DATE: Dual Prompt Learning with Information Bottleneck for Graph Out-of-Distribution Generalization

Jiayi Zeng
2310536@stu.neu.edu.cn
Northeastern University
Shenyang, China

Tao Ren
rent@swc.neu.edu.cn
Northeastern University
Shenyang, China

Changhu Wang
wangch156@g.ucla.edu
University of California, Los Angeles
Los Angeles, United States

Yifan Wang*
yifanwang@uibe.edu.cn
University of International Business and Economics
Beijing, China

Wei Ju
juwei@pku.edu.cn
Peking University
Beijing, China

Zhipeng Sun
2410599@stu.neu.edu.cn
Northeastern University
Shenyang, China

Xiao Luo
xiaoluo@cs.ucla.edu
University of California, Los Angeles
Los Angeles, United States

Abstract

This paper studies the problem of graph out-of-distribution generalization, which aims to enhance the performance of graph neural networks (GNNs) under distribution shifts. Existing approaches usually learn graph representations from a causal graph, which may not explicitly utilize environment information. Furthermore, they could suffer from performance degradation when confusing semantics related to target labels and environments. In this paper, we propose a novel framework named Dual Prompt Learning with Information Bottleneck (DATE) for graph out-of-distribution generalization. The core of our DATE is to utilize dual prompts to extract task-oriented semantics and model distribution shifts, respectively. In particular, we first pre-train a GNN using contrastive learning with pretext tokens introduced. More importantly, we not only introduce a task-oriented prompt based on LLMs to generate environment-invariant representations, but also learn the environment-oriented prompts to simulate subgraphs in different environments. To optimize our prompts, we introduce a graph information bottleneck framework, which minimizes the mutual information between environment-invariant representations and environment semantics with the most semantics preserved. Extensive experiments on various benchmark datasets are further conducted to validate the effectiveness of our DATE against various state-of-the-art approaches.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '25, Dublin, Ireland

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755795

CCS Concepts

• **Mathematics of computing** → **Graph algorithms**; • **Computing methodologies** → **Learning latent representations**.

Keywords

OOD generalization; graph neural networks; prompt-learning; information bottleneck; LLMs

ACM Reference Format:

Jiayi Zeng, Tao Ren, Changhu Wang, Yifan Wang, Wei Ju, Zhipeng Sun, and Xiao Luo. 2025. DATE: Dual Prompt Learning with Information Bottleneck for Graph Out-of-Distribution Generalization. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755795

1 Introduction

Graph serves as an efficient data structure for modeling intricate relationships in real-world contexts, such as protein interactions [5], molecular networks [16], and social networks [3, 35, 44]. Learning from graph-structured data, particularly through graph neural networks (GNNs), aims to encode the underlying dependencies within a latent space and has become a *de facto* standard in numerous applications, including drug discovery [11], physics simulation [36], and recommender systems [23, 43, 45, 46, 49]. However, these models assume that test and training samples are selected independently from an identical distribution (i.i.d.). When GNNs are deployed in new environments with out-of-distribution (OOD) graphs, they often suffer from poor performance and unstable predictions.

Recently, there has been growing research interest in enhancing the OOD generalization of GNNs under different environments, mainly from two perspectives. Data-driven approaches focus on manipulating the input data, which generates diverse graph instances to enhance generalization. These approaches encompass strategies like data augmentation [15, 37] and distribution augmentation [27, 42]. In contrast, learning-driven approaches emphasize

adjusting the model training process by incorporating specific objectives and constraints for optimization. And nootable methods include graph (causal) invariant learning [48, 61], model regularization [66], and new model architecture [14, 55]. Nevertheless, these methods still remain unsatisfactory due to the failure to handle the nuanced variations in the distribution of graph structural properties and contextual information present in training data.

The judgment of existing limitations arises from two key observations. **Firstly**, *the environment information in graph OOD generalization has not been fully explored.* Although acquiring or annotating environment labels incurs extra costs, it has been demonstrated that generalization without this supplementary information is theoretically impossible [26]. Existing methods often rely on a pre-defined environment codebook to enhance the OOD generalization of GNNs in unseen environments [13, 56]. **Secondly**, *fail to exploit distribution shifts in a manner specifically tailored for graph data.* Compared to OOD scenarios in Euclidean spaces, environments within graph data are characterized by diverse factors. Notably, distribution shifts in graphs are fundamentally reflected in the changes of structural associations. However, current methods do not effectively account for or address these shifts in the context of graph data.

Fortunately, benefiting from massive labeled and unlabeled data, recent advancements underscore the promising potential of pre-trained models in capturing rich prior knowledge and improving domain generalization [1]. Based on the pre-trained model, further prompt tuning provides an alternative method and focuses on transforming the input data space for model adaptation [22]. Particularly in the case of GNNs, on one hand, while traditional OOD generalization methods typically rely on encoders trained exclusively on graph data, we argue that integrating foundation models with prompt engineering offers a more effective approach to bridge the semantic gap [7, 33]. On other hand, although some works attempt to design prompt frameworks for graphs, most of them introduce randomly initialized parameters as virtual prompts [9, 39]. By incorporating environment information as an additional self-prompt, we can better capture underlying environment-specific patterns in graph data without predefining scales.

Towards this end, in this paper, we propose Dual Prompt Learning with Information Bottleneck (DATE) for graph OOD generalization, which follows the "pre-train and prompt" paradigm to guide the foundation model towards uncovering invariant features across all environments. Specifically, in the pre-training stage, we introduce a pretext token-augmented graph contrastive learning approach, enabling the model to assimilate structure associations from the graph data. Then, during the prompt tuning stage, our model employs the large language models (LLMs) as the enhancer, supplying the pre-trained model with the essential contextual prompt to extract the environment-invariant feature. Furthermore, given the pre-collected environment, we exploit a novel self-prompted learning mechanism to infer the underlying graph environment and a graph information bottleneck (GIB) constraint is proposed to eliminate spurious features and preserve environment-invariant features to achieve generalization. Extensive experiments conducted on various public datasets further validate the superiority of our DATE in comparison to extensive baselines.

To summarize, the primary contributions of this paper are outlined as follows:

- *Problem Connection.* We propose a new perspective that connects prompt learning with graph out-of-distribution generalization, which pioneers a path of leverage language models to address this problem.
- *Novel Methodology.* Our DATE is a graph information bottleneck framework, which not only introduces a task-oriented prompt to guide environment-invariant representation learning, but also leverages environment-oriented prompts to simulate distribution shifts.
- *Extensive Experiments.* We perform thorough experiments on multiple public datasets to systematically evaluate the framework. Empirical results demonstrate the superior performance of our DATE in graph OOD generalization.

2 Related Work

Graph OOD Generalization. The performance degradation challenge in graph machine learning under distribution shifts is primarily addressed by the field of Graph Out-of-Distribution generalization through three main strategies. The first, Data Augmentation and Distribution Alignment, utilizes methods like instance reweighting [50, 65] and the extraction of causal-based invariant subgraphs [32]. A second strategy, Invariant Model Design, focuses on acquiring causal-aware representations with domain-invariant encoders [54, 64]. The third, Optimization Strategies, leverages techniques such as distributionally robust objectives and self-supervised learning, exemplified by EERM. Complementing these methods, current evaluation frameworks [12, 17] are in place to assess model performance across these varied distribution shifts.

Causality & Invariant Learning. Addressing the performance degradation of GNNs under distribution shifts [25, 31] necessitates moving beyond standard empirical risk minimization. Two prominent paradigms, causal inference and invariant learning, offer principled approaches to enhance OOD generalization on graphs. Causal inference provides tools to model the underlying data generating processes, identify confounding factors, and estimate true causal effects, often employing techniques like backdoor adjustment to mitigate bias. In parallel, invariant learning seeks to identify predictive features, representations, or substructures that remain stable across diverse environments or domains [6, 20]. The core idea is that while spurious correlations may vary across distributions, true causal relationships tend to be invariant [19]. For graphs, this involves learning representations robust to feature or topological shifts by minimizing risk variance across environments [41] or identifying invariant subgraphs [6, 19].

