ENHANCING GRAPH INVARIANT LEARNING FROM A NEGATIVE INFERENCE PERSPECTIVE

Anonymous authors
Paper under double-blind review

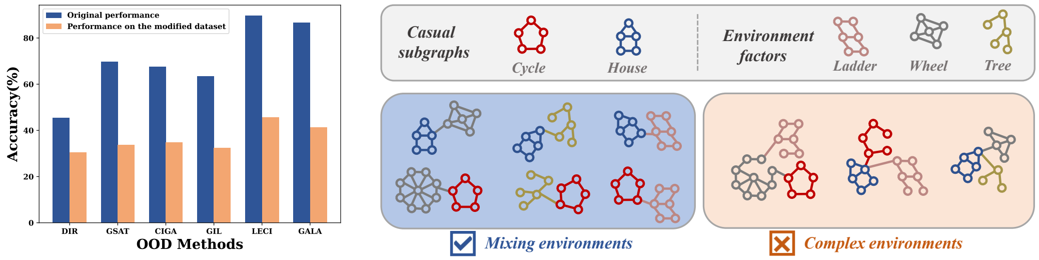
ABSTRACT

The out-of-distribution (OOD) generalization challenge is a longstanding problem in graph learning. Through studying the fundamental cause of data distribution shift, i.e., the changes of environments, significant progress has been achieved in addressing this issue. However, we observe that existing works still fail to effectively address complex environment shifts. Previous practices place excessive attention on extracting causal subgraphs, inevitably treating spurious subgraphs as environment variables. While spurious subgraphs are controlled by environments, the space of environment changes encompass more than the scale of spurious subgraphs. Therefore, existing efforts have a limited inference space for environments, leading to failure under severe environment changes. To tackle this issue, we propose a negative inference graph OOD framework (NeGo) to broaden the inference space for environment factors. Inspired by the successful practice of prompt learning in capturing underlying semantics and causal associations in large language models, we design a negative prompt environment inference to extract underlying environment information. We further introduce the environment-enhanced invariant subgraph learning method to effectively exploit inferred environment embedding, ensuring the robust extraction of causal subgraph in the environment shifts. Lastly, we conduct a comprehensive evaluation of NeGo on real-world datasets and synthetic datasets across domains. NeGo outperforms baselines on nearly all datasets, which verify the effectiveness of our framework. Our source code is available at `https://anonymous.4open.science/r/NeGo-E4C1`.

1 INTRODUCTION

Graph Neural Networks (GNNs) have emerged as the predominant approach for encoding graph data Kipf & Welling (2016); Xu et al. (2018), delivering notable achievements in various research fields including molecular property prediction Jumper et al. (2021); Yang et al. (2022), recommendation systems Wu et al. (2022b); Gao et al. (2022), and traffic flow forecasting Liang et al. (2018); Zhou et al. (2020). However, as real-world data is evolving with complex patterns, the challenge of data distribution shift has become a major obstacle for GNNs Gui et al. (2022); Ji et al. (2022); Wang et al. (2023); Zhou et al. (2022b); Zou et al. (2023). Therefore, various studies concentrate on improving the Out-Of-Distribution (OOD) generalization ability of graph learning models Chen et al. (2024; 2022); Gui et al. (2024); Miao et al. (2022); Sui et al. (2022); Li et al. (2022); Wu et al. (2022c).

Recently, environment-centered invariant learning methods achieved impressive OOD generalization performance with the aim of inferring underlying environment factors in data Chen et al. (2024); Gui et al. (2024); Xia et al. (2023); Yuan et al. (2023). Those efforts demonstrate that the changes of environment are the fundamental reason for the shift of data distribution Grice & White (1961); Liu et al. (2021); Peters et al. (2016). However, existing approaches still lack the ability to decouple causal subgraphs from complex environments. As shown in Fig. 1(a), we double the scale of spurious substructures in the *SPURIOUS-MOTIF(0.5)*, and observe a significant decrease in the performance of current methods when they are re-conducted on this modified dataset. The reason lies in that current methods, even those claiming to model environments, focus much of their attention on extracting causal subgraphs Chen et al. (2022); Wu et al. (2022a;c). This results in the model being able to extract causal subgraphs only in known environments, leading to failures in unseen



(a) Degradation in performance.

(b) Prediction failure in complex environments.

Figure 1: The motivation of our work. (a) We double the scale of spurious substructures in the *SPURIOUS-MOTIF(0.5)* Ying et al. (2019), and observe a significant decrease in the performance of current methods when they are re-conducted on this modified dataset. (b) The OOD methods, which treat spurious subgraphs as the environments, fail to address the shift of complex environments.

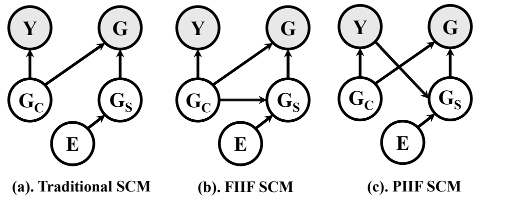
complex environments. Therefore, this poses a challenging research question: *how to broaden the inference space of environments, enabling model to handle complex environment shifts.*

We argue this limitation arises from the positive learning paradigm that focuses solely on extracting causal subgraphs as its primary objective. In contrast, negative inference paradigm, modeling the sample space except the invariant subgraph as environments, has the potential to broaden the perception scope of environment. As shown in Fig. 1(b), the positive inference can only infer the specific *ladder*, *wheel*, and *tree* as environment variables, while the negative inference approach can infer all variable space except the *cycle* and *house* as environments. However, the inaccessibility of environment information pose challenges to implementing such negative inference. Specifically, *(1) how to formulate the negative inference learning to achieve environment awareness*, and *(2) how to utilize environment information for facilitating causal invariant learning*.

In this work, we propose a novel **Ne**gative inference **G**raph **O**OD framework (NeGo). NeGo aims to achieve causal invariant learning against complex environment shifts by a negative inference. **Firstly**, we design a negative prompt learning framework for inferring underlying environment factors. We model all other class samples, i.e., extra-class samples, as the environment space for the current graph. This design enables the model to capture a broader scale of environments, no more limiting to in-sample spurious subgraphs. **Secondly**, we introduce an environment-enhanced invariant learning strategy to effectively utilize inferred environment variables. Specifically, we design an interactive decoding scheme that utilizes an attention-based residual connection architecture to encapsulate environment embedding into node representations. Different from traditional approaches that neglect the information of environment variables during subgraph extraction Chen et al. (2024); Gui et al. (2024), our design incorporates the underlying environment patterns into the process of invariant subgraph learning. **Lastly**, we conduct a comprehensive evaluation of NeGo on real-world datasets across domains, and synthetic datasets. NeGo outperforms baselines on nearly all datasets. Our **contributions** can be summarized as follows:

- We observe that existing environment-centered OOD practices encounter difficulties in handling complex environment shifts. Through a comprehensive investigation, we identify that limited environment awareness space of positive inference is the main reason to restrict the generalization capacity of existing OOD approaches.
- We propose a novel invariant learning framework with negative inference NeGo. To be specific, we design an innovative environment inference strategy via negative inference, which effectively broadens the inference space of environment factors. Moreover, we introduce an attention-based residual connection to offer our model with the ability to resist complex environment shifts.
- We conduct extensive experiments on both synthetic and real-world datasets with distribution shifts to evaluate the performance of NeGo. The results from both visualization and quantitative analysis indicate that our framework successfully achieves accurate prediction in complex environmental scenarios, a performance not accomplished by existing methods.

2 Background



(a). Traditional SCM (b). FIIF SCM (c). PIIF SCM

Figure 2: Illustrations of three structural causal models (SCMs).

Preliminaries. A graph is denoted as $G = (\mathcal{X}, \mathcal{A}) \in \mathcal{G}$, where $\mathcal{G}$ is the observed graph dataset. $\mathcal{A} \in \mathbb{R}^{N \times N}$ represents the adjacency matrix and $\mathcal{X} \in \mathbb{R}^{N \times d}$ denotes node features, where N indicates the number of nodes and d is the feature dimension. Each graph is associated with a corresponding label Y. From the perspective of causal theory, the graph data can be partitioned into a spurious subgraph G_S and a causal subgraph G_C, where G_C directly determines its label Y. The spurious subgraph G_S is controlled by the spurious variable C, while the causal subgraphs G_C is controlled by the causal invariant factor C, as shown in Fig. 2. Based on the different interdependencies among C, S and Y, structural causal models (SCMs) can be further classified into *Full Informative Invariant Features (FIIF)* and *Partially Informative Invariant Features (PIIF)* Ahuja et al. (2021); Chen et al. (2022).

Problem definition. Our work aims to address the limitations of existing approaches in handling complex data distribution shifts. We specifically focus on broadening the inference scope of environments, enabling the network to handle intricate scenarios of environment shifts. Additionally, our framework is required to effectively tackle both FIIF and PIIF assumptions.

Comparisons to recent environment-centered OOD practices. Environment-centered studies Chen et al. (2024); Gui et al. (2024); Li et al. (2022); Wu et al. (2022a); Yang et al. (2022) consider that the data distribution shifts stem from the changes of environments. To tackle the limitation of existing works failing to handle the shifts of complex environments, we propose a negative inference to broaden the inference space for environments. Our approach, which represents a pioneering practice in utilizing negative inference, is distinct from all existing practices in this field. GALA Chen et al. (2024) utilized proxy prediction mechanism to infer environment label. The negative samples mentioned in Chen et al. (2024) serve as proxies for spurious subgraphs, while our negative inference aim to capture broader environment variables beyond spurious subgraphs. Therefore, GALA essentially follows the positive inference process with the main goal of extracting the causal subgraph, failing to infer the entire environment space. LECI Gui et al. (2024) focuses on studying the variations of spurious substructures to model the environment variables. Such environment inference strategy still relies on a positive inference with narrow cognitive space of the environments.

