# E²Gen: Evidential Energy-Based Generation for Fair Federated Graph Learning

**Yuxin Wu**[1,2,*], **Jingbo Wang**[1,*], **Zitong Shi**[3,*‡], **Zihan Tan**[1], **Qiqi Lin**[4]
**Xuankun Rong**[1], **Liangtao Zheng**[5], **Wenke Huang**[6], **Mang Ye**[1,†]
[1]Wuhan University [2]Renmin University of China
[3]National University of Singapore [4]Tsinghua University
[5]University of California, San Diego [6]Nanyang Technological University

## Abstract

Federated Graph Learning is rapidly evolving as a privacy-preserving collaborative approach for decentralized graph data. However, severe fairness challenges are increasingly undermining federated systems by systematically degrading predictions for structurally disadvantaged minority nodes. The inherent vulnerabilities and missing topological contexts in Federated Graph Learning are deeply entangled, making traditional federated fairness methods and simple oversampling less effective. In our work, we propose an effective Evidential Energy-based Generation framework for Fair Federated Graph Learning ($E^2Gen$). At the client level, it explicitly identifies structurally deficient nodes using a multi-axis metric, synthesizing targeted representations via conditional energy-based models, and selects reliable samples through an evidential quality gate. At the server level, the local performance disparities uploaded by each client are evaluated to construct a fairness gap assessment, making the global model absorb equitable improvements by further adjusting the aggregation weights. Our method can handle high topological heterogeneity, does not require strict generative normalization, and is effective under both homophilic and heterophilic graph structures. Extensive results on various settings of federated graph scenarios under severe fairness challenges validate the effectiveness of this approach.

## 1 Introduction

Federated Learning (FL) (McMahan et al., 2017) has fundamentally transformed decentralized machine learning. It allows multiple clients to collaboratively train a shared model while keeping sensitive data strictly local, making it a highly effective solution for privacy-preserving applications (Huang et al., 2026; Shi et al., 2025b). This paradigm naturally extends to graph-structured data, forming the basis of Federated Graph Learning (FGL) (Tan et al., 2024; Shi et al., 2025a; Qian et al., 2025). While FGL successfully leverages distributed graphs in domains like healthcare, finance, and social networks, its decentralized nature introduces severe fairness challenges. In the federated context, fairness requires the global model to deliver equitable predictive performance across diverse demographic groups. When this principle is violated, the model systematically degrades predictions for minority entities, potentially leading to discriminatory outcomes in high-stakes risk assessments.

With the objective of mitigating these disparities, fairness-aware mechanisms in traditional FL have been widely studied (Li et al., 2019; Mohri et al., 2019). Certain methods prioritize minority groups by directly reweighting local objective functions. Some approaches dynamically adjust server aggregation weights or utilize personalized learning to balance performance across heterogeneous clients (Li et al., 2021b). However, these methods often struggle to perform effectively in FGL environments

*Equal contribution. ‡Project Leader. †Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

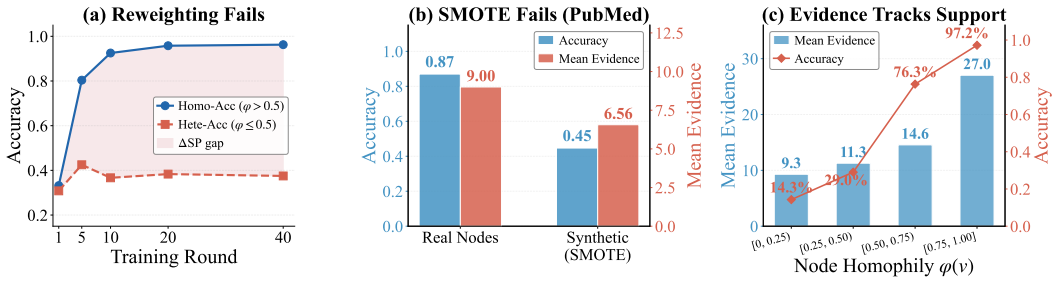


Figure 2: Empirical observations on the inherent vulnerabilities of disadvantaged nodes in FGL. **(a) Reweighting Fails:** Using the class-level reweighting method BoostFGL, the accuracy gap between advantaged (homophilous, $\varphi > 0.5$) and disadvantaged ($\varphi \leq 0.5$) nodes still persists and widens to 0.59 at round 40. **(b) SMOTE Fails:** Traditional latent interpolation (e.g., FedSig) generates unreliable representations, causing accuracy to plummet from 0.87 to 0.45 and mean evidence to drop from 9.00 to 6.56. **(c) Evidence Tracks Reliability:** The quantified mean evidence monotonically increases alongside node homophily ($\varphi$) and accuracy, confirming it as a valid metric for support deficiency. For detailed experimental setups and parameter configurations, please refer to Appendix B.

due to the unique properties of graph data. Unlike traditional machine learning that assumes independent data instances, graph learning fundamentally relies on message passing mechanisms (Kipf and Welling, 2016), where a node's embedding is recursively updated by aggregating information from its neighbors. This interconnected nature dictates that a node's learning quality is deeply entangled with its local context. Consequently, within this message-passing paradigm, disadvantaged nodes face critical inherent vulnerabilities. They lack the high-quality local patterns and informative connections necessary to aggregate meaningful messages. Since standard FL fairness techniques operate solely by adjusting local loss functions or server aggregation weights and remain inherently agnostic to these underlying node-level deficits, they cannot mitigate these vulnerabilities, thus proving ineffective for graph-based fairness issues.

Based on the aforementioned discussion, we review the challenges existing in FGL fairness. First of all, to address performance disparities, many existing fairness-aware FGL methods heavily modify the local objective function by reweighting losses or assigning higher gradient penalties to underperforming nodes (Wu et al., 2025; Chen et al., 2026). However, amplifying a poor learning signal cannot yield meaningful features if the underlying informative context is fundamentally missing. As empirically demonstrated in Figure 2(a), relying solely on reweighting mechanisms (e.g., BoostFGL) fails to bridge the performance gap between advantaged and disadvantaged nodes. This inability to create new meaningful representations inherently limits the effectiveness of reweighting strategies. This leads to our first question: 1) *How can we explicitly identify disadvantaged nodes and synthetically generate targeted representations to compensate for their inherent vulnerabilities?*

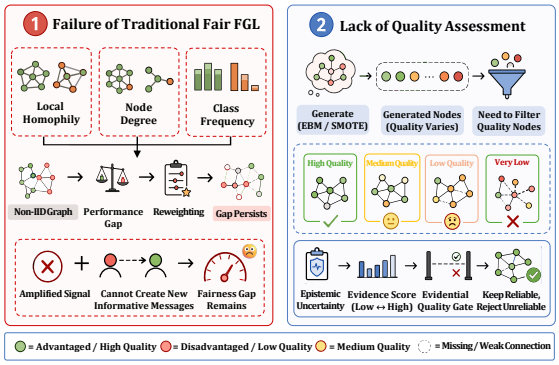


Figure 1: **Problem Illustration.** We describe the challenges in fair FGL: **I)** Traditional reweighting methods merely amplify existing signals without creating new informative messages, failing to close the fairness gap. **II)** Generation-based methods lack quality assessment, necessitating an evidential quality gate to filter out unreliable generated nodes.

To address this first question, we introduce the **Deficit Score**, a comprehensive diagnostic metric that decomposes a node's inherent vulnerabilities along three complementary axes: **Local Homophily, Node Degree, and Class Frequency**. Instead of relying on a single coarse indicator like local training loss (Tang et al., 2020; Zhu et al., 2020), this metric evaluates the multi-dimensional context along these axes to produce a calibrated aggregate score. By elevating fairness diagnosis from macro performance tracking to explicit, fine-grained evaluation, the Deficit Score provides a stable, reproducible signal to pinpoint precisely which disadvantaged nodes require targeted augmentation. We empirically validate the necessity and effectiveness of this fused multi-axis formulation through

comprehensive ablation studies, as detailed in Section 5.3. Subsequently, to generate higher-quality synthetic representations for these identified nodes, we turn to **energy-based models (EBMs)** and explore their potential. Energy is fundamentally an unnormalized probability likelihood (LeCun et al., 2006), offering a flexible modeling approach that is not constrained by strict normalization. The core strength of EBMs lies in their architectural versatility, allowing them to be integrated with virtually any model architecture. In our work, we combine energy functions with Graph Convolutional Networks (GCN) (Kipf and Welling, 2016) to construct an Energy-based GCN. This design preserves the GCN's capability to capture complex relational information while benefiting from the flexibility of energy-based modeling to stably generate plausible synthetic latent representations. By unifying the Deficit Score and this generative capability, we establish a robust, targeted augmentation foundation to effectively compensate for the inherent vulnerabilities of disadvantaged nodes.

Secondly, while generating synthetic data presents a potential solution to mitigate these inherent vulnerabilities, directly injecting these representations into local training introduces severe risks of negative transfer. For instance, some approaches adapt oversampling techniques like GraphSMOTE (Zhao et al., 2021) to the federated latent space (Bi et al., 2024). However, applying simple linear interpolation within a highly non-IID topological space often generates out-of-distribution noise or unreliable representations. As empirically shown in Figure 2(b), samples generated by simple interpolation suffer from severe semantic degradation and high epistemic uncertainty, which we quantify as a sharp decrease in predictive "evidence". Compounding this issue, standard softmax classifiers are notoriously overconfident on borderline synthetic samples (Guo et al., 2017), rendering them incapable of reliably assessing generation quality. This brings up our second question: 2) *How can we accurately quantify the epistemic uncertainty of generated graph representations to ensure only reliable samples are incorporated into the training process?*

To solve this second question regarding quality control (cf. Jiang et al., 2026b,c), we introduce **Evidential Quality Gating** (Sensoy et al., 2018). Motivated by the theoretical framework of subjective logic, we adopt an evidence vector $\boldsymbol{e} \in \mathbb{R}_+^K$ for a $K$-class classification task. Mathematically, this evidence vector is used to parameterize a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{p}|\boldsymbol{\alpha})$ over the categorical probabilities, where the concentration parameters are defined as $\boldsymbol{\alpha} = \boldsymbol{e} + \mathbf{1}$. Unlike standard softmax outputs that only provide overconfident point estimates, this evidence-parameterized distribution can be utilized to explicitly quantify the model's epistemic uncertainty—essentially capturing the model's "lack of knowledge" regarding an unfamiliar representation. As validated in Figure 2(c), this theoretical evidence metric exhibits a strong positive correlation with a node's actual reliability, making it an ideal diagnostic signal. Therefore, we evaluate every EBM-generated sample through this gate, dynamically downweighting or rejecting synthetic representations that exhibit low total evidence. Finally, to align local improvements with global fairness, we introduce a fairness gap-driven server aggregation mechanism that proportionally rewards clients who successfully close their local performance disparities. By seamlessly integrating the Deficit Score, the local EBM generative foundation, and the Evidential Quality Gate, we propose $\mathbf{E}^2$**Gen**, an **E**vidential **E**nergy-based **Gen**eration framework for Fair FGL.[1] Our principal contributions are summarized as follows.

- We introduce the **Deficit Score**, a fused diagnostic metric that explicitly quantifies node-level inherent vulnerabilities across complementary multi-dimensional axes. Furthermore, we empirically demonstrate that traditional loss reweighting and SMOTE-like latent interpolation fail to reliably mitigate these fundamental vulnerabilities.
- We propose $\mathbf{E}^2$**Gen**, a novel generative framework for fair FGL. It leverages conditional EBMs to synthesize targeted representations for disadvantaged nodes and employs an evidential quality gate to strictly filter out unreliable generations, ensuring robust and fair local training.
- We evaluate $\mathbf{E}^2$**Gen** on seven benchmarks covering both homophilic and heterophilic graphs. The results demonstrate that $\mathbf{E}^2$**Gen** establishes new state-of-the-art performance, significantly outperforming existing standard and fairness-aware FGL baselines across various non-IID settings.

[1] Code is available at https://github.com/Anti1i/E-Gen.

## 2 Related Work

### 2.1 Federated Graph Learning

Federated Graph Learning (FGL) enables collaborative training on graph-structured data while preserving privacy (Wan et al., 2025a; Guo et al., 2026). Recent research primarily addresses missing cross-client links and structural heterogeneity to improve generalization (Zhang et al., 2021; Li et al., 2024a; Tan et al., 2025b,a). However, the inherent heterogeneity and complex topological dependencies (Tan et al., 2026; Liu et al., 2026) often create significant disparities in the informative context available to different node groups, making fairness a critical but under-explored challenge.

