FedIGL: Federated Invariant Graph Learning for Non-IID Graphs

Lingren Wang[1] **Wenxuan Tu**[2*] **Jiaxin Wang**[3] **Xiong Wang**[2]
Jieren Cheng[2*] **Jingxin Liu**[3]
[1]School of Information and Communication Engineering, Hainan University
[2]School of Computer Science and Technology, Hainan University
[3]School of Cyberspace Security, Hainan University
{twx,992730}@hainanu.edu.cn

Abstract

Federated Graph Learning (FGL) effectively facilitates cross-domain graph model training by enabling decentralized learning across multiple domains, while ensuring data privacy through local data storage and communication of model updates instead of raw data. Existing approaches usually assume shared generic knowledge (e.g., prototypes, spectral features) via aggregating local structures statistically to alleviate structural heterogeneity. However, imposing overly strict assumptions about the presumed correlation between structural features and the global objective often fails in generalizing to local tasks, leading to suboptimal performance. To tackle this issue, we propose a **Fed**erated **I**nvariant **G**raph **L**earning (**FedIGL**) framework based on invariant learning, which effectively disrupts spurious correlations and further mines the invariant factors across different distributions. Specifically, a server-side global model is trained to capture client-agnostic subgraph patterns shared across clients, whereas client-side models specialize in client-specific subgraph patterns. Subsequently, without compromising privacy, we propose a novel Bi-Gradient Regularization strategy that introduces gradient constraints to guide the model in identifying client-agnostic and client-specific subgraph patterns for better graph representations. Extensive experiments on graph-level clustering and classification tasks demonstrate the superiority of FedIGL against its competitors.

1 Introduction

Graph Neural Networks (GNNs) research [50, 33, 9, 11, 38, 36, 8] is rapidly growing due to the ability of GNNs to learn representations from graph-structured data. In practice, centralizing large amounts of real-world graph data for training is prohibitive due to privacy concerns and regulatory restrictions [26, 47, 35]. Federated Graph Learning (FGL), a growing distributed learning paradigm, offers a potential solution to this challenge while preserving data privacy [24, 23]. Nonetheless, the non-IID problem remains a major challenge in FGL, as graph data from different distributions usually vary significantly [40].

Existing approaches typically rely on the assumption that generic knowledge learned from training on non-IID clients can be effectively reconstructed across clients to enable collaborative training [53]. The shared knowledge includes consensus prototypes [22, 48], generic spectral knowledge [32], and structure encoder parameters [31], which are introduced as shared knowledge representations to facilitate the execution of graph-level learning tasks. Despite the enormous success, existing methods overly rely on statistical correlation, misleadingly assuming that the robust representations learned from prior knowledge are widely applicable. The correlation between generic knowledge

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

and the target is not necessarily task-related, and such spurious correlations embedded in the learned representations often fail to generalize in real-world scenarios. Furthermore, these approaches typically upload the learned knowledge or prototypes to the server, which may lead to potential data leakage. Therefore, it is critical to promote inter-client negotiation in the framework of federated graph learning without compromising data privacy.

An intuitive solution is to exploit factors that remain consistently stable and effective across clients to mitigate the impact of spurious correlations that merely reflect statistical commonality. In other words, the global model should be capable of identifying invariant factors across clients. Empirical observations indicate that graphs with different distributions often share common subgraph patterns [54], even when their distributions differ significantly, as shown in Fig. 1. This observation inspires the following idea: if we can identify subgraph patterns that are shared across different distributions, these common patterns can serve as a foundation for inter-client collaboration and improve the generalization of the global model. In contrast, distribution-specific patterns should be retained locally on each client to prevent them from negatively impacting the global representation. This insight prompts us to consider two fundamental questions: (1) How to discover invariant subgraph patterns across different distributions in FGL? (2) How can one extract invariant subgraph patterns in a privacy-preserving manner, considering that FL prohibits data sharing across clients? To the best of our knowledge, both questions remain largely unexplored.

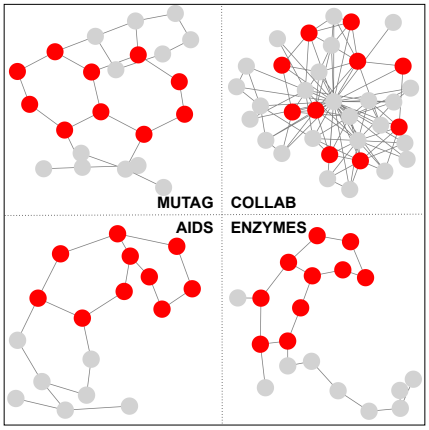


Figure 1: Illustration of shared subgraphs in different distributions in TUDatasets [30], where the red nodes highlight the common structure.

To tackle these challenges, we propose a **Fed**erated **I**nvariant **G**raph **L**earning (FedIGL) , which aims to identify invariant subgraph patterns shared across clients while preserving local privacy. To address the first question, inspired by invariant graph learning [20, 47], which focuses on improving generalization to out-of-distribution graphs, we design a Federated Subgraph Generator (FSG) to extract client-agnostic and client-specific subgraphs. The generated client-agnostic and client-specific subgraphs are used to promote negotiation among clients and maintain the heterogeneity inherent to each client, respectively. To address the second question, we introduce a novel Bi-Gradient Regularization strategy that imposes consistency and diversity constraints on gradients. It is effectively guides the generator to learn disentangled subgraph patterns while ensuring data privacy is not compromised. After obtaining the client-specific subgraphs, we design a local model for each client that is excluded from collaborative training and is trained solely on these subgraphs. FedIGL is encouraged to discover invariant subgraph patterns across data distributions, thereby mitigating client drift caused by graph heterogeneity.

- To the best of our knowledge, this is the first work leveraging invariant learning in federated graph learning to enhance generalization under non-IID client settings.
- We propose the Bi-Gradient Regularization strategy, which can coordinate the clients to learn disentangled subgraph patterns without compromising data privacy.
- We conduct extensive experiments to verify both our theoretical results and the superiority of FedIGL, which consistently outperforms existing approaches on graph-level classification and clustering tasks.

2 Related Work

Federated Graph Learning. FGL enables distributed training of GNNs across multiple parties, facilitating collaborative learning on graph-structured data without compromising data privacy [43, 41, 10, 14, 46, 49, 19, 18, 37, 12]. Due to significant differences in client distribution and graph structure across domains, low inter-graph similarity hinders unified processing [44, 7, 13, 4, 5]. Existing methods mitigate structural heterogeneity by leveraging shared, pre-trained representations from cross-domain client models [42]. Examples include prototype-based structures [40, 53], spectral

feature alignment [32], and shared structural encoder parameters [31]. FedSSP [32] shares generalized spectral knowledge with a personalized module to adapt to client-specific graph structures, while FedGCN [22] leverages multi-source clustering to generate global consensus representations, enhancing its ability to handle complex graph structures. Despite their success, most methods rely heavily on assumed shared knowledge, which limits adaptability to diverse distributions. Since this knowledge is learned via pre-trained shared parameters unrelated to task causality, distribution shifts can degrade representation quality and harm model performance [45, 39, 17, 11].

Invariant Graph Learning (IGL). Invariant Learning is a class of learning methods focused on distribution generalization or robust modeling [2, 1]. Its main idea is to learn representations or predictive functions that remain stable and effective across different environments or data distributions. As previously discussed, although graphs from different distributions are heterogeneous, they share certain common subgraph patterns. Building upon these findings, IGL has emerged as a prominent research direction in recent years [20, 47, 29]. The main idea of IGL is designing a subgraph generator to partition a graph into two components: the invariant subgraph, which captures structures that are consistent across different distributions, and the environment-specific subgraph, which represents structures that are present only in particular distributions. In this paper, we extend invariant graph learning to federated learning, where each client with a distinct distribution is treated as a separate environment.

3 Preliminaries

Federated Learning (FL). Given a local dataset $(x, y) \sim P_k$ for each client k, where P_k denotes a client-specific data distribution, the goal of standard FL approaches [27, 21] is to learn a global model that minimizes the empirical risk across all client distributions, defined as:

$$\theta^* = \arg\min_{\theta} \sum_{k=1}^{K} \mathbb{E}_{(x,y)\sim P_k}\, \ell(f_\theta(x), y), \tag{1}$$

where $\ell(\cdot)$ is a task-specific loss function and $f_\theta(\cdot)$ is the global model parameterized by θ. In practice, FL methods typically decompose this global objective into a weighted sum of local empirical losses and perform independent optimization on each client. However, independently optimizing local objectives often leads to suboptimal convergence [6, 41], particularly when client distributions exhibit significant heterogeneity. This is because, under non-IID conditions, the local updates from different clients may follow divergent optimization directions [44, 25, 28, 16, 3].

