TriFusion-IDS: A Multimodal Graph-Tabular-Text Contrastive Framework for Cross-Dataset Intrusion Detection

Qinxin Zhao[1], **Sheng Zhong** [1*]

[1]National Key Lab for Novel Software Technology, Nanjing University
652024320007@smail.nju.edu.cn, zhongsheng@nju.edu.cn

Abstract

Traditional Intrusion Detection Systems (IDS) are typically trained in specific network environments, and their performance often degrades significantly when deployed in new environments with different attack categories. To address this challenge, we propose and define the task of cross-dataset intrusion detection and design a novel multimodal contrastive learning framework named TriFusion-IDS. This framework represents network traffic from three complementary dimensions: a graph view to capture structural communication patterns, a tabular view to model statistical features, and a textual view to define the semantics of attacks. TriFusion-IDS fuses the graph and tabular representations and aligns them with textual descriptions in a shared embedding space using a CLIP-style contrastive loss function. This semantics-based alignment mechanism enables the model to overcome the effects of zero-shot categories and thus generalize to new network environments. Our extensive experiments on several mainstream datasets demonstrate that this method significantly outperforms existing baselines in cross-dataset intrusion detection scenarios.

Introduction

Intrusion Detection Systems (IDS) play a pivotal role in defending computer networks against malicious activities such as denial-of-service (DoS) attacks, brute-force intrusions, and data exfiltration (Talukder et al. 2024). As cyber-attacks grow in scale and sophistication, IDS have become indispensable for maintaining the integrity, availability, and confidentiality of digital infrastructures (Uddin et al. 2025). With the widespread adoption of large-scale and dynamic network infrastructures—such as IoT systems, cloud computing platforms, and 5G-enabled environments—the need for intelligent, adaptive, and robust intrusion detection systems has become increasingly critical to ensure comprehensive network security (Schmitt 2023).

With the rapid development of Graph Neural Networks (GNNs) (Kipf 2016; Jing, Hong, and Tao 2024; Yu et al. 2024), graph-based IDS have also achieved continuous progress. Despite recent advancements, traditional IDS approaches face a fundamental challenge when deployed in real-world scenarios: the zero-shot generalization problem (Sarhan et al. 2023). Existing models typically require a large number of labeled examples from the target deployment environment to maintain performance. However, collecting and annotating network traffic data is labor-intensive, privacy-sensitive, and often infeasible in dynamic or adversarial settings (Rao and Mane 2021). More critically, modern attackers constantly invent novel strategies that may not be represented in the training data, making it impractical to anticipate every threat beforehand.

In this paper, we introduce the task of cross-dataset intrusion detection, which aims to detect unseen attack types in a target network dataset without relying on any labeled examples from that dataset. Instead of assuming access to labeled target-domain traffic, we assume that each attack class is ac-

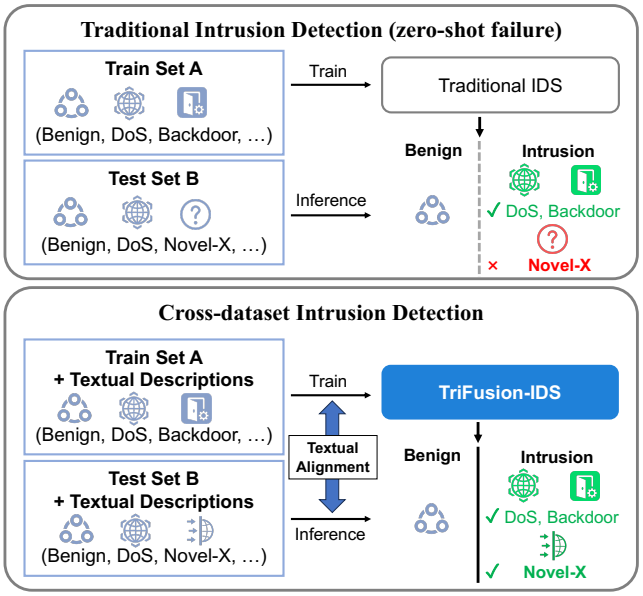

Figure 1: Traditional IDS is trained only on labeled traffic from a source domain, yet fails to recognize the novel attack *Novel-X* when deployed in an unlabeled target domain. Our proposed TriFusion-IDS aligns novel target attack categories with natural-language descriptions, enabling accurate cross-dataset intrusion detection.

*Corresponding Author.

companied by a natural language description capturing its characteristics, underlying mechanisms, or behavioral signatures. The model is trained on a source dataset with both traffic samples and class definitions, and at test time, it must infer the correct attack type for previously unseen traffic instances in the target dataset by aligning them with the semantic descriptions of novel classes. By leveraging such textual definitions, the model is enabled to perform effective anomaly-based intrusion detection across different datasets and previously unseen attack types.

