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

Anonymous Author(s)
Affiliation
Address
email

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. The code is anonymously available at https://anonymous.4open.science/r/E-Gen-A2C6.

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 (Kairouz and McMahan, 2021). This paradigm naturally extends to graph-structured data, forming the basis of Federated Graph Learning (FGL) (Zhang et al., 2021a). 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 (Fallah et al., 2020; Li et al., 2021b). However, these methods often struggle to perform effectively in FGL environments due to the unique properties of graph data. Unlike traditional machine learning that assumes independent data instances, graph learning fundamentally relies on

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

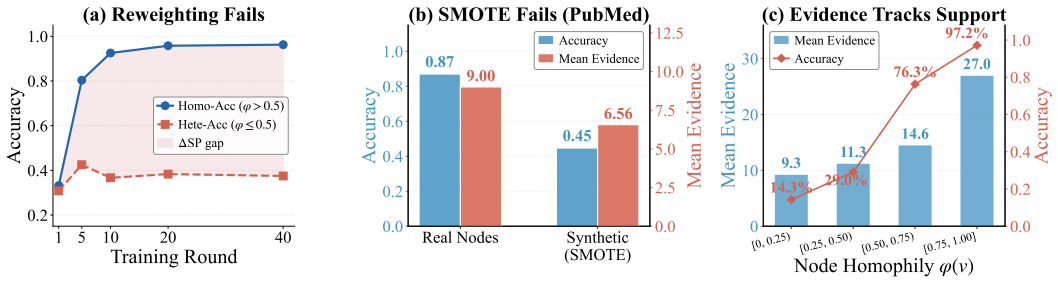

Figure 2: Empirical observations on the inherent vulnerabilities of disadvantaged nodes in FGL. **(a) Reweighting Fails:** Using the class-level reweighting method BoostFGL, the accuracy gap between advantaged (homophilous, $\varphi > 0.5$) and disadvantaged ($\varphi \leq 0.5$) nodes still persists 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 (φ) and accuracy, confirming it as a valid metric for support deficiency. For detailed experimental setups and parameter configurations, please refer to Appendix B.

message passing mechanisms (Gilmer et al., 2017), 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?*

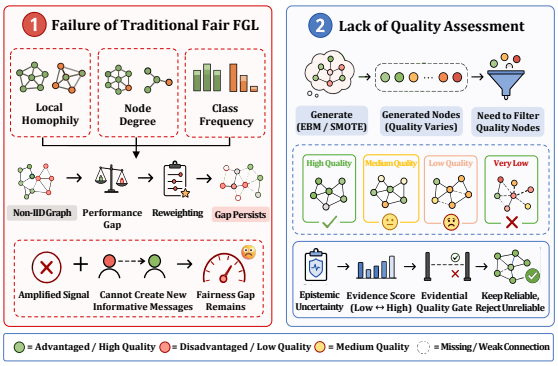

Figure 1: **Problem Illustration.** We describe the challenges in fair FGL: **I)** Traditional reweighting methods merely amplify existing signals without creating new informative messages, failing to close the fairness gap. **II)** Generation-based methods lack quality assessment, necessitating an evidential 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; Song and Kingma, 2021), 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; Hendrycks and Gimpel, 2016), 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, we introduce **Evidential Quality Gating** (Sensoy et al., 2018). Motivated by the theoretical framework of subjective logic (Jsang, 2018), 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 down-weighting 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. 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.

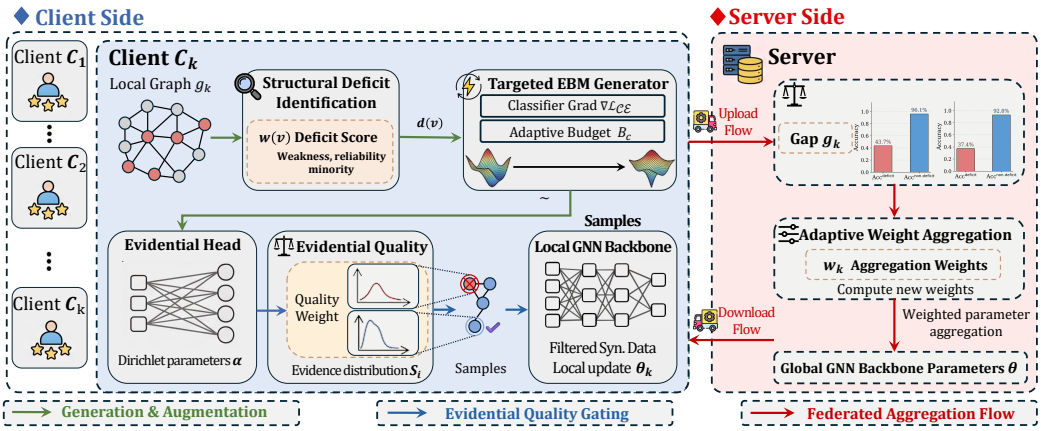

Figure 3: Architecture illustration of E^{2}Gen. We use green, blue, and red arrows to represent the three main flows of our method: Generation & Augmentation, Evidential Quality Gating, and Federated Aggregation Flow, respectively. Best viewed in color. Zoom in for details.

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$_{\uparrow 0.97}$
E²Gen (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$_{\uparrow 0.38}$	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$_{\downarrow 1.43}$
E²Gen (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**$_{\uparrow 3.19}$

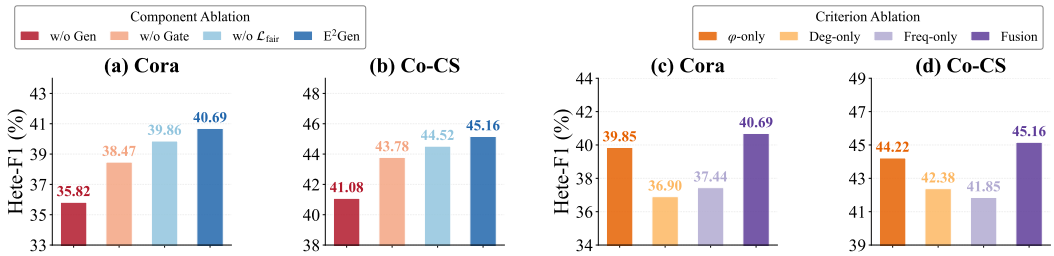

Figure 4: **Ablation Analysis** of components and criteria for $\mathbf{E}^2$Gen on the Cora and Co-CS datasets. The results illustrate the fairness performance (Hete-F1) under Component Ablation (a-b) and Criterion Ablation (c-d).