MODE: Subtractive Causal Representation via Manifold Orthogonality on Graphs

Xiao Han, Mengyao Zhou, Wei Wei*

***Abstract*—Graph out-of-distribution (OOD) generalization remains a central challenge in graph learning. Existing invariant learning methods often fail under strong feature interdependencies, where causal and trivial features become topologically entangled on collapsed latent manifolds, hindering effective disentanglement. To address this, we present MODE (Manifold Orthogonal Decoupling for Environments), which utilizes a subtractive representation learning approach. Rather than directly extracting causal features, the framework identifies and subsequently removes trivial features. Theoretically, the method employs Radon-Nikodym reweighting to achieve Conditional Barycenter Alignment of trivial features. This geometric alignment facilitates orthogonality in the tangent space between causal and trivial features, allowing for their decoupling. The causal subspace is then recovered by subtracting the orthogonal trivial subspace. Experimental results indicate that MODE performs competitively compared to established baselines across several datasets.**

***Index Terms*—out-of-distribution, graph neural network, manifold theory, barycenter alignment**

I. Introduction

Graph Neural Networks (GNNs) [1], [2] have achieved remarkable success in graph data mining tasks, ranging from node classification [3] to graph classification [4]. These models typically rely on the assumption that training and testing sets follow Independent and Identically Distributed (i.i.d) hypothesis [5]. However, in real-world scenarios, uncontrollable data collection processes often violate this assumption, leading to distribution shifts, known as Out-Of-Distribution (OOD) shifts [6]. Under such shifts, the statistical correlations often fail to generalize because models rely on spurious correlations [7] during training.

To formulate the problem, we distinguish two feature types in graph data. Causal Features C are structural patterns intrinsically determining the label Y, exemplified by molecular functional groups. In contrast, Trivial Features T represent environmental context like base structures or graph density.

* Corresponding author.

Xiao Han and Wei Wei are with the School of Artificial Intelligence, Beihang University, and the Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, Beijing 100191, China. (e-mail: hx2210@buaa.edu.cn; weiw@buaa.edu.cn).

Mengyao Zhou is with the Academy of Mathematics and Systems Science, Chinese Academy of Sciences and also with the University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhoumengyao@amss.ac.cn).

This work was supported by the National Natural Science Foundation of China (Grant No. 62276013, 62141605, 62050132), the Fundamental Research Funds for the Central Universities, and the Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing.

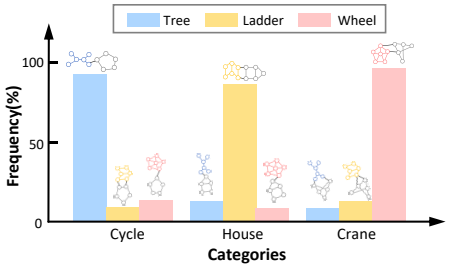


Fig. 1: The distribution of Trivial Features across different graph categories. The distinct dominance of specific background structures (e.g., Tree) within specific classes (e.g., Cycle) highlights the spurious correlation introduced by selection bias.

Although theoretically label-independent, T may induce statistical shortcuts under selection bias. Consider the motif identification task in Figure 1, which classifies graphs by motifs such as *Cycle* or *House*. Here, the motif acts as C while the base structure, e.g., *Tree* or *Ladder*, acts as T. While *Tree* bases should be uniformly distributed, a biased training set where *Tree* predominantly accompanies *Cycle* leads the model to learn a spurious correlation. Consequently, the model predicts labels based on background structures rather than semantic content. This shortcut generalizes poorly to OOD data, such as *Cycle* on *Ladder* bases. This failure indicates a topological entanglement on the latent manifold, where causal semantics are geometrically coupled with environmental noise.

The prevailing strategy for mitigating spurious correlations is Invariant Learning [8]–[10], a paradigm focused on extracting causal features stable across environments. Although successful in standard settings, these extraction methods face severe limitations under strong feature interdependency. Geometrically, this phenomenon indicates that the latent distribution of graph data does not occupy the full product space but collapses onto a low-dimensional diagonal submanifold. Within this topologically entangled structure, causal features like specific motifs are geometrically coupled with trivial features like background graphs. The motif identification task in Figure 1 exemplifies this issue: when the *Cycle* motif consistently appears with the *Tree* base, the trivial feature exhibits pseudo-invariance, rendering the true cause indistinguishable via standard invariance principles. As evidenced by Figure 2, the accuracy of current methods drops to near-random levels of 20% to 25% as the interdependency approaches 1.0. Consequently, isolating causal features within a collapsed manifold

becomes an ill-posed problem, as any geodesic traversal along the causal axis necessitates a simultaneous displacement along the environmental axis.