To address this task, we propose TriFusion-ID, a novel contrastive learning framework that aligns multimodal representations of network traffic and attack definitions in a shared embedding space. Our key insight is that traffic samples can be represented from three complementary perspectives: (1) a graph view, capturing structural communication patterns through per-sample traffic graphs; (2) a tabular view, modeling statistical features such as flow duration, packet counts, and protocol types; and (3) a textual view, reflecting semantic descriptions of attack classes. We encode the graph and tabular features through separate neural encoders and project their fused representation into a joint latent space. Simultaneously, a pretrained Transformer-based text encoder maps the class definitions into the same space. The model is trained via a CLIP-style contrastive loss to align traffic instances with their corresponding textual definitions, enabling semantic generalization to unseen classes in the target domain.

Our contributions can be summarized as follows:

- We formally define the task of cross-dataset intrusion detection, a realistic and practically important setting where the model must detect novel attack types in unseen network environments using only their textual definitions.
- We propose TriFusion-ID, a novel three-stream contrastive learning architecture that integrates graph-based structural patterns, tabular traffic features, and semantic class descriptions into a unified embedding space. To the best of our knowledge, this is the first work that introduces a CLIP-style contrastive alignment mechanism for aligning traffic behavior with natural language attack definitions, enabling semantic generalization without labeled target samples.
- We conduct comprehensive experiments on two widely-used intrusion detection datasets, NF-UNSW-NB15 and NF-CSE-CIC-IDS2018, demonstrating that our approach significantly outperforms strong baselines in the cross-dataset settings.

Related Work

Graph-based Intrusion Detection

Modern intrusion detection systems (IDS) have evolved from classical signature- and anomaly-based approaches to leveraging deep learning, particularly graph neural networks (GNNs) (Lo et al. 2021; Jing et al. 2023), to model complex traffic behavior (Zhong et al. 2024). Recent research demonstrates that representing traffic as communication graphs (nodes as IPs/ports and edges as session flows) benefits detection by encoding structural and relational patterns beyond traditional tabular features (Tran and Park 2024). Recent works capture topological and edge-level features for IDS scenarios, such as E-GraphSAGE (Lo et al. 2021) and Anomal-E (Caville et al. 2022). However, these graph-based approaches typically remain supervised, relying on labeled data in both source and target domains, and thus struggle with generalizing to unseen attack types in new environments.

Multimodal Contrastive Learning

Contrastive learning frameworks like CLIP have revolutionized multimodal alignment, training visual and textual encoders in a shared embedding space using the InfoNCE loss (Radford et al. 2021). Contrastive learning has inspired a lot of work in the field of graph learning (Yu and Yu 2025; Liu, Yu, and Luo 2025). GraphCLIP (Zhu et al. 2025) further extends this paradigm by aligning graph-structured data with textual information in general settings. However, there is a lack of frameworks that bring this CLIP-style alignment into network security, particularly combining traffic graphs, tabular flow features, and textual attack definitions. Our TriFusion-ID is, to our knowledge, the first effort in IDS to build such a three-way cross-modal contrastive framework, enabling detection of unseen attacks via semantic matching with class definitions.

Problem Definition

We aim to address the task of Cross-Dataset Zero-Shot Intrusion Detection. The task assumes a source domain dataset $D_S = \{(x_i^S, y_i^S)\}_{i=1}^{N_S}$. Here, each network traffic sample x_i^S is specifically defined as a composite structure, containing both graph-structured information that describes the communication relationships between source and destination endpoints, and a feature vector summarizing statistical traffic behavior. Each sample is associated with a label y_i^S from the set of **seen** classes, Y_S, and a corresponding text description. In the more challenging testing phase, the model must process a target domain dataset, D_T, where the ground-truth labels of its samples may belong either to the known set Y_S or to a new set of **unseen** classes, Y_U, which never appeared during training. By definition, these two class sets are mutually exclusive, i.e., $Y_S \cap Y_U = \emptyset$.

Our objective is to learn a unified prediction function f that takes a target domain traffic sample x^T—composed of its graph structure and feature vector—and the set of all possible class descriptions, $T_S \cup T_U$, as input. From these, it must predict the most probable class label $\hat{y} \in Y_S \cup Y_U$. This requires the model not only to recognize previously learned attacks but also to possess strong generalization capabilities, enabling it to identify and distinguish novel attack types solely by aligning traffic behavior with semantic descriptions, thereby effectively addressing emerging real-world threats.