Graph Prompt Tuning. Graph Prompt Tuning enables pre-trained GNNs to downstream tasks through three primary directions. Feature-space perturbation methods, represented by GPF [9], introduce learnable feature perturbations for parameter-efficient adaptation. Subgraph similarity unification frameworks, such as GraphPrompt [28], align pre-training and downstream tasks through subgraph-based prompts. Task reconstruction approaches, exemplified by GPPT [40], reformulate node classification as edge prediction using virtual label nodes. To address challenges on complex graph data, DAGPrompT [8] enhances the GNN encoder and incorporates hop-specific prompts in node and graph classification tasks.

In comparison with these works, we leverage dual prompt tuning to enhance graph out-of-distribution generalization.

3 Preliminary and Problem Definition

Notations. We denote $G = (\boldsymbol{X}, \boldsymbol{A}) \in \mathbb{G}$ as an attributed graph with n nodes and m edges, respectively. The node feature matrix can be represented as $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, where row vector $\boldsymbol{x}_v \in \mathbb{R}^d$ corresponds to the feature of node v with dimension d. And we leverage the adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ to describe the structure of graph, where $\boldsymbol{A}_{uv} = 1$ if an edge $(u, v) \in G$ exists between two nodes, otherwise, $\boldsymbol{A}_{uv} = 0$. In the graph classification scenerio, we assign a label $y \in \mathbb{Y}^c$ for each graph, where c is the number of classes.

Problem Definition. Let $\mathcal{D} = \{(G_i^e, y_i^e)\}_{e\in\mathcal{E}_{all}}$ be a dataset which can be separated as a training set $\mathcal{D}_{tr}$ and test set $\mathcal{D}_{te}$, where $\mathcal{E}_{all}$ represents the environments and the two sets originate from distinct distributions, i.e., $P_{tr}(G^e, y^e) \neq P_{te}(G^{e'}, y^{e'}), e \in \mathcal{E}_{tr}, e' \in \mathcal{E}_{te}, \mathcal{E}_{te} = \mathcal{E} \setminus \mathcal{E}_{tr}$. In contrast to Euclidean space, where distribution shifts predominantly manifest in feature vectors, graph space distribution shifts may involve more complex factors, including topological transformations. Meanwhile, as demonstrated in benchmarks like GOOD [12] and DrugOOD [18], environment labels for graph datasets are often readily accessible. Thus, bypassing the environmental inference, we directly incorporate this information for more effective environmental exploitation. Given $\mathcal{D}_{tr}$ with pre-collected environment labels, our objective is to learn a powerful GNN model, $f^*(\cdot) = h^*(g^*(\cdot)) : \mathbb{G} \to \mathbb{Y}$, capable of performing well across all possible environments:

$$f^* = \arg\min_{f} \sup_{e\in\mathcal{E}_{all}} \mathcal{R}(f|e), \tag{1}$$

where the GNN model composed of two parts, $g(\cdot)$ and $h(\cdot)$, which denote the graph encoder and corresponding classifier. Under the environment e, the empirical risk can be defined as $\mathcal{R}(f|e) = \mathbb{E}^e[\mathcal{L}(f(G^e), y^e)]$, and $\mathcal{L}(\cdot,\cdot)$ denotes a loss function.

4 Method

The overview of our proposed environment-aware framework for graph OOD generalization is illustrated in Figure 1. Our DATE comprises three core components. The pre-training phase adopts an asymmetric graph contrastive learning-based approach (see Section 4.1) to capture the entire structure associations for the task. Then, based on the pre-trained GNN, we incorporate the environment-aware prompt tuning framework (see Section 4.2), which involves two essential aspects. On the one hand, given the invariant description of the graph domain (e.g., the task and class description), we inject the text prompt embedding generated with LLM for prompt tuning and derive the environment-invariant feature. On the other hand, we design a set of learnable prompt tokens for environmental factors and extract the corresponding node-centred subgraph as the environment graph to capture the data distribution shift. To enhance generalization, an Information Bottleneck constraint, which is detailed in Section 4.3, is implemented. This constraint concurrently serves two objectives: it maximizes the mutual information (MI) between the derived environment-invariant feature and the graph environment, while simultaneously minimizing the MI connecting this feature to the environment label.

4.1 Graph Neural Network Pre-training

In this section, we examine the initial pre-training phase of the graph encoder. Although a variety of graph-based pretext tasks can be employed for pre-training, previous research on multi-task learning has shown that task diversity may introduce interference, which in turn results in suboptimal pre-training outcomes [47, 63]. To mitigate this issue, we introduce a graph contrastive learning approach that incorporates the pretext tokens to reformulate the task input during pre-training, enabling the process to synergize downstream tasks and alleviate task interference [62]. Specifically, we put forth two pretext tokens $\{\boldsymbol{p}^* \in \mathbb{R}^d\}, * = \{1, 2\}$, with each pretext token added to the node feature, expressed as:

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\},\ \boldsymbol{X}^* = \{\boldsymbol{x}_1 + \boldsymbol{p}^*, \ldots, \boldsymbol{x}_n + \boldsymbol{p}^*\}, \tag{2}$$

where the modified feature $\boldsymbol{X}^*$ replaces the original feature to generate two views of the graph. To avoid the trivial solution, we apply parameter-unshared simases encoders to conduct the two pretext tokens [58]. Note that different from the traditional graph contrastive learning with augmentation [60], we add the pretext tokens [9] and assume that this perturbation does not affect the model predictions, the contrastive loss in batch $\mathcal{B}$ can be:

$$\mathcal{L}_{cl} = -\frac{1}{|\mathcal{D}_{tr}|} \sum_{G_i\in\mathcal{D}_{tr}} \log\left(\frac{\exp(\text{sim}(\boldsymbol{z}_{i,1}, \boldsymbol{z}_{i,2})/\tau)}{\sum_{G_{i'}\in\mathcal{B}} \exp(\text{sim}(\boldsymbol{z}_{i,1}, \boldsymbol{z}_{i',2})/\tau)}\right), \tag{3}$$

where $\boldsymbol{z}_{i,*} = f(\boldsymbol{X}_i^*, \boldsymbol{A}_i) \in \mathbb{R}^{d'}$ denote the encoded graph representation with dimension d' in two views, $\text{sim}(\cdot,\cdot)$ corresponds to the similarity function and τ denote the temperature parameter. We add the cross-entropy loss $\text{CE}(\cdot)$ for graph classification to further synergize the downstream generalization task.

$$\mathcal{L}_{ce} = \sum_{i\in\mathcal{D}_{tr}} \text{CE}(y_i, h(\boldsymbol{z}_i)),\ \boldsymbol{z}_i = \frac{1}{2}(\boldsymbol{z}_{i,1} + \boldsymbol{z}_{i,2}). \tag{4}$$

The pre-training stage yields the optimal model parameters $\Theta^* = \arg\min_\Theta(\mathcal{L}_{cl}+\mathcal{L}_{ce})$, which are fixed during the downstream prompt tuning to facilitate the knowledge transfer.

4.2 Dual Prompt Tuning for Representation Disentanglement

To effectively capture the generalizable prior knowledge across environments, we introduce an environment-aware prompt tuning framework that contains both environment-invariant prompt injection and environment-disturbed self-prompt tuning on the frozen pre-trained GNN model.

Task-oriented Prompts for Invariant Learning. Since the LLM is capable of extracting the semantics from textual descriptions across environments, we leverage the LLM to assist soft prompt generation. The text description $\boldsymbol{t}$ can be divided into: *dataset overview*, *attributes and label*, *task details*, and *environment description*.