### 2.2 Fairness in Federated Learning

To tackle performance disparities across demographic groups, traditional FL fairness methods typically employ adaptive objective reweighting (Li et al., 2019), minimax optimization (Mohri et al., 2019), or personalized models (Li et al., 2021b). However, these approaches struggle in FGL scenarios where disparities are deeply rooted in structural biases, such as sparse connections and heterophilic neighborhoods. While recent FGL fairness studies (Chen et al., 2026; Wu et al., 2025) exist, they primarily focus on modifying local loss weights or gradients. These reweighting mechanisms cannot synthesize missing structural information for vulnerable nodes, leaving a critical gap in repairing structural support (Agarwal et al., 2021; Dong et al., 2022).

### 2.3 Energy-based Model

The Energy-Based Model (EBM) models the unnormalized probability density function by defining an energy function (LeCun et al., 2006; Wan et al., 2025c, 2026). Its key advantage lies in its structural flexibility, which effectively mitigates the mode collapse issues common in other generative paradigms (Du and Mordatch, 2019). In the graph domain, EBMs have successfully been applied to tasks like substructure-preserving molecule design (Liu et al., 2021), demonstrating powerful modeling capabilities in discrete topological spaces. We leverage this generative capability to synthesize plausible representations for structurally deficient nodes in a highly non-IID environment.

## 3 Preliminary

### 3.1 Federated Graph Learning

We follow the standard Subgraph Federated Learning paradigm (He et al., 2021), where $K$ clients $\mathcal{C} = \{c_k\}_{k=1}^{K}$ collaboratively train a shared GNN $f_\theta$ for the node classification task. Each client $c_k$ holds a private local graph $\mathcal{G}_k = (\mathcal{V}_k, \mathcal{E}_k)$ with node features $\mathbf{X}_k$ and labels $\mathbf{Y}_k$. In each communication round, the server distributes the global model weights. Clients then optimize $f_\theta$ locally over their labeled nodes and upload the updated weights back to the server for standard aggregation mechanisms (e.g., FedAvg (McMahan et al., 2017)).

### 3.2 Energy-based Model

An Energy-Based Model (EBM) (LeCun et al., 2006) is a flexible framework that captures the underlying data distribution by associating a scalar energy value with each data point. For a continuous latent representation $h$, the unnormalized probability density function is defined as:

$$p_\psi(h) = \frac{\exp(-E_\psi(h))}{Z_\psi} \tag{1}$$

where $E_\psi(h) : \mathbb{R}^d \to \mathbb{R}$ is the energy function parameterized by a neural network, and $Z_\psi = \int \exp(-E_\psi(h))dh$ is the intractable partition function. To bypass the direct computation of $Z_\psi$, EBMs are typically trained via Contrastive Divergence (Hinton, 2002). This objective shapes the energy landscape by assigning low energy to real samples and high energy to synthesized ones, which are generated by navigating towards low-energy regions via Langevin dynamics (Du and Mordatch, 2019).

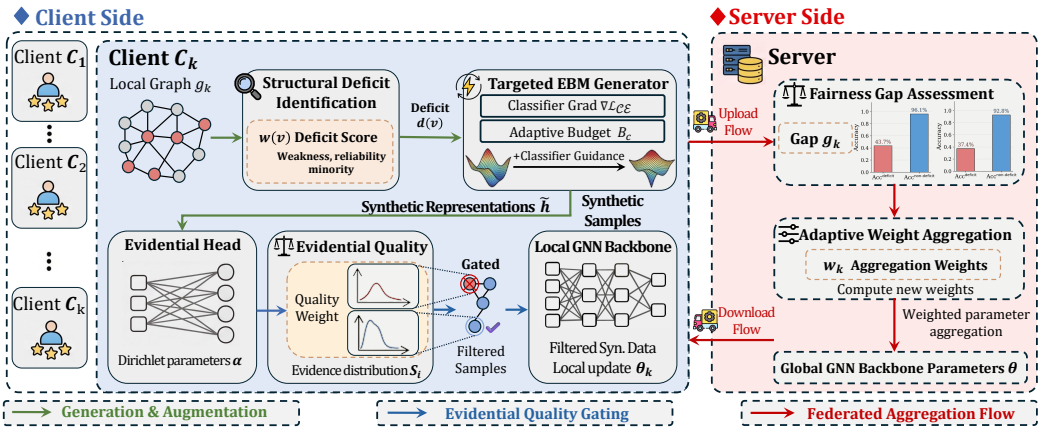


Figure 3: Architecture illustration of E$^2$Gen. We use green, blue, and red arrows to represent the three main flows of our method: Generation & Augmentation, Evidential Quality Gating, and Federated Aggregation Flow, respectively. Best viewed in color. Zoom in for details.

### 3.3 Evidential Deep Learning

Evidential Deep Learning (EDL) explicitly quantifies epistemic uncertainty to overcome the over-confidence issues of standard Softmax classifiers, especially on out-of-distribution samples. For a $C$-class classification problem, the model outputs a non-negative evidence vector $\mathbf{e} \in \mathbb{R}^C_{\geq 0}$, which parameterizes a Dirichlet distribution with parameters $\alpha_c = e_c + 1$. The predicted probability $p_c$ for class $c$ and the epistemic uncertainty $u$ are computed as:

$$p_c = \frac{\alpha_c}{S}, \quad u = \frac{C}{S} \tag{2}$$

where $S = \sum_{c=1}^{C} \alpha_c$ is the total accumulated evidence. A higher total evidence $S$ indicates a confident prediction, whereas a low $S$ yields high uncertainty $u$, naturally acting as a fail-safe metric to flag unreliable synthetic instances.

## 4 Methodology

### 4.1 Framework Overview

Figure 3 illustrates the overall **E$^2$Gen** framework. On the client side, a local conditional energy-based model (EBM) generates synthetic representations for structurally disadvantaged nodes, which we define as **deficit nodes**. To maintain semantic correctness during local training, an evidential quality gate filters these generated samples. On the server side, the system evaluates the fairness gap between deficit and non-deficit nodes across clients. It then dynamically adjusts aggregation weights to guide the global model toward more equitable performance.

### 4.2 Structural Deficit Identification

**Motivation.** Our empirical observations (Section 1) show that fairness issues in federated graph learning (FGL) largely arise from a multifaceted lack of training support. Current fairness-aware methods (Dai and Wang, 2021) usually rely on predefined sensitive attributes or coarse global metrics. These approaches often overlook specific topological vulnerabilities, such as noisy aggregation (Wan et al., 2025b; Shi et al., 2026; Jiang et al., 2026a), sparse connections, and label imbalance, which put individual nodes at a structural disadvantage. To resolve this disparity, we first need a way to quantify it.

**Definition 4.1 (Structural Deficit Score).** We define a continuous deficit score $w(v) \in [0, 1]$ for each node $v$. Rather than a pure vulnerability metric, $w(v)$ acts as a deficit-priority score with reliability calibration to capture a node's structural deficiency while avoiding topological noise. It combines three structural dimensions:

$$w(v) = \underbrace{\big(1 - \phi(v)\big)}_{\text{structural weakness}} \times \underbrace{\big(1 - e^{-\deg(v)/\tau_d}\big)}_{\text{reliability}} \times \underbrace{m(v)}_{\text{class minority}}, \tag{3}$$

Here, $\phi(v) = \frac{|\{(v,u)\in\mathcal{E}_k : y_v = y_u\}|}{\deg(v)}$ represents the local homophily ratio. The term $(1 - e^{-\deg(v)/\tau_d})$ serves as a reliability calibration factor. Since homophily estimates for extremely low-degree nodes are highly unstable and prone to high variance, this term explicitly suppresses unreliable topological estimates, preventing the generative model from being misguided by local noise. $m(v) = \frac{\max_c f_c}{f_{y_v}}$ (scaled to $[0, 1]$) measures class rarity, where $f_c$ is the frequency of class $c$ in the client's local graph.If a node's score $w(v)$ exceeds the median of all nonzero scores, we label it a deficit node ($d(v) = 1$). This data-driven approach pinpoints exactly where we need to apply client-side augmentation, ensuring we allocate our generative budget to the most topologically vulnerable yet reliably measurable regions.

### 4.3 Deficit-Targeted Energy-Based Generation

**Motivation.** Once deficit nodes are identified, an intuitive solution is to synthesize representations to compensate for their lack of training support. However, in heterogeneous federated graph environments, traditional generative paradigms like GANs or VAEs often suffer from training instability and mode collapse when dealing with highly discrete topological features. Energy-Based Models (EBMs) present a flexible alternative. By directly modeling the unnormalized probability density (the energy landscape) of the GNN's hidden representations, EBMs can smoothly and stably capture complex structural distributions without strict normalization constraints.

**Conditional Energy-Based Model.** For each client $k$, we build a local conditional EBM $E_\psi$ over the hidden representations $\mathbf{h} = f_\theta^{(1)}(\mathbf{X}, \mathcal{E})$ from the first GNN layer. To ensure structurally aware generation, the energy function is conditioned on both the class label $y$ and the deficit status $d$:

$$E_\psi(\mathbf{h} \mid y, d) = \mathrm{MLP}\big([\mathbf{h} \parallel \mathbf{e}_y \parallel \mathbf{e}_d]\big). \tag{4}$$

The EBM is optimized via Contrastive Divergence (CD) (Du and Mordatch, 2019). This objective explicitly shapes the energy landscape by assigning low energy to real representations drawn from the data distribution $p_{\text{data}}$ and high energy to synthesized anomalies from $p_\psi$:

$$\mathcal{L}_{\text{EBM}} = \mathbb{E}_{\mathbf{h}\sim p_{\text{data}}}\big[E_\psi(\mathbf{h} \mid y, d)\big] - \mathbb{E}_{\tilde{\mathbf{h}}\sim p_\psi}\big[E_\psi(\tilde{\mathbf{h}} \mid y, d)\big] + \lambda_{\text{reg}}\mathbb{E}\big[E_\psi^2\big]. \tag{5}$$

Negative samples $\tilde{\mathbf{h}}$ are warm-started from real representations with small Gaussian perturbations and refined by Langevin MCMC; the full negative-sample initialization, step-size schedule, EBM warm-up, and stability tricks (gradient clipping, energy regularization) are deferred to Section D.1.

**Classifier-Guided Langevin Generation.** To synthesize representations that are both distributionally plausible and semantically accurate, we employ Langevin dynamics (Welling and Teh, 2011; Dhariwal and Nichol, 2021) augmented with classifier guidance. The sampling process iteratively navigates the feature space toward low-energy and class-consistent regions:

$$\mathbf{h}_{t+1} = \mathbf{h}_t - \eta\nabla_{\mathbf{h}}E_\psi(\mathbf{h}_t \mid y, d) - \gamma\nabla_{\mathbf{h}}\mathcal{L}_{\text{CE}}\big(f_\theta(\mathbf{h}_t), y\big) + \sigma\boldsymbol{\epsilon}_t. \tag{6}$$

To explicitly bridge the fairness gap, we initialize the Markov chains exclusively from **deficit seed representations** ($d = 1$). Furthermore, we allocate an **adaptive budget** $B_c$ for each class $c$, which scales dynamically with its average structural weakness $\bar{w}_c$ and the class imbalance ratio:

$$B_c = B_{\max} \cdot \big(0.5 + 0.3 \cdot \bar{w}_c + 0.2 \cdot \min(\sqrt{n_{\max}/n_c},\, 3)\big). \tag{7}$$

By concentrating generative efforts exclusively on structurally vulnerable and minority-class regions, this targeted augmentation effectively repairs the local representation space, laying a balanced foundation before the global aggregation phase.

### 4.4 Evidential Quality Gating and Local Objective

**Motivation.** Although EBM sampling ensures that the synthesized representations conform to a valid structural distribution, Langevin dynamics may inadvertently push samples across complex decision boundaries. If we directly inject these samples into local training, it risks negative transfer. Furthermore, standard Softmax classifiers are known to be overconfident on out-of-distribution or borderline samples. To guarantee semantic correctness, we must rigorously quantify the **epistemic uncertainty** of the generated representations. Evidential learning perfectly fulfills this role: it not only provides classification probabilities but also quantifies the model's confidence. In scenarios of

extreme label skew where a client almost entirely lacks a certain class, the evidential model naturally yields low total evidence (high uncertainty) for that class, thereby acting as a fail-safe to confidently gate out unreliable generations.

