

LIGHT: LLM-GUIDED GRAPH EXPERT ROUTING FOR SEMI-SUPERVISED DOMAIN GENERALIZATION

Anonymous authors

Paper under double-blind review

ABSTRACT

Although graph neural networks (GNNs) have shown remarkable performance in graph machine learning, their effectiveness in practice often suffers from realistic challenges including distribution shifts and label scarcity. Towards this end, this paper studies the problem of semi-supervised domain generalization, which aims to improve the performance of GNNs on unseen graphs using both labeled and unlabeled data. We propose a novel approach named LLM-Guided Graph Expert Routing (LIGHT) for semi-supervised domain generalization. The core idea of LIGHT is to distill the knowledge from LLM-as-a-judge to determine context-aware routing weights for a multi-hop graph mixture-of-experts framework. In particular, our LIGHT employs diverse graph experts that explore neighborhood information at varying depths. More importantly, we leverage LLMs to provide judgments of the most reliable graph experts are for crucial nodes, which provide context-aware routing guidance with high generalizability for knowledge distillation. To further address label scarcity, we introduce an expert-aware dynamic pseudo-labeling strategies that selects reliable nodes for additional training. Extensive experiments on various benchmark datasets validate the effectiveness of the proposed LIGHT in comparison with competing approaches. Our source code can be found at <https://anonymous.4open.science/r/LIGHT-A817>.

1 INTRODUCTION

Graphs serve as a fundamental data structure for capturing intricate relational and structural dependencies. With their proven effectiveness and computational efficiency in graph-based machine learning, Graph Neural Networks (GNNs) have become a *de facto* approach across a variety of tasks. One fundamental task is node classification, which endeavors to predict the label of each node within a graph based on its features and relational context, has been widely applied in various domains, i.e., social network analysis Bhagat et al. (2011), biochemistry Yue et al. (2020) and recommender systems Wu et al. (2022b). However, GNNs inherently rely on the assumption that training and testing data are independently and identically distributed (i.i.d.) Pan & Yang (2010), making them susceptible to the out-of-distribution (OOD) problem that induces significant performance drops under *distribution shifts*.

To prevent the generalization failure of GNNs, recent efforts have focused on the invariance principle from causality and invariant learning has emerged as a promising strategy for generalizable graph learning. Actually, GNNs’ OOD generalization (domain generalization) is impaired by spurious associations between non-essential graph patterns and target labels Chen et al. (2024b); Ju et al. (2024b). The rationale of invariant graph learning is to identify the invariant subgraph that consistently contributes to the prediction under distribution shift Li et al. (2022b); Wu et al. (2022a); Liu et al. (2023a). Despite the effectiveness, these methods often assume fully labeled source domains, whereas in real-world scenarios, the source domain typically contains only a small number of labeled samples alongside a large amount of unlabeled data Qiao et al. (2023b); Dan et al. (2024a;b).

In this work, we focus on a more realistic problem of semi-supervised domain generalization on graphs. Nonetheless, straightly adapting the above methods remains non-trivial, primarily arising from two key challenges: **First**, *The availability of diverse training environments under label scarcity setting*. Effective disentanglement of invariant components from spurious correlations re-

quires exposure to a variety of training environments [Xia et al. \(2023\)](#); [Li et al. \(2023\)](#); [Wang et al. \(2024b\)](#). However, in real-world settings, unseen domains often involve a broad spectrum of latent factors, whose quantity and nature are difficult to specify and replicate. This issue is further exacerbated when labeled data are scarce, making it difficult to reliably simulate the necessary diversity for invariant learning. **Second**, *The heterogeneity of graph data could degrade generalization ability*. Contrary to regular Euclidean data, graphs inherently possess diverse and intricate structures, with each node connected to different neighbors, resulting in highly divergent neighborhood structures and node degrees. The message-passing mechanism in GNNs tends to emphasize signals from nodes or subgraphs with high degrees and strong modularity [Liu et al. \(2023b\)](#); [Sun et al. \(2024\)](#). This often induces a simplicity bias, where the model focuses on prominent yet potentially spurious patterns, at the expense of capturing the true invariant substructures critical for robust prediction.

Fortunately, the remarkable generalization capabilities of Large Language Models (LLMs) provide new avenues for addressing these challenges. By harnessing their extensive pre-trained knowledge and semantic reasoning abilities, LLMs can offer valuable guidance to complement the learning process of GNNs. On the one hand, some prior works have utilized LLMs as enhancers to enrich node or graph representations [Chien et al. \(2022\)](#); [Jin et al. \(2023\)](#); [He et al. \(2024\)](#); [Liu et al. \(2024\)](#); [Li et al. \(2024a\)](#). On the other hand, some approaches have attempted to directly employ LLMs as predictors [Zhao et al. \(2023\)](#); [Fatemi et al. \(2024\)](#); [Tang et al. \(2024\)](#); [Perozzi et al. \(2024\)](#); [Chen et al. \(2024a\)](#); [Wang et al. \(2024a\)](#). However, LLMs are fundamentally designed to generate discrete textual outputs in a decoder-only manner, which limits their ability to accurately represent complex graph semantics [Wei et al. \(2022\)](#); [Jiang & Luo \(2025\)](#). Meanwhile, directly generating predictions from LLMs often leads to instability and unrelated outputs required for graph-based tasks.

Towards the end, in this paper, we propose a novel approach named **LLM-guided Graph Mixture-of-expert Routing (LIGHT)** for semi-supervised domain generalization on graphs, which aims to utilize LLMs as a context-aware routing to select appropriate invariant subgraph experts under the label scarcity setting. Specifically, given the input graph, we first transform the context of the node from different hops (i.e., 0-hop, 1-hop and 2-hop) into prompts, which are fed into pre-trained LLMs. Then, instead of directly predicting node labels, our LIGHT employs a graph mixture-of-expert (MoE) framework [Wang et al. \(2023\)](#), where each expert is dedicated to handling distinct data distributions, thereby facilitating better alignment with invariant correlations. Based on this, we generate predictions from each graph expert and LLMs are employed to dynamically select the most contextually aligned output. Finally, pseudo-labels are assigned to the remaining unlabeled nodes, and a flexible confidence threshold informed by expert outputs is utilized to identify high-confidence pseudo-labeled nodes. Therefore, benefiting from the zero-shot capabilities of LLMs, the framework could generalize effectively to unseen domains under the semi-supervised scenario.

The contribution of our LIGHT can be summarized as follows:

- ❶ *New Perspective*. We study an underexplored yet practical problem of semi-supervised domain generalization on graphs and pioneer a path of leveraging zero-shot capabilities of LLMs to address this problem.
- ❷ *Novel Methodology*. We propose a novel LLM-Guided MoE framework LIGHT, which transforms the multi-hop ego-subgraphs of each node into LLM prompts to guide the selection of GNN expert as the invariant predictor, and introduces a flexible expert-enhanced pseudo-labeling strategy to supervise the training of unlabeled nodes.
- ❸ *Extensive Experiments*. We conduct extensive experiments on multiple public datasets to evaluate the effectiveness of the proposed framework. The results consistently demonstrate the superiority of LIGHT in semi-supervised graph domain generalization.

2 RELATED WORK

Graph Domain Generalization. Graph Domain Generalization (GDG) has emerged as a critical research area to address the prevalent issue of performance degradation when GNNs encounter OOD data [Li et al. \(2022a\)](#); [Mai et al. \(2025\)](#). Early efforts focused on self-supervised pretraining (e.g. Pretraining-GNN), disentanglement (e.g. DisenGCN [Ma et al. \(2019a\)](#)), and adversarial training (e.g. GraphAT) [Zhang et al. \(2025\)](#). Recently, causal inference with representative methods including CIGA [Chen et al. \(2022\)](#) and StableGNN [Polina Andreeva & Bochenina \(2022\)](#). Data augmentation techniques have matured and label semantic preservation (e.g. LiSA [Yu et al. \(2023\)](#)). Complex multicomponent frameworks, such as MLDGG [Tian et al. \(2025\)](#) which integrates meta-learning

and causal reasoning, and TRACI Zhao et al. (2025a) which employs topological adversarial learning and prototypical mixup, signify a shift towards holistic solutions. However, these methods often assume fully labeled source domains, which limits their applicability in real-world scenarios where label annotations are scarce or expensive. In our paper, we consider the graph domain generalization task under the semi-supervised scenario.

Semi-supervised Learning on Graphs. Semi-Supervised Learning on Graphs (GSSL) has emerged as a crucial paradigm for learning from graph-structured data where labeled instances are scarce but unlabeled data and structural information are abundant Zhao et al. (2025b). Traditional GSSL includes Label Propagation Algorithms Kipf & Welling (2017); Ju et al. (2024a) and graph regularization techniques that enforce smoothness in predictions, while current emphasizes enhancing robustness and scalability through regularization Wu et al. (2025); Parveen et al. (2024), pseudo-labeling, and sophisticated Graph Contrastive Learning coupled with advanced data augmentation strategies Du et al. (2025); Wang et al. (2025); Ma et al. (2019b). Build on various kind on GNNs architecture, advanced SSL paradigms emerged, self-training strategies like A3-GCN Abdolali et al. (2025) and graph contrastive learning method like DGI Zhu et al. (2020), GRACE Hu et al. (2024) and AFECL Li et al. (2024b) for unsupervised representation learning. However, how to effectively perform semi-supervised learning under distribution shift remains underexplored.