Environment inference with negative prompt. Our negative prompter is proposed to achieve a broader inference scale of environments, which is inspired by the success of prompt learning in language models Brown et al. (2020); Gao et al. (2020). Prompt learning is designed to capture underlying semantic knowledge in language data, which improves the generalization ability of models by introducing appropriate prompt tokens to guide the network learn desired answers Rao et al. (2022); Sordoni et al. (2024); White et al. (2023); Sun et al. (2023). For example, in the semantic emotion classification task, the language model constructs a template such as "`the emotion expressed by this sentence is [class]`", where `[class]` is trained to learn real label. In a similar way, our framework can be viewed as constructing a set of text prompts such as "`the underlying environments of current sample are [answer]`", where `[answer]` can be guided to capture the real environment states. Different from random data augmentation techniques Han et al. (2022); Li et al. (2021); Lu et al. (2024); Rong et al. (2019); Wang et al. (2021); You et al. (2020); Zhao et al. (2021) and distributionally robust optimization (DRO) methods Staib & Jegelka (2019); Wu et al. (2024); Zhu et al. (2021), our prompt-based approach not only broadens the scale of environment inference but also deepens the understanding of underlying data generation process. Existing methods always expand the inference boundary of the model by incorporating stochastic perturbations. However, the introduction of randomness prevents the model from capturing the underlying semantics and hinders its ability to deepen the understanding of generation process. In contrast, our prompt-based approach allows us to deeply study the underlying casual correlation of variables, which is the reason we adopt the technique of prompt learning. More discussion about related works can be found in Appendix B.

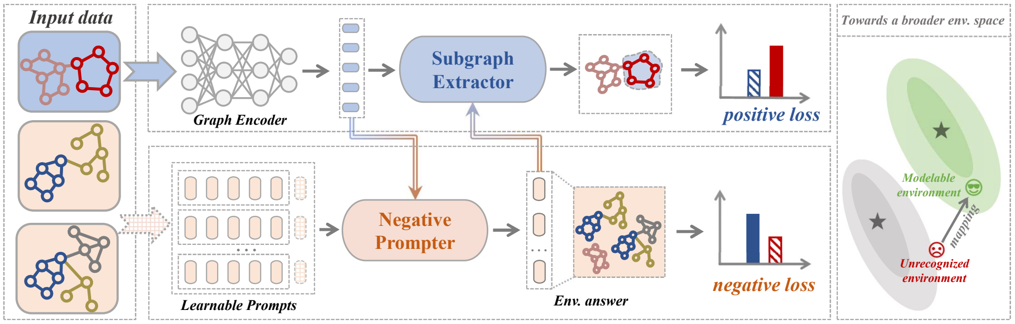


Figure 3: The architecture of NeGo. We implements an environment-enhanced graph learning framework in which the environment is extracted through a negative prompt mechanism. The training process is guided by both a positive loss and a negative loss, aiming to broaden the modeling space for the environment.

3 Graph OOD generalization via environment inference

Existing environment-centered practices aim to enable the networks with the ability to resist data distribution shifts. However, our empirical observations indicate that these approaches are insufficient in handling complex environment shifts. To address this issue, we conduct a theoretical analysis of these methods, and identify that their failures stem from the limited environment inference space of positive inference by treating spurious subgraphs as environment variables. Further, we propose a promising method based on negative inference.

3.1 Limited Environment Cognitive Space for Positive Inference

From the perspective of causal theory Pearl (2009; 2010), the variables of generating the graph data include causal subgraph G_C and spurious subgraph G_S, where G_S is controlled by environment variable E. As shown in Fig. 2, $G_C \to Y$ demonstrates a stable casual relationship from G_C to Y in the data generation process. Consequently, the distribution shift between the training data and the test data can be attributed to the shifts of environment E, which can be formally expressed as $\mathbb{P}_{train}(\mathcal{G}, E) \neq \mathbb{P}_{test}(\mathcal{G}, E)$. Modeling environment variables becomes crucial for tackling OOD generalization issue Chen et al. (2024); Gui et al. (2024); Xia et al. (2023); Yuan et al. (2023). With the observed training dataset $\mathcal{G}$, environment-centered approaches strive to learn the distribution of the environment factor E,

$$\mathbb{P}(E|\mathcal{G}) = \frac{\mathbb{P}(\mathcal{G}, E)}{\mathbb{P}(\mathcal{G})} = \frac{\mathbb{P}(\mathcal{G}|E)\mathbb{P}(E)}{\int\limits_{E} \mathbb{P}(\mathcal{G}|E)\mathbb{P}(E)dE}. \tag{1}$$

The prior distribution $\mathbb{P}(E)$ and the likelihood $\mathbb{P}(\mathcal{G}|E) = \prod_{i=1}^{N} \mathbb{P}(G_i|E)$ make the numerator theoretically computable. However, due to the uncertainty in the scale of environments E, the denominator of Eq. 1 involving integration becomes intractable. To tackle this issue, existing works presuppose an distribution shift boundary based on environment mixing assumption Li et al. (2022).

Assumption 3.1. If K different environment labels can be extracted from the observed dataset $\mathcal{G}$, they are formulated by K independent D-dimensional Gaussian distributions $\mathcal{N}(\mu_i, I)$, where $\mu_i \in \mathbb{R}^{1\times D}$. Therefore, environment variables can be modeled from a vector space perspective, allowing us to approximate the environment space by exploring the mixture space of vectors.

Given the Assumption 3.1, we can model the environments codebook $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_K) \in \mathbb{R}^{K\times D}$. This environment codebook serves as a proxy for the environment space, representing the entire environment space through the mixture of vectors. This principle can be expressed formally as Proposition 3.2.

Proposition 3.2. *The scale of environments is modeled as a mixing space of extracted environment variables. As a result, the new data G_i is associated with the environment state $E_i \sim \mathcal{N}(e_i \cdot \boldsymbol{\mu}, I)$, where $e_i \in \mathbb{R}^{1\times K}$ representing the mixing weight.*

Proposition 3.2 indicates that the latent variables $\boldsymbol{e} = (e_1, e_2, ..., e_N) \in \mathbb{R}^{N\times K}$ be regarded as the proxy factor for the environment variable E, directly determining the observed data $\mathcal{G}$ generation process. The posterior probability of the environments $\mathbb{P}(E|\mathcal{G})$ is then transformed into,

$$\mathbb{P}(\boldsymbol{e}|\mathcal{G}) = \frac{\mathbb{P}(\mathcal{G}, \boldsymbol{e})}{\mathbb{P}(\mathcal{G})} = \frac{\prod_{i=1}^{N} \mathbb{P}(e_i)\mathbb{P}(G_i|e_i)}{\sum_{e} \mathbb{P}(\boldsymbol{e}) \prod_{i=1}^{N} \mathbb{P}(G_i|e_i)} = \frac{\prod_{i=1}^{N} \mathbb{P}(e_i)}{\sum_{e} \mathbb{P}(\boldsymbol{e})}. \tag{2}$$

The finite space of $\boldsymbol{e}$ allows for the approximation strategy to be feasible. However, the limited scale of $\boldsymbol{e}$ may limit the capacity of model to effectively counter complex environment shifts, which is verified by our empirical results. We next delve into the reason contributing to this limitation. We first present a definition of the *basis* and *base environments* of the environment space, similar to the concepts of basis and base vectors in the vector space.

Definition 3.3. *Let $\boldsymbol{E}_b = \{E_1, ..., E_K\}$ be the basis of environment space, and each element E_i within it is referred to as the base environment. The linear combination of base environments can completely describe the entire environment space.*

Actually, the environment mixing assumption fundamentally relies on the expectation that extracted environment codebook can cover the basis $\boldsymbol{E}_b$. However, we observe that such goal cannot be achieved by existing methods. Given a graph G, current environment-centered methods aim to decompose it into causal subgraph G_C and spurious subgraph G_S. The spurious subgraph G_S is inferred as the environment variable. Although G_S is controlled by environment factor ($E \to G_S$), the space of environment changes encompass more than the scale of spurious subgraphs. For example, consider the substructure G_C that is causally associated with one graph-level property l, but the variants of such G_C act as environment factors for other properties. Existing methods that treat spurious subgraphs as environments cannot capture such scenario.

Theorem 3.4. *Given an observed graph dataset $\mathcal{G}$, the inference process, considering G_S as the environment factor, fails to capture the basis $\boldsymbol{E}_b$ that can represent the entire environment space.*

Theorem 3.4 indicates that the mixing of $\boldsymbol{\mu}$ is unable to encompass the entire environment space. Therefore, the existing environment-centered methods have a narrow understanding space of the environments, which leads to the network incapable to extract the causal graph from the complex environments. Detailed proof can be found in Appendix C.1. Therefore, the limitations of existing works are attributed to the narrow inference space of the model for environment variables.