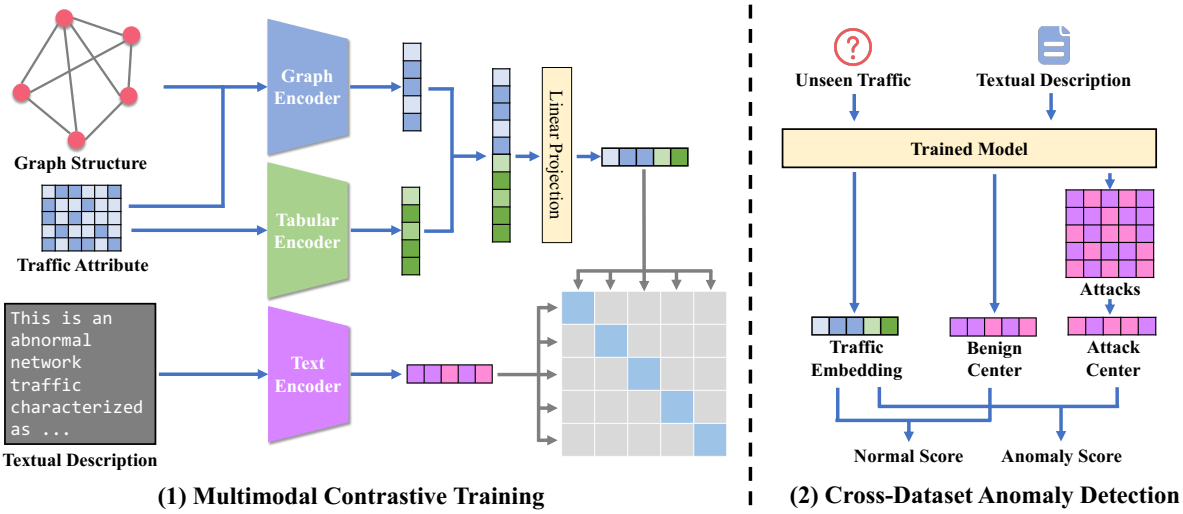

Figure 2: Overview of the proposed TriFusion-IDS framework.

Method

Framework Overview

Our core hypothesis is that a robust intrusion detection model should be able to understand the intrinsic consistency between the structural behavior (graph), statistical features (tabular), and semantic essence (text) of network traffic. The TriFusion-IDS framework is founded on this idea, aligning these three modalities into a unified embedding space through multimodal contrastive learning. The framework consists of three parallel encoder branches (f_G, f_{Tab}, f_T), a fusion module for integrating traffic information, and a contrastive learning head to drive the model's learning process.

As shown in Figure 2, the entire learning process is divided into the following two stages. The first stage is **multimodal representation learning**. On the source domain D_S, we train the model to learn an aligned multimodal embedding space using our proposed contrastive learning objective. In this space, the representation of a traffic sample is semantically close to the representation of its corresponding attack's text description. The second stage is **cross-Dataset inference**. On the target domain D_T, we freeze the trained model. By calculating the similarity between an unknown traffic sample and a pre-defined "normal center" and "anomaly center," we perform binary classification directly to determine if it is anomalous traffic.

Multimodal Feature Encoding

Graph Structural Encoding To capture the complex contextual relationships between traffic flows, we construct a global Flow-Relation Graph $G = (V, E)$ for the entire dataset, rather than modeling each flow sample individually. In this graph, each vertex $v_i \in V$ represents an individual network flow, and an edge $(v_i, v_j) \in E$ is added if two distinct flows, v_i and v_j, share key network entities (e.g., they have the same source or destination IP address).

On this global graph, we employ GraphSAGE (Hamilton et al. 2017) as the graph encoder f_G. GraphSAGE iteratively updates a node's representation by first aggregating information from its neighbors and then combining it with the node's own representation. At layer k, for each node v_i, we first aggregate the representations of its neighbors $\mathcal{N}(v_i)$ from the previous layer:

$$a_{v_i}^{(k)} = \text{AGG}\left(\{h_u^{(k-1)}, \forall u \in \mathcal{N}(v_i)\}\right) \tag{1}$$

where AGG is an aggregation function (e.g., mean, max, or LSTM). Next, the node's new representation $h_{v_i}^{(k)}$ is updated by combining its previous representation $h_{v_i}^{(k-1)}$ with the aggregated neighborhood vector $a_{v_i}^{(k)}$:

$$h_{v_i}^{(k)} = \sigma\left(W^{(k)} \cdot \text{Concat}\left(h_{v_i}^{(k-1)}, a_{v_i}^{(k)}\right)\right) \tag{2}$$

where $W^{(k)}$ is a learnable weight matrix. After K iterations, we take the output of the final layer, $h_{v_i}^{(K)}$, as the graph feature z_G^i for the traffic sample.

To enhance model robustness and prevent overfitting, we follow the idea from GraphCLIP (Zhu et al. 2025) and introduce minor perturbations to the graph to create an augmented view. Consequently, for any node v_i in the graph, we use GraphSAGE to extract its context-aware embeddings from both the original and the perturbed graphs. This generates two graph features for each traffic sample i: the original graph feature $z_{G,\text{orig}}^i$ representing the original structure, and the perturbed graph feature $z_{G,\text{perturb}}^i$ from the augmented view.