3 PRELIMINARIES

Notations. Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ denote a graph, where $\mathcal{V}$ is the node set with N nodes and $\mathcal{E}$ is the edge set. The adjacency matrix is denoted as $\mathbf{A} \in \{0, 1\}^{N \times N}$, where $\mathbf{A}_{uv} = 1$ if there is an edge $(u, v) \in \mathcal{E}$, otherwise $\mathbf{A}_{uv} = 0$. Besides, the node feature matrix can be denoted as $\mathbf{X} \in \mathbb{R}^{N \times d}$, where d is the feature dimension. From the data-generating perspective, the input graph can be decomposed into a set of overlapping ego-graphs Wu et al. (2022a; 2024b) and the k -hop ego-graph of node v can be defined as $\mathcal{G}_v^{(k)} = \{\mathbf{X}_v^{(k)}, \mathbf{A}_v^{(k)}\}$, where feature matrix $\mathbf{X}_v^{(k)}$ and adjacency matrix $\mathbf{A}_v^{(k)}$ is induced by the node v as well as its k -hop neighborhood respectively.

Semi-supervised Domain Generalization on Graphs. Given the graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, we denote $\mathcal{V}^s \subset \mathcal{V}$ as the labeled node set, where each node $v \in \mathcal{V}^s$ is assigned a label $\mathbf{Y}_v \in \{0, 1\}^C$, with C being the number of classes, and $\mathbf{Y}_{v,c} = 1$ if node v belong to c -th class, otherwise $\mathbf{Y}_{v,c} = 0$. The rest $\mathcal{V}^u = \mathcal{V} \setminus \mathcal{V}^s$ is an unlabeled node set. The distribution shifts on the graph indicate that the distribution of ego-subgraphs and their labels differs between the training and testing data, namely $p_{tr}(\mathcal{G}_v, y_v) \neq p_{te}(\mathcal{G}_v, y_v)$. The objective is to learn an optimal predictor $f^*(\cdot)$ given the source graph $\mathcal{G}$ with limited labels such that it performs well on the unseen target graph.

$$f^* = \arg \min_f \mathbb{E}_{p_{tr}}[\mathcal{L}(f(\mathcal{G}_v), \mathbf{Y}_v)], \quad (1)$$

where $\mathcal{L}(\cdot)$ is the loss function. The learned model is then utilized to infer the node label for the target graph. Note that here the source and target graphs share the same label space.

4 METHODOLOGY

4.1 FRAMEWORK OVERVIEW

We introduce an LLM-guided MoE framework for semi-supervised domain generalization on graphs. The high-level idea of our LIGHT is to leverage the zero-shot capability of LLMs to dynamically select appropriate GNN experts based on multi-hop ego-subgraph prompts, thereby enhancing the generalization to unseen domains. In particular, we first decouple the node context into different hops and feed them into an MoE framework. Based on this, we extract the context of each node from different hops into prompts and integrate them into LLMs to adaptively route inputs to the most relevant experts with high generalizability for knowledge distillation. Finally, we utilize an expert-enhanced pseudo-labeling to alleviate the label scarcity. The overview of our LIGHT is illustrated in Figure 1. More details are introduced as follows.

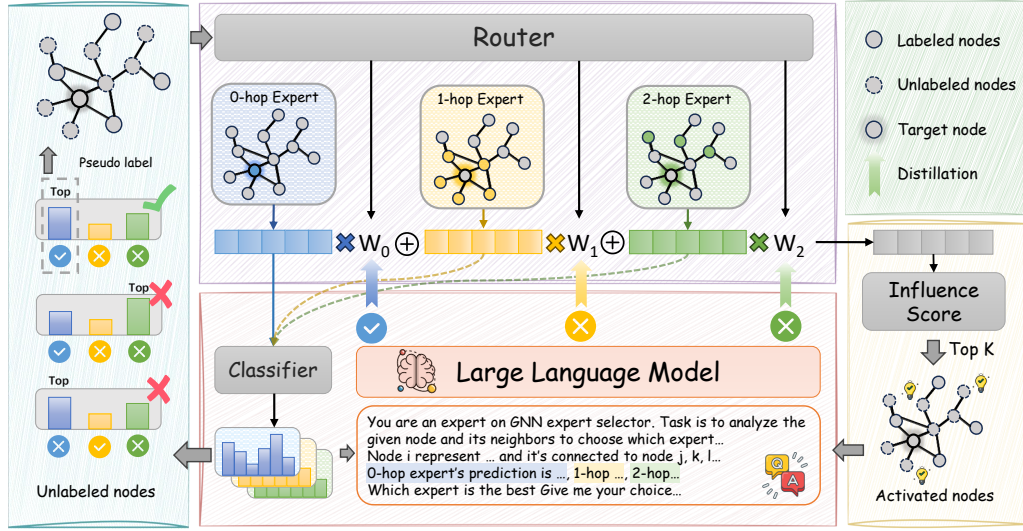


Figure 1: The overview framework of our proposed LIGHT. We decouple the node context into different hops and feed them into a mixture-of-expert framework. Then, we transform the context into prompts and input them into an LLM. The LLMs serve as a teacher router and we distill the knowledge from LLM-as-judge. Finally, we introduce expert-aware dynamic pseudo-labeling to select reliable nodes for additional training.

4.2 MULTI-HOP GRAPH MIXTURE-OF-EXPERT FOR DIVERSE NEIGHBORHOOD EXPLORATION

Traditional GNNs often suffer from oversmoothing or overfitting when aggregating multi-hop neighbors, particularly under distribution shifts. These issues can be attributed to the indiscriminate aggregation of neighborhood information across varying hops. This practice can lead to the generation of spurious correlations or redundant patterns both of which are sensitive to domain-specific biases. Theorem 4.1 demonstrates that the mixture-of-experts (MoE) model achieves a lower prediction error compared to the individual model.

Theorem 4.1. Suppose the true label for node v is 1. If only a single expert is available (i.e., $K = 1$) and this expert is not aligned with the domain-specific distribution of node v , then the prediction error satisfies the lower bound:

$$\mathbb{P}(\hat{Y}_v \neq 1) = \mathbb{P}(Q_{v,1} < 0) > 0.$$

In contrast, if K experts are available, the prediction error is upper bounded as:

$$\mathbb{P}(\hat{Y}_v \neq 1) \leq \frac{C_0}{K},$$

where C_0 is a constant that depends on the margin ν and the sub-Gaussian parameter σ^2 .

The proof of this bound is provided in Appendix B. Notably, different hops capture different levels of contextual semantics: lower hops (e.g., 1-hop) typically reflect local homophilic patterns, while higher hops encode global and heterophilic dependencies. Therefore, we decouple information aggregation by hop and generate a set of candidate predictions by designing a graph MoE framework Cai et al. (2024); Wang et al. (2023), where hop-specific experts generate candidate predictions that are adaptively integrated based on node context for robust performance under distribution shift.

For the k -hop ego-subgraph $\mathcal{G}_v^{(k)}$, we calculate positive pointwise mutual information (PPMI) Zhuang & Ma (2018) between nodes to fully explore the structural information. Specifically, we employ a random walk to sample a set of paths on $\mathcal{A}$, and the frequency matrix F between

nodes can be calculated as the co-occurrence on the paths. Then, the PPMI between nodes can be:

$$\begin{aligned}\tilde{P}_{ij} &= \frac{F_{ij}}{\sum_{i,j} F_{ij}}, \tilde{P}_i = \frac{\sum_j F_{ij}}{\sum_{i,j} F_{ij}}, \tilde{P}_j = \frac{\sum_i F_{ij}}{\sum_{i,j} F_{ij}}, \\ P_{ij} &= \max\{\log(\frac{\tilde{P}_{ij}}{\tilde{P}_i \times \tilde{P}_j}), 0\}.\end{aligned}\quad (2)$$

where $\tilde{P}_{ij}$ denotes the probability that node v_i within the contextual window of node v_j and P_{ij} is the PPMI between them. Therefore, a larger P_{ij} indicates a higher co-occurrence frequency between nodes than if they were independent. We treat $P_v^{(k)}$ as the new adjacency matrix for the k -hop ego-subgraph $\mathcal{G}_v^{(k)}$ and the representation of node v can be extracted as:

$$Z_v^{(k)} = \text{GNN}_k(\mathbf{X}_v^{(k)}, P_v^{(k)}), \quad (3)$$

where $\text{GNN}_k(\cdot)$ is the GNN with L layers. Note that in our work, we employ K experts sharing the same architecture but parameterized independently. We utilize a MoE routing mechanism to assign expert weights, followed by a classifier that produces the final prediction:

$$\hat{Y}_v = f\left(\sum_{k=1}^K \omega_v^{(k)} Z_v^{(k)}\right), \quad (4)$$

where $f(\cdot)$ is a learnable classifier and ω_v denote the weight of the different expert.

4.3 NEIGHBORHOOD-AWARE PROMPT ENGINEERING

Since the distribution shifts on the graph entail a broad range of structural variations, directly learning a routing function for ω_v as in traditional MoE frameworks becomes unreliable. We leverage the zero-shot capability of LLMs as a judge to enable more informed and adaptive expert selection. In particular, for each node v , we first extract K ego-subgraphs corresponding to different hops, effectively capturing multi-hop structural contexts. Then, we take each ego-graph $\mathcal{G}_v$ as input of the expert and provide the context to describe the expert ($g : \mathcal{G}_v \rightarrow \mathbf{W}$) and the goal of tasks ($q : \mathbf{W} \rightarrow \mathbf{W}$), where $\mathbf{W}$ is the sequence of discrete tokens for the LLMs.

Expert Context Provision. Given a node v , the ego-subgraph for each expert and its context can be presented as: *You are an expert on GNN experts selector, given GNN experts: $\langle \text{expert} \rangle \dots \backslash n$, with node v and its neighborhood $\{\text{graph}\}$ represent node connection relationships... $\backslash n$* , where $\langle \text{expert} \rangle$ is the placeholder for expert inputs and $\{\text{graph}\}$ is the ego-subgraph information.