3.2 The Enhancement of Negative Inference

Negative inference has a major advantage in effectively expanding the cognitive boundary of models. For example, the positive inference can only infer the specific *ladder*, *wheel*, and *tree* as environment variables, as shown in Fig. 1(b), while the negative inference approach can infer all variable space except the *cycle* and *house* as environments. While the ultimate objective is still to extract invariant subgraphs, the negative inference mechanism prioritizes inferring the environment space, empowering the model with the capability to adapt to complex environment shifts. From the perspective of information theory, the training objective of negative inference can be formalized as,

$$\max I(E; G_C|\overline{Y}) = \max I(E; \mathcal{G} - G_C|Y) = \max I(E; \mathcal{G}|\overline{Y}) - I(E; G_C|Y). \tag{3}$$

Theorem 3.5. *The learning objective of negative inference paradigm (Eq. 3) encompasses a broader cognitive space for environments, with its upper limit being the ground-truth environment distribution.*

Theorem 3.5 emphasizes that the negative inference paradigm enables a broader-scale environment inference space by cooperatively modeling both intra-class spurious subgraphs and extra-class samples. Detailed proofs can be found in Appendix C.2.

4 Graph Invariant Learning with Negative Inference

In this section, we introduce a novel negative inference graph OOD framework `NeGo` to address the limitation of existing efforts in handling complex environments shifts. Specifically, `NeGo` is developed to design a negative inference learning task to capture underlying environments, and leverage inferred environment embeddings to enhance graph invariant learning.

4.1 Negative Prompt Learning for Environment Inference

Negative inference focuses on indirectly extracting invariant subgraph by investigating the information beyond the causal factors. This leads to the problem that the space of variables beyond causal information is infinite-dimensional. Given the insight from Theorem 3.5, we decouple the process of modeling the environment through negative inference into two components: the extraction of intra-class spurious subgraphs and the inference of extra-class samples. Modeling spurious subgraphs is relatively straightforward and extensively studied. The crucial challenge lies in achieving a comprehensive understanding of extra-class sample space.

Formally, let the prior distribution of extra-class samples for G be denoted as $\mathbb{P}(\overline{Y})$, where G is with the label Y. We introduce a variational estimate of the environment variables denoted as $\mathbb{Q}_\phi(E|G)$ (a.k.a., f_ϕ), where ϕ is the parameterized network. Denoting KL-divergence as $\mathrm{KL}(\cdot||\cdot)$, the training of $\mathbb{Q}_\phi$ is to implement the first term of Eq. 3, which can be formalized as following optimization,

$$\min_\phi \mathbb{E}[\mathrm{KL}(\mathbb{Q}_\phi(E|G))||\mathbb{P}(\overline{Y})]. \tag{4}$$

Inspired by the success of prompt learning in capturing underlying semantic and causal associations in large language models Floridi & Chiriatti (2020); Sordoni et al. (2024), we introduce a negative prompter to achieve this goal. Specifically, given a sample G belonging to class l, the negative prompter treats all extra-class samples as environments. Designing appropriate prompt tokens to guild effective learning is the primary question that needs to be addressed when employing the concept of prompt learning.

Given the proven efficacy of learnable prompts in various practices, we design class-specific learnable prompt tokens $\boldsymbol{P} = [\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, ..., \boldsymbol{v}^{(L)}]$, where $\boldsymbol{v}^{(i)} \in \mathbb{R}^{1\times d}$ and L is the number of classes. The class-specific design manner aims to capture the extra-class sample space for each graph, in order to achieve the objective defined by Eq. 4. The negative prompter $f_\phi(\cdot)$ is guided to learn the prompt answers $\boldsymbol{A}_N \in \mathbb{R}^{L\times d}$ by interacting the encoded graph embedding $\boldsymbol{Z}_G \in \mathbb{R}^{1\times d}$ and the learnable prompts $\boldsymbol{P}$,

$$\boldsymbol{A}_N = f_\phi(\boldsymbol{Z}_G, \boldsymbol{P}). \tag{5}$$

The negative prompter f_ϕ is parameterized the cross-attention network in Transformer decoder Vaswani et al. (2017), where $\boldsymbol{Z}_G$ is obtained by a GNN backbone encoder $h_\psi(\cdot)$. For a sample G belonging to class l, such negative prompts answers $\boldsymbol{A}_N$ should satisfy the following two properties:

- The prompts answers $\boldsymbol{A}_N$ should produce a *low match* with graphs whose labels are l.
- The prompts answers $\boldsymbol{A}_N$ should produce a *high match* with graphs whose labels are not l.

With the explanation in the language models, our *negative prompt mechanism* involves designing prompt tokens to learn the desired `[answer]` of "`the underlying environments of current sample are [answer]`". These two properties guide $f_\phi(\cdot)$ to learn a positive answer when interacting with each extra-class sample and a negative answer when interacting with each intra-class sample. Therefore, the training objective of our *negative prompt mechanism* can be formulated as,

$$\mathcal{L}_{naga} = \mathbb{E}[\mathrm{KL}(\mathbb{P}(\overline{Y})||\mathbb{Q}_\phi(E|G))] = -\mathbb{E}[\log \mathbb{P}_\phi(\bar{Y}|G, \boldsymbol{P}) - \log \mathbb{P}_\phi(Y|G, \boldsymbol{P})]. \tag{6}$$

The environment variables we infer are class-specific, in contrast to the global environment factors constructed by previous methods. Our design is intuitively reasonable, as a specific subgraph may be perceived by one class as causal information, while its minor variations are perceived by other classes as environments. Moreover, it is worth noting that we do not overlook the inference of the environments (spurious subgraphs) within intra-class samples. Given that the intra-class environments are always intertwined with causal factors, we incorporate the inference of intra-class environment variables into the discovery of the causal subgraph, which is provided in the next subsection.

4.2 Environment-enhanced Graph Invariant Learning

While inferring environment variables is a crucial step in understanding the data generation process, the ultimate goal of graph learning is to achieve casual invariant prediction. Thus, the next challenge to address is the disentanglement of the causal subgraph from environments. Existing methods often neglect the design of a graph-tailored environment exploitation algorithm, which can lead to the failure in extracting causal subgraphs when environment becomes complex Gui et al. (2024).

We propose an environment-enhanced invariant learning mechanism that leverages perceived latent environment embeddings to achieve the extraction of causal subgraphs with resistance to complex environment disturbances. Different from the negative prompter that investigates the extra-class sample space, we concentrate on the disentanglement of causal invariant substructures within the intra-class samples in this subsection.

Let the marginal distribution of the causal subgraph G_C be $\mathbb{P}(G_C)$. We introduce a variational estimation of the subgraph extraction $\mathbb{Q}_\xi(G_C|G, E)$ (a.k.a., g_ξ), where ξ is the parameterized networks. The model can make casual invariant predictions of the label distribution $\mathbb{P}_\theta(Y|G_C)$ (a.k.a., g_θ), only when the causal graph is accurately extracted from complex environments. The learning objective for environment-enhanced graph invariant learning $\mathbb{P}_\theta \circ \mathbb{Q}_\xi(\cdot)$ is to implement the second term of Eq. 3, which can be formalized as following optimization,

$$\min_{\xi,\theta} \mathbb{E}[\mathrm{KL}(\mathbb{Q}_\xi(G_C|G,E)||\mathbb{P}(G_C)) - \log \mathbb{P}_\theta(Y|G_C)]. \tag{7}$$

The environment embedding $\boldsymbol{A}_N \in \mathbb{R}^{L\times d}$ is inferred at the graph level, but the extraction of substructures often requires node-level operations. Therefore, the primary focus of environment-enhanced invariant learning is to propagate the perceived environment embedding $\boldsymbol{A}_N$ to individual nodes. We design an interaction-decoding module $g_{\xi_1}(\cdot)$ to address this issue.

Specifically, $g_{\xi_1}(\cdot)$ consists of three families of learnable parameters, i.e., $W^Q, W^K, W^V \in \mathbb{R}^{d\times d}$. $g_{\xi_1}(\cdot)$ takes the node-level representation $\boldsymbol{Z} \in \mathbb{R}^{N\times d}$ encoded by the GNN encoder $h_\psi(\cdot)$ and the environment embedding $\boldsymbol{A}_N \in \mathbb{R}^{L\times d}$ obtained by negative prompt as inputs. Three hidden state matrices are calculated by,

$$\boldsymbol{Z}^Q = \boldsymbol{Z}W^Q,\ \boldsymbol{A}^K = \boldsymbol{A}_N W^K,\ \boldsymbol{A}^V = \boldsymbol{A}_N W^V. \tag{8}$$

The node embedding with environment information obtained through residual connections is,

$$\boldsymbol{Z}_E = \mathrm{softmax}(\frac{\boldsymbol{Z}^Q(\boldsymbol{A}^K)^T}{\sqrt{d}})\boldsymbol{A}^V + \boldsymbol{Z}. \tag{9}$$

We exploit a subgraph extractor $G_C = g_{\xi_2}(\boldsymbol{Z}_E)$ to realize invariant subgraph discovery. Then, G_C is encoded by $h_\psi(\cdot)$ to obtain the causal representation for prediction. This representation is passed through an MLP layer g_θ to model the distribution of Y. Therefore, the training objective of environment-enhanced invariant learning is,

$$\mathcal{L}_{posi} = -\mathbb{E}[\log \mathbb{P}_{\xi,\theta}(Y|G_C)] = -\mathbb{E}[\log \mathbb{P}_\theta(Y|G_C) + \log \mathbb{P}_{\xi_1,\xi_2}(G_C|G, \boldsymbol{A}_N)]. \tag{10}$$

4.3 Optimization and Theoretical Analysis

Our NeGo achieves a graph learning framework with a wider space of environment inference. This is accomplished through two sequential approaches, first focusing on constructing the learning task for negative inference, and then leveraging the environment embeddings obtained from negative inference to enhance graph causal invariant learning. Thus, the training objective of our NeGo is,

$$\mathcal{L} = \mathcal{L}_{nega} + \mathcal{L}_{posi}. \tag{11}$$

The training process of NeGo is provided in Alg. 1. It is worth noting that the two sub-challenges addressed by NeGo are not independent but closely interconnected. The environment negative inference mechanism assists the network in comprehending the distribution shift of data, while the causal invariant learning with environment enhancement empowers the network to accurately extract causal invariant subgraphs even in complex environments. Therefore, the former serves as a foundation for the latter. This design reflects the principle that understanding data generation process is crucial to enhance the generalization of models. We also provide theoretical evidence supporting the ability of NeGo to effectively address both *FIIF* and *PIIF* under both cases of $H(G_C|Y) < H(G_S|Y)$ and $H(G_C|Y) > H(G_S|Y)$, where detailed proof is provided in Appendix C.3.

