Towards Robust Text-Attributed Federated Graph Learning: Multimodal Threats and Defense

Zitong Shi[1], **Guancheng Wan**[1], **Wenke Huang**[1*], **Yuxin Wu**[2], **Quan Zhang**[3], **Mang Ye**[1*]
[1]National Engineering Research Center for Multimedia Software,
School of Computer Science, Wuhan University, Wuhan, China
[2]Renmin University of China
[3]ByteDance Inc.
{zitongshi, wenkehuang, yemang}@whu.edu.cn

Abstract

Text-Attributed Graphs (TAGs) are graphs where both nodes and edges are associated with text attributes. To leverage their semantic richness, recent efforts have integrated large language models (LLMs) with graph neural networks, leading to the development of GraphLLMs. However, many real-world datasets remain inaccessible, and processing text-attributed graphs while ensuring privacy and efficiency remains a challenge. To address this, we place TAGs within a federated environment, referred to as TAG-FGL. Despite its potential, TAG-FGL remains largely underexplored in the face of adversarial threats. In this work, we introduce GTAE, a novel attack framework that cascades influence-guided topological perturbations and embedding-level text refinements to generate transferable, modality-agnostic adversarial inputs. To defend against these threats, we propose STRUM, a defense strategy that combines local adversarial training with robustness-aware aggregation, enhancing resilience at both the node and system levels. Extensive experiments on five real-world datasets with diverse model backbones demonstrate that GTAE significantly degrades model performance, while STRUM consistently improves robustness.

Introduction

Text-attributed Graphs (TAGs) are a type of graph data where both nodes and edges are associated with text attributes. Recently, they have become increasingly important in social media analysis and biochemical research (Li et al. 2023; Fang et al. 2024). To leverage the semantic richness of text-attributed graphs, recent works have proposed integrating large language models into graph neural networks, giving rise to GraphLLMs. However, the centralized processing of these graphs raises concerns over data privacy and scalability, especially with sensitive or large-scale datasets. For instance, in healthcare applications, patient data associated with medical graphs cannot be centrally stored due to privacy regulations like HIPAA, as this may lead to the exposure of sensitive information. Therefore, introducing federated settings is crucial, as it allows for distributed training across multiple clients without sharing sensitive data, ensuring privacy while enabling collaborative learning for improved model performance (Yang et al. 2019; Mammen 2021). To address the challenges of processing text-attributed graphs while maintaining privacy and efficiency, Federated Graph Learning (FGL) has emerged as a promising solution (Luo and Wu 2022; Wan et al. 2025b). This combination of text and graph data in a federated setting, which we refer to as TAG-FGL, provides a powerful framework for privacy-preserving graph learning while leveraging the richness of textual attributes.

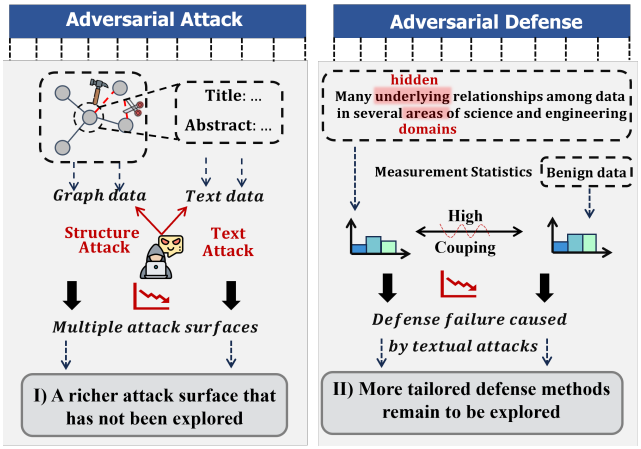


Figure 1: Problem illustration. We describe the challenges TAG-FGL encounters under adversarial attacks: **I)** TAG-FGL introduces additional unexplored attack surfaces. **II)** The multi-modal attack makes some traditional defense methods ineffective.

Despite these modeling advantages, GraphLLMs introduce new security risks due to their reliance on high-dimensional and multi-modal inputs. The integration of complex textual and structural features increases the attack surface. These risks are further amplified in federated settings, where the decentralized training paradigm enables malicious clients to launch attacks on the model without the server's awareness (Zhang et al. 2020; Moradi and Samwald 2021; Wang et al. 2022). Such a setup hinders anomaly detection and reduces the effectiveness of centralized defense mechanisms. Notably, while extensive studies have explored adversarial defense in traditional FL and structure-attributed

*Co-corresponding authors

graph learning (Jiang and Li 2022; Yang et al. 2024), several practical systems have also implemented corresponding defenses. These include Google's deployment of robust aggregation in Gboard (Zhang et al. 2023a) and Meta's structure-denoising GNN models for content recommendation (Zhang and Zitnik 2020). In contrast, little attention has been given to securing text-attributed federated graph learning (TAG-FGL). Given that real-world data is often sensitive and cannot be shared directly, investigating the vulnerabilities and defenses of GraphLLMs in federated settings is both a timely and pressing research direction.

In this work, we aim to develop effective adversarial attack strategies for TAG-FGL, alongside robust defense mechanisms to mitigate such threats. Most adversarial attacks operate during the inference stage, where the attacker constructs malicious yet imperceptible samples that mislead the model into underperforming on key metrics, ultimately preventing its deployment. Compared to conventional FGL, where existing adversarial attacks predominantly focus on the structural modality by manipulating graph connections or injecting specific subgraphs, text-attributed FGL introduces a richer and more vulnerable attack surface that remains largely unexplored (Lei et al. 2024; Lyu et al. 2025). We classify potential adversarial threats into two main modalities: structural and textual. Attacks in both categories can simultaneously alter node representations and interfere with prompt construction, ultimately leading to erroneous predictions from large language models (Shi et al. 2025c; Rong et al. 2025a; Ye et al. 2025; Wan et al. 2025a). In particular, textual attack present distinct challenges due to the nuanced nature of linguistic perturbations, which can subtly manipulate model behavior while preserving semantic plausibility. Despite their destructive potential, such attacks have received little attention in prior work. These observations lead to a central question: *(1) How can we design effective adversarial attacks in TAG-FGL that exploit both structural and textual vulnerabilities?*

To mitigate such adversarial threats at inference time, several federated learning defenses have been explored, including differential privacy, cryptographic protocols, and input anomaly detection. However, these approaches are not well-suited to the multimodal and semantics-aware nature of TAG-FGL. For instance, differential privacy adds random noise but struggles to defend against carefully crafted, semantically plausible text perturbations. Cryptographic methods ensure secure inference but do not prevent attackers from submitting malicious inputs. Input sanitization relies on detecting distributional shifts, which are often imperceptible in language data or minor structural edits. Moreover, most adversarial training methods assume prior knowledge of attack patterns and are limited to the training stage, making them ineffective against evasion attacks. These limitations highlight the lack of effective defenses tailored to real-world, inference-time adversarial risks in text-attributed GraphLLMs, motivating the need for lightweight, modality-aware robustness strategies. These limitations reveal a significant gap in the current defense landscape and raise a pressing question: *(2) How can we build effective and generalizable defense mechanisms tailored to the multimodal nature of TAG-FGL?*

To address the identified questions, we instantiate these ideas with a Graph-Text Attack Engine (GTAE) and a Structure and Text Robustness Unification Method (STRUM). Our unified framework combines multi-modal attack strategies with modality-aware and robustness-aware defense mechanisms. Specifically, GTAE adopts a dual-view attack pipeline: structural attacks manipulate critical edges based on influence-driven heuristics, while textual attacks craft adversarial inputs through embedding-space perturbations that preserve surface semantics. On the defense side, STRUM injects adversarial perturbations into graph embeddings during local training and incorporates textual augmentation to enhance resilience across modalities (Ye et al. 2022). Furthermore, to improve global robustness during federated aggregation, STRUM employs a robustness-aware weighting scheme that assigns higher weights to clients with smaller accuracy drops under local adversarial evaluation.