4 Methodology

In this section, we introduce our proposed method in detail, whose framework is shown in Fig. 2. First, we define our optimization objective. Then, we present the FSG identifying client-invariant subgraph patterns. Finally, we propose the Bi-Gradient Regularization strategy for objective optimization and provide its theoretical analysis. Algorithmic details can be found in Appendix A.

4.1 Problem Formulation

In the federated optimization objective in Eq. (1), when client data are not identically distributed, each client's optimization direction tends to push the global model along different update trajectories [40]. This leads to gradient conflicts, hindering convergence to a globally optimal solution. To mitigate this, our goal is to guide the global model to focus on features that are shared across all client distributions, while excluding distribution-specific features from global optimization [27, 6]. This strategy helps prevent conflicting updates and promotes more stable convergence. We therefore aim to incorporate client-agnostic subgraphs, namely invariant subgraphs, into global model training to better accommodate distributional shifts [20]. In contrast, client-specific subgraphs, which capture environment-specific patterns, are processed by dedicated local models that remain discrepancies to each client. In this paper, we decouple the model into two distinct components: a global model $f_g(\cdot)$, which participates in federated aggregation, and a local model $f_c(\cdot)$, which remains client-specific.

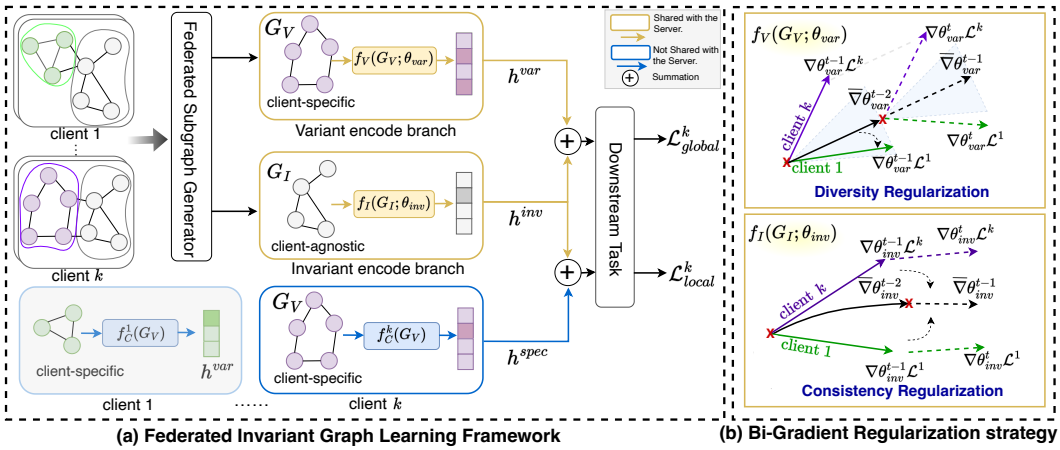


Figure 2: Architecture illustration of FedIGL. The left box (a) represents the process by which clients obtain client-agnostic and client-specific subgraphs through the federated subgraph generator (FSG). The yellow and blue boxes denote parameters that are globally shared and not shared, respectively. The right box (b) shows that diversity regularization penalizes overly similar gradients in the variant encoder. Consistency regularization encourages stability and agreement in the optimization trajectory of the invariant encoder across rounds.

Based on this formulation, the optimization objective of FedIGL is defined as follows:

$$\min_{\theta_g,\{\theta_c^k\}_{k=1}^K} \sum_{k=1}^{K} \mathbb{E}_{(x,y)\sim P_k} \left[\ell\left(h\left(f_g(x;\theta_g),\ f_c(x;\theta_c^k)\right),\ y\right)\right], \tag{2}$$

where θ_c^k denotes the local model parameters of client k, $\ell(\cdot,\cdot)$ is the loss function, and $h(\cdot,\cdot)$ is a fusion function, such as concatenation or addition, used to integrate the outputs of the global and local models for downstream tasks. We decompose the overall optimization objective into two stages: global model optimization and local model optimization, which will be elaborated upon in the following sections.

4.2 Discovering Invariant and Variant Subgraphs

For the global model, we first employ a FSG to decompose each graph into client-agnostic and client-specific subgraphs. These subgraphs are then encoded separately, and their resulting representations are combined and fed into the downstream task for training.

Similar to prior work [15, 54], we implement the FSG using a graph neural network (GNN). Given graph G with n nodes and its adjacency matrix $\mathbf{A} = \{0,1\}^{n\times n}$, where $\mathbf{A}_{i,j} = 1$ represents that there exists an edge between node i and j, and $\mathbf{A}_{i,j} = 0$ otherwise. The FSG first generates a mask matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ for $\mathbf{A}$:

$$\mathbf{M}_{ij} = \mathrm{MLP}\big(\mathrm{CONCAT}(\mathbf{Z}_i, \mathbf{Z}_j)\big), \quad \mathbf{Z} = \mathrm{GNN}(G), \tag{3}$$

where $\mathrm{MLP}(\cdot)$ is a multilayer perceptron and $\mathrm{CONCAT}(\cdot,\cdot)$ is the concatenation operation, $\mathbf{Z}_i$ denotes the representation of the i-th node in graph G. We use MLP instead of the inner product to generate the matrix $\mathbf{M}$, since in federated learning, the inner product may leak structural information by enabling edge reconstruction [22]. Then, we can obtain the adjacency $\mathbf{A}_I$ and $\mathbf{A}_V$ corresponding to the client-agnostic subgraph G_I and the client-specific subgraph G_V:

$$\mathbf{A}_I = Top_t(\mathbf{M} \odot \mathbf{A}), \quad \mathbf{A}_V = \mathbf{A} - \mathbf{A}_I, \tag{4}$$

where $Top_t(\cdot)$ select the elements in $\mathbf{M}$ whose size is the top percent t, and t is a hyperparameter. After obtaining G_I and G_V, we adopt two branches, invariant and variant encode branch, to encode the obtained client-agnostic and client-specific subgraphs representation, respectively:

$$\mathbf{h}^{inv} = \mathcal{R}\big(f_I(G_I;\theta_{inv})\big), \quad \mathbf{h}^{var} = \mathcal{R}\big(f_V(G_V;\theta_{var})\big), \tag{5}$$

where $\mathcal{R}(\cdot)$ is used to obtain the graph-level representation, and $f(\cdot)$ is the graph encoder. Then, we use the sum of $\mathbf{h}^{inv}$and $\mathbf{h}^{var}$ as the representation of graph G to train the global model. The global model loss in client k is then defined as:

$$\mathcal{L}^k_{global}(\theta_{FSG}, \theta_{inv}, \theta_{var}) = \ell(\mathbf{h}^{inv} + \mathbf{h}^{var}, y), \tag{6}$$

where θ_{FSG} is the parameters of the FSG.

For the client model, we define a private encoder $f_C(\cdot)$ for each client to encode the client-specific subgraph. Then, we take G_V from the subgraph generator as the input of the local model, and we can get the client-specific subgraph representation in client k:

$$\mathbf{h}^{spec}_k = \mathcal{R}\big(f^k_C(G_V)\big). \tag{7}$$

It should be emphasized that $\mathbf{h}^{spec}$ and $\mathbf{h}^{var}$, though derived from the same client-specific subgraph, are encoded using distinct encoders. After obtaining the client-specific subgraph representation, we use the sum of $\mathbf{h}^{inv}$ and $\mathbf{h}^{spec}$ as the overall graph representation to train the local model in client k. The local model loss is defined as:

$$\mathcal{L}^k_{local} = \ell(\mathbf{h}^{inv} + \mathbf{h}^{spec}, y). \tag{8}$$

Note that we fix the global model when training the local model. That is, $\mathbf{h}^{inv}$ obtained from the global model does not participate in gradient calculation and only optimizes the local model in minimizing $\mathcal{L}_{local}$.