Table 1: The ROC-AUC performance of NeGo on four real-world datasets in chemical research field. ID val and OOD val represent the results of OOD test set using the in-distribution and out-of-distribution validation sets, respectively Gui et al. (2024). The best results are shown in **bold** and the second best results are underlined.

Model	GOOD-HIV-scaffold		GOOD-HIV-size		DrugOOD-assay		DrugOOD-size	
	ID val	OOD val	ID val	OOD val	ID val	OOD val	ID val	OOD val
ERM	69.61±1.32	70.37±1.19	61.66±2.45	57.31±1.06	70.03±0.16	72.18±0.18	62.97±0.26	63.29±0.33
IRM	73.35±2.30	70.89±0.29	58.52±0.86	60.86±2.78	71.56±0.32	72.69±0.29	63.24±0.26	63.46±0.23
V-Rex	71.73±3.51	71.18±0.69	58.39±1.54	60.10±2.09	70.22±0.86	72.32±0.58	63.87±0.42	64.11±0.39
IB-IRM	67.56±2.31	66.25±0.93	57.45±0.74	56.65±1.22	69.34±0.48	71.32±0.76	64.03±0.61	64.59±0.70
DIR	65.84±1.71	68.59±3.70	59.69±1.59	60.85±0.52	67.29±0.73	69.70±0.65	63.85±0.65	64.73±0.54
GSAT	71.55±3.58	71.39±1.41	60.92±1.00	60.61±1.19	71.01±0.54	72.26±0.45	65.12±0.38	65.67±0.45
CAL	73.48±2.64	72.38±1.03	62.83±1.26	62.58±1.04	71.89±0.92	71.23±1.13	63.85±0.49	64.22±0.74
CIGA	66.25±2.89	71.47±1.29	58.24±3.78	62.56±1.76	67.68±1.14	70.54±0.59	64.14±0.66	64.83±0.79
GIL	70.89±1.60	70.23±1.23	61.74±1.76	61.29±1.34	70.45±0.89	70.73±1.36	64.91±0.51	65.43±0.64
LECI	74.04±0.65	74.43±1.69	64.83±2.59	65.44±1.78	72.67±0.46	73.45±0.17	**65.93±0.43**	66.49±0.60
GALA	73.85±1.10	74.02±1.34	63.99±1.54	64.45±2.26	72.83±0.73	73.23±0.29	65.23±0.72	65.84±0.52
NeGo	**75.21±0.73**	**75.87±1.02**	**65.23±1.74**	**65.92±1.82**	**73.20±0.18**	**73.94±0.25**	65.49±0.73	**66.91±0.84**

Theorem 4.1. *Given the FIIF or PIIF assumptions under both cases when $H(G_C|Y) < H(G_S|Y)$ and $H(G_C|Y) > H(G_S|Y)$, the causal subgraph G_C can be extracted by optimizing Eq. 11.*

5 Experiments

We conduct extensive experiments to evaluate the effectiveness of NeGo in addressing the out-of-distribution generalization issue. Specifically, we analyze the effectiveness of NeGo by answering the following questions. **Q1**. Does our approach effectively address the issue unresolved in existing works? **Q2**. Is our framework sufficiently interpretable? **Q3**. Does each component in our NeGo effectively enhance the generalization capacity? **Q4**. Does our framework operate with high efficiency?

5.1 Baselines and Datasets

Baselines. We choose four representative OOD methods and seven graph-specific OOD approaches for comparison. Representative OOD frameworks consist of ERM, IRM Arjovsky et al. (2019), V-Rex Krueger et al. (2021), and IB-IRM Ahuja et al. (2021). The Empirical Risk Minimization (ERM) baseline is a vanilla GNN with ERM objective, which is trained by using the same settings with Gui et al. (2024). Graph OOD approaches includes DIR Wu et al. (2022c), GSAT Miao et al. (2022), CAL Sui et al. (2022), CIGA Chen et al. (2022), GIL Li et al. (2022), LECI Gui et al. (2024) and GALA Chen et al. (2024). Detailed baselines is given in Appendix B.5.

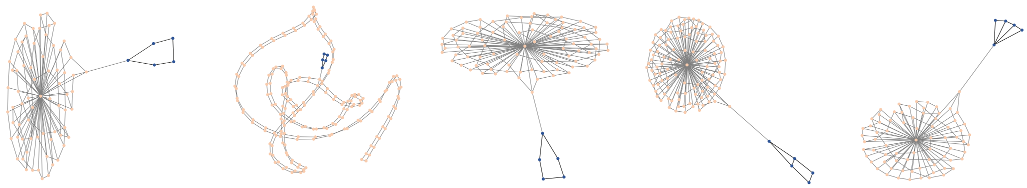


Figure 4: The causal subgraphs extracted by NeGo on the modified dataset in Fig. 1(a).

Datasets. We adopt two synthetic datasets with distribution shift and six real-world scenario shift datasets from various domains. Synthetic datasets include GOOD-Motif Wu et al. (2022c) and GOOD-CMNIST Gui et al. (2022). In molecular property prediction fields, we select the scaffold and size splits of GOOD-HIV dataset Gui et al. (2022); Wu et al. (2018) and the assay and size splits of DrugOOD LBAP-core-ic50 dataset Ji et al. (2022). We also choose two social sentiment graph datasets with distribution shift a, including GOOD-SST2 and GOOD-Twitter Yuan et al. (2022). Detailed descriptions about datasets can be found in Appendix B.4.

5.2 Implementation Details

We implement our Nego and parts of baselines with PyTorch 1.10.1 on a server with NVIDIA A100-PCIE-40GB. All experiments are repeated with 10 different random seeds of [1,2,3,4,5,6,7,8,9,10]. The reported results include the mean and standard deviation obtained from these 10 runs. During the training stage, we employ the Adam optimizer. We set the maximum number of training epochs to 200. The batch size of training is set as 32 except for GOOD-CMNIST, which uses a batch size of 64. For GOOD-Motif, GOOD-CMNIST and GOODSST2, the learning rate is set to 5×10^{-4}. For GOOD-HIV, GOOD-Twitter, and DrugOOD, we exploit a learning rate of 10^{-4}. Additionally, we utilize a weight decay of 10^{-4} to help with regularization and prevent overfitting. The experiment setup of all baselines is same as Gui et al. (2024).

Table 2: The accuracy of NeGo on two sentiment graph datasets.

Model	GOOD-SST2		GOOD-Twitter	
	ID val	OOD val	ID val	OOD val
ERM	78.37±2.64	80.41±0.69	54.93±0.96	57.04±1.70
IRM	79.73±1.45	80.17±1.52	55.27±1.19	57.72±1.03
V-Rex	79.31±1.40	80.33±1.09	56.46±0.93	56.37±0.76
IB-IRM	78.93±1.23	80.22±0.55	54.23±1.21	56.73±1.02
DIR	77.65±0.71	81.50±0.55	55.32±1.85	56.81±0.91
GSAT	79.25±1.09	80.46±0.38	55.09±0.66	56.07±0.53
CAL	81.20±1.21	82.34±0.67	56.77±0.86	57.82±0.44
CIGA	80.37±1.46	82.93±0.75	57.51±1.36	57.19±1.15
GIL	81.43±1.02	83.31±0.50	58.21±1.24	57.82±1.18
LECI	**82.93±0.22**	83.44±0.27	59.35±1.44	59.64±0.15
GALA	82.60±0.66	82.98±0.42	59.03±0.65	60.45±1.36
NeGo	82.72±0.51	**84.16±0.23**	**60.82±0.22**	**61.25±0.70**

5.3 Result Comparison and Analysis

We comprehensively evaluate the OOD performance of NeGo on both real-world and synthetic datasets to answer **Q1**. Tab. 1 and 2 present the performance of NeGo on chemical and sentiment graph datasets. Tab. 3 showcases the performance of our framework on two synthetic datasets. Compared to existing methods, our method achieves optimal performance on almost all datasets. Besides, the performance of environment-centered OOD methods, such as LECI and GALA, often achieves suboptimal or even optimal results on various datasets. This demonstrate the effectiveness of modeling environment factors in addressing data distribution shifts.

