SPECTRAL TOPOLOGY DIFFUSION ON MEMORY RESIDUALS FOR ANOMALY DETECTION IN DYNAMIC GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Anomaly detection in continuous-time dynamic graphs (CTDGs) aims to identify nodes whose behavioral patterns deviate from the normal majority within an evolving edge stream. Two critical yet underexplored requirements complicate practical application: a model trained on a labeled source domain should generalize to unseen target domains without fine-tuning, and it should maintain stable discrimination across diverse temporal gaps from short-interval bursts to extended-dormancy horizons. Existing methods based on self-supervised contrastive learning or reconstruction objectives tend to overfit domain-specific distributions and couple anomaly scoring with the immediate temporal context, limiting transferability and temporal robustness. We propose TopoDiff, a topology-aware diffusion framework that models the generative distribution of memory residuals, defined as the difference between a node's posterior and prior memory representations after an interaction. For normal interactions, these residuals empirically exhibit structured, topology-conditioned patterns that remain more consistent across domains and time scales than raw embeddings, whereas anomalous events produce significant deviations. A stable topology projection built on a Chebyshev polynomial basis over the spectrum of the normalized graph Laplacian yields a positive definite spectral operator, whose inverse is efficiently solved via batched conjugate gradient on event-induced subgraphs. A denoising diffusion process conditioned on the projected topology and filtered residuals learns the normal residual distribution in a latent space. Experiments demonstrate that TopoDiff achieves strong zero-shot cross-domain generalization and stable performance across short-term and long-term horizons.

1 INTRODUCTION

Anomaly detection in continuous-time dynamic graphs (CTDGs) aims to identify nodes whose behavioral patterns deviate from the normal majority within an evolving edge stream (Ares-Robledo et al., 2026; Hua et al., 2026). This task is central to applications ranging from financial fraud detection in transaction networks to identifying malicious accounts in social platforms and spotting anomalous purchasing behaviors in user-product systems (Qiao et al., 2025; Huang et al., 2025). In these scenarios, anomalous events are rare and heterogeneous, and the underlying graph topology evolves continuously, demanding models that process interactions as they arrive and maintain discriminative power across diverse temporal horizons and unseen domains without target annotations (Tailhardat et al., 2026).

Recent advances in dynamic graph representation learning have produced powerful architectures for encoding evolving interactions. Event-driven models maintain per-node memory modules updated at each interaction and combine them with temporal graph attention or recurrent mechanisms to produce time-aware embeddings. Building on these encoders, a growing line of work targets anomaly detection specifically (Kumar et al., 2019; Rossi et al., 2020; Xu et al., 2021). SLADE designs self-supervised tasks that minimize representation drift and reconstruct long-term interaction patterns from short-term ones, flagging nodes whose reconstruction fails (Lee et al., 2024). MHisCL leverages multi-step historical contrastive learning to capture smooth long-term evolutionary patterns of

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

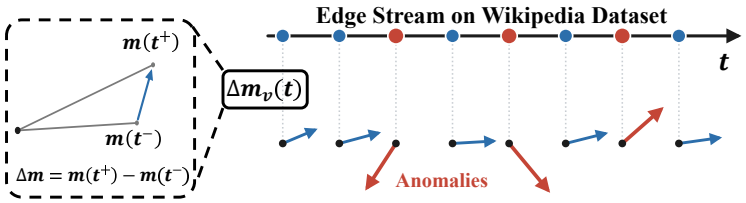


Figure 1: Illustration of memory residuals. Normal interactions (blue) produce small, structured residuals, whereas anomalous events (red) generate significant deviations from this distribution.

normal nodes (Fu et al., 2025b). CTDG-SSM introduces a state-space model with HiPPO-based topology-aware memory for long-range spatio-temporal propagation (Raghuvanshi et al., 2026).

Despite these advances, two critical yet underexplored requirements limit the practical applicability of existing methods. The first is cross-domain generalization (Liu et al., 2025; Zheng et al., 2025; Fu et al., 2025a). A model trained on a labeled source domain is expected to transfer to unseen targets without fine-tuning. This capability is crucial because the validity of anomaly labels is often short-lived in rapidly evolving systems, rendering frequent retraining ineffective. The second is temporal robustness (Chen et al., 2023; Dong et al., 2024; Ares-Robledo et al., 2026). The model should maintain stable discrimination across diverse temporal gaps, from rapid successive events driven by short-term memory to interactions after prolonged dormancy relying on long-term representations.

Existing methods fall short on both fronts. Self-supervised contrastive objectives, as adopted by SLADE and MHisCL, learn representations aligned with the training domain distribution, making them susceptible to domain shift when transferred. Reconstruction-based scoring ties anomaly decisions to the immediate temporal context, degrading performance when the relevant evidence spans long time scales. State-space approaches such as CTDG-SSM address long-range dependency modeling but are designed for predictive tasks rather than anomaly detection, lacking a principled mechanism for characterizing distributional deviations. Moreover, these methods typically couple anomaly scoring with domain-specific learned parameters, limiting zero-shot transferability.

We propose TopoDiff, a topology-aware diffusion framework that addresses these gaps by regularizing a supervised classifier with the generative distribution of memory residuals. As illustrated in Figure 1, a memory residual is the difference between a node's posterior and prior memory representations after an interaction. For normal interactions, these residuals empirically exhibit structured, topology-conditioned patterns that remain more consistent across domains and time scales than raw embeddings, whereas anomalous events produce significant deviations.

The main contributions are summarized as follows:

- We propose a stable topology projection built on a Chebyshev polynomial basis over the spectrum of the normalized graph Laplacian, yielding a provably positive definite spectral operator. Its inverse is solved efficiently via batched conjugate gradient on event-induced subgraphs, avoiding the construction of dense Laplacian matrices and ensuring scalability.
- We design a denoising diffusion process conditioned on the projected topology and filtered residuals to learn the expected normal residual distribution. The diffusion model operates in the latent space of encoded innovations and is conditioned on prior memory states, endpoint topology context, and continuous-time encodings, ensuring that the diffusion prior captures the normal variation given the pre-interaction context.
- We derive a three-component anomaly score from denoising error, reconstruction discrepancy, and topology energy ratio, jointly capturing diverse facets of anomalous behavior.
- Extensive experiments on multiple datasets demonstrate that TopoDiff outperforms strong baselines across standard, zero-shot, as well as short-term and long-term evaluation settings. Our code and all parameters are available in an anonymous repository[1].

[1] https://anonymous.4open.science/r/topo_diff-1D7C/

2 Related Works

Anomaly Detection on CTDG. Anomaly detection in continuous-time dynamic graphs relies on two intertwined capabilities: learning faithful temporal representations and designing effective anomaly scoring mechanisms. Early CTDG representation methods build node states through recurrent or point-process formulations: JODIE (Kumar et al., 2019) couples dual RNNs with a projection operator to forecast future embeddings, DyRep (Trivedi et al., 2019) models interaction dynamics via temporal point processes at multiple time scales, and TGN (Rossi et al., 2020) unifies memory modules with graph-based operators for efficient event-stream processing. Attention-based approaches such as TGAT (Xu et al., 2021) replace recurrence with temporal kernel-weighted neighborhood aggregation, enabling inductive inference on unseen nodes. More recently, CTDG-SSM (Raghuvanshi et al., 2026) derives a HiPPO-based state-space model that compresses the full event history into topology-aware memory states, achieving strong performance on tasks requiring long-range temporal reasoning. These representation backbones have been adapted for anomaly detection through various scoring strategies. SAD (Tian et al., 2023) couples a temporal graph encoder with a time-equipped memory bank and pseudo-label contrastive learning to exploit unlabeled samples. SLADE (Lee et al., 2024) trains a network on representation drift minimization and interaction pattern extrapolation to detect behavioral deviations in edge streams. MHisCL (Fu et al., 2025b) aligns a node's current state with its multi-step historical states through contrastive learning, detecting anomalies whose representations deviate from smooth long-term evolutionary patterns. GeneralDyG (Yang et al., 2025) extracts temporal ego-graph features for generalizable detection across diverse dynamic graphs.

Diffusion Models on Graphs. Diffusion probabilistic models have achieved remarkable success in continuous domains such as image synthesis (Ho et al., 2020; Song et al., 2021). Extending these generative frameworks to discrete and irregular graph structures has recently attracted growing attention. GRAND interprets graph neural networks as discretizations of diffusion partial differential equations, establishing a principled connection between message passing and continuous diffusion processes (Chamberlain et al., 2021). For graph generation, discrete-state continuous-time diffusion preserves the combinatorial nature of graph data while enabling flexible trade-offs between sample quality and computational efficiency (Xu et al., 2024). Topology-aware generation has been further advanced by HOG-Diff, which guides the reverse diffusion process with higher-order structural priors through a coarse-to-fine curriculum (Huang & Birdal, 2026), and MoDiff, which incorporates motif-level spectral information to capture local structural patterns in temporal graphs (Xu & Ma, 2025). Social graph generation has benefited from prior-based discrete diffusion models that redefine the forward process for node-level connection probability learning with social context priors (Yin et al., 2025). More recently, information-geometric frameworks reinterpret diffusion sampling trajectories as parametric curves on Riemannian manifolds to improve structural fidelity (Lu et al., 2026). However, all these methods focus on static or snapshot-based graph generation and do not address anomaly detection in continuous-time event streams.

3 Problem Formulation

Continuous-Time Dynamic Graphs. A CTDG is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{S}, \mathcal{T}, \mathcal{X})$, where $\mathcal{V}$ denotes the node set, $\mathcal{T} \subseteq \mathbb{R}^+$ is the continuous time domain, and $\mathcal{S} = \{(v_i, v_j, t, \mathbf{x}_{ij}^t)\}$ is a chronologically ordered event stream. Each tuple records an interaction from source node v_i to target node v_j at time $t \in \mathcal{T}$ with edge features $\mathbf{x}_{ij}^t \in \mathcal{X}$. We use $\mathcal{S}_{<t}$ to denote the prefix of the event stream consisting of all events strictly before time t.

Event-Induced Subgraph. For an event $(v_i, v_j, t, \mathbf{x}_{ij}^t)$, we construct an event-induced subgraph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ over the active nodes in the temporal neighborhood of the interacting endpoints v_i and v_j up to time t. The node set $\mathcal{V}_t$ contains all nodes that have participated in recent interactions with v_i or v_j, and $\mathcal{E}_t$ collects the corresponding historical edges. This subgraph captures the local topological context available at the moment of the event and serves as the training data for anomaly scoring.