**Evidential Quality Gating.** Following the framework of evidential deep learning (Sensoy et al., 2018), we augment our backbone GNN with an evidential classification head, which models the predictive distribution as a Dirichlet distribution parameterized by $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha} = \mathbf{e} + \mathbf{1}, \quad p(y = c \mid \mathbf{x}) = \frac{\alpha_c}{S}, \quad S = \textstyle\sum_{c=1}^{C} \alpha_c \tag{8}$$

where $\mathbf{e}$ represents the evidence derived from the network output, and $S$ indicates the total accumulated evidence. The evidential head is supervised by a Type-II MLE (digamma) loss with an annealed Dirichlet KL regularizer, used as an auxiliary calibration term on top of class-balanced cross-entropy; the full objective is given in Section D.2. For each synthetic sample $(\tilde{\mathbf{h}}_i, \tilde{y}_i)$, we design a continuous evidence gate that computes a quality weight $q_i$ to evaluate both its label agreement and intrinsic evidence quality as follows:

$$q_i = \underbrace{p(\hat{y}_i = \tilde{y}_i)}_{\text{label agreement}} \times \underbrace{\sigma\Big(\frac{S_i - \tau}{\delta}\Big)}_{\text{evidence quality}} \tag{9}$$

where $\tau$ is an evidence threshold and $\delta$ is a temperature scaling factor.

**Local Training Objective.** The final local training objective harmonizes the real structural data with the high-quality synthetic data. For real training nodes, we compute the cross-entropy loss $\mathcal{L}_{\text{real}}$ utilizing class-balanced and deficit-aware sample weights $\bar{w}(v)$:

$$\mathcal{L}_{\text{real}} = \frac{1}{|\mathcal{V}_{\text{train}}|} \sum_{v \in \mathcal{V}_{\text{train}}} \bar{w}(v) \cdot \ell_{\text{CE}}\big(f_\theta(\mathbf{x}_v, \mathcal{E}),\ y_v\big) \tag{10}$$

For EBM-generated samples, the loss $\mathcal{L}_{\text{syn}}$ is re-weighted by the evidential quality score $q_i$:

$$\mathcal{L}_{\text{syn}} = \frac{\sum_i q_i \cdot \text{invfreq}(\tilde{y}_i) \cdot \ell_{\text{CE}}\big(f_\theta(\tilde{\mathbf{h}}_i),\ \tilde{y}_i\big)}{\sum_i q_i \cdot \text{invfreq}(\tilde{y}_i)} \tag{11}$$

To further reduce deficit–non-deficit disparity during local optimization, we add a lightweight fairness regularizer $\mathcal{L}_{\text{fair}} = \mathcal{L}_{\text{parity}}(\mathcal{V}_{\text{def}}, \mathcal{V}_{\text{nondef}})$, instantiated as a soft, one-sided loss-disparity penalty that lifts the deficit group with the non-deficit term stop-gradient detached (full form in Section D.3). The total local loss is integrated as:

$$\mathcal{L}_k = \mathcal{L}_{\text{real}} + \lambda_{\text{syn}} \alpha_r \mathcal{L}_{\text{syn}} + \lambda_{\text{fair}} \mathcal{L}_{\text{fair}}, \tag{12}$$

where $\alpha_r$ is a warmup coefficient. By strictly gating the synthetic data based on evidential uncertainty, we ensure that only semantically reliable representations contribute to the local update.

### 4.5 Fairness-Aware Server Aggregation

**Energy Disparity Aggregation.** Standard FedAvg aggregation weights clients purely by data size, ignoring heterogeneous fairness conditions. We adopt a fairness gap-driven aggregation strategy. At round $r$, each client first computes a scalar fairness gap $g_k = \max(0, Acc_k^{\text{non-deficit}} - Acc_k^{\text{deficit}})$. The server then amplifies client $k$ by $\phi_k = 1 + 0.5\, g_k / (\max_j g_j + \epsilon)$ and combines it with the data-size prior $n_k$. The final normalized aggregation weight is

$$\alpha_k = \frac{n_k\, \phi_k}{\sum_j n_j\, \phi_j}, \qquad \mathbf{w}^{r+1} = \sum_{k=1}^{K} \alpha_k\, \mathbf{w}_k^r. \tag{13}$$

Here, $g_k$ is only a high-level scalar performance statistic, rather than raw node features or local samples. This mechanism adaptively upweights clients with larger deficit–nondeficit gaps and helps the global model absorb fairer improvements while remaining stable.

## 5 Experiments

We conducted experiments under both homophilic and heterophilic settings on seven widely-used node classification benchmarks to validate the superiority of our proposed $\mathbf{E}^2$Gen.

Table 1: Comparison with **state-of-the-art** methods for fair federated graph learning over seven benchmark datasets under Non-IID Louvain partitioning. The upper table presents overall predictive performance, and the lower table reports fairness metrics. The best and second-best results are highlighted with bold and underline, respectively. Please see additional analysis in Section 5.2.

| Methods | Cora | | CiteSeer | | PubMed | | Coauthor-CS | | Coauthor-Phy | | Chameleon | | Squirrel | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{A}$ | $\mathcal{F}$ | $\mathcal{A}$ | $\mathcal{F}$ | $\mathcal{A}$ | $\mathcal{F}$ | $\mathcal{A}$ | $\mathcal{F}$ | $\mathcal{A}$ | $\mathcal{F}$ | $\mathcal{A}$ | $\mathcal{F}$ | $\mathcal{A}$ | $\mathcal{F}$ |
| *Standard FL baselines* | | | | | | | | | | | | | | |
| FedAvg | 82.64 | 82.13 | 72.06 | 65.82 | 86.05 | 85.64 | 91.11 | 88.14 | 94.88 | 93.27 | 53.64 | 52.46 | 40.60 | 39.54 |
| MOON | 82.37$_{\downarrow 0.27}$ | 81.84$_{\downarrow 0.29}$ | 70.66$_{\downarrow 1.40}$ | 66.38$_{\uparrow 0.56}$ | 85.67$_{\downarrow 0.38}$ | 85.23$_{\downarrow 0.41}$ | 91.02$_{\downarrow 0.09}$ | 87.96$_{\downarrow 0.18}$ | 94.82$_{\downarrow 0.06}$ | 93.12$_{\downarrow 0.15}$ | 54.05$_{\uparrow 0.41}$ | 52.64$_{\uparrow 0.18}$ | 41.73$_{\uparrow 1.13}$ | 40.68$_{\uparrow 1.14}$ |
| *Federated Graph Learning baselines* | | | | | | | | | | | | | | |
| FedGTA | 82.28$_{\downarrow 0.36}$ | 81.77$_{\downarrow 0.36}$ | 73.24$_{\uparrow 1.18}$ | 69.28$_{\uparrow 3.46}$ | 85.99$_{\downarrow 0.06}$ | 85.53$_{\downarrow 0.11}$ | 91.08$_{\downarrow 0.03}$ | 88.04$_{\downarrow 0.10}$ | 94.93$_{\uparrow 0.05}$ | 93.32$_{\uparrow 0.05}$ | 54.05$_{\uparrow 0.41}$ | 52.08$_{\downarrow 0.38}$ | 41.17$_{\uparrow 0.57}$ | 39.97$_{\uparrow 0.43}$ |
| FedSage+ | 81.38$_{\downarrow 1.26}$ | 80.49$_{\downarrow 1.64}$ | 70.73$_{\downarrow 1.33}$ | 64.88$_{\downarrow 0.94}$ | 77.14$_{\downarrow 8.91}$ | 74.28$_{\downarrow 11.36}$ | 89.46$_{\downarrow 1.65}$ | 84.13$_{\downarrow 4.01}$ | 94.37$_{\downarrow 0.51}$ | 92.68$_{\downarrow 0.59}$ | 54.47$_{\uparrow 0.83}$ | 54.49$_{\uparrow 2.03}$ | 35.15$_{\downarrow 5.45}$ | 29.37$_{\downarrow 10.17}$ |
| AdaFGL | 81.92$_{\downarrow 0.72}$ | 81.54$_{\downarrow 0.59}$ | 71.25$_{\downarrow 0.81}$ | 65.49$_{\downarrow 0.33}$ | 86.30$_{\uparrow 0.25}$ | 85.79$_{\uparrow 0.15}$ | 90.71$_{\downarrow 0.40}$ | 87.60$_{\downarrow 0.54}$ | 94.92$_{\uparrow 0.04}$ | 93.31$_{\uparrow 0.04}$ | 55.09$_{\uparrow 1.45}$ | 52.93$_{\uparrow 0.47}$ | 41.26$_{\uparrow 0.66}$ | 40.26$_{\uparrow 0.72}$ |
| FedTAD | 82.46$_{\downarrow 0.18}$ | 81.65$_{\downarrow 0.48}$ | 70.29$_{\downarrow 1.77}$ | 61.20$_{\downarrow 4.62}$ | 85.87$_{\downarrow 0.18}$ | 85.36$_{\downarrow 0.28}$ | 90.73$_{\downarrow 0.38}$ | 87.68$_{\downarrow 0.46}$ | 94.54$_{\downarrow 0.34}$ | 92.89$_{\downarrow 0.38}$ | 55.30$_{\uparrow 1.66}$ | 52.27$_{\downarrow 0.19}$ | 42.39$_{\uparrow 1.79}$ | 41.29$_{\uparrow 1.75}$ |
| *Fairness-aware FGL baselines* | | | | | | | | | | | | | | |
| FairFGL | 76.80$_{\downarrow 5.84}$ | 76.84$_{\downarrow 5.29}$ | 67.11$_{\downarrow 4.95}$ | 61.89$_{\downarrow 3.93}$ | 83.15$_{\downarrow 2.90}$ | 82.85$_{\downarrow 2.79}$ | 88.96$_{\downarrow 2.15}$ | 85.64$_{\downarrow 2.50}$ | 93.47$_{\downarrow 1.41}$ | 91.36$_{\downarrow 1.91}$ | 52.60$_{\downarrow 1.04}$ | 51.16$_{\downarrow 1.30}$ | 40.79$_{\uparrow 0.19}$ | 40.64$_{\uparrow 1.10}$ |
| BoostFGL | 82.28$_{\downarrow 0.36}$ | 81.73$_{\downarrow 0.40}$ | 70.44$_{\downarrow 1.62}$ | 64.07$_{\downarrow 1.75}$ | 86.59$_{\uparrow 0.54}$ | 86.13$_{\uparrow 0.49}$ | 91.08$_{\downarrow 0.03}$ | 88.20$_{\uparrow 0.06}$ | 94.99$_{\uparrow 0.11}$ | 93.41$_{\uparrow 0.14}$ | 52.18$_{\downarrow 1.46}$ | 52.02$_{\downarrow 0.44}$ | 40.98$_{\uparrow 0.38}$ | 39.06$_{\downarrow 0.48}$ |
| *Oversampling baselines* | | | | | | | | | | | | | | |
| RandomOS | 80.04$_{\downarrow 2.60}$ | 79.74$_{\downarrow 2.39}$ | 69.70$_{\downarrow 2.36}$ | 64.13$_{\downarrow 1.69}$ | 85.97$_{\downarrow 0.08}$ | 85.60$_{\downarrow 0.04}$ | 90.37$_{\downarrow 0.74}$ | 87.34$_{\downarrow 0.80}$ | 94.25$_{\downarrow 0.63}$ | 92.35$_{\downarrow 0.92}$ | 56.13$_{\uparrow 2.49}$ | 55.59$_{\uparrow 3.13}$ | 42.20$_{\uparrow 1.60}$ | 41.24$_{\uparrow 1.70}$ |
| BalancedOS | 81.56$_{\downarrow 1.08}$ | 81.08$_{\downarrow 1.05}$ | 72.80$_{\uparrow 0.74}$ | 66.90$_{\uparrow 1.08}$ | 85.31$_{\downarrow 0.74}$ | 84.96$_{\downarrow 0.68}$ | 90.90$_{\downarrow 0.21}$ | 87.98$_{\downarrow 0.16}$ | 94.75$_{\downarrow 0.13}$ | 93.14$_{\downarrow 0.13}$ | 54.89$_{\uparrow 1.25}$ | 53.32$_{\uparrow 0.86}$ | 42.01$_{\uparrow 1.41}$ | 41.49$_{\uparrow 1.95}$ |
| FedSig | 82.10$_{\downarrow 0.54}$ | 81.60$_{\downarrow 0.53}$ | 73.10$_{\uparrow 1.04}$ | 69.23$_{\uparrow 3.41}$ | 86.45$_{\uparrow 0.40}$ | 86.00$_{\uparrow 0.36}$ | 90.88$_{\downarrow 0.23}$ | 87.70$_{\downarrow 0.44}$ | 94.74$_{\downarrow 0.14}$ | 93.09$_{\downarrow 0.18}$ | 53.43$_{\downarrow 0.21}$ | 51.08$_{\downarrow 1.38}$ | 40.70$_{\uparrow 0.10}$ | 40.51$_{\uparrow 0.97}$ |
| **E$^2$Gen (Ours)** | **82.91**$_{\uparrow 0.27}$ | **82.25**$_{\uparrow 0.12}$ | **73.76**$_{\uparrow 1.70}$ | **70.24**$_{\uparrow 4.42}$ | **86.97**$_{\uparrow 0.92}$ | **86.57**$_{\uparrow 0.93}$ | **91.25**$_{\uparrow 0.14}$ | **88.78**$_{\uparrow 0.64}$ | **95.01**$_{\uparrow 0.13}$ | **93.58**$_{\uparrow 0.31}$ | **58.21**$_{\uparrow 4.57}$ | **57.78**$_{\uparrow 5.32}$ | **43.14**$_{\uparrow 2.54}$ | **43.04**$_{\uparrow 3.50}$ |