To further investigate whether our method can effectively tackle environment shifts, we evaluate the performance of our framework in the the complex environments scenario illustrated in Fig. 1(a). NeGo achieves 87.34% and 80.29% on the original and adjusted dataset, respectively. There is only a minor decrease in performance, suggesting that our method effectively tackles the limitations encountered by existing methods in handling complex environments. To answer the **Q2**, we visually represent the causal subgraphs extracted by NeGo on the modified dataset in Fig. 1(a). As depicted in Fig. 4, our method consistently extracts the ground-truth subgraph. The visualized results further validate the effectiveness of our proposed negative inference method. By modeling extensive extra-class samples as environments, our approach offers undeniable advantages in handling complex environment shifts.

Table 3: The accuracy of NeGo on two synthetic datasets, where GOOD-Motif has a structure shift and GOOD-CMNIST has a feature shift.

Model	GOOD-Motif		GOOD-CMNIST	
	basis	size	color	covariate
ERM	60.93±11.11	56.63±7.12	26.64±2.37	57.56±9.59
IRM	64.94±4.85	54.52±3.27	29.63±2.06	58.11±5.14
V-Rex	61.59±6.58	55.85±9.42	27.13±2.90	48.78±7.81
IB-IRM	63.45±5.42	52.76±4.67	28.95±1.98	50.56±6.62
DIR	34.39±2.02	43.11±2.78	22.53±2.56	44.67±0.00
GSAT	62.27±8.79	50.03±5.71	35.02±2.78	68.22±7.23
CAL	59.45±3.34	51.27±2.50	28.87±1.80	52.59±2.76
CIGA	37.81±2.42	51.87±5.15	25.06±3.07	56.78±2.99
GIL	68.48±2.46	63.61±2.75	47.32±2.27	57.61±2.98
LECI	**84.56±2.22**	71.43±1.96	51.80±2.71	**83.20±5.89**
GALA	80.95±1.31	70.45±1.30	52.68±2.40	81.23±3.29
NeGo	83.96±1.90	**72.65±1.47**	**53.28±1.79**	82.43±1.73

Table 5: The training efficiency of NeGo with other baselines on DrugOOD-size (s/epoch).

Models	GSAT	DIR	CIGA	LECI	GALA	NeGo
Training Time	51.6	52.6	54.2	59.1	62.3	58.7

5.4 Ablation Studies

To answer **Q3**, we investigate each component of NeGo. Specifically, we conduct ablation studies to explore the effectiveness of negative prompter and interactive decoding component. Tab. 4 shows that the performance drops significantly when there is either no negative prompter or interactive decoding component. NeGo-NoPro refers to the framework that eliminates negative prompter and negative loss, which causes the most performance drop. Therefore, the negative inference mechanism plays a vital role in enhancing the capability of environment perception. This further validates the rationale of our motivation for incorporating negative inference. NeGo-NoEnv indicates that the casual subgraphs are extracted directly using node embedding without integrating inferred environment information. The performance decline emphasizes the significance of environment utilizing strategies overlooked by existing works.

5.5 Efficiency Analysis

To address Q4, we explore the training efficiency of NeGo from both theoretical and practical perspectives. The time complexity of NeGo is $\mathcal{O}(\mathcal{V}\times d^2+\mathcal{V}\times d\times h+\mathcal{E}\times d)$, where $|\mathcal{V}|$ represents the number of nodes, $|\mathcal{E}|$ denotes the number of edges, d is the feature dimension, and h represents the number of cross-attention heads. Our method has linear time complexity with high training efficiency. We empirically compare the training efficiency of NeGo with other baselines on DrugOOD-size dataset as shown in Tab. 5. Compared with some earlier invariant learning methods (DIR and GSAT), the minor increase in running time of our menthod brings out the substantial performance boost. Additionally, our approach demonstrates greater competitiveness in both training efficiency and performance compared to existing environment-centered methods.

Table 4: Ablation studies of NeGo.

Model	**DrugOOD**	**GOOD**
	(assay)	**(Twitter)**
NeGo-NoPro	70.37	58.41
NeGo-NoEnv	71.71	59.17
NeGo	73.20	60.82

6 Conclusion and Future Work

In this work, we propose a negative inference graph OOD framework NeGo to handle complex environment shift in OOD scenarios. Our NeGo aims to comprehensively infer the entire environmental space by explicitly modeling the extra-class environment that has been significantly overlooked in prior research. By inheriting the successful practices of prompt learning in language modeling, we fist design a negative prompter to realize extra-class environment awareness. We then introduce an environment-enhanced invariant learning strategy to eliminate spurious subgraphs from the data. This strategy effectively leverages the inferred environment variables to enhance the ability to remove irrelevant information. Extensive experiments on real-world datasets across domains and synthetic datasets validate the effectiveness of NeGo.

Future work. Our design can effectively solve the existing challenges, but there still exist a limitation. The negative prompter in our approach learns class-specific environment embeddings by considering all extra-class samples as environment variables. This results in our method relying on the class information of the dataset. With a larger number of classes, the model is better equipped to capture and recognize complex underlying environment factors. When the dataset is limited to a binary classification task, environment factors always present within the in-class samples. In this case, our negative prompter may have reduced capability to expand the environment inference space. The reason for this limitation is that the model is sensitive to the characteristics of dataset. Actually, we can realize that environment variables are often shareable across datasets. Therefore, it is a promising research direction to study cross-task graph OOD work to capture broader environmental information. In the future, we aim to investigate transferable multi-task graph out-of-distribution generalization learning, which is not discussed in existing works.

REFERENCES

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. *Advances in Neural Information Processing Systems*, 35:22131–22148, 2022.

Yongqiang Chen, Yatao Bian, Kaiwen Zhou, Binghui Xie, Bo Han, and James Cheng. Does invariant graph learning via environment augmentation learn invariance? *Advances in Neural Information Processing Systems*, 36, 2024.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

Chen Gao, Xiang Wang, Xiangnan He, and Yong Li. Graph neural networks for recommender system. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 1623–1625, 2022.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

H Paul Grice and Alan R White. Symposium: The causal theory of perception. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 35:121–168, 1961.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. Good: A graph out-of-distribution benchmark. *Advances in Neural Information Processing Systems*, 35:2059–2073, 2022.

Shurui Gui, Meng Liu, Xiner Li, Youzhi Luo, and Shuiwang Ji. Joint learning of label and environment causal independence for graph out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Long-Kai Huang, Tingyang Xu, Yu Rong, Lanqing Li, Jie Ren, Ding Xue, et al. Drugood: Out-of-distribution (ood) dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. *arXiv preprint arXiv:2201.09637*, 2022.

Tianrui Jia, Haoyang Li, Cheng Yang, Tao Tao, and Chuan Shi. Graph invariant learning with subgraph co-mixup for out-of-distribution generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8562–8570, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. Learning invariant graph representations for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35: 11828–11841, 2022.

Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8886–8895, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, and Yu Zheng. Geoman: Multi-level attention networks for geo-sensory time series prediction. In *IJCAI*, volume 2018, pp. 3428–3434, 2018.

Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, 34:6155–6170, 2021.

Bin Lu, Ze Zhao, Xiaoying Gan, Shiyu Liang, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. Graph out-of-distribution generalization with controllable data augmentation. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International Conference on Machine Learning*, pp. 15524–15543. PMLR, 2022.

Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5115–5124, 2017.

Judea Pearl. Causal inference in statistics: An overview. 2009.

Judea Pearl. Causal inference. *Causality: objectives and assessment*, pp. 39–58, 2010.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

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

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18082–18091, 2022.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

Alessandro Sordoni, Eric Yuan, Marc-Alexandre Côté, Matheus Pereira, Adam Trischler, Ziang Xiao, Arian Hosseini, Friederike Niedtner, and Nicolas Le Roux. Joint prompt optimization of stacked llms using variational inference. *Advances in Neural Information Processing Systems*, 36, 2024.

Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Yongduo Sui, Xiang Wang, Jiancan Wu, Min Lin, Xiangnan He, and Tat-Seng Chua. Causal attention for interpretable and generalizable graph classification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1696–1705, 2022.

Xiangguo Sun, Hong Cheng, Jia Li, Bo Liu, and Jihong Guan. All in one: Multi-task prompting for graph neural networks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2120–2131, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312: 135–153, 2018.

Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, and Bryan Hooi. Mixup for node and graph classification. In *Proceedings of the Web Conference 2021*, pp. 3663–3674, 2021.

Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5764–5773, 2019.

Zihao Wang, Yongqiang Chen, Yang Duan, Weijiang Li, Bo Han, James Cheng, and Hanghang Tong. Towards out-of-distribution generalizable predictions of chemical kinetics properties. *arXiv preprint arXiv:2310.03152*, 2023.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

Junkang Wu, Jiawei Chen, Jiancan Wu, Wentao Shi, Xiang Wang, and Xiangnan He. Understanding contrastive learning via distributionally robust optimization. *Advances in Neural Information Processing Systems*, 36, 2024.

Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective. *arXiv preprint arXiv:2202.02466*, 2022a.

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.

Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. *arXiv preprint arXiv:2201.12872*, 2022c.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Yutong Xia, Yuxuan Liang, Haomin Wen, Xu Liu, Kun Wang, Zhengyang Zhou, and Roger Zimmermann. Deciphering spatio-temporal graph forecasting: A causal lens and treatment. *arXiv preprint arXiv:2309.13378*, 2023.

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

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

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823, 2020.

Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5782–5799, 2022.