Memory State and Memory Residual. A memory module maintains a per-node state vector $\mathbf{m}_v(t) \in \mathbb{R}^d$ that summarizes the historical interactions of node v up to time t. Upon each event, the memory is updated. We denote the memory state immediately before and af-

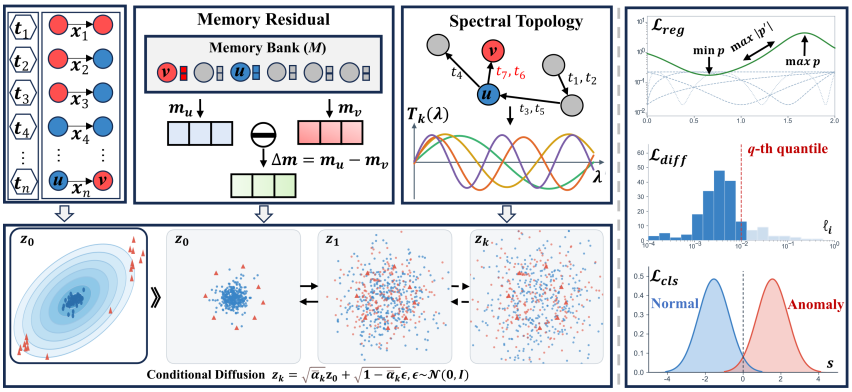


Figure 2: TopoDiff Framework. The model processes continuous edge streams to compute memory residuals representing instantaneous state changes. A stable topology projection built on Chebyshev polynomials captures structural context over events. A conditional denoising diffusion process then learns the normal residual distribution in the latent space, producing complementary anomaly signals from denoising residuals, reconstruction errors, and topology energy ratios to distinguish anomalous behaviors.

ter the update at time t as $\mathbf{m}_v(t^-)$ and $\mathbf{m}_v(t^+)$, respectively, and define the memory residual as $\Delta\mathbf{m}_v(t) = \mathbf{m}_v(t^+) - \mathbf{m}_v(t^-)$, which captures the instantaneous state change induced by the interaction.

Anomaly Detection. The task is dynamic node classification on CTDGs. Given the event history $\mathcal{S}_{<t}$, the objective is to predict a binary label $y \in \{0, 1\}$ for an incoming event $(v_i, v_j, t, \mathbf{x}_{ij}^t)$, where $y = 1$ indicates that the source node v_i exhibits anomalous behavior. The model learns a scoring function f that maps the temporal embedding $\mathbf{z}_{v_i}(t)$, the memory residual $\Delta\mathbf{m}_{v_i}(t)$, and the local topology $\mathcal{G}_t$ to an anomaly score. We impose two requirements on f: (1) cross-domain generalization, where f trained on a source domain should transfer to unseen target domains without fine-tuning; and (2) temporal robustness, where f should maintain stable discrimination across both short-term and long-term temporal horizons.

4 Methodology

In this section, we elaborate on the proposed TopoDiff framework for anomaly detection in CTDGs, as illustrated in Figure 2.

4.1 Temporal Encoder and Memory Residual

For each event (v_i, v_j, t), the encoder constructs the event-induced subgraph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ by collecting the temporal neighbors of the endpoints v_i and v_j from the history $\mathcal{S}_{<t}$. Node and edge features within $\mathcal{G}_t$ are projected to a common hidden dimension d through linear layers, and multi-head temporal graph attention computes the temporal embedding $\mathbf{z}_{v_i}(t) \in \mathbb{R}^d$ for each endpoint. The temporal embedding is then fused with the prior memory states of both endpoints via a linear layer, and a GRU cell transitions the memory from the prior $\mathbf{m}_{v_i}(t^-)$ to the posterior $\mathbf{m}_{v_i}(t^+)$.

We define the *memory residual* as the instantaneous state change induced by the interaction:

$$\Delta\mathbf{m}_{v_i}(t) = \mathbf{m}_{v_i}(t^+) - \mathbf{m}_{v_i}(t^-). \tag{1}$$

We posit that normal residuals follow a structured distribution conditioned on the local topology, while anomalous events produce significant deviations from this distribution. This motivates modeling the normal residual distribution via a denoising diffusion process and detecting anomalies by quantifying how poorly an observed residual is explained by the learned distribution.

4.2 CHEBYSHEV SPECTRAL PROJECTION

We introduce a spectrally stable graph operator that injects topological awareness into anomaly scoring while being invertible by construction. The operator is built on the normalized Laplacian $\mathbf{L}_t = \mathbf{I} - \mathbf{D}_t^{-1/2}\mathbf{A}_t\mathbf{D}_t^{-1/2}$ of the event-induced subgraph $\mathcal{G}_t$, whose historical edges are symmetrized. A Chebyshev polynomial basis is defined over the shifted operator $\tilde{\mathbf{L}}_t = \mathbf{L}_t - \mathbf{I}$, whose spectrum lies in $[-1, 1]$:

$$\begin{aligned} &T_0(\tilde{\mathbf{L}}_t)\mathbf{x} = \mathbf{x}, \quad T_1(\tilde{\mathbf{L}}_t)\mathbf{x} = \tilde{\mathbf{L}}_t\mathbf{x}, \\ &T_k(\tilde{\mathbf{L}}_t)\mathbf{x} = 2\tilde{\mathbf{L}}_t\, T_{k-1}(\tilde{\mathbf{L}}_t)\mathbf{x} - T_{k-2}(\tilde{\mathbf{L}}_t)\mathbf{x}, \quad k \geq 2. \end{aligned} \tag{2}$$

Each application of $\tilde{\mathbf{L}}_t$ to a vector requires only a single sparse aggregation pass over the edges $\mathcal{E}_t$, avoiding any dense matrix construction. A learnable polynomial filter is formed as $q(\mathbf{L}_t)\mathbf{x} = \sum_{k=0}^{K-1} c_k\, T_k(\tilde{\mathbf{L}}_t)\mathbf{x}$ with trainable coefficients $\{c_k\}$. We assemble it into a positive-definite operator:

$$p(\mathbf{L}_t) = \epsilon\mathbf{I} + q(\mathbf{L}_t)^2, \quad \epsilon > 0. \tag{3}$$

Since $\mathbf{L}_t$ is symmetric, $q(\mathbf{L}_t)$ is symmetric for any real coefficients $\{c_k\}$, making $q(\mathbf{L}_t)^2$ positive semidefinite. Adding $\epsilon\mathbf{I}$ guarantees strict positive definiteness and thus the existence and uniqueness of $p(\mathbf{L}_t)^{-1}$ without constraining $\{c_k\}$ (see Proof C.1).

The inverse projection $\mathbf{y} = p(\mathbf{L}_t)^{-1}\mathbf{s}$ for an input signal $\mathbf{s}$ over $\mathcal{V}_t$ is approximated by parallel columnwise conjugate gradient with a fixed number of iterations. Each iteration only applies $p(\mathbf{L}_t)$ to direction vectors through sparse aggregation passes (see Proof C.2).

Beyond the inverse-projected signal, we extract a multiscale topology context $\mathbf{h}^{\text{ms}} = \frac{1}{K}\sum_{k=0}^{K-1} T_k(\tilde{\mathbf{L}}_t)\mathbf{s}$ by averaging all Chebyshev basis terms, where $\mathbf{s}$ denotes the node temporal embeddings. The combined topology context $\mathbf{h}^{\text{topo}} = p(\mathbf{L}_t)^{-1}\mathbf{s} + \mathbf{h}^{\text{ms}}$ is pooled at the source and target nodes to produce the endpoint topology context $\mathbf{h}_{ij}^{\text{topo}}$, which conditions the downstream diffusion process.

4.3 RESIDUAL DIFFUSION

The memory residuals $\Delta\mathbf{m}_{v_i}(t)$ and $\Delta\mathbf{m}_{v_j}(t)$ are scattered to the active nodes in $\mathcal{V}_t$ by averaging endpoint residuals at each node, then filtered through $p(\mathbf{L}_t)$ to produce a topology-filtered residual. The filtered residuals at the source and target endpoints, denoted $\tilde{\Delta}_{v_i}(t)$ and $\tilde{\Delta}_{v_j}(t)$, are pooled and concatenated with the raw residuals to encode a latent representation:

$$\mathbf{z}_0 = \text{MLP}\Big(\Delta\mathbf{m}_{v_i}(t) \parallel \Delta\mathbf{m}_{v_j}(t) \parallel \widetilde{\Delta}_{ij}(t)\Big), \tag{4}$$

where $\widetilde{\Delta}_{ij}(t) = \frac{1}{2}(\tilde{\Delta}_{v_i}(t) + \tilde{\Delta}_{v_j}(t))$ and $\mathbf{z}_0 \in \mathbb{R}^{d_z}$. The concatenation provides the encoder with the raw state changes at both endpoints and their topology-filtered response.

The diffusion process is conditioned on a vector built from the prior memory states, the endpoint topology context, and a continuous time encoding:

$$\mathbf{c} = \text{MLP}\big(\mathbf{m}_{v_i}(t^-) \parallel \mathbf{m}_{v_j}(t^-) \parallel \mathbf{h}_{ij}^{\text{topo}} \parallel \phi(t)\big). \tag{5}$$

By conditioning on the prior state rather than the posterior, the diffusion model learns the distribution of residuals given the historical context. Anomalous events then manifest as residuals that are poorly explained by this conditional model.

We adopt a variance-preserving DDPM objective with a cosine noise schedule $\bar{\alpha}_k$. The forward process corrupts $\mathbf{z}_0$ with Gaussian noise:

$$\mathbf{z}_k = \sqrt{\bar{\alpha}_k}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_k}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{6}$$

A conditional MLP noise predictor $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta(\mathbf{z}_k, k, \mathbf{c})$ estimates the injected noise from the corrupted latent, the diffusion step embedding, and the condition vector. During inference, a deterministic function of the endpoint indices, timestamp, diffusion step, and latent coordinate replaces random sampling, ensuring reproducible scores for identical inputs.

4.4 Anomaly Scoring and Optimization

The anomaly score contains two differently normalized diffusion-error components and one topology-response component.

The *denoising score matching* component measures how poorly the learned diffusion model explains the observed latent:

$$s_1^{(k)} = \frac{1}{d_z} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_k, k, \mathbf{c}) \right\|_2^2 . \tag{7}$$

A high $s_1^{(k)}$ indicates that the residual is unlikely under the learned normal residual distribution.

The *reconstruction* component evaluates the latent recovered from the noisy observation:

$$\hat{\mathbf{z}}_0 = \frac{\mathbf{z}_k - \sqrt{1-\bar{\alpha}_k}\,\hat{\boldsymbol{\epsilon}}}{\sqrt{\bar{\alpha}_k}}, \quad s_2^{(k)} = \frac{1}{d_z} \sum_{r=1}^{d_z} \frac{(z_{0,r} - \hat{z}_{0,r})^2}{\max\{\sigma_r^2, \sigma_{\min}^2\}}, \tag{8}$$

where σ_r^2 is an exponential moving estimate of the variance of latent coordinate r, and $\sigma_{\min}^2 > 0$ ensures numerical stability. At inference, s_1 and s_2 are obtained by averaging $s_1^{(k)}$ and $s_2^{(k)}$ over a fixed set of diffusion steps (see Proof C.4).

The *topology response ratio* compares the filtered and raw endpoint residual magnitudes:

$$s_3 = \frac{\|\tilde{\Delta}_{v_i}(t)\|_2^2 + \|\tilde{\Delta}_{v_j}(t)\|_2^2}{\max\left\{\|\Delta\mathbf{m}_{v_i}(t)\|_2^2 + \|\Delta\mathbf{m}_{v_j}(t)\|_2^2, \eta\right\}}, \tag{9}$$

where $\eta > 0$ absorbs the common per-coordinate normalization. Since $p(\mathbf{L}_t)$ is not constrained to be a contraction, s_3 measures topology response rather than energy retention. With the increasing aggregation below, larger s_3 is treated as stronger anomalous evidence. Its event-level discrimination is an empirical property (see Proof C.3).