Tabular Feature Encoding Each traffic sample i is also associated with a d-dimensional tabular feature vector $F_i \in \mathbb{R}^d$, which includes statistical information such as flow duration and packet size. We use a Multi-Layer Perceptron (MLP) as the tabular encoder f_{Tab}, which maps the high-dimensional raw features into a compact and informative low-dimensional representation:

$$z_{Tab}^i = \text{MLP}(F_i) \tag{3}$$

Textual Semantic Encoding To ensure that the textual representations not only contain attack definitions but also guide the model to learn the fundamental differences between "normal" and "abnormal," we design structured prompts for each attack class c_k as its text input t_k. These prompts explicitly emphasize the nature of the traffic, preparing the model for the subsequent cross-dataset anomaly detection task. For example:

- **Benign**: `"This is normal network traffic with no malicious activity detected."`
- **DoS**: `"This is abnormal network traffic characterized as a DoS attack,..."`

We use a pre-trained **BERT** model as the text encoder f_T. To obtain a more comprehensive representation of the entire prompt's semantics, we apply **Mean Pooling** to the last hidden state vectors of all tokens from BERT's output. The resulting pooled vector serves as the semantic embedding z_T for the attack class, rich with contextual and contrastive information:

$$z_T = \text{Mean-Pooling}(\text{BERT}(t_k)) \tag{4}$$

Cross-Modal Fusion and Contrastive Learning

Traffic Feature Fusion To gain a comprehensive understanding of network traffic, we must fuse its structural and statistical information. For each sample i, we concatenate the two structural embeddings from the graph encoder (original and perturbed) with the tabular feature embedding, respectively, to form two unified traffic representations:

$$z^i_{\text{traffic,orig}} = \text{Concat}(z^i_{G,\text{orig}}, z^i_{Tab}) \tag{5}$$

$$z^i_{\text{traffic,perturb}} = \text{Concat}(z^i_{G,\text{perturb}}, z^i_{Tab}) \tag{6}$$

Contrastive Alignment To enable comparison between representations from different modalities in the same space, we first pass the two fused traffic representations and the text representation through separate, independent non-linear projection heads. This maps them into the final contrastive space, yielding the original traffic embedding $e_{\text{traffic,orig}}$, the perturbed traffic embedding $e_{\text{traffic,perturb}}$, and the text embedding e_{text}. Next, we construct our training objective based on the **InfoNCE** contrastive loss function. We calculate the contrastive loss for the original and perturbed features separately. In a mini-batch of size N, for an original traffic embedding $e_{i,\text{orig}}$ and its matching text description e_j, the loss is:

$$\mathcal{L}_{\text{orig}} = -\log \frac{\exp(\text{sim}(e_{i,\text{orig}}, e_j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(e_{i,\text{orig}}, e_k)/\tau)} \tag{7}$$

Similarly, for the perturbed embedding $e_{i,\text{perturb}}$, the loss is:

$$\mathcal{L}_{\text{perturb}} = -\log \frac{\exp(\text{sim}(e_{i,\text{perturb}}, e_j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(e_{i,\text{perturb}}, e_k)/\tau)} \tag{8}$$

The final total loss is a weighted sum of these two losses, balanced by a hyperparameter λ:

$$\mathcal{L}_{\text{CL}} = \lambda \mathcal{L}_{\text{orig}} + (1-\lambda)\mathcal{L}_{\text{perturb}} \tag{9}$$

Cross-Dataset Inference and Anomaly Detection

After the model is trained, we leverage the aligned embedding space for zero-shot anomaly detection. The core idea of this stage is to transform the multi-class classification problem into a binary judgment: whether a given traffic flow is "normal" or "anomalous."

First, we pre-compute the text embeddings for all candidate attack classes (including both Y_S and Y_U), denoted as $\{e^k_{\text{text}}\}_{c_k \in Y_S \cup Y_U}$. We then construct a generalized anomaly center c_{anomaly} by taking the mean of these attack embeddings:

$$c_{\text{anomaly}} = \frac{1}{|Y_S \cup Y_U|} \sum_{c_k \in Y_S \cup Y_U} e^k_{\text{text}} \tag{10}$$

This anomaly center semantically represents the common characteristics of all known and unknown attacks. Concurrently, we use the text embedding of the "Benign" class directly as the normal center c_{normal}.

For an unseen test sample x^T from the target domain, we first compute its traffic embedding e^T_{traffic} (using only the original graph features during inference). The final decision is made by comparing the similarity of this traffic embedding to the two centers:

$$\hat{y} = \begin{cases} 1, & \text{if } \text{sim}(e^T_{\text{traffic}}, c_{\text{anomaly}}) > \text{sim}(e^T_{\text{traffic}}, c_{\text{normal}}) \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where $\hat{y} = 1$ means anomaly, and $\hat{y} = 0$ means normal. This process cleverly transforms the task of specific attack identification into a measurement of a flow's "anomalousness," significantly enhancing the model's generalization ability and robustness when facing entirely unknown attacks.

Experiments

Experimental Settings

Datasets In our experiments, we utilize the enhanced NetFlow v2 version (Sarhan, Layeghy, and Portmann 2022) of three network flow datasets from recent research, which are notable for their diversity of modern attacks and accurate modeling of real-world network traffic (Caville et al. 2022). The datasets are UNSW-NB15 (Moustafa and Slay 2015), ToN-IoT (Koroniotis et al. 2019), CSE-CIC-IDS2018 (Sharafaldin, Lashkari, and Ghorbani 2018).