Haonan Yuan, Qingyun Sun, Xingcheng Fu, Ziwei Zhang, Cheng Ji, Hao Peng, and Jianxin Li. Environment-aware dynamic graph learning for out-of-distribution generalization. *arXiv preprint arXiv:2311.11114*, 2023.

Tong Zhao, Yozen Liu, Leonardo Neves, Oliver Woodford, Meng Jiang, and Neil Shah. Data augmentation for graph neural networks. In *Proceedings of the aaai conference on artificial intelligence*, volume 35, pp. 11015–11023, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022a.

Yangze Zhou, Gitta Kutyniok, and Bruno Ribeiro. Ood link prediction generalization capabilities of message-passing gnns in larger test graphs. *Advances in Neural Information Processing Systems*, 35:20257–20272, 2022b.

Zhengyang Zhou, Yang Wang, Xike Xie, Lianliang Chen, and Hengchang Liu. Riskoracle: A minute-level citywide traffic accident forecasting framework. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 1258–1265, 2020.

Jia-Jie Zhu, Wittawat Jitkrittum, Moritz Diehl, and Bernhard Schölkopf. Kernel distributionally robust optimization: Generalized duality theorem and stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 280–288. PMLR, 2021.

Deyu Zou, Shikun Liu, Siqi Miao, Victor Fung, Shiyu Chang, and Pan Li. Gdl-ds: A benchmark for geometric deep learning under distribution shifts. *arXiv preprint arXiv:2310.08677*, 2023.

A Broader impacts

Graph learning models are widely used to support scientific research and social development, such as molecular discovery, recommendation systems, and smart cities. However, with the increasing complexity of data scale and application scenarios, the distribution shifts between training and test data have become a significant obstacle in the development of graph learning. In light of this, our work aims to address the issue of data distribution shifts in the model and promote the broader application of graph learning in various fields. Therefore, our work aims to develop a model with the out-of-distribution generalization ability and thereby promote the widespread application of graph learning in various fields.

We ensure the full ethical compliance of our work, and all the datasets we utilize are publicly available. Our work does not involve human subjects and does not introduce any potential negative social impacts or issues related to privacy and fairness.

B Related works

B.1 OOD Generalization.

Out-of-Distribution (OOD) generalization learning refers to the task of adapting a model that has been trained on a specific distribution to effectively process data from a potentially different distribution. This study holds significant importance because the issue of data distribution shifts is a common occurrence in the real world. External factors, such as changes in environmental conditions, technological advancements, or evolving user preferences, can lead to shifts in the data distribution. Various approaches can be employed for OOD generalization, including data augmentation Rong et al. (2019); Wang et al. (2021); You et al. (2020), domain adaptation Wang & Deng (2018), and causal invariant learning Sui et al. (2022); Wu et al. (2022c). Jia et al. Jia et al. (2024) innovatively proposes a mixup-based environment modeling framework, IGM, to enhance graph invariant learning. IGM focuses on expanding the environment space through generation (mixing), while our NeGo aims to mine environmental space as much as possible from the novel perspective of negative learning. Piao et al. (2024) et al. creatively proposes a hierarchical environment inference paradigm to enhance graph invariant learning methods. This work focuses on generating sample-level hierarchical environments to expand the modeling of the environment space. Unlike this method, our NeGo focuses on class-level environment augmentation, collaborating with extra-class environment modeling and inter-class invariant learning to achieve global inference of environment space.

Among them, causal invariant learning demonstrates impressive performance in various fields, due to its powerful interpretability Chen et al. (2022); Li et al. (2022); Miao et al. (2022); Wu et al. (2022c). Our NeGo is aligned with this research line, as an environment-centered invariant learning method based on causal theory. However, in the field of graph learning, most existing invariant learning methods focus on extracting the causal graph to achieve invariant learning. This strategy limits the inference space of the environments to the dimension of spurious subgraphs, which hinders the ability of models to capture the complex environment states. In this work, we propose an invariant learning mechanism based on negative inference to address this limitation.

B.2 Prompt Learning

Prompt learning is proposed in NLP models to infer underlying semantic and potential causal associations in linguistic data. Many effective prompt methods has developed with the introduction of large language models, including some hand-crafted prompts Brown et al. (2020), discrete prompts Gao et al. (2020); Shin et al. (2020), and learnable prompts design Li & Liang (2021). There have been various works on the interaction of computer vision and natural language processing fields, e.g., text-to-image retrieval text-to-image retrieval Wang et al. (2019), visual question answeringAntol et al. (2015); Rao et al. (2022); Zhou et al. (2022a) and so on.

In recent years, prompt learning has also been developed in the graph learning field, including multi-task learning framework Sun et al. (2023). Our approach is the pioneering effort to apply prompt learning to address the challenge of graph OOD generalization issue.

B.3 Comparisons to Previous Graph OOD Works

Environment-centered studies Chen et al. (2024); Gui et al. (2024); Li et al. (2022); Wu et al. (2022a); Yang et al. (2022) consider that the data distribution shifts stem from the changes of environments. Therefore, these practices enable the model to withstand data distribution shifts by inferring environment variables. Concretely, the networks are often trained with the objective of equipping models to effectively handle mixed environments scenarios. However, this design allows the networks to make narrow inference about the environments, and makes the networks unable to handle with distribution shifts in complex environments. We attribute this limitation of inference scale to the shortcomings of positive inference, which is proved both empirically and theoretically. Therefore, we propose a negative inference mechanism to broaden the inference space for environments, without relying on the mixed environments hypothesis.

Our approach, which represents a pioneering practice in utilizing negative inference, is distinct from all existing practices in this field. DIR Wu et al. (2022c) aims to identify causal patterns that are stable across different distributions and filter out spurious patterns that are unstable. This work is a classic work in the early application of causal theory to address the challenge of graph OOD generalization. It focuses on obtaining invariant subgraphs with a positive inference manner. GIL Li et al. (2022) aims to capture the invariant relationships between predictive graph structural information and labels in a mixture of latent environments through jointly optimizing three mutually promoting modules. This method relies on the mixing environment hypothesis and has limited inference space for environments. CIGA Chen et al. (2022) build three Structural Causal Models (SCMs) to characterize the distribution shifts that could happen on graphs: one is to model the graph generation process, and the other two are to model two possible interactions between invariant and spurious features during the graph generation, i.e., FIIF and PIIF. This work provides a fresh perspective on existing research on out-of-distribution generalization based on causality. However, it still falls within the framework of positive inference, aiming to extract causal subgraphs. GALA Chen et al. (2024) utilized proxy prediction mechanism to infer environment label. It is worth noting that the negative samples mentioned in this work are different from our negative inference, and their design is also to improve performance under the mixed environments hypothesis. Thus, it essentially follows a positive inferring process for environment variables. LECI Gui et al. (2024) primarily focused on spurious substructures space to model the environment variables. Such environment inference strategy still relies on a positive inference with narrow cognitive space of the environments.

Algorithm 1: The training process of `NeGo`

```
Input: training data $\mathcal{G}$, negative prompts $\boldsymbol{P}$.
Initial: the GNN encoder $h_\psi$, the negative prompter $f_\phi$, environment-enhanced invariant
learning mechanism $g_\xi$, final predictor $g_\theta$, learnable prompt tokens $\boldsymbol{P}$, the number of epochs $K$.
for $i = 1$ to $K$ do
    $\boldsymbol{Z}_G = h_\psi(G)$
    $\boldsymbol{A}_N = f_\phi(\boldsymbol{Z}_G, \boldsymbol{P})$
    $\boldsymbol{Z}^Q = \boldsymbol{Z}W^Q,\ \boldsymbol{A}^K = \boldsymbol{A}_N W^K,\ \boldsymbol{A}^V = \boldsymbol{A}_N W^V$
    $\boldsymbol{Z}_E = \mathrm{softmax}(\frac{\boldsymbol{Z}^Q(\boldsymbol{A}^K)^T}{\sqrt{d}})\boldsymbol{A}^V$
    $Y = g_\theta(G_C),\ \ G_C = g_{_2}(\boldsymbol{Z}_E + \boldsymbol{Z})$
    Optimizing:
    $\mathcal{L}_{naga} = \mathbb{E}[\mathrm{KL}(\mathbb{P}(\overline{Y})||\mathbb{Q}_\phi(E|G))] = -\mathbb{E}[\log \mathbb{P}_\phi(\bar{Y}|G, \boldsymbol{P}) - \log \mathbb{P}_\phi(Y|G, \boldsymbol{P})]$
    $\mathcal{L}_{posi} = -\mathbb{E}[\log \mathbb{P}_{\xi,\theta}(Y|G_C)] = -\mathbb{E}[\log \mathbb{P}_\theta(Y|G_C) + \log \mathbb{P}_{\xi_1,\xi_2}(G_C|G, \boldsymbol{A}_N)]$
    $\min\limits_{\psi,\phi,\theta,\xi,\mathbf{P}} \mathcal{L} = \mathcal{L}_{nega} + \mathcal{L}_{posi}$
end for
Return $h_\psi$, $f_\phi$, $g_\xi$, $g_\theta$ and $\mathbf{P}$
```

B.4 Datasets

We adopt two synthetic datasets with distribution shift and six real-world scenario shift datasets from various domains. **Synthetic datasets** include GOOD-Motif Wu et al. (2022c) and GOOD-CMNIST Gui et al. (2022). In **molecular property prediction fields**, we select the scaffold and size splits of

Table 6: Statistics on the number of graphs in the datasets.

