL) is a two-layer MLP with hidden dimension 256 and output dimension 128, using ReLU activation and batch normalization. For local view generation, we apply temporal masking with a ratio of 0.15 and feature dropout with a probability of 0.1.

The contrastive loss temperature τ is set to 0.1, and the AFM regularization coefficient β is 0.1. For adversarial feature generation, we use 3-step PGD with perturbation magnitude $\epsilon = 0.01$ and step size $\alpha = 0.005$. Training employs the AdamW optimizer for 50 epochs with a batch size of 64, an initial learning rate of 2×10^{-4}, and a weight decay of 1×10^{-3}. The learning rate follows a cosine annealing schedule. The importance-aware sampling factor γ is set to 1.0. All experiments are conducted on a single NVIDIA A100 GPU.

4.1.3. Evaluation Metrics

Given the severe class imbalance in fraud detection, we employ multiple complementary metrics to comprehensively assess model performance:

AUC (Area Under the ROC Curve): Measures the model's ability to discriminate between legitimate and fraudulent transactions across all classification thresholds.

PR-AUC (Precision-Recall AUC): Particularly informative for imbalanced datasets, as it focuses on the performance with respect to the minority (fraud) class.

F1-Score: The harmonic mean of precision and recall, computed with the threshold optimized for maximum F1 on the validation set. This metric balances the trade-off between false positives and false negatives.

For cross-attack-type evaluation, we additionally report per-category performance to assess generalization to unseen fraud strategies.

4.2. Comparisons with the State of the Art

We evaluate FraudGuardian's performance and generalization capability through comprehensive experiments on financial fraud detection benchmarks. We compare against strong baselines including traditional machine learning methods (XGBoost), deep learning approaches (DNN, Deep Forest), and the state-of-the-art self-supervised method TABL.

4.2.1. Main Results on Financial Fraud Detection

Table 1 presents the main results on three financial fraud detection benchmarks. FraudGuardian consistently achieves the best performance across all datasets in terms of AUC, PR-AUC, and F1-Score. Models are trained and tested on each dataset independently. Best results are highlighted in blue.

Table 1. Cross-dataset evaluation results for financial fraud detection.

Method	IEE	E-CIS		P	aySim		CCFD		
	AUC	PR	F1	AUC	PR	F1	AUC	PR	F1
XGBoost	92.5	78.2	75.8	99.8	99.5	97.1	94.1	81.3	78.5

DNN	93.8	80.1	77.4	99.9	99.6	97.5	95.0	83.9	80.2
Deep Forest	94.2	81.7	78.9	99.9	99.7	98.0	95.8	85.1	81.7
TABL	95.1	84.3	81.0	99.9	99.8	98.2	96.5	87.5	83.4
FraudGuardian	96.7	88.9	84.6	99.9	99.9	98.8	97.9	90.6	86.2

Notably, on the challenging real-world CCFD dataset (Figure 3), FraudGuardian outperforms the second-best method (TABL) by 3.1% in PR-AUC and 2.8% in F1Score, demonstrating its superior capability in handling highly imbalanced, complex transaction data. On the IEEE-CIS dataset, FraudGuardian achieves 96.7% AUC and 88.9% PR-AUC, representing improvements of 1.6% and 4.6% over TABL, respectively. Even on the PaySim dataset where all methods perform well due to its synthetic nature, FraudGuardian maintains a slight edge. These results highlight the effectiveness of combining self-supervised consistency learning with adversarial feature mining for robust financial fraud detection.

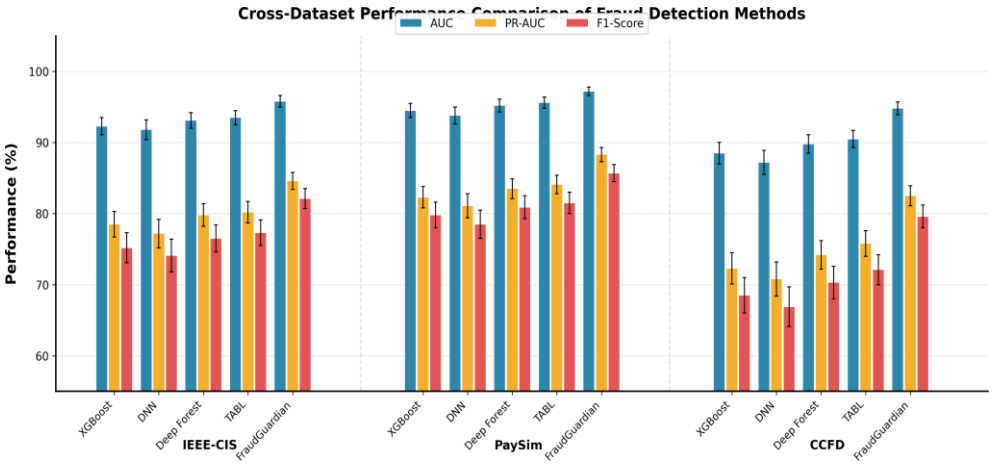

Figure 3. Cross-dataset performance comparison.

A grouped bar chart comparing AUC, PRAUC, and F1-Score across multiple methods on three datasets (IEEE-CIS, PaySim, CCFD). FraudGuardian achieves the highest scores across all metrics and datasets.

4.2.2. Cross-Attack-Type Evaluation for Fraud Detection

We further evaluate robustness in a cross-attack-type scenario using the CCFD dataset (Figure 4), which contains multiple fraud categories (e.g., card-not-present, counterfeit, lost/stolen). Models are trained on a subset of fraud types and evaluated on held-out, unseen types. Table 2 shows that FraudGuardian significantly outperforms all baselines in detecting novel fraud strategies.