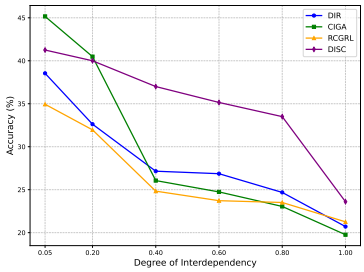


Fig. 2: Impact of feature interdependency on model accuracy. The performance collapse under strong interdependency demonstrates the failure of existing methods to disentangle causal features from collapsed latent manifolds.

To address these limitations, we propose subtractive causal representation. Instead of direct extraction, we isolate the causal features by geometrically freezing identifiable trivial features to peel away environmental noise. Specifically, we introduce **M**anifold **O**rthogonal **D**ecoupling for **E**nvironments (**MODE**), which employs a reweighting mechanism based on the Radon-Nikodym derivative [11] to enforce conditional barycenter alignment on the latent Riemannian manifold.

As illustrated in Fig. 3, latent distributions under strong interdependency collapse onto low-dimensional submanifolds where class-specific Riemannian barycenters ($\mu_{\text{blue}}, \mu_{\text{red}}$) misalign along environmental axes. This entanglement induces a coupled variation $\langle \nabla \mathbf{z}_C, \nabla \mathbf{z}_T \rangle$, where causal semantic shifts necessitate simultaneous trivial displacements, hindering generalization. MODE resolves this via a Radon-Nikodym measure change (Fig. 3b), aligning weighted centers to a common centroid μ^*. This alignment creates a differential asymmetry in the tangent bundle: trivial tangent vectors $\mathbf{v}_T$ approach zero, while causal vectors $\mathbf{v}_C$ remain dynamic. This enforces statistical orthogonality, allowing MODE to recover the robust causal subspace by subtracting the stationary trivial subspace.

The contributions of this work are summarized as follows:

- We propose a novel Subtractive Causal Representation paradigm, shifting the focus from extracting causal features to identifying and eliminating trivial ones via geometric constraints.
- We establish a theoretical framework based on Manifold Orthogonality, proving that enforcing Conditional Barycenter Alignment via Radon-Nikodym reweighting leads to tangent space disentanglement between causal and trivial features.
- We instantiate this theory into the MODE framework. Extensive experiments demonstrate its superiority, particularly in challenging scenarios with strong interdependency where traditional invariant learning fails.

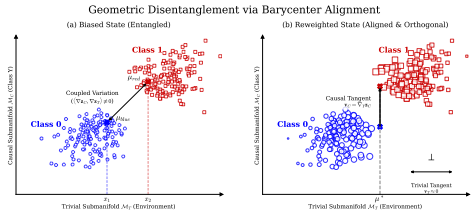


Fig. 3: Geometric intuition of MODE. (a) Entangled State: The data distribution collapses onto a diagonal submanifold due to spurious correlations. The misalignment between class-specific barycenters ($\mu_{\text{blue}} \neq \mu_{\text{red}}$) along the trivial axis results in inseparable feature coupling. (b) Aligned State: Radon-Nikodym reweighting aligns the weighted barycenters to μ^*. This constraint forces the trivial tangent vectors $\mathbf{v}_T$ to vanish, establishing geometric orthogonality with the causal vectors $\mathbf{v}_C$ and enabling effective disentanglement.

II. Related Work

Invariant Learning on Graphs. Invariant learning has established itself as a leading paradigm for graph OOD generalization. Within this domain, statistical penalty methods such as DIR [9] and CIGA [12] focus on minimizing prediction variance, whereas intervention-based approaches like RCGRL [13] and DISC [14] utilize instrumental variables or counterfactuals to mitigate confounders.