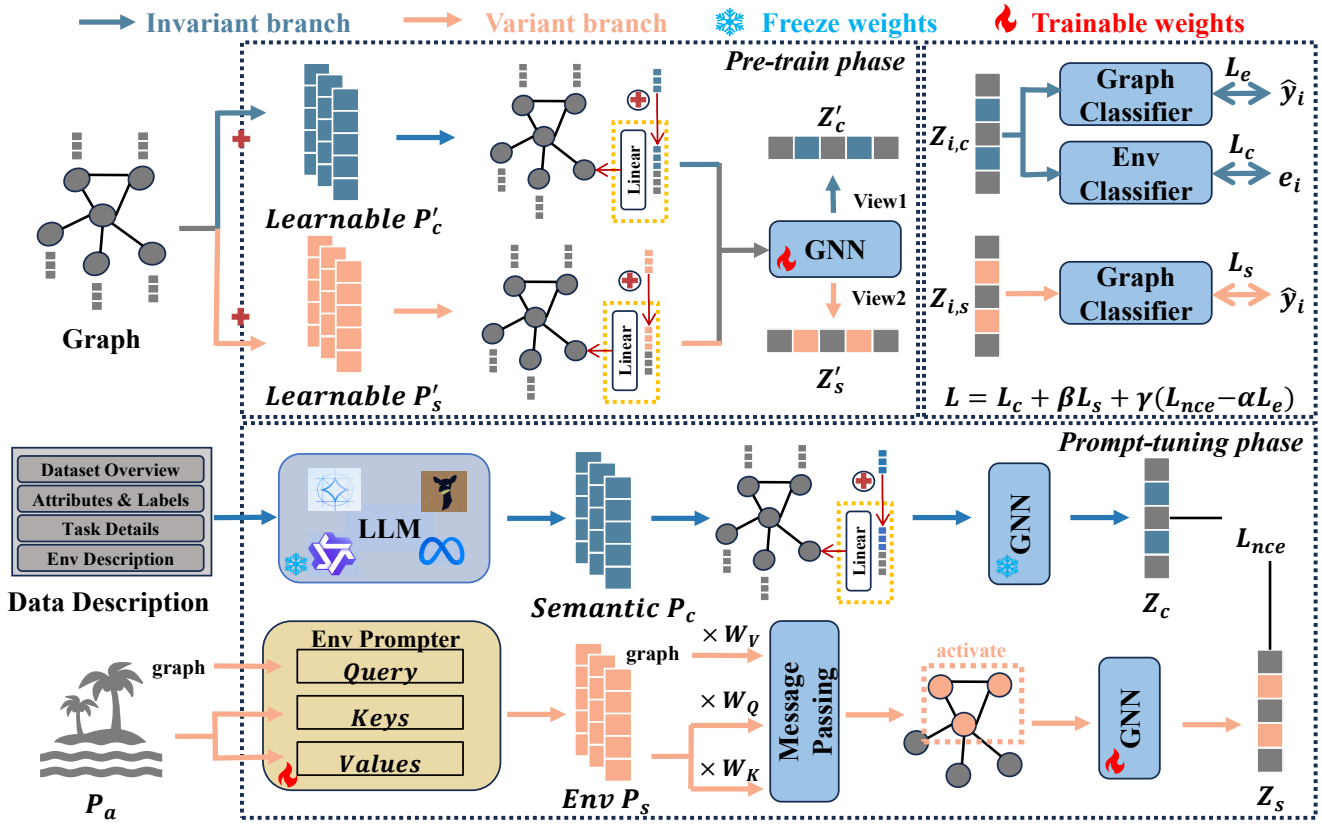

Figure 1: This figure illustrates the architecture of our proposed DATE framework, which is trained via a two-stage process. The initial pre-training stage introduces a learnable prompt. Subsequently, during the fine-tuning stage, we froze the pre-trained GNN model while replacing this learnable prompt with the environment-invariant prompt injection and environment perturbation self-prompt tuning.

> **Description Example:** You are an expert in drug discovery and machine learning. The task is graph OOD generalization. *Dataset Overview*: The DrugOOD(assay) dataset is derived from the ChEMBL database ...\n *Attribute and Label*: In DrugOOD, a molecule is represented as a graph where graph.x represents the features of the atoms (nodes) in the molecule, including atom type, valence, charge, and hybridization. The label, graph.y represents the ground truth property or label associated with the molecule, which is the target variable for prediction tasks. It describe the activity in a specific assay, assessing the binding affinity between the drug and the target. ...\n *Task Details*: The primary task is to predict whether the drug-target pair will have a high or low affinity ...\n *Environment Description*: Graphs in the same assay are assigned to the same environment.

By learning a continuous vector from $\boldsymbol{t}$ via LLM, the general textual embedding can be denoted as $\boldsymbol{w} = \text{LLM}(\boldsymbol{t}) \in \mathbb{R}^{d_L}$, where d_L is the dimension of LLM token embedding. We align the token embedding from LLM to the feature space of pre-trained GNN and inject the extracted general knowledge as prompt without tuning the LLM, which can be defined as:

$$\boldsymbol{p}_c = f_{MLP}^c(\boldsymbol{w}), \quad \boldsymbol{X}_c = \{\boldsymbol{x}_1 + \boldsymbol{p}_c, \ldots, \boldsymbol{x}_n + \boldsymbol{p}_c\}, \tag{5}$$

where $\boldsymbol{p}_c \in \mathbb{R}^d$, $f_{MLP}^c(\cdot)$ can be implemented as an MLP in practice, $\boldsymbol{X}_c$ denote the prompted feature. And the environment-invariant feature can be $\boldsymbol{z}_c = g(\boldsymbol{X}_c, \boldsymbol{A}) \in \mathbb{R}^{d'}$.

Environment-oriented Prompts for Subgraph Mining. Given the pre-collected environment labels, we focus on inferring the diverse graph patterns as the underlying graph environment. Instead of partitioning the original graph to obtain the causal subgraph [13], we design an environment-disturbed subgraph extractor based on the diverse states of the environment factor. Specifically, we initialize a set of learnable prompts for all environments $\mathcal{E}_{tr}$, denoted as $\boldsymbol{P}_a = \{\boldsymbol{p}_a^e\}_{e=1}^{|\mathcal{E}_{tr}|}$. For each node in the graph, we propagate the environment embedding to disturb the feature:

$$\begin{aligned} \boldsymbol{X}_s &= \boldsymbol{X} + \boldsymbol{P}_s, \ \boldsymbol{P}_s = \text{softmax}\left(\frac{\boldsymbol{X}^Q (\boldsymbol{P}_a^K)^T}{\sqrt{d}}\right)\boldsymbol{P}_a^V, \\ \boldsymbol{X}^Q &= \boldsymbol{X}\boldsymbol{W}^Q, \boldsymbol{P}_a^K = \boldsymbol{P}_a\boldsymbol{W}^K, \boldsymbol{P}_a^V = \boldsymbol{P}_a\boldsymbol{W}^V, \end{aligned} \tag{6}$$

where $\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V \in \mathbb{R}^{d\times d}$ are the trainable weight matrices, $\boldsymbol{P}_s$ denoted the propagated prompt output. In general, we assume that nodes sharing similar prompts should have stronger associations in the environment-disturbed subgraph [56]. Thus, the environment correlation matrix $\boldsymbol{S}$ after prompt propagation can be computed as:

$$\boldsymbol{S}_{uv} = \text{sim}(\boldsymbol{p}_{s,u}, \boldsymbol{p}_{s,v}), \tag{7}$$

where $\boldsymbol{p}_{s,v}$ denote the propagated prompt for node v. Considering t-step association, the environment-disturbed subgraph can be extracted based on the correlation matrix $\boldsymbol{S}$:

$$\boldsymbol{A}_s = \boldsymbol{S} \odot \mathbb{I}(\boldsymbol{A}^t), \tag{8}$$

where $\odot$ denotes the Hadamard product between matrix, $\mathbb{I}$ is an indicator function that assigns 1 if the element is greater than 0, and 0 otherwise. The environment-disturbed feature can be encoded as $\boldsymbol{z}_s = f(\boldsymbol{X}_s, \boldsymbol{A}_s) \in \mathbb{R}^{d'}$.