To simulate a realistic scenario, we conduct a rigorous cross-dataset training and evaluation protocol. Our strategy is to train on one dataset and then test separately on the other two. For instance, after a model is trained on NF-UNSW-NB15-v2 (whose attack classes are considered the seen classes Y_S), we perform two independent sets of tests: one on the NF-ToN-IoT-v2 dataset and another on the NF-CSE-CIC-IDS2018-v2 dataset. In both testing scenarios, the test set comprises samples from Y_S as well as the native attacks from the target test set (i.e., the unseen classes Y_U). This approach allows us to systematically evaluate the model's generalization capabilities across different target environments.

Baselines To comprehensively evaluate our proposed TriFusion-IDS, we compare it against a wide range of representative baselines across four categories. All supervised

Method	UNSW-NB15 → ToN-IoT				UNSW-NB15 → CSE-CIC-IDS			
	10%		100%		10%		100%	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
♠ GCN (Kipf 2016)	35.50	26.20	36.69	27.17	94.39	87.22	95.56	88.66
♠ GIN (Xu et al. 2018)	35.62	26.61	34.58	26.07	87.23	58.73	86.93	58.52
♠ GAT (Veličković et al. 2017)	47.42	44.45	49.94	46.96	91.20	71.10	93.98	73.60
♠ GraphSAGE (Hamilton et al. 2017)	36.18	27.09	36.63	27.39	91.39	88.34	91.65	88.75
♣ MLP	43.62	39.46	42.47	38.14	93.07	84.24	92.17	83.22
♣ TabNet (Arik and Pfister 2021)	64.10	46.52	63.83	46.19	88.77	62.27	89.13	62.46
♡ E-GraphSAGE (Lo et al. 2021)	41.94	36.81	44.38	38.51	87.43	52.18	89.22	54.26
♡ Anomal-E (Caville et al. 2022)	64.10	41.02	62.10	39.44	87.96	50.25	86.10	48.88
♢ GraphCLIP (Zhu et al. 2025)	54.03	53.80	52.22	51.61	89.28	73.78	87.87	72.02
Ours	**65.37**	**59.82**	**63.95**	**57.93**	**96.75**	**91.74**	**95.72**	**90.39**

Table 1: Comparison with state-of-the-arts on UNSW-NB15 dataset (%). ♠ indicates graph-based method, ♣ indicates tabular-based method, ♡ indicates specialized NIDS method, ♢ indicates graph contrastive learning method.

Method	ToN-IoT → UNSW-NB15				ToN-IoT → CSE-CIC-IDS			
	10%		100%		10%		100%	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
♠ GCN (Kipf 2016)	86.18	57.90	87.80	59.27	85.53	72.82	87.15	73.81
♠ GIN (Xu et al. 2018)	83.87	45.82	83.24	44.53	68.58	60.41	67.69	60.02
♠ GAT (Veličković et al. 2017)	66.95	40.19	67.64	40.62	51.09	41.63	51.79	42.24
♠ GraphSAGE (Hamilton et al. 2017)	62.15	43.13	63.28	44.12	75.97	61.30	76.34	61.87
♣ MLP	62.35	45.08	63.69	47.30	53.06	48.33	54.82	49.79
♣ TabNet (Arik and Pfister 2021)	90.20	60.30	92.97	62.91	88.99	69.35	91.81	71.63
♡ E-GraphSAGE (Lo et al. 2021)	50.62	34.56	51.94	35.89	89.19	63.62	89.72	64.23
♡ Anomal-E (Caville et al. 2022)	91.24	60.68	88.46	58.43	89.96	78.98	87.38	76.92
♢ GraphCLIP (Zhu et al. 2025)	94.95	56.24	**94.50**	55.72	86.60	74.03	86.12	73.18
Ours	**96.04**	**62.58**	93.71	**59.98**	**96.09**	**89.60**	**93.97**	**87.82**

Table 2: Comparison with state-of-the-arts on ToN-IoT dataset (%). ♠ indicates graph-based method, ♣ indicates tabular-based method, ♡ indicates specialized NIDS method, ♢ indicates graph contrastive learning method.

baselines are trained conventionally and are evaluated on their ability to detect anomalies in the cross-dataset setting.

- Graph-based Methods: To assess the efficacy of using graph structure alone, we include prominent Graph Neural Networks (GNNs): GCN (Kipf 2016), GAT (Veličković et al. 2017), GIN (Xu et al. 2018), and GraphSAGE (Hamilton et al. 2017).
- Tabular-based Methods: Representing approaches that rely solely on statistical flow features, we compare against: MLP and TabNet (Arik and Pfister 2021).
- Specialized NIDS Methods: We include two advanced graph-based models designed specifically for network intrusion and anomaly detection: E-GraphSAGE (Lo et al. 2021) and Anomal-E (Caville et al. 2022).
- Graph Contrastive Learning Methods: As our most critical baseline, we implement GraphCLIP (Zhu et al. 2025), which directly aligns graph-level representations with textual attack descriptions via contrastive learning. This method shares our zero-shot learning paradigm but omits tabular features.