| Methods | Cora | | CiteSeer | | PubMed | | Coauthor-CS | | Coauthor-Phy | | Chameleon | | Squirrel | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{H}$ | $\mathcal{B}$ | $\mathcal{H}$ | $\mathcal{B}$ | $\mathcal{H}$ | $\mathcal{B}$ | $\mathcal{H}$ | $\mathcal{B}$ | $\mathcal{H}$ | $\mathcal{B}$ | $\mathcal{H}$ | $\mathcal{B}$ | $\mathcal{H}$ | $\mathcal{B}$ |
| *Standard FL baselines* | | | | | | | | | | | | | | |
| FedAvg | 34.19 | 79.79 | 32.56 | 67.59 | 51.81 | 85.84 | 41.44 | 86.08 | 39.85 | 92.87 | **57.83** | 58.57 | 40.91 | 42.28 |
| MOON | 34.71$_{\uparrow 0.52}$ | 80.07$_{\uparrow 0.28}$ | 32.72$_{\uparrow 0.16}$ | 67.78$_{\uparrow 0.19}$ | 51.49$_{\downarrow 0.32}$ | 85.79$_{\downarrow 0.05}$ | 42.82$_{\uparrow 1.38}$ | 86.01$_{\downarrow 0.07}$ | 39.52$_{\downarrow 0.33}$ | 93.21$_{\uparrow 0.34}$ | 56.70$_{\downarrow 1.13}$ | 57.57$_{\downarrow 1.00}$ | 40.67$_{\downarrow 0.24}$ | 41.42$_{\downarrow 0.86}$ |
| *Federated Graph Learning baselines* | | | | | | | | | | | | | | |
| FedGTA | 34.99$_{\uparrow 0.80}$ | 80.04$_{\uparrow 0.25}$ | 32.69$_{\uparrow 0.13}$ | 67.71$_{\uparrow 0.12}$ | 51.74$_{\downarrow 0.07}$ | 85.68$_{\downarrow 0.16}$ | 42.87$_{\uparrow 1.43}$ | 85.54$_{\downarrow 0.54}$ | 39.35$_{\downarrow 0.50}$ | 92.90$_{\uparrow 0.03}$ | 55.18$_{\downarrow 2.65}$ | 56.02$_{\downarrow 2.55}$ | 43.07$_{\uparrow 2.16}$ | 44.14$_{\uparrow 1.86}$ |
| FedSage+ | 36.81$_{\uparrow 2.62}$ | 79.32$_{\downarrow 0.47}$ | 32.61$_{\uparrow 0.05}$ | 67.59$_{\downarrow 0.00}$ | 49.07$_{\downarrow 2.74}$ | 83.51$_{\downarrow 2.33}$ | 43.51$_{\uparrow 2.07}$ | 85.97$_{\downarrow 0.11}$ | 39.42$_{\downarrow 0.43}$ | 91.84$_{\downarrow 1.03}$ | 50.75$_{\downarrow 7.08}$ | 52.29$_{\downarrow 6.28}$ | 30.46$_{\downarrow 10.45}$ | 34.36$_{\downarrow 7.92}$ |
| AdaFGL | 36.38$_{\uparrow 2.19}$ | 79.46$_{\downarrow 0.33}$ | 34.52$_{\uparrow 1.96}$ | 68.24$_{\uparrow 0.65}$ | 52.00$_{\uparrow 0.19}$ | 85.88$_{\uparrow 0.04}$ | 42.13$_{\uparrow 0.69}$ | 86.36$_{\uparrow 0.28}$ | 41.69$_{\uparrow 1.84}$ | 93.24$_{\uparrow 0.37}$ | 46.14$_{\downarrow 11.69}$ | 48.76$_{\downarrow 9.81}$ | 38.10$_{\downarrow 2.81}$ | 39.91$_{\downarrow 2.37}$ |
| FedTAD | 36.43$_{\uparrow 2.24}$ | 80.02$_{\uparrow 0.23}$ | 32.49$_{\downarrow 0.07}$ | 67.46$_{\downarrow 0.13}$ | 51.41$_{\downarrow 0.40}$ | 84.71$_{\downarrow 1.13}$ | 41.69$_{\uparrow 0.25}$ | 85.15$_{\downarrow 0.93}$ | 40.39$_{\uparrow 0.54}$ | 92.89$_{\uparrow 0.02}$ | 55.68$_{\downarrow 2.15}$ | 56.64$_{\downarrow 1.93}$ | 43.84$_{\uparrow 2.93}$ | 44.52$_{\uparrow 2.24}$ |
| *Fairness-aware FGL baselines* | | | | | | | | | | | | | | |
| FairFGL | 34.89$_{\uparrow 0.70}$ | 76.13$_{\downarrow 3.66}$ | 31.82$_{\downarrow 0.74}$ | 63.93$_{\downarrow 3.66}$ | 53.01$_{\uparrow 1.20}$ | 82.69$_{\downarrow 3.15}$ | 41.94$_{\uparrow 0.50}$ | 83.80$_{\downarrow 2.28}$ | 37.01$_{\downarrow 2.84}$ | 89.70$_{\downarrow 3.17}$ | 54.32$_{\downarrow 3.51}$ | 53.99$_{\downarrow 4.58}$ | 42.65$_{\uparrow 1.74}$ | 43.66$_{\uparrow 1.38}$ |
| BoostFGL | 35.23$_{\uparrow 1.04}$ | 79.90$_{\uparrow 0.11}$ | 31.56$_{\downarrow 1.00}$ | 67.46$_{\downarrow 0.13}$ | 52.27$_{\uparrow 0.46}$ | 85.78$_{\downarrow 0.06}$ | 42.47$_{\uparrow 1.03}$ | 86.32$_{\uparrow 0.24}$ | 40.82$_{\uparrow 0.97}$ | 92.93$_{\uparrow 0.06}$ | 54.46$_{\downarrow 3.37}$ | 56.26$_{\downarrow 2.31}$ | 40.67$_{\downarrow 0.24}$ | 41.61$_{\downarrow 0.67}$ |
| *Oversampling baselines* | | | | | | | | | | | | | | |
| RandomOS | 37.02$_{\uparrow 2.83}$ | 79.05$_{\downarrow 0.74}$ | 32.60$_{\uparrow 0.04}$ | 66.84$_{\downarrow 0.75}$ | 52.27$_{\uparrow 0.46}$ | 85.17$_{\downarrow 0.67}$ | 41.18$_{\downarrow 0.26}$ | 84.74$_{\downarrow 1.34}$ | 41.01$_{\uparrow 1.16}$ | 92.15$_{\downarrow 0.72}$ | 57.34$_{\downarrow 0.49}$ | 58.73$_{\uparrow 0.16}$ | 41.30$_{\uparrow 0.39}$ | 43.66$_{\uparrow 1.38}$ |
| BalancedOS | 37.22$_{\uparrow 3.03}$ | 79.71$_{\downarrow 0.08}$ | 35.35$_{\uparrow 2.79}$ | 68.57$_{\uparrow 0.98}$ | 50.68$_{\downarrow 1.13}$ | 84.62$_{\downarrow 1.22}$ | 42.97$_{\uparrow 1.53}$ | 86.00$_{\downarrow 0.08}$ | 40.88$_{\uparrow 1.03}$ | 93.28$_{\uparrow 0.41}$ | 56.82$_{\downarrow 1.01}$ | 58.38$_{\downarrow 0.19}$ | 40.10$_{\downarrow 0.81}$ | 42.08$_{\downarrow 0.20}$ |
| FedSig | 33.87$_{\downarrow 0.32}$ | 80.84$_{\uparrow 1.05}$ | 33.04$_{\uparrow 0.48}$ | 68.01$_{\uparrow 0.42}$ | 52.96$_{\uparrow 1.15}$ | 85.47$_{\downarrow 0.37}$ | 41.82$_{\downarrow 0.38}$ | 86.10$_{\uparrow 0.02}$ | 42.75$_{\uparrow 2.90}$ | 93.75$_{\uparrow 0.88}$ | 55.56$_{\downarrow 2.27}$ | 56.20$_{\downarrow 2.37}$ | 39.70$_{\downarrow 1.21}$ | 40.85$_{\downarrow 1.43}$ |
| **E$^2$Gen (Ours)** | **40.69**$_{\uparrow 6.50}$ | **81.70**$_{\uparrow 1.91}$ | **36.76**$_{\uparrow 4.20}$ | **70.01**$_{\uparrow 2.42}$ | **54.40**$_{\uparrow 2.59}$ | **86.11**$_{\uparrow 0.27}$ | **45.16**$_{\uparrow 3.72}$ | **87.08**$_{\uparrow 1.00}$ | **43.28**$_{\uparrow 3.43}$ | **93.85**$_{\uparrow 0.98}$ | 57.45$_{\downarrow 0.38}$ | **58.77**$_{\uparrow 0.20}$ | **44.78**$_{\uparrow 3.87}$ | **45.47**$_{\uparrow 3.19}$ |

## 5.1 Experimental Setup

**Datasets.** We evaluate **E**$^2$Gen across three distinct scenarios encompassing seven datasets: Citation Networks (Yang et al., 2016), Co-authorship Networks (Shchur et al., 2018), and Wikipedia Page-link Networks (Rozemberczki et al., 2021). Detailed information, structural statistics, and dataset split settings are provided in Section A.

**Comparison Methods.** We compare **E**$^2$Gen with 11 state-of-the-art methods: (1) Standard FL: FedAvg (McMahan et al., 2017), MOON (Li et al., 2021a). (2) Standard FGL: FedGTA (Li et al., 2024b), FedSage+ (Zhang et al., 2021), AdaFGL (Li et al., 2024a), FedTAD (Zhu et al., 2024). (3) Fairness-aware FGL: FairFGL (Wu et al., 2025), BoostFGL (Chen et al., 2026). (4) Oversampling methods: RandomOS, BalancedOS, FedSig (Bi et al., 2024). Detailed descriptions can be found in section C.

**Network Structure.** Following common FGL approaches, all evaluated methods utilize a standard 2-layer GCN backbone. For our **E**$^2$Gen, the local EBM uses a 3-layer MLP trained for exactly 3 epochs per round via contrastive divergence.

**Federated Setting.** Following the standard FGL protocol (Chen et al., 2026), we partition each graph into $K$=5 client subgraphs using Louvain community detection (Blondel et al., 2008), which naturally induces non-IID structural and label distributions across clients.

**Implement Details & Evaluation Metric.** All experiments run for 100 communication rounds with 3 local training epochs per round (hardware and hyperparameter tuning details are deferred to Section A). We evaluate the global model across all clients from two perspectives. For overall predictive performance, we report **Accuracy** ($\mathcal{A}$) and **Macro-F1** ($\mathcal{F}$). For fairness evaluation, we measure **Hete-F1** ($\mathcal{H}$) and **Balanced Accuracy** ($\mathcal{B}$) to assess the equity across different structural or demographic groups (Brodersen et al., 2010).

### 5.2 Experimental Results

**Performance Comparison.** Table 1 presents the comprehensive comparison. $\mathbf{E}^2$Gen establishes new state-of-the-art results, achieving the best performance on **all 14 predictive metrics** (upper table) and **13 out of 14 fairness metrics** (lower table). It demonstrates consistent superiority over standard FGL methods on predictive tasks (e.g., +4.57pp Acc on Chameleon, +2.54pp Acc on Squirrel over FedAvg).

Regarding fairness evaluation, $\mathbf{E}^2$Gen yields substantial improvements over baselines, such as +6.50pp and +4.20pp in $\mathcal{H}$ on Cora and CiteSeer, respectively. Fairness-aware methods like FairFGL show limited effectiveness due to coarse-grained reweighting, and frequently sacrifice overall predictive accuracy for marginal fairness gains. BoostFGL's model-predicted difficulty is also less reliable than $\mathbf{E}^2$Gen's topology-derived deficit scores. Furthermore, oversampling baselines (e.g., FedSig) degrade on heterophilic graphs since interpolation between dissimilar neighbors produces incoherent features. In contrast, $\mathbf{E}^2$Gen's energy-based latent generation naturally produces diverse, high-quality samples, successfully optimizing both overall accuracy and structural fairness simultaneously.