4.3 Bi-Gradient Regularization strategy

After introducing the overall FedIGL framework, we now provide detailed insights into the optimization strategies for the global and local models. In this section, we focus on how to optimize the global model to achieve disentangled and effective feature learning across distributed clients. We begin with the global optimization objective for a specific client k:

$$min\ \mathcal{L}^k_{global}(\theta_{FSG}, \theta_{inv}, \theta_{var}). \tag{9}$$

In this objective, directly minimizing $\mathcal{L}^k_{global}$ for each client k independently would degenerate to conventional federated learning objectives. This approach fails to enforce collaboration among clients for learning client-invariant and client-specific representations. Accordingly, we delve deeper into optimizing the global model with respect to this objective.

Assume the invariant subgraph generator is ideal, such that it can extract the same semantic subgraph pattern from different distributed graphs. Then, the feature representations of such subgraphs should be consistent across all clients. Structural and feature differences can be reflected by GNN gradients, as proved in [48]. Consequently, if the encoder f_I is applied uniformly across clients, the optimization gradients with respect to θ_{inv} should also align across any clients k, k':

$$\nabla_{\theta_{inv}} \mathcal{L}^k{}_{global} = \nabla_{\theta_{inv}} \mathcal{L}^{k'}{}_{global}, \tag{10}$$

where $\nabla_{\theta_{inv}} \mathcal{L}^k{}_{global}$ represents the gradient of invariant branch encoder f_I when optimizing the global loss Eq.(6) for client k. Similarly, each client-specific subgraph captures features unique to each distribution. Therefore, the feature difference should drive different optimization directions for the variant branch encoder f_V with the same parameters. More precisely, the gradient between any clients should satisfy:

$$|| \nabla_{\theta_{var}} \mathcal{L}^k - \nabla_{\theta_{var}} \mathcal{L}^{k'} ||\geq \varepsilon, \tag{11}$$

where ε is a predefined margin enforcing gradient diversity across clients for the variant branch. This inequality indicates that the optimization direction of heterogeneous features from different clients on the same encoder should have significant differences.

Building upon these observations, we design a novel **Bi-Gradient Regularization strategy**. The central idea is to regulate the update directions of the invariant and variant encoders, such that 1) the invariant branch encoder gradients across clients are encouraged to be consistent; 2) the variant branch encoder gradients are enforced to be diverse (repulsion beyond margin ε).

We achieve this by using the global aggregated gradients from the previous round as reference directions. In each round t, all clients' current gradients are compared with the previous global aggregation gradients, and regularization is applied accordingly. The total loss for the global model is then defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{global} + \lambda \,||\, \nabla_{\theta^t_{inv}} \mathcal{L}_{global} - \overline{\nabla}_{\theta^{t-1}_{inv}} \,||_2 + \beta \max(0, \varepsilon - ||\, \nabla_{\theta^t_{var}} \mathcal{L}_{global} - \overline{\nabla}_{\theta^{t-1}_{var}} \,||_2), \quad (12)$$

where λ and β are hyperparameters controlling the strength of alignment and diversity regularization, $\overline{\nabla}_{\theta^{t-1}_{inv}}$ and $\overline{\nabla}_{\theta^{t-1}_{var}}$ is the aggregated gradients with the FedAvg [27] method. The second term, **consistency regularization**, encourages stability and agreement in the invariant encoder's optimization trajectory across rounds. The third term, **diversity regularization**, penalizes overly similar gradients in the variant encoder, ensuring meaningful divergence between client-specific representations.

4.4 Theoretical Analysis

We provide a theoretical analysis of the FedIGL framework with invariant and variant branch encoders with bi-gradient regularization. Our objective is to demonstrate that (i) the invariant encoder achieves convergence with gradient alignment regularization; (ii) the variant encoder yields distinguishable representations across clients due to repulsive regularization.

Let $\mathcal{C} = \{1, \ldots, K\}$ be the set of clients, and $\mathcal{D}_k$ denote the local dataset of client k. Let $\theta = (\theta_{\text{inv}}, \theta_{\text{var}}, \theta_G)$ denote the parameters of f_I, f_V and FSG, respectively. Define the task loss of client k as $\mathcal{L}_k(\theta)$. We denote the gradients as:

$$g_k^t = \nabla_{\theta_{\text{inv}}} \mathcal{L}_k(\theta^t), \quad h_k^t = \nabla_{\theta_{\text{var}}} \mathcal{L}_k(\theta^t),$$
$$\bar{g}^{t-1} = \frac{1}{K} \sum_{k=1}^{K} g_k^{t-1}, \quad \bar{h}^{t-1} = \frac{1}{K} \sum_{k=1}^{K} h_k^{t-1}.$$

The total optimization objective at round t is:

$$\min_{\theta} \sum_{k=1}^{K} \mathcal{L}_k(\theta) + \mathcal{R}_{\text{inv}} + \mathcal{R}_{\text{var}}, \quad (13)$$

where $\mathcal{R}_{\text{inv}}$ and $\mathcal{R}_{\text{var}}$ are the consistency and diversity regularization, respectively. In the federated setting, clients do not have access to the current global mean gradients $\bar{g}^t$, $\bar{h}^t$ at round t when performing local updates. Therefore, we compute the regularization terms using the previous round's statistics $\bar{g}^{t-1}$, $\bar{h}^{t-1}$, which are broadcast to clients at the start of each communication round. Nevertheless, the theoretical analysis below evaluates the variance and convergence behavior with respect to the true global means $\bar{g}^t$ and $\bar{h}^t$, as is standard in federated optimization literature.

Proposition 4.1 (Convergence of Invariant Encoder)**.** *Assume each $\mathcal{L}_k(\theta)$ is convex and L-smooth, and the learning rate $\eta \leq 1/L$. Under the* FEDAVG *scheme with gradient alignment $\mathcal{R}^t_{inv}$, the global invariant parameters θ^t_{inv} satisfy:*

$$\min_{t=1}^{T} \mathbb{E}\left[\left\| \nabla_{\theta_{inv}} \mathcal{L}_{global}(\theta^t) \right\|^2 \right] \leq \mathcal{O}\left(\frac{1}{T} \right). \quad (14)$$

Proposition 4.2 (Representation Separation for Variant Encoder)**.** *Let h_k^t denote the gradient of the variant encoder for client k. When $\mathcal{R}^t_{var}$ is minimized, then for all $k \neq k'$, the representations remain sufficiently separated:*

$$\|h_k^t - h_{k'}^t\| \geq \epsilon. \quad (15)$$

This guarantees per-client distinguishability in the variant branch.

Theorem 4.1 (Global Objective Convergence with Gradient Regularization)**.** *Let the total loss be $\mathcal{L}_{global}(\theta) + \mathcal{R}_{inv} + \mathcal{R}_{var}$. If each $\mathcal{L}_k(\theta)$ is convex and L-smooth, and $\eta = \mathcal{O}(1/\sqrt{T})$, then:*

$$\min_{t=1}^{T} \mathbb{E}\left[\left\| \nabla \mathcal{L}_{global}(\theta^t) \right\|^2 \right] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}} \right). \quad (16)$$

Indicating convergence. Furthermore, the learned representation satisfies: (i) invariant alignment across clients in θ_{inv}, and (ii) per-client diversity in θ_{var}.

Based on the above formulation, we provide a theoretical justification for the effectiveness of our Bi-gradient regularization strategy. Specifically, Proposition 4.1 shows that the gradient alignment regularization term $\mathcal{R}_{\text{inv}}$ reduces the variance of client gradients in the invariant encoder, thus promoting convergence in federated optimization. Proposition 4.2 demonstrates that the repulsive regularization $\mathcal{R}_{\text{var}}$ enforces diversity among the variant gradients, enabling the model to capture client-specific characteristics. Together, these results support the disentanglement of invariant and variant subgraph patterns in a federated setting. Theorem 4.1 further establishes the convergence guarantee of our overall optimization procedure under bi-gradient regularization, quantifying the trade-off between alignment and diversity in gradient space. The proof of the above proposition and theorem see Appendix B.

5 Experiments

In this section, we conduct extensive experiments on graph-level classification and clustering tasks in various cross-dataset and cross-domain scenarios to validate the superiority of FedIGL. The following research questions need to be validated. **(RQ1)** Can FedIGL achieve better performance compared to SOTA baselines? **(RQ2)** Does FedIGL converge under the constraints of bi-gradient optimization? **(RQ3)** How does each of the strategies we propose contribute to the final performance? **(RQ4)** How about the hyperparameter sensitivity of FedIGL?