- We are the first to systematically investigate adversarial attacks and defenses in TAG-FGL, a previously underexplored yet practically important setting that combines both structural and textual vulnerabilities.
- We propose a novel attack framework that incorporates both structural and textual modalities: structural attack are introduced through influence-guided edge perturbations, while textual attack are generated via embedding-space optimization that preserves semantic consistency.
- To counter these threats, we design a modality-aware defense strategy that enhances structural robustness by injecting learned perturbations into node embeddings during training, while augmenting textual inputs with semantically consistent adversarial variants.
- Extensive experiments across three real-world TAG-FGL benchmarks demonstrate the high effectiveness and transferability of our method.

Related Work

Federated Graph Learning

Federated Graph Learning (FGL) combines the principles of Federated Learning (FL) (Han et al. 2023; Huang et al. 2023; Rong et al. 2025b) and Graph Neural Networks (GNNs) (Rong et al. 2019; Zhou et al. 2020), enabling multiple clients to collaboratively learn graph-based models while preserving data privacy. Recent FGL frameworks (Zhang et al. 2022; Jiang et al. 2023b; Shi et al. 2025a) focus on tackling graph heterogeneity, communication efficiency, and generalization across domains. However, the distributed nature and topological complexity inherent to FGL expose models to serious security risks, including adversarial attacks. While prior work has extensively studied adversarial vulnerabilities in centralized GNNs (Xu et al. 2020; Kairouz et al. 2021; Sun et al. 2020), defense mechanisms in federated settings remain largely underdeveloped, especially when it comes to text-attributed graphs. To the best of our knowledge, we are the first to systematically investigate adversarial attack and defense strategies tailored to text-attributed federated graph learning.

Adversarial Robustness in Federated Learning

Federated learning systems are inherently vulnerable to adversarial threats due to limited global observability and heterogeneous data. For instance, PELTA generates adversarial examples on individual clients by optimizing gradients to reduce inference accuracy, while NA2 identifies influential samples to craft input perturbations that mislead classification (Chen, Zhang, and Zheng 2024). To mitigate such threats, defense techniques such as differential privacy, input anomaly detection, secure aggregation protocols, and encrypted inference have been proposed (Pillutla, Kakade, and Harchaoui 2022a; Alotaibi and Rassam 2023). These approaches typically aim to obscure or validate local input to prevent malicious manipulation (Pillutla, Kakade, and Harchaoui 2022b). However, the high-dimensional and multimodal nature of TAG-FGL, where textual and structural information are tightly entangled, can significantly undermine their effectiveness (Li et al. 2024; Shang and Huang 2024). Among existing paradigms, adversarial training has shown promise in other domains but remains underexplored in TAG-FGL, despite its potential to enhance robustness across both modalities (Wang et al. 2024, 2025).

LLM-based Graph Learning

Large Language Models (LLMs) have recently been integrated with graph learning systems to exploit both graph topology and natural language features. GraphLLMs typically encode structural information through node embeddings (Tian et al. 2024; Wan et al. 2025c), hop-based sequences (Fang et al. 2024), or hybrid approaches that align graph and language representations using shared embedding spaces (Chen et al. 2024). These methods often freeze the LLM during training, injecting graph signals through token prompts or adapters (Zhang et al. 2023b; Shi et al. 2025b; Wan et al. 2025a; Li et al. 2025). Such designs enable expressive multimodal representations but introduce a broader attack surface, as both structure and text can be subtly manipulated. Although recent studies have begun to examine GraphLLM robustness in centralized environments (Fang et al. 2024; Ren et al. 2024), little is known about their behavior in federated scenarios.

Preliminaries

Federated Graph Learning. Following the general paradigms of traditional text-attributed graph learning and federated graph learning, we consider a federated setting with K clients (indexed by k), where each client owns its local graph data represented as $\mathcal{G}_k = (\mathcal{V}_k, \mathcal{E}_k, \mathcal{X}_k, \mathcal{T}_k)$. Here, $\mathcal{V}_k$ and $\mathcal{E}_k$ denote the sets of nodes and edges, respectively. Each node $v_{i,k} \in \mathcal{V}_k$ is associated with a textual description $t_{i,k} \in \mathcal{T}_k$. Each feature vector $\mathbf{x}_{i,k} \in \mathcal{X}_k$ is computed from its corresponding text via an encoding function $\phi : \mathcal{T} \to \mathbb{R}^d$. The adjacency matrix of the local graph is denoted as $\mathcal{A}_k \in \{0,1\}^{|\mathcal{V}_k|\times|\mathcal{V}_k|}$, where $\mathcal{A}_{ij,k} = 1$ indicates that nodes $v_{i,k}$ and $v_{j,k}$ are connected.

LLM-based Graph Learning. For each node $v_{i,k} \in \mathcal{V}_k$, its textual description $t_{i,k} \in \mathcal{T}_k$ is tokenized into a sequence

$$X_{i,k} = \{x^1_{i,k}, x^2_{i,k}, \ldots, x^m_{i,k}\}, \tag{1}$$

where each $x^j_{i,k}$ denotes a token embedding. To incorporate structural context, a graph encoder f_G computes node representations:

$$h_{i,k} = f_G(\mathcal{G}_k, v_{i,k}), \tag{2}$$

while a set of neighbor embeddings $\{h_{j,k}\}$ is used to form a local structure summary. These embeddings are concatenated with the textual tokens to form an extended prompt:

$$Z_{i,k} = (h_{j_1,k}, \ldots, h_{j_K,k}, x^1_{i,k}, \ldots, x^m_{i,k}), \tag{3}$$

The language model Φ_θ then generates a target sequence $Y_{i,k} = (y^1_{i,k}, \ldots, y^R_{i,k})$ in an auto-regressive fashion, where each token $y^r_{i,k}$ is predicted based on the input $Z_{i,k}$ and the previously generated tokens $y^{<r}_{i,k}$, enabling the model to leverage both structural and textual signals during inference.