Despite their effectiveness in standard settings, these methods encounter severe limitations under strong interdependency where the data manifold collapses onto a low-dimensional trajectory. In this regime, statistical methods falter because causal and trivial features become geometrically fused, which precludes effective separation. Similarly, intervention strategies become intractable as the collapsed structure hinders the independent manipulation of variables. Specifically, deriving valid instruments or realistic counterfactuals proves impossible when the requisite feature combinations are absent from the latent distribution. Distinct from these paradigms, MODE prioritizes structural reconstruction. By enforcing conditional barycenter alignment, we repair the collapsed geometry to establish the necessary orthogonality for disentanglement.

III. Theoretical Analysis

To deeply understand the failure of invariant learning under strong interdependency, we conducted a theoretical analysis from the perspective of geometric disentanglement on the latent manifold. We reformulate the mechanism of our proposed MODE as utilizing the Radon-Nikodym derivative to induce a measure change, thereby achieving orthogonal decoupling in the tangent bundle.

A. Topological Entanglement

Let $\mathcal{G}$ be the input graph space. We assume the encoder Φ maps graphs into a high-dimensional latent Riemannian manifold $\mathcal{M} \subseteq \mathbb{R}^d$. Ideally, this manifold should be locally

homeomorphic to the Cartesian product of a causal submanifold and a trivial submanifold:

$$\mathcal{M} \cong \mathcal{M}_C \times \mathcal{M}_T. \tag{1}$$

Here, $\mathcal{M}_C$ stands for semantic content invariant to environmental shifts, where the label Y is a function value defined primarily on coordinates of $\mathcal{M}_C$, while $\mathcal{M}_T$ is the environmental noise.

The Geometric Essence of Interdependency: In a biased training set with strong spurious correlations, the observed joint distribution $P(z_C, z_T)$ does not span the full product space. Instead, due to selection bias, the distribution collapses onto a low-dimensional Diagonal Submanifold $\mathcal{S}_{train} \subset \mathcal{M}_C \times \mathcal{M}_T$. This topological entanglement implies that geodesic movement along $\mathcal{M}_C$ (changing the semantic class Y) inevitably forces a displacement along $\mathcal{M}_T$ (changing the environment). Consequently, the tangent vectors of causal and trivial features become coupled, i.e., $\langle \nabla z_C, \nabla z_T \rangle \neq 0$, causing the ordinary machine learning models often to capture spurious correlations.

B. Theoretical Justification of MODE

MODE addresses this collapse by introducing a reweighting mechanism to recover the independent product structure. We analyze this process through measure theory and differential geometry.

1) Measure Change via Radon-Nikodym Derivative: We view the reweighting process as a measure transformation. Let P be the probability measure on the original training manifold. The learned sample weights $W(z)$ define a new induced measure P_W. From a measure-theoretic perspective, W acts as the Radon-Nikodym derivative of P_W with respect to P:

$$\frac{dP_W}{dP}(z) = W(z). \tag{2}$$

The optimization is thus transferred to the new measure space $(\mathcal{M}, P_W)$. Since $W(z)$, implemented via bounded networks is strictly positive and finite, P_W and P are mutually absolutely continuous. This leads to our first theoretical guarantee:

Theorem 1 (Topological Invariance)**.** *The reweighting operation $P \to P_W$ induced by $W(z)$ preserves the underlying topological structure of the manifold. Specifically, the support and connectivity of the distribution remain invariant, ensuring that the causal mechanism inherent in the graph topology is not destroyed but merely density-shifted.*

Proof: Let $(\mathcal{M}, \Sigma)$ be the measurable space of the latent manifold. The reweighted measure P_W is defined via the Radon-Nikodym derivative $W(z)$. In our implementation, $W(z)$ is parameterized by a neural network, satisfying the boundedness condition: There exist constants

$$0 < \epsilon < M < \infty,$$

such that

$$\epsilon \leq W(z) \leq M,$$

for all $z \in \mathcal{M}$. To prove topological invariance, we establish that P and P_W are *mutually absolutely continuous* that is denoted as $P \ll P_W$ and $P_W \ll P$. Consider any measurable set $A \in \Sigma$. First, since $W(z) \leq M$, we have

$$P_W(A) = \int_A W(z) dP(z) \leq M \cdot P(A).$$

Thus,

$$P(A) = 0 \implies P_W(A) = 0, \text{ that is } P_W \ll P.$$

Conversely, since $W(z) \geq \epsilon > 0$, we have

$$P_W(A) \geq \epsilon \cdot P(A).$$

Thus,

$$P_W(A) = 0 \implies P(A) = 0, \text{ that is } P \ll P_W.$$

Since P and P_W share the same set of null measure sets, their topological supports coincide:

$$\text{supp}(P_W) = \text{supp}(P).$$

■

The support defines the valid geometric region of the data manifold and preserves its topological properties. P and P_W are mutually absolutely continuous and share the same support. Therefore, the reweighting does not change the topological support of the distribution.