The components are aggregated using nonnegative normalized weights $w_r \geq 0$ and $\sum_{r=1}^{3} w_r = 1$:

$$s = w_1 \log(1 + s_1) + w_2 \log(1 + s_2) + w_3 \log(1 + s_3). \tag{10}$$

A parallel classification head f_{cls} maps the temporal embedding $\mathbf{z}_{v_i}(t)$ to a classification logit, while s is projected to a diffusion logit via $g(s) = \text{softplus}(\kappa) \cdot s + b$. The final anomaly probability is

$$\hat{y} = \sigma(f_{\text{cls}}(\mathbf{z}_{v_i}(t)) + \beta\, g(s)), \tag{11}$$

where σ is the sigmoid function and $\beta \geq 0$ controls the contribution of the diffusion signal (The relevant proof is in Appendix C.5).

5 Experiments

5.1 Experiment Setup

We evaluate TopoDiff on dynamic node anomaly detection across six CTDG datasets spanning social (Kumar et al., 2019), financial (Kumar et al., 2016), and user-product (McAuley & Leskovec, 2013) domains. Beyond standard supervised evaluation, we design three additional experimental settings to probe capabilities that distinguish practical anomaly detection systems. A zero-shot cross-domain experiment tests whether a model trained on one domain can generalize to entirely unseen domains without fine-tuning. A short-term and long-term temporal experiment assesses whether the model maintains stable discrimination for events occurring in rapid succession versus those reactivated after extended dormant periods. See Supplementary Materials for more parameters and results.

Baseline models. We compare the performance of TopoDiff against the following state-of-the-art dynamic graph algorithms, namely, JODIE (Kumar et al., 2019), DyRep (Trivedi et al., 2019), TGAT (Xu et al., 2021), TGN (Rossi et al., 2020), SAD (Tian et al., 2023), SLADE (Lee et al., 2024), GeneralDyG (Yang et al., 2025), MHisCL (Fu et al., 2025b), and CTDG-SSM (Raghuvanshi et al., 2026). For static methods, we also consider models DOMINANT (Ding et al., 2019), AnomalyDAE (Fan et al., 2020), and CONAD (Xu et al., 2022) for comparison, where TGAT is employed as a temporal preprocessing step to produce time-aware node embeddings following established practice (Tian et al., 2023).

Table 1: AUROC and AUPRC (in mean ± standard deviation %) values comparison on three benchmark datasets. TLE denotes a time limit of 8 h exceeded. The best performing method is in bold and the second-best method is in underlining.

Methods	Wikipedia		Mooc		Amazon	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
DOMINANT	73.51±2.9	1.42±0.4	50.00±0.2	0.94±0.1	50.01±0.2	15.41±0.1
AnomalyDAE	81.85±1.3	1.82±0.1	50.00±0.1	0.94±0.2	50.20±0.2	15.52±0.1
CONAD	76.21±2.5	0.96±0.3	50.00±0.1	0.94±0.2	50.10±0.1	15.68±0.1
JODIE	87.15±0.7	1.41±0.1	66.72±0.3	2.13±0.1	50.70±0.2	15.73±0.2
Dyrep	84.77±0.3	1.29±0.1	66.34±0.1	2.23±0.2	50.80±0.1	15.71±0.1
TGAT	82.36±1.3	3.34±0.2	68.32±0.9	2.43±0.1	50.49±0.1	15.80±0.1
TGN	88.01±0.4	1.32±0.1	64.56±0.4	1.98±0.1	51.79±0.2	16.19±0.2
SAD	86.82±0.2	1.86±0.2	67.05±0.5	3.24±0.2	52.18±0.3	15.49±0.1
SLADE	87.19±0.1	3.66±0.7	63.07±0.7	1.51±0.1	53.24±0.1	16.72±0.2
GeneralDyG	87.56±0.1	0.14±0.1	TLE	TLE	TLE	TLE
MHisCL	<u>88.61±0.2</u>	2.00±0.2	<u>77.84±0.4</u>	3.47±0.2	<u>55.25±0.1</u>	<u>17.21±0.1</u>
CTDG-SSM	84.84±2.1	<u>3.97±1.1</u>	68.56±0.4	<u>4.49±0.2</u>	50.02±0.1	15.61±0.1
TopoDiff	**90.28±0.1**	**5.65±2.8**	**78.38±0.1**	**6.38±0.2**	**60.33±0.1**	**22.29±0.4**

5.1.1 Experiment Details

The experiments uniformly split the datasets into 70% for training, 15% for validation, and 15% for testing, with the final results reported on the test set. All baselines conduct experiments using the best parameters from the original papers or open-source codes TopoDiff employs Adam (Kingma & Ba, 2015) as the optimizer, and a batch size of 256 during training. Experimental setup comprises an AMD EPYC 7J13 CPU, an NVIDIA GeForce RTX 4090 GPU, and 64GB of RAM.

5.2 Anomaly Detection Results

In Table 1, we present the results with AUROC and AUPRC (see A.5 for more results). It can be seen that on social interaction benchmarks such as Wikipedia and Mooc, our method consistently outperforms state-of-the-art baselines due to the model's ability in characterizing the topology-conditioned distribution of memory residuals via a Chebyshev spectral operator and a denoising diffusion process. Furthermore, TopoDiff exhibits notably strong discrimination on the challenging Amazon, highlighting its ability to capture structural invariants that remain robust across heterogeneous anomaly patterns and domains.

5.3 Zero-Shot and Cross-Domain Study

To assess cross-domain generalization, we conduct a zero-shot experiment in which models pre-trained on a source dataset are directly transferred to previously unseen target datasets without any target-domain fine-tuning. In Table 2, we present the results on seven transfer pairs covering social-to-financial, user-product-to-social, and user-product-to-financial migrations (see A.6 for more results and settings). It can be seen that TopoDiff achieves the highest AUROC across all seven pairs and the highest AUPRC on six of them, with the only exception being Mooc to Bitcoin-OTC where CTDG-SSM slightly outperforms in AUPRC. Notably, on the challenging Amazon-to-financial transfers, TopoDiff achieves 75.33 and 75.11 in AUROC while CTDG-SSM degrades to near-random performance around 50, indicating severe overfitting to domain-specific distributions. Furthermore, TopoDiff exhibits stable cross-domain transferability across all source-target combinations, highlighting that the normal residual distribution conditioned on spectral topology encodes structural invariants that are preserved across heterogeneous domains and time scales.

Table 2: Zero-shot experiments on the dynamic node classification task. A → B means use A's pre-trained model to detect the B dataset. The best performing method is in bold.

Methods	TopoDiff		CTDG-SSM		MHisCL	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Wikipedia → Bitcoin-Alpha	**72.66**	**18.76**	55.29	8.26	65.59	11.44
Wikipedia → Bitcoin-OTC	**71.58**	**22.89**	55.92	13.12	65.25	15.40
Mooc → Bitcoin-Alpha	**56.51**	**11.54**	54.91	8.84	50.85	5.59
Mooc → Bitcoin-OTC	**69.43**	17.53	67.51	**19.37**	57.47	15.70
Amazon → Wikipedia	**78.33**	**1.84**	50.01	0.18	66.95	0.95
Amazon → Bitcoin-Alpha	**75.33**	**20.57**	50.01	6.51	59.58	12.27
Amazon → Bitcoin-OTC	**75.11**	**23.19**	50.02	9.12	56.16	16.28

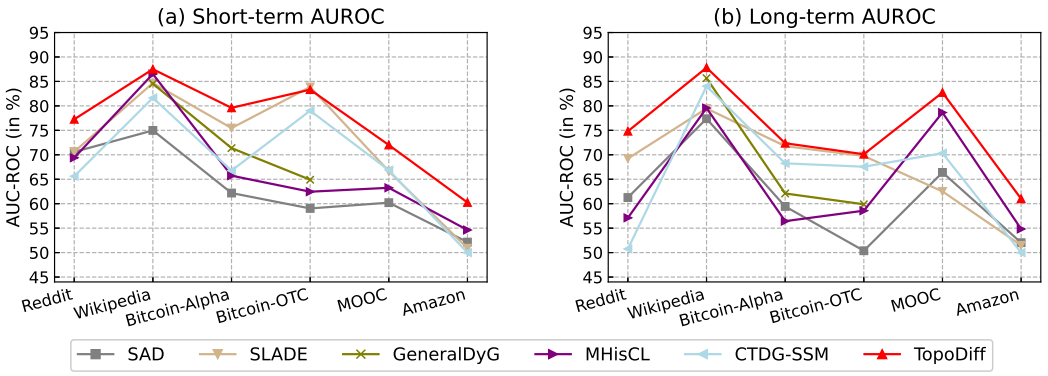


Figure 3: Short-term and long-term experiment.

5.4 Short-Term and Long-Term Study

A practically useful anomaly detector must maintain stable discrimination regardless of when the relevant evidence was last observed. To assess this capability, we partition the test set into short-term and long-term subsets based on the elapsed time $\Delta t = \min(\Delta t_u, \Delta t_v)$ since the last activity of each interacting endpoint, using the 20th percentile as the dividing threshold. The short-term subset captures interactions where at least one endpoint was recently active, testing the model's ability to leverage immediate temporal context. The long-term subset consists of interactions reactivated after extended dormancy, probing whether compressed historical representations remain discriminative.

Figure 3 presents the AUROC on both subsets across all six datasets. TopoDiff achieves the best performance on eleven of the twelve evaluations. On the short-term subset, TopoDiff leads on Reddit, Wikipedia, Bitcoin-Alpha, Mooc, and Amazon, while SLADE narrowly outperforms on Bitcoin-OTC. On the long-term subset, TopoDiff attains the highest AUROC on all six datasets, with particularly large margins on Mooc. Beyond absolute performance, the temporal consistency of TopoDiff is noteworthy. The short-to-long AUROC change on Reddit is only 2.44%, whereas MHisCL drops by 12.28% and CTDG-SSM by 14.84% on the same dataset. More strikingly, TopoDiff improves from short-term to long-term on Mooc by 10.73%, on Wikipedia by 0.32%, and on Amazon by 0.75%, suggesting that the spectral residual distribution becomes better calibrated as more topological context accumulates. This stability confirms that the memory residual distribution conditioned on spectral topology encodes invariants that transcend the temporal distance to the most recent interaction.

Table 3: Model parameters and inference resource usage.

Methods	Parameters	Inference (ms)	Memory (MB)
TGAT	3,033,606	21.57	562.35
TGN	1,275,678	10.09	609.21
JODIE	267,758	5.76	542.40
DyRep	979,150	10.35	606.05
SAD	1,769,988	10.85	743.73
SLADE	1,053,056	5.02	660.72
MHisCL	982,657	7.44	713.27
GeneralDyG	133,379	267.64	630.09
CTDG-SSM	57,275	3.83	848.50
TopoDiff	1,848,326	7.69	644.87

Table 4: Experiments on GPU memory and inference time.