5.1 Experiment Setup

Benchmark Datasets. We employed a total of 19 diverse datasets across multiple domains to conduct comprehensive evaluations on both classification and clustering tasks. These domains include Small Molecules (e.g., MUTAG, BZR, COX2, DHFR, PTC_MR, AIDS, BZR_MD, and NCI1), Bioinformatics (e.g., DD, PROTEINS, OHSU, and Peking_1), Synthetic (SYNTHETIC), Social Networks (e.g., COLLAB, IMDBMULTI, and IMDB-BINARY), and Computer Vision (e.g., Letter-high, Letter-low, and Letter-med). Regarding classification tasks, We follow the settings in [32], which include six distinct experimental designs: (1) cross-dataset setting utilizing seven small molecule datasets (SM), and (2)-(6) settings that incorporate both cross-dataset and cross-domain aspects, based on datasets from two different domains (BIO-SM, SM-CV) and three different domains (BIO-SM-SN, BIO-SN-CV, SM-SN-CV). For clustering tasks, we adopt the protocols in [22], including five types of non-IID settings: (1) 2 clusters within the same domain (SM), (2) 3 clusters within the same domain (SN), (3) 15 clusters within the same domain (CV), (4) 2 clusters across two domains (SM-BIO), and (5) 2 clusters across three domains (SM-BIO-SY). The dataset and experimental implementation details are provided in Appendix C.1.

Baseline Methods. In both classification and clustering tasks, we compare FedIGL with two classical federated learning methods, FedAvg [27] and FedProx [21]. Additionally, we include four state-of-the-art federated graph learning methods: FedSage [52], GCFL [48], FedStar [31], and FedSSP [32]. For clustering tasks specifically, we also compare with FedGCN [22].

Implementation Details. To ensure fair comparisons, all methods, including FedIGL and baselines, were implemented in PyTorch and executed on the same NVIDIA GeForce RTX 3090 GPU. For graph-level structure embeddings, we use a three-layer Graph Isomorphism Network (GIN) [51] with a hidden dimension of 64 and batch size of 128 [34]. Model optimization is performed using the Adam optimizer with a learning rate of 1e-3. Dropout is set to 0.5 and weight decay to 5e-4 to improve generalization.

5.2 Experimental Results

Performance Comparison (RQ1) . Tab. 1 and Tab. 2 present a comparison of the performance of FedIGL against SOTA methods on graph-level classification and clustering tasks. In the classification task, FedIGL achieves the best results in 5 out of 6 classification settings and ranks second in the remaining one, demonstrating the most robust overall performance. Notably, under the single-domain SM setting, FedIGL improves over the FedSSP about a 4.3% relative gain. Moreover, although FedSSP slightly outperforms FedIGL on SM-CV, FedIGL shows more pronounced advantages in the more challenging multi-domain scenarios, indicating better generalization under stronger distribution shifts. In the clustering task, FedIGL ranks among the top methods across the five clustering settings,

demonstrating more stable clustering quality under non-IID setting. Notably, on SN and CV, FedGCN achieves a slight success on specific tasks but remains suggesting that the two methods emphasize cluster alignment consistency and clustering decision correctness, respectively. Existing methods often perform well on specific tasks but remain vulnerable to spurious correlations, which limits their generalization.

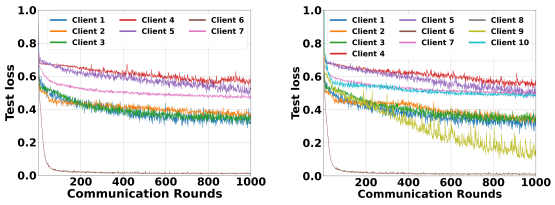
Table 1: Comparison with state-of-the-art methods on one cross-dataset and five cross-domain settings for classification tasks. The best is marked with **boldface** and the second best is with underline.

Methods	Single-domain	Double-domain		Multi-Domain		
	SM	BIO-SM	SM-CV	BIO-SM-SN	BIO-SN-CV	SM-SN-CV
FedAvg (ASTAT17)	74.12 ± 2.10	67.82 ± 1.63	81.21 ± 1.00	67.31 ± 2.56	70.93 ± 2.91	75.33 ± 1.06
FedProx (arXiv18)	69.35 ± 3.36	67.27 ± 4.17	70.02 ± 2.27	63.89 ± 4.33	69.32 ± 1.75	67.15 ± 2.25
FedSage (NeurIPS21)	75.61 ± 1.16	72.60 ± 3.18	76.23 ± 0.49	70.84 ± 0.88	69.69 ± 1.11	73.36 ± 0.86
GCFL (NeurIPS21)	77.71 ± 1.53	72.05 ± 2.20	72.64 ± 0.71	70.43 ± 1.39	67.91 ± 2.15	71.79 ± 0.21
FedStar (AAAI23)	78.63 ± 2.11	72.71 ± 1.22	78.84± 1.07	72.60 ± 2.45	69.51 ± 0.84	75.94 ± 0.40
FedSSP (NeurIPS24)	79.62 ± 2.23	73.66 ± 2.34	**84.29 ± 0.68**	72.37 ± 2.18	75.07 ± 2.70	79.12 ± 1.23
FedIGL(ours)	**83.07 ± 1.76**	**77.02 ± 1.32**	83.14 ± 0.28	**75.25 ± 1.13**	**78.50 ± 0.44**	**79.23 ± 1.28**

Table 2: Comparison with state-of-the-art methods on three cross-dataset and two cross-domain settings for clustering tasks. Please note that the FGL methods marked with the symbol * have been adapted from classification to clustering tasks, as was done in previous studies. The best is marked with **boldface** and the second best is with underline.

