ON THE GENERALIZABILITY OF NON-CAUSAL SUBGRAPHS: ADDRESSING CUE DISCORDANCE FOR GRAPH OOD GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph neural networks often exploit spurious correlations in training environments, leading to poor out-of-distribution generalization. Existing graph OOD methods typically seek causal or invariant subgraphs, implicitly treating non-causal structures as poorly generalizable by-products of overfitting. However, in our analysis and experiments, we find that non-causal subgraphs can also exhibit nontrivial cross-sample generalization and provide predictive cues that conflict with those encoded by causal subgraphs. We characterize this phenomenon as *cue discordance*, where the causal subgraph supports the ground-truth class while the spurious favors a different one. Empirically, we show that cue discordance is widespread and markedly more common among misclassified samples; moreover, increasing the proportion of discordant samples consistently reduces accuracy while increasing classification loss. We further theoretically establish the Discordance-Conditioned Causal Contrast (D3C) criterion, proving that under cue discordance, causal and non-causal subgraphs exhibit a non-trivial relative risk separation. Building on these findings, we propose Discordance-Oriented Causal Subgraph Discrimination (DCS), which uses full-graph classification risk as an observable proxy for latent discordance strength and adaptively enforces a stronger predictive advantage of candidate causal subgraphs over their non-causal complements, enabling efficient and robust causal subgraph identification. Extensive experiments on 8 datasets demonstrate consistent improvements over existing graph OOD methods. Code is available at `https://anonymous.4open.science/r/DCS-18A1`.

1 INTRODUCTION

Graph neural networks (GNNs) Kipf & Welling (2017); Veličković et al. (2018); Xu et al. (2018) have achieved remarkable progress across diverse graph tasks, but their out-of-distribution (OOD) generalization remains challenging, largely due to the tendency to exploit spurious structures which are statistically correlated with labels in the training environments, rather than stable causal structures truly causal to the labels, leading to substantial performance degradation under distribution shifts Hu et al. (2020); Gui et al. (2022). Existing approaches typically seek to identify label-relevant causal or invariant subgraphs based on their transferability across samples from diverse environments Chen et al. (2022; 2023); Yao et al. (2024a;b). In contrast, non-causal subgraphs, which are more prone to overfitting sample- or environment-specific correlations, are commonly considered less generalizable.

However, this view may be overly simplistic. Our analysis reveals that non-causal subgraphs can also exhibit nontrivial cross-sample *generalization*. Specifically, as shown in Figure 1, spurious subgraphs (blue) may provide predictive cues conflicting with those encoded by causal subgraphs (orange and green), with the causal subgraph supporting the ground-truth class while the spurious subgraph favors a different class, a phenomenon we refer to as ***cue discordance***. Current works mainly attribute predictive ambiguity to the presence of diverse causal patterns associated with the same class Liu et al. (2025); Sun et al. (2026). Our analysis, however, reveals that a substantial fraction of conflicts instead arises from spurious subgraphs correlated with **other classes**, thereby inducing cue discordance that biases predictions toward incorrect and make optimization difficult.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

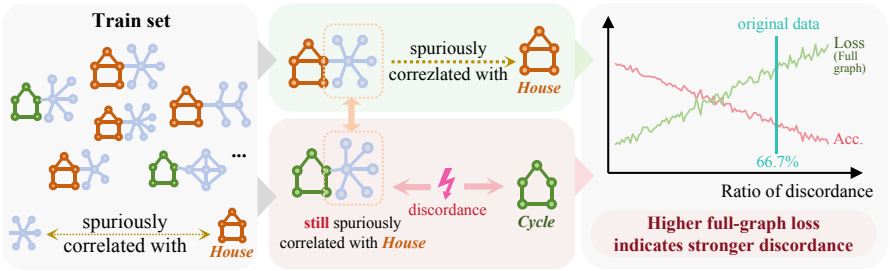


Figure 1: **Cue discordance captures the conflict between causal and spurious predictive cues.** When causal and spurious subgraphs support the same class, their cues are difficult to distinguish. When the spurious subgraph instead favors a different class, cue discordance arises. Increasing discordance leads to lower accuracy and higher loss, indicating that it contributes to performance degradation and that full-graph loss can proxy latent discordance strength.

To systematically assess the effect of cue discordance on model behavior, we construct fixed-size test subsets with controlled proportions of discordant samples on synthetic dataset Motif-basis Gui et al. (2022). As shown in Figure 1, increasing the prevalence of cue discordance leads to a consistent decline in predictive performance, accompanied by a corresponding increase in classification loss. We further quantify the prevalence of discordant samples in the original dataset and find that cue discordance indeed constitutes a substantial fraction of the data (66.7%). These observations suggest that such discordance contributes markedly to performance degradation and the loss brings hidden discordance to light and captures their severity.

In light of the revealed discordance, we further establish a theoretical characterization of this phenomenon. Under widely adopted assumptions in previous works, we prove that, when causal and non-causal subgraphs induce cue discordance, the risk of the causal subgraph is separated from that of the non-causal subgraph by a non-trivial margin. We formalize this relative risk separation as the Discordance-Conditioned Causal Contrast (D3C) criterion, which provides a principled grounding for distinguishing causal structures from spurious counterparts in revealed discordance.

Motivated by these theoretical and empirical findings, we propose a Discordance-Oriented Causal Subgraph discrimination (DCS) framework for graph OOD generalization. We use the relative full-graph classification risk as an observable proxy for latent discordance strength, assigning greater emphasis to samples exhibiting stronger signs of discordance. Guided by the D3C criterion, we enforce a stronger predictive advantage of the candidate subgraph over its non-causal complement as discordance intensifies. This enables effective causal subgraph identification without requiring environment annotations and complex data augmentation. Our contributions are summarized as follows:

- We highlight cue discordance between causal and spurious structures as a previously underexplored source of graph OOD degradation, and empirically demonstrate such discordance are prevalent and closely associated with performance degradation and increased classification loss.
- We theoretically establish the Discordance-Conditioned Causal Contrast (D3C) criterion, proving that, under such discordance, causal and non-causal subgraphs exhibit a non-trivial relative risk separation, providing a principled grounding for distinguishing causal subgraphs from spurious.
- We propose a Discordance-Oriented Causal Subgraph discrimination (DCS) framework that enforces the D3C criterion through discordance-aware relative subgraph discrimination, enabling robust graph OOD generalization without complex environment exploration or data augmentation.
- We conduct extensive experiments on 8 datasets, demonstrating DCS consistently improves graph OOD performance over existing methods and validates the discordance-oriented characterization.

2 Preliminaries

Notations: Let $G := (\mathcal{V}, \mathcal{E})$ be an undirected graph with n nodes and m edges, represented by its adjacency matrix A and node feature matrix $X \in \mathbb{R}^{n \times d}$ with d feature dimensions. We write G_c

and G_s for the invariant (causal) and spurious subgraphs of G. $g_\theta : \mathcal{G} \mapsto \mathcal{G}$ and $h_\phi : \mathcal{G} \mapsto \mathcal{Y}$ denote the subgraph extractor and predictor with parameter θ, ϕ respectively, and Z represents the extracted subgraph Z given by $Z = g_\theta(G)$. $\widehat{Y}_Z$ is the predicted label given by $\widehat{Y}_Z = h_\phi(Z)$ while Y is the true label. We denote the training and testing set as $\mathcal{D}_{tr} = \{G^e\}_{e\in\mathcal{E}_{tr}}, \mathcal{D}_{te} = \{G^e\}_{e\in\mathcal{E}_{te}}, \mathcal{E}_{tr} \neq \mathcal{E}_{te}$.

Problem Definition: Random train–test splits assume i.i.d. data, but real-world variations in collection environments and processing methods often violate this assumption. Splitting can result in distributional differences between the training and test sets. The out-of-distribution problem refers to the situation where the test set is drawn from a distribution different from that of the training set.

In this study, we focus on OOD generalization in graph classification. Given a collection of graph datasets $\mathcal{D} = \{G^e\}_{e\in\mathcal{E}_{tr}\subseteq\mathcal{E}_{all}}$, the objective of OOD generalization on graphs is to learn an optimal GNN model $f^*(\cdot) : \mathcal{G} \to \mathcal{Y}$ with data from training environments $\mathcal{D}_{tr} = \{G^e\}_{e\in\mathcal{E}_{tr}}$ that effectively generalizes across all (unseen) environments:

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

where $\mathcal{R}(f \mid e) \;=\; \mathbb{E}^e_{G,Y}\big[\ell\big(f(G),\, Y\big)\big]$ is the risk of predictor $f(\cdot)$ in environment e , and $\ell(\cdot,\cdot)$ denotes the loss function. Specifically, $f(\cdot) = g_\theta \circ h_\phi$ in subgraph-based methods.