Methodology

Overview

In this section, we introduce the core methodology of GTAE and STRUM. GTAE is a dual-modality adversarial framework that jointly manipulates graph structure and textual attributes to mislead model predictions. For structural attacks, each malicious client independently trains a local surrogate GCN and applies influence-guided edge perturbations to reduce classification confidence. For textual attacks, adversarial node descriptions are crafted through optimization in the embedding space, generating semantically plausible yet deceptive inputs that induce label flipping. On the defense side, STRUM employs modality-aware adversarial training, injecting structure-aware perturbations into node representations and augmenting training with adversarial text variants to enhance resilience against multimodal threats. Additionally, we introduce a robustness-aware aggregation strategy that assigns higher weights to clients demonstrating stronger local robustness under adversarial evaluation, thus improving global model stability and reliability.

Influence-Guided Topological Perturbation

For the structural component of GTAE, every malicious client C_k first trains a local GCN surrogate g_k on its private graph $\mathcal{G}_k$. Let $\mathcal{W}_k$ be the learned weights and h the hidden size. Given a target node $v_{i,k}$ with ground-truth label $y_{i,k}$, the attacker searches a perturbed adjacency $\tilde{A}_k$ that maximally decreases the surrogate's confidence on $y_{i,k}$ while staying within a small edge-flipping budget $B_{i,k}$:

$$\begin{aligned} F(x, \tilde{\mathcal{A}}_k) &= \log \hat{p}^x_{i,k}(\tilde{\mathcal{A}}_k, \mathcal{X}_k; \mathcal{W}_k), \\ \tilde{\mathcal{A}}^*_k &= \underset{\tilde{A}_k \in \mathcal{P}(\mathcal{A}_k,\ B_{i,k})}{\arg\max} \Big[F(c^\star) - F(y_{i,k})\Big]. \end{aligned} \tag{4}$$

where $\mathcal{P}(\mathcal{A}_k,\ v_{i,k},\ B_{i,k})$ denotes the set of adjacency matrices that differ from $\mathcal{A}_k$ by at most $B_{i,k}$ edge additions or removals incident to node $v_{i,k}$, $\hat{p}^c_{i,k}$ represents the softmax probability output of the surrogate model g_k for class c; $c^\star = \arg\max_{c \neq y_{i,k}} \hat{p}^c_{i,k}(\mathcal{A}_k, \mathcal{X}_k; \mathcal{W}_k)$ denotes the most competitive incorrect class; and $\mathcal{X}_k$ is the node-feature matrix of client k, where each row $\mathcal{X}_k[i] = \mathbf{x}_{i,k}$ corresponds to

the encoded feature of node $v_{i,k}$. This objective aims to maximize the logit margin between the most confusing incorrect class and the true class by strategically modifying the local graph structure, thereby inducing targeted misclassification.

Node-wise Greedy Perturbation Execution

Once the optimization objective in Eq. 4 is defined, the attacker performs a node-wise greedy search to approximate the optimal solution. For each test node $v_{i,k} \in \mathcal{V}_k$, the local attacker in client C_k initializes a perturbation budget $B_{i,k} = \deg(v_{i,k})$, proportional to the node's degree in $\mathcal{A}_k$. The attacker iteratively selects edge perturbations that maximize the following gain function:

$$\begin{aligned} F(x, A) &= \log \hat{p}^x_{i,k}(A, \mathcal{X}_k; \mathcal{W}_k), \\ \Delta_{ij} &= \Big[F(c^\star, A_k \oplus e_{ij}) - F(y_{i,k}, A_k \oplus e_{ij})\Big]. \end{aligned} \tag{5}$$

where e_{ij} denotes an edge flip between node $v_{i,k}$ and a candidate $v_{j,k}$, and $\oplus$ denotes XOR applied to the adjacency matrix to toggle the edge. At each step, the attacker selects the edge flip with the highest Δ_{ij}, updating the adjacency matrix iteratively until the budget $B_{i,k}$ is reached or no further gain is possible. The surrogate model g_k is trained once on the original graph $\mathcal{G}_k$, and its weights $\mathcal{W}_k$ are reused throughout the attack, following a static surrogate strategy. This greedy, influence-guided method enables efficient node-level misclassification under strict perturbation constraints.

Embedding-Driven Lexical Perturbation Strategy

Semantic Candidate Generation. For the textual modality in GTAE, each malicious client C_k targets the natural-language description $\mathbf{x}_{i,k} = \{w^1_{i,k}, w^2_{i,k}, \ldots, w^m_{i,k}\}$ of node $v_{i,k}$, aiming to craft an adversarial input $\tilde{\mathbf{x}}_{i,k}$ that induces a label flip while preserving semantic plausibility.

To this end, we define a local neighborhood $\mathcal{S}(w^j_{i,k})$ for each word $w^j_{i,k}$, consisting of semantically similar synonyms retrieved from a pre-trained embedding space. The attacker first selects a candidate set of replaceable words:

$$\mathcal{I}_{i,k} = \Big\{ j \mid w^j_{i,k} \in \mathcal{T}_{i,k},\ \text{POS}(w^j_{i,k}) \in \mathcal{C},\ \text{len}(w^j_{i,k}) > 2 \Big\}, \tag{6}$$

where $\mathcal{C}$ denotes a predefined set of content-bearing part-of-speech tags. The attacker constructs perturbed sequences $\tilde{\mathbf{x}}_{i,k}$ by replacing tokens $w^j_{i,k}$ at positions $j \in \mathcal{I}_{i,k}$ with alternative candidates from $\mathcal{S}(w^j_{i,k})$, and evaluates the change in prediction confidence. Inputs that lead to misclassification are selected for further refinement.

Gradient-Guided Text Refinement. To further improve attack success while maintaining semantic plausibility, GTAE performs a continuous optimization in the embedding space. Let $\mathbf{E}_{\text{orig}}$ and $\mathbf{E}_{\text{adv}} \in \mathbb{R}^{|\mathcal{I}_{i,k}| \times d}$ denote the word embeddings of the original and perturbed tokens, respectively. The perturbation direction is defined as:

$$\boldsymbol{\theta}_{i,k} = \mathbf{E}_{\text{adv}} - \mathbf{E}_{\text{orig}}, \tag{7}$$

which encodes the high-dimensional modification applied to selected words. To guide this optimization in preserving semantic similarity, we employ a Gaussian smoothing-based gradient estimation technique. Given a similarity function $f_{\text{sim}}(\cdot)$, the directional gradient is estimated as:

$$\hat{\nabla}_{\boldsymbol{\theta}} f_{\text{sim}} = \frac{f_{\text{sim}}(\boldsymbol{\theta} + \delta \mathbf{u}) - f_{\text{sim}}(\boldsymbol{\theta})}{\delta} \cdot \mathbf{u}, \quad \mathbf{u} \sim \mathcal{N}(0, \mathbf{I}), \tag{8}$$

where δ is a small positive constant that controls the granularity of sampling, and $\mathbf{u}$ is drawn from a standard Gaussian distribution. This estimated gradient is then combined with an ℓ_1-inspired regularizer to guide the refinement of perturbations while promoting sparsity:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \cdot \left(\hat{\nabla}_{\boldsymbol{\theta}} f_{\text{sim}} + \lambda \cdot \text{sign}(\boldsymbol{\theta}) \right), \tag{9}$$

where η is the learning rate and λ controls the strength of ℓ_1-style regularization, encouraging sparse yet effective modifications. Through iterative updates of the perturbation direction, GTAE crafts adversarial examples that preserve semantic consistency while substantially degrading model performance. This embedding-level refinement is particularly effective in circumventing token-level defenses and enhancing cross-model transferability.