**Graph Structures Adaptability.** $\mathbf{E}^2$Gen naturally adapts to different graph structures. On homophilic graphs, the synthetic data weight ($\lambda_{\text{syn}}$) tends to be higher (0.5–1.5), as the EBM can reliably generate class-consistent representations. On heterophilic graphs, lower $\lambda_{\text{syn}}$ values (0.03–0.5) reflect the increased difficulty of generation. The evidence gate complements this $\lambda_{\text{syn}}$-based modulation by performing per-sample, continuous re-weighting on a calibrated evidence signal: the empirical filter intensity scales smoothly with dataset complexity (from $9\%$ floored on Coauthor-Phy to $89\%$ on Squirrel), as quantified in Section F.

### 5.3 Diagnostic Analysis

**Key Components.** We conducted an ablation study on the core components of $\mathbf{E}^2$Gen using the Cora and Co-CS datasets. The results, demonstrating the effectiveness of each module, are presented in Figure 4(a) and (b). Generative augmentation (*w/o Gen*) contributes most significantly by compensating for fundamental structural scarcity. The evidential quality gate (*w/o Gate*) prevents negative transfer by filtering noisy synthetic samples, while the fairness objective (*w/o* $\mathcal{L}_{fair}$) is crucial to fully align the global model toward equitable performance. Further analyses isolating the gate from the generator and comparing the EBM against alternative generator architectures (Mixup, CVAE, GAN) are provided in Section E.

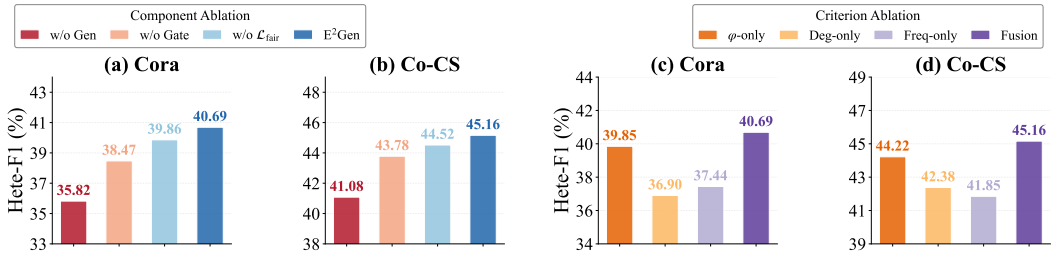


Figure 4: **Ablation Analysis** of components and criteria for $\mathbf{E}^2$Gen on the Cora and Co-CS datasets. The results illustrate the fairness performance (Hete-F1) under Component Ablation (a-b) and Criterion Ablation (c-d).

**Deficit Score Formulation.** We compared our full composite metric (*Fusion*) with single-factor diagnostic variants, including local homophily (*φ-only*), node degree (*Deg-only*), and class frequency (*Freq-only*). The results are presented in Figure 4(c) and (d). Relying on a single topological feature is insufficient to accurately identify disadvantaged nodes, leading to notable performance drops. While local homophily serves as a relatively strong indicator, it still falls short of our multi-axis approach. In contrast, the *Fusion* strategy consistently achieves the highest fairness performance, providing a more accurate and stable diagnostic signal for targeted structural repair.

**Hyper Parameters and Scalability.** We investigate $\mathbf{E}^2$Gen's sensitivity to key loss-weighting hyperparameters, including the synthetic data weight ($\lambda_{\text{syn}}$), the evidential auxiliary weight ($\lambda_{\text{evi}}$), and the fairness weight ($\lambda_{\text{fair}}$). As detailed in Appendix E.3, our framework demonstrates strong robustness across homophilic and medium-scale datasets. Across extensive hyperparameter ranges, the performance fluctuations remain remarkably low (less than 0.62% in both Accuracy and Macro-F1). This indicates that $\mathbf{E}^2$Gen is generally insensitive to these loss-weighting choices, and principled default values can achieve near-optimal performance without exhaustive tuning. Scalability-wise,

$\mathbf{E}^2$Gen introduces a moderate overhead ($\sim$35% wall-clock time increase per round vs. FedAvg on Coauthor-Physics). This overhead is amortized across communication rounds and remains highly efficient, as the EBM generation and evidential gating operate entirely in the low-dimensional latent space ($d$=64–256) rather than the high-dimensional input feature space.

## 6 Conclusion and Discussion

In this paper, we innovatively explore the fundamental problem of structural fairness disparities in federated graph learning. We propose a novel framework called E²Gen, an Evidential Energy-based Generation approach for Fair FGL. At the client level, we identify vulnerable nodes using a multi-axis Deficit Score and synthesize targeted representations via conditional energy-based models, utilizing an evidential quality gate to strictly filter out unreliable samples. At the server level, we evaluate the local fairness gaps between deficit and non-deficit nodes and determine aggregation weights dynamically to reward clients who successfully mitigate their disparities. Despite introducing additional computational overhead, this method yields stable and robust fairness gains across multiple homophilic and heterophilic scenarios. We hope this work offers a novel generative perspective for future research on fair federated graph learning.

## Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China under Grants 62361166629 and 623B2080, the Independent Research Project of the National Key Laboratory of Big Data and Decision, the National Research Foundation (NRF), Singapore, under Grant ID NRF-PA2025-NTU, and Nanyang Technological University (NTU). The supercomputing system at the Supercomputing Center of Wuhan University supported the numerical calculations in this paper.

## References

Agarwal, C., Lakkaraju, H., and Zitnik, M. (2021). Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in artificial intelligence*, pages 2114–2124. PMLR.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Bi, B., Zhang, Z., Qiao, P., Yuan, Y., and Wang, G. (2024). Fedsig: a federated graph augmentation for class-imbalanced node classification. In *International Conference on Database Systems for Advanced Applications*, pages 474–490. Springer.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE.

Chen, Z., Yang, K., Li, X., Sun, H., Zhang, Z., Li, J., Dai, Q., Li, R.-H., and Wang, G. (2026). Boostfgl: Boosting fairness in federated graph learning. *arXiv preprint arXiv:2601.16496*.

Dai, E. and Wang, S. (2021). Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 680–688.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Dong, Y., Liu, N., Jalaian, B., and Li, J. (2022). Edits: Modeling and mitigating data bias for graph neural networks. In *Proceedings of the ACM web conference 2022*, pages 1259–1269.

Du, Y. and Mordatch, I. (2019). Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Guo, H., Tan, Z., Huang, W., Yang, B., and Ye, M. (2026). FedFST: Mitigating spectral catastrophic forgetting in federated graph continual learning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 1404–1414.

He, C., Balasubramanian, K., Ceyani, E., Yang, C., Xie, H., Sun, L., He, L., Yang, L., Yu, P. S., Rong, Y., et al. (2021). Fedgraphnn: A federated learning system and benchmark for graph neural networks. *arXiv preprint arXiv:2104.07145*.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Huang, W., Liao, Y., Lei, W., Wan, G., Rong, X., Wen, C., Li, H., Ye, M., Wu, Q., and Du, B. (2026). Federated clinical concept and disease semantic learning for congenital heart disease diagnosis. *npj Digital Medicine*, 9:567.

Jiang, W., Liang, K., Huang, W., Zhang, X., Xu, Z., Wan, G., Tan, C., Fan, F. X., and Wang, J. (2026a). Unveiling and mitigating untargeted poisoning attacks on federated knowledge graph embedding. In *Proceedings of the ACM Web Conference 2026*, pages 2569–2580.

Jiang, W., Liang, K., Rong, X., Zhou, J., Zhong, Z., Wan, G., and Wang, J. (2026b). PurMM: Attention-guided test-time backdoor purification in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35562–35570.

Jiang, W., Rong, X., Zhai, Y., Feng, D., and Wang, H. (2026c). ReFrame: Evidence-guided test-time safety alignment in multimodal large language models. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F., et al. (2006). A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Li, Q., He, B., and Song, D. (2021a). Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722.

Li, T., Hu, S., Beirami, A., and Smith, V. (2021b). Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368. PMLR.

Li, T., Sanjabi, M., Beirami, A., and Smith, V. (2019). Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*.

Li, X., Wu, Z., Zhang, W., Sun, H., Li, R.-H., and Wang, G. (2024a). Adafgl: A new paradigm for federated node classification with topology heterogeneity. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 2517–2530. IEEE.

Li, X., Wu, Z., Zhang, W., Zhu, Y., Li, R.-H., and Wang, G. (2024b). Fedgta: Topology-aware averaging for federated graph learning. *arXiv preprint arXiv:2401.11755*.

Liu, J., Tan, Z., Wan, G., Huang, W., Li, H., and Ye, M. (2026). FedSDR: Federated graph learning with structural noise detection and reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3379–3389.

Liu, M., Yan, K., Oztekin, B., and Ji, S. (2021). Graphebm: Molecular graph generation with energy-based models. *arXiv preprint arXiv:2102.00546*.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. Pmlr.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Mohri, M., Sivek, G., and Suresh, A. T. (2019). Agnostic federated learning. In *International conference on machine learning*, pages 4615–4625. PMLR.

Qian, J., Wan, G., Huang, W., Zhang, G., Wu, Y., Du, B., and Ye, M. (2025). GHOST: Generalizable one-shot federated graph learning with proxy-based topology knowledge retention. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 50047–50065. PMLR.

Rozemberczki, B., Allen, C., and Sarkar, R. (2021). Multi-scale attributed node embedding. *Journal of Complex Networks*, 9(2):cnab014.

Sensoy, M., Kaplan, L., and Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Shchur, O., Mumme, M., Bojchevski, A., and Günnemann, S. (2018). Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*.

Shi, Z., Wan, G., Huang, W., Wu, Y., Zhang, Q., and Ye, M. (2026). Towards robust text-attributed federated graph learning: Multimodal threats and defense. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 25383–25391.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 55046–55064. PMLR.

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*.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., and Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.

Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

Tan, Z., Huang, S., Wan, G., Huang, W., Li, H., and Ye, M. (2025a). S2FGL: Spatial spectral federated graph learning. *arXiv preprint arXiv:2507.02409*.

Tan, Z., Wan, G., Huang, W., Li, H., Zhang, G., Yang, C., and Ye, M. (2025b). FedSPA: Generalizable federated graph learning under homophily heterogeneity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15464–15475.

Tan, Z., Wan, G., Huang, W., Yang, B., and Ye, M. (2026). FedPRE: Robust federated graph learning against topological corruption. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1390–1400.

Tan, Z., Wan, G., Huang, W., and Ye, M. (2024). FedSSP: Federated graph learning with spectral knowledge and personalized preference. *Advances in Neural Information Processing Systems*, 37:34561–34581.

Tang, X., Yao, H., Sun, Y., Wang, Y., Tang, J., Aggarwal, C., Mitra, P., and Wang, S. (2020). Investigating and mitigating degree-related biases in graph convoltuional networks. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1435–1444.

Wan, G., Ran, F., Zhang, R., Huang, W., Rong, X., Zhang, G., Wu, Y., Du, B., and Ye, M. (2025a). MOTION: Multi-sculpt evolutionary coarsening for federated continual graph learning. In *Advances in Neural Information Processing Systems*, volume 38.

Wan, G., Shang, X., Wu, Y., Zhang, G., Bi, J., Zheng, L., Lin, X., Liu, Y., Ma, Y., Huang, W., and Du, B. (2025b). HYPERION: Fine-grained hypersphere alignment for robust federated graph learning. In *Advances in Neural Information Processing Systems*, volume 38.

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

Wan, G., Wu, Y., Shi, Z., Huang, W., Du, B., and Ye, M. (2026). Dual-view energy-topological backdoor defense for federated graph learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688.

Wu, Z., Pang, B., Li, X., Zhu, Y., Su, D., Fan, B., Li, R.-H., Wang, G., and Zhou, C. (2025). Towards unbiased federated graph learning: Label and topology perspectives. *arXiv preprint arXiv:2504.09963*.

Yang, Z., Cohen, W., and Salakhudinov, R. (2016). Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pages 40–48. PMLR.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, K., Yang, C., Li, X., Sun, L., and Yiu, S. M. (2021). Subgraph federated learning with missing neighbor generation. *Advances in neural information processing systems*, 34:6671–6682.

Zhao, T., Zhang, X., and Wang, S. (2021). Graphsmote: Imbalanced node classification on graphs with graph neural networks. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 833–841.

Zhu, J., Yan, Y., Zhao, L., Heimann, M., Akoglu, L., and Koutra, D. (2020). Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in neural information processing systems*, 33:7793–7804.

Zhu, Y., Li, X., Wu, Z., Wu, D., Hu, M., and Li, R.-H. (2024). Fedtad: Topology-aware data-free knowledge distillation for subgraph federated learning. *arXiv preprint arXiv:2404.14061*.

# A Datasets and Experimental Details

The statistics of the datasets used in our experiments are provided in Table 2.