2) Conditional Barycenter Alignment and Tangent Bundle Orthogonal Decoupling: In the original measure P, due to the spurious correlation of C and Y, the distributions of trivial features for different classes are misaligned on $\mathcal{M}_T$, that is,

$$\mathbb{E}_P[z_T | Y = i] \neq \mathbb{E}_P[z_T | Y = j].$$

MODE enforces *Riemannian Barycenter Alignment* under the induced measure P_W by balancing the cumulative distributions of trivial features:

$$\mathbb{E}_{P_W}[z_T | Y = i] \approx \mathbb{E}_{P_W}[z_T | Y = j] \approx \mu^*, \tag{3}$$

where μ^* is the global geometric barycenter on $\mathcal{M}_T$.

The core innovation of MODE is that this barycenter alignment creates a *Differential Asymmetry* in the tangent bundle $T\mathcal{M}$ under the measure P_W:

- **Trivial Invariance:** Due to alignment, the trivial representation z_T becomes stationary with respect to label Y in expectation. Its tangent vector vanishes: $\nabla_Y^{P_W} z_T \approx \mathbf{0}$.
- **Causal Variance:** To predict accurately, z_C must vary with Y. Its tangent vector remains significant: $\nabla_Y^{P_W} z_C \neq \mathbf{0}$.

Theorem 2 (Geometric Orthogonality)**.** *If the induced measure P_W satisfies the Conditional Barycenter Alignment condition for trivial features, then in the tangent bundle of $(\mathcal{M}, P_W)$, the trivial features z_T and causal features z_C become statistically orthogonal:*

$$\langle \nabla_Y^{P_W} z_T, \nabla_Y^{P_W} z_C \rangle \approx \langle \mathbf{0}, \nabla_Y^{P_W} z_C \rangle = 0. \tag{4}$$

By exploiting this geometric orthogonality, MODE identifies the trivial subspace as the component that can be frozen via reweighting, and subtracts it to recover the robust causal representation.

IV. METHOD

Guided by our theoretical analysis of manifold orthogonality, we propose **MODE** (Manifold Orthogonal Decoupling for Environments). As illustrated in Figure 4, MODE functions as a geometric filter that extracts robust causal representations by identifying and subtracting the trivial submanifolds. The framework is composed of two primary components: the **Topological Decomposition Module**, which projects the input graph onto latent causal and trivial submanifolds, and the **Geometric Disentanglement Module**, which enforces barycenter alignment and tangent orthogonality via Radon-Nikodym reweighting.

A. Topological Decomposition Module

The objective of this module is to perform the decomposition of the input graph $\mathbf{G} = \{A, X\}$ on the latent manifold $\mathcal{M} \cong \mathcal{M}_C \times \mathcal{M}_T$. We aim to learn a projection function that separates the graph structure into a causal subgraph $\mathcal{G}_C$ and a trivial subgraph $\mathcal{G}_T$.

We employ a learnable soft masking mechanism [15] as the coordinate projection operator. First, a GNN-based encoder Ψ maps the graph to a high-dimensional latent space to capture global topology:

$$H = \Psi(A, X), \quad H = \{h_1^\top, \ldots, h_N^\top\} \in \mathbb{R}^{N \times d}, \tag{5}$$

$$e_{ij} = (h_i || h_j), \tag{6}$$

where H and e_{ij} represent node and edge embeddings, respectively, with $||$ denoting concatenation. To define the submanifolds, we utilize two separate Multilayer Perceptrons (MLPs) [16] to predict the probability of each element belonging to $\mathcal{G}_C$ or $\mathcal{G}_T$:

$$\begin{aligned} \alpha_i^c, \alpha_i^t &= \mathrm{SoftMax}\left(\mathrm{MLP}_{\text{node}}(h_i)\right), \\ M_{\text{node}}^c &= \{\alpha_i^c\}_{i=1}^N, \quad M_{\text{node}}^t = \{\alpha_i^t\}_{i=1}^N, \end{aligned} \tag{7}$$

$$\begin{aligned} \beta_{ij}^c, \beta_{ij}^t &= \mathrm{SoftMax}\left(\mathrm{MLP}_{\text{edge}}(e_{ij})\right), \\ M_{\text{edge}}^c &= \{\beta_{ij}^c\}, \quad M_{\text{edge}}^t = \{\beta_{ij}^t\}. \end{aligned} \tag{8}$$