Dataset	Wikipedia	Reddit	Bitcoin-Alpha	Bitcoin-OTC	Mooc	Amazon
GPU (MB)	23.81	25.54	19.47	21.52	22.59	338.35
Inference (ms)	74.54	26.29	33.54	32.88	44.51	267.09

5.5 Performance Analysis

Computational efficiency is a critical consideration for deploying anomaly detectors on high-throughput event streams. Table 3 reports the number of learnable parameters, average inference latency per event, and peak system memory consumption of TopoDiff and all dynamic baselines measured on the same hardware, using a dummy graph with a feature dimension of 172. It can be seen that TopoDiff achieves an inference latency of 7.69 milliseconds per event, which is faster than TGN, DyRep, SAD, and TGAT, and is orders of magnitude faster than GeneralDyG. The parameter count of TopoDiff is comparable to that of SAD and significantly smaller than TGAT, while the peak memory usage is lower than SLADE, MHisCL, SAD , and CTDG-SSM. The efficiency stems from two design choices: the stable topology projection built on a Chebyshev polynomial basis avoids the construction of dense Laplacian matrices and its inverse is solved via batched conjugate gradient on compact event-induced subgraphs, and the denoising diffusion process operates in a low-dimensional latent space rather than the full hidden dimension, keeping the overhead of conditional noise prediction manageable. These results demonstrate that the superior detection accuracy of TopoDiff across standard, zero-shot, and temporal robustness evaluations does not come at the cost of prohibitive computational overhead, making it suitable for real-time anomaly detection in continuously evolving graphs.

We additionally evaluate the GPU memory usage and per-batch inference latency of TopoDiff, using a fixed batch size of 256. Table 4 presents the measurements on all six datasets. The resource usage grows proportionally with the event-induced subgraph size rather than the full graph, confirming that the spectral operator and conjugate gradient solver scale efficiently. These results demonstrate that TopoDiff remains practical for large-scale continuous-time dynamic graphs.

6 Conclusion and Future Work

We proposed TopoDiff, a topology-aware diffusion framework that formulates anomaly detection in continuous-time dynamic graphs as learning the normal distribution of memory residuals. In particular, we introduced a stable spectral operator built on a Chebyshev polynomial basis over the graph Laplacian spectrum, efficiently inverted via batched conjugate gradient on compact event-induced subgraphs. We leveraged this operator to condition a denoising diffusion process. Extensive experiments on multiple datasets demonstrated significant gains in standard detection accuracy, zero-shot cross-domain generalization, and temporal robustness across short-term and long-term horizons.

Future directions include investigating adaptive diffusion step scheduling based on interaction frequency, and applying TopoDiff to real-time detection in large-scale networks.

AI USE STATEMENT

In this work, we used generative AI tools for grammatical corrections and phrasing improvements during the writing process. We have not used generative AI tools for generating novel scientific ideas, designing experiments, or analyzing data, and AI-assisted generation of code, figures, or datasets is not applicable to this work. Additionally, we used generative AI tools for refining sentence structure and word choice in the manuscript. We have reviewed all AI-assisted work. Specifically, all AI-suggested edits to grammar and phrasing were manually examined and approved by the authors to ensure that the original meaning and scientific rigor were preserved; no AI-generated content was adopted without author verification. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of the experimental results, we have open sourced the source code and provided a complete description of the dataset, experimental parameters, proofs, and instructions of the environment Settings in the supplementary material.

REFERENCES

Fernando Ares-Robledo, Helena Rifà-Pous, and Robert Clarisó. Graph neural networks for anomaly detection: A systematic review of dynamic temporal approaches. *Artificial Intelligence Review*, 59(5):129, 2026.

Ben Chamberlain, James Rowbottom, Maria I. Gorinova, Michael Bronstein, Stefan Webb, and Emanuele Rossi. GRAND: Graph neural diffusion. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 1407–1418, 2021.

Nan Chen, Zemin Liu, Bryan Hooi, Bingsheng He, Rizal Fathony, Jun Hu, and Jia Chen. Consistency training with learnable data augmentation for graph anomaly detection with limited supervision. In *The Twelfth International Conference on Learning Representations*, 2023.

Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pp. 594–602, 2019.

Xiangyu Dong, Xingyi Zhang, Lei Chen, Mingxuan Yuan, and Sibo Wang. SpaceGNN: Multi-space graph neural network for node anomaly detection with extremely limited labels. In *The Thirteenth International Conference on Learning Representations*, 2024.

Haoyi Fan, Fengbin Zhang, and Zuoyong Li. Anomalydae: Dual autoencoder for anomaly detection on attributed networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5685–5689, 2020.

Changlong Fu, Zhenli He, Xiong Zhang, Cheng Xie, Xin Jin, and Yun Yang. DR-GGAD: Dual residual centering for mitigating anomaly non-discriminativity in generalist graph anomaly detection. In *The Fourteenth International Conference on Learning Representations*, 2025a.

Yun Fu, Che Zhou, Jintang Li, and Liang Chen. Long-term evolutionary patterns matter: Self-supervised anomaly detection on dynamic graphs. *Knowledge-Based Systems*, 311:113049, 2025b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Fengrui Hua, Yiyan Qi, Zikai Wei, Yuxing Tian, Chengjin Xu, Xiaojun Wu, Jia Li, and Jian Guo. BAG: Benchmarking anomaly detection on dynamic graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(17):14892–14900, 2026.

Haoqi Huang, Ping Wang, Jianhua Pei, Jiacheng Wang, Shahen Alexanian, and Dusit Niyato. Deep learning advancements in anomaly detection: A comprehensive survey. *IEEE Internet of Things Journal*, 12(21):44318–44342, 2025.

Yiming Huang and Tolga Birdal. HOG-diff: Higher-order guided diffusion for graph generation. In *International Conference on Learning Representations*, volume 2026, pp. 111572–111614, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Srijan Kumar, Francesca Spezzano, V. S. Subrahmanian, and Christos Faloutsos. Edge weight prediction in weighted signed networks. In *2016 IEEE 16th International Conference on Data Mining*, pp. 221–230, 2016.

Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1269–1278, 2019.

Jongha Lee, Sunwoo Kim, and Kijung Shin. SLADE: Detecting dynamic anomalies in edge streams without labels via self-supervised learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1506–1517, 2024.

Yunhui Liu, Jiashun Cheng, Yiqing Lin, Qizhuo Xie, Jia Li, Fugee Tsung, Hongzhi Yin, Tao Zheng, Jianhua Zhao, and Tieke He. Towards anomaly-aware pre-training and fine-tuning for graph anomaly detection. In *The Fourteenth International Conference on Learning Representations*, 2025.

Yuhui Lu, Wenjing Liu, and Kun Zhan. Information-geometric adaptive sampling for graph diffusion. In *Forty-Third International Conference on Machine Learning*, 2026.

Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 897–908, 2013.

Hezhe Qiao, Hanghang Tong, Bo An, Irwin King, Charu Aggarwal, and Guansong Pang. Deep graph anomaly detection: A survey and new perspectives. *IEEE Transactions on Knowledge and Data Engineering*, 37(9):5106–5126, 2025.

Ayushman Raghuvanshi, Thummaluru Siddartha Reddy, Sundeep Prabhakar Chepuri, and Mahesh Chandran. Learning long range spatio-temporal representations over continuous time dynamic graphs with state space models. In *Forty-third International Conference on Machine Learning*, 2026.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. *ArXiv preprint arXiv:2006.10637*, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Lionel Tailhardat, Raphael Troncy, and Yoan Chabot. Anomaly detection using knowledge graphs: A survey for network management and cybersecurity application. *ACM Comput. Surv.*, 58(15), August 2026.

Sheng Tian, Jihai Dong, Jintang Li, Wenlong Zhao, Xiaolong Xu, Baokun Wang, Bowen Song, Changhua Meng, Tianyi Zhang, and Liang Chen. SAD: Semi-Supervised Anomaly Detection on Dynamic Graphs. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 2306–2314, 2023.

Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. Dyrep: Learning representations over dynamic graphs. In *International Conference on Learning Representations*, 2019.

Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. A temporal kernel approach for deep learning with continuous-time information. In *International Conference on Learning Representations*, 2021.

Yuwei Xu and Chenhao Ma. MoDiff - graph generation with motif-aware diffusion model. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 3437–3448, 2025.

Zhe Xu, Ruizhong Qiu, Yuzhong Chen, Huiyuan Chen, Xiran Fan, Menghai Pan, Zhichen Zeng, Mahashweta Das, and Hanghang Tong. Discrete-state continuous-time diffusion for graph generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 79704–79740, 2024.

Zhiming Xu, Xiao Huang, Yue Zhao, Yushun Dong, and Jundong Li. Contrastive attributed network anomaly detection with data augmentation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 444–457, 2022.

Xiao Yang, Xuejiao Zhao, and Zhiqi Shen. A generalizable anomaly detection method in dynamic graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22001–22009, 2025.

Shu Yin, Dongpeng Hou, Lianwei Wu, Xianghua Li, and Chao Gao. A prior-based discrete diffusion model for social graph generation. In *Thirty-Fourth International Joint Conference on Artificial Intelligence*, volume 4, pp. 3570–3578, 2025.

Lecheng Zheng, Dongqi Fu, Zihao Li, and Jingrui He. OwlEye: Zero-shot learner for cross-domain graph data anomaly detection. In *The Fourteenth International Conference on Learning Representations*, 2025.

A Additional Experimental Results

A.1 Loss Functions

The classification loss is the standard binary cross entropy over a batch of N events:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right]. \tag{12}$$

The diffusion loss trains the normal residual distribution and is computed exclusively on normal samples ($\mathcal{B}_n$) to prevent contamination by anomalous residuals. A trimmed weighting scheme down-weights potential outliers within the nominally normal set:

$$\mathcal{L}_{\text{diff}} = \frac{\sum_{i\in\mathcal{B}_n} w_i\,\ell_i}{\sum_{i\in\mathcal{B}_n} w_i}, \quad w_i = \min\left(1, \frac{\tau}{\ell_i}\right), \tag{13}$$

where $\ell_i = \|\boldsymbol{\epsilon}_i - \hat{\boldsymbol{\epsilon}}_i\|_2^2$ is the per-sample denoising error and τ is the q-th quantile of $\{\ell_i\}_{i\in\mathcal{B}_n}$. A spectral regularizer penalizes sharp responses in $p(\lambda)$ over a grid of eigenvalues $\lambda \in [0, 2]$, encouraging smooth spectral behavior:

$$\mathcal{L}_{\text{reg}} = \log\left(1 + \frac{\max_\lambda p(\lambda)\cdot\max_\lambda |p'(\lambda)|}{\min_\lambda p(\lambda)^2}\right), \tag{14}$$

where $p'(\lambda)$ is approximated by finite differences. This prevents the learned polynomial from overfitting to noise. The total training objective is:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{cls}} + \lambda_2\mathcal{L}_{\text{diff}} + \lambda_3\mathcal{L}_{\text{reg}}. \tag{15}$$

where λ_1, λ_2, and λ_3 are balancing hyperparameters. The classification loss $\mathcal{L}_{\text{cls}}$ updates both the encoder and the classification head, while $\mathcal{L}_{\text{diff}}$ and $\mathcal{L}_{\text{reg}}$ train only the TopoDiff plug-in parameters (the relevant proof is in Appendix C.6).

A.2 Ablation Experimental Results

We conduct an ablation study to better evaluate the contributions of different components in TopoDiff, as detailed below:

- **W/$\mathcal{L}_{\text{cls}}$:** Remove the loss function $\mathcal{L}_{\text{cls}}$.
- **W/$\mathcal{L}_{\text{diff}}$:** Remove the loss function $\mathcal{L}_{\text{diff}}$.
- **W/$\mathcal{L}_{\text{reg}}$:** Remove the loss function $\mathcal{L}_{\text{reg}}$.
- **W/Diffusion:** Remove the Diffusion Model.
- **W/Subgraph:** Remove the event-induced subgraph.