Note that the k -th expert input is the description w.r.t. k -hop neighborhood, i.e., the 1-hop expert can be *1-hop: direct connected neighbors providing classification signals*. Ego-subgraph information can be described in an incident manner, i.e., *Node v represents..., and it is connected to node $u \dots \backslash n$* .

Task Description. To guide the expert selection process, we design a task-specific prompt where the model is asked to choose the most appropriate expert for node classification based on the expert context. Specifically, the prompt includes not only the task description but also a set of alternative expert choices, i.e., *Which expert does node v classification used? Please give out your choice on the following experts: $\{\text{ans}\}$* , where $\{\text{ans}\}$ represent the set of experts, namely $\{1, 2 \dots, K\}$.

Based on this, the context information of nodes can be formulated as prompts $(g(\mathcal{G}_v), q(\mathbf{Q}), \mathcal{R}_v)$, where $\mathbf{Q}$ denote the question text for the task. Note that we also incorporate the individual predictions $\mathcal{R}_v$ from each expert as the context (see Section 4.2).

4.4 LLM-GUIDED CONTEXT-AWARE ROUTING WITH KNOWLEDGE DISTILLATION

Instead of relying on LLMs to directly generate predictions, which often results in unreliable outputs or incorrect formats, we propose to leverage LLMs as a judge instead of a predictor and employ knowledge distillation to transfer the LLM-guided routing behavior to a lightweight student model.

LLM Judge for Context-aware Routing. We generate the predictions from different experts and incorporate them into the LLM judge for contextual-aware routing, which can be formulated as:

$$\mathcal{R}_v = \{\mathbf{R}_v^{(k)}\}_{k=1}^K, \mathbf{R}_v^{(k)} = f\left(\sum_{k'=1}^K e_v^{(k)}(k') Z_v^{(k')}\right), \quad (5)$$

where $e_v^{(k)}$ is initialized as a one-hot vector with the k -th entry set to 1. Then, the resulting predictions are subsequently incorporated into the LLMs, which selects the most plausible expert k^* :

$$k^* = \text{LLM}(g(\mathcal{G}_v), q(\mathbf{Q}), \mathcal{R}_v). \quad (6)$$

We adopt a label smoothing strategy Müller et al. (2019) within our MoE framework by assigning lower weights to non-selected experts. Accordingly, the weight vector $e_v^{(k)}$ can be updated as:

$$e_v^{(k)} = \begin{cases} \alpha, & \text{if } k = k^*, \\ \frac{1-\alpha}{K-1}, & \text{if } k \neq k^*, \end{cases} \quad (7)$$

where $\alpha \in (0, 1)$ is the smoothing coefficient.

Knowledge Distillation with Influence Maximization. Note that performing LLM inference for each node is computationally inefficient, we distill a lightweight routing function from the LLM to enable efficient expert selection. Inspired by the influence maximization (IM) problem in the social network Kempe et al. (2003), we aim to select a set of anchor nodes $\mathcal{S}$ that maximizes the utility of LLMs in a cost-effective manner, and the formation is:

$$\max_{\mathcal{S}} |\sigma(\mathcal{S})| \text{ s.t. } \mathcal{S} \subseteq \mathcal{V}, |\mathcal{S}| = B, \quad (8)$$

where $\sigma(\mathcal{S})$ denotes the set of nodes that can be activated by $\mathcal{S}$, defined as:

$$\sigma(\mathcal{S}) = \bigcup_{v \in \mathcal{V}, I_v(\mathcal{S}) > \theta} \{v\}, \quad I_v(\mathcal{S}) = \max_{u \in \mathcal{S}} I_v(u), \quad (9)$$

where $I_v(\mathcal{S})$ denotes the maximum influence from the anchor nodes, θ is the activation threshold. In practice, we introduce $\mathbf{P}_{u,v}$ and the influence quality $I_v(u)$ from anchor node u is measured as:

$$I_v(u) = \mathbf{P}_{u,v} \sum_{k=1}^K E_u^{(k)}, \quad (10)$$

where we use entropy $E_u^{(k)} = -\sum_{c=1}^C \mathbf{R}_{u,c}^{(k)} \log \mathbf{R}_{u,c}^{(k)}$ as the influence score. We employ a greedy node selection to obtain $\mathcal{S}$ and the details can be seen in Appendix 1. Based on $\mathcal{S}$, we distill knowledge from the LLMs to obtain a lightweight routing function, and the teaching loss can be:

$$\mathcal{L}_{dl} = -\sum_{u \in \mathcal{S}} \text{CE}(e_u^{(k^*)}, h(\mathbf{Z}_u)), \quad \mathbf{Z}_u = \mathbf{Z}_u^{(1)} \parallel \dots \parallel \mathbf{Z}_u^{(K)}, \quad (11)$$

where $h(\cdot)$ denotes the routing function and $\text{CE}(\cdot)$ is the cross-entropy loss. Note that to effectively address domain shifts, we perform knowledge re-distillation from the LLMs for each target domain prior to inference, and the MoE framework in Equation 3 can be updated into:

$$\hat{\mathbf{Y}}_v = f\left(\sum_{k=1}^K h(\mathbf{Z}_v) \mathbf{Z}_v^{(k)}\right), \quad (12)$$

4.5 EXPERT-AWARE DYNAMIC PSEUDO-LABELING

For each unlabeled node v , we generate the sharpening pseudo-label $\tilde{\mathbf{Y}}_v$ from the MoE framework, which is then used for the node classification. In particular, we use the θ -percentile of the entropy value $u_v^{(k)}$ from each expert to establish a dynamic threshold τ , which can be defined as:

$$\tau = \text{Percentile}_{\theta}\left(\{u_v^{(k)}\}_{k=1}^K\right). \quad (13)$$

Given this dynamic threshold, we filter out noisy pseudo labels and leave only the high-quality samples from the unlabeled data. In formula:

$$\mathcal{V}_{\text{selected}} = \{(v, u_v^{(k^*)}) | u_v^{(k^*)} \leq \tau\}. \quad (14)$$

We leverage the high-quality samples to fine-tune the model, thereby improving its performance and promoting better generalisation under domain shifts. The unsupervised classification loss can be:

$$\mathcal{L}_u = \sum_{v \in \mathcal{V}_{\text{selected}}} \text{CE}\left(\tilde{\mathbf{Y}}_v, f\left(\sum_{k=1}^K h(\mathbf{Z}_v) \mathbf{Z}_v^{(k)}\right)\right). \quad (15)$$

Table 1: The model performance comparison on six domain generalization tasks with source label rate as 5%. A: ACMv9; C:Citationv1; D: DBLPv7. $A \Rightarrow C$ represents that A is the source graph and C is the target graph. The same applies to other tasks.

Methods	$A \Rightarrow C$		$A \Rightarrow D$		$C \Rightarrow A$		$C \Rightarrow D$		$D \Rightarrow A$		$D \Rightarrow C$	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
MLP	41.3 $\pm$ 1.15	35.8 $\pm$ 0.72	42.8 $\pm$ 0.88	36.3 $\pm$ 0.77	39.4 $\pm$ 0.57	33.7 $\pm$ 0.58	43.7 $\pm$ 0.69	36.7 $\pm$ 0.55	37.3 $\pm$ 0.32	30.8 $\pm$ 0.37	39.4 $\pm$ 0.99	32.8 $\pm$ 0.99
GCN	54.4 $\pm$ 1.52	52.0 $\pm$ 1.62	56.9 $\pm$ 2.33	53.4 $\pm$ 2.81	54.1 $\pm$ 1.40	52.3 $\pm$ 1.98	58.9 $\pm$ 0.99	54.5 $\pm$ 1.55	50.1 $\pm$ 2.14	48.0 $\pm$ 3.28	56.0 $\pm$ 1.24	51.9 $\pm$ 1.49
GSAGE	49.3 $\pm$ 2.18	46.4 $\pm$ 2.06	51.8 $\pm$ 1.35	47.4 $\pm$ 1.62	46.8 $\pm$ 2.56	45.0 $\pm$ 2.78	51.7 $\pm$ 1.95	48.1 $\pm$ 1.97	41.7 $\pm$ 2.17	37.4 $\pm$ 4.59	45.4 $\pm$ 2.11	39.3 $\pm$ 3.45
GAT	55.1 $\pm$ 3.22	50.8 $\pm$ 1.45	55.3 $\pm$ 2.52	51.8 $\pm$ 2.60	50.0 $\pm$ 1.20	45.6 $\pm$ 2.36	55.4 $\pm$ 2.73	49.2 $\pm$ 2.59	44.8 $\pm$ 2.74	38.3 $\pm$ 4.84	50.4 $\pm$ 3.35	42.0 $\pm$ 4.46
GIN	64.6 $\pm$ 2.47	56.0 $\pm$ 2.73	60.0 $\pm$ 2.09	51.3 $\pm$ 3.99	57.1 $\pm$ 1.19	54.4 $\pm$ 2.57	62.0 $\pm$ 1.05	56.8 $\pm$ 1.40	51.9 $\pm$ 2.00	45.4 $\pm$ 2.16	60.2 $\pm$ 3.05	53.0 $\pm$ 2.10
CaNet	62.3 $\pm$ 3.11	54.9 $\pm$ 2.98	61.2 $\pm$ 3.77	51.0 $\pm$ 2.00	58.6 $\pm$ 1.01	51.3 $\pm$ 0.76	63.0 $\pm$ 2.31	57.5 $\pm$ 1.64	52.4 $\pm$ 5.09	43.8 $\pm$ 1.77	57.2 $\pm$ 4.01	48.3 $\pm$ 3.67
MARIO	65.5 $\pm$ 4.07	60.9 $\pm$ 4.31	66.3 $\pm$ 5.85	61.4 $\pm$ 4.05	57.0 $\pm$ 4.16	53.5 $\pm$ 5.10	63.9 $\pm$ 4.89	53.7 $\pm$ 5.56	51.5 $\pm$ 4.13	42.9 $\pm$ 3.06	63.9 $\pm$ 5.28	55.7 $\pm$ 4.57
LDAT	69.3 $\pm$ 0.99	70.3 $\pm$ 1.30	65.0 $\pm$ 2.24	67.0 $\pm$ 1.53	65.8 $\pm$ 0.89	63.1 $\pm$ 1.65	70.7 $\pm$ 1.57	68.4 $\pm$ 1.11	60.6 $\pm$ 2.81	60.3 $\pm$ 2.65	64.7 $\pm$ 2.00	62.3 $\pm$ 1.87
SGDA	71.8 $\pm$ 1.85	70.0 $\pm$ 1.83	63.3 $\pm$ 3.01	61.6 $\pm$ 2.15	66.2 $\pm$ 1.60	65.3 $\pm$ 1.74	68.5 $\pm$ 1.40	65.8 $\pm$ 1.29	56.9 $\pm$ 1.01	57.4 $\pm$ 0.73	62.3 $\pm$ 0.77	61.8 $\pm$ 0.85
Ours	75.7 $\pm$ 0.89	73.5 $\pm$ 0.91	70.7 $\pm$ 0.50	68.5 $\pm$ 0.57	70.6 $\pm$ 0.51	71.3 $\pm$ 0.50	73.3 $\pm$ 0.58	72.3 $\pm$ 0.53	64.5 $\pm$ 0.38	64.3 $\pm$ 0.50	73.6 $\pm$ 0.61	69.0 $\pm$ 1.71