Implementation Details

Our experiments were conducted on a server equipped with an NVIDIA GeForce RTX 4090 GPU (24GB VRAM), running the Ubuntu 22.04 operating system. The model was implemented based on the PyTorch framework built using PyTorch Geometric (PyG) library. During the training process, we used the Adam optimizer with an initial learning rate set to 1e-3. To comprehensively evaluate the model's generalization capability, we utilized three mainstream datasets: UNSW-NB15, ToN-IoT, and CSE-CIC-IDS. We performed pairwise cross-dataset testing among these three datasets, which involved training the model on one dataset and then testing it separately on the other two, resulting in a total of 6 independent experimental scenarios.

Comparison with State-of-arts

To comprehensively validate the effectiveness of our proposed TriFusion-IDS framework, we conducted a systematic comparison against a series of representative baseline methods. These baselines include graph-only, tabular-only, and advanced models specifically designed for network in-

Method	CSE-CIC-IDS → UNSW-NB15				CSE-CIC-IDS → ToN-IoT			
	10%		100%		10%		100%	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
♠ GCN (Kipf 2016)	80.23	45.39	80.11	44.82	55.94	55.39	55.46	54.71
♠ GIN (Xu et al. 2018)	79.05	53.13	81.24	56.20	55.09	54.99	58.12	58.08
♠ GAT (Veličković et al. 2017)	94.36	63.27	94.20	63.13	63.20	63.19	62.71	62.53
♠ GraphSAGE (Hamilton et al. 2017)	92.35	62.94	92.27	63.21	63.15	63.14	62.87	63.42
♣ MLP	74.71	51.16	72.32	48.95	63.61	63.59	61.26	61.61
♣ TabNet (Arik and Pfister 2021)	95.93	56.23	98.72	58.19	64.54	64.53	**67.11**	**66.78**
♡ E-GraphSAGE (Lo et al. 2021)	85.61	48.09	85.04	46.99	53.94	53.53	53.46	52.62
♡ Anomal-E (Caville et al. 2022)	92.81	62.87	91.48	61.44	64.20	63.91	62.49	61.65
♢ GraphCLIP (Zhu et al. 2025)	90.99	61.27	92.52	62.79	50.25	49.06	51.25	50.35
Ours	**96.02**	**64.68**	**93.80**	**63.32**	**66.65**	**64.86**	64.82	63.21

Table 3: Comparison with state-of-the-arts on CSE-CIC-IDS dataset (%). ♠ indicates graph-based method, ♣ indicates tabular-based method, ♡ indicates specialized NIDS method, ♢ indicates graph contrastive learning method.

Inference Strategy	IDS → NB15		IDS → IoT	
	Acc	F1	Acc	F1
w/o Graph Module	94.34	62.63	55.17	51.95
w/o Tabular Module	95.23	53.25	53.59	52.61
w/o Perturbation	94.87	61.14	48.85	49.57
w/o Prompting	93.23	51.03	47.49	46.04
TriFusion-IDS	96.02	64.68	66.65	64.86

Table 4: Ablation study results. The best performance is emphasized in **bold**.

trusion detection. All methods were evaluated under six different cross-dataset (train → test) scenarios, with detailed performance results recorded in Table 1, 2, 3. It can be observed as the followings.

(1) Experimental results show that our TriFusion-IDS model consistently outperformed all baseline methods across all key metrics in all six cross-dataset testing scenarios. This demonstrates the model's excellent generalization and robustness in cross-domain intrusion detection, with its significant Macro F1 advantage indicating higher reliability in handling imbalanced data, especially for unseen attacks.

(2) In contrast, graph-based baselines, including general GNNs (like GCN, GAT) and specialized anomaly detection models, showed severe generalization limits in cross-dataset scenarios, with accuracy dropping as low as 35.50% in some tests. This is fundamentally due to their reliance on traditional supervised learning, which depends heavily on attack labels seen during training and thus fails to handle the zero-shot challenge in new environments.

(3) Similarly, tabular methods relying solely on flow statistics (e.g., MLP, TabNet) also hit a performance bottleneck. For instance, TabNet's high accuracy (64.10%) contrasted sharply with its low Macro F1 (46.52%), exposing its failure to balancedly identify all attacks, especially unseen ones. This confirms statistical patterns alone are insufficient to capture the complex nature of attacks; lacking a deep understanding of traffic structure and attack semantics, these models cannot achieve robust generalization.

Ablation Study

We conducted a series of ablation studies to validate the necessity of each key component in our TriFusion-IDS framework, with results presented in Table 4. The experiments confirm that the complete model achieves the best performance and that removing any component degrades performance. This demonstrates that our multimodal fusion, graph perturbation, and semantic alignment strategies each make an indispensable contribution to the final result.

The importance of multimodal fusion was evident. Removing the Graph Module (w/o Graph Module) caused a significant accuracy drop from 66.65% to 55.17% in the CSE → ToN-IoT scenario, highlighting the criticality of structural patterns. Similarly, removing the Tabular Module (w/o Tabular Module) led to a sharp decline in the Macro F1 score from 64.68% to 53.25% in the CSE → UNSW-NB15 test, proving that graph and tabular features provide essential, complementary information.