Table 5: AUROC and AUPRC (in mean ± standard deviation %) values comparison.

Methods	Wikipedia		Mooc		Amazon	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
TopoDiff	**90.28±0.02**	**5.65±2.87**	**78.38±0.11**	**6.38±0.27**	**60.33±0.08**	**22.29±0.45**
W/$\mathcal{L}_{\text{cls}}$	71.58±0.15	0.58±0.01	52.75±0.06	1.67±0.01	59.90±0.06	20.94±0.10
W/$\mathcal{L}_{\text{diff}}$	90.20±0.04	1.87±0.05	77.36±0.15	5.32±0.04	60.27±0.55	19.23±0.16
W/$\mathcal{L}_{\text{reg}}$	90.19±0.03	1.87±0.06	78.20±0.14	5.76±0.06	60.18±0.16	19.29±0.11
W/Diffusion	90.09±0.09	1.91±0.09	78.03±0.07	5.56±0.06	51.79±7.04	15.61±3.22
W/Subgraph	89.91±0.06	3.91±0.45	78.21±0.14	5.47±0.07	50.01±2.66	14.94±0.72

Table 6: AUROC and AUPRC (in mean ± standard deviation %) values comparison.

Methods	Reddit		Bitcoin-Alpha		Bitcoin-OTC	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
TopoDiff	**73.21±0.06**	**0.59±0.51**	**78.87±0.03**	**20.24±0.13**	**77.09±0.05**	**22.44±0.02**
W/$\mathcal{L}_{\text{cls}}$	58.57±0.17	0.16±0.01	63.86±0.03	14.10±0.02	46.32±0.10	11.12±0.04
W/$\mathcal{L}_{\text{diff}}$	66.28±2.87	0.25±0.06	78.72±0.03	20.05±0.04	64.83±6.08	14.90±3.96
W/$\mathcal{L}_{\text{reg}}$	67.88±1.78	0.29±0.08	78.57±0.03	19.84±0.05	67.90±4.78	16.70±3.54
W/Diffusion	67.85±2.77	0.28±0.06	78.81±0.04	19.51±0.06	65.37±6.27	14.28±3.66
W/Subgraph	66.01±4.13	0.24±0.11	77.26±0.02	19.03±0.03	77.01±0.02	22.18±0.03

Tables 5 and 6 report the experimental results. The classification loss $\mathcal{L}_{\text{cls}}$ is the most critical component. Removing it causes the largest degradation across all datasets, with AUROC dropping by 18.70 points on Wikipedia, 25.63 points on Mooc, 14.64 points on Reddit, and 30.77 points on Bitcoin-OTC. This confirms that the supervised classification objective provides the essential discriminative foundation upon which the diffusion-based anomaly scoring builds.

The diffusion model and its associated loss $\mathcal{L}_{\text{diff}}$ contribute most significantly to AUPRC, which is the more informative metric when anomalies are extremely rare. On Amazon, removing the diffusion model reduces AUPRC from 22.29 to 15.61, a 6.68-point drop, and on Bitcoin-OTC, the reduction reaches 8.16 points from 22.44 to 14.28. The impact is also pronounced on Reddit, where removing $\mathcal{L}_{\text{diff}}$ degrades AUROC by 6.93 points. These results validate that the denoising diffusion process conditioned on spectral topology effectively captures the normal residual distribution, and that deviations from this distribution provide complementary anomaly signals beyond what the classifier alone can learn.

The topology regularizer $\mathcal{L}_{\text{reg}}$ provides consistent improvements, particularly on Reddit and Bitcoin-OTC where its removal reduces AUROC by 5.33 and 9.19 points respectively. This confirms that the spectral response sharpness constraint prevents the Chebyshev polynomial coefficients from degenerating and ensures the stability of the topology projection.

The event-induced subgraph construction is especially important for Amazon and Reddit, where its removal reduces AUROC by 10.32 and 7.20 points respectively. This demonstrates that the local topological context captured by $\mathcal{G}_t$ is essential for the spectral operator to produce meaningful

topology-conditioned representations. On Bitcoin-OTC, the subgraph removal has minimal impact (0.08 AUROC drop), suggesting that the relatively dense trust-rating topology already provides sufficient structural information even without localized subgraph extraction.

A.3 Parameter Sensitivity Experimental Results

We study the sensitivity of TopoDiff to four hyperparameters: the classification loss weight λ_1, the diffusion loss weight λ_2, the topology regularizer weight λ_3, and the diffusion output weight β. Each is varied across multiple candidate values with all others fixed at their defaults.

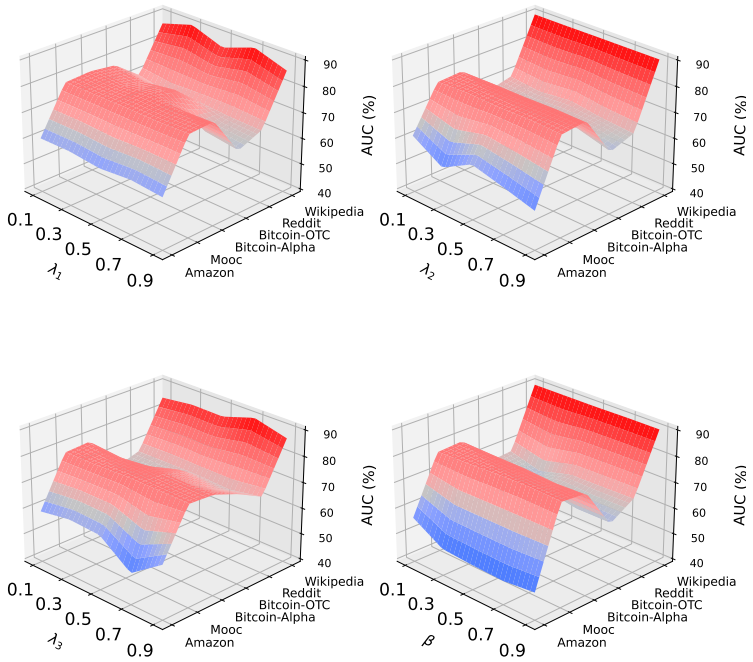


Figure 4: Parameter experiments (AUROC) on different weights.

Figures 4 and 5 report the experimental results. Among the four parameters, λ_1 has the largest impact on performance. Its variation produces AUROC fluctuations of up to 13.63 points on Bitcoin-OTC and 4.48 points on Wikipedia, confirming that the classification loss anchors the discriminative capability of the model. The diffusion loss weight λ_2 shows high stability on Wikipedia where the AUROC varies by less than 0.1 point, but exhibits a 7.20-point range on Amazon where the extreme anomaly scarcity amplifies the sensitivity to the classification-diffusion training balance. The topology regularizer λ_3 is most influential on Bitcoin-OTC, where insufficient or excessive regularization causes AUROC drops of up to 10.14 points, reflecting the delicate trade-off between spectral smoothness and discriminative topological expressiveness. The diffusion output weight β yields stable performance on Wikipedia and Bitcoin-Alpha with fluctuations below 1.1 points, while Amazon shows sensitivity of up to 8.15 points due to the interaction between the noise schedule and severe class imbalance. Across all parameters and datasets, the default configuration consistently achieves near-optimal performance, demonstrating that TopoDiff does not require extensive hyperparameter tuning for practical deployment.

A.4 Few-Data Experimental Results

To evaluate data efficiency, we reduce the training proportion from 70% down to 30% on all six datasets while keeping the test set fixed. Figures 6 and 7 report the AUROC and AUPRC of TopoDiff against seven baselines. TopoDiff consistently achieves the highest AUROC across all training proportions and all datasets. The degradation from 70% to 30% is remarkably small on Bitcoin-OTC

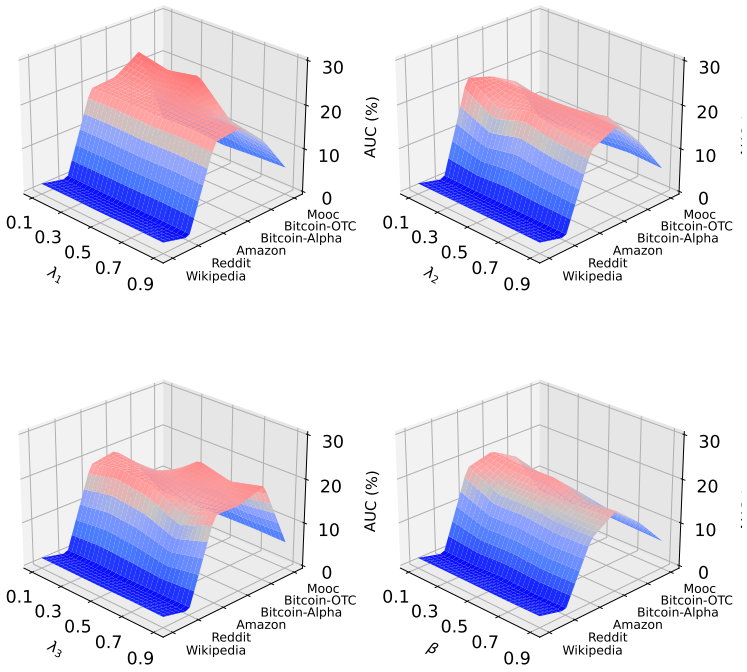


Figure 5: Parameter experiments (AUPRC) on different weights.

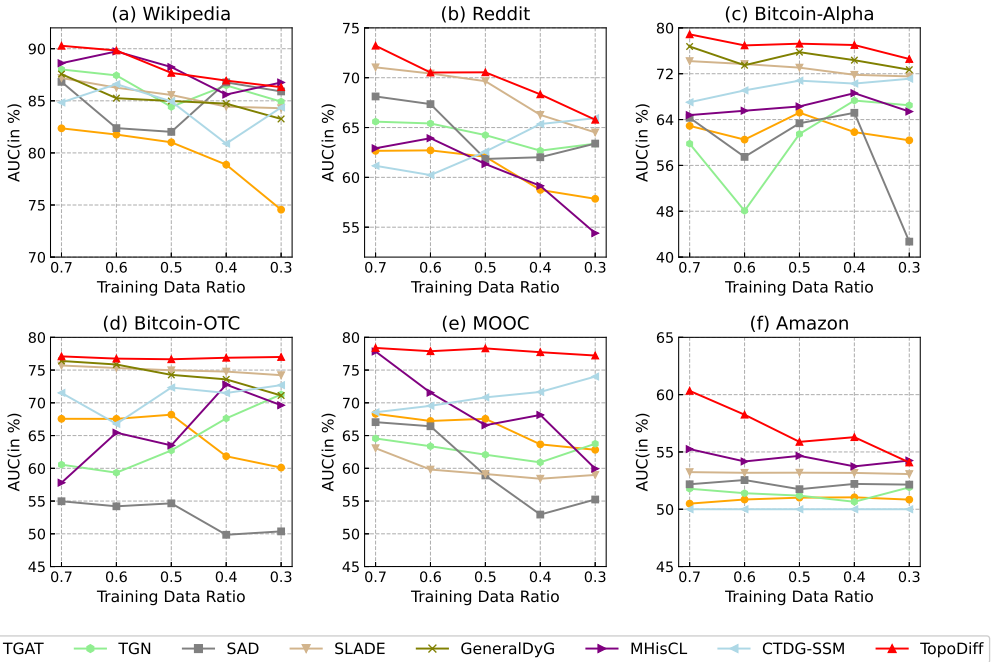


Figure 6: AUROC values under few-data experiment.