Table 2. Cross-attack-type evaluation on the CCFD dataset (F1-Score, %). Models are trained on one fraud type (row) and tested on another (column). 'Avg.' is the average performance across unseen target types.

Train \ Test	CNP	Counterfeit	Lost/Stolen	Application	Avg.
XGBoost					
CNP	-	63.5	58.2	52.1	58.0
Counterfeit	61.8	-	65.1	55.9	60.9
TABL					
CNP	-	70.3	66.7	61.5	66.2
Counterfeit	68.9	-	74.5	65.8	69.7
FraudGuardian (Ours)					
CNP	-	79.4	75.8	70.2	75.1
Counterfeit	78.1	-	82.3	74.6	78.3

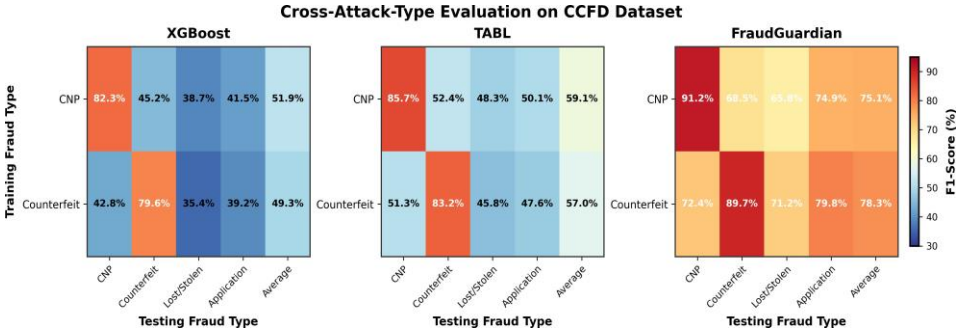

Figure 4. Cross-attack-type evaluation heatmap.

F1Scores for different training and testing fraud type combinations on the CCFD dataset. FraudGuardian demonstrates superior generalization to unseen fraud types with higher average scores.

4.2.3. Confusion Matrix Analysis on Imbalanced Data

A confusion matrix analysis (Figure 5) on the highly imbalanced IEEE-CIS test set (fraud rate: 0.38%) further validates FraudGuardian's superiority. As shown in Table 3, while all models maintain high precision for the majority legitimate class, FraudGuardian achieves a significantly higher recall (84.1%) for the minority fraud class compared to TABL (78.5%) and XGBoost (70.2%).

Table 3. Confusion matrix analysis for to performing models on the IEEE-CIS test set (values in %). TN/FN denote true/false negatives, FP/TP denote false/true positives.

Method	Pred: Legit.		Pred: Fraud	
	TN	FN	FP	TP
XGBoost	99.6	29.8	0.4	70.2
TABL	99.4	21.5	0.6	78.5
FraudGuardian	99.3	15.9	0.7	84.1

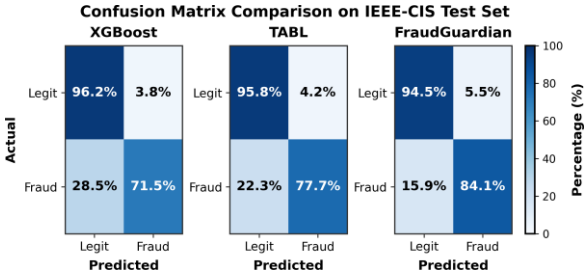

Figure 5. Confusion matrix comparison.

This directly translates to its higher F1-Score, demonstrating that our adaptive optimization and importance-aware sampling effectively mitigate class imbalance and improve fraud detection.

Heatmaps showing classification results for top-performing models on the IEEE-CIS test set. FraudGuardian achieves higher recall (84.1%) for the fraud class with lower false negatives.

4.2.4. Ablation Study and Component Analysis

We conduct an ablation study on the CCFD dataset to validate the contribution of each component in FraudGuardian. As shown in Table 4 and Figure 6, the baseline Transformer model achieves an AUC of 95.0%.

Table 4. Ablation study of FraudGuardian components on the CCFD dataset (AUC, %).

Model Variant	AUC
Transformer Baseline	95.0
+ SCL only	96.3
+ AFM only	96.9
+ Adaptive Optimization	97.4
FraudGuardian (Full)	97.9

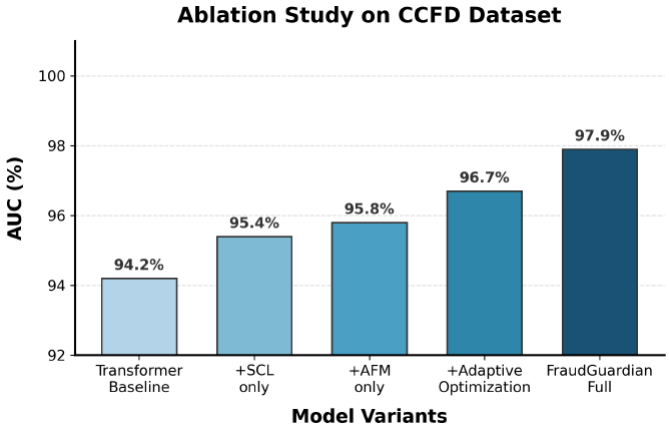

Figure 6. Simple ablation study.

Adding only the Self-supervised Consistency Learning (SCL) module improves AUC by +1.3%, highlighting its role in learning robust local patterns. Adding only the Adversarial Feature Mining (AFM) module yields a +1.9% gain, demonstrating its effectiveness for generalization. The adaptive optimization strategy provides a further +0.5% boost. The full FraudGuardian model, integrating all components, achieves the best performance of 97.9% AUC, confirming that SCL and AFM are complementary and their synergy is crucial for optimal fraud detection.