Domain	Metric	FedAvg ASTAT17	FedProx arXiv18	FedSage* NeurIPS21	GCFL* NeurIPS21	FedStar* AAAI23	FedSSP* NeurIPS24	FedGCN AAAI25	FedIGL Ours
SM	ACC	35.3 ± 1.1	69.4 ± 3.4	55.6 ± 1.4	61.1 ± 1.8	58.9 ± 2.4	76.4 ± 0.5	75.9 ± 0.8	**77.0 ± 1.1**
	NMI	10.2 ± 1.5	8.4 ± 2.9	12.2 ± 1.3	8.7 ± 2.4	12.0 ± 1.2	12.9 ± 3.0	24.9 ± 3.0	**25.7 ± 1.5**
	ARI	8.4 ± 0.9	9.5 ± 2.1	7.6 ± 0.6	9.4 ± 2.4	0.1 ± 0.8	34.1 ± 2.3	31.1 ± 3.4	**36.3 ± 1.4**
	F1	52.7 ± 1.3	48.2 ± 2.5	50.2 ± 1.0	43.3 ± 1.6	49.7 ± 2.8	68.2 ± 2.5	67.1 ± 1.5	**69.4 ± 2.5**
SN	ACC	61.5 ± 2.2	71.4 ± 2.8	53.3 ± 1.9	52.1 ± 2.3	51.7 ± 2.7	73.6 ± 2.3	66.6 ± 2.3	**73.7 ± 0.3**
	NMI	15.6 ± 1.7	11.6 ± 2.0	14.8 ± 1.4	12.5 ± 2.3	13.7 ± 2.8	11.8 ± 3.4	**30.4 ± 6.6**	23.1 ± 1.2
	ARI	10.3 ± 1.2	12.3 ± 2.7	11.6 ± 2.8	13.2 ± 2.3	12.4 ± 1.9	31.9 ± 1.7	**34.1 ± 5.3**	32.6 ± 2.1
	F1	58.1 ± 2.8	53.7 ± 2.2	49.3 ± 2.0	52.3 ± 1.6	50.7 ± 2.3	**64.9 ± 2.6**	50.7 ± 2.4	62.7 ± 2.7
CV	ACC	33.8 ± 0.7	38.6 ± 2.1	10.1 ± 1.4	13.7 ± 2.0	12.4 ± 2.7	34.2 ± 2.6	34.6 ± 2.8	**39.2 ± 1.1**
	NMI	15.9 ± 2.3	12.1 ± 2.7	30.5 ± 1.7	17.7 ± 2.4	22.4 ± 2.5	12.0 ± 2.4	**34.2 ± 1.4**	27.3 ± 1.8
	ARI	10.5 ± 2.0	12.7 ± 2.1	13.6 ± 1.8	14.3 ± 2.7	15.3 ± 2.1	33.1 ± 3.3	19.3 ± 1.8	**38.1 ± 2.6**
	F1	49.7 ± 1.5	54.5 ± 2.5	10.4 ± 1.7	13.2 ± 1.4	11.6 ± 1.9	35.3 ± 1.5	31.6 ± 3.1	**36.2 ± 2.4**
SM-BIO	ACC	54.3 ± 2.9	67.1 ± 1.9	57.4 ± 2.2	60.1 ± 1.8	59.5 ± 1.6	72.3 ± 2.1	69.2 ± 0.6	**73.6 ± 1.0**
	NMI	11.8 ± 2.1	7.9 ± 2.3	5.2 ± 2.1	4.7 ± 2.4	5.3 ± 1.6	11.3 ± 1.2	14.0 ± 2.7	**23.0 ± 1.6**
	ARI	7.2 ± 3.0	8.7 ± 1.6	4.2 ± 2.7	3.2 ± 2.3	3.8 ± 2.0	**32.4 ± 2.1**	17.5 ± 3.1	30.7 ± 2.5
	F1	48.7 ± 2.4	44.3 ± 2.5	49.9 ± 0.5	47.3 ± 1.5	51.7 ± 2.2	**63.7 ± 1.8**	59.1 ± 0.9	59.4 ± 3.8
SM-BIO-SY	ACC	56.1 ± 2.8	68.3 ± 1.1	57.6 ± 1.9	59.1 ± 2.0	57.9 ± 2.6	75.9 ± 0.4	68.6 ± 1.3	**76.7 ± 2.2**
	NMI	12.5 ± 1.4	8.0 ± 2.2	20.6 ± 1.9	14.4 ± 2.2	15.7 ± 2.4	13.9 ± 1.4	13.5 ± 2.1	**22.7 ± 1.8**
	ARI	7.8 ± 0.9	8.9 ± 2.1	17.6 ± 2.4	13.7 ± 2.8	16.1 ± 3.0	**34.8 ± 2.6**	17.2 ± 3.6	34.6 ± 1.6
	F1	52.1 ± 1.7	46.7 ± 2.0	49.4 ± 1.7	52.3 ± 1.9	52.3 ± 2.2	**69.7 ± 2.3**	59.4 ± 3.8	67.5 ± 1.9

Convergence Analysis (RQ2) . We visualize the graph classification testing loss with respect to the communication rounds to show the convergence of FedIGL clients under the non-IID setting, as shown in Fig. 3. Despite the inherent heterogeneity of data across clients, our method demonstrates a smooth and monotonic decrease in the global objective over communication rounds. This behavior aligns with the theoretical guarantees of federated optimization. In particular, our introduction of bi-gradient regularization further stabilizes the learning dynamics by mitigating the divergence caused by client drift, leading to faster and more consistent convergence. More non-IID settings in Appendix C.4.



(a) SM (b) SM-BIO

Figure 3: The loss curves of each clisnt on the SM and SM-BIO of non-IID setting.

Ablation Study (RQ3) . We perform ablation studies to assess how the proposed Bi-Gradient Regularization strategy contributes to the overall performance. Tab. 3 reports results of FedIGL and its variants under non-IID settings: (i) removing both consistency regularization (CR) and diversity regularization (DR), (ii) enabling only one of them (CR or DR), and (iii) enabling both. An intuitive observation is that FedIGL performs best when CR and DR are enabled simultaneously. Each component on its own still delivers consistent gains across most datasets. On the SM-BIO-SY clustering benchmark, introducing CR alone leads to negligible gains over the baseline. In contrast, enabling DR alone yields a noticeable improvement of 2.54%. This indicates that under multi-domain shift, encouraging representation diversity better mitigates domain bias and leads to more stable clustering. Overall, the empirical trends align with our theoretical analysis, supporting the design of jointly integrating CR and DR. Additional results are provided in Appendix C.2.

Table 3: The ablation study covers both classification and clustering tasks. A checkmark (✓) indicates inclusion of the strategy, while a cross (✗) indicates its exclusion. Our non-IID settings include single-domain, double-domain, and multi-domain scenarios, corresponding to the SM, SM-BIO, SM-BIO-SN for the classification task and SM-BIO-SY for the clustering task, respectively.

CR	DR	SM		SM-BIO		SM-BIO-SN (SY)	
		Classification	Clustering	Classification	Clustering	Classification	Clustering
✗	✗	78.21	72.45	74.67	69.73	70.66	68.33
✓	✗	79.84	73.04	75.83	70.34	71.24	68.36
✗	✓	80.55	75.39	76.01	71.28	72.61	70.87
✓	✓	**83.07**	**77.04**	**77.02**	**73.71**	**75.25**	**76.71**

Hyper-Parameter Study (RQ4) . We investigate the sensitivity of several hyper-parameters in our method, including the invariance regularization strength λ, variance regularization strength β, divergence parameter ε, and invariant subgraph ratio τ. Under the SM dataset setting, we evaluate the performance of FedIGL across various hyperparameter combinations on classification and clustering tasks, as shown in Fig. 4. The results indicate that the optimal values of λ, β, and ε differ by task. For classification, $\lambda = 0.05$, $\beta = 0.2$, and $\varepsilon = 0.1$ provide the best performance balance; for clustering, $\lambda = 0.2$, $\beta = 0.25$, and $\varepsilon = 0.15$ perform best. This likely reflects distinct requirements for graph representations, suggesting that client subgraph invariance and variability are influenced by downstream tasks. The parameter τ governs the proportion of invariant subgraphs; a large τ may include excessive variant structures, while a small τ may limit structural capture. In our experiments, $\tau = 0.25$ balances shared and client-specific structures effectively in both tasks. Additionally, we conduct further hyper-parameter sensitivity analyses under non-IID settings in Appendix C.3.

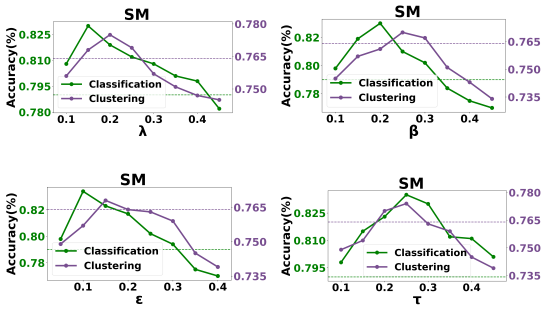


Figure 4: Hyperparameter Sensitivity analysis on SM setting. The x-axis represents the four hyperparameters: λ, β, ε, and τ, with the left and right y-axis representing the classification and clustering accuracy, respectively.

6 Conclusion

In this work, we present a novel Federated Invariant Graph Learning framework from a fresh perspective, aimed at capturing invariant subgraph structures to mitigate client distribution shifts. We propose a Bi-Gradient Regularization strategy applying consistency regularization to the invariant subgraph encoder and diversity regularization to the variant one, which enhances graph representation quality, stability and model performance. Overall, as a pioneering study, FedIGL provides valuable insights for addressing the graph structural differences associated with client distributional heterogeneity and is supported by extensive experimental and theoretical analysis. While our approach intuitively protects client privacy by avoiding the sharing of prototype structures, future work will further explore stringent measures for model privacy preservation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) (Grant No. 62562026, 62506102), the Key Research and Development Program of Hainan Province (Grant No. ZDYF2024GXJS014, ZDYF2023GXJS163), the Hainan Province Graduate Innovation Research Project (No. Qhyb2023-104), and the Natural Science Foundation of Hainan University (Grant No. XJ2400009401).

References

[1] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 3438–3450, 2021.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019.

[3] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *Proceedings of the International Conference on Machine Learning*, pages 528–539, 2020.

[4] Bowen Deng, Lele Fu, Jialong Chen, Sheng Huang, Tianchi Liao, Zhang Tao, and Chuan Chen. Towards understanding parametric generalized category discovery on graphs. In *Proceedings of the International Conference on Machine Learning*, 2025.