4.3 Graph Information Bottleneck for Prompt Learning

Inspired by the GIB principle [51], which compresses the graph representation to learn minimal sufficient information for efficient label prediction, we employ the GIB to facilitate the above environment-aware prompt tuning process for graph OOD generalization [6, 13, 24, 57]. Specifically, we seek to maximize the Shannon MI between the encoded environment-invariant features $\boldsymbol{z}_c$ and the environment-disturbed features $\boldsymbol{z}_s$, while minimizing the MI between the $\boldsymbol{z}_c$ and the environment label e of the graph. The objective from the GIB perspective can be formulated as:

$$\max_{\boldsymbol{z}_c} I(\boldsymbol{z}_c, \boldsymbol{z}_s) - \alpha I(\boldsymbol{z}_c, e), \tag{9}$$

where α denotes the Lagrange multiplier. In practice, since the intractability of MI, we transfer to optimize the lower/upper bound of two constraints. For the first term, we turn to Noise Contrastive Estimation (NCE)-based MI lower bound [30], which is a well-known MI estimator, maximizing $I(\boldsymbol{z}_c, \boldsymbol{z}_s)$ can be defined as:

$$\mathcal{L}_{nce} = -\frac{1}{|\mathcal{D}_{tr}|}\sum_{G_i\in\mathcal{D}_{tr}} \log\left(\frac{\exp(\text{sim}(\boldsymbol{z}_{i,c}, \boldsymbol{z}_{i,s})/\tau)}{\sum_{G_{i'}\in\mathcal{B}}\exp(\text{sim}(\boldsymbol{z}_{i,c}, \boldsymbol{z}_{i',s})/\tau)}\right), \tag{10}$$

For the second term, we reformulate the MI as:

$$I(\boldsymbol{z}_c, e) = \iint p(e, \boldsymbol{z}_c)\log\frac{p(e|\boldsymbol{z}_c)}{p(e)}ded\boldsymbol{z}_c. \tag{11}$$

Since $p(e|\boldsymbol{z}_c)$ is intractable, we introduce a variational approximation $q(e|\boldsymbol{z}_c)$ and leverage Kullback-Leibler (KL) divergence D_{KL} to measure their distance. Then we have $D_{KL}(p(e|\boldsymbol{z}_c)||q(e|\boldsymbol{z}_c)) \geq 0$, which leads to:

$$\begin{aligned} I(\boldsymbol{z}_c, e) &\geq \iint p(\boldsymbol{z}_c, e)\log q(e|\boldsymbol{z}_c)d\boldsymbol{z}_c de + H(e) \\ &\geq \iint p(\boldsymbol{z}_c, e)\log q(e|\boldsymbol{z}_c)d\boldsymbol{z}_c de, \end{aligned} \tag{12}$$

where $H(e) = -\int p(e)\log p(e)de \geq 0$ denote the entropy of e. Thus, we reduce to maximizing the cross-entropy loss to minimize the $I(\boldsymbol{z}_c, e)$, formulated as:

$$\mathcal{L}_e = \sum_{i\in\mathcal{D}_{tr}} \text{CE}(e_i, h_e(\boldsymbol{z}_{i,c})), \tag{13}$$

where $h_e(\cdot)$ denotes the environment classifier and e_i is the pre-collected environment label of the graph.

4.4 Overall Optimization

To ensure the encoded environment-invariant feature and environment-disturbed feature, we further add the training loss for the two prompt tuning processes, which can be written as:

$$\mathcal{L}_c = \sum_{i\in\mathcal{D}_{tr}} \text{CE}(y_i, h(\boldsymbol{z}_{i,c})),\ \mathcal{L}_s = \sum_{i\in\mathcal{D}_{tr}} \text{CE}(y_i, h(\boldsymbol{z}_{i,s})). \tag{14}$$

Finally, we formalize the overall objective as follows:

$$\mathcal{L} = \mathcal{L}_c + \beta\mathcal{L}_s + \gamma(\mathcal{L}_{nce} - \alpha\mathcal{L}_e), \tag{15}$$

where β and γ are the hyperparameters which are used to control the weights of each part of the loss, respectively.

4.5 Theoretical Analysis

In this subsection, we present the theoretical analysis of DATE. Specifically, we highlight the importance of environment-aware prompt tuning through the generalization error bound, characterized by the excess risk of the test set. We first introduce some notations and definitions. We aim to learn a powerful GNN model $f^*(\cdot) = h^*(g^*(\cdot))$, where $g^*(\cdot)$ is the graph encoder, and $h^*(\cdot)$ is the classifier. For simplicity, we assume two classes, $y \in \{-1, 1\}$, and define the ±1 loss as $\ell(\cdot) = \ell_{\pm1}(\hat{y}, y) = -y\hat{y}$. Let $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$ denote the training and test sets, respectively. Here, we define $\mathcal{F}_O$ as the function class without environment-aware prompt tuning, and $\mathcal{F}_P$ as the function class with it. It is crucial to note that the function class $\mathcal{F}_O$ is a subset of $\mathcal{F}_P$. The risk function of the trained model $f(\cdot)$ is defined as

$$L(f) = \mathbb{E}_{(G,y)\sim\mathcal{D}_{tr}}[\ell(f(G), y)], \tag{16}$$

and the test risk as

$$L_{te}(f) = \mathbb{E}_{(G,y)\sim\mathcal{D}_{te}}[\ell(f(G), y)]. \tag{17}$$

The empirical risks for the training and test sets are given by

$$\hat{L}(f) = \frac{1}{|\mathcal{D}_{tr}|}\sum_{(G,y)\in\mathcal{D}_{tr}} \ell(f(G), y) \tag{18}$$

and

$$\hat{L}_{te}(f) = \frac{1}{|\mathcal{D}_{te}|}\sum_{(G,y)\in\mathcal{D}_{te}} \ell(f(G), y). \tag{19}$$

The empirical risk minimizers for $\mathcal{F}_O$ and $\mathcal{F}_P$ are defined as

$$\hat{f}_O = \underset{f\in\mathcal{F}_O}{\text{argmin}}\,\hat{L}(f), \quad \hat{f}_P = \underset{f\in\mathcal{F}_P}{\text{argmin}}\,\hat{L}(f). \tag{20}$$

Using the empirical risk minimizers, we can further define the excess risk of $\hat{f}_O$ as follows:

$$R_O = L_{te}(\hat{f}_O) - \inf_{f\in\mathcal{F}} L_{te}(f), \tag{21}$$

and the excess risk of $\hat{f}_P$ as

$$R_P = L_{te}(\hat{f}_P) - \inf_{f\in\mathcal{F}} L_{te}(f), \tag{22}$$

where $\mathcal{F}$ represents the function class containing all possible models. The function $f^* = \text{argmin}_{f\in\mathcal{F}} L_{te}(f)$ is the optimal model that minimizes the test risk. In other words, the function f^* could effectively handle the training and test sets distribution shift with respect

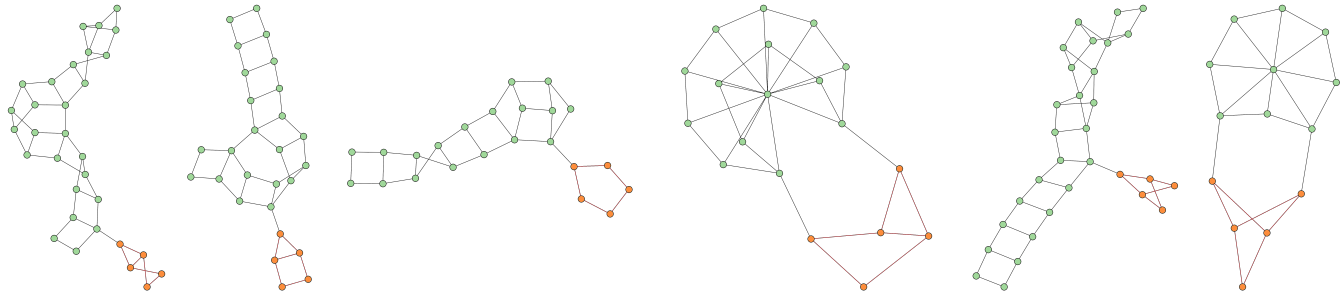

Figure 2: Environment-oriented subgraph mining visualization. This figure shows a selection of results from extracting Environment-oriented Subgraphs on the GOOD-Motif. The motif graph is represented by green nodes, the base graph by orange nodes, and the subgraphs extracted using the Environment-oriented method are indicated by red edges.

to the environment. The excess risk R_O quantifies the generalization error of the corresponding model trained without environment-aware prompt tuning, while R_P quantifies the generalization error of the model trained with the prompt tuning.