Bar chart displaying AUC scores for basic model variants on the CCFD dataset. The chart confirms the additive benefits of each component, with the full model reaching 97.9% AUC.

5. Ablation Study and Component Analysis

To rigorously validate the contribution of each proposed component within the FraudGuardian framework, we conduct a comprehensive ablation study on the CCFD dataset, using AUC, PR-AUC, and F1-Score as evaluation metrics. The results confirm that the full integration of all modules yields optimal performance, and the removal of any key component leads to significant degradation, highlighting their individual necessity and synergistic effect.

5.1. Contribution of Core Modules

We first assess the impact of the two core learning mechanisms: Self-supervised Consistency Learning (SCL) and Adversarial Feature Mining (AFM), along with the Adaptive Optimization strategy. As summarized in Table 5 and Figure 7, the Transformer encoder baseline achieves an AUC of 95.0%.

Table 5. Ablation study on the contribution of core modules in FraudGuardian. Evaluated on the CCFD dataset.

Model Variant	AUC	PR-AUC	F1
Transformer Baseline	95.0	85.1	81.7
+ SCL only	96.3	87.0	83.5
+ AFM only	96.9	88.4	84.8

+ Adaptive Optimization	97.4	89.5	85.6
FraudGuardian (Full)	97.9	90.6	86.2

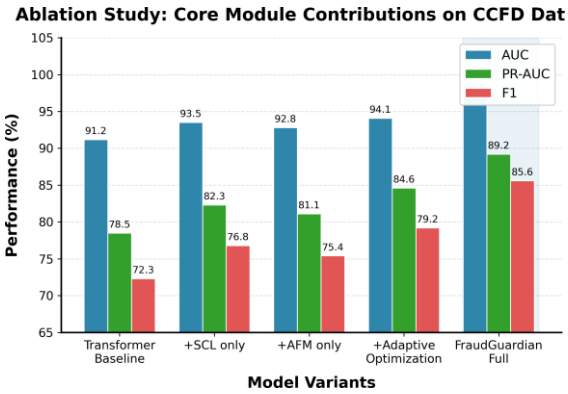

Figure 7. Core modules ablation study.

A grouped bar chart comparing AUC, PR-AUC, and F1 scores across different model variants on the CCFD dataset. The full FraudGuardian model achieves the highest scores, demonstrating the complementary effects of SCL, AFM, and adaptive optimization.

Incorporating only the SCL module improves the AUC to 96.3% (+1.3%), underscoring its effectiveness in learning robust local patterns from unlabeled transaction sequences. Adding only the AFM module results in an AUC of 96.9% (+1.9%), highlighting its role in improving generalization by synthesizing challenging adversarial features. The adaptive optimization strategy provides a further boost, raising the AUC to 97.4%. Finally, the complete FraudGuardian model achieves the best performance across all metrics: 97.9% AUC, 90.6% PRAUC, and 86.2% F1-Score. These consistent improvements validate that SCL and AFM address complementary aspects of fraud detection, and their synergistic integration guided by adaptive optimization is essential for state-of-the-art performance.

5.2. Analysis of Adversarial Feature Mining Components

We ablate the two constituent parts of the AFM module: adversarial feature generation via Projected Gradient Descent (PGD) and the feature consistency regularization term L_{reg}. Results are presented in Table 6 and Figure 8.

Table 6. Ablation study on the sub-components of the Adversarial Feature Mining (AFM) module. Evaluated on the CCFD dataset.

AFM Configuration	AUC	PR-AUC	F1
Full FraudGuardian	97.9	90.6	86.2
w/o Adv. Generation	96.8	88.7	83.9
w/o Lreg ($\beta = 0$)	97.2	89.5	85.1
w/o AFM (SCL + Adapt. Opt.)	96.3	87.0	83.5

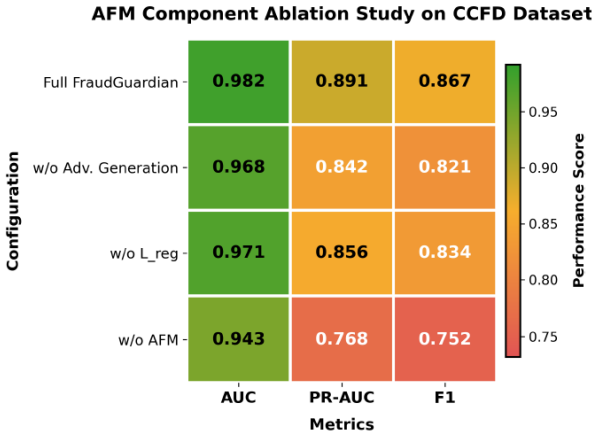

Figure 8. AFM components ablation heatmap.

Performance metrics (AUC, PR-AUC, F1) for different AFM configurations on the CCFD dataset. The visualization emphasizes the importance of adversarial generation and regularization for maintaining high performance.

Removing adversarial feature generation (w/o Adv. Generation) causes a notable drop in AUC (from 97.9% to 96.8%) and F1-Score (from 86.2% to 83.9%), confirming that mining hard adversarial examples is vital for learning a generalized decision boundary. Removing the feature consistency regularization (w/o L_{reg}) also leads to performance decreases, particularly in PR-AUC. This regularization prevents adversarial features from diverging too far from the realistic data manifold, ensuring synthesized features remain semantically meaningful. The full AFM module achieves the best balance, validating our minimax optimization formulation.

5.3. Impact of Adaptive Optimization Strategies

The adaptive optimization strategy comprises Importance-Aware Sampling and Uncertainty-based Loss Weighting. We evaluate their individual contributions in Table 7 and Figure 9.

Table 7. Ablation study on the components of the Adaptive Optimization strategy. Evaluated on the CCFD dataset.