Dataset	Training	ID validation	ID test	OOD validation	OOD test
GOOD-HIV-Scaffold	24682	4112	4112	4113	4108
GOOD-HIV-Size	26169	4112	4112	2773	3961
GOOD-SST2-Length	24744	5301	5301	17206	17490
GOOD-Twitter-Length	2590	554	554	1785	1457
GOOD-CMNIST-Color	42000	7000	7000	7000	7000
GOOD-Motif-Basis	18000	3000	3000	3000	3000
GOOD-Motif-Size	18000	3000	3000	3000	3000
DrugOOD-assay	34179	11314	11683	19028	19032
DrugOOD-size	36597	12153	12411	17660	16415

GOOD-HIV dataset Gui et al. (2022); Wu et al. (2018) and the assay and size splits of DrugOOD LBAP-core-ic50 dataset Ji et al. (2022). We also choose two **social sentiment graph datasets** with distribution shifts, including GOOD-SST2 and GOOD-Twitter Yuan et al. (2022). Detailed statistics on the number of graphs in those datasets are provided in Tab. 6.

- **GOOD-Motif** is a synthetic dataset designed for studying structure shifts. Each graph in the dataset is created by connecting a base graph and a motif, where the label is determined by the motif. This accessible ground-truth substructure brings a lot of convenience to the invariant subgraph learning with interpretability. This dataset include five label-irrelevant base graphs (wheel, tree, ladder, star, and path) and three label-determining motifs (house, cycle, and crane) are used to generate the graphs in the dataset. In environment-centered invariant learning, such base graphs can be seen as environment factors and such motifs are be consider as the casual factors.
- **GOOD-CMNIST** is a semi-synthetic dataset that has been purposefully created to evaluate node feature shifts. It comprises graphs constructed from hand-written digits extracted from the MNIST database, with the transformation applied using superpixel techniques Monti et al. (2017).
- **GOOD-HIV** is a compact and real-world molecular dataset that has been derived from Wu et al. (2018). It comprises molecular graphs, where atoms represent nodes and chemical bonds represent edges. The primary task associated with this dataset is to predict a molecule's potential for inhibiting HIV replication. Its distribution shift scenario is developed into two, i.e., the scaffold, and the size of nodes in a molecular graph.
- **DrugOOD(LBAP-core-ic50)** is utilized in the Ligand-based Affinity Prediction (LBAP) task, where the core noise level and IC50 measurement type serve as domain features. Its distribution shift scenario is developed into three, i.e., the scaffold, the size, and the assay.
- **GOOD-SST2** is a real-world social sentiment dataset derived from natural language. This dataset represents each sentence as a graph, where individual words are treated as nodes, and their corresponding word embeddings serve as node features. The primary task in this dataset involves binary classification, aiming to predict the sentiment polarity of each sentence.
- **GOOD-Twitter** is a real-world natural language sentiment dataset that shares the same transformation process as the SST2 dataset. The classification task of this dataset involves predicting one of three sentiment polarities for each sentence. Similar to the GOOD-SST2 dataset, the sentence lengths are chosen as the domains.

B.5 Baselines

We choose four representative OOD methods and seven graph-specific OOD approaches for comparison. The representative OOD frameworks we select consist of ERM, IRM Arjovsky et al. (2019), V-Rex Krueger et al. (2021), and IB-IRM Ahuja et al. (2021). The Empirical Risk Minimization (ERM) baseline is a vanilla GNN with ERM objective, which is trained using the same settings with Gui et al. (2024). Graph OOD approaches includes DIR Wu et al. (2022c), GSAT Miao et al. (2022), CAL Sui et al. (2022), CIGA Chen et al. (2022), GIL Li et al. (2022), LECI Gui et al. (2024) and GALA Chen et al. (2024).

- **DIR** Wu et al. (2022c) is an early work using causal theory to address the distribution shifts issue in graph data. This work provides detailed theoretical proofs that demonstrate the feasibility of extracting invariant subgraph from graph data.
- **GSAT** Miao et al. (2022) employ information bottleneck theory to select causal subgraphs under onlythe FIIF assumption. The proposed stochastic attention mechanism in this paper is highly robust in extracting casual subgraphs, and has emerged as a backbone model in numerous methods. Actually, the subgraph extractor used in our work is also inspired by GSAT.
- **CAL** Sui et al. (2022) is guided by the backdoor adjustment principle derived from causal theory. It encourages the Graph Neural Networks (GNNs) to focus on exploiting causal features while disregarding shortcut connections.
- **CIGA** Chen et al. (2022) is the first graph OOD method considering both Fully Informative Invariant Feature (FIIF) and Partially Informative Invariant Feature (PIIF) assumptions. This work presents an OOD algorithm for graphs that is provably generalizable under different types of distribution shifts.
- **GIL** is designed to capture invariant graph patterns in a mixture of underlying environments and handle the distribution shift issue. This work introduces a GNN-based subgraph generator to identify potentially invariant subgraphs from the complex interaction between invariant and variant patterns.
- **LECI** comprehensively reviews existing OOD approaches and identifies the current causal-subgraph discovery challenges. This work jointly optimize label and environment causal independence to achieve powerful causal subgraphs learning.
- **GALA** designs an additional assistant model to enhance model with more powerful OOD generalization ability without explicit environment labels. Theoretical proofs establish that GALA possesses robust out-of-distribution generalization capabilities under the FIIF and PIIF assumptions.

C Theory and discussions

C.1 Proof of Theorem 3.4

Theorem C.1. *Given an observed graph dataset $\mathcal{G}$, the inference process, considering G_S as the environment factor, fails to capture the basis $\boldsymbol{E}_b$ that can represent the entire environment space.*

Proof. The basis $\boldsymbol{E}_b$ represents a set of fundamental components or features that can accurately represent the entire environment space. These components capture the essential variations, patterns, and characteristics present in the environment. However, if the inference process fails to capture this basis, it implies that the process is unable to fully understand and model the complexities of the environment. Thus, we next investigate that whether the environment variable inferred from G_S covers such base environments. We consider two SCMs hypotheses FIIF and PIIF as shown in Fig. 2.

Under the FIIF assumption, $Y \perp G_S | G_C$, we have $P(Y, G_S|G_C) = P(Y|G_C) \cdot P(G_S|G_C)$. This conditional independence assumption leads to an equivalent expression: $P(Y|G) = P(Y|G_S, G_C) = P(Y|G_C)$. Therefore, the process of extracting the causal subgraph G_C is equivalent to the process of modeling the spurious correlations G_S. Traditional positive casual learning methods are capable of handling the FIIF assumption.

Under the PIIF assumption, $Y \not\perp G_S | G_C$, we have $P(Y, G_S|G_C) \neq P(Y|G_C) \cdot P(G_S|G_C)$. Furthermore, we can obtain $P(Y|G) = P(Y|G_S, G_C) \neq P(Y|G_C)$. Thus, the process of extracting the causal subgraph G_C cannot be used to infer the labels of samples. More formally, using mutual information theory, we derive the following,

$$I(Y; G_S|G_C) = H(Y|G_C) - H(Y|G_S, G_C) > 0, \tag{12}$$

$$H(Y|G_C) > H(Y|G_S, G_C). \tag{13}$$

This indicates that, given the causal subgraph G_C, the uncertainty of Y is higher than when both the spurious subgraph G_S and the causal subgraph G_C are given. This suggests that the spurious subgraph G_S contains additional information about Y.

Therefore, the causal subgraph $\hat{G}_C$ learned by the model with the positive learning manner contains components of the spurious subgraph, i.e., $G_S \cap \hat{G}_C \neq \emptyset$. At this point, if we can obtain the basis for the environment space, the model should be able to infer the spurious subgraph G_S and treat it as part of the environment E. The extracted causal subgraph $\hat{G}_C$ should be able to effectively remove the spurious subgraph, i.e., $G_S \cap \hat{G}_C = \emptyset$. This clearly contradicts the PIIF assumption, indicating that the model currently lacks the capability to obtain a basis for the environmental space. Therefore, simply inferring the causal subgraph with a postive manner is not sufficient to address the PIIF assumption. Since $E \to G_S$, modeling the spurious subgraph G_S requires modeling and understanding its root E. Existing methods that simply model $G - G_C$ also lack the capability to address the PIIF assumption.

C.2 Proof of Theorem 3.5

Theorem C.2. *The learning objective of negative inference paradigm (Eq. 3) encompasses a broader cognitive space for environments, with its upper limit being the ground-truth environment distribution.*

Proof. The optimization of Eq. 3 enables a broader scale environment inference space by cooperatively modeling intra-class spurious subgraphs and extra-class samples. Given that $\max -I(E; G_C|Y) = \max I(E; G_S|Y)$, maximizing $I(E; G_S|Y)$ implements the inference process for intra-class spurious subgraphs. Consider $I(E; \mathcal{G}|\bar{Y}) = \sum\limits_{y_i \in \bar{Y}} I(E; G^{(i)})$, maximizing $I(E; \mathcal{G}|\overline{Y})$ implements the modeling of extra-class sample space. The optimization procedure of $\max I(E; \mathcal{G}|\overline{Y})$ indicates that all other extra-class samples $\{G^{(i)}|y_i \in \bar{Y}\}$ are modeled as environment variables when making environment inference on samples with label Y. Therefore, the optimization process for Eq. 3 encompasses a broader cognitive space for environments, with its upper limit being the ground-truth environment distribution.