Multimodal Attack Composition. GTAE jointly coordinates structure-level and text-level perturbations to induce stronger misclassification. While topological attacks corrupt the relational inductive bias by manipulating informative edges, lexical attacks craft semantically plausible yet adversarial node descriptions. By simultaneously perturbing both modalities, the attacker enhances model confusion in both structural and semantic dimensions, leading to compounded degradation in prediction reliability.

Stage 1 (structure attack) delivers an influence-guided structural flip, yielding a poisoned adjacency $\tilde{\mathcal{A}}_k$. *Stage 2* (text attack) then treats $\tilde{\mathcal{A}}_k$ as frozen context and refines a lexical adversary that maximises the misclassification margin under the already-corrupted topology. Formally, letting $\boldsymbol{\theta}_{i,k}$ be the embedding-space token perturbation, the second-stage objective for node $v_{i,k}$ is:

$$\begin{aligned} F(x) &= \log \hat{p}^x_{i,k}\big(\tilde{\mathcal{A}}_k, \mathbf{E}_k + \boldsymbol{\theta}\big), \\ \boldsymbol{\theta}^*_{i,k} &= \underset{\|\boldsymbol{\theta}\|_0 \leqslant \rho}{\arg\max} \Big[F(c^\star) - F(y_{i,k}) - \beta\, f_{\text{sim}}\big(\mathbf{E}_k, \mathbf{E}_k + \boldsymbol{\theta}\big) \Big], \end{aligned} \tag{10}$$

The optimization objective in Eq. 10 aims to identify a sparse embedding-space perturbation $\boldsymbol{\theta}^*_{i,k}$ that maximizes the classification margin between the most competitive incorrect class $c^\star$ and the ground-truth label $y_{i,k}$ under the structure-poisoned context $\tilde{\mathcal{A}}_k$, while simultaneously minimizing semantic drift. Specifically, the first term encourages label flipping by increasing the logit difference in favor of $c^\star$, and the second term, weighted by a hyperparameter β, acts as a regularizer that penalizes deviations in semantic similarity between the original and perturbed embeddings. The constraint $\|\boldsymbol{\theta}\|_0 \leqslant \rho$ ensures that only a limited number of tokens are modified, thereby promoting minimal and interpretable textual changes.

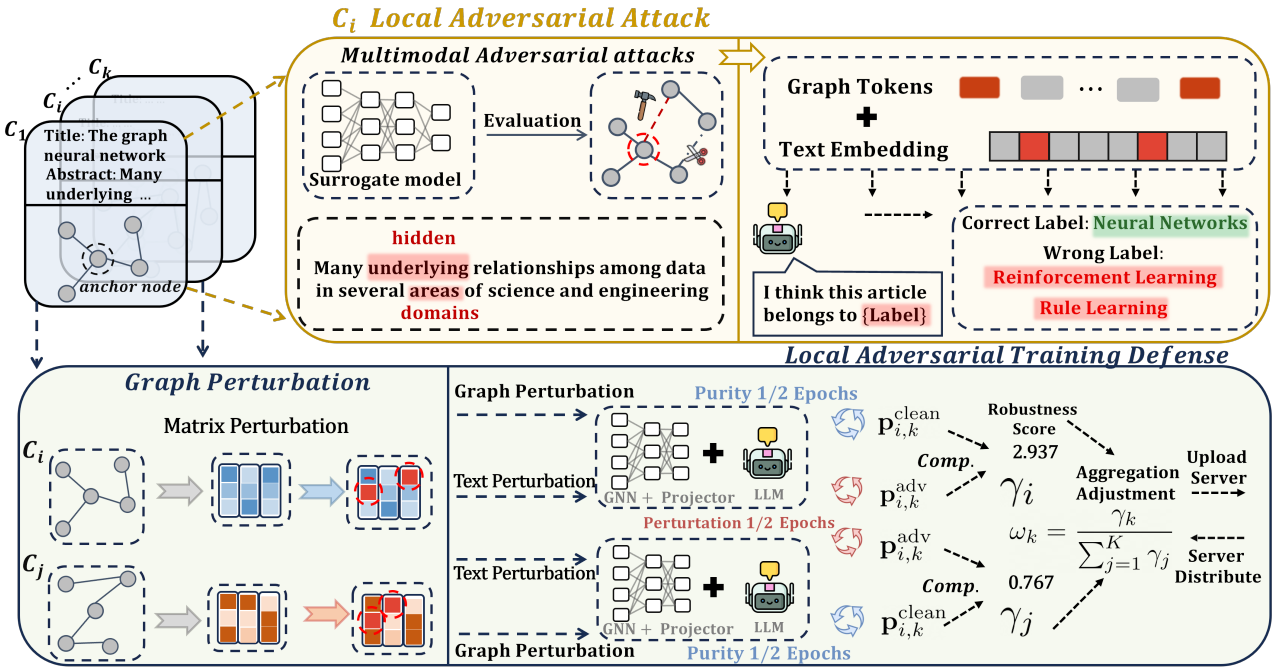


Figure 2: Architecture illustration of our method. The upper part illustrates *Local Multimodal Adversarial Attacks*, where each malicious client perturbs both graph topology and textual features. The lower part shows *Local Adversarial Training Defense*, where benign clients apply modality-aware robustness strategies against structural and textual perturbations. Best viewed in color. Zoom in for details.

Modality-Aware Local Adversarial Training

To counter the threat of modality-specific attacks, STRUM equips each client with a localized adversarial training scheme tailored to the structure-text duality of graph data. Specifically, we adopt a perturbation-injection mechanism during local updates to improve the model's robustness in both topological and linguistic dimensions.

Structure-Side Iterative Injection. Inspired by the FreeLB algorithm (Zhu et al. 2019), we simulate edge-level perturbations in the representation space without explicitly modifying the graph structure. Concretely, given a local GCN encoder g_k with parameters $\mathcal{W}_k$, we introduce a trainable perturbation matrix $\boldsymbol{\xi}_k \in \mathbb{R}^{|\mathcal{V}_k| \times h}$ to the node embeddings during training. Let $\mathbf{H}_k = g_k(\mathcal{A}_k, \mathcal{X}_k)$ be the original hidden representations, the perturbed outputs become:

$$\mathbf{H}_k^{\text{adv}} = g_k(\mathcal{A}_k, \mathcal{X}_k + \boldsymbol{\xi}_k), \tag{11}$$

where $\boldsymbol{\xi}_k$ is iteratively optimized to maximize the local classification loss:

$$\boldsymbol{\xi}_k^* = \arg \max_{\|\boldsymbol{\xi}_k\|_2 \leqslant \epsilon} \mathcal{L}_{\text{CE}}(g_k(\mathcal{A}_k, \mathcal{X}_k + \boldsymbol{\xi}_k), \mathbf{y}_k). \tag{12}$$

The resulting perturbations are added during local model updates, implicitly improving robustness to edge-level manipulations by encouraging the encoder to generalize under embedding-space shifts.