Data Generating Process in our work follows prior study Chen et al. (2023); Gui et al. (2023); Li et al. (2022); Miao et al. (2022); Yao et al. (2024a;b) and is founded on Structural Causal Models Peters et al. (2016), which elucidate how latent factors give rise to observable graph properties. We model graph creation via a function $f_{\text{gen}} : \mathcal{Z} \to \mathcal{G}$. where $\mathcal{Z} \subseteq \mathbb{R}^n$ is the latent space and $\mathcal{G}$ the space of graphs. Under this SCM perspective, generation decomposes into three stages—first producing the causal subgraph G_c, then the spurious subgraph G_s, and finally the full observed graph G (omitting noise and task-irrelevant factors for simplicity):

$$G_c \;:=\; f^{G_c}_{\text{gen}}(C), \quad G_s \;:=\; f^{G_s}_{\text{gen}}(S), \quad G \;:=\; f^{G}_{\text{gen}}(G_c,\, G_s), \tag{2}$$

Let c, s be latent codes for causal and spurious factors. The graph G splits into a causal subgraph G_c (from c) and a spurious subgraph G_s(from s). c drives target Y, while s vary across environments.

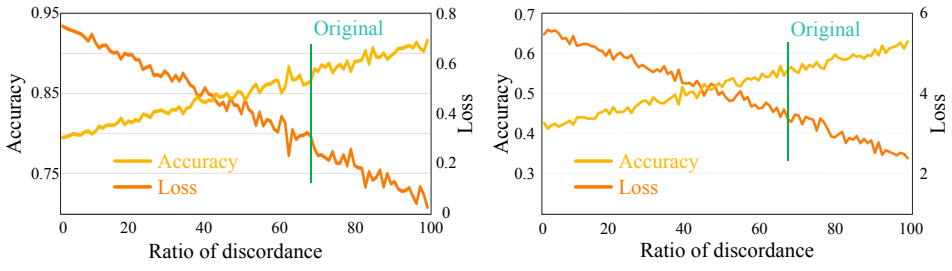
Distributional Shift can lead to performance degradation of the model on the test set. Depending on the type of the shift, it can be categorized into concept shift, covariate shift, etc Quinonero-Candela et al. (2022). Covariate shift refers to the change in the distribution $P(X)$ between the training and test splits, whereas concept shift refers to the change in the conditional distribution $P(Y \mid X)$. These two shift types typically call for different methodologies. In this study, we focus on concept shift, which often engenders *spurious correlations* between non-causal features and labels. These correlations can be easily exploited by models as the *shortcut* Wu et al. (2021), critically degrading model performance and hindering generalization Chang et al. (2020); Du et al. (2023).

3 Cue Discordance Reveals Causal–Spurious Discrimination

3.1 Empirical Observation: How Cue Discordance Affects Model Prediction

In this section, we conduct an in-depth investigation into how cue discordance affects model behavior. In real-world datasets, spurious correlations are often highly complex and difficult to observe, as multiple latent factors may induce correlations of varying strengths. Therefore, we adopt the synthetic Motif-basis dataset Gui et al. (2022) and the CMNIST handwritten-digit dataset Arjovsky et al. (2020), where spurious correlations can be explicitly introduced and systematically controlled. In both datasets, labels are determined by specific causal structures, while task-irrelevant factors—the basis structure/digit color—are spuriously correlated with the labels. This allows us to explicitly determine whether the causal and spurious components of each sample support the same or different labels. We construct test sets with a fixed number of samples and gradually vary the proportion of conflicting samples from 0% to 100%. As shown in Figure 2, the green vertical lines indicate the discordance ratios in the original test sets. From these experiments, we obtain these observations:

• Even without artificially increasing the discordance ratio, a substantial fraction of samples already exhibit disagreement between causal and spurious structures in the original dataset, suggesting that cue discordance are not rare corner cases but naturally occur under spurious correlations.



(a) Motif-basis (b) CMNIST-color

Figure 2: Effect of cue discordance on model behavior. We vary the proportion of conflicting samples from 0% to 100% on (a) Motif-basis and (b) CMNIST-color while keeping the test-set size fixed. As the discordance ratio increases, accuracy consistently decreases while the full-graph classification loss increases. The green vertical lines indicate the conflict ratios in the original datasets.

• As the proportion of discordant samples increases, model accuracy consistently decreases on both datasets. More importantly, for the model trained on the original data, discordant samples account for 66.4% of the entire test set, but 89.5% of the misclassified samples, indicating that cue discordance can be an important source of model degradation under distribution shifts.

• **Full-graph classification loss provides an observable signal of discordance strength.** As the discordance ratio increases, the classification loss also rises consistently. Since the causal–spurious decomposition is generally unavailable in practice, this observation suggests that full-graph loss can serve as a practical proxy for the severity of cue discordance, further discussion is in Appendix A.

3.2 Theoretical Analysis: Discordance-Conditioned Causal Contrast

Observations in Section 3.1 motivate a further question: *when discordance, where is the difference between causal and non-causal parts?* In this section, we show that cue discordance exposes a non-trivial relative risk separation between them, which we formalize as the Discordance-Conditioned Causal Contrast (D3C) criterion, providing a principled basis for distinguishing causal structures from their non-causal counterparts.

Consider an input graph decomposed as $G = (G_c, G_s)$, where G_c denotes the causal subgraph and $G_s = G/G_c$ the remaining non-causal structure. For simplicity and clarity, in the following proofs, we use C and S to denote G_c and G_s, respectively. Let

$$\eta_k(C) := P(Y = k \mid C), \qquad k \in \{1, \ldots, K\}, \tag{3}$$

denote the conditional class posterior given C, and let $q_C(\cdot \mid C)$ and $q_S(\cdot \mid S)$ denote the predictive distributions based on C and S, respectively. Then we define the cue discordance as follows:

Definition 1 (Cue discordance)**.** Define the classes favored by causal and non-causal structures as:

$$a(C) = \arg\max_k q_C(k \mid C), \qquad b(S) = \arg\max_k q_S(k \mid S). \tag{4}$$

A cue discordance occurs when $a(C) \neq b(S)$, and we denote the set of such samples by

$$\mathcal{H} = \{(C, S) : a(C) \neq b(S)\}, \quad p_{\mathcal{H}} = P\big((C, S) \in \mathcal{H}\big) \tag{5}$$

Then we make the following assumptions which serve as premise in many existing works:

Assumption 1. The causal subgraph C is the sole causal determinant of the label Y, such that

$$Y \perp S \mid C, \quad \text{i.e.,} \quad P(Y \mid C, S) = P(Y \mid C). \tag{6}$$

Assumption 2. Let $a^*(C) = \arg\max_k \eta_k(C)$ denote the unique Bayes-optimal class induced by the causal subgraph. We assume that the causal predictor is consistent with the Bayes decision,

$$a(C) = \arg\max_k q_C(k \mid C) = a^*(C), \tag{7}$$

while allowing imperfect estimation of the true conditional distribution. Specifically, its conditional excess cross-entropy risk is bounded as follows for some $\varepsilon_C \geq 0$:

$$D_{\mathrm{KL}}(\eta(C) \,\|\, q_C(\cdot \mid C)) \leq \varepsilon_C. \tag{8}$$

Discussion of the validity of these assumptions is provided in the Appendix B.1. With these assumptions, we establish a qualitative separation between causal and non-causal subgraphs:

Lemma 1 (Posterior Advantage under discordance)**.** *Under Assumption 2, for every* $(C, S) \in \mathcal{H}$*,*

$$\Delta(C, S) = \eta_{a(C)}(C) - \eta_{b(S)}(C) > 0. \tag{9}$$

The proof of this lemma is provided in Appendix B.3. This lemma establishes the qualitative result that cue discordance necessarily exposes a strict predictive advantage of the causal mechanism over the spurious one. Notably, this result does not require any fixed margin assumption. Since our empirical objective is based on predictive loss, we further establish the corresponding result under cross-entropy risk. Given the definitions of q_C and q_S, the conditional cross-entropy risks are:

$$\mathcal{R}^C_{\mathrm{CE}} = \mathbb{E}_{Y|C,S}\left[-\log q_C(Y \mid C)\right], \quad \mathcal{R}^S_{\mathrm{CE}} = \mathbb{E}_{Y|C,S}\left[-\log q_S(Y \mid S)\right]. \tag{10}$$

Lemma 2 (Cross-Entropy Risk Margin)**.** *Under Assumptions 1, 2, for every* $(C, S) \in \mathcal{H}$*, we have:*

$$\mathcal{R}^S_{\mathrm{CE}} - \mathcal{R}^C_{\mathrm{CE}} \geq \frac{1}{2}\Delta(C, S)^2 - \varepsilon_C. \tag{11}$$

The proof of this lemma is provided in Appendix B.4. This lemma shows that posterior discordance induces a lower bound on the relative predictive-risk gap, up to the approximation error ε_C, providing an justification for using relative predictive loss as a signal for causal-subgraph discovery.