[5] Bowen Deng, Tong Wang, Lele Fu, Sheng Huang, Chuan Chen, and Tao Zhang. THESAURUS: contrastive graph clustering by swapping fused gromov-wasserstein couplings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16199–16207, 2025.

[6] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv:2010.01412*, 2020.

[7] Lele Fu, Bowen Deng, Sheng Huang, Tianchi Liao, Shirui Pan, and Chuan Chen. Less is more: Federated graph learning with alleviating topology heterogeneity from a causal perspective. In *Proceedings of the International Conference on Machine Learning*, 2025.

[8] Lei Gong, Wenxuan Tu, Sihang Zhou, Long Zhao, Zhe Liu, and Xinwang Liu. Deep fusion clustering network with reliable structure preservation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7792–7803, 2024.

[9] Renxiang Guan, Wenxuan Tu, Siwei Wang, Jiyuan Liu, Dayu Hu, Chang Tang, Yu Feng, Junhong Li, Baili Xiao, and Xinwang Liu. Structure-adaptive multi-view graph clustering for remote sensing data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16933–16941, 2025.

[10] Ming Hu, Yue Cao, Anran Li, Zhiming Li, Chengwei Liu, Tianlin Li, Mingsong Chen, and Yang Liu. Fedmut: Generalized federated learning via stochastic mutation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12528–12537, 2024.

[11] Yaowen Hu, Wenxuan Tu, Yue Liu, Miaomiao Li, Wenpeng Lu, Zhigang Luo, Xinwang Liu, and Ping Chen. Divide-then-rule: A cluster-driven hierarchical interpolator for attribute-missing graphs. In *Proceedings of the ACM International Conference on Multimedia*, pages 10905–10914, 2025.

[12] Yaowen Hu, Wenxuan Tu, Yue Liu, Xinhang Wan, Junyi Yan, Taichun Zhou, and Xinwang Liu. Scalable attribute-missing graph clustering via neighborhood differentiation. In *Proceedings of the International Conference on Machine Learning*, 2025.

[13] Sheng Huang, Lele Fu, Tianchi Liao, Bowen Deng, Chuanfu Zhang, and Chuan Chen. Fedbg: Proactively mitigating bias in cross-domain graph federated learning using background data. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 5408–5416, 2025.

[14] Wenke Huang, Guancheng Wan, Mang Ye, and Bo Du. Federated graph semantic and structural learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2023.

[15] Tianrui Jia, Haoyang Li, Cheng Yang, Tao Tao, and Chuan Shi. Graph invariant learning with subgraph co-mixup for out-of-distribution generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8562–8570, 2024.

[16] Bo Jiang, Ziyan Zhang, Doudou Lin, Jin Tang, and Bin Luo. Semi-supervised learning with graph learning-convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11313–11320, 2019.

[17] Siyang Jiang, Rui Fang, Hsi-Wen Chen, Wei Ding, and Ming-Syan Chen. Dual alignment framework for few-shot learning with inter-set and intra-set shifts. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2025.

[18] Siyang Jiang, Xian Shuai, and Guoliang Xing. Artfl: Exploiting data resolution in federated learning for dynamic runtime inference via multi-scale training. In *Proceedings of the ACM/IEEE International Conference on Information Processing in Sensor Networks*, pages 27–38, 2024.

[19] Siyang Jiang, Hao Yang, Qipeng Xie, Chuan Ma, Sen Wang, Zhe Liu, Tao Xiang, and Guoliang Xing. Towards compute-efficient byzantine-robust federated learning with fully homomorphic encryption. *Nature Machine Intelligence*, pages 1–12, 2025.

[20] Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. Learning invariant graph representations for out-of-distribution generalization. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 11828–11841, 2022.

[21] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of the Conference on Machine Learning and Systems*, pages 429–450, 2020.

[22] Jingxin Liu, Jieren Cheng, Renda Han, Wenxuan Tu, Jiaxin Wang, and Xin Peng. Federated graph-level clustering network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18870–18878, 2025.

[23] Jingxin Liu, Renda Han, Wenxuan Tu, Haotian Wang, Junlong Wu, and Jieren Cheng. Federated node-level clustering network with cross-subgraph link mending. In *Proceedings of the International Conference on Machine Learning*, 2025.

[24] Meng Liu, Yue Liu, Ke Liang, Wenxuan Tu, Siwei Wang, Sihang Zhou, and Xinwang Liu. Deep temporal graph clustering. In *Proceedings of the International Conference on Learning Representations*, 2024.

[25] Suyuan Liu, Siwei Wang, Ke Liang, Junpu Zhang, Zhibin Dong, Tianrui Liu, En Zhu, Xinwang Liu, and Kunlun He. Alleviate anchor-shift: Explore blind spots with cross-view reconstruction for incomplete multi-view clustering. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 87509–87531, 2024.

[26] Yue Liu, Xihong Yang, Sihang Zhou, Xinwang Liu, Zhen Wang, Ke Liang, Wenxuan Tu, Liang Li, Jingcan Duan, and Cancan Chen. Hard sample aware network for contrastive deep graph clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8914–8922, 2023.

[27] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

[28] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *Proceedings of the International Conference on Machine Learning*, pages 7721–7735, 2021.

[29] Yanhu Mo, Xiao Wang, Shaohua Fan, and Chuan Shi. Graph contrastive invariant learning from the causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8904–8912, 2024.

[30] Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. *arXiv:2007.08663*, 2020.

[31] Yue Tan, Yixin Liu, Guodong Long, Jing Jiang, Qinghua Lu, and Chengqi Zhang. Federated learning on non-iid graphs via structural knowledge sharing. In *Proceedings of the AAAI conference on artificial intelligence*, pages 9953–9961, 2023.

[32] Zihan Tan, Guancheng Wan, Wenke Huang, and Mang Ye. Fedssp: Federated graph learning with spectral knowledge and personalized preference. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 34561–34581, 2024.

[33] Wenxuan Tu, Renxiang Guan, Sihang Zhou, Chuan Ma, Xin Peng, Zhiping Cai, Zhe Liu, Jieren Cheng, and Xinwang Liu. Attribute-missing graph clustering network. pages 15392–15401, 2024.

[34] Wenxuan Tu, Qing Liao, Sihang Zhou, Xin Peng, Chuan Ma, Zhe Liu, Xinwang Liu, Zhiping Cai, and Kunlun He. Rare: Robust masked graph autoencoder. *IEEE Transactions on Knowledge and Data Engineering*, 36(10):5340–5353, 2024.

[35] Wenxuan Tu, Bin Xiao, Xinwang Liu, Sihang Zhou, Zhiping Cai, and Jieren Cheng. Revisiting initializing then refining: An incomplete and missing graph imputation network. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):3244–3257, 2025.

[36] Wenxuan Tu, Sihang Zhou, Xinwang Liu, Zhiping Cai, Yawei Zhao, Yue Liu, and Kunlun He. Wage: Weight-sharing attribute-missing graph autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5760–5777, 2025.

[37] Wenxuan Tu, Sihang Zhou, Xinwang Liu, Chunpeng Ge, Zhiping Cai, and Yue Liu. Hierarchically contrastive hard sample mining for graph self-supervised pre-training. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16748–16761, 2024.

[38] Wenxuan Tu, Sihang Zhou, Xinwang Liu, Xifeng Guo, Zhiping Cai, En Zhu, and Jieren Cheng. Deep fusion clustering network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9978–9987, 2021.

[39] Wenxuan Tu, Sihang Zhou, Xinwang Liu, Yue Liu, Zhiping Cai, En Zhu, Changwang Zhang, and Jieren Cheng. Initializing then refining: A simple graph attribute imputation network. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3494–3500, 2022.

[40] Guancheng Wan, Wenke Huang, and Mang Ye. Federated graph learning under domain shift with generalizable prototypes. In *Proceedings of the AAAI conference on artificial intelligence*, pages 15429–15437, 2024.

[41] Haozhao Wang, Song Guo, Bin Tang, Ruixuan Li, Yutong Yang, Zhihao Qu, and Yi Wang. Heterogeneity-aware gradient coding for tolerating and leveraging stragglers. *IEEE Transactions on Computers*, 71(4):779–794, 2021.

[42] Haozhao Wang, Yichen Li, Wenchao Xu, Ruixuan Li, Yufeng Zhan, and Zhigang Zeng. Dafkd: Domain-aware federated knowledge distillation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 20412–20421, 2023.