4.6 SUMMARIZATION

To ensure each expert captures distinct aspects of the ego-subgraphs, we propose a diversity objective with distance correlation Székely et al. (2007); Wang et al. (2020) to encourage divergence:

$$\mathcal{L}_{ind} = \sum_{k=1}^K \sum_{k'=k+1}^K dCor(\mathbf{Z}^{(k)}, \mathbf{Z}^{(k')}), \quad (16)$$

where $\mathbf{Z}^{(k)}$ is the representation of all nodes and $dCor(\cdot)$ is the distance correlation function:

$$dCor(\mathbf{Z}^{(k)}, \mathbf{Z}^{(k')}) = \frac{dCov(\mathbf{Z}^{(k)}, \mathbf{Z}^{(k')})}{\sqrt{dVar(\mathbf{Z}^{(k)}) \cdot dVar(\mathbf{Z}^{(k')})}}, \quad (17)$$

where $dVar(\cdot)$ and $dCov(\cdot)$ is the distance variance and covariance respectively. Finally, we combine supervised classification loss $\mathcal{L}_s$ from labeled node, the knowledge distillation loss $\mathcal{L}_{dl}$ and the expert diversity regularization $\mathcal{L}_{ind}$. The overall objective of our LIGHT is:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_u + \lambda_{dl}\mathcal{L}_{dl} + \lambda_{ind}\mathcal{L}_{ind}, \quad (18)$$

where λ_{dl} and λ_{ind} are the balance factors that control the weights of the losses. The overall algorithm of our proposed LIGHT can be found in the Appendix B.

5 EXPERIMENT

5.1 SETUP

Datasets. We conduct extensive experiments from ArnetMiner Tang et al. (2008), i.e., ACMv9, Citationv1, and DBLPv7. These datasets are widely used for evaluating graph learning models, where each paper, represented as a node, is to be categorized into one of five predefined research areas, including Artificial Intelligence, Computer Vision, Database, Information Security, and Networking. The citation relationship is represented as an edge. These datasets provide diverse graph structures and data contributions, making them suitable to evaluate our method.

Compared Baselines. To comprehensively evaluate our proposed LIGHT, we compare it against a comprehensive suite of baseline methods including GCN Kipf & Welling (2017), GraphSAGE (GSAGE) Hamilton et al. (2017), GAT Vrahatis et al. (2024), GIN Xu et al. (2019), CaNet Wu et al. (2024a), MARIO Zhu et al. (2024), LDAT Wang et al. (2022), and SGDA Qiao et al. (2023a). Note that we adopt the same experimental settings with baselines to ensure a fair comparison.

Implementational Details. All experiments are conducted on a single NVIDIA RTX3090 GPU with 24GB memory. We utilize PyTorch and PyG to implement our framework. We utilize a random splitting strategy to ensure diversity. In the source domain, 5% of node labels are used for supervised training. The MoE network is composed of multiple experts and a router network. All experiments are conducted with five fixed random seeds for reproducibility. Our LIGHT is trained using the Adam optimizer with a learning rate of 1e-3 and L2 weight decay of 2e-3. We utilize Qwen2.5-7B as our default LLMs and attempt the Gemma3 series model in our experiments as well. We report the performance on the target domain nodes in terms of both Micro F1-score and Macro F1-score.

Table 2: Ablation studies on the Citations benchmark (MoE/LLM/DIV/HQ as peers).

Models	$A \Rightarrow C$		$A \Rightarrow D$		$C \Rightarrow A$		$C \Rightarrow D$		$D \Rightarrow A$		$D \Rightarrow C$	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
LIGHT w/o MoE	43.3	37.1	44.5	38.5	42.1	34.3	46.8	37.8	36.0	31.5	38.2	33.9
LIGHT w/o LLM	72.0	66.6	66.7	58.5	66.1	63.3	69.7	64.7	51.0	43.4	66.0	61.7
LIGHT w/o DIV	74.9	72.7	68.6	66.9	67.5	68.1	71.8	71.2	60.3	57.5	68.0	61.8
LIGHT w/o HQ	70.0	66.0	67.5	62.2	64.4	60.4	71.1	69.2	57.5	52.0	57.2	50.1
LIGHT	75.7	73.5	70.7	68.5	70.6	71.3	73.3	72.3	64.5	64.3	73.6	71.6

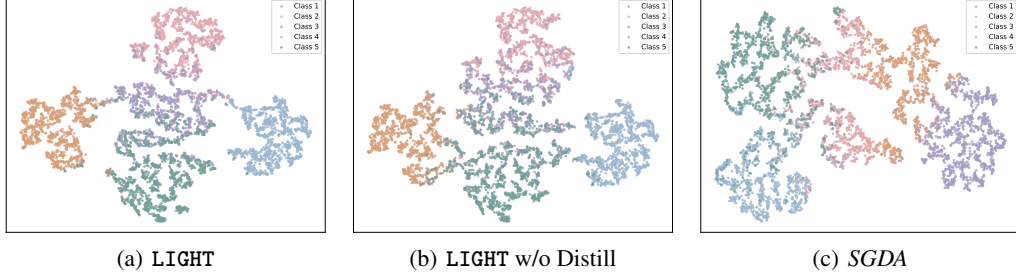


Figure 2: Analysis of different t-SNE visualizations on $A \Rightarrow C$ task. (a), (b) respectively represent the performance of our method before and after knowledge distillation, (c) represents the effect of the baseline method SGDA.

5.2 EMPIRICAL PERFORMANCE

Performance Comparison. The compared results can be found in Table 1, across all six domain generalization tasks involving transfers between ACMv9 (A), Citationv1 (C), and DBLPv7 (D) datasets, even with a scarce 5% source label rate. $\triangleright$ *Firstly*, our LIGHT consistently achieves state-of-the-art Micro-F1 and Macro-F1 scores. For instance, on the $A \Rightarrow C$ task, our method achieves 75.7% Micro-F1, surpassing the strongest specialized baseline SGDA (71.8%) by 3.9% and exhibiting larger improvements over standard GNNs like GIN. The performance increase can be attributed to two factors as follows: (I) Introduction of our LLM-based guidance, which provides the judgments of the most graph experts for strong generalizability with knowledge distillation. (II) Introduction of our mixture-of-experts framework, which provides adaptive neighborhood aggregation for topological exploration. $\triangleright$ *Secondly*, LLM guidance dynamically selects experts, especially for nodes with high uncertainty, providing an adaptability which is absent in baselines. This tailored expertise, combined with a semi-supervised learning component that leverages high-quality pseudo-labels identified through a novel process considering both expert agreement and LLM-derived confidence, facilitates effective knowledge distillation and transfer from limited labeled data. The consistent high performance across all tasks, which involve datasets with varied structural properties and data distributions, underscores the robustness and generalizability of LIGHT.

Ablation Study. To validate the individual contributions of key components in our LIGHT, we introduce the following model variants: (I) LIGHT w/o MoE, which removes the MoE framework and utilizes one single expert; (II) LIGHT w/o LLM, which removes the knowledge distillation from LLMs; (III) LIGHT w/o DIV, which removes the divergence-prompting objective in our framework; (IV) LIGHT w/o HQ, which removes our expert-aware dynamic pseudo-labeling. Results are shown in Table 2. We can observe that removing the MoE framework led to the most substantial performance decline, underscoring its fundamental role. Disabling LLM-guided expert selection caused Micro-F1 scores decreasing by about 7.3% and Macro-F1 by 12.9%, highlighting the LLM’s contribution to refining expert routing. The exclusion of the high-quality pseudo-labeling strategy resulted in a considerable reduction in performance, averaging a 7.0% decrease in Micro-F1 and an 11.6% decrease in Macro-F1, emphasizing its importance in leveraging unlabeled data. Finally, removing the MoE’s diversity/coherence loss led to Micro-F1 scores dropping by an average of 2.9% and Macro-F1 by 4.0%, indicating its value in refining expert representations. These findings collectively affirm the crucial role each component plays in the proposed method’s overall effectiveness.