The contributions of our key strategies were also clear. Removing graph perturbation (w/o Perturbation) severely impacted generalization, with accuracy plummeting to 48.85% in the CSE → ToN-IoT test. Furthermore, removing text prompting (w/o Prompting) caused a substantial drop in the Macro F1 score to 51.03% in the CSE → UNSW-NB15 scenario, validating our core hypothesis that aligning traffic with semantic descriptions is crucial for zero-shot detection.

Impact of Center Strategy

To validate the effectiveness of our inference strategy, we compared three different zero-shot inference strategies, with the results presented in Table 5. The data clearly demonstrates that our proposed Anomaly Center strategy significantly outperforms the other methods across all cross-dataset scenarios. For instance, in the CSE-CIC-IDS → UNSW-NB15 scenario, our method achieves an accuracy of 96.02% and a Macro F1 score of 64.68%, comprehensively surpassing both Multi-Class Similarity and Single

Inference Strategy	IDS → NB15		IDS → IoT	
	Acc	F1	Acc	F1
Multi-Class	91.03	59.68	46.66	45.90
Single Prompt	93.01	62.98	54.26	53.27
Anomaly Center	**96.02**	**64.68**	**66.65**	**64.86**

Table 5: Impact of different center strategies. The best performance is emphasized in **bold**.

Anomaly Prompt approaches. This advantage is even more pronounced in the more challenging CSE-CIC-IDS → ToN-IoT scenario, where our method's accuracy leads the next best by over 12 percentage points. This superiority stems from the fact that the Anomaly Center, by averaging the semantic descriptions of all attacks, constructs a generalized and robust concept of "abnormality." This allows the model to determine a flow's semantic distance from the abstract concepts of "normal" and "abnormal" rather than relying on a fragile match with a specific attack description. In contrast, the Multi-Class Similarity method underperforms due to its over-reliance on exact matching, while the Single Anomaly Prompt, though more general, is limited by its impoverished semantic information. Therefore, the experiment confirms that transforming the multi-class task into a binary anomaly judgment based on a generalized semantic center is key to improving the model's cross-dataset detection capabilities.

Impact of Perturbation Factor λ

To investigate the role of the perturbation factor λ in balancing the losses from the original and perturbed graphs, we conducted a series of parameter sensitivity analyses. As shown in Figure 3, the model's performance exhibits a consistent trend with respect to the value of λ: as λ increases from 0.1, performance generally improves, reaching a peak at λ=0.5, and then declines as λ continues to increase. This result indicates that assigning equal weight to the losses from both the original and perturbed graphs (i.e., λ=0.5) achieves the optimal balance. This allows for full utilization of the true structural information from the original graph while enhancing the model's robustness and generalization ability through the graph augmentation strategy. Therefore, setting λ to 0.5 is the ideal choice for achieving optimal performance, which also validates the critical importance of equally leveraging the original and augmented views in our contrastive learning framework.

Visualization Analysis

To intuitively evaluate the quality of the feature representations learned by Anomal-E baseline and our model, we utilized the t-SNE technique to reduce the dimensionality of the test sample feature vectors for visualization, as shown in Figure 4. We compared the feature distributions of our method and the baseline model in two cross-dataset scenarios: CSE → UNSW and CSE → ToN. It is clear from the figure that the features extracted by the baseline model cause the benign and attack samples to be severely mixed in the

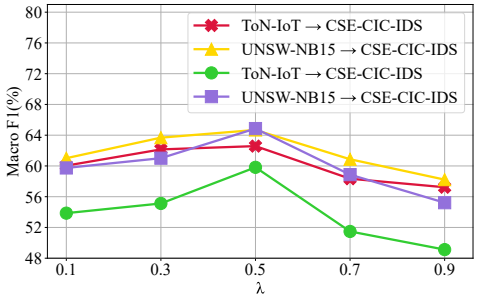

Figure 3: Sensitivity study results for the perturbation factor.

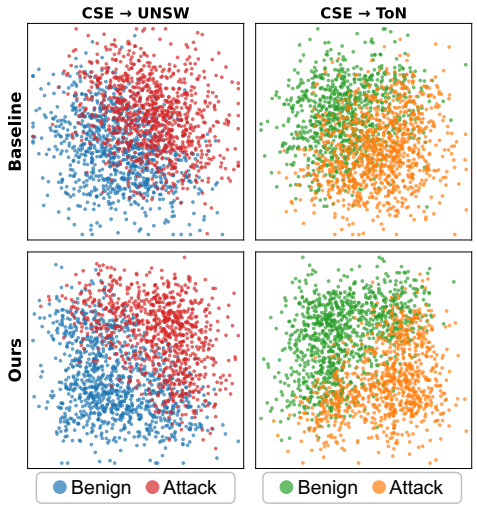

Figure 4: Scatter visulization of the extracted features by baseline and ours.

feature space, making them almost indistinguishable, which visually explains its poor classification performance. In contrast, the features learned by our TriFusion-IDS model exhibit significant intra-class compactness and inter-class separability. The benign and attack samples each form clearer and more distinct clusters, even in challenging cross-dataset scenarios. This visualization result provides strong evidence that our method can learn more discriminative and generalizable feature representations, thereby effectively distinguishing unknown attacks from normal traffic.