[43] Haozhao Wang, Zhihao Qu, Song Guo, Ningqi Wang, Ruixuan Li, and Weihua Zhuang. Losp: Overlap synchronization parallel with local compensation for fast distributed training. *IEEE Journal on Selected Areas in Communications*, 39(8):2541–2557, 2021.

[44] Haozhao Wang, Haoran Xu, Yichen Li, Yuan Xu, Ruixuan Li, and Tianwei Zhang. Fedcda: Federated learning with cross-rounds divergence-aware aggregation. In *Proceedings of the International Conference on Learning Representations*, 2023.

[45] Haozhao Wang, Peirong Zheng, Xingshuo Han, Wenchao Xu, Ruixuan Li, and Tianwei Zhang. Fednlr: Federated learning with neuron-wise learning rates. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*, pages 3069–3080, 2024.

[46] Lingren Wang, Wenxuan Tu, Jieren Cheng, Jianan Wang, Xiangyan Tang, and Chenchen Wang. Discovering maximum frequency consensus: Lightweight federated learning for medical image segmentation. In *Proceedings of the ACM International Conference on Multimedia*, 2025.

[47] Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *Proceedings of the International Conference on Learning Representations*, 2022.

[48] Han Xie, Jing Ma, Li Xiong, and Carl Yang. Federated graph classification over non-iid graphs. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 18839–18852, 2021.

[49] Qipeng Xie, Siyang Jiang, Linshan Jiang, Yongzhi Huang, Zhihe Zhao, Salabat Khan, Wangchen Dai, Zhe Liu, and Kaishun Wu. Efficiency optimization techniques in privacy-preserving federated learning with homomorphic encryption: A brief survey. *IEEE Internet of Things Journal*, 11(14):24569–24580, 2024.

[50] Fangqiang Xu, Wenxuan Tu, Fan Feng, Malitha Gunawardhana, Jiayuan Yang, Yun Gu, and Jichao Zhao. Dynamic position transformation and boundary refinement network for left atrial segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention*, pages 209–219, 2024.

[51] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv:1810.00826*, 2018.

[52] Ke Zhang, Carl Yang, Xiaoxiao Li, Lichao Sun, and Siu Ming Yiu. Subgraph federated learning with missing neighbor generation. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 6671–6682, 2021.

[53] Yinlin Zhu, Xunkai Li, Zhengyu Wu, Di Wu, Miao Hu, and Rong-Hua Li. Fedtad: Topology-aware data-free knowledge distillation for subgraph federated learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2024.

[54] Xiang Zhuang, Qiang Zhang, Keyan Ding, Yatao Bian, Xiao Wang, Jingsong Lv, Hongyang Chen, and Huajun Chen. Learning invariant molecular representation in latent discrete space. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 78435–78452, 2023.

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: Our main contributions include: (i) We pioneer the application of invariant learning within federated graph learning to effectively mitigate spurious relational knowledge shared; (ii) We introduce a novel Bi-Gradient Regularization strategy that facilitates collaborative learning of disentangled subgraph representations while ensuring data privacy; (iii) We rigorously validate the effectiveness of our framework through comprehensive empirical evaluations and theoretical analyses. These contributions are clearly articulated in both the abstract and introduction sections.

   Guidelines:

   - The answer NA 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 NA 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 terms of computational efficiency in Section6.

   Guidelines:

   - The answer NA 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: [Yes]

   Justification: We analyze the issue of structural heterogeneity in federated graph learning from the perspective of invariant learning. The assumptions of the federated generator are introduced in Section 4.2, and the theoretical analysis of the optimization strategies in the federated graph learning framework is presented in Section 4.4.

   Guidelines:

   - The answer NA 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: We provide all the details of our method in Section5.1 and present the necessary hyper-parameters in Section2 to ensure the reproducibility of our method. Furthermore,we provide the code of our proposed method in the supplementary material.

   Guidelines:

   - The answer NA 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: We include the code of our proposed method in the supplementary material. The necessary environments and data preparation procedures are provided in the GitHub repository of our method.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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 recommen

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We specify all the necessary hyper-parameters in Section2 and Appendix C.2.

   Guidelines:

   - The answer NA 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: We don't report the error bar in our experimental results because the previous works do not report the error bar and we follow them. All of our experimental results are averaged over 5 runs of 5 different seeds.

Guidelines:

- The answer NA 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 the compute workers in Section2, and list the params and time of execution in 4 and 3.

   Guidelines:

   - The answer NA 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: We conduct the research with the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA 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: [Yes]

Justification: We provide the broader impacts of our work in Section 6.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: Our work does not pose such risks.

Guidelines:

- The answer NA 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 datasets, models, and code involved in our paper are open source.

Guidelines:

- The answer NA 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: [No]

    Justification: Our paper does not release new assets

    Guidelines:

    - The answer NA 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: [No]

    Justification: We do not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA 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: [No]

    Justification: We do not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA 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 rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: We do not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Algorithm

Algorithm 1 Optimization process of FedIGL

Input: Maximum epoch T; Number of clients K; The distributed Non-IID datasets $\{\mathcal{D}_k\}_{k=1}^{K}$; hyper-parameters t, λ, β, ε.
Output: Model trained with FedIGL.
Initialize $\{\theta_{FSG}, \theta_{inv}, \theta_{var}\}$.
for $t = 1$ **to** T **do**
  for $k = 1$ **to** K **do**
    Obtain G_I, G_V for each graph $G \in \mathcal{D}_k$ with t by Eq. (4).
    Obtain subgraph representation $\mathbf{h}^{inv}, \mathbf{h}^{var}$ for G_I, G_V.
    Calculate $\mathcal{L}_{global}^{k}$ for client k with λ, β, ε by Eq. (12).
    Update $\{\theta_{FSG}, \theta_{inv}, \theta_{var}\}$ by Stochastic Gradient Descent.
    Fix the parameters of $\{\theta_{FSG}, \theta_{inv}, \theta_{var}\}$.
    Obtain client-specific subgraph representation $\mathbf{h}^{spec}$ with G_V.
    Calculate $\mathcal{L}_{local}^{k}$ for client k by Eq. (8).
    Update local model f_C^k for client k by Stochastic Gradient Descent.
  end for
  Aggregate $\{\theta_{FSG}, \theta_{inv}, \theta_{var}\}$.
end for

B Proofs in Section 4.4

Lemma B.1 (Gradient Variance Reduction for Invariant Branch)**.** *Suppose each $L_k(\theta)$ is L-smooth and the initial variance of local invariant gradients satisfies $\mathbb{E}\|g_k^t - \bar{g}^t\|^2 \leq \sigma^2$. Then, applying the regularization $\mathcal{R}_{inv}$ leads to exponential decay in gradient variance:*

$$\mathbb{E}\|g_k^t - \bar{g}^t\|^2 \leq (1 - \lambda_1\eta)^t\sigma^2,$$

assuming $0 < \eta \leq 1/L$ and $\lambda_1 > 0$.

Proof. Using L-smoothness, we apply the standard gradient descent update:

$$g_k^t = g_k^{t-1} - \eta\nabla^2_{\theta_{\text{inv}}}\mathcal{L}_k(\theta^{t-1}) + \mathcal{O}(\eta^2).$$

Applying the regularization $\mathcal{R}_{\text{inv}}$ effectively forces each g_k^t to align with the global average $\bar{g}^{t-1}$, thereby reducing the variance. The gradient variance evolves as:

$$\mathbb{E}\|g_k^t - \bar{g}^t\|^2 \leq (1 - \lambda_1\eta)^t\sigma^2.$$

□

B.1 Proofs of Proposition 4.1

Proof. With convexity and smoothness, we apply standard convergence results for Stochastic Gradient Descent (SGD) with variance reduction. By Lemma B.1, the variance of g_k^t decreases over time, which helps stabilize the FedAvg updates. Using the descent lemma and unbiased gradients, we have:

$$\mathcal{L}_{\text{global}}(\theta^{t+1}) \leq \mathcal{L}_{\text{global}}(\theta^t) - \eta\left\|\nabla_{\theta_{\text{inv}}}\mathcal{L}_{\text{global}}(\theta^t)\right\|^2 + \eta^2\mathcal{L}\sigma^2.$$