Lemma 1 establishes a strictly positive posterior advantage on every non-degenerate discordant sample, while Lemma 2 translates this advantage into a predictive-risk separation up to the causal approximation error ε_C. However, this pointwise advantage may become arbitrarily small near the Bayes decision boundary, limiting its discriminative value in practice. We therefore introduce a standard multiclass Tsybakov margin condition Audibert & Tsybakov (2007); Mammen & Tsybakov (1999); Tsybakov (2004) to control the probability mass of such near-boundary cases. This allows us to move from a merely positive separation to a high-probability, non-trivial margin between the causal and spurious mechanisms.

Assumption 3 (Multiclass Tsybakov Margin Condition)**.** Let $a(C)$ denote the unique Bayes-optimal class induced by the causal subgraph, and define the multiclass posterior margin as

$$m(C) := \eta_{a(C)}(C) - \max_{k \neq a(C)} \eta_k(C). \tag{12}$$

Under the unique Bayes decision in Assumption 2, we have $m(C) > 0$. Following the multiclass extension of the Tsybakov margin condition, we assume that there exist constants $A \geq 1$ and $\alpha > 0$ such that, for all $t \in (0, 1]$,

$$P\left(m(C) \leq t\right) \leq A t^{\alpha}. \tag{13}$$

Theorem 1. *Combining Lemmas 1 and 2 with Assumption 3, for any* $\varepsilon \in (0, 1)$ *such that* $\gamma_\varepsilon := \left(\frac{\varepsilon p_{\mathcal{H}}}{A}\right)^{1/\alpha}$ *lies in the valid margin range:*

$$P(\Delta(C, S) > \gamma_\varepsilon \mid \mathcal{H}) \geq 1 - \varepsilon.$$

Moreover, if $\varepsilon_C < \frac{1}{2}\gamma_\varepsilon^2$*, , letting* $\delta_\varepsilon := \frac{1}{2}\gamma_\varepsilon^2 - \varepsilon_C > 0$*, we have*

$$P\big(\mathcal{R}^C_{\mathrm{CE}} + \delta_\varepsilon < \mathcal{R}^S_{\mathrm{CE}} \mid \mathcal{H}\big) \geq 1 - \varepsilon.$$

The proof of this theorem is provided in Appendix B.5. This theorem establishes a population-level relative risk margin between the causal and non-causal subgraphs under cue discordance. However, the conditional risks $\mathcal{R}^C_{\mathrm{CE}}$ and $\mathcal{R}^S_{\mathrm{CE}}$ are not directly observable from individual training samples. We therefore establish the following result, which connects the population-level risk ordering to a tractable sample-wise surrogate objective.

Proposition 1. *Let* $\ell_C = -\log q_C(Y \mid C), \ell_S = -\log q_S(Y \mid S)$*. For any margin* $\delta \geq 0$*,*

$$\left[\delta + \mathcal{R}^C_{\mathrm{CE}} - \mathcal{R}^S_{\mathrm{CE}}\right]_+ \leq \mathbb{E}\left[\left[\delta + \ell_C - \ell_S\right]_+ \mid C, S\right]. \tag{14}$$

Hence, $\mathcal{L}_{\mathrm{rel}} = \left[\delta + \ell_C - \ell_S\right]_+$ *provides a stochastic sample-wise surrogate whose conditional expectation upper-bounds violations of the desired relative-risk ordering* $\mathcal{R}^C_{\mathrm{CE}} + \delta \leq \mathcal{R}^S_{\mathrm{CE}}$*.*

The proof of this proposition is provided in Appendix B.6. Theorem 1 and Proposition 1 establish that cue discordance defines a discriminative regime in which causal subgraphs exhibit a non-trivial predictive advantage over spurious structures with high probability, and that this population-level advantage can be operationalized through a sample-wise relative-loss surrogate. This result also connects directly to the empirical observation in Section 3.1.

The above theoretical results provide an identification principle: under cue discordance, the true causal component should exhibit a predictive-risk advantage over its non-causal counterpart. Therefore searching for a candidate subgraph whose predictive advantage over its complement best satisfies this relative-risk criterion. The theoretical guarantee characterizes the ideal causal–spurious decomposition. Building on these empirical observations and theoretical results, we formalize the ***Discordance-Conditioned Causal Contrast*** **(D3C) criterion**. It states that, under cue discordance, causal subgraphs exhibit a non-trivial predictive risk advantage over their non-causal counterparts. Formally, D3C characterizes the following discordance-conditioned risk separation:

$$\min_{Z(\cdot)} \mathbb{E}_{(G,Y)\sim P_{\mathrm{tr}}} \Big[w(G) \left[\delta + \ell_\theta(Y \mid Z) - \ell_\theta\big(Y \mid Z^-\big) \right]_+ \Big], \tag{15}$$

where $Z \subseteq G$ denotes a candidate subgraph and $Z^- = G \setminus Z(G)$ its complementary structure. The weight $w(G)$ measures the cue discordance of G, receiving greater emphasis in enforcing the relative causal-spurious margin.

4 METHODOLOGY

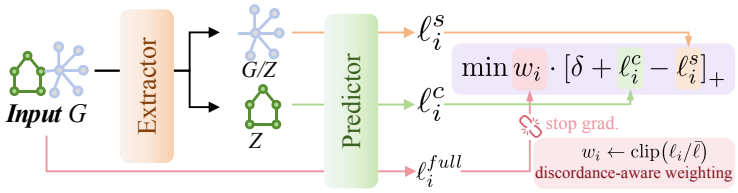


Figure 3: Overview of the DCS framework. The extractor partitions an input graph G into a candidate subgraph Z and its complement $Z^- = G \setminus Z$ evaluated by a shared predictor. Full-graph loss provides a stop-gradient discordance-aware weight w_i. Guided by D3C criterion, DCS enforces a margin-based predictive advantage of Z over Z^-, with greater emphasis on stronger discordance.

4.1 FRAMEWORK

We develop an extract–predict paradigm for graph generalization. Let $\mathcal{D}_{\mathrm{tr}} = (G_i, y_i)i = 1^N$ denote the training set, where $G_i = (A_i, X_i)$ consists of an adjacency matrix A_i and node attributes X_i. Our framework consists of a subgraph extractor g_θ and a shared predictor h_ϕ.

Given an input graph $G = (A, X)$, the extractor first computes node representations as $H = g_\theta(G) \in \mathbb{R}^{n\times d}$. For each observed edge $(u, v) \in \mathcal{E}(G)$, its edge score is computed from the representations, followed by a Top-r selection:

$$a_{uv} = \mathrm{MLP}_\theta\big([H_u \| H_v]\big), s_{uv} = \begin{cases} \sigma((a_{uv} + a_{vu})/2)\,, & \text{undirected,} \\ \sigma(a_{uv})\,, & \text{directed,} \end{cases}$$

$$M_{uv} = s_{uv}\, \mathbf{1}[(u, v) \in \mathrm{Top}_r(s)]\,.$$

The learned mask induces a selected candidate rationale Z and its complementary view Z^-:

$$Z = (A \odot M, X), \qquad Z^- = (A \odot (\mathbf{1} - M), X)$$

Here, Z is referred to as a candidate causal subgraph, while Z^- retains the residual information.

A shared predictor evaluates all three graph views:

$$z_v = \mathrm{Readout}\,(h_\phi(v))\,, \qquad p_v = \mathrm{softmax}\,(c_\phi(z_v))\,, \qquad v \in Z, Z^-, G.$$

The shared parameterization enables a direct comparison between the selected and complementary views. The selected causal subgraph is used for final prediction, while the complementary view supports relative discrimination and the full-graph loss provides a proxy of latent cue discordance.

4.2 Discordance-oriented Relative Optimization

Motivated by the D3C criterion, we design a Discordance-oriented Causal Subgraph (DCS) discrimination framework, which uses the full-graph classification loss to characterize latent discordance strength and adaptively enlarges the predictive advantage of the candidate causal subgraph.