and Mooc, while baselines exhibit severe instability: MHisCL collapses by 17.90 points on Mooc and SAD by 21.61 points on Bitcoin-Alpha. On Amazon where anomalies are extremely scarce, TopoDiff retains 54.08 AUROC at 30%, outperforming MHisCL by a comparable margin and all other baselines by more than 2 points. The AUPRC results corroborate this advantage: TopoDiff

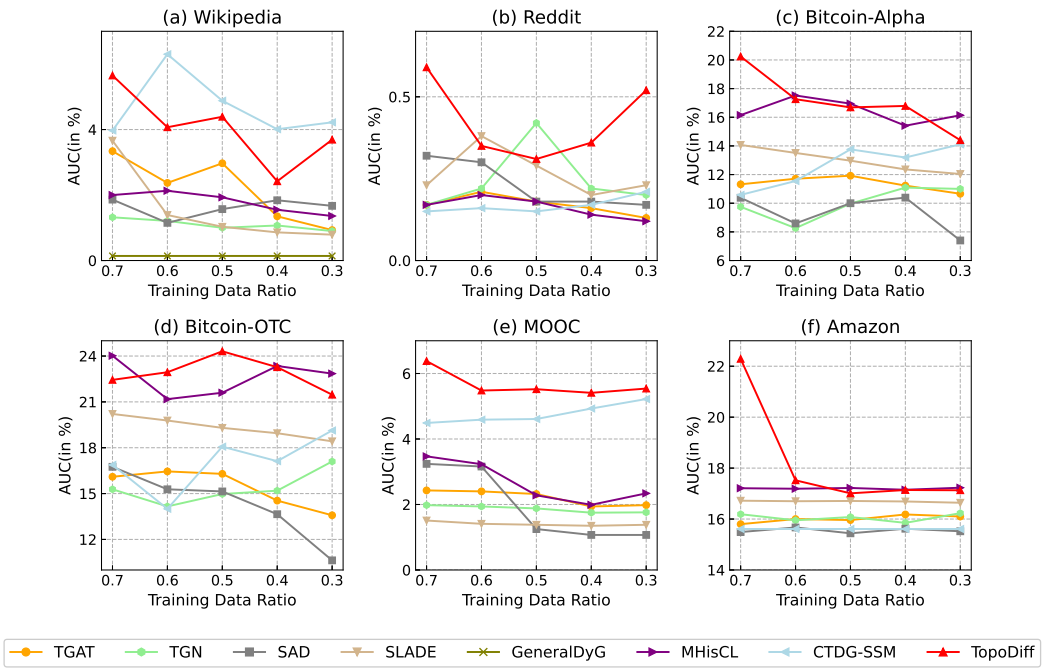


Figure 7: AUPRC values under few-data experiment.

Table 7: AUROC and AUPRC (in mean ± standard deviation %) values comparison on three benchmark datasets. TLE denotes a time limit of 8 h exceeded. The best performing method is in bold and the second-best method is in underlining.

Methods	Reddit		Bitcoin-Alpha		Bitcoin-OTC	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
DOMINANT	57.89±0.6	0.13±0.1	64.52±0.1	10.26±0.1	57.56±1.9	14.16±0.8
AnomalyDAE	55.55±0.4	0.13±0.1	67.02±0.7	12.08±0.2	62.50±1.2	15.06±0.3
CONAD	57.70±0.4	0.12±0.2	62.10±0.3	10.01±0.2	57.73±1.9	14.69±0.7
JODIE	53.73±2.0	0.11±0.2	70.30±2.4	12.77±0.9	67.68±2.9	16.97±2.2
Dyrep	61.34±1.0	0.14±0.1	71.48±1.5	13.73±0.1	70.20±1.7	18.59±1.2
TGAT	62.66±1.5	0.17±0.1	62.90±0.3	11.32±0.1	67.56±0.5	16.10±0.2
TGN	65.59±0.6	0.17±0.2	59.79±5.8	9.75±2.0	60.56±9.3	15.28±3.7
SAD	68.12±2.5	0.32±0.1	64.31±1.5	10.38±0.3	54.97±0.2	16.75±0.2
SLADE	71.04±0.4	0.23±0.1	74.20±0.1	14.07±0.1	75.70±0.1	20.21±0.1
GeneralDyG	TLE	TLE	76.75±0.6	3.73±0.1	76.39±0.4	6.34±0.1
MHisCL	62.92±2.4	0.17±0.1	64.78±8.6	16.15±0.7	57.80±4.6	**24.02±0.4**
CTDG-SSM	61.16±5.1	0.15±0.1	67.02±0.1	10.58±0.1	71.52±0.1	16.88±0.1
TopoDiff	**73.21±0.1**	**0.59±0.5**	**78.87±0.1**	**20.24±0.1**	**77.09±0.1**	22.44±0.1

achieves the highest AUPRC at nearly all settings, with the most pronounced lead on Bitcoin-Alpha (14.40 versus 16.14 for MHisCL at 30%) and Mooc (5.54 versus 5.22 for CTDG-SSM at 30%). These results confirm that the normal residual distribution learned through spectral topology and diffusion captures structural invariants that generalize from limited labeled data.

A.5 Additional Anomaly Detection Results

We additionally report the anomaly detection results on Reddit, Bitcoin-Alpha, and Bitcoin-OTC under an alternative evaluation protocol in Table 7. TopoDiff achieves the highest AUROC on

all three datasets, reaching 73.21 on Reddit, 78.87 on Bitcoin-Alpha, and 77.09 on Bitcoin-OTC, consistently surpassing the second-best methods SLADE (71.04), GeneralDyG (76.75), and GeneralDyG (76.39) by 2.17, 2.12, and 0.70 percentage points respectively. In terms of AUPRC, TopoDiff leads on Reddit (0.59 versus 0.32 for SAD) and Bitcoin-Alpha (20.24 versus 16.15 for MHisCL), while on Bitcoin-OTC, MHisCL attains the highest AUPRC of 24.02 compared to 22.44 for TopoDiff. Notably, GeneralDyG achieves competitive AUROC on Bitcoin-Alpha and Bitcoin-OTC but incurs a time limit exceeded error on Reddit, and its AUPRC values (3.73 and 6.34) remain substantially lower than those of TopoDiff, indicating that its detection confidence is poorly calibrated despite reasonable ranking ability. Static graph methods (DOMINANT, AnomalyDAE, CONAD) uniformly yield near-random AUPRC on Reddit, confirming that static reconstruction objectives are fundamentally inadequate for capturing the evolving interaction patterns that distinguish anomalous behavior. These results further corroborate that the topology-conditioned residual distribution modeled by TopoDiff provides robust and complementary anomaly signals across diverse evaluation settings.

A.6 Additional Zero-Shot and Cross-Domain Study

Zero-shot setting. The model evaluation on the target domain is strictly restricted to a single forward inference pass without any parameter updates or gradient computations, fundamentally eliminating the suspicion of target-domain fine-tuning. To address the heterogeneity in edge feature dimensions across different domains, such as the feature space discrepancies between social networks and financial transaction networks, the model dynamically reconstructs the edge feature projection layer according to the input feature dimension of the target domain after loading the source-domain pretrained weights. Meanwhile, for scenarios where the target domain node scale exceeds that of the source domain, the memory module dynamically expands the node state tensor via zero-padding, ensuring that new nodes possess a uniform initial prior representation before their first interaction.

Table 8: Zero-shot experiments. A $\rightarrow$ B means use A's pre-trained model to detect the B dataset. The best performing method is in bold.

Methods	TopoDiff		CTDG-SSM		MHisCL	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Bitcoin-OTC $\rightarrow$ Bitcoin-Alpha	**77.18**	**18.77**	74.74	12.89	60.09	12.01

Table 9: Zero-shot experiments on GeneralDyG. A $\rightarrow$ B means use A's pre-trained model to detect the B dataset.

Methods	GeneralDyG	
	AUROC	AUPRC
Wikipedia $\rightarrow$ Bitcoin-Alpha	50.01	4.02
Wikipedia $\rightarrow$ Bitcoin-OTC	50.01	8.14
Bitcoin-OTC $\rightarrow$ Bitcoin-Alpha	69.20	3.86

We further evaluate the cross-domain generalization capability of TopoDiff by conducting an additional zero-shot transfer experiment from Bitcoin-OTC to Bitcoin-Alpha. Table 8 reports the results of TopoDiff against CTDG-SSM and MHisCL. TopoDiff achieves the highest performance on both evaluation metrics, demonstrating its ability to transfer the learned normal residual distribution across related but distinct financial trust networks.

A.7 Additional Short-Term and Long-Term Study

We further evaluate the temporal robustness of TopoDiff by reporting the AUPRC on the short-term and long-term subsets, as shown in Figure 8. As a metric that is more sensitive to rare anomalies than AUROC, AUPRC provides a stricter assessment of detection calibration under extreme class

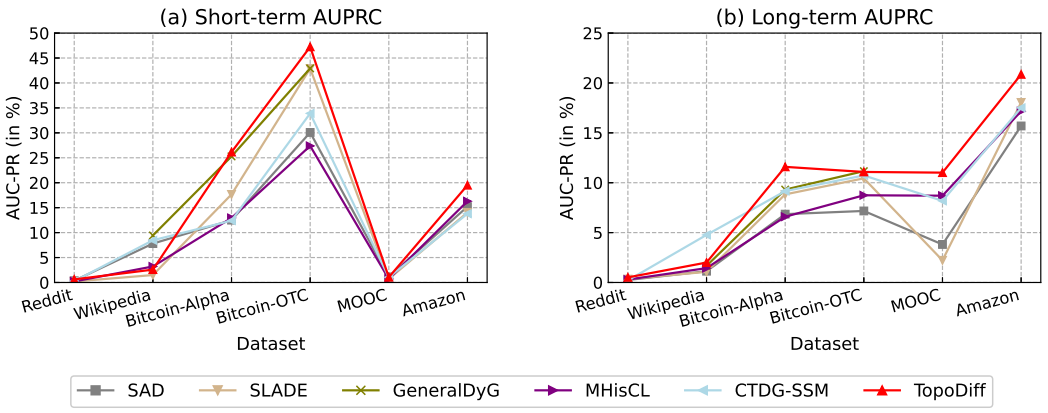


Figure 8: Short-term and long-term experiment (AUPRC).

imbalance. TopoDiff achieves the highest AUPRC on ten out of the twelve subset-dataset combinations, with the only exceptions being the Wikipedia subsets where CTDG-SSM and GeneralDyG perform competitively. The advantages of TopoDiff are particularly pronounced on the financial transaction networks and Amazon, where it substantially outperforms all baselines in both temporal horizons. In contrast, most baselines suffer severe degradations in AUPRC as the elapsed time increases, revealing that they assign high confidence to false positives when immediate temporal context is unavailable. These results confirm that the three-component anomaly score derived from the spectral topology and diffusion process maintains well-calibrated discrimination regardless of the temporal distance to the most recent interaction.