Case Study. To further investigate the qualitative impact of our proposed method, particularly the LLM-guided expert selection distillation, we present a case study visualizing the learned node embeddings from a representative domain generalization task (e.g., $A \Rightarrow C$). The t-SNE [van der Maaten](#)

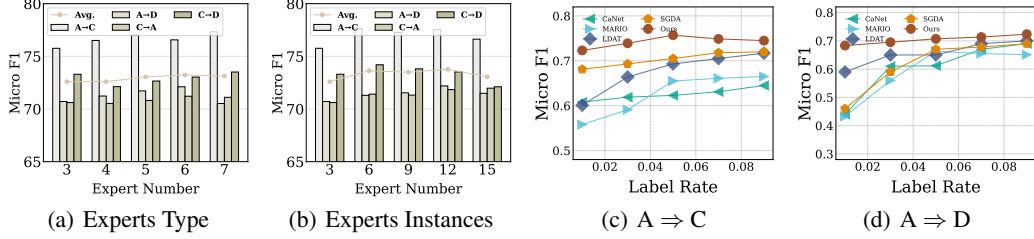


Figure 3: (a) shows the Micro F1 score for varying expert types, and (b) illustrates results when expert instances increase proportionally. Sub-figures (c) and (d) compare performance under different label rates on $A \Rightarrow C$ and $A \Rightarrow D$ tasks respectively.

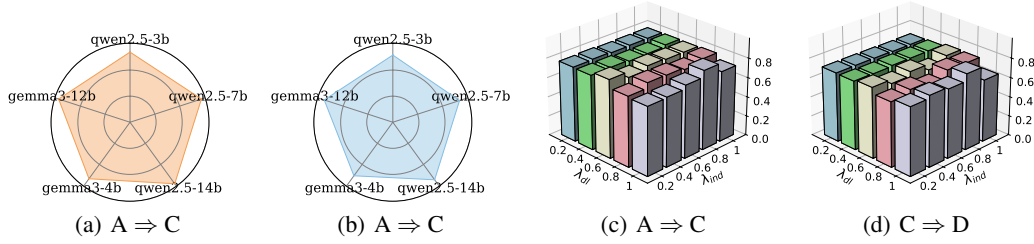


Figure 4: Effect analysis of LLMs type and loss weight effect impact analyze on micro-f1 score.

& Hinton (2008) results in Figure 2 illustrates these visualizations for three distinct model configurations: (a) our full proposed method including LLM distillation (ours w/ distill), (b) an ablation of our method without the LLM distillation component (ours w/o distill), and (c) the SGDA baseline, a relevant graph domain adaptation method. Node colors correspond to their ground-truth classes. By providing high-level, context-aware guidance to the MoE’s gating mechanism, the LLM helps the model assign nodes to the most appropriate specialized experts, leading to embeddings that better align with the true class distinctions. The noticeable degradation in cluster quality when distillation is removed corroborates this, further emphasizing the value added by the LLM component.

Parameter Analysis. In this part, we conduct parameter analysis. The number of experts influences performance in Figure 3 (a), where increasing the diversity of expert types or the number of instances per expert type in Figure 3 (b) can yield improvements, though gains diminish beyond an optimal point. Regarding data scarcity, while higher label rates generally lead to better outcomes, the proposed method demonstrates strong performance even with low label rates, outperforming alternatives in semi-supervised domain generalization. The results in Figure 4 (c) and (d) shows the choice of LLM for guidance affects outcomes, with larger or more capable LLMs generally offering better guidance, although the performance gap between mid-sized (e.g., 7b) and larger (e.g., 12-14b) models was not substantial, while smaller models (e.g., 3-4b) showed a noticeable performance decrease. Further analysis is provided in the Appendix F. Finally, the model’s performance is sensitive to the loss weight hyperparameters (λ_{dl} , λ_{ind}), which control the balance between unsupervised classification, knowledge distillation, and expert diversity regularization losses; their optimal values vary across some tasks, tuning as illustrated by the results in Figure 3 (c) and (d).

6 CONCLUSION

In this work, we have introduced a novel framework **LIGHT** for semi-supervised domain generalization on graphs. Our **LIGHT** leverages the zero-shot ability of LLMs to steer a MoE architecture, where each expert specializes in different graph topological neighborhoods. This allows for dynamic, context-aware selection of the most pertinent expert, guided by multi-hop ego-subgraph prompts transformed for LLM comprehension. Furthermore, we have incorporated an expert-enhanced pseudo-labeling strategy, which leverages high-confidence pseudo-labels to effectively supervise the model’s training in label-scarce environments. Extensive empirical evaluations on multiple public benchmark datasets demonstrate the superiority of our method in generalizing across different graph domains, significantly outperforming existing state-of-the-art methods.

REFERENCES

- Maryam Abdolali, Romina Zakerian, Behnam Roshanfekr, Fardin Ayar, and Mohammad Rahmati. Adapt, agree, aggregate: Semi-supervised ensemble labeling for graph convolutional networks, 2025. URL <https://arxiv.org/abs/2503.17842>.
- Smriti Bhagat, Graham Cormode, and S Muthukrishnan. Node classification in social networks. *Social network data analytics*, pp. 115–148, 2011.
- Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts. *arXiv preprint arXiv:2407.06204*, 2024.
- Runjin Chen, Tong Zhao, Ajay Jaiswal, Neil Shah, and Zhangyang Wang. Llaga: Large language and graph assistant. In *Proceedings of the International Conference on Machine Learning*, pp. 7809–7823, 2024a.
- Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, Kaili Ma, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. In *Advances in Neural Information Processing Systems*, 2022.
- Zhengyu Chen, Teng Xiao, Kun Kuang, Zheqi Lv, Min Zhang, Jinluan Yang, Chengqiang Lu, Hongxia Yang, and Fei Wu. Learning to reweight for generalizable graph neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8320–8328, 2024b.
- Eli Chien, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, Jiong Zhang, Olgica Milenkovic, and Inderjit S Dhillon. Node feature extraction by self-supervised multi-scale neighborhood prediction. In *Proceedings of the International Conference on Learning Representations*, 2022.
- Jun Dan, Weiming Liu, Mushui Liu, Chunfeng Xie, Shunjie Dong, Guofang Ma, Yanchao Tan, and Jiazheng Xing. Hogda: Boosting semi-supervised graph domain adaptation via high-order structure-guided adaptive feature alignment. pp. 11109–11118, 2024a.
- Jun Dan, Weiming Liu, Xie Xie, Hua Yu, Shunjie Dong, and Yanchao Tan. Tfgda: Exploring topology and feature alignment in semi-supervised graph domain adaptation through robust clustering. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 50230–50255, 2024b.
- Ally Yalei Du, Eric Huang, and Dravyansh Sharma. Tuning algorithmic and architectural hyperparameters in graph-based semi-supervised learning with provable guarantees, 2025. URL <https://arxiv.org/abs/2502.12937>.
- Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. Talk like a graph: Encoding graphs for large language models. In *Proceedings of the International Conference on Learning Representations*, 2024.
- William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, pp. 1025–1035, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.
- Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. Harnessing explanations: Llm-to-lm interpreter for enhanced text-attributed graph representation learning. In *Proceedings of the International Conference on Learning Representations*, 2024.
- Yulan Hu, Sheng Ouyang, Jingyu Liu, Ge Chen, Zhirui Yang, Junchen Wan, Fuzheng Zhang, Zhongyuan Wang, and Yong Liu. Graph ranking contrastive learning: A extremely simple yet efficient method, 2024. URL <https://arxiv.org/abs/2310.14525>.
- Dapeng Jiang and Xiao Luo. Marrying llms with dynamic forecasting: A graph mixture-of-expert perspective. pp. 396–410, 2025.
- Bowen Jin, Yu Zhang, Yu Meng, and Jiawei Han. Edgeformers: Graph-empowered transformers for representation learning on textual-edge networks. In *Proceedings of the International Conference on Learning Representations*, 2023.

- Wei Ju, Siyu Yi, Yifan Wang, Qingqing Long, Junyu Luo, Zhiping Xiao, and Ming Zhang. A survey of data-efficient graph learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI '24*, 2024a. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/896. URL <https://doi.org/10.24963/ijcai.2024/896>.
- Wei Ju, Siyu Yi, Yifan Wang, Zhiping Xiao, Zhengyang Mao, Hourun Li, Yiyang Gu, Yifang Qin, Nan Yin, Senzhang Wang, et al. A survey of graph neural networks in real world: Imbalance, noise, privacy and ood challenges. *arXiv preprint arXiv:2403.04468*, 2024b.
- David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the International ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 137–146, 2003.
- Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL <https://openreview.net/forum?id=SJU4ayYgl>.
- Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. Ood-gnn: Out-of-distribution generalized graph neural network. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–14, 01 2022a. doi: 10.1109/TKDE.2022.3193725.
- Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. Out-of-distribution generalization on graphs: A survey. *arXiv preprint arXiv:2202.07987*, 2022b.
- Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. Invariant node representation learning under distribution shifts with multiple latent environments. *ACM Transactions on Information Systems*, 42(1):1–30, 2023.
- Yuhan Li, Peisong Wang, Zhixun Li, Jeffrey Xu Yu, and Jia Li. Zerog: Investigating cross-dataset zero-shot transferability in graphs. In *Proceedings of the International ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1725–1735, 2024a.
- Yujun Li, Hongyuan Zhang, and Yuan Yuan. Edge contrastive learning: An augmentation-free graph contrastive learning model, 2024b. URL <https://arxiv.org/abs/2412.11075>.
- Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks. In *Proceedings of the International Conference on Learning Representations*, 2024.
- Yang Liu, Xiang Ao, Fuli Feng, Yunshan Ma, Kuan Li, Tat-Seng Chua, and Qing He. Flood: A flexible invariant learning framework for out-of-distribution generalization on graphs. In *Proceedings of the International ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1548–1558, 2023a.
- Zemin Liu, Trung-Kien Nguyen, and Yuan Fang. On generalized degree fairness in graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4525–4533, 2023b.
- Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. Disentangled graph convolutional networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4212–4221. PMLR, 09–15 Jun 2019a. URL <https://proceedings.mlr.press/v97/ma19a.html>.
- Jiaqi Ma, Weijing Tang, Ji Zhu, and Qiaozhu Mei. A flexible generative framework for graph-based semi-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3276–3285, 2019b.
- Junjie Mai, Chongzhi Gao, and Jun Bao. Domain generalization through data augmentation: A survey of methods, applications, and challenges. *Mathematics*, 13(5), 2025. ISSN 2227-7390. doi: 10.3390/math13050824. URL <https://www.mdpi.com/2227-7390/13/5/824>.

- Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 4694–4703, 2019.
- Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.
- Alisha Parveen, Tikam Dakal, Pankaj Yadav, and Abhishek Kumar. *Semi-Supervised Learning in Bioinformatics*. 01 2024. ISBN 9780128096338. doi: 10.1016/B978-0-323-95502-7.00097-X.
- Bryan Perozzi, Bahare Fatemi, Dustin Zelle, Anton Tsitsulin, Mehran Kazemi, Rami Al-Rfou, and Jonathan Halcrow. Let your graph do the talking: Encoding structured data for llms. *arXiv preprint arXiv:2402.05862*, 2024.
- Egor Shikov Polina Andreeva and Claudie Bochenina. Attributed labeled bter-based generative model for benchmarking of graph neural networks. In *Proceedings of the 17th International Workshop on Mining and Learning with Graphs (MLG)*, 2022.
- Ziyue Qiao, Xiao Luo, Meng Xiao, Hao Dong, Yuanchun Zhou, and Hui Xiong. Semi-supervised domain adaptation in graph transfer learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI '23*, 2023a. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/253. URL <https://doi.org/10.24963/ijcai.2023/253>.
- Ziyue Qiao, Xiao Luo, Meng Xiao, Hao Dong, Yuanchun Zhou, and Hui Xiong. Semi-supervised domain adaptation in graph transfer learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 2279–2287, 2023b.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- Xin Sun, Liang Wang, Qiang Liu, Shu Wu, Zilei Wang, and Liang Wang. Dive: subgraph disagreement for graph out-of-distribution generalization. In *Proceedings of the International ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2794–2805, 2024.
- Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. 2007.
- Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 491–500, 2024.
- Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 990–998, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1402008. URL <https://doi.org/10.1145/1401890.1402008>.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, Andrés

- György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrin, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Pöder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL <https://arxiv.org/abs/2503.19786>.
- Qin Tian, Chen Zhao, Minglai Shao, Wenjun Wang, Yujie Lin, and Dong Li. Mldgg: Meta-learning for domain generalization on graphs. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, pp. 1361–1372, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712456. doi: 10.1145/3690624.3709188. URL <https://doi.org/10.1145/3690624.3709188>.
- Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL <http://jmlr.org/papers/v9/vandermaaten08a.html>.
- Aristidis G. Vrahatis, Konstantinos Lazaros, and Sotiris Kotsiantis. Graph attention networks: A comprehensive review of methods and applications. *Future Internet*, 16(9), 2024. ISSN 1999-5903. doi: 10.3390/fi16090318. URL <https://www.mdpi.com/1999-5903/16/9/318>.
- Duo Wang, Yuan Zuo, Fengzhi Li, and Junjie Wu. Llms as zero-shot graph learners: Alignment of gnn representations with llm token embeddings. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 5950–5973, 2024a.
- Haotao Wang, Ziyu Jiang, Yuning You, Yan Han, Gaowen Liu, Jayanth Srinivasa, Ramana Kompella, Zhangyang Wang, et al. Graph mixture of experts: Learning on large-scale graphs with explicit diversity modeling. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 50825–50837, 2023.
- Qixun Wang, Yifei Wang, Hong Zhu, and Yisen Wang. Improving out-of-distribution generalization by adversarial training with structured priors. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.
- Qixun Wang, Yifei Wang, Yisen Wang, and Xianghua Ying. Dissecting the failure of invariant learning on graphs. In *Proceedings of the Conference on Neural Information Processing Systems*, 2024b.

- Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. Disentangled graph collaborative filtering. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1001–1010, 2020.
- Zaitian Wang, Pengfei Wang, Kunpeng Liu, Pengyang Wang, Yanjie Fu, Chang-Tien Lu, Charu C. Aggarwal, Jian Pei, and Yuanchun Zhou. A comprehensive survey on data augmentation, 2025. URL <https://arxiv.org/abs/2405.09591>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 24824–24837, 2022.
- Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective. In *Proceedings of the International Conference on Learning Representations*, 2022a.
- Qitian Wu, Nie Fan, Chenxiao Yang, Tianyi Bao, and Junchi Yan. Graph out-of-distribution generalization via causal intervention. In *The Web Conference*, 2024a.
- Qitian Wu, Fan Nie, Chenxiao Yang, Tianyi Bao, and Junchi Yan. Graph out-of-distribution generalization via causal intervention. In *Proceedings of the Web Conference*, pp. 850–860, 2024b.
- Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, 55(5):1–37, 2022b.
- Xin Wu, Jingjing Xu, Kuan Li, Jianping Yin, and Jian Xiong. Fldtmatch: Improving unbalanced data classification via deep semi-supervised learning with self-adaptive dynamic threshold. *Mathematics*, 13(3), 2025. ISSN 2227-7390. doi: 10.3390/math13030392. URL <https://www.mdpi.com/2227-7390/13/3/392>.
- Donglin Xia, Xiao Wang, Nian Liu, and Chuan Shi. Learning invariant representations of graph neural networks via cluster generalization. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 45602–45613, 2023.
- Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL <https://openreview.net/forum?id=ryGs6iA5Km>.
- Junchi Yu, Jian Liang, and Ran He. Mind the label shift of augmentation-based graph ood generalization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11620–11630, 2023. doi: 10.1109/CVPR52729.2023.01118.
- Xiang Yue, Zhen Wang, Jingong Huang, Srinivasan Parthasarathy, Soheil Moosavinasab, Yungui Huang, Simon M Lin, Wen Zhang, Ping Zhang, and Huan Sun. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics*, 36(4):1241–1251, 2020.
- Kexin Zhang, Shuhan Liu, Song Wang, Weili Shi, Chen Chen, Pan Li, Sheng Li, Jundong Li, and Kaize Ding. A survey of deep graph learning under distribution shifts: from graph out-of-distribution generalization to adaptation, 2025. URL <https://arxiv.org/abs/2410.19265>.
- Haiteng Zhao, Shengchao Liu, Ma Chang, Hannan Xu, Jie Fu, Zhihong Deng, Lingpeng Kong, and Qi Liu. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 5850–5887, 2023.
- Yusheng Zhao, Changhu Wang, Xiao Luo, Junyu Luo, Wei Ju, Zhiping Xiao, and Ming Zhang. Traci: A data-centric approach for multi-domain generalization on graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12):13401–13409, Apr. 2025a. doi: 10.1609/aaai.v39i12.33463. URL <https://ojs.aaai.org/index.php/AAAI/article/view/33463>.

- Ziwen Zhao, Yixin Su, Yuhua Li, Yixiong Zou, Ruixuan Li, and Rui Zhang. A survey on self-supervised graph foundation models: Knowledge-based perspective, 2025b. URL <https://arxiv.org/abs/2403.16137>.
- Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *CoRR*, abs/2006.04131, 2020. URL <https://arxiv.org/abs/2006.04131>.
- Yun Zhu, Haizhou Shi, Zhenshuo Zhang, and Siliang Tang. Mario: Model agnostic recipe for improving ood generalization of graph contrastive learning. In *Proceedings of the ACM Web Conference 2024, WWW '24*, pp. 300–311, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645322. URL <https://doi.org/10.1145/3589334.3645322>.
- Chenyi Zhuang and Qiang Ma. Dual graph convolutional networks for graph-based semi-supervised classification. In *Proceedings of the Web Conference*, pp. 499–508, 2018.

A LARGE LANGUAGE MODEL (LLM) USAGE STATEMENT

We use the LLM as a general-purpose assistant tool. Specifically, the LLM assists in (i) checking grammar and improving clarity of text descriptions, and (ii) suggesting alternative phrasings for some sections. No parts of the paper are generated entirely by the LLM. All research ideas, experiments, model designs, and results are conceived, implemented, and analyzed solely by the authors. The LLM does not contribute to the development of the methodology, experiments, or analysis presented in this paper. We confirm that the use of the LLM is limited to minor writing support and does not constitute a substantive contribution that would qualify it as a co-author.