For a mini-batch $\mathcal{B}$ of B graphs, we define the cross-entropy loss of each view as $\ell^v = -\log p_v(y \mid v)$. The predictor is supervised using both the selected rationale and the full graph:

$$\mathcal{L}_{\text{task}} = \sum_{i=1}^{B} \left(\ell_i^c + \ell_i^{\text{full}}\right), \tag{16}$$

The selected-view loss preserves label-predictive information in the extracted rationale, while the full-graph loss provides a model-dependent estimate of cue discordance. Specifically, we compute the discordance weight from the detached full-graph loss:

$$d_i = \text{sg}\left(\ell_i^{\text{full}}\right), \qquad w_i = \text{clip}\left(\frac{d_i}{B^{-1}\sum_{j=1}^{B} d_j + \varepsilon}, w_{\min}, w_{\max}\right),$$

where $\text{sg}(\cdot)$ denotes stop-gradient. We set $\varepsilon = 10^{-8}$, $w_{\min} = 0.5$, and $w_{\max} = 2.0$. Batch normalization makes w_i represent relative prediction discordance within the current mini-batch, while clipping prevents a small number of extreme samples from dominating optimization. Stop-gradient further prevents the relative objective from trivially manipulating its own discordance weights.

The role of w_i follows directly from the motivation of D3C criterion. Cue discordance does not by itself determine whether a particular structure is causal. Instead, discordant samples are assigned greater importance because they are more likely to expose disagreement between competing predictive mechanisms, making the relative advantage between the selected and complementary structures more informative. Thus, all samples participate in training, while samples with greater cue discordance receive stronger emphasis in causal–spurious discrimination. We operationalize the theoretical relative-risk ordering through the following discordance-weighted sample-wise surrogate:

$$\mathcal{L}_{\text{rel}} = \frac{1}{B}\sum_{i=1}^{B} \underbrace{w_i}_{\text{discordance evaluation}} \cdot \underbrace{[\delta + \ell_i^c - \ell_i^s]_+}_{\text{relative causal advantage}} \qquad [t]_+ = \max(0, t). \tag{17}$$

where $\delta > 0$ denotes the desired predictive margin. This objective is the empirical realization of the surrogate established in Proposition 1. The margin δ can be viewed as the practical finite-sample counterpart of the positive relative-risk margin established in Theorem 1 and Proposition 1.

We also include a compactness regularization term $\mathcal{L}_{\text{comp}} = -\left[(1-M)\log(1-M) + M\log M\right]$ from previous works to encourage the selected subgraph to be compact Ying et al. (2019); Luo et al. (2020). Together with $\mathcal{L}_{\text{task}}$, the overall objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{rel}\mathcal{L}_{\text{rel}} + \lambda_{comp}\mathcal{L}_{\text{comp}} \tag{18}$$

We employ a warm-up strategy before activating the relative margin objective. Specifically, during the first $T_{\text{warm}} = 50$ epochs, the model is optimized without $\mathcal{L}_{\text{rel}}$, allowing the extractor and predictor to first acquire a reliable predictive representation from the selected and full-graph views.

5 Experiments

5.1 Datasets, Metrics and Baselines

Datasets and Metrics. We adopt widely used benchmark GraphOOD Gui et al. (2022) for graph OOD generalization, including Motif, CMNIST, HIV, SST2, and Twitter, covering synthetic, superpixel, molecular, and text graphs. We follow the concept-shift setting, where labels can be spuriously correlated with structural or feature domains, weaker correlations in validation, and no spurious correlations in testing. We report ROC-AUC for binary tasks and accuracy for the others. All experiments are repeated three times with different random seeds.

Baselines. We compare our method with several competitive baselines, including empirical risk minimization (ERM), four traditional out-of-distribution baselines including IRM Arjovsky et al.

Table 1: Results on synthetic, semi-synthetic, and real-world datasets over 3 runs. For Motif, SP Rate denotes the spurious correlation probability.

Method	Motif						CMNIST		HIV		SST2	Twitter
	Basis-0.6	Basis-0.7	Basis-0.8	Basis-0.9	Basis-0.95	Size-0.95	Color	Background	Scaffold	Size	Length	Length
IRM	91.85(0.24)	91.25(0.29)	90.48(0.08)	85.38(0.40)	78.67(0.39)	67.11(1.02)	42.68(0.53)	32.00(0.18)	70.89(1.08)	72.48(1.25)	73.89(1.72)	48.15(0.82)
ERM	91.67(0.02)	91.25(0.39)	90.63(0.16)	85.92(0.21)	78.66(0.22)	67.77(0.36)	43.45(0.22)	32.05(0.27)	67.72(1.08)	71.34(3.42)	72.44(0.91)	43.75(0.68)
VREX	91.52(0.13)	91.28(0.37)	90.45(0.04)	85.40(0.25)	79.19(0.65)	68.90(1.34)	42.74(0.33)	32.27(0.77)	68.85(1.04)	67.13(6.95)	72.50(0.81)	50.85(1.29)
Coral	91.68(0.33)	91.12(0.33)	90.50(0.08)	85.75(0.11)	79.51(0.63)	67.57(1.05)	42.95(0.46)	31.17(0.23)	71.27(0.59)	70.40(3.49)	73.05(0.57)	49.27(1.13)
DANN	91.57(0.05)	91.22(0.22)	90.27(0.14)	85.97(0.51)	79.25(0.67)	68.86(1.52)	43.44(0.73)	31.79(0.55)	71.74(1.68)	66.38(7.13)	73.75(0.69)	51.50(0.64)
GIL	89.78(3.21)	80.33(0.32)	78.72(1.72)	69.72(0.94)	68.60(0.31)	48.82(1.74)	23.98(2.32)	24.16(2.13)	67.17(0.95)	71.17(1.42)	73.77(0.42)	52.43(1.29)
DIR	88.70(2.66)	87.82(1.08)	87.43(1.39)	84.80(2.50)	80.20(2.45)	54.96(3.46)	30.07(5.81)	17.93(3.17)	71.85(1.12)	71.20(5.20)	69.73(3.05)	52.49(2.74)
GSAT	74.25(3.26)	72.87(2.14)	68.49(1.97)	66.70(3.55)	65.83(3.65)	59.81(3.30)	32.28(0.92)	14.20(0.38)	70.84(0.14)	67.17(7.39)	73.64(0.87)	47.33(0.52)
CIGA	84.90(0.37)	82.28(5.96)	83.57(8.83)	77.10(6.41)	76.93(2.34)	69.73(0.49)	25.01(1.40)	17.08(4.04)	71.81(1.82)	72.02(2.87)	65.75(2.66)	48.97(2.54)
GALA	–	–	–	–	–	–	44.57(0.49)	43.88(0.36)	71.47(1.23)	70.25(3.21)	76.45(0.98)	52.14(0.67)
LECI	92.22(0.36)	89.77(0.53)	89.25(0.21)	81.05(0.47)	73.43(2.28)	66.50(1.57)	50.61(0.73)	50.98(0.44)	68.65(1.18)	54.02(2.23)	76.25(0.83)	51.72(0.74)
iMoLD	89.27(0.73)	88.46(0.31)	87.11(0.14)	83.44(0.79)	67.10(0.88)	66.48(1.23)	49.63(0.42)	49.72(0.36)	72.04(0.34)	67.87(0.82)	72.39(0.41)	50.34(1.65)
EQuAD	88.74(0.86)	88.29(0.21)	88.21(0.37)	84.22(0.47)	72.37(0.69)	68.32(0.53)	50.49(0.68)	49.97(0.36)	72.18(0.37)	68.41(0.13)	69.34(0.54)	52.33(0.71)
LIRS	88.97(0.56)	88.94(0.36)	88.89(0.65)	84.74(0.39)	73.15(1.23)	69.34(0.47)	51.08(0.29)	50.32(0.37)	72.46(0.72)	68.69(0.88)	69.79(0.36)	53.46(1.82)
UIL	87.63(0.53)	88.91(0.29)	88.83(0.74)	84.77(0.47)	79.54(1.31)	69.57(0.53)	49.84(0.34)	47.21(0.77)	71.45(0.96)	67.34(0.23)	71.07(0.44)	52.65(0.93)
PrunE	90.45(0.55)	90.76(0.74)	89.25(0.93)	87.63(0.78)	77.63(1.74)	67.84(0.76)	52.37(0.58)	51.65(0.72)	72.03(0.43)	69.73(0.79)	72.39(1.01)	53.17(1.54)
DCS	**92.34**(0.33)	**91.77**(0.26)	**91.23**(0.45)	**90.49**(0.73)	**84.86**(0.76)	**72.08**(1.04)	**54.13**(1.13)	**54.34**(1.07)	**72.56**(3.24)	**74.47**(1.69)	**78.71**(1.94)	**56.67**(1.79)

Table 2: Ablation study (%). $\mathcal{L}_{\text{comp}}$ denotes the compactness loss; CDW denotes Discordance Weighting w_i; w/o RC removes Relative Contrast between ℓ_i^c and ℓ_i^s; w/o margin removes the non-trivial separation margin δ; w/o $\mathcal{L}_{\text{rel}}$ removes complete relative discrimination objective.