C.3 Proof of Theorem 4.1

Theorem C.3. *Given the FIIF or PIIF assumptions under both cases when $H(G_C|Y) < H(G_S|Y)$ and $H(G_C|Y) > H(G_S|Y)$, the causal subgraph G_C can be extracted by optimizing Eq. 11.*

Proof. Given that PIIF shifts in the absence of environment labels are more challenging Chen et al. (2024), our work focuses on the ability of NeGo on the PIIF assumption, namely PIIF implies that the causal variable G_C indirectly influences the spurious variable G_S through the mediator Y. In the following analysis, we analyze the two specific scenarios under PIIF assumption, i.e., $H(G_C|Y) < H(G_S|Y)$ and $H(G_C|Y) > H(G_S|Y)$. NeGo aims to comprehensively capture the underlying environment space by inferring the extra-class sample space and the intra-class spurious subgraphs. The learning objective of extracting casual subgraph G_C can be rewritten as follows,

$$\underset{\forall e_i, e_j \in E}{\arg\max}_{\hat{G}_C} (I(\hat{G}_C^{e_i}, \hat{G}_C^{e_j}|C) - I(\hat{G}_C, \bar{G}|Y)) = \underset{\forall e_i, e_j \in E}{\arg\max}_{\hat{G}_C} (-I(\hat{G}_C, \bar{G}|Y) + I(\hat{G}_C^{e_i}, \hat{G}_C^{e_j}|Y)), \quad (14)$$

where $\hat{G}_C^{e_i}$ denotes the extracted causal subgraph under any environmental scenario e_i. The first term represents the constraint of negative inference, meaning that NeGo models all extra-class samples as environmental space. The second term represents the constraint of positive causal inference, meaning that the causal subgraph extracted under any environmental condition remains consistent, and is most useful for label prediction. Next, we will demonstrate that NeGo can address the two scenarios of the PIIF assumption.

For the case of $I(G_C; Y) > H(G_C) - H(G_S)$, we can get following derivation,

$$H(G_S|Y) > H(G_C|Y), \quad (15)$$

$$H(G_S) - I(G_S; Y) > H(G_C) - I(G_C; Y), \quad (16)$$

$$H(G_S) - H(G_C) + I(G_C; Y) > I(G_S; Y) > 0, \quad (17)$$

$$I(G_C; Y) > H(G_C) - H(G_S). \quad (18)$$

Table 7: Comparison of existing methods on addressing OOD generalization issue.

Methods	SCMs	$H(G_C\|Y) < H(G_S\|Y)$ $\&H(G_C\|Y) > H(G_S\|Y)$	Inferred Environment Space
DIR	FIIF	×	Spurious subgraphs
GSAT	FIIF	×	Spurious subgraphs
CIGA	FIIF & PIIF	×	Spurious subgraphs
GALA	FIIF & PIIF	√	Spurious subgraphs
LECI	FIIF & PIIF	√	Spurious subgraphs
NeGo	FIIF & PIIF	√	Intra-class spurious subgraphs and extra-class sample space

We can get that inferring G_C from Y is more effective and seamless compared to simply separating causal and spurious substructures based on entropy differences. Thus, our positive inference approach, $\underset{\forall e_i,e_j \in E}{\arg\max} I(\hat{G}_C^{e_i}, \hat{G}_C^{e_j}|Y)$, is sufficient to achieve the decoupling of G_C from the label Y.

For the case of $I(G_C;Y) < H(G_C) - H(G_S)$, we get $I(G_C;Y) < H(G_C) - H(G_S)$. This means that we need to consider entropy differences in the data composition to assess the differences between causal and spurious relationships. In other words, positive inference $\underset{\forall e_i,e_j \in E}{\arg\max} I(\hat{G}_C^{e_i}, \hat{G}_C^{e_j}|Y)$ alone may result in $\hat{G}_C$ containing spurious subgraph information, meaning $G_S \in \hat{G}_C$. Fortunately, our negative inference strategy can further refines $\hat{G}_C$ by considering entropy differences $H(G_C) - H(G_S)$ to better distinguish between causal and spurious relationships. Specifically, our G_C is also subject to this constraint through a negative inference approach to learn $\hat{G}_S$,

$$G_C \in G - \arg\max(I(Y|\hat{G}_S) - I(\hat{G}_S|\bar{Y})). \tag{19}$$

D Additional experiment results

In this section, we will discuss more interpretable results and the training efficiency of our framework.

D.1 More Interpretability Results

We provide more visual results to discuss the interpretability of NeGo. Fig. 5 presents th casual subgraphs extracted by NeGo on the modified dataset in Fig. 1(a). Our NeGo can accurately extract the causal subgraph from the complex spurious information. However, it is worth acknowledging that in some complex environments, our method may not only extract the ground-truth causal subgraph but also include some spurious substructures. Actually, this does not affect the accuracy of final forecasting.

D.2 Case Studies

We also explore whether incorporating prompt learning can enhance the model's performance, rather than our overall negative prompt framework. To this end, we develop a variant of our NeGo framework, referred to as PoGo, which incorporates the positive prompt practice. We evaluate the effectiveness (ROC-AUC) of PoGo on four distribution shift datasets. We present the final performance by averaging the results from two runs conducted on an NVIDIA H100 PCIe 80 GB with different random seeds. As shown in Fig. 6, the performance of PoGo is competitive with recent successful practices like LECI and GALA, demonstrating that the design of positive prompt can still obtain excellent generalization. However, our framework of negative prompt shows superior performance.

We further investigate the reason of such performance of positive prompt practice PoGo. We modify PoGo by masking the $\mathcal{L}_{posi}$ (the original Negative Loss $\mathcal{L}_{naga}$), obtaining PoGo (w/o. $\mathcal{L}_{posi}$). With all other configurations remaining the same, we observe a significant decrease in the performance of PoGo (w/o. $\mathcal{L}_{posi}$). Our analysis is as follows: although both $\mathcal{L}_{posi}$ and $\mathcal{L}_{pred}$ are positive losses

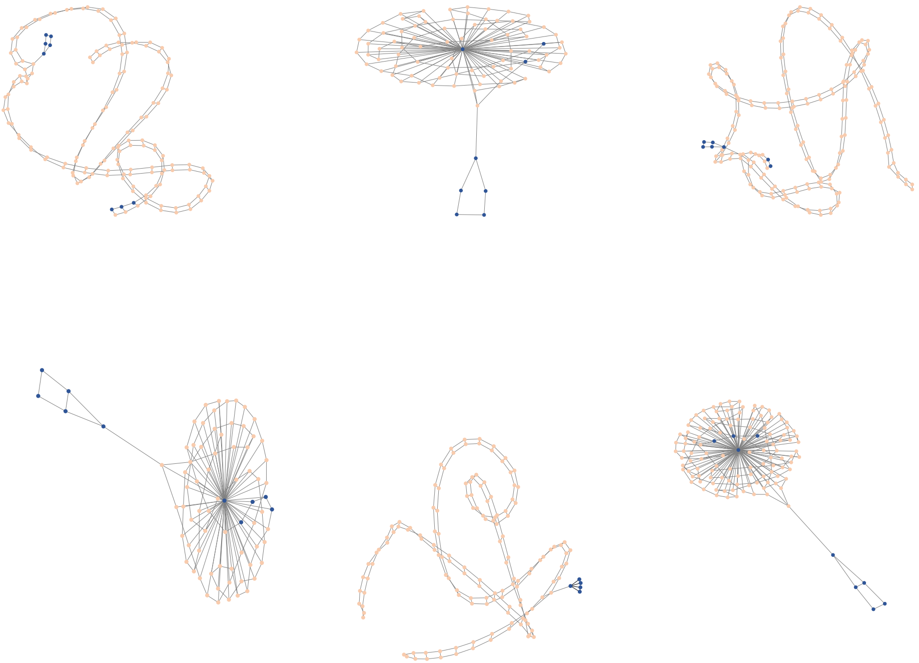


Figure 5: The causal subgraphs extracted by NeGo on the modified dataset in Fig. 1(a).

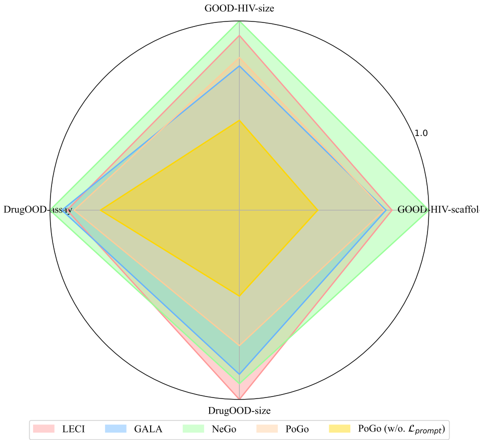


Figure 6: To explore the role of prompt learning, we develop a variant of our NeGo framework, referred to as PoGo, which incorporates the positive prompt practice.

in PoGo, we argue they serve different purposes and convey distinct information. $\mathcal{L}_{prompt}$, as a guidance strategy for the positive prompt, guides the prompt module to learn more potential environment semantics, while $\mathcal{L}_{pred}$ enhances prediction accuracy. Without prompt guidance $\mathcal{L}_{prompt}$, the advantage of prompt learning is not released. Therefore, we argue that positive prompt may also enhance the model to capture a broader scale of environments. A more in-depth investigation will be left for our future work.