A.8 Additional Performance Analysis

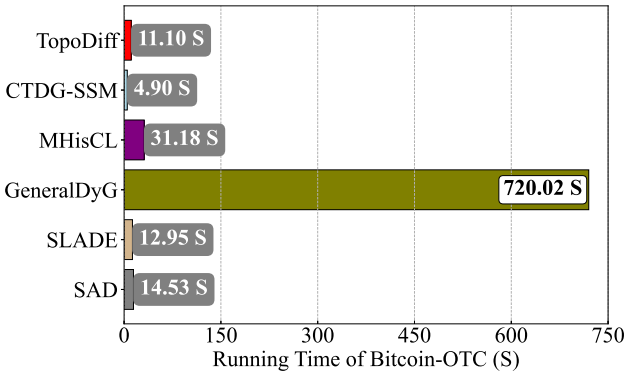


Figure 9: Comparison of training time in Bitcoin-OTC.

We additionally evaluate the training efficiency of TopoDiff by comparing its total training time against five dynamic baselines, as shown in Figure 9. Despite incorporating a diffusion process and spectral projection, TopoDiff maintains a highly competitive training overhead. It trains significantly faster than GeneralDyG, MHisCL and SAD whose complex temporal sampling and multi step contrastive mechanisms incur substantial costs. While CTDG-SSM requires less training time, they suffer severe performance degradation in cross domain and few data settings. These results demonstrate that TopoDiff achieves a superior trade off between detection accuracy and training efficiency.

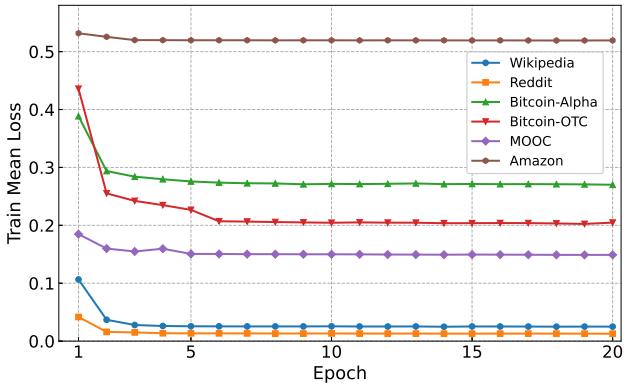


Figure 10: Training loss value evolution.

A.9 Training Loss Analysis

To verify the optimization stability of TopoDiff, we record the training loss over twenty epochs across all six datasets. Figure 10 exhibit a consistent pattern of rapid initial descent followed by a smooth and stable convergence phase. The absence of significant fluctuations throughout the training process indicates that the joint loss formulation yields well behaved gradients, preventing the denoising diffusion process and the spectral topology projection from destabilizing the primary classification objective. Furthermore, the steady convergence observed on the highly imbalanced Amazon dataset demonstrates that the conditional diffusion modeling effectively captures the underlying data distribution without being distracted by rare anomalies.

B Experimental Parameters

B.1 Datasets

Table 10: Statistics of six datasets.

Dataset	#Nodes	#Edges	#Anomalies	Snapshots	Timespan
Wikipedia	9,227	157,474	0.1378%	152,757	30 days
Reddit	10,984	672,447	0.0544%	669,065	30 days
Mooc	7,074	411,749	0.9874%	345,600	30 days
Bitcoin-Alpha	3,783	24,186	3.6136%	1,647	1,901 days
Bitcoin-OTC	5,881	35,592	7.2151%	35,427	1,901 days
Amazon	330,317	568,454	14.4315%	3,168	4,767 days

In the experimental evaluation, we select six representative CTDG datasets to comprehensively examine the performance. They cover three typical application scenarios: three social networks (Wikipedia, Reddit, and Mooc) Kumar et al. (2019), two financial transaction networks (Bitcoin-Alpha and Bitcoin-OTC) Kumar et al. (2016), and one user-product review network (Amazon) McAuley & Leskovec (2013). The key statistics of each dataset are summarized in Table 10, and their specific compositions and anomaly definitions are described below.

Wikipedia records user edits on encyclopedia pages. Each edit constitutes a timestamped interaction, with users as source nodes and pages as destination nodes. Users blocked after editing are labeled as anomalous nodes. **Reddit** contains user posting activities in subreddits. Each post represents an interaction between a user and a subreddit, modeled similarly with users and subreddits; users subject to platform moderation actions are regarded as anomalous. **Mooc** originates from student operation logs in massive open online courses. Each action is treated as an interaction between a student and a course component, and students who drop out of the course are annotated as anomalous. **Bitcoin-Alpha** and **Bitcoin-OTC** are trust rating networks among users, where nodes represent users and directed edges indicate one user's trust rating of another; anomalous nodes are

determined based on abnormal rating patterns. **Amazon** constructs a review graph, where each review record from a user to a product, accompanied by helpfulness vote counts.

C PROOFS

This section provides the guarantees needed for the spectral construction, the conjugate-gradient approximation, and the anomaly score. The analysis applies to every active graph on which the spectral module is evaluated, including a batched union of event neighborhoods.

Let $\mathbf{A}_t = \mathbf{A}_t^\top$ be the nonnegative adjacency matrix obtained after symmetrizing the historical edges, and let $\mathbf{D}_t$ be its degree matrix. Inverse-degree entries are set to zero for isolated nodes. We define

$$\mathbf{L}_t = \mathbf{I} - \mathbf{D}_t^{-1/2}\mathbf{A}_t\mathbf{D}_t^{-1/2}, \qquad \tilde{\mathbf{L}}_t = \mathbf{L}_t - \mathbf{I}. \tag{16}$$

For $K \geq 1$, real coefficients $\{c_k\}_{k=0}^{K-1}$, and $\epsilon > 0$, write

$$\mathbf{Q}_t = q(\mathbf{L}_t) = \sum_{k=0}^{K-1} c_k\, T_k(\tilde{\mathbf{L}}_t), \qquad \mathbf{P}_t = p(\mathbf{L}_t) = \epsilon\mathbf{I} + \mathbf{Q}_t^2. \tag{17}$$

C.1 SPECTRAL STABILITY AND INVERTIBILITY

The spectra of $\mathbf{L}_t$ and $\tilde{\mathbf{L}}_t$ satisfy

$$\sigma(\mathbf{L}_t) \subseteq [0, 2], \qquad \sigma(\tilde{\mathbf{L}}_t) \subseteq [-1, 1]. \tag{18}$$

For every $k \geq 0$,

$$\|T_k(\tilde{\mathbf{L}}_t)\|_2 \leq 1. \tag{19}$$

Moreover, $\mathbf{P}_t$ is symmetric positive definite. If $C_1 = \sum_{k=0}^{K-1} |c_k|$, then

$$\|\mathbf{P}_t^{-1}\|_2 \leq \epsilon^{-1}, \qquad \kappa(\mathbf{P}_t) \leq 1 + \frac{C_1^2}{\epsilon}. \tag{20}$$

Proof. For any vector $\mathbf{v}$, the quadratic form of the normalized Laplacian can be written as

$$\mathbf{v}^\top \mathbf{L}_t \mathbf{v} = \frac{1}{2} \sum_{\substack{u,w: \\ d_u>0,\, d_w>0}} (\mathbf{A}_t)_{uw} \left(\frac{v_u}{\sqrt{d_u}} - \frac{v_w}{\sqrt{d_w}} \right)^2 + \sum_{u: d_u=0} v_u^2. \tag{21}$$

It is nonnegative. Applying $(a-b)^2 \leq 2a^2 + 2b^2$ also gives $\mathbf{v}^\top \mathbf{L}_t \mathbf{v} \leq 2\|\mathbf{v}\|_2^2$. This proves Eq. (18).

Since $|T_k(x)| \leq 1$ for $x \in [-1, 1]$, the spectral theorem gives $\|T_k(\tilde{\mathbf{L}}_t)\|_2 \leq 1$ and hence

$$\|\mathbf{Q}_t\|_2 \leq \sum_{k=0}^{K-1} |c_k| = C_1. \tag{22}$$

The matrix $\mathbf{Q}_t$ is symmetric. Therefore, for every nonzero $\mathbf{v}$,

$$\mathbf{v}^\top \mathbf{P}_t \mathbf{v} = \epsilon\|\mathbf{v}\|_2^2 + \|\mathbf{Q}_t\mathbf{v}\|_2^2 > 0. \tag{23}$$

Every eigenvalue of $\mathbf{P}_t$ lies in $[\epsilon, \epsilon + C_1^2]$, proving Eq. (20). □

The same bounds control the topology context. For a signal $\mathbf{S}$, let

$$\mathbf{Y} = \mathbf{P}_t^{-1}\mathbf{S}, \qquad \mathbf{H}^{\mathrm{ms}} = \frac{1}{K} \sum_{k=0}^{K-1} T_k(\tilde{\mathbf{L}}_t)\mathbf{S}. \tag{24}$$

Then

$$\|\mathbf{Y}\|_F \leq \epsilon^{-1}\|\mathbf{S}\|_F, \qquad \|\mathbf{H}^{\mathrm{ms}}\|_F \leq \|\mathbf{S}\|_F. \tag{25}$$

These bounds hold in exact arithmetic. Rounding error can still accumulate in a high order recurrence.

C.2 CONJUGATE-GRADIENT CONVERGENCE

Let $\mathbf{Y}^\star = \mathbf{P}_t^{-1}\mathbf{S}$, and let $\mathbf{Y}^{(m)}$ be obtained after m independent columnwise conjugate-gradient iterations. In exact arithmetic,

$$\left\|\mathbf{Y}^{(m)} - \mathbf{Y}^\star\right\|_{\mathbf{P}_t,F} \le 2\left(\frac{\sqrt{\kappa_t}-1}{\sqrt{\kappa_t}+1}\right)^m \left\|\mathbf{Y}^{(0)} - \mathbf{Y}^\star\right\|_{\mathbf{P}_t,F}, \tag{26}$$

where

$$\|\mathbf{X}\|^2_{\mathbf{P}_t,F} = \mathrm{tr}(\mathbf{X}^\top \mathbf{P}_t \mathbf{X}), \qquad \kappa_t = \kappa(\mathbf{P}_t) \le 1 + \frac{C_1^2}{\epsilon}. \tag{27}$$

Proof. Proposition C.1 makes $\mathbf{P}_t$ symmetric positive definite, so the standard conjugate-gradient energy-norm bound applies to every right-hand-side column. Squaring the columnwise bounds, summing them, and taking the square root gives Eq. (26). □

For n_t active nodes, e_t stored edges, feature dimension d, polynomial order K, and m iterations, sparse Laplacian multiplication and the Chebyshev recurrence give

$$\mathcal{O}((m+1)K(e_t + n_t)d) \tag{28}$$

time. No dense $n_t \times n_t$ Laplacian or inverse is constructed. A fixed iteration count is an approximation parameter. Positive definiteness guarantees convergence, but does not guarantee that a particular count attains a prescribed error. The approximation can be checked using $\|\mathbf{P}_t\mathbf{Y}^{(m)} - \mathbf{S}\|_F / \|\mathbf{S}\|_F$.