Variant	**Motif-basis**	**Motif-size**	**CMNIST-bg**	**HIV-size**
w/o $\mathcal{L}_{\text{comp}}$	81.63	69.54	50.51	72.39
w/o CDW	80.64	67.53	49.37	70.06
w/o RC	79.27	64.75	47.87	69.59
w/o margin	81.26	68.93	48.82	71.54
w/o $\mathcal{L}_{\text{rel}}$	74.23	60.57	43.29	69.32
DCS	**84.86**	**72.08**	**54.13**	**74.47**

Table 3: Edge-level causal subgraph quality comparison on Motif-basis with ground-truth edges. We report edge accuracy, recall, precision, and F1 score against the ground-truth causal subgraph. Results are obtained by selecting same ratio across all methods.

Method	Edge Acc.	Recall	Prec.	F1
GIL	0.892	0.40	0.72	0.51
GSAT	0.904	0.44	0.74	0.55
LECI	0.916	0.49	0.76	0.60
DCS	**0.934**	**0.57**	**0.79**	**0.66**

(2020), VREx Krueger et al. (2021), and Coral Sun & Saenko (2016), and eight graph-specific OOD baselines including IRM-based methods DIR Wu et al. (2021), GIL Li et al. (2022), iMoLD Zhuang et al. (2023) and IRM-free methods CIGA Chen et al. (2022), GALA Chen et al. (2023), LECI Gui et al. (2023), EQuAD Yao et al. (2024a), LIRS Yao et al. (2024b), UILSui et al. (2025), PrunEYao et al. (2025) as well as interpretable graph rationalization method GSAT Miao et al. (2022).

5.2 Main Results

Dealing with Spurious Correlation. We evaluate DCS to examine its ability to mitigate the adverse effects of cue discordance on the Motif dataset, where three motifs (house, cycle, and crane) are grafted onto three basis graphs (wheel, tree, and ladder). In the basis and size domains, labels are spuriously correlated with the basis type and size, respectively. As shown in Table 1, stronger spurious correlations increasingly encourage models to exploit shortcuts, resulting in degraded performance. In contrast, our method consistently outperforms all baselines and exhibits the smallest drop. Under the strongest correlation setting, with training ratios of (0.99, 0.97, 0.95) and validation/test ratios of 0.7/0, our method still improves over SOTA methods by 4.66% and 2.35% on the basis and size domains respectively, demonstrating its effectiveness handling discordance.

Real World Scenarios. To evaluate real-world generalization, we conduct experiments on the real-world datasets. DCS consistently performs well under diverse shifts, outperforming OOD baselines by 2.69%, 2.26%, 3.21%, and 3.27% on CMNIST-Bg, SST2, Twitter, and HIV-size, respectively.

5.3 Ablation Study

We conduct ablation studies on four representative datasets. As shown in Table 2, removing any component consistently degrades performance, with eliminating $\mathcal{L}_{\text{rel}}$ causing the largest drop. RC (Relative Contrast) and CDW (Cue Discordance Weighting) also contribute substantially, confirming the importance of relative subgraph contrast and cue-discordance-aware weighting. The margin term and $\mathcal{L}_{\text{comp}}$ provide further consistent gains, further improving the causal subgraph identification.

Table 4: Sensitivity analysis of the relative discrimination coefficient λ_{rel} and margin δ on representative datasets. Only one hyperparameter is varied at a time.

Dataset	λ_{rel} 0.25	0.5	1.0	2.0	4.0	δ 0	0.1	0.2	0.3	0.4
Motif-basis	84.26	84.86	84.07	83.27	83.53	81.23	84.86	84.27	84.33	82.69
CMNIST-color	53.30	53.79	54.13	54.06	53.21	48.96	50.47	54.13	54.54	53.67

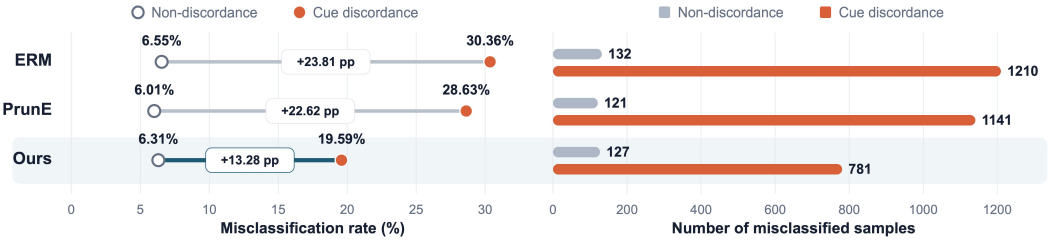


Figure 4: Misclassification rate and number of errors on cue-discordant and non-discordant samples in Motif-basis. Our method markedly reduces errors on cue-discordant samples.

5.4 Delve into Cue Discordance

Mitigating Cue Discordance. To investigate whether our method effectively mitigates the impact of cue discordance, we compare the misclassification rates on discordant and non-discordant samples. As shown in Figure 4, methods exhibit similar error rates on non-discordant samples, while cue discordance causes a substantial error increase for ERM/PrunE. In contrast, our method reduces the discordance-induced error gap from 23.81 and 22.62 pp (percentage points) to 13.28 pp, indicating that the improvement of our method mainly comes from its stronger robustness to cue discordance.

Subgraph Quality. We evaluate the quality of the extracted causal subgraphs on Motif-basis with ground-truth notation. As in Table 3, DCS consistently achieves the best edge-level performance across accuracy, recall, precision, and F1 score. These results suggest that DCS not only improves OOD prediction, but also identifies causal substructures more faithfully under the D3C criterion.

Sensitivity Study. As shown in Table 4, DCS remains stable over a broad range of λ_{rel} and δ, indicating limited sensitivity to precise hyperparameter choices. A proper margin generally performs better than 0. Empirically, we recommend slightly larger values of both parameters for datasets with more classes, as spurious cues are more likely to support different and exhibit more discordance.

6 Related Work

OOD generalization remains a critical challenge in graph learning, as models trained on specific distributions often fail to generalize to unseen ones. A prominent line of work builds on invariant risk minimization (IRM) Arjovsky et al. (2020) to learn causally relevant representations that remain stable across environments Yang et al. (2022); Zhuang et al. (2023); Li et al. (2022; 2024a); Wu et al. (2021); Liu et al. (2022); Chen et al. (2022); however, these methods typically rely on observed or inferred environment labels. Other approaches avoid explicit environment supervision by exploiting Infomax or Information Bottleneck principles to extract causal features Yao et al. (2024b;a); Miao et al. (2022); Wang et al. (2024); Wu et al. (2021); Gui et al. (2023); Chen et al. (2022), yet they only partially resolve the distinction between causal and spurious features. Recent studies explore multiple causal subgraphs, yet overlook the generalizability of non-causal ones Liu et al. (2025); Sun et al. (2026). In contrast, our method exploits cue discordance to derive a relative-risk criterion for efficient causal subgraph identification. More discussion is provided in the Appendix C.

7 Conclusion

In this work, we revisit graph OOD generalization through cue discordance between causal and spurious subgraphs. We reveal its prevalence and close association with prediction errors and degradation. We establish the D3C criterion and propose DCS for efficient discordance-aware causal subgraph identification. Extensive experiments demonstrate the effectiveness and robustness.

AI USE STATEMENT

In this work, we did not use generative AI tools for research idea generation, methodology or experimental design, theoretical development, mathematical claims or proofs, code implementation, data processing, or result interpretation. We used generative AI tools only for language polishing, grammar checking, and spelling correction. All AI-assisted edits were carefully reviewed by the authors to ensure that they did not alter the technical meaning, scientific claims, or conclusions of the manuscript. The authors take full responsibility for the final content of this work, including any text produced or revised with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility. All datasets used in this work are publicly available. Detailed experimental settings, dataset descriptions, evaluation protocols and implementation details are provided in the main paper, appendix and mentioned papers, together with complete derivations and proofs of the theoretical results. Part of the implementation is already publicly available at `https://anonymous.4open.science/r/DCS-18A1`, and the complete codebase will be released upon acceptance.