Averaging over T rounds gives the $\mathcal{O}(1/T)$ rate. □

B.2 Proofs of Proposition 4.2

Proof. The penalty $\max(0, \epsilon - \|\bar{h}^{t-1} - h_k^t\|)^2$ pushes each h_k^t to be at least ϵ away from the mean. If $\|\bar{h}^{t-1} - h_k^t\| < \epsilon$, the penalty is active, increasing the loss. At the optimum, this penalty is zero, implying that $\|\bar{h}^{t-1} - h_k^t\| \geq \epsilon$, assuming that the mean of all clients' gradients remains close to the previous gradients. And thus pairwise $\|h_k^t - h_{k'}^t\| \geq \epsilon$. Consequently, the absolute difference between the gradients of any two clients exceeds the penalty term. □

B.3 Proofs of Theorem 4.1

Proof. The total loss includes smooth convex functions and squared penalties. Using standard convergence bounds for smooth objectives with gradient regularization and a diminishing step size $\eta = \mathcal{O}(1/\sqrt{T})$, we get the convergence rate of $\mathcal{O}(1/\sqrt{T})$ for gradient norms. The variance reduction and margin-enforcing terms ensure stable updates for both branches. □

Response: On Computational Complexity. The per-round per-client computation in FedIGL mainly involves two parts: 1. The **Federated Subgraph Generator (FSG)**, which consists of a L_1-layer GNN with complexity $O\big(L_1(|E|d + |V|d^2)\big)$ to encode node features, and an edge-wise MLP scorer with complexity $O(|E|d^2)$ for selecting invariant edges; 2. The **dual-branch GNN encoder**, each branch having L_2 layers, resulting in a total complexity of $O\big(L_2(|E|d + |V|d^2)\big)$. Here $|V|$ and $|E|$ denote the number of nodes and edges in the client's local graph, respectively; d is the feature dimensionality of each node; and L_1, L_2 represent the number of GNN layers in the Federated Subgraph Generator and dual-branch encoder, respectively. Hence, the overall per-client cost per round is: $O\big((L_1 + L_2)(|E|d + |V|d^2) + |E|d^2\big)$, which remains linear in the graph size and thus comparable to standard GNN-based federated learning methods. The extra overhead from edge scoring is lightweight and only applied once per round.

C Additional Experiments

C.1 Experiment Dataset

Evaluation Metrics In classification tasks, we employ Accuracy (ACC) to assess the performance of the method. Regarding clustering tasks, we utilize widely-adopted clustering result evaluation metrics, namely Accuracy (ACC), Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and F1 Score (F1). These metrics provide multi-faceted evaluations of the clustering results. Specifically, larger values of these metrics correspond to better performance. They imply more efficient data partitioning and a more accurate capture of the underlying data structure, thereby demonstrating the superiority of the clustering method in organizing data and uncovering its inherent characteristics.

Table 4: A superscript "1" in the upper-right corner indicates that the dataset is only used for classification tasks, "2" indicates that the dataset is only used for clustering tasks, and the absence of a superscript indicates that the dataset is used for both classification and clustering tasks.

Datasets	Domain	Classes	Graphs	A.Nodes	A.Edges
MUTAG	SM	2	188	17.93	19.79
BZR		2	405	35.75	38.36
COX2		2	467	41.22	43.45
DHFR		2	756	42.43	44.54
PTC_MR		2	344	14.29	14.69
AIDS		2	2000	15.69	16.20
NCI1[1]		2	4110	29.87	32.30
BZR_MD[2]		2	306	21.30	225.06
DD[2]	BIO	2	1178	284.32	715.66
PROTEINS		2	1113	39.06	72.82
OHSU[1]		2	79	82.01	199.66
Peking_1[1]		2	85	39.31	77.35
SYNTHETIC[2]	SY	2	300	100.00	196.00
COLLAB[2]	SN	3	5000	74.49	2457.78
IMDB-MULTI		3	1500	13.00	65.94
IMDB-BINARY		2	1000	19.77	96.53
Letter-high	CV	15	2250	4.67	4.50
Letter-low		15	2250	4.68	3.13
Letter-med		15	2250	4.67	3.21

C.2 Ablation Study

We present additional ablation experiments across multiple non-IID settings, as shown in Tab.5. The results demonstrate that the combination of two optimization strategies significantly outperforms the use of individual strategies, thereby validating the effectiveness of our proposed optimization framework. Notably, our method not only excels in specific settings but also exhibits consistent performance across a wide range of scenarios, highlighting its robustness and adaptability to varying conditions. The proposed approach effectively handles diverse datasets and domain configurations, yielding high-quality graph representations that deliver superior performance in the classification tasks.

Table 5: Ablation study of key components, namely Consistency Regularization (CR) and Diversity Regularization (DR), of FedIGL on double-domain and multi-domain settings (SM-CV, BIO-SN-CV and SM-SN-CV) in the classification.

CR	DR	SM-CV	BIO-SN-CV	SM-SN-CV
✗	✗	78.35	73.89	69.82
✓	✗	79.76	74.95	72.31
✗	✓	80.43	75.11	71.69
✓	✓	**83.14**	**78.50**	**79.23**

C.3 Hyper-Parameter Study

We investigate the sensitivity of several hyperparameters in our method, including the invariance regularization strength λ, the variance regularization strength β, the divergence parameter ε, and the invariant subgraph ratio τ. The hyperparameter tuning results across the non-IID settings are presented in Fig.5, with the following key observations:(1) The value of τ is primarily influenced by cross-domain distribution shifts, rather than downstream tasks, with the optimal range identified between [0.2, 0.3]. (2) The value of λ is notably task-dependent. For clustering tasks, the optimal range lies within [0.2, 0.3], while for classification tasks, it is within [0.1, 0.2]. (3) The optimal values for β and ε are found within the ranges of [0.2, 0.3] and [0.1, 0.2], respectively. These results provide crucial theoretical insights into hyperparameter optimization and serve as a strong foundation for adapting our model to heterogeneous data distributions in real-world applications.

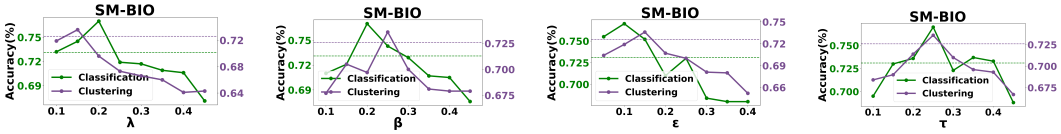


Figure 5: Hyperparameter Sensitivity analysis. The x-axis represents the four hyperparameters: λ, β, ε, and τ, with the left and right y-axis representing the classification and clustering accuracy, respectively. The dashed lines in the figure represent the highest test accuracy of the baseline method under the settings of the SM-BIO.

C.4 Convergence Analysis

Fig. 6 provides additional information on the relationship between graph classification loss and communication rounds in three non-IID settings. The experimental results indicate that, as the communication rounds progress, each client's loss curve exhibits a smooth, consistent decline, substantiating the effectiveness of our method in promoting model convergence. Further convergence experiments under the non-IID setting reinforce the superiority of the proposed bi-gradient regularization strategy in cross-dataset and cross-domain scenarios, demonstrating enhanced training stability and stronger generalization capability.

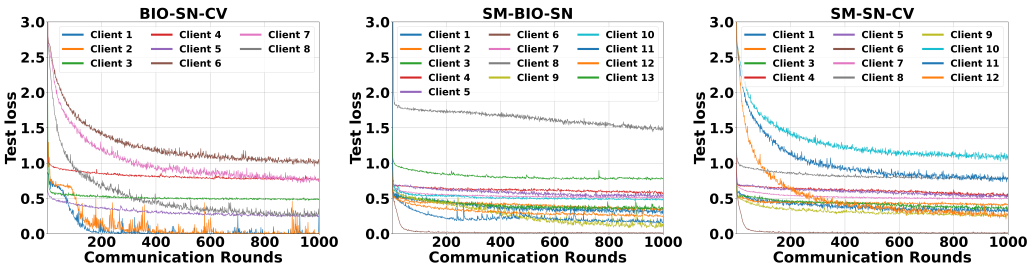


Figure 6: Loss trends of individual clients under three dataset settings: BIO-SN-CV, SM-BIO-SN, and SM-SN-CV.