Text-Side Lexical Adversarial Augmentation. For textual robustness, we synthesize adversarial variants of natural-language descriptions via synonym-level perturbation. Given a node description $\mathbf{x}_{i,k}$, we identify a candidate token set $\mathcal{I}_{i,k}$ and retrieve semantic-preserving substitutes $\mathcal{S}(w_{i,k}^j)$ from a pre-trained embedding space. A greedy search is performed to generate perturbations $\tilde{\mathbf{x}}_{i,k}$ that cause maximal prediction deviation. The adversarial text is then mixed into training via a balanced augmentation scheme:

$$\mathcal{L}_{\text{text}} = \alpha\, \mathcal{L}_{\text{CE}}(\mathbf{x}_{i,k}, y_{i,k}) + (1-\alpha)\, \mathcal{L}_{\text{CE}}(\tilde{\mathbf{x}}_{i,k}, y_{i,k}), \tag{13}$$

where $\alpha \in [0,1]$ balances clean and adversarial supervision. This unified augmentation improves the model's ability to resist both structural manipulation and semantic noise prior to global aggregation.

Robustness-Aware Federated Aggregation

After completing the first half of local training under clean input conditions, each client C_k introduces adversarial perturbations (structural and textual) during the second half. Let $\mathbf{p}_{i,k}^{\text{clean}}$ and $\mathbf{p}_{i,k}^{\text{adv}} \in \mathbb{R}^C$ denote the prediction distributions before and after perturbation for node $v_{i,k}$, and $\mathcal{V}_k^{\text{eval}}$ be a held-out local evaluation set. The robustness score γ_k is defined as the average prediction consistency:

$$\gamma_k = \frac{1}{|\mathcal{V}_k^{\text{eval}}|} \sum_{v_{i,k} \in \mathcal{V}_k^{\text{eval}}} \cos\left(\mathbf{p}_{i,k}^{\text{clean}}, \mathbf{p}_{i,k}^{\text{adv}}\right), \tag{14}$$

where cosine similarity is used to measure prediction alignment across clean and adversarial conditions. A higher γ_k indicates greater local resilience to perturbation.

Weighted Model Aggregation. During global aggregation, the server computes a normalized robustness-aware weight ω_k for each client:

$$\omega_k = \frac{\gamma_k}{\sum_{j=1}^{K} \gamma_j}, \tag{15}$$

Backbone	Methods	Structure Attack	Text Attack	Cora		Pubmed		Ogbn-Arxiv		Amz-Computers		Amz-Sports	
				Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR
LLaGA	Vanilla	✗	✗	$81.81_{\pm 0.56}$	––	$86.21_{\pm 0.73}$	––	$62.61_{\pm 0.41}$	––	$70.81_{\pm 0.38}$	––	$80.26_{\pm 0.52}$	––
	SGA	✓	✗	$68.22_{\pm 0.94}$	13.59	$70.39_{\pm 0.62}$	15.82	$53.47_{\pm 0.79}$	9.14	$58.41_{\pm 0.88}$	12.40	$74.53_{\pm 0.81}$	5.73
	FGA	✓	✗	$65.88_{\pm 0.96}$	15.93	$66.54_{\pm 0.97}$	19.67	$50.77_{\pm 0.84}$	11.84	$55.93_{\pm 0.74}$	14.88	$70.86_{\pm 0.85}$	9.40
	HLBB	✗	✓	$62.41_{\pm 0.47}$	19.40	$65.01_{\pm 0.53}$	21.20	$47.19_{\pm 0.38}$	15.42	$53.69_{\pm 0.42}$	17.12	$67.92_{\pm 0.33}$	12.34
	TextHoaxer	✗	✓	$59.58_{\pm 0.53}$	22.23	$62.28_{\pm 0.59}$	23.93	$44.91_{\pm 0.40}$	17.70	$56.20_{\pm 0.50}$	14.61	$65.14_{\pm 0.42}$	15.12
	GTAE (ours)	✓	✓	$\mathbf{57.34_{\pm 0.61}}$	**24.47**	$\mathbf{48.59_{\pm 0.59}}$	**37.62**	$\mathbf{41.72_{\pm 0.63}}$	**20.89**	$\mathbf{50.38_{\pm 0.66}}$	**20.43**	$\mathbf{64.89_{\pm 0.71}}$	**15.37**
GraphPrompter	Vanilla	✗	✗	$65.74_{\pm 0.62}$	––	$91.25_{\pm 0.55}$	––	$61.16_{\pm 0.48}$	––	$71.45_{\pm 0.41}$	––	$71.98_{\pm 0.49}$	––
	SGA	✓	✗	$63.56_{\pm 0.97}$	2.18	$84.78_{\pm 1.06}$	6.47	$60.57_{\pm 1.02}$	0.59	$60.66_{\pm 0.99}$	10.79	$66.42_{\pm 1.00}$	5.56
	FGA	✓	✗	$57.35_{\pm 0.96}$	8.39	$79.93_{\pm 1.01}$	11.32	$56.89_{\pm 0.94}$	4.27	$57.34_{\pm 0.87}$	14.11	$66.92_{\pm 0.99}$	5.06
	HLBB	✗	✓	$59.78_{\pm 0.45}$	5.96	$79.11_{\pm 0.51}$	12.14	$53.13_{\pm 0.38}$	8.03	$61.02_{\pm 0.47}$	10.43	$62.98_{\pm 0.50}$	9.00
	TextHoaxer	✗	✓	$56.89_{\pm 0.50}$	8.85	$76.05_{\pm 0.55}$	15.20	$50.01_{\pm 0.46}$	11.15	$58.31_{\pm 0.52}$	13.14	$59.53_{\pm 0.51}$	12.45
	GTAE (ours)	✓	✓	$\mathbf{51.90_{\pm 0.69}}$	**13.84**	$\mathbf{67.54_{\pm 0.64}}$	**23.71**	$\mathbf{39.36_{\pm 0.53}}$	**21.80**	$\mathbf{53.83_{\pm 0.61}}$	**17.62**	$\mathbf{51.85_{\pm 0.67}}$	**20.13**