REFERENCES

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

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant Rationalization. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1448–1458. PMLR, 2020.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, M. A. 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*, 36:71486–71519, 2023.

Yanrui Du, Jing Yan, Yan Chen, Jing Liu, Sendong Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Bing Qin. Less Learn Shortcut: Analyzing and Mitigating Learning of Spurious Feature-Label Correlation, 2023.

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*, 36:3945–3978, 2023.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open Graph Benchmark: Datasets for Machine Learning on Graphs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22118–22133. Curran Associates, Inc., 2020.

Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In *Conference On Learning Theory*, pp. 1882–1886. PMLR, 2018.

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 *Proceedings of the 38th 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.

Haoyang Li, Xin Wang, Zeyang Zhang, Haibo Chen, Ziwei Zhang, and Wenwu Zhu. Disentangled Graph Self-supervised Learning for Out-of-Distribution Generalization. In *Forty-First International Conference on Machine Learning*, 2024a.

Xiner Li, Shurui Gui, Youzhi Luo, and Shuiwang Ji. Graph Structure Extrapolation for Out-of-Distribution Generalization. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 27846–27874, Vienna, Austria, 2024b. PMLR.

Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. ZIN: When and How to Learn Invariance Without Environment Partition? *Advances in Neural Information Processing Systems*, 35:24529–24542, 2022.

Bowen Liu, Haoyang Li, Shuning Wang, Shuo Nie, and Shanghang Zhang. Subgraph aggregation for out-of-distribution generalization on graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):18763–18771, 2025.

Gang Liu, Tong Zhao, Jiaxin Xu, Tengfei Luo, and Meng Jiang. Graph Rationalization with Environment-based Augmentations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1069–1078, 2022.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized Explainer for Graph Neural Network. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19620–19631. Curran Associates, Inc., 2020.

Subha Maity, Yuekai Sun, and Moulinath Banerjee. Minimax optimal approaches to the label shift problem in non-parametric settings. *Journal of Machine Learning Research*, 23(346):1–45, 2022.

Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and Generalizable Graph Learning via Stochastic Attention Mechanism. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 15524–15543. PMLR, 2022.

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.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2022. ISBN 978-0-262-54587-7.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*, 2019.

Yongduo Sui, Qitian Wu, Jiancan Wu, Qing Cui, Longfei Li, Jun Zhou, Xiang Wang, and Xiangnan He. Unleashing the Power of Graph Data Augmentation on Covariate Distribution Shift, 2023.

Yongduo Sui, Jie Sun, Shuyao Wang, Zemin Liu, Qing Cui, Longfei Li, and Xiang Wang. A Unified Invariant Learning Framework for Graph Classification. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.1, KDD 2025, Toronto, ON, Canada, August 3-7, 2025*, pp. 1301–1312, Toronto ON Canada, 2025. ACM. ISBN 979-8-4007-1245-6.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In Gang Hua and Hervé Jégou (eds.), *Computer Vision – ECCV 2016 Workshops*, pp. 443–450, Cham, 2016. Springer International Publishing. ISBN 978-3-319-49409-8.

Jerry Sun, Mohamed Abubakr Hassan, Yaoyu Zhang, Wanying Zhang, and Chi-Guhn Lee. Diverse and sparse mixture-of-experts for causal subgraph–based out-of-distribution graph learning. *International Conference on Learning Representations*, 2026:79573–79597, 2026.

Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *International Conference on Learning Representations*, 2018.

Luzhi Wang, Dongxiao He, He Zhang, Yixin Liu, Wenjie Wang, Shirui Pan, Di Jin, and Tat-Seng Chua. GOODAT: Towards Test-Time Graph Out-of-Distribution Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):15537–15545, 2024.

Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering Invariant Rationales for Graph Neural Networks. In *International Conference on Learning Representations*, 2021.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks? In *International Conference on Learning Representations*, 2018.

Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. Learning Substructure Invariance for Out-of-Distribution Molecular Representations. In *Advances in Neural Information Processing Systems*, 2022.

Tianjun Yao, Yongqiang Chen, Zhenhao Chen, Kai Hu, Zhiqiang Shen, and Kun Zhang. Empowering Graph Invariance Learning with Deep Spurious Infomax, 2024a.

Tianjun Yao, Yongqiang Chen, Kai Hu, Tongliang Liu, Kun Zhang, and Zhiqiang Shen. Learning Graph Invariance by Harnessing Spuriosity. In *The Thirteenth International Conference on Learning Representations*, 2024b.

Tianjun Yao, Haoxuan Li, Yongqiang Chen, Tongliang Liu, Le Song, Eric P. Xing, and Zhiqiang Shen. Pruning Spurious Subgraphs for Graph Out-of-Distribution Generalization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating Explanations for Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Junchi Yu, Jian Liang, and Ran He. Finding Diverse and Predictable Subgraphs for Graph Domain Generalization, 2022.

Junchi Yu, Jian Liang, and Ran He. Mind the Label Shift of Augmentation-based Graph OOD Generalization, 2023.

Xiang Zhuang, Qiang Zhang, Keyan Ding, Yatao Bian, Xiao Wang, Jingsong Lv, Hongyang Chen, and Huajun Chen. Learning Invariant Molecular Representation in Latent Discrete Space, 2023.

A Full-Graph Loss as a Practical Signal of Cue Discordance

Prior work has shown that models exploiting spurious correlations tend to incur higher errors on samples where such correlations fail (Sagawa* et al., 2019). In our setting, cue discordance represents an analogous conflict between causal and spurious predictive cues: when the two support different classes, their competing evidence tends to reduce confidence in the ground-truth prediction and increase the full-graph classification loss. We therefore do not interpret full-graph loss as an

exact estimator of cue discordance, since it may also reflect generic sample difficulty, label ambiguity, or optimization error. Rather, it serves as a practical soft proxy that enriches samples where causal–spurious disagreement is more likely to occur. This interpretation is further supported by our controlled experiments, where increasing the proportion of discordant samples consistently leads to higher full-graph loss. Accordingly, DCS uses the normalized loss only to softly reweight relative subgraph discrimination, rather than to explicitly identify discordant samples.

B Theoretical Analysis and Proofs

B.1 Dicussion about assumptions

Assumption 1 (Causal Sufficiency)**.** The causal subgraph C is the sole causal determinant of the label Y, such that

$$Y \perp S \mid C, \quad \text{i.e.,} \quad P(Y \mid C, S) = P(Y \mid C). \tag{19}$$

This assumption does not require S to be marginally independent of Y; S may still be strongly correlated with the label in the training distribution and thus serve as a predictive shortcut. This assumption naturally follows from the definition of causal subgraph and serves as a fundamental premise in many existing graph OOD generalization methods Gui et al. (2023); Wu et al. (2021); Li et al. (2022); Chen et al. (2022).

Assumption 2 (Approximate Causal Prediction)**.** Let $a^*(C) = \arg\max_k \eta_k(C)$ denote the unique Bayes-optimal class induced by the causal subgraph. We assume that the causal predictor is consistent with the Bayes decision,

$$a(C) = \arg\max_k q_C(k \mid C) = a^*(C), \tag{20}$$

while allowing imperfect estimation of the true conditional distribution. Specifically, its conditional excess cross-entropy risk is bounded as follows for some $\varepsilon_C \geq 0$:

$$D_{\mathrm{KL}}(\eta(C) \,\|\, q_C(\cdot \mid C)) \leq \varepsilon_C. \tag{21}$$

This assumption does not require the causal predictor to exactly recover the true conditional distribution. Instead, it only requires consistency with the Bayes-optimal decision while allowing a bounded posterior approximation error. Since cross-entropy is a strictly proper loss, the excess conditional cross-entropy risk relative to the Bayes predictor is exactly characterized by $D_{\mathrm{KL}}(\eta(C) \| q_C(\cdot \mid C))$. Thus, ε_C quantifies the deviation of the learned causal predictor from the population optimum. When $\varepsilon_C = 0$, this assumption reduces to the realizable setting $q_C(\cdot \mid C) = \eta(C)$.