B THEORETICAL ANALYSIS

In this section, we present a theoretical analysis of the proposed **LIGHT** framework. We assume the existence of a set of K experts, each specialized in learning a particular domain. Each domain may be subject to potential distribution shifts, making domain-specific modeling essential. For notational simplicity, we focus on a binary classification setting with $C = 2$ classes and assume the presence of two distinct domains. Suppose that for the first $K/2$ experts, the domain-specific distribution is aligned with the first domain, while the remaining $K/2$ experts are aligned with the second domain. Assume that the routing weights $\omega_v^{(k)}$ satisfy the following condition: for the correct expert k ,

$$\omega_v^{(k)} = \frac{2}{K}, \quad \text{and for all incorrect experts } k' \neq k, \quad \omega_v^{(k')} = 0,$$

which indicates that the routing mechanism assigns equal weight to the correct expert while ignoring all others. Let the final prediction for node v be defined as:

$$\hat{Y}_v = f \left(\sum_{k=1}^K \omega_v^{(k)} \mathbf{Z}_v^{(k)} \right),$$

where $\mathbf{Z}_v^{(k)}$ is the representation of node v obtained from the k -th expert, $\omega_v^{(k)}$ denotes the routing weight for expert k , and $f(\cdot)$ is a learnable classifier. We further assume that the classifier $f(\cdot)$ can be expressed as a logistic (sigmoid) transformation of a linear combination of expert outputs:

$$f \left(\sum_{k=1}^K \omega_v^{(k)} \mathbf{Z}_v^{(k)} \right) = \sigma \left(\sum_{k=1}^K \omega_v^{(k)} \mathbf{W}^\top \mathbf{Z}_v^{(k)} \right),$$

where $\sigma(\cdot)$ is the sigmoid activation function and $\mathbf{W}$ is the weight matrix of the classifier. Under this formulation, the term $\mathbf{W}^\top \mathbf{Z}_v^{(k)}$ can be interpreted as the logit output of the k -th expert for node v . Due to the monotonicity of the sigmoid function, a larger logit corresponds to a higher predicted probability of class 1. The following lemma establishes a threshold-based decision rule for the aggregated expert logits.

Lemma B.1. *There exists a threshold $\tau \in \mathbb{R}$ such that:*

$$\hat{Y}_v = \begin{cases} 1, & \text{if } \sum_{k=1}^K \omega_v^{(k)} \mathbf{W}^\top \mathbf{Z}_v^{(k)} > \tau, \\ 0, & \text{otherwise.} \end{cases}$$

In the multi-class setting ($C > 2$), a similar formulation applies by replacing the sigmoid function with the softmax function, and defining a class-specific threshold for each output dimension.

With this lemma, we can analyze the performance of the proposed **LIGHT** framework. Specifically, we aim to show that a mixture-of-experts (MoE) model can outperform a single-expert model in terms of classification accuracy. Without loss of generality, we assume that the threshold τ in Lemma B.1 is set to 0. Let the logit output of the k -th expert for node v be denoted as

$$Q_{v,k} = \mathbf{W}^\top \mathbf{Z}_v^{(k)},$$

and let its expected value be $\mu_{v,k} = \mathbb{E}[Q_{v,k}]$. For a given node v , suppose expert k is the correct expert (i.e., best aligned with the domain-specific distribution of v). Then, the expected logit $\mu_{v,k}$ should exceed a positive margin $\nu > 0$, while any incorrect expert $k' \neq k$ should have mean

logit output $\mu_{v,k'} = 0$. This reflects the fact that only the correct expert can effectively model the distributional characteristics of node v . We further assume that the logits of different experts are independent and sub-Gaussian distributed, i.e.,

$$\mathbb{E}[\exp t(Q_{v,k} - \mu_{v,k})] \leq \exp \frac{t^2 \sigma^2}{2}, \quad \forall t \in \mathbb{R},$$

where σ^2 is a constant. With these assumptions, we can analyze the prediction error of the MoE model.

Proof of Theorem 4.1. We begin by analyzing the case where only a single expert is available, i.e., $K = 1$. Suppose the true label for node v is 1. The prediction error of the model is then given by

$$\mathbb{P}(\hat{Y}_v \neq 1) = \mathbb{P}(Q_{v,1} < 0).$$

We aim to show that this probability is strictly greater than zero, i.e., $\mathbb{P}(Q_{v,1} < 0) > 0$.

Since the expert is misaligned with the domain-specific distribution of node v , we assume $\mathbb{E}[Q_{v,1}] = 0$. Define the following events:

$$\mathcal{A}_1 = \{Q_{v,1} < 0\}, \quad \mathcal{A}_2 = \{Q_{v,1} \geq 0\}, \quad \mathcal{A}_3 = \{Q_{v,1} = 0\}.$$

Because $Q_{v,1}$ is not almost surely constant, we must have $\mathbb{P}(\mathcal{A}_3) < 1$. If $\mathbb{P}(Q_{v,1} < 0) = 0$, then necessarily $\mathbb{P}(Q_{v,1} > 0) > 0$, which contradicts the assumption that $\mathbb{E}[Q_{v,1}] = 0$. Therefore, it must be that $\mathbb{P}(Q_{v,1} < 0) > 0$, and the prediction error is strictly positive in the single-expert case.

Now consider the case with K experts. Suppose expert k is correctly aligned with the domain of node v , and thus has expected logit output $\mu_{v,k} = \nu > 0$, while all other experts $k' \neq k$ have zero-mean outputs, $\mu_{v,k'} = 0$. Then the expected value of the aggregated logit is

$$\mathbb{E} \left[\sum_{k=1}^K \omega_v^{(k)} Q_{v,k} \right] = \sum_{k=1}^K \omega_v^{(k)} \mu_{v,k} = \nu.$$

Assuming the expert outputs $Q_{v,k}$ are independent and sub-Gaussian with variance proxy σ^2 , we apply Hoeffding's inequality to bound the tail probability:

$$\mathbb{P} \left(\sum_{k=1}^K \omega_v^{(k)} Q_{v,k} < 0 \right) \leq \exp \left(-\frac{C\nu^2}{2\sigma^2/K} \right).$$

Therefore, the prediction error of the MoE model satisfies:

$$\mathbb{P}(\hat{Y}_v \neq 1) \leq \exp \left(-\frac{C\nu^2 K}{2\sigma^2} \right) \leq \frac{C_0}{K},$$

for some constant $C_0 > 0$. This completes the proof. $\square$

C GREEDY ANCHOR NODE SELECTION

Algorithm 1 Greedy Anchor Node Selection

Input: PPMI matrix P , Influence Score $I_v(u)$ from node v to node u , budget B .

Output: Anchor Node set S

1. $S = \emptyset$
 2. **for** $t = 1, 2, \dots, B$ **do**
 - (a) **for** $v \in \mathcal{V}_{train} \setminus S$ **do**
 - i. update $\sigma(S \cup \{v\})$ based on Eq. (9)
 - (b) **end for**
 - (c) $v^* = \arg \max_{v \in \mathcal{V}_{train} \setminus S} (|\sigma(S \cup \{v\})| - |\sigma(S)|)$
 - (d) $S = S \cup \{v^*\}$
 3. **end for**
 4. **return** S
-

D BASELINE DETAILS

To comprehensively evaluate our method, we compare it against a comprehensive suite of baseline methods. These baselines are carefully selected to cover a range of approaches: (i) a fundamental neural network model that processes node features independently of graph structure; (ii) standard Graph Neural Network architectures trained via empirical risk minimization on the source domain, representing common graph representation learning techniques without explicit domain generalization mechanisms; and (iii) contemporary methods specifically designed for graph out-of-distribution generalization or domain adaptation, which employ diverse strategies to tackle distribution shifts. Following previous works [Kipf & Welling \(2017\)](#); [Qiao et al. \(2023a\)](#), all GNN-based baselines are trained using the 5% labeled data from the source domain for the specified cross-domain generalization tasks. To assess performance against established graph representation learning techniques, we include several widely-adopted architectures. We begin with an MLP which serves to quantify the performance achievable without considering the intrinsic graph structure.

E FURTHER ANALYSIS OF ABLATION EXPERIMENT

To meticulously validate the individual contributions of the key components within our proposed method, we conducted a series of ablation experiments. We systematically removed or simplified core modules: (1) the Mixture of Experts architecture (w/o MoE), reverting to a more standard single MLP encoder. Without the expert mechanism, all other modules based on it are also removed; (2) the LLM-guided expert selection distillation (w/o LLM); (3) the MoE’s internal diversity/coherence loss (w/o DIV); and (4) our expert-aware dynamic pseudo-labeling strategy for semi-supervised learning (w/o HQ). The performance of these variants was evaluated on the same six domain generalization tasks, with results and the extent of performance degradation detailed in Table 3.

Table 3: Ablation studies on the Citations benchmark with component presence indicators.

Components				$A \Rightarrow C$		$A \Rightarrow D$		$C \Rightarrow A$		$C \Rightarrow D$		$D \Rightarrow A$		$D \Rightarrow C$	
PPMI	LLM	DIV	HQ	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
		✓		71.3	66.0	65.9	58.8	65.7	63.1	69.1	63.6	50.4	42.6	65.7	61.3
	✓			68.8	64.4	65.7	60.1	62.8	58.7	68.2	67.3	55.8	49.7	54.6	49.2
	✓	✓		69.6	65.3	67.0	62.0	63.7	60.2	71.0	68.7	56.7	51.2	56.4	49.9
	✓		✓	73.2	72.0	68.2	66.1	66.7	67.4	70.4	70.0	58.9	55.1	66.1	59.8
✓		✓		72.0	66.6	66.7	58.5	66.1	63.3	69.7	64.7	51.0	43.4	66.0	61.7
✓	✓			69.3	65.7	66.2	61.0	63.1	59.1	69.8	68.5	56.3	50.7	55.0	49.9
✓	✓	✓		70.0	66.0	67.5	62.2	64.4	60.4	71.1	69.2	57.5	52.0	57.2	50.1
✓	✓		✓	74.9	72.7	68.6	66.9	67.5	68.1	71.8	71.2	60.3	57.5	68.0	61.8
✓	✓	✓	✓	75.7	73.5	70.7	68.5	70.6	71.3	73.3	72.3	64.5	64.3	73.6	71.6

Impact of MoE Architecture (MoE). The results show removing the Mixture of Experts framework leads to the most substantial performance decline across all tasks and metrics, with Micro-F1 scores dropping by an average of approximately 29.2% and Macro-F1 by 34.5%. For instance, on $A \Rightarrow C$, the Micro-F1 score plummets by 32.4 points. Notice without MoE Architecture, all other experts related mechanism are disabled and our framework revert to simple MLP-based method. This drastically reduced performance underscores that the ability of MoE to employ specialized experts and adaptively combine their insights is fundamental to our model’s success, exceeding the capabilities of a monolithic encoder structure in these challenging domain generalization scenarios.