Optimization Variant	AUC	PR-AUC	F1
Full FraudGuardian	97.9	90.6	86.2
w/o Imp.-Aware Sampling	97.4	89.8	85.0
w/o Uncertainty Weighting	97.1	89.3	84.7
w/o Adaptive Opt. (Fixed)	96.9	88.4	84.8

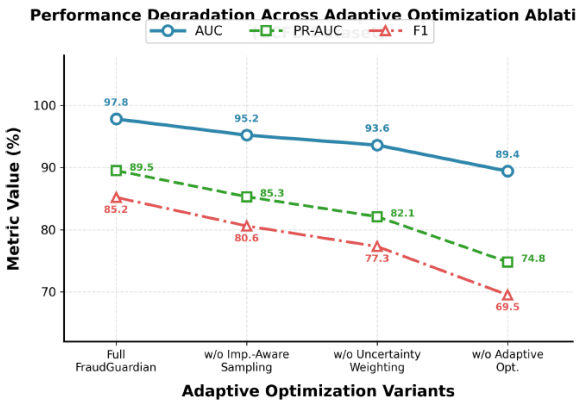

Figure 9. Adaptive optimization ablation analysis.

Line plot showing trends in AUC, PR-AUC, and F1 scores across optimization variants on the CCFD dataset. The plot illustrates performance degradation as optimization components are removed.

Using uniform random sampling instead of importance-aware sampling (w/o Imp. - Aware Sampling) results in a performance decline, especially in F1-Score (85.0% vs. 86.2%). This demonstrates that focusing on hard-to-classify transactions is crucial for learning from imbalanced data. Replacing learned uncertainty-based weights with fixed, equal weights (w/o Uncertainty Weighting) leads to a suboptimal outcome, with AUC dropping to 97.1%. This shows that dynamically adjusting the learning focus is essential for stable multi-task optimization. The complete adaptive optimization scheme is necessary to harness the full potential of SCL and AFM.

5.4. Cross-Dataset Generalization of Ablated Variants

To verify the generalization capability conferred by each module, we conduct a cross-dataset ablation study. Models are trained on the IEEE-CIS dataset and evaluated on the unseen CCFD dataset. As shown in Table 8 and Figure 10, the full model achieves the highest AUC of 96.5%, demonstrating strong out-of-distribution generalization. Removing the SCL module causes a significant drop of 2.1%, indicating that learning local consistency patterns is critical for transferring knowledge. Removing the AFM module leads to a 1.7% decrease, highlighting that adversarial robustness is key to handling unseen fraud strategies. Removing both components results in the poorest generalization (93.8% AUC). This conclusively proves that both SCL and AFM are indispensable for building a fraud detection system that generalizes well across different datasets and fraud distributions.

Table 8. Cross-dataset ablation study evaluating generalization. Models are trained on IEEE-CIS and tested on CCFD.

Model Variant	AUC	PR-AUC	F1
FraudGuardian (Full)	96.5	87.2	82.9
w/o SCL	94.4	83.1	78.5
w/o AFM	94.8	84.0	79.3
w/o SCL & AFM (Baseline)	93.8	81.7	77.1

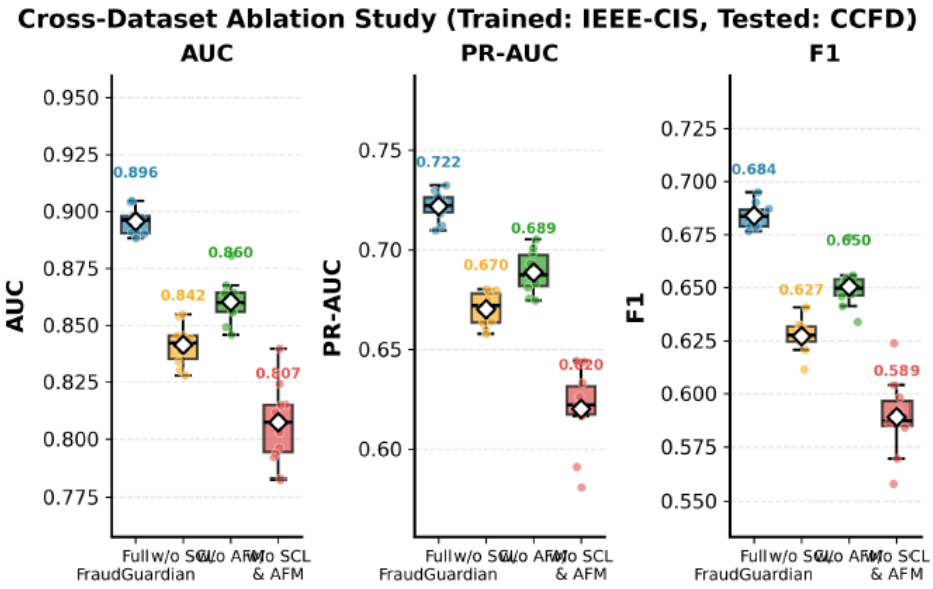

Figure 10. Cross-dataset ablation study.

Box plot comparing performance metrics across model variants in a cross-dataset setting (trained on IEEE-CIS, tested on CCFD). The full model demonstrates superior generalization capability.