(a) Attack Results

Backbone	Methods	Structure Defense	Text Defense	Cora		Pubmed		Ogbn-Arxiv		Amz-Computers		Amz-Sports	
				Acc.	Imp.	Acc.	Imp.	Acc.	Imp.	Acc.	Imp.	Acc.	Imp.
LLaGA	GTAE (ours)	✗	✗	$57.34_{\pm 0.61}$	––	$48.59_{\pm 0.59}$	––	$41.72_{\pm 0.63}$	––	$50.38_{\pm 0.66}$	––	$64.89_{\pm 0.71}$	––
	ProGNN	✓	✗	$74.26_{\pm 0.59}$	16.92	$62.78_{\pm 0.52}$	14.19	$49.87_{\pm 0.48}$	8.15	$61.78_{\pm 0.51}$	11.40	$74.27_{\pm 0.55}$	9.38
	SimPGCN	✓	✗	$73.14_{\pm 0.53}$	15.80	$60.47_{\pm 0.51}$	11.88	$50.49_{\pm 0.46}$	8.77	$61.45_{\pm 0.50}$	11.07	$72.12_{\pm 0.54}$	7.23
	SAFER	✗	✓	$78.35_{\pm 0.60}$	21.01	$64.12_{\pm 0.53}$	15.53	$53.25_{\pm 0.45}$	11.53	$64.37_{\pm 0.48}$	13.99	$75.87_{\pm 0.57}$	10.98
	STRUM (ours)	✓	✓	$\mathbf{81.47_{\pm 0.55}}$	**24.13**	$\mathbf{84.49_{\pm 0.52}}$	**35.90**	$\mathbf{60.23_{\pm 0.49}}$	**18.51**	$\mathbf{68.71_{\pm 0.50}}$	**18.33**	$\mathbf{78.53_{\pm 0.53}}$	**13.64**
GraphPrompter	GTAE (ours)	✗	✗	$51.90_{\pm 0.69}$	––	$67.54_{\pm 0.64}$	––	$39.36_{\pm 0.53}$	––	$53.83_{\pm 0.61}$	––	$51.85_{\pm 0.67}$	––
	ProGNN	✓	✗	$51.37_{\pm 0.55}$	−0.53	$82.04_{\pm 0.50}$	14.50	$44.83_{\pm 0.44}$	5.47	$58.34_{\pm 0.52}$	4.51	$62.22_{\pm 0.48}$	10.37
	SimPGCN	✓	✗	$50.48_{\pm 0.47}$	−1.42	$82.71_{\pm 0.52}$	15.17	$43.89_{\pm 0.42}$	4.53	$57.95_{\pm 0.50}$	4.12	$59.33_{\pm 0.49}$	7.48
	SAFER	✗	✓	$56.83_{\pm 0.50}$	4.93	$86.41_{\pm 0.53}$	18.87	$49.05_{\pm 0.48}$	9.69	$61.96_{\pm 0.51}$	8.13	$64.58_{\pm 0.50}$	12.73
	STRUM (ours)	✓	✓	$\mathbf{61.47_{\pm 0.54}}$	**9.57**	$\mathbf{89.14_{\pm 0.52}}$	**21.60**	$\mathbf{60.15_{\pm 0.50}}$	**20.79**	$\mathbf{67.24_{\pm 0.51}}$	**13.41**	$\mathbf{67.84_{\pm 0.53}}$	**15.99**

(b) Defense Results

Table 1: Node classification performance under both attack and defense scenarios. All results are averaged over five independent runs. The best results are highlighted in bold and underlined, respectively. The default baseline model used in all experiments is Llama-3.1-8B. Additional experimental settings and extended results are provided in Appendix B.

which biases the update rule to favor clients with higher empirical robustness. The aggregated model parameters $\mathcal{W}_{\text{global}}$ are then computed as:

$$\mathcal{W}_{\text{global}} \leftarrow \sum_{k=1}^{K} \omega_k \cdot \mathcal{W}_k. \tag{16}$$

By explicitly incorporating client-side robustness estimation into the optimization loop, STRUM adaptively suppresses the influence of vulnerable or compromised clients, thereby improving the global model's overall resilience.

Experiment

This section presents a thorough evaluation of our proposed attack (GTAE) and defense (STRUM) strategies, focusing on the following key questions.

- **Q1:** How severely can GTAE degrade performance in TAG-FGL scenarios?
- **Q2:** Is STRUM sufficiently robust to preserve accuracy in the presence of adversarial attacks?
- **Q3:** Do both the structural and textual components of GTAE contribute significantly to the attack success?
- **Q4:** How does the proposed framework perform under varying hyper-parameter settings?

Experimental Setup

Datasets We leverage five datasets to evaluate the validity of proposed method: Cora (Mammen 2021), Pubmed (Shchur et al. 2018), Ogbn-Arxiv (Hu et al. 2020), Ama-Computers, Ama-Sports (McAuley et al. 2015). Detailed information about the datasets is provided in Appendix A.

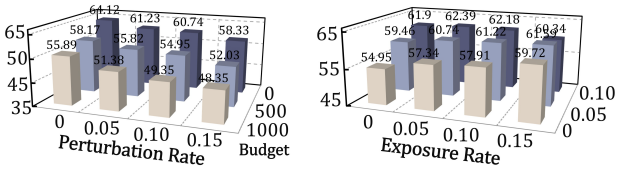
Implemention Details We adopt LLaMA-2-7B (Grattafiori et al. 2024) as the default backbone language model, while Mistral-7B (Jiang et al. 2023a) and ChatGLM-6B (GLM et al. 2024) are included for cross-model evaluation. The graph encoder is implemented as a 4-layer GAT with 1024 hidden units. Model optimization utilizes the AdamW algorithm with an initial learning rate of 1×10^{-5} (increased to 1×10^{-4} for translator-based variants), cosine learning rate decay, a weight decay of 0.05, and gradient clipping set to 0.1. Training is performed with a batch size of 12 and two steps of gradient accumulation (Kingma 2014). We follow a single-shot federated learning setup with a default configuration of five clients. Each client is trained locally for up to 5 epochs, and no additional global fine-tuning is applied. The final performance is reported on the test set using micro-averaged accuracy. Additional implementation details can be found in Appendix B.

Structure Attack	Text Attack	Cora		Pubmed		Ogbn-Arxiv	
		Acc.	ASR	Acc.	ASR	Acc.	ASR
✗	✗	81.81	–	86.21	–	62.61	–
✓	✗	63.85	18.06	68.19	18.02	51.27	11.34
✗	✓	67.31	14.60	53.79	32.42	44.38	18.23
✓	✓	**57.34**	**24.47**	**48.59**	**37.62**	**41.72**	**20.89**

Table 2: **Ablation on key components** for GTAE on Cora, Pubmed and Ogbn-Arxiv under Non-IID-Louvain.

Structure Defense	Text Defense	Cora		Pubmed		Ogbn-Arxiv	
		Acc.	Imp.	Acc.	Imp.	Acc.	Imp.
✗	✗	57.34	–	48.59	–	41.72	–
✓	✗	67.88	10.54	56.42	7.83	48.49	6.77
✗	✓	74.27	16.93	77.15	28.56	57.38	15.66
✓	✓	**81.47**	**24.13**	**84.49**	**35.90**	**60.23**	**18.51**

Table 3: **Ablation on key components** for STRUM on Cora, Pubmed and Ogbn-Arxiv under Non-IID-Louvain.



(a) Attack (b) Defense

Figure 3: **Three-dimensional bar chart** delineates the performance of our method under different hyperparameter settings in both the attack (left) and defense (right) scenarios.