Impact of LLM-guided Expert Selection (LLM). Disabling the LLM-based distillation for expert selection results in a significant performance drop, with Micro-F1 scores decreasing by an average of approximately 7.3% and Macro-F1 by 12.9%. The degradation is particularly notable on tasks like $D \Rightarrow A$ (13.5 points drop in Micro-F1). This highlights the contribution of the LLM’s semantic guidance in refining the MoE’s expert routing, especially for uncertain nodes. Without this external, high-level reasoning, the MoE’s autonomous gating mechanism is demonstrably less effective at tailoring expertise to specific node contexts for optimal generalization.

Impact of MoE Diversity/Coherence Loss (DIV). Removing the MoE’s internal diversity loss results in a noticeable, albeit comparatively smaller, performance decrease, with Micro-F1 scores dropping by an average of 2.9% and Macro-F1 by 4.0%. While more modest, this degradation across tasks indicates that this regularization term plays a valuable role in refining the representations learned by each expert. By encouraging intra-expert coherence or better-defined specializations, the diverse neighborhood exploration also avoid the routing network get into local optimal solution.

Impact of Expert-aware Dynamic Pseudo-labeling (HQ). The exclusion of our specialized expert-aware dynamic pseudo-labeling strategy also leads to a considerable reduction in performance, averaging a 7.0% decrease in Micro-F1 and an 11.6% decrease in Macro-F1. This emphasizes the importance of this sophisticated semi-supervised learning component, which selectively leverages unlabeled source data. By identifying reliable pseudo-labels through a process informed by expert consistency and LLM insights, our method effectively learns from the scarce 5% labeled data and mitigates error propagation common in naive pseudo-labeling approaches.

Impact of Positive Pointwise Mutual Information (PPMI). We replace $P_v^{(k)}$ to the original graph adjacency matrix. This results of the overall performance has decreased by about 1-2%. The reason is that without such a model architecture highly sensitive to topological structures, the differentiation in feature extraction among heterogeneous experts diminishes. Consequently, it becomes more challenging for the router to assign nodes to appropriate experts.

F FURTHER ANALYSIS OF PARAMETER SENSITIVITY

Number of Experts. We further investigated the influence of the number of experts on the model’s performance. Figure 3 (a) and (b) illustrates two scenarios: *Varying Expert Types*: The results from (a) shows how performance changes as the diversity and numbers of expert types increases. The results suggest that increasing the variety of specialized experts can merely positively impact the Micro F1 score up to a certain point by providing more capacity or stability to the specialized processing. *Varying Expert Instances*: The results from (b) demonstrates the effect of increasing the number of instances for each type of expert. This indicates that having more instances of the same expert type can also contribute to performance improvements, after which the gains may diminish or performance could even degrade.

Label Rate. Figure 3 (c) and (d) examines our model’s performance and robustness to the scarcity of labeled data by varying the label rates in the source domain. The experiments were conducted on the $A \Rightarrow C$ and $A \Rightarrow D$ tasks. As expected, a higher label rate generally leads to better performance. However, the proposed method demonstrates strong performance with low label rates, outperforming other compared methods, highlighting its effectiveness in semi-supervised domain generalization scenarios.

Loss Weights. The impact of different loss weight hyperparameters ($\lambda_u, \lambda_{dl}, \lambda_{ind}$) on the Micro-F1 score was analyzed, as shown in the heatmaps in Figure 4 (c) and (d) for tasks $A \Rightarrow C, C \Rightarrow D$. These hyperparameters control the balance between the unsupervised classification loss, the knowledge distillation loss, and the expert diversity regularization loss, respectively. The results indicate that the optimal weights can vary depending on the specific domain generalization task. Generally, the model’s performance is sensitive to these weights, and careful tuning is required to achieve the best results. For instance, in the $A \Rightarrow C$ task, specific ranges for λ_u, λ_{dl} , and λ_{ind} yield higher Micro-F1 scores, as depicted by the color intensity in the heatmap. Similar trends, though with potentially different optimal regions, are observed for the $C \Rightarrow D$ and $D \Rightarrow C$ tasks.

Smoothing Coefficient and Activation Threshold. Figure 5 presents an analysis of the smoothing coefficient (α) used in the label smoothing strategy for expert selection and the activation threshold (θ) used in the expert-enhanced pseudo-labeling process. These experiments were performed on the $A \Rightarrow C, A \Rightarrow D, C \Rightarrow A$ and $C \Rightarrow D$ tasks. The smoothing coefficient α determines the softness of the target distribution provided by the LLM for training the routing function. The activation threshold θ is used to filter noisy pseudo-labels by selecting high-confidence samples. The results suggest that model performance is influenced by the interplay of these two parameters. Most of optimal performance is achieved within specific ranges of $\alpha = 0.8$ and $\theta = 0.05$, indicating that both appropriate smoothing of LLM guidance and careful selection of pseudo-labels are important for the model’s success.

LLM Types. We further explored the effect of using different types of LLMs for the LLM-guided expert selection, as illustrated in Figure 4 (a) and (b) for the $A \Rightarrow C$ and $A \Rightarrow D$ tasks. The results compare the performance achieved using various LLMs, such as qwen2.5-3b, qwen2.5-7b, qwen2.5-14b Qwen et al. (2025), gemma3-4b, and gemma3-12b Team et al. (2025). The results indicate that the choice of LLM can impact the model’s performance. Larger or more capable LLMs (e.g., qwen2.5-14b or gemma-12b in some cases) tend to provide better guidance, leading to

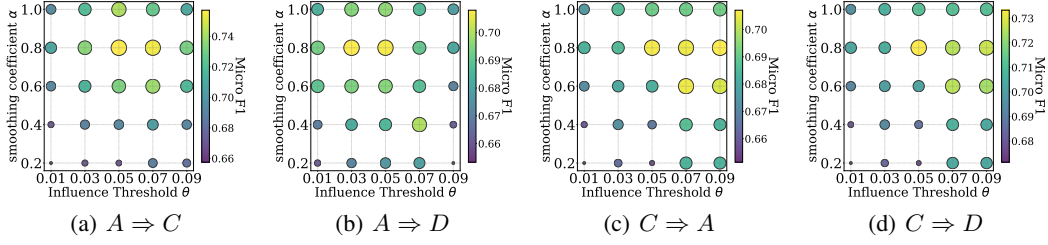


Figure 5: Effect analysis of smoothing coefficient α and influence threshold θ .

improved results, though the best-performing LLM might vary depending on the specific task. We observe 3-4% of performance degradation with small size model like 3-4b, while the performance gap between the 7b model and 12-14b model is not significant.

G LIMITATIONS

While the proposed framework demonstrates promising results in semi-supervised domain generalization on graphs, we acknowledge certain limitations that offer avenues for future research.

Firstly, our empirical evaluation was conducted on established benchmark datasets in academic research. Although these datasets are standard for assessing graph learning models, the performance of LIGHT on graphs with substantially different characteristics, such as highly dynamic graphs or graphs from vastly different domains (e.g., biological networks or social networks with unique structures), may vary. Further investigation across a broader spectrum of graph types and real-world applications would be beneficial to ascertain the generalizability of our approach.

Secondly, the use of LLMs and graph data inherently touches upon broader considerations of privacy and fairness. Depending on the nature of the graph data and the information processed by the LLM, potential privacy implications could arise. Similarly, fairness in node classification could be affected by biases present in either the graph data or the LLM. Future research could explicitly address these aspects by incorporating privacy-preserving techniques or fairness-aware learning objectives within our framework.

Addressing these limitations could further enhance the robustness, applicability, and societal alignment of the LIGHT framework and similar approaches leveraging LLMs for graph-based machine learning.

H SOCIETAL IMPACTS

Our method primarily focuses on advancing graph neural networks for semi-supervised domain generalization, aiming to improve their performance on unseen graphs with limited labeled data. The core idea involves using Large Language Models to guide a mixture-of-experts framework, enhancing knowledge distillation and addressing label scarcity through dynamic pseudo-labeling. The successful implementation of this approach could lead to:

Improved Machine Learning Model Generalizability: The LIGHT framework’s ability to generalize effectively to unseen domains, even with scarce labeled data, can lead to more robust and adaptable machine learning models across various applications. This is particularly relevant in fields where data distribution shifts are common and obtaining labeled data is challenging or expensive.

Enhanced Performance in Various Domains: The paper highlights node classification as a fundamental task with wide applications in areas such as social network analysis, biochemistry, and recommender systems. Improvements in this area, as demonstrated by the LIGHT framework, could translate to more accurate predictions and better decision-making in these domains.

Advancements in Semi-supervised Learning: The research contributes to the field of semi-supervised learning on graphs by pioneering the use of LLM zero-shot capabilities to address domain

generalization under label scarcity. This could inspire new methodologies for leveraging LLMs in data-efficient graph learning.

Better Understanding of Complex Relational Data: By effectively capturing intricate relational and structural dependencies in graph data, this research can contribute to a deeper understanding of complex systems and networks.

We also acknowledge Potential Negative Societal Impacts:

Potential for Misuse: As with many advancements in machine learning, there's a general risk that improvements in GNNs and LLMs could be adapted for malicious purposes if not carefully governed.

Fairness Considerations: Biases present in either the graph data or the LLM could affect fairness in node classification. If the training data or the LLM itself contains biases against certain groups, the deployed technology could make decisions that unfairly impact these groups.