C.3 GLOBAL TOPOLOGY-RESPONSE CHARACTERIZATION

For a nonzero graph residual signal $\mathbf{X} \in \mathbb{R}^{n_t \times d}$, define

$$R(\mathbf{X}) = \frac{\|\mathbf{P}_t\mathbf{X}\|_F^2}{\|\mathbf{X}\|_F^2}. \tag{29}$$

Let $\mathbf{L}_t = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, $\widehat{\mathbf{X}} = \mathbf{U}^\top\mathbf{X}$, and

$$\pi_\ell = \frac{\|\widehat{\mathbf{X}}_{\ell,:}\|_2^2}{\sum_j \|\widehat{\mathbf{X}}_{j,:}\|_2^2}. \tag{30}$$

Then

$$R(\mathbf{X}) = \sum_{\ell=1}^{n_t} \pi_\ell p(\lambda_\ell)^2, \qquad \epsilon^2 \le R(\mathbf{X}) \le (\epsilon + C_1^2)^2. \tag{31}$$

Proof. Functional calculus gives

$$\mathbf{P}_t = \mathbf{U}\,\mathrm{diag}\left(p(\lambda_1), \ldots, p(\lambda_{n_t})\right)\mathbf{U}^\top. \tag{32}$$

Orthogonal invariance of the Frobenius norm proves the expansion in Eq. (31). Its bounds follow from $p(\lambda_\ell) \in [\epsilon, \epsilon + C_1^2]$. □

Section C.3 characterizes a graph-wide response. The event-level topology component is localized. Let $\mathbf{\Delta}_\mathcal{B}$ collect the endpoint residuals in a batch, let $\mathbf{S}_t$ scatter and average them on active nodes, and let $\mathbf{R}_e$ select the two endpoints of event e. After absorbing the common per-coordinate factor into η, the unclipped response has the form

$$\rho_{3,e} = \frac{\|\mathbf{R}_e\mathbf{P}_t\mathbf{S}_t\mathbf{\Delta}_\mathcal{B}\|_F^2}{\max\{\|\mathbf{\Delta}_e\|_F^2, \eta\}}, \qquad s_{3,e} = \min\{\rho_{3,e}, s_{\max}\}. \tag{33}$$

Endpoint selection does not generally commute with $\mathbf{P}_t$, and the denominator is event-specific rather than graph-wide. Therefore, Eq. (31) does not establish an unconditional ordering of event-level scores. The interpretation that a larger s_3 provides stronger anomalous evidence is a learned response-separation hypothesis, consistent with the increasing score aggregation.

C.4 Relation between diffusion score components

At diffusion step k, the reconstruction in Eq. (8) satisfies

$$\mathbf{z}_0 - \hat{\mathbf{z}}_0 = \sqrt{\frac{1-\bar{\alpha}_k}{\bar{\alpha}_k}}\,(\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon})\,. \tag{34}$$

Let $\nu_r = \max\{\sigma_r^2, \sigma_{\min}^2\}$ and suppose $0 < \nu_{\min} \le \nu_r \le \nu_{\max}$. Then

$$\frac{1-\bar{\alpha}_k}{\bar{\alpha}_k \nu_{\max}} s_1^{(k)} \le s_2^{(k)} \le \frac{1-\bar{\alpha}_k}{\bar{\alpha}_k \nu_{\min}} s_1^{(k)}. \tag{35}$$

Proof. Substituting

$$\mathbf{z}_k = \sqrt{\bar{\alpha}_k}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_k}\boldsymbol{\epsilon} \tag{36}$$

into the expression for $\hat{\mathbf{z}}_0$ gives Eq. (34). Consequently,

$$s_2^{(k)} = \frac{1-\bar{\alpha}_k}{\bar{\alpha}_k}\frac{1}{d_z}\sum_{r=1}^{d_z}\frac{(\hat{\epsilon}_r - \epsilon_r)^2}{\nu_r}. \tag{37}$$

Bounding each $1/\nu_r$ between $1/\nu_{\max}$ and $1/\nu_{\min}$ proves Eq. (35). □

Thus, $s_1^{(k)}$ and $s_2^{(k)}$ are unnormalized and variance-normalized views of the same noise-prediction error. Averaging over several fixed diffusion steps changes their relative weighting, but not the per-step identity. During training, Gaussian noise defines the DDPM corruption expectation. The deterministic inference probe preserves the algebraic identity and reproducibility, but is not an unbiased Monte Carlo estimate of that expectation.

C.5 Monotonicity of the anomaly score

Assume $s_r \ge 0$, $w_r \ge 0$, $\sum_{r=1}^{3} w_r = 1$, and $\beta \ge 0$. For

$$s = \sum_{r=1}^{3} w_r \log(1 + s_r) \tag{38}$$

and

$$\hat{y} = \sigma\left(f_{\text{cls}}(\mathbf{z}_{v_i}(t)) + \beta\,(\text{softplus}(\kappa)s + b)\right), \tag{39}$$

$\hat{y}$ is nondecreasing in every s_r. It is strictly increasing when $\beta w_r > 0$ and the corresponding component is not at a clipping boundary.

Proof. Let ℓ denote the argument of the sigmoid. Away from a clipping boundary,

$$\frac{\partial \hat{y}}{\partial s_r} = \hat{y}(1-\hat{y})\,\beta\,\text{softplus}(\kappa)\,\frac{w_r}{1+s_r} \ge 0. \tag{40}$$

The upper clipping of the topology ratio is itself nondecreasing, so it cannot reverse this ordering. □

This proposition explains why the topology component must use the larger-response-is-more-anomalous interpretation under the stated aggregation. If $\beta = 0$, the diffusion score does not affect the final prediction.

For completeness, the optimization objective used in training is

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right], \tag{41}$$

$$\mathcal{L}_{\text{reg}} = \log\left(1 + \frac{\max_\lambda p(\lambda)\max_\lambda |p'(\lambda)|}{\min_\lambda p(\lambda)^2}\right), \tag{42}$$

and

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{cls}} + \lambda_2\mathcal{L}_{\text{diff}} + \lambda_3\mathcal{L}_{\text{reg}}. \tag{43}$$

The derivative $p'(\lambda)$ is approximated by finite differences over a uniform grid on $[0, 2]$, and $\mathcal{L}_{\text{diff}}$ is defined below.

C.6 Effect of detached trimming weights

For a nonempty normal-labeled subset $\mathcal{B}_n$, let $\ell_i = s_{1,i}^{(k)} \geq 0$, let $\tau \geq \xi > 0$ be a detached loss quantile, and define

$$a_i = \min\left(1, \frac{\tau}{\max\{\ell_i, \xi\}}\right), \qquad \mathcal{L}_{\text{diff}} = \frac{\sum_{i\in\mathcal{B}_n} a_i \ell_i}{\sum_{i\in\mathcal{B}_n} a_i}. \tag{44}$$

Each event contributes at most τ to the numerator. Since a_i and τ are detached, the gradient with respect to the diffusion parameters θ is

$$\nabla_\theta \mathcal{L}_{\text{diff}} = \frac{\sum_{i\in\mathcal{B}_n} a_i \nabla_\theta \ell_i}{\sum_{i\in\mathcal{B}_n} a_i}. \tag{45}$$

Proof. If $\ell_i < \xi$, then $\tau \geq \xi$ gives $a_i = 1$ and $a_i \ell_i < \tau$. If $\ell_i \geq \xi$, then $a_i \ell_i = \min\{\ell_i, \tau\} \leq \tau$. Detachment treats the weights and their denominator as constants during backpropagation, which gives Eq. (45). □

The trimming rule reduces the contribution of large-loss nominal events. It does not by itself guarantee robustness to arbitrary label contamination. The spectral regularizer used with this loss discourages sharp responses of $p(\lambda)$, but does not impose a low-pass response, monotonicity, or contraction.

D Visualizations

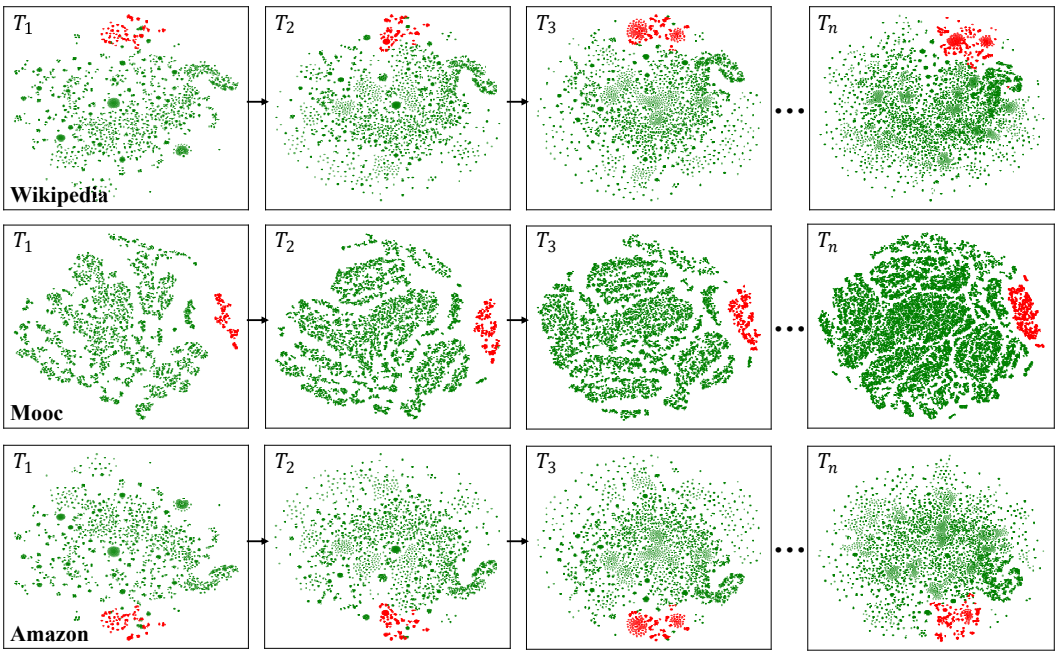


Figure 11: Visualization of T-SNE on different datasets.

We further analyze the quality of the learned node representations by visualizing their T-SNE projections on Wikipedia, Mooc and Amazon at multiple time steps. Figure 11 reveals a distinct structural evolution where normal nodes gradually aggregate into dense and cohesive clusters as interactions accumulate over time. Meanwhile, the anomalous nodes remain distinctly separated from these normal clusters across all temporal snapshots. This persistent spatial isolation confirms that the denoising diffusion process successfully models the topology conditioned normal residual distribution, ensuring that anomalous interactions are mapped to distant regions in the latent space. The uniform separation patterns observed across social, educational and commercial datasets highlight the domain invariant nature of the proposed spectral topology projection.

E LIMITATIONS

Despite the promising empirical results, our work presents a few limitations that outline directions for future research. First, the integration of the denoising diffusion process and the batched conjugate gradient solver introduces additional computational overhead per event. Although event-induced subgraph construction and batched operations significantly mitigate this cost, deploying the framework in ultra-high-throughput streaming environments with strict millisecond-level latency constraints may require further acceleration techniques, such as diffusion distillation. Second, our approach fundamentally relies on historical memory residuals and the local topological context. For cold-start nodes involved in their very first interactions, or in extremely sparse graphs where local neighborhoods are largely disconnected, the memory states and topological signals may lack sufficient information, potentially bounding the detection accuracy in these specific scenarios. Finally, while TopoDiff achieves strong zero-shot generalization on unseen target domains without fine-tuning, the initial training on the source domain still requires careful selection of hyper-parameters, including the diffusion steps, polynomial order, and loss balancing weights, which may incur non-trivial validation costs when establishing the source model for entirely new applications.