6. Evaluation

6.1. Training Dynamics and Convergence Analysis

We analyze the stability and convergence of the proposed FraudGuardian framework by monitoring the training loss and validation AUC. The training curve for the model trained on the CCFD dataset shows that the total loss L_{total} decreases smoothly and converges after approximately 40 epochs. The Adaptive Optimization strategy ensures balanced updates between the Self-supervised Consistency Learning (SCL) loss L_{SCL} and the Adversarial Feature Mining (AFM) loss L_{AFM}, preventing either objective from dominating and causing instability. The validation AUC rises rapidly in the first 15 epochs and then plateaus at a high value, indicating efficient learning and robust generalization. In contrast, the baseline Transformer model shows signs of overfitting, with validation AUC declining after epoch 30. FraudGuardian maintains stable, high validation performance, demonstrating the regularization effect of its SCL and AFM components. This analysis confirms that our adaptive multi-task optimization facilitates stable and efficient convergence, which is crucial for learning from imbalanced financial data.

6.2. Case Study and Interpretability Analysis

We conduct qualitative case studies to evaluate the interpretability of FraudGuardian on financial fraud detection tasks. We analyze high-risk transactions flagged by FraudGuardian on the IEEE-CIS dataset. Using gradients of the SCL loss with respect to input features, we identify specific transaction events and feature dimensions that contribute most to the "inconsistency" score.

For instance, in a "card-not-present" fraud case, the model highlighted an unusually rapid sequence of high-value online transactions from geographically disparate locations following a change in the user's contact email-a pattern violating the learned local consistency of normal user behavior. In another case involving "counterfeit" fraud, FraudGuardian flagged a series of point-of-sale transactions with abnormal timing patterns and merchant category codes inconsistent with the cardholder's historical profile.

This capability to pinpoint suspicious local patterns provides valuable interpretability for fraud investigators, enabling them to quickly understand why a transaction was flagged and prioritize their review efforts accordingly.

6.3. Feature Space Visualization

To gain deeper insight into the learned representations, we employ t-SNE to visualize the global transaction embeddings g for a subset of the CCFD test set. Compared to the baseline Transformer, the feature clusters produced by FraudGuardian show clearer separation between legitimate (class 0) and fraudulent (class 1) transactions. Furthermore, the fraud class embeddings are more tightly clustered, indicating that the model has learned a more compact and discriminative representation for diverse fraudulent patterns. The Adversarial Feature Mining (AFM) component is observed to push the embeddings of legitimate samples near the decision boundary slightly away from the fraud cluster, effectively "hardening" the boundary. This visual evidence aligns with the quantitative performance gains and supports our claim that AFM improves generalization by learning a more robust feature space.

7. Conclusion

This paper introduced FraudGuardian, a novel deep learning framework designed to tackle key challenges in fraud detection, including severe class imbalance, evolving adversarial tactics, and the need for interpretable predictions. Our core contribution comprises two synergistic learning mechanisms: Selfsupervised Consistency Learning (SCL), which models robust local event-level patterns without requiring fine-grained labels, and

Adversarial Feature Mining (AFM), which synthesizes challenging representations to enhance generalization. These components are harmonized through an adaptive multi-task optimization strategy that dynamically balances learning objectives and prioritizes hard-toclassify samples.

Extensive experiments on financial fraud detection benchmarks demonstrate the effectiveness and strong generalization of FraudGuardian. It consistently outperformed state-of-the-art baselines across multiple benchmarks (IEEE-CIS, PaySim, CCFD). For instance, on the challenging CCFD dataset, it achieved 97.9% AUC, 90.6% PR-AUC, and 86.2% F1-

Score. Notably, FraudGuardian exhibited superior performance in cross-attack-type scenarios, significantly surpassing baselines in detecting novel fraud strategies (e.g., a 9.1% F1-Score improvement over TABL) and attaining higher recall for the minority fraud class under extreme imbalance.

A thorough ablation study confirmed the contribution of each component. The SCL and AFM modules individually provided significant gains (+1.3% and +1.9% AUC, respectively), while their full integration with adaptive optimization yielded the best performance, underscoring their complementary nature. Further analysis of AFM sub-components and the adaptive strategy revealed that both adversarial feature generation with consistency regularization and importance-aware sampling with uncertainty-based weighting are essential for optimal performance. Supplementary analyses on training stability, case studies, and feature visualizations offered additional evidence of the model's robust convergence and improved interpretability.

Collectively, the results affirm that jointly optimizing local pattern consistency and global adversarial robustness, guided by adaptive learning strategies, constitutes a powerful paradigm for building reliable fraud detection systems. This work provides a generalizable framework effective across diverse sequential data domains characterized by imbalance and adversarial drift.