Conclusion

This paper addresses the cross-dataset generalization problem in IDS by proposing TriFusion-IDS, a novel multimodal framework that fuses graph, tabular, and textual data. By aligning network traffic with natural language descriptions, our model effectively identifies unseen attacks based on their semantics. Extensive cross-dataset experiments, supported by ablation and visualization analyses, validate our approach, demonstrating superior generalization and robustness over strong baselines. TriFusion-IDS thus offers a promising direction for developing adaptive intrusion detection systems capable of handling ever-evolving threats.

Acknowledgments

This research is supported, in part, by NSFC-61872176, NSFC-62272215, Leading-edge Technology Program of Jiangsu NSF under Grant BK20202001, and National Key R&D Program of China under Grants 2020YFB1005900.

References

Arik, S. Ö.; and Pfister, T. 2021. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 6679–6687.

Caville, E.; Lo, W. W.; Layeghy, S.; and Portmann, M. 2022. Anomal-E: A self-supervised network intrusion detection system based on graph neural networks. *Knowledge-based systems*, 258: 110030.

Hamilton, W.; Ying, Z.; Leskovec, J.; and et al. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Jing, Y.; Hong, S.-H.; and Tao, D. 2024. Deep Graph Mating. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jing, Y.; Yuan, C.; Ju, L.; Yang, Y.; Wang, X.; and Tao, D. 2023. Deep Graph Reprogramming. In *CVPR*.

Kipf, T. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907*.

Koroniotis, N.; Moustafa, N.; Sitnikova, E.; and Turnbull, B. 2019. Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset. *Future Generation Computer Systems*, 100: 779–796.

Liu, Z.; Yu, H.; and Luo, X. 2025. Federated Graph Anomaly Detection via Disentangled Representation Learning. In *Proceedings of the ACM on Web Conference 2025*, 1216–1224.

Lo, W. W.; Layeghy, S.; Sarhan, M.; Gallagher, M.; and Portmann, M. 2021. E-graphsage: A graph neural network based intrusion detection system for iot. *arXiv preprint arXiv:2103.16329*.

Moustafa, N.; and Slay, J. 2015. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 military communications and information systems conference (MilCIS)*, 1–6. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rao, D.; and Mane, S. 2021. Zero-shot learning approach to adaptive Cybersecurity using Explainable AI. *arXiv preprint arXiv:2106.14647*.

Sarhan, M.; Layeghy, S.; Gallagher, M.; and Portmann, M. 2023. From zero-shot machine learning to zero-day attack detection. *International Journal of Information Security*, 22(4): 947–959.

Sarhan, M.; Layeghy, S.; and Portmann, M. 2022. Towards a standard feature set for network intrusion detection system datasets. *Mobile networks and applications*, 27(1): 357–370.

Schmitt, M. 2023. Securing the digital world: Protecting smart infrastructures and digital industries with artificial intelligence (AI)-enabled malware and intrusion detection. *Journal of Industrial Information Integration*, 36: 100520.

Sharafaldin, I.; Lashkari, A. H.; and Ghorbani, A. A. 2018. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, 1: 108–116.

Talukder, M. A.; Islam, M. M.; Uddin, M. A.; Hasan, K. F.; Sharmin, S.; Alyami, S. A.; and Moni, M. A. 2024. Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. *Journal of big data*, 11(1): 33.

Tran, D.-H.; and Park, M. 2024. FN-GNN: A novel graph embedding approach for enhancing graph neural networks in network intrusion detection systems. *Applied Sciences*, 14(16): 6932.

Uddin, M. A.; Aryal, S.; Bouadjenek, M. R.; Al-Hawawreh, M.; and Talukder, M. A. 2025. A dual-tier adaptive one-class classification IDS for emerging cyberthreats. *Computer Communications*, 229: 108006.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Yu, H.; Wen, J.; Sun, Y.; Wei, X.; and Lu, J. 2024. CA-GNN: A competence-aware graph neural network for semi-supervised learning on streaming data. *IEEE Transactions on Cybernetics*.

Yu, H.; and Yu, H. 2025. Enhancing Zero-Shot Knowledge Graph Relation Prediction through Large Language Models and Contrastive Learning. In *Companion Proceedings of the ACM on Web Conference 2025*, 1490–1494.

Zhong, M.; Lin, M.; Zhang, C.; and Xu, Z. 2024. A survey on graph neural networks for intrusion detection systems: Methods, trends and challenges. *Computers & Security*, 141: 103821.

Zhu, Y.; Shi, H.; Wang, X.; Liu, Y.; Wang, Y.; Peng, B.; Hong, C.; and Tang, S. 2025. Graphclip: Enhancing transferability in graph foundation models for text-attributed graphs. In *Proceedings of the ACM on Web Conference 2025*, 2183–2197.