Theorem 4.1 (Generalization Error Bound). *Considering a two-class classification problem, where $\mathcal{Y} = \{\pm 1\}$, we define the ± 1 loss, $\ell_{\pm 1}(\hat{y}, y) = -y\hat{y}$. Then, we have*

$$R_O - R_P \geq \mathbb{E}\left[\hat{L}_{te}(\hat{f}_O) - \hat{L}_{te}(\hat{f}_P)\right] - 2\sqrt{\frac{2\log\left(2\Pi_{\mathcal{F}_P}(n)\right)}{n}}. \tag{23}$$

where

$$\Pi_{\mathcal{F}}(n) = \max\left\{\left|\left\{\left(f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)\right) : f \in \mathcal{F}_P\right\}\right| : \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}\right\}. \tag{24}$$

is the VC-dimension of $\mathcal{F}_P$ on $\mathcal{X}$.

Theorem 4.1 establishes a generalization error bound for models trained with and without environment-aware prompt tuning. The bound is determined by the difference in excess risks, $\mathbb{E}\left[\hat{L}_{te}(\hat{f}_O) - \hat{L}_{te}(\hat{f}_P)\right]$. Since $\mathcal{F}_O \subseteq \mathcal{F}_P$, $\mathbb{E}\left[\hat{L}_{tr}(\hat{f}_O) - \hat{L}_{tr}(\hat{f}_P)\right] > 0$. On test data, $\mathbb{E}\left[\hat{L}_{te}(\hat{f}_O) - \hat{L}_{te}(\hat{f}_P)\right]$ is expected to be even larger, as the environment-aware prompt tuning in $\mathcal{F}_P$ promotes better generalization compared to $\mathcal{F}_O$. The bound also incorporates the VC-dimension of $\mathcal{F}_P$ on $\mathcal{X}$, which reflects model complexity. Typically, the VC-dimension depends on the number of model parameters [4]. When the number of parameters is comparable to the sample size, the VC-dimension remains small. In conclusion, Theorem 4.1 provides a theoretical guarantee that models with environment-aware prompt tuning generalize better than those without.

5 Experiments

5.1 Experiment Setup

Datasets. We evaluate our proposed DATE across a diverse set of datasets, including synthetic, semi-synthetic, and real-world datasets, to assess its performance under various shift scenarios. We employ synthetic and semi-synthetic datasets to analyze structure and feature shifts performance. The synthetic dataset, GOOD-Motif, follows the GOOD benchmark and consists of graphs with base and motif subgraphs, where the motif shape defines the classification label. We test two splits: base split for generalization to unseen base subgraphs and size split for generalization from small to large graphs. For feature shift sanity checks, we use the semi-synthetic dataset GOOD-CMNIST from the GOOD benchmark, made of colored MNIST digits transformed via the superpixel technique to test generalization to unseen colors. Real-world performance is evaluated across datasets such as GOOD-HIV for molecular property prediction with scaffold and size splits, DrugOOD LBAP-core-ic50 for assay splits, and natural language processing with GOOD-SST2 and the similarly structured GOOD-Twitter dataset.

Baselines. We compare our DATE against several baselines: Empirical Risk Minimization (ERM); Adaptive Structure Aware Pooling (ASAP) [34], a graph pooling method; four traditional OOD baselines—Invariant Risk Minimization (IRM) [2], Variance Risk Extrapolation (VREx) [21], Domain Adversarial Neural Network (DANN) [10], and Correlation Alignment (Coral) [38]; four graph-specific OOD baselines—Discovering Invariant Rationales (DIR) [52], Graph Stochastic Attention Mechanism (GSAT) [29], Causally Invariant Graph Augmentation (CIGA) [6], and Graph Invariant Learning (GIL) [24]; and two recent graph OOD works: Empowering Graph Invariance Learning with Deep Spurious Infomax (EQuAD) [59] and Joint Learning of Label and Environment Causal Independence for Graph Out-of-Distribution Generalization (LECI) [13].

Implementation. Following previous research, we implement the standard three-layer GIN [53] as our GNN model. We enhance the GIN by incorporating virtual nodes to address the size-shift challenges presented by GOOD-Motif (size), GOOD-CMNIST, and real-world datasets. This additional node aggregates global information that is otherwise difficult to capture. We employed a 200-epoch training on DATE, with the initial 150 epochs serving as a pre-training phase, followed by the subsequent 50 epochs dedicated to prompt-tuning. The temperature parameter for LLMs was set to zero to eliminate output randomness, thereby mitigating instability during the training process.

5.2 Performance Comparison

We conduct a rigorous examination of the OOD performance of our proposed DATE across both real-world and synthetic datasets.

Table 1: Performance comparison on real-world datasets, evaluated using accuracy for the GOOD-SST2 and GOOD-Twitter datasets, and ROC-AUC for the GOOD-HIV and DrugOOD datasets.

Methods	GOOD-SST2		GOOD-Twitter		GOOD-HIV (scaffold)		GOOD-HIV (size)		DrugOOD (assay)	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
ERM	78.37 (2.64)	80.41 (0.69)	54.93 (0.96)	57.04 (1.70)	69.61 (1.32)	70.37 (1.19)	61.66 (2.45)	57.31 (1.06)	70.03 (0.16)	72.18 (0.18)
IRM	79.73 (1.45)	80.17 (1.52)	55.27 (1.19)	57.72 (1.03)	73.35 (2.30)	70.89 (0.29)	58.52 (0.86)	60.86 (2.78)	71.56 (0.32)	72.69 (0.29)
VREx	79.31 (1.40)	80.33 (1.09)	56.46 (0.39)	56.37 (0.76)	71.73 (3.51)	71.18 (0.69)	58.39 (1.54)	60.10 (0.29)	70.22 (0.68)	72.32 (0.58)
Coral	78.24 (3.26)	80.97 (1.07)	56.57 (0.42)	56.41 (1.76)	71.19 (2.92)	71.12 (2.92)	61.48 (4.76)	62.01 (1.05)	70.09 (0.82)	72.26 (0.54)
DANN	78.74 (0.82)	80.36 (0.61)	55.52 (1.27)	55.71 (1.23)	69.88 (3.66)	72.25 (1.59)	61.37 (0.53)	60.04 (1.11)	69.83 (0.95)	72.23 (0.26)
ASAP	78.51 (2.26)	80.44 (0.59)	56.10 (0.56)	56.37 (1.30)	69.97 (2.91)	68.44 (0.49)	61.08 (2.26)	61.54 (2.53)	68.02 (1.22)	71.73 (0.39)
DIR	77.65 (0.71)	81.50 (0.55)	55.32 (1.58)	56.81 (0.91)	65.84 (1.71)	68.59 (3.70)	59.69 (1.69)	60.85 (0.52)	67.29 (0.73)	69.70 (0.65)
GSAT	79.25 (1.09)	80.46 (0.38)	55.09 (0.66)	57.00 (0.53)	71.55 (3.58)	71.39 (1.41)	60.24 (3.88)	62.56 (1.76)	71.01 (0.54)	72.06 (0.45)
CIGA	80.37 (1.46)	81.20 (0.75)	57.51 (1.36)	57.19 (1.15)	66.25 (2.89)	71.47 (1.29)	58.24 (3.78)	62.56 (1.76)	67.68 (1.14)	70.54 (0.59)
LECI	82.93 (0.22)	83.44 (0.27)	59.35 (0.15)	59.64 (0.15)	74.04 (0.65)	74.43 (1.69)	64.83 (2.59)	65.44 (1.78)	72.67 (0.46)	73.45 (0.17)
EQuAD	83.11 (1.97)	82.18 (0.81)	58.71 (1.55)	57.65 (1.37)	75.22 (2.85)	70.97 (2.19)	63.62 (3.72)	62.10 (1.83)	73.20 (1.23)	71.31 (0.52)
Ours	**85.32 (0.33)**	**84.21 (0.51)**	**62.13 (1.27)**	**61.79 (1.09)**	**79.27 (1.89)**	**75.03 (1.12)**	**77.31 (0.78)**	**66.55 (0.68)**	**74.31 (1.07)**	**74.54 (0.62)**

Table 2: Comparison of performance on synthetic datasets. The results are measured by accuracy on OOD validation set.