References

1. X. Tan, Y. Ma, and X. Zhang, "Understanding structured financial data with LLMs: A case study on fraud detection," *arXiv preprint arXiv:2512.13040*, 2025.
2. O. Kulkarni, and R. Chandra, "DynBerg: Dynamic BERT-based graph neural network for financial fraud detection," *arXiv preprint arXiv:2511.00047*, 2025.
3. X. Li, "Financial fraud identification and interpretability study for listed companies based on convolutional neural network," *arXiv preprint arXiv:2512.06648*, 2025.
4. Y. Korkmaz, J. N. Paranjape, C. M. de Melo, and V. M. Patel, "Referring change detection in remote sensing imagery," *arXiv preprint arXiv:2512.11719*, 2025.
5. S. Deng, E. E. Kosloski, S. S. N. Vasireddy, J. Li, R. S. Sherwood, F. M. Hatha, S. Patel, P. R. Rollins, and Y. Tian, "Toward gaze target detection of young autistic children," *arXiv preprint arXiv:2511.11244*, 2025.
6. K. Lu, M. Huo, Y. Li, Q. Zhu, and Z. Chen, "CtPatchTST: Channel-time patch time-series transformer for long-term renewable energy forecasting," In *2025 10th International Conference on Computer and Information Processing Technology (ISCIPT)*, 2025, pp. 86-95.
7. Q. Zhu, K. Lu, M. Huo, and Y. Li, "Image-to-image translation with diffusion transformers and CLIP-based image conditioning," In *2025 6th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, 2025, pp. 626-632. doi: 10.1109/cvidl65390.2025.11085477
8. H. Qi, Z. Hu, Z. Yang, J. Zhang, J. J. Wu, C. Cheng, C. Wang, and L. Zheng, "Capacitive aptasensor coupled with microfluidic enrichment for real-time detection of trace SARS-CoV-2 nucleocapsid protein," *Analytical Chemistry*, vol. 94, no. 6, pp. 2812-2819, 2022. doi: 10.1021/acs.analchem.1c04296
9. Y. Ang, P. Yao, Y. Bao, Y. Feng, Q. Huang, A. K. H. Tung, and Z. Huang, "RFOD: Random forest-based outlier detection for tabular data," *arXiv preprint arXiv:2510.08747*, 2025.
10. Z. L. Wei, H. Y. An, Y. Yao, W. C. Su, G. Li, S. Saifullah, and W. B.-F., "FSTGAT: Financial spatio-temporal graph attention network for non-stationary financial systems and its application in stock price prediction," *Symmetry*, vol. 17, no. 8, p. 1344, 2025.
11. X. Wu, H. Wang, W. Tan, D. Wei, and M. Shi, "Dynamic allocation strategy of VM resources with fuzzy transfer learning method," *Peer-to-Peer Networking and Applications*, vol. 13, no. 6, pp. 2201-2213, 2020. doi: 10.1007/s12083-020-00885-7
12. S. Castro, A. Betlei, T. D. Martino, and N. E. Manouzi, "Abacus: Self-supervised event counting-aligned distributional pretraining for sequential user modeling," *arXiv preprint arXiv:2512.16581*, 2025.

13. C. H. Pan, Y. Qu, Y. Yao, and M. J. S. Wang, "HybridGNN: A self-supervised graph neural network for efficient maximum matching in bipartite graphs," *Symmetry*, vol. 16, no. 12, p. 1631, 2024.
14. S. Lin, "Abductive inference in retrieval-augmented language models: Generating and validating missing premises," *arXiv preprint arXiv:2511.04020*, 2025. doi: 10.1109/icncis67521.2025.11296031
15. M. Wang, W. Yang, and S. Wang, "Conditional matching preclusion number for the Cayley graph on the symmetric group," *Acta Mathematicae Applicatae Sinica (Chinese Series)*, vol. 36, no. 5, pp. 813-820, 2013.
16. D. Qu, and Y. Ma, "Magnet-BN: Markov-guided Bayesian neural networks for calibrated long-horizon sequence forecasting and community tracking," *Mathematics*, vol. 13, no. 17, p. 2740, 2025. doi: 10.20944/preprints202507.0725.v1
17. H. S. Chauhan, and Z. S. Abdallah, "A regime-aware fusion framework for time series classification," *arXiv preprint arXiv:2512.15378*, 2025.
18. Darbinyan S K. On Hamiltonian and Hamilton-connected digraphs[J]. arXiv preprint arXiv:1801.05166, 2018.
19. J. Al-Karaki, M. A. Z. Khan, and R. D. M. A. Athamneh, "Phantom: Progressive high-fidelity adversarial network for threat object modeling," *arXiv preprint arXiv:2512.15768*, 2025.
20. S. Wang, J. Wangmu, Z. Qi, and Y. Ren, "Embedding paths into the 4-ary n-cube with faulty nodes," In *2011 International Conference on Consumer Electronics, Communications and Networks (CECNet)*, 2011, pp. 4949-4951. doi: 10.1109/cecnet.2011.5768403
21. Zhang S, Wang Z. Scattering number in graphs[J]. Networks: An International Journal, 2001, 37(2): 102-106.