Attack Results (Q1)

As summarised in Table 1, we report results on two representative backbones, LLaGA and GraphPrompter. **(1) GTAE is consistently the most destructive multimodal attack.** For example, when LLaGA is used on PubMed, GTAE attains an attack-success rate of 37.62%, exceeding the second-strongest baseline by 13.69%. **(2) Strong transferability across datasets and models.** The same level of degradation appears on heterogeneous graphs such as Ama-Sports and also on the structurally different backbone GraphPrompter, which demonstrates that combining topological and lexical perturbations produces a transferable, modality-agnostic threat that consistently amplifies the vulnerability of TAG-FGL systems.

Defense Results (Q2)

We evaluate the robustness of our defense method STRUM under the strongest multimodal attack GTAE across both backbones. **(1) STRUM achieves superior defense performance across all datasets.** On the PubMed dataset with LLaGA, STRUM recovers accuracy to 93.14%, significantly outperforming the best baseline SAFER by over 6.35 percentage points. A similar trend is observed on other graphs such as Ogbn-Arxiv and Ama-Computers, demonstrating the general effectiveness of our modality-aware defense. **(2) Robustness generalises to different model architectures.** STRUM maintains competitive robustness even when applied to a fundamentally different backbone, GraphPrompter. For instance, on Ama-Sports, it achieves 81.84% accuracy, outperforming SAFER and other GNN-based baselines. These results confirm that our defense mechanism not only mitigates multimodal threats but also generalises well across varied model structures.

Key Components (Q3)

We perform an ablation study to validate the necessity of each component in our attack and defense designs. **(1) Both attack channels are indispensable.** Removing either structure or text attack from GTAE leads to noticeable performance recovery. In contrast, enabling both components results in the highest attack success rates. For instance, on PubMed and Cora, the attack success rates reach 37.62% and 27.47%, respectively, demonstrating the complementary effect of the two modalities. **(2) Defense benefits from multimodal design.** On the defense side, equipping both structural and textual modules significantly enhances robustness across datasets. For example, the accuracy on Cora increases from 57.34% without defense to 81.47% with full defense, validating the effectiveness of perturbation-aware training and robustness-weighted aggregation.

Hyper-Parameters (Q4)

As shown in Figure 3, we conduct a hyperparameter ablation on the PubMed dataset using GraphTranslator as the backbone. In the attack setting (left), higher structure perturbation rates and budgets lead to greater accuracy drops. For instance, accuracy declines from 64.12% to 48.35% as the structure exposure rate increases from 0 to 0.15. In the defense setting (right), combining moderate structural exposure and text perturbation yields the best robustness. Accuracy peaks at 61.9% when the exposure rate is 0 and the text perturbation is 0.10. These results confirm that balanced adversarial augmentation across both modalities is crucial to robust TAG-FGL training.

Conclusion

In this paper, we investigate the security vulnerabilities of Text-Attributed Federated Graph Learning (TAG-FGL) by proposing a novel multimodal adversarial attack method, GTAE, and a corresponding defense framework, STRUM. GTAE coordinates structural and textual perturbations in a staged pipeline, revealing a previously underexplored compound threat that is highly transferable across datasets and backbones. To mitigate this, STRUM integrates modality-aware adversarial training with robustness-aware aggregation, effectively improving the system's resilience against both edge manipulation and textual substitution. Extensive experiments on real-world datasets confirm the effectiveness and generality of both our attack and defense designs. We hope this work inspires future research on trustworthy federated learning under multimodal and graph-centric settings.

Acknowledgements

This research is supported by the National Natural Science Foundation of China under Grant (62361166629), the Major Project of Science and Technology Innovation of Hubei Province (2024BCA003, 2025BEA002), the Innovative Research Group Project of Hubei Province under Grants (2024AFA017) and the National Natural Science Foundation of China under Grants (623B2080). The supercomputing system at the Supercomputing Center of Wuhan University supported the numerical calculations in this paper.

References

Alotaibi, A.; and Rassam, M. A. 2023. Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. *Future Internet*, 15(2): 62.

Chen, J.; Zhang, X.; and Zheng, H. 2024. Query-Efficient Adversarial Attack Against Vertical Federated Graph Learning. In *Attacks, Defenses and Testing for Deep Learning*, 75–96. Springer.

Chen, Z.; Mao, H.; Li, H.; Jin, W.; Wen, H.; Wei, X.; Wang, S.; Yin, D.; Fan, W.; Liu, H.; et al. 2024. Exploring the potential of large language models (llms) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 25(2): 42–61.

Fang, Y.; Fan, D.; Zha, D.; and Tan, Q. 2024. Gaugllm: Improving graph contrastive learning for text-attributed graphs with large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 747–758.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Han, S.; Park, S.; Wu, F.; Kim, S.; Zhu, B.; Xie, X.; and Cha, M. 2023. Towards attack-tolerant federated learning via critical parameter analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4999–5008.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33: 22118–22133.

Huang, H.; Zhang, L.; Sun, C.; Fang, R.; Yuan, X.; and Wu, D. 2023. Distributed pruning towards tiny neural networks in federated learning. In *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*, 190–201. IEEE.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; et al. 2023a. Mistral 7B.

Jiang, B.; and Li, Z. 2022. Defending against backdoor attack on graph nerual network by explainability. *arXiv preprint arXiv:2209.02902*.

Jiang, Z.; Xu, Y.; Xu, H.; Wang, Z.; Liu, J.; Chen, Q.; and Qiao, C. 2023b. Computation and communication efficient federated learning with adaptive model pruning. *IEEE Transactions on Mobile Computing*, 23(3): 2003–2021.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2): 1–210.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lei, R.; Hu, Y.; Ren, Y.; and Wei, Z. 2024. Intruding with words: Towards understanding graph injection attacks at the text level. *Advances in Neural Information Processing Systems*, 37: 49214–49251.

Li, M.; Zhang, P.; Xing, W.; Zheng, Y.; Zaporojets, K.; Chen, J.; Zhang, R.; Zhang, Y.; Gong, S.; Hu, J.; et al. 2025. Using Large Language Models to Tackle Fundamental Challenges in Graph Learning: A Comprehensive Survey. *arXiv preprint arXiv:2505.18475*.

Li, X.; Wu, Z.; Wu, J.; Cui, H.; Jia, J.; Li, R.-H.; and Wang, G. 2024. Graph learning in the era of llms: A survey from the perspective of data, models, and tasks. *arXiv preprint arXiv:2412.12456*.

Li, Y.; Li, Z.; Wang, P.; Li, J.; Sun, X.; Cheng, H.; and Yu, J. X. 2023. A survey of graph meets large language model: Progress and future directions. *arXiv preprint arXiv:2311.12399*.

Luo, J.; and Wu, S. 2022. Adapt to adaptation: Learning personalization for cross-silo federated learning. In *IJCAI: proceedings of the conference*, volume 2022, 2166. NIH Public Access.