Assumption 3 (Multiclass Tsybakov Margin Condition)**.** Let $a(C)$ denote the unique Bayes-optimal class induced by the causal subgraph, and define the multiclass posterior margin as

$$m(C) := \eta_{a(C)}(C) - \max_{k \neq a(C)} \eta_k(C). \tag{22}$$

Under the unique Bayes decision in Assumption 2, we have $m(C) > 0$. Following the multiclass extension of the Tsybakov margin condition, we assume that there exist constants $A \geq 1$ and $\alpha > 0$ such that, for all $t \in (0, 1]$,

$$P\left(m(C) \leq t\right) \leq A t^{\alpha}. \tag{23}$$

This condition controls the probability mass near the Bayes decision boundary: rather than requiring a uniform separation between the most probable and competing classes, it only requires samples with an arbitrarily small posterior margin to become increasingly rare. This assumption has been widely adopted in statistical classification works Kpotufe & Martinet (2018); Maity et al. (2022) (refer to the Appendix B.2 for more details)

B.2 Relation to the Tsybakov margin condition.

The Tsybakov margin condition Audibert & Tsybakov (2007); Mammen & Tsybakov (1999); Tsybakov (2004) is a standard low-noise assumption in statistical learning theory, introduced to characterize the probability mass near the Bayes decision boundary. Its multiclass extensions measure this ambiguity through the posterior gap between the Bayes-optimal class and its closest competitor.

Rather than imposing a uniform hard margin, it only requires samples with arbitrarily small posterior gaps to occur with sufficiently low probability. This condition has been widely adopted in theoretical analyses of classification to derive excess-risk and fast-rate guarantees, and thus provides a standard and mild regularity assumption for controlling near-boundary samples Kpotufe & Martinet (2018); Maity et al. (2022).

We briefly review the standard Tsybakov margin condition and clarify its relation to Assumption 3. Consider first binary classification, where $Y \in \{0, 1\}$ and

$$\eta(x) := P(Y = 1 \mid X = x).$$

The Bayes decision boundary is characterized by $\eta(x) = 1/2$. A commonly used form of the Tsybakov margin, or low-noise, condition assumes that there exist constants $A > 0$ and $\alpha > 0$ such that

$$P_X(0 < |2\eta(X) - 1| \le t) \le At^\alpha, \qquad \forall\, t \in (0, 1]. \tag{24}$$

Equivalently, up to a rescaling of the constant A, the condition is often written in terms of $|\eta(X) - 1/2|$. It requires that the probability mass of samples whose class posterior is close to the Bayes decision boundary decreases polynomially with the margin threshold t. A larger α therefore corresponds to less probability mass near the decision boundary, whereas the condition does not require a uniformly positive margin for every sample.

For multiclass classification with $Y \in \{1, \dots, K\}$, let

$$\eta_k(x) := P(Y = k \mid X = x), \qquad a(x) := \arg\max_k \eta_k(x),$$

and suppose that the Bayes-optimal class $a(x)$ is unique. The natural multiclass posterior margin is

$$m(x) := \min_{k \neq a(x)} \left[\eta_{a(x)}(x) - \eta_k(x)\right]. \tag{25}$$

Since $a(x)$ is the posterior-maximizing class, this quantity is equivalently the gap between the largest and the second-largest class posterior. A standard multiclass extension of the Tsybakov condition takes the form

$$P_X(0 < m(X) \le t) \le At^\alpha, \qquad \forall\, t \in (0, 1]. \tag{26}$$

Thus, as in the binary case, the condition controls the amount of probability mass concentrated near the Bayes decision boundary, now measured through the posterior separation between the Bayes-optimal class and its strongest competing class.

Assumption 3 is precisely this multiclass Tsybakov condition specialized to the causal representation C. In particular, taking $X = C$, we have

$$a(C) = \arg\max_k \eta_k(C)$$

and

$$m(C) = \min_{k \neq a(C)} \left[\eta_{a(C)}(C) - \eta_k(C)\right],$$

so that Assumption 3 becomes

$$P(0 < m(C) \le t) \le At^\alpha.$$

Therefore, our assumption introduces no additional form of margin condition beyond the standard multiclass low-noise assumption; it simply applies the condition to the class posterior induced by the causal subgraph rather than to the original input.

This specialization is particularly useful for our analysis. For a cue discordance sample $(C, S) \in \mathcal{H}$, the class preferred by the non-causal mechanism satisfies $b(S) \neq a(C)$. Hence,

$$\Delta(C, S) = \eta_{a(C)}(C) - \eta_{b(S)}(C) \ge m(C). \tag{27}$$

Consequently, the Tsybakov condition controls the probability of cue discordance samples for which the causal advantage $\Delta(C, S)$ is arbitrarily small. Importantly, it does not impose a deterministic margin $\Delta(C, S) \ge \gamma$ uniformly over all samples. Instead, it provides a probabilistic low-noise characterization: samples with an arbitrarily small causal posterior advantage occupy increasingly small probability mass.

B.3 Proof of Lemma 1

Here is the proof of the lemma:

Proof. By Assumption 1, we have $P(Y \mid C, S) = P(Y \mid C) = \eta(C)$.

From Assumption 2, $a(C)$ is the unique Bayes-optimal class under $\eta(C)$. For a discordant sample, $b(S) \neq a(C)$, and therefore

$$\eta_{a(C)}(C) > \eta_{b(S)}(C), \tag{28}$$

which gives $\Delta(C, S) > 0$, completing the proof.

□

B.4 Proof of Lemma 2

Here is the proof of the lemma:

Proof. Using Assumptions 1 and 2, we have

$$P(Y \mid C, S) = P(Y \mid C) = \eta(C).$$

Therefore, the difference between the conditional cross-entropy risks is

$$\begin{aligned} \mathcal{R}^S_{\mathrm{CE}} - \mathcal{R}^C_{\mathrm{CE}} &= D_{\mathrm{KL}}\left(\eta(C) \| q_S(\cdot \mid S)\right) - D_{\mathrm{KL}}\left(\eta(C) \| q_C(\cdot \mid C)\right) \\ &\geq D_{\mathrm{KL}}\left(\eta(C) \| q_S(\cdot \mid S)\right) - \varepsilon_C. \end{aligned} \tag{29}$$

For notational simplicity, let $a = a(C)$ and $b = b(S)$. Since b is the class preferred by the spurious predictor,

$$q_S(b \mid S) \geq q_S(a \mid S). \tag{30}$$

Moreover, for a discordant sample, $a \neq b$, and by Lemma 1,

$$\eta_a(C) - \eta_b(C) = \Delta(C, S) > 0. \tag{31}$$

We now lower-bound the ℓ_1 distance between $\eta(C)$ and $q_S(\cdot \mid S)$. Restricting the sum to the coordinates corresponding to a and b, we obtain

$$\begin{aligned} \left\|\eta(C) - q_S(\cdot \mid S)\right\|_1 &= \sum_{k=1}^{K} \left|\eta_k(C) - q_S(k \mid S)\right| && (32) \\ &\geq \left|\eta_a(C) - q_S(a \mid S)\right| + \left|\eta_b(C) - q_S(b \mid S)\right| && (33) \\ &\geq \left|\left(\eta_a(C) - q_S(a \mid S)\right) - \left(\eta_b(C) - q_S(b \mid S)\right)\right| && (34) \\ &= \left|\eta_a(C) - \eta_b(C) + q_S(b \mid S) - q_S(a \mid S)\right|. && (35) \end{aligned}$$

Since $\eta_a(C) - \eta_b(C) = \Delta(C, S) > 0$ and $q_S(b \mid S) - q_S(a \mid S) \geq 0$, the quantity inside the absolute value is nonnegative. Hence,

$$\left\|\eta(C) - q_S(\cdot \mid S)\right\|_1 \geq \eta_a(C) - \eta_b(C) = \Delta(C, S). \tag{36}$$

Applying Pinsker's inequality then gives

$$\begin{aligned} D_{\mathrm{KL}}\left(\eta(C) \| q_S(\cdot \mid S)\right) &\geq \frac{1}{2}\left\|\eta(C) - q_S(\cdot \mid S)\right\|_1^2 \\ &\geq \frac{1}{2}\Delta(C, S)^2. \end{aligned} \tag{37}$$

Combining the above results, we have

$$\mathcal{R}^S_{\mathrm{CE}} - \mathcal{R}^C_{\mathrm{CE}} \geq \frac{1}{2}\Delta(C, S)^2 - \varepsilon_C, \tag{38}$$

which completes the proof. □

B.5 Proof of Theorem 1

Here is the proof of the theorem:

Proof. Define the Bayes margin induced by the causal subgraph as

$$m(C) := \min_{k \neq a(C)} \left[\eta_{a(C)}(C) - \eta_k(C) \right]. \tag{39}$$