- **Citation Networks** (Cora, CiteSeer, PubMed): These datasets consist of interconnected research papers where nodes represent studies and edges denote citation relationships. They are widely used homophilic benchmarks ($\bar{\phi} \geq 0.74$) for node classification tasks, providing valuable topology for evaluating standard graph learning paradigms (Yang et al., 2016).
- **Co-authorship Networks** (Coauthor-CS, Coauthor-Physics): Derived from the Microsoft Academic Graph (Sinha et al., 2015), nodes in these datasets represent authors, and edges signify co-author relationships. They exhibit highly homophilic structural characteristics ($\bar{\phi} \geq 0.81$) and are typically used to predict the respective research fields of authors (Shchur et al., 2018).
- **Wikipedia Page-link Networks** (Chameleon, Squirrel): In these datasets, nodes represent Wikipedia pages on specific topics, and edges represent mutual links between them. Unlike the previous datasets, these are highly heterophilic graphs ($\bar{\phi} \leq 0.23$), where nodes tend to connect with differing classes, providing a rigorous testbed for structurally diverse and challenging settings (Rozemberczki et al., 2021).

Table 2: Dataset statistics. $\bar{\phi}$ denotes the average node homophily ratio.

| Dataset | Nodes | Edges | Classes | $\bar{\phi}$ | Type |
|---|---|---|---|---|---|
| Cora | 2,708 | 10,556 | 7 | 0.81 | Homophilic |
| CiteSeer | 3,327 | 9,104 | 6 | 0.74 | Homophilic |
| PubMed | 19,717 | 88,648 | 3 | 0.80 | Homophilic |
| Coauthor-CS | 18,333 | 163,788 | 15 | 0.81 | Homophilic |
| Coauthor-Phy | 34,493 | 495,924 | 5 | 0.93 | Homophilic |
| Chameleon | 2,277 | 36,101 | 5 | 0.23 | Heterophilic |
| Squirrel | 5,201 | 217,073 | 5 | 0.22 | Heterophilic |

**Dataset Split:** For the Citation and Co-authorship networks, we use a train/validation/test split of 20%/40%/40%. For the highly heterophilic Wikipedia datasets, we employ a 48%/32%/20% split.

**Hardware and Hyperparameter Tuning:** All experiments are conducted on NVIDIA RTX 4090 GPUs. To ensure a fair comparison, we employ the Optuna framework (Akiba et al., 2019) for hyperparameter search, conducting 50 trials per dataset for our method. All baselines are tuned with equivalent effort.

# B Case Study Configurations and Analysis

This section provides the comprehensive experimental configurations and dataset statistics corresponding to the empirical observations presented in Figure 2.

**Specific Case Study Details:**

- **Figure 2(a) - Reweighting Fails:** Conducted on the Cora dataset (2,708 nodes, 7 classes). We apply BoostFGL, a state-of-the-art class-level reweighting baseline. The accuracy for homophilous ($\varphi > 0.5$) and heterophilous ($\varphi \leq 0.5$) nodes is recorded at rounds 1, 5, 10, 20, and 40. **Result Analysis:** The results reveal that the performance gap widens significantly (reaching 0.59 at round 40) despite the reweighting mechanism. This demonstrates that merely adjusting local loss weights is insufficient to compensate for the inherent vulnerabilities and lack of informative context in disadvantaged nodes.
- **Figure 2(b) - SMOTE Fails:** Conducted on the PubMed dataset (19,717 nodes, 3 classes). We evaluate FedSig, which utilizes SMOTE-like latent oversampling. We record the mean evidence and accuracy of real nodes versus the synthetically interpolated nodes. **Result Analysis:** The comparison shows a drastic drop in both accuracy (from 0.87 to 0.45) and mean evidence (from 9.00 to 6.56) for synthetic samples. This indicates that simple linear interpolation in a highly non-IID topological space produces unreliable, out-of-distribution representations, which introduces negative transfer rather than meaningful augmentation.
- **Figure 2(c) - Evidence Tracks Reliability:** Conducted on the Cora dataset using standard FedAvg with our Evidential GCN. Nodes are grouped into four homophily quantiles: $[0, 0.25)$, $[0.25, 0.50)$, $[0.50, 0.75)$, and $[0.75, 1.00]$. **Result Analysis:** The analysis illustrates a strong positive correlation, where mean evidence monotonically increases alongside actual node accuracy and homophily.

This confirms that the epistemic uncertainty (quantified as evidence) effectively tracks the actual structural reliability of a node, validating its use as a robust diagnostic signal for generating quality control.

## C Details of the Comparison Methods

- **FedAvg** (McMahan et al., 2017): The fundamental federated learning algorithm where the server computes a weighted average of local model updates. It serves as a generic standard baseline, addressing decentralized training but without considering graph topology or fairness constraints.
- **MOON** (Li et al., 2021a): A model-contrastive federated learning framework designed to mitigate the issue of non-IID data. It corrects local training by maximizing the agreement between representations learned by the local model and the global model, though it does not explicitly utilize graph structural information.
- **FedGTA** (Li et al., 2024b): A federated graph learning method that introduces a topology-aware aggregation mechanism. It leverages the local graph structure and node degree distributions to dynamically assign aggregation weights, improving node classification performance under non-IID graph topologies.
- **FedSage+** (Zhang et al., 2021): A spatial-based FGL framework designed to address missing cross-client edges. It trains a missing neighbor generator to synthesize cross-client edges and missing nodes, enriching the local graph structure before applying Graph Neural Networks to mitigate the structural information loss in isolated subgraphs.
- **AdaFGL** (Li et al., 2024a): An adaptive federated graph learning framework that dynamically adjusts the aggregation strategy and local training process. It handles structural heterogeneity by adapting to the varying graph distributions across different clients.
- **FedTAD** (Zhu et al., 2024): A topology-aware knowledge distillation or decoupling framework for FGL. It focuses on extracting and transferring topology-specific knowledge to improve the global model's generalization ability across clients with diverse and heterogeneous graph structures.
- **FairFGL** (Wu et al., 2025): A fairness-aware federated graph learning approach specifically designed to mitigate bias. It incorporates fairness constraints into the federated training process to ensure equitable performance across different node groups or sensitive attributes, addressing the structural bias inherent in local graphs.
- **BoostFGL** (Chen et al., 2026): A fairness-oriented FGL method that employs a boosting-like mechanism. It sequentially re-weights or focuses on hard-to-train and underrepresented nodes across clients, thereby improving both the overall fairness and the performance of the global graph neural network on minority groups.
- **RandomOS**: A standard data-level oversampling baseline applied in the federated setting. It randomly duplicates minority class nodes or underrepresented samples in the local graphs to achieve a balanced training distribution before local model updates.
- **BalancedOS**: A structured oversampling approach that selectively oversamples minority instances based on specific class distributions. It aims to achieve a strictly balanced training set within each client to prevent the local models from collapsing into majority-class predictions.
- **FedSig** (Bi et al., 2024): A state-of-the-art oversampling method tailored for federated learning. It strategically generates synthetic samples or augments features for minority instances across clients to address severe class or structural imbalance, enhancing model robustness and fairness without compromising data privacy.

## D Implementation Details of Method Components

This section provides the full specifications of three components whose high-level descriptions in Sections 4.4 and 4.5 are deliberately condensed: (i) the fairness regularizer $\mathcal{L}_{\text{fair}}$, (ii) the evidential head training objective, and (iii) the conditional EBM training and sampling protocol.

### D.1 EBM Training and Classifier-Guided Sampling

**Architecture.** The conditional energy $E_\psi(\mathbf{h} \mid y, d)$ in Equation (4) is a 3-layer MLP ($\mathrm{LeakyReLU}(0.2)$, hidden width 128) over the concatenation of (a) the first-layer GNN representation $\mathbf{h}$, (b) a learnable class embedding $\mathbf{e}_y \in \mathbb{R}^{32}$, and (c) a learnable binary deficit embedding $\mathbf{e}_d \in \mathbb{R}^{16}$.

**Negative-sample initialization.** The Markov chain that produces $\tilde{\mathbf{h}}$ in Eq. (5) is *warm-started from real data*, rather than from a uniform or Gaussian prior, to avoid wasted exploration in regions far from the data manifold. Concretely, for each training batch we initialize

$$\mathbf{h}_0 = \mathbf{h}_{\text{real}} + \sigma_{\text{init}}\,\boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{14}$$

with $\sigma_{\text{init}} = 0.1$. During the targeted-generation phase (Equation (6)), the seeds are restricted to *deficit-flagged* training nodes of the target class ($d = 1$), which is the explicit structural-fairness intervention.

**Langevin schedule.** We run $T = 20$ steps for CD training and $T = 40$ steps for downstream generation, with a fixed step size $\eta = 10^{-2}$ and Langevin noise scale $\sigma = 5 \times 10^{-3}$. We additionally support a *dataset-adaptive* noise scale $\sigma_{\text{adapt}} = \sigma \cdot \mathrm{clip}\big(1 + \log(n_{pc}/100),\, 0.5,\, 5\big)$, where $n_{pc}$ is the average number of training nodes per class on the client; this avoids over- or under-exploration on graphs of very different scales.

**Classifier guidance.** For generation we follow Eq. (6) with guidance scale $\gamma \in [0, 1]$ multiplying the gradient of the cross-entropy loss against the target class. The classifier branch uses a dedicated linear synthetic-head that maps the conv1 representation directly to logits, so as to remove the train/synthetic distribution mismatch caused by graph-aggregated forward paths. The synthetic head is updated jointly with the GNN through $\mathcal{L}_{\text{syn}}$ in Eq. (11).

**Stability tricks.** Three mechanisms make the EBM training stable: (i) the L2 energy regularizer $\lambda_{\text{reg}}\,\mathbb{E}[E_\psi^2]$ in Equation (5) with $\lambda_{\text{reg}} = 10^{-3}$; (ii) global gradient clipping at norm 1.0 on both EBM and GNN parameters; and (iii) an EBM *warm-up* of $R_{\text{warm}} = 15$ federated rounds, during which only the GNN is optimized so that the conv1 representation has stabilized before the EBM begins to fit it. We further linearly ramp $\lambda_{\text{syn}}$ from 0 to its target value over the next $R_{\text{ramp}} = 10$ rounds; once $r > R_{\text{warm}} + R_{\text{ramp}}$ the schedule reaches the steady-state setting reported in our main experiments.

**Reproducibility.** All hyperparameters above use the default values listed in our released code; the per-dataset variations of $\lambda_{\text{syn}}$, $\lambda_{\text{evi}}$, and $\lambda_{\text{fair}}$ are characterized in Section E.3, where the maximum performance fluctuation across 45 runs is below 0.62% in both Accuracy and Macro-F1.

### D.2 Evidential Head Training Objective

For each real training node $v$, the evidential GNN produces non-negative evidence $\mathbf{e}_v = \mathrm{softplus}(\mathrm{logits}_v)$ and Dirichlet concentrations $\boldsymbol{\alpha}_v = \mathbf{e}_v + \mathbf{1}$ with $S_v = \sum_c \alpha_{v,c}$ as in Equation (8). We adopt the Type-II MLE (digamma) formulation of Sensoy et al. (2018), augmented with an annealed Kullback–Leibler (KL) regularizer that suppresses evidence on incorrect classes:

$$\mathcal{L}_{\text{evi}} = \frac{1}{|\mathcal{V}_{\text{train}}|} \sum_v \Big[ \sum_{c=1}^{C} y_{v,c}\big(\psi(S_v) - \psi(\alpha_{v,c})\big) + \lambda_t\, \mathrm{KL}\big(\mathrm{Dir}(\tilde{\boldsymbol{\alpha}}_v) \,\|\, \mathrm{Dir}(\mathbf{1})\big) \Big], \tag{15}$$

where $y_{v,c}$ is the one-hot label, $\psi(\cdot)$ is the digamma function, $\tilde{\boldsymbol{\alpha}}_v = \mathbf{1} + (1 - \mathbf{y}_v) \odot (\boldsymbol{\alpha}_v - \mathbf{1})$ removes evidence from the target class so that the KL only penalizes *misleading* evidence, and $\lambda_t = \min(1,\, t/T_{\text{anneal}})$ is a linear annealing coefficient with $T_{\text{anneal}} = 25$ rounds.

To reconcile the evidential objective with classification accuracy under sparse client labels, our default setting uses a class-balanced cross-entropy as the primary signal and treats Eq. (15) as an auxiliary calibration term:

$$\mathcal{L}_{\text{real}} = \mathcal{L}_{\text{CE}}^{\text{bal}} + \lambda_{\text{evi}}\, \mathcal{L}_{\text{evi}}, \tag{16}$$

where the class-balanced weights are $w_c = \big(n_{\max}/n_c\big)^p$ with $p = \min(0.8,\, 0.3 + 0.1 \log_2(n_{\max}/n_{\min}))$. The auxiliary weight $\lambda_{\text{evi}}$ corresponds to `evi_aux_weight` in our codebase; its sensitivity is studied in Section E.3. At inference, predictions still use the Dirichlet mean $\alpha_c/S$, so calibrated uncertainty is preserved.