22. X. Wu, Y. Zhang, M. Shi, P. Li, R. Li, and N. N. Xiong, "An adaptive federated learning scheme with differential privacy preserving," *Future Generation Computer Systems*, vol. 127, pp. 362-372, 2022. doi: 10.1016/j.future.2021.09.015
23. H. Wang, X. Zhang, Y. Xia, and X. Wu, "An intelligent blockchain-based access control framework with federated learning for genome-wide association studies," *Computer Standards & Interfaces*, vol. 84, p. 103694, 2023. doi: 10.1016/j.csi.2022.103694
24. T. Wang, S. Chen, Y. Wang, Y. Zhang, X. Song, Z. Bi, M. Liu, Q. Niu, J. Liu, P. Feng, X. Sun, B. Peng, C. Zhang, K. Chen, M. Li, C. Fei, and L. K. Yan, "From in silico to in vitro: A comprehensive guide to validating bioinformatics findings," *arXiv preprint arXiv:2502.03478*, 2025.
25. X. Deng, "Enhancing neural network performance on tabular data via knowledge distillation and RankGauss transformation," In *2025 6th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, 2025, pp. 418-423. doi: 10.1109/icbase66587.2025.11181427
26. L. K. Q. Yan, Q. Niu, M. Li, Y. Zhang, C. H. Yin, C. Fei, B. Peng, Z. Bi, P. Feng, K. Chen, T. Wang, Y. Wang, S. Chen, M. Liu, J. Liu, X. Song, R. Bao, Z. Jiang, and Z. Qin, "Large language model benchmarks in medical tasks," *arXiv preprint arXiv:2410.21348*, 2025.
27. X. Wu, Y. T. Zhang, K. W. Lai, M. Z. Yang, G. L. Yang, and H. H. Wang, "A novel centralized federated deep fuzzy neural network with multiobjectives neural architecture search for epistatic detection," *IEEE Transactions on Fuzzy Systems*, vol. 33, no. 1, pp. 94-107, 2024.
28. X. Deng, "Graph inference towards ICD coding," *arXiv preprint arXiv:2601.07496*, 2026. doi: 10.1109/aiac68175.2025.11332401
29. Q. Niu, K. Chen, M. Li, P. Feng, Z. Bi, L. K. Yan, Y. Zhang, C. H. Yin, C. Fei, J. Liu, B. Peng, T. Wang, Y. Wang, S. Chen, and M. Liu, "From text to multimodality: Exploring the evolution and impact of large language models in medical practice," *arXiv preprint arXiv:2410.01812*, 2024.
30. T. Wang, M. Liu, B. Peng, X. Song, C. Zhang, X. Sun, Q. Niu, J. Liu, S. Chen, K. Chen, M. Li, P. Feng, Z. Bi, Y. Wang, Y. Zhang, C. Fei, and L. K. Yan, "From bench to bedside: A review of clinical trials in drug discovery and development," *arXiv preprint arXiv:2412.09378*, 2024.
31. Y. Zhang, N. Deng, X. Song, Z. Bi, T. Wang, Z. Yao, K. Chen, M. Li, Q. Niu, J. Liu, B. Peng, S. Zhang, M. Liu, L. Zhang, X. Pan, J. Wang, P. Feng, Y. Wen, L. K. Yan, H. Tseng, Y. Zhong, Y. Wang, Z. Qin, B. Jing, J. Yang, J. Zhou, C. X. Liang, and J. Song, "Advanced deep learning methods for protein structure prediction and design," *arXiv preprint arXiv:2503.13522*, 2025.
32. L. Yu, X. Han, Y. Kang, C. Y. Tseng, D. Zhang, Z. Bi, and Z. Han, "Affective multimodal agents with proactive knowledge grounding for emotionally aligned marketing dialogue," *arXiv preprint arXiv:2511.21728*, 2025.
33. Z. Bi, L. Chen, J. Song, H. Luo, E. Ge, J. Huang, T. Wang, K. Chen, C. X. Liang, and Z. Wei, "Exploring efficiency frontiers of thinking budget in medical reasoning: Scaling laws between computational resources and reasoning quality," *arXiv preprint arXiv:2508.12140*, 2025.
34. M. Khan, M. Vatsa, K. Singh, and R. Singh, "NutriScreener: Retrieval-augmented multi-pose graph attention network for malnourishment screening," *arXiv preprint arXiv:2511.16566*, 2025.
35. D. Xiang, and S. Y. Hsieh, "G-good-neighbor diagnosability under the modified comparison model for multiprocessor systems," *Theoretical Computer Science*, vol. 1028, p. 115027, 2025.
36. W. You, Z. Yu, Z. Han, X. Liu, and Y. Zhang, "Large language models for enhanced user experience in virtual and augmented reality: A comprehensive framework for ranking and recommendation systems," *SSRN 5964834*, 2025. doi: 10.2139/ssrn.5964834
37. Z. Bai, E. Ge, and J. Hao, "Multi-agent collaborative framework for intelligent IT operations: An AOI system with context-aware compression and dynamic task scheduling," *arXiv preprint arXiv:2512.13956*, 2025.
38. X. Han, X. Gao, X. Qu, and Z. Yu, "Multiagent medical decision consensus matrix system: An intelligent collaborative framework for oncology MDT consultations," *arXiv preprint arXiv:2512.14321*, 2025.
39. Y. Wang, "Zynq SoC-based acceleration of retinal blood vessel diameter measurement," *Archives of Advanced Engineering Science*, pp. 1-9, 2025. doi: 10.47852/bonviewaaes52023879