By Assumption 3,

$$P\left(0 < m(C) \leq t\right) \leq At^{\alpha}. \tag{40}$$

For every discordant sample $(C, S) \in \mathcal{H}$, we have $b(S) \neq a(C)$. Therefore, since $m(C)$ is the minimum posterior gap between the Bayes-optimal class and any competing class, we have $m(C) > 0$ and:

$$\Delta(C, S) = \eta_{a(C)}(C) - \eta_{b(S)}(C) \geq m(C). \tag{41}$$

Hence,

$$\mathcal{H} \cap \{0 < \Delta(C, S) \leq t\} \subseteq \{0 < m(C) \leq t\}. \tag{42}$$

It follows that

$$P\left(0 < \Delta(C, S) \leq t \mid \mathcal{H}\right) = \frac{P\left(\mathcal{H} \cap \{0 < \Delta(C, S) \leq t\}\right)}{P(\mathcal{H})} \tag{43}$$

$$\leq \frac{P\left(0 < m(C) \leq t\right)}{p_{\mathcal{H}}} \tag{44}$$

$$\leq \frac{A}{p_{\mathcal{H}}} t^{\alpha}. \tag{45}$$

Setting

$$t = \gamma_{\varepsilon} = \left(\frac{\varepsilon p_{\mathcal{H}}}{A}\right)^{1/\alpha} \tag{46}$$

gives

$$P\left(0 < \Delta(C, S) \leq \gamma_{\varepsilon} \mid \mathcal{H}\right) \leq \varepsilon. \tag{47}$$

By Lemma 1, $\Delta(C, S) > 0$ for every discordant sample. Therefore,

$$P\left(\Delta(C, S) > \gamma_{\varepsilon} \mid \mathcal{H}\right) \geq 1 - \varepsilon. \tag{48}$$

Furthermore, by Lemma 2, for every $(C, S) \in \mathcal{H}$,

$$\mathcal{R}^{S}_{\mathrm{CE}} - \mathcal{R}^{C}_{\mathrm{CE}} \geq \frac{1}{2}\Delta(C, S)^2 - \varepsilon_C. \tag{49}$$

Thus, for any $(C, S) \in \mathcal{H}$ satisfying $\Delta(C, S) > \gamma_{\varepsilon}$,

$$\mathcal{R}^{S}_{\mathrm{CE}} - \mathcal{R}^{C}_{\mathrm{CE}} > \frac{1}{2}\gamma_{\varepsilon}^2 - \varepsilon_C = \delta_{\varepsilon}. \tag{50}$$

Equivalently,

$$\mathcal{R}^{C}_{\mathrm{CE}} + \delta_{\varepsilon} < \mathcal{R}^{S}_{\mathrm{CE}}. \tag{51}$$

Since this event holds whenever $\Delta(C, S) > \gamma_{\varepsilon}$, we finally obtain

$$P\left(\mathcal{R}^{C}_{\mathrm{CE}} + \delta_{\varepsilon} < \mathcal{R}^{S}_{\mathrm{CE}} \mid \mathcal{H}\right) \geq 1 - \varepsilon. \tag{52}$$

□

B.6 Proof of Proposition 1

Here is the proof of the proposition:

Proof. Let

$$\phi(x) = [x]_+. \tag{53}$$

Since ϕ is convex, Jensen's inequality gives

$$\begin{aligned} \mathbb{E}\left[[\delta + \ell_C - \ell_S]_+ \middle| C, S\right] &= \mathbb{E}\left[\phi(\delta + \ell_C - \ell_S) | C, S\right] \\ &\geq \phi\left(\mathbb{E}[\delta + \ell_C - \ell_S \mid C, S]\right). \end{aligned} \tag{54}$$

By the definitions of the conditional cross-entropy risks,

$$\mathbb{E}[\ell_C \mid C, S] = \mathcal{R}^C_{\mathrm{CE}}, \qquad \mathbb{E}[\ell_S \mid C, S] = \mathcal{R}^S_{\mathrm{CE}}. \tag{55}$$

Therefore,

$$\mathbb{E}\left[[\delta + \ell_C - \ell_S]_+ \middle| C, S\right] \geq \left[\delta + \mathcal{R}^C_{\mathrm{CE}} - \mathcal{R}^S_{\mathrm{CE}}\right]_+, \tag{56}$$

which finishs the proof. □

C Related Work

To address distribution shifts in graph, many methods adopt the framework of Invariant Risk Minimization Arjovsky et al. (2020). Invariant Risk Minimization (IRM) is designed to enhance generalization across diverse environments by identifying and leveraging invariant features that remain stable and consistent across various data distributions. Intuitively, data within the same environment group share similar non-critical information, such as image backgrounds and basis graph. It is assumed that the shift between the training and test distributions should be reflected among these environments. Consequently, to tackle generalization challenges in test environments, these methods integrate environment shift information from training environments into neural networks, driving the networks to remain invariant to the shifts between training and test distributions. However, in real-world datasets, environment labels are often unavailable, and manual annotation is prohibitively expensive. Consequently, existing methods typically attempt to infer environment labels based on differences among samples. DIR Wu et al. (2021) generates distinct interventional distributions by dividing and recombining samples to discover invariant subgraphs. GIL Li et al. (2022) and OOD-GCL Li et al. (2024a) employs unsupervised clustering to infer environmental labels and leverages the invariant principle to identify causal subgraphs. MoleOOD Yang et al. (2022) employs environment inference techniques to facilitate the learning of invariant graph-level representations. GOODHSE Piao et al. (2024) try to generate hierarchical semantic environments for each graph and guide the model to learn the diversity of environments within the same hierarchy while maintaining consistency across different hierarchies.

Due to reliance on environment labels of IRM, the predicted environment labels may not necessarily align with the real environmental distribution of the dataset. Moreover, recent studies have demonstrated that, in the absence of auxiliary information, it is theoretically impossible to accurately infer the environment from which a sample originates Lin et al. (2022); Chen et al. (2023). Consequently, many graph out-of-distribution generalization methods try to explore IRM-free approaches to address this limitation. GREA Liu et al. (2022) identifies subgraph structures called rationales by environment replacement to create virtual data points to improve generalizability and interpretability. LECI Gui et al. (2023) identifies causal subgraphs by introducing the Label and Environment Causal Independence assumption, which states that causal subgraphs should be independent of environment labels, while non-causal subgraphs should be independent of prediction labels. iMoLD Zhuang et al. (2023) identifies invariant representations by quantifying residual vectors and applying feature perturbation and alignment. EQuAD Yao et al. (2024a) and LIRS Yao et al. (2024b) adopt the Infomax principle to extract causal features. EQuAD adopts self-supervised learning to learn spurious features first, followed by learning invariant features by unlearning spurious features. LIRS takes an indirect approach by first learning the spurious features and then removing them from the ERM-learned features. GSAT Miao et al. (2022) and GOODAT Wang et al. (2024) incorporates graph information bottleneck theory to identify the rationale. CIGA Chen et al. (2022) derives an information-theoretic objective for the featurizer to identify subgraphs that

maximally preserve invariant intra-class information across a set of different (unknown) environments. AIA Sui et al. (2023) employs an adversarial augmenter to explore OOD data by generating new environments while maintaining stable feature consistency. GALA Chen et al. (2023) learns invariant graph representations guided by an ERM-based environment assistant model. In addition, several approaches attempt to tackle out-of-distribution generalization from the perspective of data augmentation. DPS Yu et al. (2022) constructs multiple label-invariant subgraphs using dedicated generators to train an invariant GNN predictor. LiSA Yu et al. (2023) generates diverse augmented environments with a consistent predictive relationship and facilitates learning an invariant GNN. G-Splice Li et al. (2024b) broadens the sample distribution through linear extrapolation of graph structure and features. UIL Sui et al. (2025) jointly exploits structural and semantic invariance by first inferring latent environments with an unsupervised clustering algorithm and then enforcing consistency of stable graph patterns across environments while learning environment-invariant representations. SuGAr Liu et al. (2025) learns multiple diverse invariant subgraphs via subgraph sampling and diversity regularization, and aggregates them to capture complementary causal patterns for improved graph OOD generalization. DiSCO Sun et al. (2026) uses a diversity- and sparsity-driven mixture-of-experts framework to capture instance-level heterogeneous causal subgraphs, where different experts identify distinct causal structures and a sparse gating mechanism adaptively selects the most relevant ones for each graph. PrunE Yao et al. (2025) proposes a pruning-based paradigm that eliminates spurious subgraphs rather than explicitly selecting invariant ones, encouraging the preservation of invariant structures through graph size regularization and probability alignment. Although these methods reduce the reliance on manually annotated environments, many of them still depend on strong causal assumptions, auxiliary objectives, or specific information-theoretic criteria.

Some of these methods were not included as baselines due to either limited empirical competitiveness or insufficient reproducibility, such as incomplete implementation details or the absence of publicly available code.