### D.3 Fairness Regularizer $\mathcal{L}_{\text{fair}}$

The parity penalty $\mathcal{L}_{\text{parity}}(\mathcal{V}_{\text{def}}, \mathcal{V}_{\text{nondef}})$ in Eq. (12) is implemented as a *soft, one-sided loss-disparity penalty* that explicitly discourages the model from optimizing non-deficit nodes at the expense of deficit ones. Let $\ell_v$ denote the per-node training loss of node $v \in \mathcal{V}_{\text{train}}$ (in our default setup, the digamma evidential loss; see Section D.2), and let $w(v) \in [0, 1]$ be the continuous deficit score from Equation (3). We define group-mean losses with $w(v)$ as a soft membership of $\mathcal{V}_{\text{def}}$ and $1 - w(v)$ as a soft membership of $\mathcal{V}_{\text{nondef}}$:

$$\bar{\ell}_{\text{def}} = \frac{\sum_v w(v)\,\ell_v}{\sum_v w(v) + \varepsilon}, \qquad \bar{\ell}_{\text{nondef}} = \frac{\sum_v (1 - w(v))\,\ell_v}{\sum_v (1 - w(v)) + \varepsilon}. \tag{17}$$

The regularizer then penalizes only the case where deficit nodes incur larger loss than non-deficit ones, with the non-deficit term *detached* from the gradient graph so that fairness is improved by *lifting up* the deficit group rather than *harming* the non-deficit group:

$$\mathcal{L}_{\text{fair}} = \left[\bar{\ell}_{\text{def}} - \text{sg}(\bar{\ell}_{\text{nondef}})\right]_+, \tag{18}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator and $[x]_+ = \max(0, x)$. The penalty is gated to zero whenever either $\sum w(v)$ or $\sum(1 - w(v))$ falls below $10^{-6}$, which prevents spurious gradients in trivially-skewed batches. In Eq. (12) we combine $\mathcal{L}_{\text{fair}}$ with a fixed coefficient $\lambda_{\text{fair}}$, whose sensitivity is reported in Section E.3.

## E Additional Ablation Studies

This appendix complements the component ablation in Figure 4(a-b) of the main paper with three further analyses: (i) isolating the contribution of the evidence gate from that of the EBM generator, (ii) comparing the EBM against alternative generator architectures under the same gate, and (iii) characterizing the sensitivity of $\mathbf{E}^2$Gen to its three loss-weighting hyperparameters. To keep each table self-contained, we report two representative homophilic benchmarks per ablation; corresponding numbers on the remaining datasets are consistent in direction and are released with our code.

### E.1 Contribution of the Gate and the Generator

We isolate the two pillars of $\mathbf{E}^2$Gen by enabling them in turn:

- **Only Gate**: no synthetic generation; the evidential head re-weights real training nodes by per-node uncertainty $w = 1 + \frac{1}{2}(K/S)$, equivalent to a simple uncertainty-aware filter *without* any generator.
- **Only Generator**: synthetic samples from the conditional EBM are mixed in unfiltered ($w = 1$ for all synthetic nodes), equivalent to a generator without any quality gate.
- **Both (ours)**: the synthetic samples produced by the EBM are passed through the evidence gate of Equation (9) and used at the gated weight.

All other settings are held fixed; we use the per-dataset optimal HP returned by Optuna for the joint setting and disable the fairness regularizer ($\lambda_{\text{fair}} = 0$) to remove a potential confound.

Table 3: Contribution of the gate and the generator. Either component alone leaves substantial performance on the table; combining them is monotonically best on both datasets and on both metrics.

| | PubMed | | Coauthor-CS | |
|---|---|---|---|---|
| Variant | Acc | MF1 | Acc | MF1 |
| Only Gate (uncertainty filter only) | 86.04 | 85.47 | 86.72 | 73.17 |
| Only Generator (unfiltered samples) | 86.20 | 85.73 | 90.29 | 87.06 |
| **Gate + Generator (ours)** | **86.68** | **86.11** | **90.42** | **87.11** |
| $\Delta$ over best single component | +0.48 | +0.38 | +0.13 | +0.05 |

### E.2 Comparison of Generator Architectures

We next ask whether the choice of an energy-based generator specifically is justified, by holding the evidence gate of Equation (9) fixed and swapping only the synthesizer. Three non-energy alternatives are considered:

- **Mixup** (Zhang et al., 2017): linear interpolation of representations from the same client, with the deficit-weighted sampling of Equation (3) to retain the targeting bias.
- **CVAE** (Sohn et al., 2015): per-client conditional variational autoencoder over the conv1 representation, sampling minor classes proportionally to deficit.
- **GAN** (Mirza and Osindero, 2014): per-client conditional GAN over the same representation space.

Architectures of the CVAE and GAN match the EBM in width (hidden $= 256$) and per-round training budget (Section D.1). Numbers in Table 4 are reported under the same HP as Table 3.

Table 4: Comparison of generator architectures under the same evidence gate. The EBM is the strongest generator on Accuracy and Macro-F1 across both datasets; non-energy alternatives are competitive but never strictly best on either metric.

| | PubMed | | CiteSeer | |
|---|---|---|---|---|
| Generator (under same gate) | Acc | MF1 | Acc | MF1 |
| Mixup | 86.37 | 85.85 | 73.76 | 68.22 |
| CVAE | 86.58 | 86.04 | 73.69 | 67.83 |
| GAN | 85.81 | 85.31 | 73.54 | 66.86 |
| **EBM (ours)** | **86.68** | **86.11** | **73.84** | **68.59** |

**Discussion.** The EBM is the top generator on both metrics across both datasets. On PubMed the lead is most pronounced, with the adversarial GAN trailing by $0.87\%$ Acc and $0.80\%$ MF1. On CiteSeer the four generators differ by less than $0.30\%$ on Accuracy — the gate is already absorbing most of the per-sample quality variance — yet the EBM still extracts an extra $0.37\%$ Macro-F1 over the strongest alternative, reflecting better recall on the minority classes that the gate alone cannot recover. Two failure modes of the non-energy alternatives also surfaced during these runs: the GAN exhibited adversarial training instability with frequent discriminator mode collapse on small-graph clients, and the CVAE displayed posterior collapse on under-represented classes, constraining the diversity of accepted samples.

### E.3 Hyperparameter Sensitivity

We evaluate the robustness of $\mathbf{E}^2$Gen with respect to its three core loss-weighting hyperparameters: the synthetic data loss weight $\lambda_{\text{syn}}$, the evidential auxiliary loss weight $\lambda_{\text{evi}}$, and the fairness loss weight $\lambda_{\text{fair}}$. Hyperparameter tuning exerts minimal impact on these datasets; the Optuna search ranges yield accuracy fluctuations of merely $2.07\%$, $1.07\%$, and $0.96\%$ for CiteSeer, Coauthor-CS, and Coauthor-Physics, respectively.

For each hyperparameter, we fix all other hyperparameters at their per-dataset optimal values and vary only the target hyperparameter across five representative values spanning its feasible range: $\lambda_{\text{syn}} \in \{0.01, 0.1, 0.5, 1.0, 2.0\}$, $\lambda_{\text{evi}} \in \{0.0, 0.05, 0.1, 0.2, 0.3\}$, and $\lambda_{\text{fair}} \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$. We report both Accuracy ($\mathcal{A}$) and Macro-F1 ($\mathcal{F}$).

**Results.** Figures 5 to 7 present the results, with each figure showing Accuracy ($\mathcal{A}$, left) and Macro-F1 ($\mathcal{F}$, right) side by side. $\mathbf{E}^2$Gen exhibits strong robustness to all three hyperparameters on these datasets:

- **Synthetic data weight** $\lambda_{\text{syn}}$ (Figure 5): Varying $\lambda_{\text{syn}}$ across two orders of magnitude ($0.01 \rightarrow 2.0$) causes at most $0.30\%$ Accuracy and $0.46\%$ Macro-F1 fluctuation. Notably, CiteSeer remains *completely invariant* ($73.39\%$ Accuracy and $67.97\%$ Macro-F1 across all five values), indicating that the evidence gate effectively filters synthetic data regardless of its loss weight.

- **Evidential auxiliary weight** $\lambda_{\text{evi}}$ (Figure 6): Performance is stable for $\lambda_{\text{evi}} \in [0.0, 0.2]$, with Accuracy varying by at most $0.19\%$ on Physics and $0.07\%$ on CiteSeer. A slight decline appears only at the extreme $\lambda_{\text{evi}} = 0.3$ on Coauthor-CS ($-0.37\%$ Accuracy, $-0.62\%$ Macro-F1) and Physics ($-0.19\%$ Accuracy, $-0.32\%$ Macro-F1), suggesting that overly strong auxiliary supervision can marginally interfere with the primary classification objective.
- **Fairness loss weight** $\lambda_{\text{fair}}$ (Figure 7): Across all three datasets, Accuracy fluctuates by at most $0.37\%$ and Macro-F1 by $0.40\%$. This demonstrates that the fairness regularizer is well calibrated and does not trade off predictive performance even when $\lambda_{\text{fair}}$ varies by two orders of magnitude ($0.001 \rightarrow 0.1$).

Overall, the maximum performance fluctuation across all 45 experiments (3 datasets $\times$ 3 hyperparameters $\times$ 5 values) is **less than** $0.62\%$ in both Accuracy and Macro-F1. This confirms that $\mathbf{E}^2$Gen is not sensitive to the choice of loss-weighting hyperparameters on homophilic and medium-scale graph datasets, and reasonable default values can achieve near-optimal performance without extensive tuning.

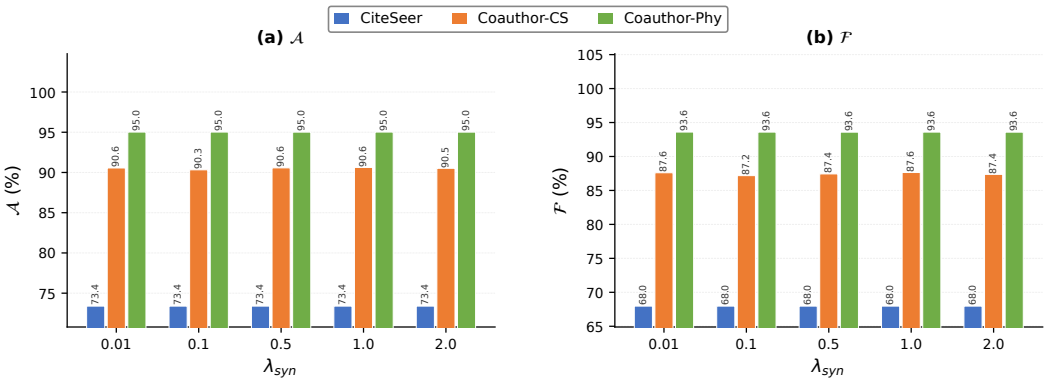


Figure 5: Sensitivity to the synthetic data loss weight $\lambda_{\text{syn}} \in \{0.01, 0.1, 0.5, 1.0, 2.0\}$. Left: Accuracy ($\mathcal{A}$); Right: Macro-F1 ($\mathcal{F}$). CiteSeer is completely invariant (73.39%/67.97%), and the maximum fluctuation across all five values and three datasets is below 0.46%.

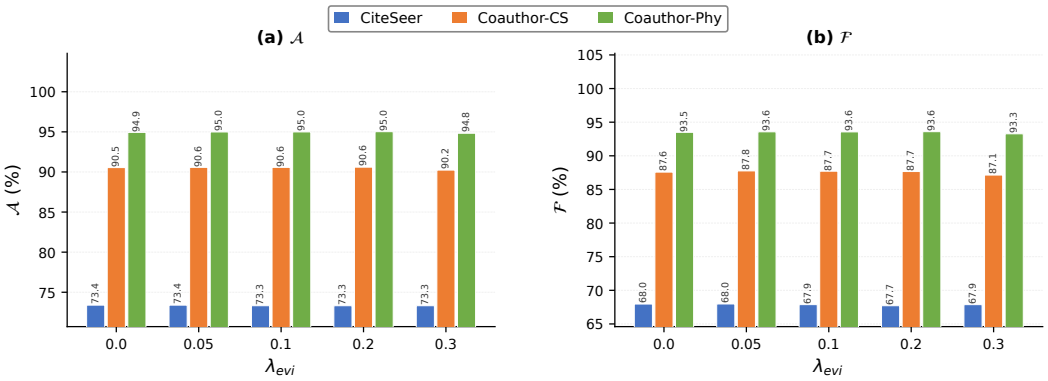


Figure 6: Sensitivity to the evidential auxiliary loss weight $\lambda_{\text{evi}} \in \{0.0, 0.05, 0.1, 0.2, 0.3\}$. Left: Accuracy ($\mathcal{A}$); Right: Macro-F1 ($\mathcal{F}$). Performance is stable for $\lambda_{\text{evi}} \leq 0.2$, with a mild decline only at the extreme value 0.3 on Coauthor-CS and Coauthor-Physics.