	GOOD-Motif		GOOD-CMNIST	
	basis	size	color	covariate
ERM	60.93 (11.11)	56.63 (7.12)	26.64 (3.27)	57.56 (9.59)
IRM	64.94 (4.85)	54.52 (3.27)	29.63 (2.06)	58.11 (5.14)
VREx	61.59 (6.58)	55.85 (9.42)	27.13 (2.09)	48.78 (7.81)
Coral	61.95 (10.36)	55.80 (4.05)	29.21 (6.87)	57.56 (9.59)
DANN	50.62 (4.71)	46.61 (3.78)	27.86 (5.07)	57.56 (9.59)
ASAP	45.00 (11.66)	42.23 (4.20)	23.53 (6.07)	57.56 (9.59)
DIR	34.39 (2.02)	43.11 (2.78)	25.32 (2.50)	44.67 (3.01)
GSAT	62.27 (8.79)	50.03 (5.71)	27.06 (5.17)	68.22 (7.23)
CIGA	37.81 (2.42)	51.87 (5.15)	25.06 (3.87	56.78 (2.99)
LECI	84.56 (2.22)	71.43 (1.96)	51.80 (2.70)	83.20 (5.89)
EQuAD	73.10 (3.32)	70.33 (4.36)	52.31 (3.20)	78.99 (1.87)
Ours	**84.77 (2.11)**	**73.28 (1.31)**	**72.01 (2.98)**	**83.53 (3.92)**

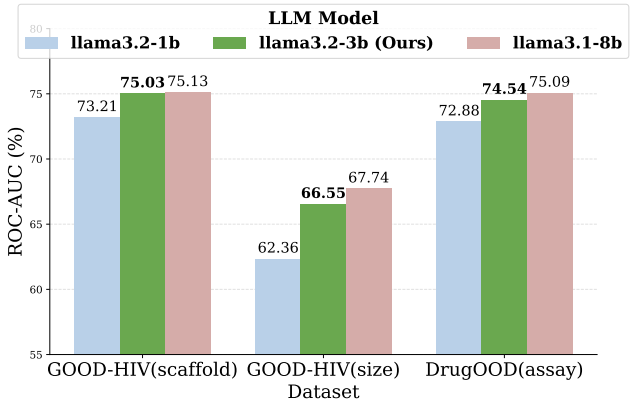

Figure 3: Comparative Performance of LLMs with varying parameter sizes on GOOD-HIV (scaffold), GOOD-HIV (size), and DrugOOD (assay) dataset.

Table 1 details the performance on four real-world datasets, for each dataset, ID and OOD denote the results on ID and OOD validation sets. Table 2 presents the accuracy on two synthetic datasets: GOOD-Motif and GOOD-CMNIST. Our DATE outperforms all baselines on the synthetic datasets under various distribution shift scenarios, including basis, size, and color shifts in GOOD-Motif, and covariate shift in GOOD-CMNIST. These results confirm the efficacy of DATE in handling both structural and feature shifts within synthetic environments. The performance improvement of our method in both real-world and synthetic environments demonstrates its potential to handle complex distribution shifts in different data modalities. The results suggest that our method is particularly effective in real world scenarios with significant distribution shifts. In synthetic data, the consistent outperformance in the face of various shifts, such as size, color, and covariate shifts, also indicates that DATE is robust against different types of distribution shifts.

Ablation Studies. We further perform ablation experiments in this section to evaluate the contribution of the components for our proposed model. In particular, we introduce six model variants as follows: (1) DATE w/o L, which removes the LLM-based prompt generation; (2) DATE w/o E, which removes environment information utilization; (3) DATE w/o I, which removes information bottleneck loss function; (4) DATE w/o V, which removes the prompts for invariant representations; (5) DATE w/o S, which removes the subgraph strategy; (6) DATE w/o C, which removes the cross-entropy loss during prompt-tuning phase.

We present the results in Table 3. These results indicate that each component contributes positively, highlighting their synergistic integration, such that the removal of any single module can disrupt the model's theoretical guarantees for generalization. Both LLM-based prompt generation (L) and environment subgraph utilization (E, S) contribute significantly to performance. Environment information (E) plays a particularly crucial role for OOD generalization; its removal leads to greater performance degradation than removing the cross-entropy loss (C). This is because lacking valid environment information fundamentally destabilizes the training process for separating invariant and variant features (similar poor

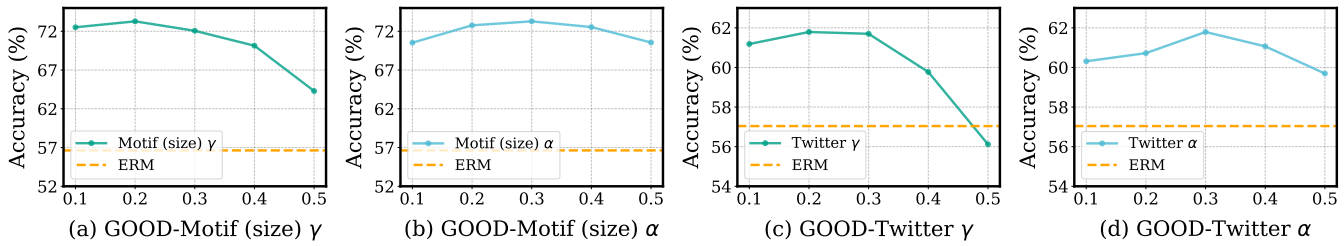

Figure 4: Parameter study on GOOD-Motif and GOOD-Twitter dataset. The impact of hyperparameter variations on the results shows an almost consistent trend.

Table 3: Results of Ablation studies in terms of ROC-AUC on different datasets. Evaluate on validation set.

Model	GOOD-HIV (scaffold)	GOOD-HIV (size)	DrugOOD (assay)
DATE w/o L	72.97 (↓ 2.06)	61.41 (↓ 5.14)	71.32 (↓ 3.22)
DATE w/o E	68.48 (↓ 6.55)	60.84 (↓ 5.73)	70.33 (↓ 4.21)
DATE w/o I	70.59 (↓ 4.44)	60.17 (↓ 5.21)	70.52 (↓ 4.02)
DATE w/o V	69.73 (↓ 5.30)	60.71 (↓ 5.84)	68.11 (↓ 6.43)
DATE w/o S	72.63 (↓ 2.40)	63.38 (↓ 3.17)	73.17 (↓ 1.37)
DATE w/o C	73.58 (↓ 1.45)	64.08 (↓ 2.47)	72.50 (↓ 2.04)
DATE (Full Model)	**75.03**	**66.55**	**74.54**

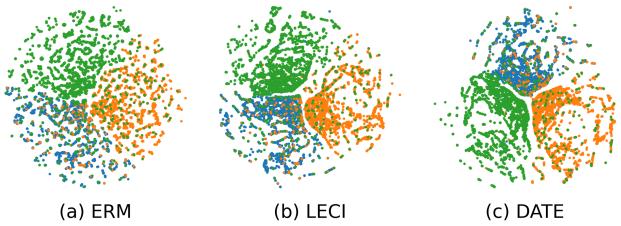

Figure 5: UMAP visualization of graph features on the GOOD-Motif dataset.

performance was observed when using randomized environment labels), whereas the inherent structure guided by LLM task prompts retains partial classification capacity even without direct CE loss optimization. The dual-branch structure (implicitly involving V and I) is effective for capturing distinct feature types. Subgraph extraction (S) further refines environment information utilization. Finally, the information bottleneck loss function (I) positively influences performance by promoting the disentanglement of feature representations, contributing to the model's robustness.