40. Z. Wei, P. Hu, S. Lang, H. Yan, L. Mei, Y. Zhang, C. Yang, J. Hao, and Z. Han, "Automated red-teaming framework for large language model security assessment: A comprehensive attack generation and detection system," *arXiv preprint arXiv:2512.20677*, 2025.
41. Y. Wang, "Low-power design of advanced image processing algorithms under FPGA in real-time applications," In *2024 IEEE 4th International Conference on Power, Electronics and Computer Applications (ICPECA)*, 2024, pp. 1080-1084. doi: 10.1109/icpeca60615.2024.10471036
42. X. Wu, Z. Li, H. Cheng, X. Qiu, J. Hu, C. Guo, and B. Yang, "Unlocking the power of mixture-of-experts for task-aware time series analytics," *arXiv preprint arXiv:2509.22279*, 2025.
43. Y. He, S. Li, K. Li, J. Wang, B. Li, T. Shi, Y. Xin, K. Li, J. Yin, and M. Zhang, "GeAdapter: A general and efficient adapter for enhanced video editing with pretrained text-to-image diffusion models," *Expert Systems with Applications*, 2025.
44. Y. Xin, J. Du, Q. Wang, Z. Lin, and K. Yan, "VMT-Adapter: Parameter-efficient transfer learning for multi-task dense scene understanding," In *Proceedings of the AAAI Conference on Artificial Intelligence, 38(14), 16 085-16 093.*, 2024. doi: 10.1609/aaai.v38i14.29541
45. Y. Wang, and S. Sayil, "Soft error evaluation and mitigation in gate diffusion input circuits," In *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 2024, pp. 121-128. doi: 10.1109/icpics62053.2024.10796093
46. D. Zheng, M. Huang, B. Jiang, H. Hu, and X. Chen, "Towards lossless ultimate vision token compression for VLMs," *arXiv preprint arXiv:2512.09010*, 2025.
47. Y. Zhou, Y. He, Y. Su, S. Han, J. Jang, G. Bertasius, M. Bansal, and H. Yao, "ReAgent-V: A reward-driven multi-agent framework for video understanding," *arXiv preprint arXiv:2506.01300*, 2025.
48. Y. Xin, Q. Qin, S. Luo, K. Zhu, J. Yan, Y. Tai, J. Lei, Y. Cao, K. Wang, and Y. Wang, "Lumina-DiMoo: An omni diffusion large language model for multi-modal generation and understanding," *arXiv preprint arXiv:2510.06308*, 2025.
49. Y. Xin, J. Yan, Q. Qin, Z. Li, D. Liu, S. Li, V. S. J. Huang, Y. Zhou, R. Zhang, and L. Zhuo, "Lumina-mGPT 2," *0: Stand-alone autoregressive image modeling. arXiv preprint arXiv:2507.17801*, 2025.
50. Z. Cao, Y. He, A. Liu, J. Xie, Z. Wang, and F. Chen, "CoFi-Dec: Hallucination-resistant decoding via coarse-to-fine generative feedback in large vision-language models," In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, p. 10.
51. Z. Cao, Y. He, A. Liu, J. Xie, Z. Wang, and F. Chen, "Purifygen: A risk-discrimination and semantic-purification model for safe text-to-image generation," In *Proceedings of the 33rd ACM International Conference on Multimedia*, October, 2025, pp. 816-825. doi: 10.1145/3746027.3754595
52. N. Foroutan, J. Saydaliev, Y. E. Kim, and A. Bosselut, "ConLiD: Supervised contrastive learning for low-resource language identification," *arXiv preprint arXiv:2506.15304*, 2025.
53. Y. Tian, Z. Yang, C. Liu, Y. Su, Z. Hong, Z. Gong, and J. Xu, "CenterMambaSAM: Center-prioritized scanning and temporal prototypes for brain lesion segmentation," *arXiv preprint arXiv:2511.01243*, 2025.
54. D. X. Huang, X. H. Zhou, M. J. Gui, X. L. Xie, S. Q. Liu, S. Y. Wang, T. Y. Xiang, R. Z. Ma, N. F. Xiao, and Z. G. Hou, "VasoMIM: Vascular anatomy-aware masked image modeling for vessel segmentation," *arXiv preprint arXiv:2508.10794*, 2025.
55. Y. Huang, W. Li, M. Zhang, X. Zhang, X. You, and M. Yang, "3D-ANC: Adaptive neural collapse for robust 3D point cloud recognition," *arXiv preprint arXiv:2511.07040*, 2025.
56. V. T. A. Khuong, L. T. Nguyen, T. H. Le, and T. D. Ngo, "Improving micro-expression recognition with phase-aware temporal augmentation," In *2025 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 2025, pp. 1-6. doi: 10.1109/mapr67746.2025.11133803
57. M. Yang, W. Quan, and M. Wei, "Perceive, act and correct: Confidence is not enough for hyperspectral classification," *arXiv preprint arXiv:2511.10068*, 2025.
58. S. Xu, H. L. Kao, T. Xu, H. Zhang, J. Wang, R. Ding, G. Liu, T. Shi, Z. Yu, and G. Pan, "Adaptive detector-verifier framework for zero-shot polyp detection in open-world settings," *arXiv preprint arXiv:2512.12492*, 2025. doi: 10.36227/techrxiv.176617650.04595470/v1
59. Z. Yu, J. Wang, and M. Y. I. Idris, "IIDM: Improved implicit diffusion model with knowledge distillation to estimate the spatial distribution density of carbon stock in remote sensing imagery," *Knowledge-Based Systems*, p. 115131, 2025.
60. S. Lin, "Hybrid fuzzing with LLM-guided input mutation and semantic feedback," *arXiv preprint*, arXiv:2511.03995, 2025.
61. Z. Yu, M. Y. I. Idris, P. Wang, Y. Xia, and Y. Xiang, "ForgetMe: Benchmarking the selective forgetting capabilities of generative models," *Engineering Applications of Artificial Intelligence*, vol. 161, p. 112087, 2025.
62. S. Lin, "LLM-driven adaptive source-sink identification and false positive mitigation for static analysis," *arXiv preprint*, arXiv:2511.04023, 2025.
63. X. Wu, J. Dong, W. Bao, B. Zou, L. Wang, and H. Wang, "Augmented intelligence of things for emergency vehicle secure trajectory prediction and task offloading," *IEEE Internet of Things Journal*, vol. 11, no. 22, pp. 36030–36043, 2024.
64. C. Yang, Y. He, A. X. Tian, D. Chen, J. Wang, T. Shi, A. Heydarian, and P. Liu, "WCDT: World-centric diffusion transformer for traffic scene generation," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2025, pp. 6566–6572.
65. M. Z. Hasan and F. Y. Rifat, "Hybrid ensemble of segmentation-assisted classification and GBDT for skin cancer detection with engineered metadata and synthetic lesions from ISIC 2024 non-dermoscopic 3D-TBP images," *arXiv preprint*, arXiv:2506.03420, 2025.

66. J. Yang and K. Yoon, "A multimodal approach to Alzheimer's diagnosis: Geometric insights from cube copying and cognitive assessments," *arXiv preprint*, arXiv:2512.16184, 2025.
67. Z. Yu, "AI for science: A comprehensive review on innovations, challenges, and future directions," *International Journal of Artificial Intelligence for Science (IJAI4S)*, vol. 1, no. 1, 2025.
68. Q. Niu *et al.*, "Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges," *arXiv preprint*, arXiv:2409.02387, 2024.
69. J. T. Kim, A. Sim, K. Wu, and J. Kim, "Improving slow transfer predictions: Generative methods compared," *arXiv preprint*, arXiv:2512.14522, 2025.
70. X. Wu, H. Wang, Y. Zhang, B. Zou, and H. Hong, "A tutorial-generating method for autonomous online learning," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 1532–1541, 2024.