## F Gate Behavior: Acceptance, Evidence, and Label Consistency

To complement the description of the evidential gate in Section 4.4, we present direct, per-sample diagnostics of its runtime behavior. All statistics in this section are pooled across clients and the last 30 federated rounds of a single seed (*seed* $= 42$), after the gate becomes active. We report **(i)** the

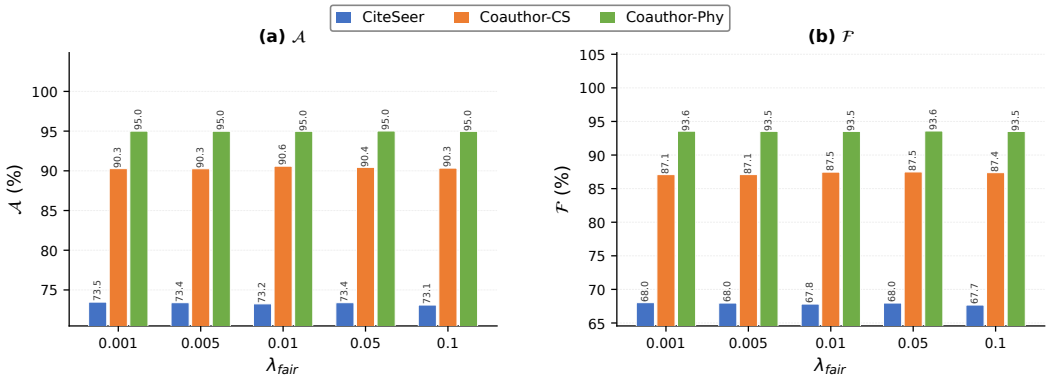


Figure 7: Sensitivity to the fairness loss weight $\lambda_{\text{fair}} \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$. Left: Accuracy ($\mathcal{A}$); Right: Macro-F1 ($\mathcal{F}$). Performance remains stable across two orders of magnitude, with fluctuations below 0.40% on every dataset.

empirical distribution of soft acceptance weights $w$ produced by the gate on seven datasets spanning the homophily spectrum, **(ii)** the distribution of the underlying evidence score $\hat{e}$ (real training nodes vs. accepted vs. floored synthetic samples), and **(iii)** a stress test of whether high evidence can coexist with disagreement with the conditioning label (Section F.1).

**Acceptance statistics across datasets (Figure 8).** We bin every synthetic sample into one of three groups according to its gate weight $w$: *High* ($w > 0.4$), *Medium* ($0.15 < w \leq 0.4$), and *Floor* ($w \leq 0.15$, the soft minimum-weight setting). The acceptance profile varies substantially across datasets, and correlates with overall data complexity rather than a single homophily threshold. On the cleanest homophilic graph (Coauthor-Phy, $\bar{\phi} = 0.93$), 54% of synthetic samples reach the high-weight regime and the mean weight is $\bar{w} = 0.42$. On the most challenging heterophilic graph (Squirrel, $\bar{\phi} = 0.22$), 89% of samples are squashed to the weight floor and $\bar{w} = 0.13$. Datasets in between (PubMed, CiteSeer, CS, Cora, Chameleon) exhibit intermediate, monotonically related profiles. These numbers reflect that the gate exerts essentially no effect on confident homophilic samples and applies near-uniform downweighting on hard heterophilic samples.

**Evidence profiles (Figure 9).** The continuous signal driving the gate is the per-sample evidence score $\hat{e} = p_{\max} \cdot (1 - K/S)$, which combines the predicted-class probability with the Dirichlet certainty ($K$ is the number of classes, $S$ the total evidence). For each of four representative datasets we overlay the empirical density of $\hat{e}$ for (a) real training nodes, (b) high-weight synthetic samples, and (c) floored synthetic samples, together with the gate threshold $\tau = 0.35$. On PubMed (homophilic, easy generation), the three densities are clearly separated: floored samples sit below $\tau$, the real-data mode straddles $\tau$, and the high-weight samples concentrate above $\tau$. On Cora (homophilic but small) the gate operates over a wider range, and on the heterophilic Chameleon and Squirrel the synthetic-sample mode lies entirely below $\tau$, leaving no high-weight green mass. Importantly, in every panel the synthetic-sample density occupies the same support as the real-training density, indicating that the gate is not an out-of-distribution rejector but an interpretable continuous re-weighter operating on a calibrated signal.

Together, Figures 8 and 9 provide direct empirical evidence that the evidential gate is doing work: its filter intensity is dataset-dependent, calibrated to the underlying evidence distribution, and observable per sample.

### F.1 Evidence–Consistency Decoupling: A Gate Stress Test

**Protocol and interpretation.** We analyze the per-sample gate diagnostics underlying Figures 8 and 9, comprising approximately 657,000 synthetic samples across seven datasets under the pooling protocol described above. For each synthetic sample, we compare the classifier's predicted label $\hat{y}_i$ with its conditioning class $\tilde{y}_i$. Since the same classifier guides sampling and supplies the label-agreement factor in Equation (9), this is a *model-consistency proxy*, not semantic ground truth. The

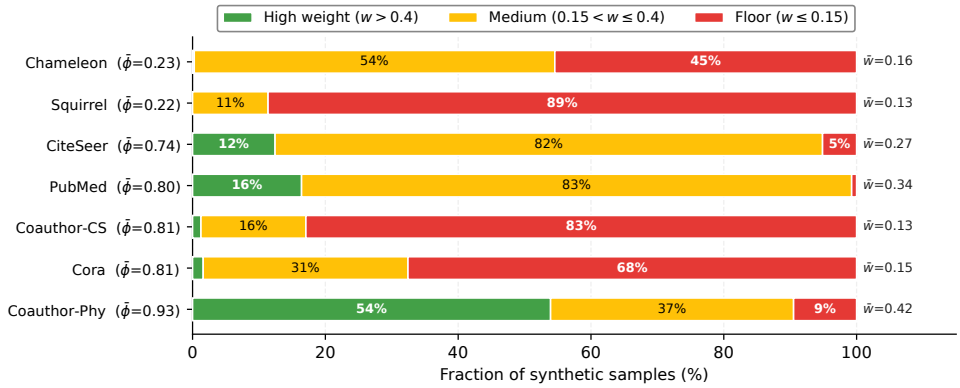


Figure 8: Per-dataset distribution of synthetic-sample gate weights $w$ (pooled across clients and the last 30 rounds, *seed* $= 42$). *High*: $w > 0.4$; *Medium*: $0.15 < w \leq 0.4$; *Floor*: $w \leq 0.15$. The right-edge label reports the mean weight $\bar{w}$. The gate's filter intensity varies smoothly with dataset complexity: from 9% floor on Coauthor-Phy to 89% floor on Squirrel. A detailed analysis can be found in section F.

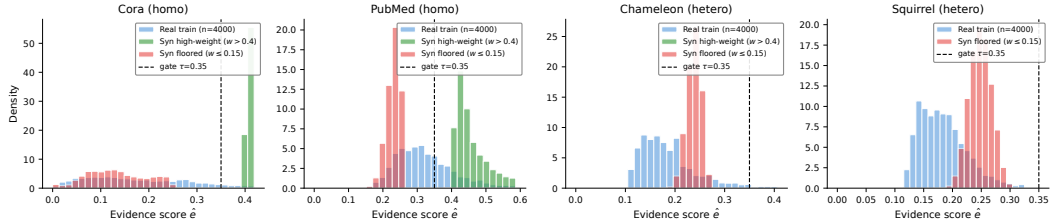


Figure 9: Empirical density of the gate's underlying evidence score $\hat{e} = p_{\max} \cdot (1 - K/S)$ for four representative datasets. Blue: real training nodes; green: high-weight synthetic samples ($w > 0.4$); red: floored synthetic samples ($w \leq 0.15$). The dashed vertical line marks the gate threshold $\tau = 0.35$. On PubMed the three modes are cleanly separated; on Squirrel and Chameleon the synthetic mode is concentrated below $\tau$, which explains the high floor rates in Figure 8. A detailed analysis can be found in section F.

downstream utility of the augmentation is evaluated separately by held-out test performance in Table 1; it is not established by this agreement diagnostic.

Within each dataset, we define high evidence as the top quartile of the logged evidence scores. We report the joint fraction of samples that have high evidence and disagree with the conditioning label, using *all synthetic samples in that dataset* as the denominator. This differs from the disagreement rate conditional on high evidence. We also compare the mean gate weights of agreeing and disagreeing samples within the high-evidence group.

Table 5: Gate stress test. The joint fraction is the share of all synthetic samples in each dataset that fall in the top evidence quartile and disagree with the conditioning class. Mean gate weights are computed within the high-evidence group; the reduction is relative to the agreeing group. The other five datasets are Cora, CiteSeer, PubMed, Chameleon, and Squirrel, each with no observed label disagreement. Dashes indicate that no weight comparison is available.

| | High evidence ∩ disagreement | Mean gate weight (high evidence) | | |
|---|---|---|---|---|
| Dataset | Joint fraction | Agreeing | Disagreeing | Reduction |
| Coauthor-CS | 2.0% | 0.210 | 0.113 | 46% |
| Coauthor-Physics | 0.5% | 0.673 | 0.147 | 78% |
| Other five datasets (each) | 0.0% | — | — | — |

**Observed decoupling.** Table 5 shows that high evidence and label disagreement can coexist, but the joint event accounts for at most $2.0\%$ of all synthetic samples on any dataset in these diagnostics. On five of the seven datasets, label agreement is $100\%$ across the logged synthetic samples. Such saturation is unsurprising given the shared classifier and limits what this proxy can reveal about sample quality. Where agreement varies, evidence and consistency are positively associated: on Coauthor-Physics, agreement increases monotonically from $78.0\%$ in the lowest evidence decile to $99.5\%$ in the highest, with correlation $r = +0.29$ between evidence and the binary agreement indicator.

**Response of the two-factor gate.** The gate in Equation (9) multiplies the probability assigned to the conditioning class by an evidence sigmoid. Because the sigmoid is bounded by one, high evidence alone cannot override a low conditioning-class probability. Consistent with this mechanism, high-evidence samples with disagreeing labels receive mean gate weights of $0.113$ on Coauthor-CS and $0.147$ on Coauthor-Physics, compared with $0.210$ and $0.673$ for their agreeing counterparts: reductions of approximately $46\%$ and $78\%$, respectively. These results support the gate's ability to downweight the observed evidence–consistency mismatch. They do not establish semantic correctness for samples on which the shared classifier agrees with the conditioning label.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction provide an accurate summary of the paper's scope and contributions, aligning well with the main content.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: While this paper does not delve into formal theoretical proofs, we hypothesize that structural fairness disparities in federated graph learning stem from inherent topological vulnerabilities and missing local contexts, which can be mitigated via targeted energy-based generation and evidential gating. We provide extensive empirical evidence to support this hypothesis, and the experimental results further corroborate its validity through consistent improvements across multiple benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper specifies the dataset splits, the number of communication rounds, local training epochs, and the use of Optuna for hyperparameter search. Detailed settings for EBM training, Langevin schedules, and other components are in section D.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our code and dataset are openly released with clear and detailed instructions, ensuring that others can readily reproduce the experimental findings reported in this paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the dataset splits, the number of communication rounds, local training epochs, and the use of Optuna for hyperparameter search in section 5.1. Detailed settings for EBM training, Langevin schedules, and other components are in section D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Given the consistently strong performance improvements observed across our experiments, we consider the empirical evidence sufficient to support our conclusions. As such, traditional significance testing and detailed uncertainty quantification were deemphasized in favor of presenting the overall magnitude of gains.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient details about the computational resources required to reproduce our experiments in Section 5.1

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in our paper conforms with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: The research presented in this paper fully complies with the NeurIPS Code of Ethics in all respects. We have carefully considered the ethical implications throughout the research process, including data usage, model behavior, and potential societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All external assets used in this paper, including code, data, and models, are appropriately cited and used in accordance with their respective licenses and terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have provided the code via an GitHub link.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve any crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve any research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The research does not involve the use of large language models in the design or implementation of the core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.