Parameter Sensitivity. We conduct a systematic examination to quantify the model's parametric sensitivity with respect to critical hyperparameter configurations, including the global regularization magnitude γ and the optimization weights α. We tested our model on GOOD-Motif and GOOD-Twitter. The results are shown in Figure 4. Experiments show that an overly large γ degrades performance due to excessive information compression, with an optimal value around 0.2. Parameter α's optimal value is around 0.3, balancing two objectives: first, to maximize the mutual information shared by environment-invariant and variant features; and second, to minimize the mutual information between the invariant features and the corresponding environment label.

Subgraph Mining Visualization. As illustrated in Figure 2, DATE accurately extracts environment-related subgraphs from complex graph structures. This extraction process is guided by environment-oriented prompts, which isolate the substructures within the graph that are most pertinent to specific environmental contexts. The efficacy of this subgraph extraction is important for the extraction of environment-invariant features. Subgraph mining disentangle environment-specific factors from the underlying core features of the graph, leveraging the information bottleneck principle to isolate and refine environment-invariant features.

LLM Scale Evaluation. To evaluate the performance of DATE with LLMs of varying scales, we conducted experiments using LLMs of varying scales (1B, 3B and 8B parameters). The results in Figure 3 indicate that larger LLMs achieve higher performance on the OOD set. Although both models remain functional, the 1B variant exhibits significantly larger performance degradations than its 3B counterpart. While performance improves with increasing scale, the 3B model presents a favorable balance between effectiveness and efficiency. Consequently, we chose this model for integration into our framework.

UMAP Visualization. We apply UMAP to the features extracted from the GOOD-Motif dataset to elucidate the feature space learned by different models. As shown in Figure 5, our DATE produces features where these categories are clearly separated with minimal overlap. These observation indicating DATE can extract higher degree of intra-class similarity and a tighter clustering of features belonging to the same category.

6 Conclusion

In this paper, we have presented DATE, a novel approach for addressing graph out-of-distribution generalization through dual prompt learning with information bottleneck. The key contributions of our work include: a contrastive pretraining strategy with pretext tokens, the introduction of dual prompts leveraging LLMs for environment-invariant representations, and a graph information bottleneck framework that optimally balances semantic preservation with environment independence. Our results on multiple public benchmark datasets further demonstrate that our proposed DATE consistently outperforms state-of-the-art methods. In future works, we plan to extend our proposed DATE to more complicated scenarios such as multimodal graphs and protein function analysis.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (62276058, 41774063), the Fundamental Research Funds for the Central Universities (N25GFZ011) and the Fundamental Research Funds for the Central Universities in UIBE (Grant No. 23QN02).

References

[1] Sravanti Addepalli, Ashish Ramayee Asokan, Lakshay Sharma, and R Venkatesh Babu. 2024. Leveraging Vision-Language Models for Improving Domain Generalization in Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23922–23932.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant Risk Minimization.

[3] Lars Backstrom and Jure Leskovec. 2011. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the International ACM Conference on Web Search & Data Mining*. 635–644.

[4] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. 2021. Deep learning: a statistical viewpoint. *Acta Numerica* 30 (2021), 87 – 201. https://api.semanticscholar.org/CorpusID:232240630

[5] Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel. 2005. Protein function prediction via graph kernels. *Bioinformatics* 21, suppl_1 (2005), i47–i56.

[6] Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, Kaili Ma, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. 2022. Learning Causally Invariant Representations for Out-of-Distribution Generalization on Graphs. In *Proceedings of the Conference on Neural Information Processing Systems*.

[7] De Cheng, Zhipeng Xu, Xinyang Jiang, Nannan Wang, Dongsheng Li, and Xinbo Gao. 2024. Disentangled Prompt Representation for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23595–23604.

[8] Eulrang Cho, Jooyeon Kim, and Hyunwoo J Kim. 2023. Distribution-Aware Prompt Tuning for Vision-Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22004–22013.

[9] Taoran Fang, Yunchao Zhang, Yang Yang, Chunping Wang, and Lei Chen. 2023. Universal prompt tuning for graph neural networks. In *Proceedings of the Conference on Neural Information Processing Systems*, Vol. 36.

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-Adversarial Training of Neural Networks.

[11] Thomas Gaudelet, Ben Day, Arian R Jamasb, Jyothish Soman, Cristian Regep, Gertrude Liu, Jeremy BR Hayter, Richard Vickers, Charles Roberts, Jian Tang, et al. 2021. Utilizing graph machine learning within drug discovery and development. *Briefings in bioinformatics* 22, 6 (2021), bbab159.

[12] Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. 2022. Good: A graph out-of-distribution benchmark. In *Proceedings of the Conference on Neural Information Processing Systems*. 2059–2073.

[13] Shurui Gui, Meng Liu, Xiner Li, Youzhi Luo, and Shuiwang Ji. 2023. Joint learning of label and environment causal independence for graph out-of-distribution generalization. In *Proceedings of the Conference on Neural Information Processing Systems*, Vol. 36.

[14] Kai Guo, Hongzhi Wen, Wei Jin, Yaming Guo, Jiliang Tang, and Yi Chang. 2024. Investigating out-of-distribution generalization of GNNs: An architecture perspective. In *Proceedings of the International ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 932–943.

[15] Xiaotian Han, Zhimeng Jiang, Ninghao Liu, and Xia Hu. 2022. G-mixup: Graph data augmentation for graph classification. In *Proceedings of the International Conference on Machine Learning*. 8230–8248.

[16] Zhongkai Hao, Chengqiang Lu, Zhenya Huang, Hao Wang, Zheyuan Hu, Qi Liu, Enhong Chen, and Cheekong Lee. 2020. ASGN: An active semi-supervised graph neural network for molecular property prediction. In *Proceedings of the International ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 731–752.

[17] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: datasets for machine learning on graphs. In *Proceedings of the Conference on Neural Information Processing Systems*. 16 pages.

[18] Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Lanqing Li, Long-Kai Huang, Tingyang Xu, Yu Rong, Jie Ren, Ding Xue, et al. 2023. Drugood: Out-of-distribution dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8023–8031.

[19] Tianrui Jia, Haoyang Li, Cheng Yang, Tao Tao, and Chuan Shi. 2024. Graph invariant learning with subgraph co-mixup for out-of-distribution generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, Article 952.

[20] Licheng Jiao, Yuhan Wang, Xu Liu, Lingling Li, Fang Liu, Wenping Ma, Yuwei Guo, Puhua Chen, Shuyuan Yang, and Biao Hou. 2024. Causal Inference Meets Deep Learning: A Comprehensive Survey. *Research* 7 (2024), 0467. arXiv:https://spj.science.org/doi/pdf/10.34133/research.0467 doi:10.34133/research.0467

[21] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *Proceedings of the International Conference on Machine Learning*. 5815–5826.

[22] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).

[23] Hourun Li, Yifan Wang, Zhiping Xiao, Jia Yang, Changling Zhou, Ming Zhang, and Wei Ju. 2025. DisCo: graph-based disentangled contrastive learning for cold-start cross-domain recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 12049–12057.

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

[25] Mingkai Lin, Wenzhong Li, Ding Li, Yizhou Chen, Guohao Li, and Sanglu Lu. 2023. Multi-Domain Generalized Graph Meta Learning. 37, 4 (2023), 4479–4487.

[26] Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. 2022. ZIN: When and how to learn invariance without environment partition?. In *Proceedings of the Conference on Neural Information Processing Systems*. 24529–24542.

[27] Gang Liu, Tong Zhao, Jiaxin Xu, Tengfei Luo, and Meng Jiang. 2022. Graph rationalization with environment-based augmentations. In *Proceedings of the International ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1069–1078.

[28] Zemin Liu, Xingtong Yu, Yuan Fang, and Xinming Zhang. 2023. GraphPrompt: Unifying Pre-Training and Downstream Tasks for Graph Neural Networks. In *Proceedings of the Web Conference*.