Lyu, Y.; Li, C.; Zhang, X.; and Zhang, T. 2025. Navigating the Black Box: Leveraging LLMs for Effective Text-Level Graph Injection Attacks. *arXiv preprint arXiv:2506.13276*.

Mammen, P. M. 2021. Federated learning: Opportunities and challenges. *arXiv preprint arXiv:2101.05428*.

McAuley, J.; Targett, C.; Shi, Q.; and Van Den Hengel, A. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 43–52.

Moradi, M.; and Samwald, M. 2021. Evaluating the robustness of neural language models to input perturbations. *arXiv preprint arXiv:2108.12237*.

Pillutla, K.; Kakade, S. M.; and Harchaoui, Z. 2022a. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70: 1142–1154.

Pillutla, K.; Kakade, S. M.; and Harchaoui, Z. 2022b. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70: 1142–1154.

Ren, X.; Tang, J.; Yin, D.; Chawla, N.; and Huang, C. 2024. A survey of large language models for graphs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6616–6626.

Rong, X.; Huang, W.; Liang, J.; Bi, J.; Xiao, X.; Li, Y.; Du, B.; and Ye, M. 2025a. Backdoor Cleaning without External Guidance in MLLM Fine-tuning. *arXiv preprint arXiv:2505.16916*.

Rong, X.; Zhang, J.; He, K.; and Ye, M. 2025b. CAN: Leveraging Clients As Navigators for Generative Replay in Federated Continual Learning. In *Forty-second International Conference on Machine Learning*.

Rong, Y.; Huang, W.; Xu, T.; and Huang, J. 2019. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*.

Shang, W.; and Huang, X. 2024. A survey of large language models on generative graph analytics: Query, learning, and applications. *arXiv preprint arXiv:2404.14809*.

Shchur, O.; Mumme, M.; Bojchevski, A.; and Günnemann, S. 2018. Pitfalls of Graph Neural Network Evaluation. *Relational Representation Learning Workshop, NeurIPS 2018*.

Shi, Z.; Wan, G.; Huang, W.; Zhang, G.; Li, H.; Yang, C.; and Ye, M. 2025a. EAGLES: Towards Effective, Efficient, and Economical Federated Graph Learning via Unified Sparsification. In *Forty-second International Conference on Machine Learning*.

Shi, Z.; Wan, G.; Huang, W.; Zhang, G.; Shao, J.; Ye, M.; and Yang, C. 2025b. Privacy-Enhancing Paradigms within Federated Multi-Agent Systems. *arXiv preprint arXiv:2503.08175*.

Shi, Z.; Wan, G.; Wang, H.; Li, R.; Huang, Z.; Zhao, W.; Xiao, Y.; Luo, X.; Yang, C.; Sun, Y.; et al. 2025c. Don't Forget the Enjoin: FocalLoRA for Instruction Hierarchical Alignment in Large Language Models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Sun, Y.; Wang, S.; Tang, X.; Hsieh, T.-Y.; and Honavar, V. 2020. Adversarial attacks on graph neural networks via node injections: A hierarchical reinforcement learning approach. In *Proceedings of the Web Conference 2020*, 673–683.

Tian, Y.; Song, H.; Wang, Z.; Wang, H.; Hu, Z.; Wang, F.; Chawla, N. V.; and Xu, P. 2024. Graph neural prompting with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19080–19088.

Wan, G.; Cheng, X.; Liu, R.; Huang, W.; Shi, Z.; Jin, P.; Zhang, G.; Du, B.; and Ye, M. 2025a. Multi-order Orchestrated Curriculum Distillation for Model-Heterogeneous Federated Graph Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Wan, G.; Shi, Z.; Huang, W.; Zhang, G.; Tao, D.; and Ye, M. 2025b. Energy-based backdoor defense against federated graph learning. In *The Thirteenth International Conference on Learning Representations*.

Wan, G.; Sun, L.; Dou, L.; Shi, Z.; Wu, F.; Jiang, E. H.; Huang, W.; Zhang, G.; Geng, H.; Tang, X.; et al. 2025c. Diagnose, Localize, Align: A Full-Stack Framework for Reliable LLM Multi-Agent Systems under Instruction Conflicts. *arXiv preprint arXiv:2509.23188*.

Wang, B.; Xu, C.; Liu, X.; Cheng, Y.; and Li, B. 2022. SemAttack: Natural textual attacks via different semantic spaces. *arXiv preprint arXiv:2205.01287*.

Wang, S.; Huang, J.; Chen, Z.; Song, Y.; Tang, W.; Mao, H.; Fan, W.; Liu, H.; Liu, X.; Yin, D.; et al. 2024. Graph machine learning in the era of large language models (llms). *ACM Transactions on Intelligent Systems and Technology*.

Wang, Y.; Dai, X.; Fan, W.; and Ma, Y. 2025. Exploring graph tasks with pure llms: A comprehensive benchmark and investigation. *arXiv preprint arXiv:2502.18771*.

Xu, H.; Ma, Y.; Liu, H.-C.; Deb, D.; Liu, H.; Tang, J.-L.; and Jain, A. K. 2020. Adversarial attacks and defenses in images, graphs and text: A review. *International journal of automation and computing*, 17: 151–178.

Yang, Q.; Liu, Y.; Chen, T.; and Tong, Y. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2): 1–19.

Yang, Y.; Li, Q.; Jia, J.; Hong, Y.; and Wang, B. 2024. Distributed backdoor attacks on federated graph learning and certified defenses. *arXiv preprint arXiv:2407.08935*.

Ye, M.; Miao, C.; Wang, T.; and Ma, F. 2022. Texthoaxer: Budgeted hard-label adversarial attacks on text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3877–3884.

Ye, M.; Rong, X.; Huang, W.; Du, B.; Yu, N.; and Tao, D. 2025. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint arXiv:2502.14881*.

Zhang, K.; Wang, Y.; Wang, H.; Huang, L.; Yang, C.; Chen, X.; and Sun, L. 2022. Efficient federated learning on knowledge graphs via privacy-preserving relation embedding aggregation. *arXiv preprint arXiv:2203.09553*.

Zhang, W. E.; Sheng, Q. Z.; Alhazmi, A.; and Li, C. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3): 1–41.

Zhang, X.; and Zitnik, M. 2020. Gnnguard: Defending graph neural networks against adversarial attacks. *Advances in neural information processing systems*, 33: 9263–9275.

Zhang, Y.; Ramage, D.; Xu, Z.; Zhang, Y.; Zhai, S.; and Kairouz, P. 2023a. Private federated learning in gboard. *arXiv preprint arXiv:2306.14793*.

Zhang, Z.; Li, H.; Zhang, Z.; Qin, Y.; Wang, X.; and Zhu, W. 2023b. Graph meets llms: Towards large graph models. *arXiv preprint arXiv:2308.14522*.

Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; and Sun, M. 2020. Graph neural networks: A review of methods and applications. *AI open*, 1: 57–81.

Zhu, C.; Cheng, Y.; Gan, Z.; Sun, S.; Goldstein, T.; and Liu, J. 2019. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*.