[29] Siqi Miao, Mia Liu, and Pan Li. 2022. Interpretable and Generalizable Graph Learning via Stochastic Attention Mechanism. *Proceedings of the International Conference on Machine Learning* (2022).

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[31] Yinhua Piao, Sangseon Lee, Yijingxiu Lu, and Sun Kim. 2024. Improving out-of-distribution generalization in graphs via hierarchical semantic environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27631–27640.

[32] Egor Shikov Polina Andreeva and Claudie Bochenia. 2022. 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)*.

[33] Xiaoru Qu, Yifan Wang, Zhao Li, and Jun Gao. 2024. Graph-enhanced prompt learning for personalized review generation. *Data Science and Engineering* 9, 3 (2024), 309–324.

[34] Ekagra Ranjan, Soumya Sanyal, and Partha Pratim Talukdar. 2019. ASAP: Adaptive Structure Aware Pooling for Learning Hierarchical Graph Representations. *arXiv preprint arXiv:1911.07979* (2019).

[35] Tao Ren, Haodong Zhang, Yifan Wang, Wei Ju, Chengwu Liu, Fanchun Meng, Siyu Yi, and Xiao Luo. 2025. MHGC: Multi-scale hard sample mining for contrastive deep graph clustering. *Information Processing & Management* 62, 4 (2025), 104084.

[36] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. 2020. Learning to simulate complex physics with graph networks. In *Proceedings of the International Conference on Machine Learning*. 8459–8468.

[37] Yongduo Sui, Qitian Wu, Jiancan Wu, Qing Cui, Longfei Li, Jun Zhou, Xiang Wang, and Xiangnan He. 2023. Unleashing the power of graph data augmentation on covariate distribution shift. In *Proceedings of the Conference on Neural Information Processing Systems*. 18109–18131.

[38] Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of Frustratingly Easy Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[39] Mingchen Sun, Kaixiong Zhou, Xin He, Ying Wang, and Xin Wang. 2022. Gppt: Graph pre-training and prompt tuning to generalize graph neural networks. In *Proceedings of the International ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1717–1727.

[40] Mingchen Sun, Kaixiong Zhou, Xin He, Ying Wang, and Xin Wang. 2022. GPPT: Graph Pre-training and Prompt Tuning to Generalize Graph Neural Networks. In *Proceedings of the International ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1717–1727.

[41] Qixun Wang, Yifei Wang, Yisen Wang, and Xianghua Ying. 2024. Dissecting the Failure of Invariant Learning on Graphs. In *Proceedings of the Conference on Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 80383–80438.

[42] Yifan Wang, Hourun Li, Ling Yue, Zhiping Xiao, Jia Yang, Changling Zhou, Wei Ju, Ming Zhang, and Xiao Luo. 2025. DANCE: Dual Unbiased Expansion with Group-acquired Alignment for Out-of-distribution Graph Fairness Learning. In *Proceedings of the International Conference on Machine Learning*.

[43] Yifan Wang, Yifang Qin, Yu Han, Mingyang Yin, Jingren Zhou, Hongxia Yang, and Ming Zhang. 2022. Ad-aug: Adversarial data augmentation for counterfactual recommendation. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 474–490.

[44] Yifan Wang, Jianhao Shen, Yiping Song, Sheng Wang, and Ming Zhang. 2022. HE-SNE: Heterogeneous event sequence-based streaming network embedding for dynamic behaviors. In *2022 International Joint Conference on Neural Networks (IJCNN)*. 1–8.

[45] Yifan Wang, Suyao Tang, Yuntong Lei, Weiping Song, Sheng Wang, and Ming Zhang. 2020. Disenhan: Disentangled heterogeneous graph attention network for recommendation. In *Proceedings of the International Conference on Information and Knowledge Management*. 1605–1614.

[46] Yifan Wang, Yangzi Yang, Shuai Li, Yutao Xie, Zhiping Xiao, Ming Zhang, and Wei Ju. 2025. GMR-Rec: Graph mutual regularization learning for multi-domain recommendation. *Information Sciences* 703 (2025), 121946.

[47] Zeyuan Wang, Qiang Zhang, HU Shuang-Wei, Haoran Yu, Xurui Jin, Zhichen Gong, and Huajun Chen. 2022. Multi-level protein structure pre-training via prompt learning. In *Proceedings of the International Conference on Learning Representations*.

[48] Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. 2022. Handling distribution shifts on graphs: An invariance perspective. In *Proceedings of the International Conference on Learning Representations*.

[49] Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. 2022. Graph neural networks in recommender systems: a survey. *Comput. Surveys* 55, 5 (2022), 1–37.

[50] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based Recommendation with Graph Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 346–353.

[51] Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. 2020. Graph information bottleneck. In *Proceedings of the Conference on Neural Information Processing Systems*. 20437–20448.

[52] Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat seng Chua. 2022. Discovering Invariant Rationales for Graph Neural Networks. In *Proceedings of the International Conference on Learning Representations*.

[53] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How powerful are graph neural networks?. In *Proceedings of the International Conference on Learning Representations*.

[54] Bencheng Yan and Chaokun Wang. 2020. GraphAE: Adaptive Embedding across Graphs. In *Proceedings of the IEEE Conference on Data Engineering*. 1958–1961.

[55] Chenxiao Yang, Qitian Wu, Jiahua Wang, and Junchi Yan. 2023. Graph neural networks are inherently good generalizers: Insights by bridging gnns and mlps. In *Proceedings of the International Conference on Learning Representations*.

[56] Kuo Yang, Zhengyang Zhou, Qihe Huang, Limin Li, Yuxuan Liang, and Yang Wang. 2024. Improving Generalization of Dynamic Graph Learning via Environment Prompt. In *Proceedings of the Conference on Neural Information Processing Systems*.

[57] Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. 2022. Learning substructure invariance for out-of-distribution molecular representations. *Advances in Neural Information Processing Systems* 35 (2022), 12964–12978.

[58] Xihong Yang, Yue Liu, Sihang Zhou, Siwei Wang, Wenxuan Tu, Qun Zheng, Xinwang Liu, Liming Fang, and En Zhu. 2023. Cluster-guided contrastive graph clustering network. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 9 pages.

[59] Tianjun Yao, Yongqiang Chen, Zhenhao Chen, Kai Hu, Zhiqiang Shen, and Kun Zhang. 2024. Empowering Graph Invariance Learning with Deep Spurious Infomax. In *Proceedings of the International Conference on Machine Learning*.

[60] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. In *Proceedings of the Conference on Neural Information Processing Systems*. 5812–5823.

[61] Junchi Yu, Jian Liang, and Ran He. 2023. Mind the label shift of augmentation-based graph ood generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11620–11630.

[62] Xingtong Yu, Zhenghao Liu, Yuan Fang, Zemin Liu, Sihong Chen, and Xinming Zhang. 2024. Generalized graph prompt: Toward a unification of pre-training and downstream tasks on graphs. *IEEE Transactions on Knowledge and Data Engineering* (2024).

[63] Xingtong Yu, Chang Zhou, Yuan Fang, and Xinming Zhang. 2024. MultiGPrompt for multi-task pre-training and prompting on graphs. In *Proceedings of the Web Conference*. 515–526.

[64] Chu Zhao, Enneng Yang, Yuliang Liang, Pengxiang Lan, Yuting Liu, Jianzhe Zhao, Guibing Guo, and Xingwei Wang. 2025. Graph representation learning via causal diffusion for out-of-distribution recommendation. In *Proceedings of the ACM on Web Conference 2025*. 334–346.

[65] Chu Zhao, Enneng Yang, Yuliang Liang, Jianzhe Zhao, Guibing Guo, and Xingwei Wang. 2025. Distributionally robust graph out-of-distribution recommendation via diffusion model. In *Proceedings of the ACM on Web Conference 2025*. 2018–2031.

[66] Qi Zhu, Natalia Ponomareva, Jiawei Han, and Bryan Perozzi. 2021. Shift-robust gnns: Overcoming the limitations of localized graph training data. In *Proceedings of the Conference on Neural Information Processing Systems*. 27965–27977.