Here, the masks M^c and M^t act as geometric projectors. We reconstruct the subgraphs via Hadamard product $\odot$:

$$\begin{aligned} \mathcal{G}_C &= \{M_{\text{edge}}^c \odot A, M_{\text{node}}^c \odot X\}, \\ \mathcal{G}_T &= \{M_{\text{edge}}^t \odot A, M_{\text{node}}^t \odot X\}. \end{aligned} \tag{9}$$

Finally, separate GNN encoders Φ_C and Φ_T map these subgraphs to their respective representations $\mathbf{z}_C \in \mathcal{M}_C$ and $\mathbf{z}_T \in \mathcal{M}_T$.

B. Geometric Disentanglement Module

This module implements the core **Subtractive Causal Representation** strategy. It operates in a sequential logic: first identifying the spurious shortcuts, then freezing them via geometric alignment, and finally subtracting them via orthogonality.

*1) **Phase 1: Spurious Identification Constraint**:* Before removing trivial features, they must be correctly identified. Although trivial features $\mathbf{z}_T$ is non-causal, it can be strongly correlated with Y in biased training data, making it highly predictive.

To effectively capture these spurious shortcuts while preventing them from dominating the predictions, we impose a constraint that renders the trivial subgraph predictive yet suboptimal. Initially, we minimize the empirical risk of the trivial features to encapsulate the environmental bias:

$$\min_{\Psi, \Phi_T, g_T} \mathcal{R}_T = L\left(g_T(\mathbf{Z}_T), \mathbf{Y}^{N^{tr}}\right), \tag{10}$$

where $L(\cdot)$ denotes the cross-entropy loss, $g_T(\cdot)$ serves as the classifier for trivial features, and $\mathbf{Z}_T = \{\mathbf{z}_T^i\}_{i=1}^{N^{tr}}$ denotes the confounding representations. Crucially, to guarantee that the causal features possess superior discriminative power compared to the trivial ones (i.e., $\mathcal{R}_T > \mathcal{R}_C$), we introduce complementary risk terms based on the prediction inversion:

$$\mathcal{R}_{rC} = L(1 - g_C(\mathbf{Z}_C), \mathbf{Y}^{N^{tr}}), \quad \mathcal{R}_{rT} = L(1 - g_T(\mathbf{Z}_T), \mathbf{Y}^{N^{tr}}),$$

where $g_C(\cdot)$ represents the classifier for causal features, and $\mathbf{Z}_C = \{\mathbf{z}_C^i\}_{i=1}^{N^{tr}}$ denotes the causal representations. Subsequently, we formulate an inequality constraint loss $\mathcal{R}_{ine}$ as follows:

$$\mathcal{R}_{ine} = \Gamma\left(\hat{\mathbf{Y}}_C^{N^{tr}}, \hat{\mathbf{Y}}_T^{N^{tr}}; \hat{\mathbf{Y}}^{N^{tr}}\right) \tag{11}$$

$$= \mathbb{1}^T\left(\mathcal{R}_{rT} \odot \sigma\left(\mathcal{R}_{rC} > \mathcal{R}_{rT}\right)\right), \tag{12}$$

where σ is the indicator function. This term enforces feature separation through a competition-suppression mechanism. By monitoring these risks, the objective imposes a penalty on the trivial branch whenever it outperforms the causal branch. Consequently, the final objective for this phase integrates bias learning with the suppression constraint:

$$\min \ \mathcal{L}_{ident} = \mathcal{R}_T + \mathcal{R}_{ine}. \tag{13}$$

This formulation ensures that $\mathcal{G}_T$ captures the dominant environmental bias, thereby providing a robust target for the subsequent geometric disentanglement.

*2) **Phase 2: Measure Change via Barycenter Alignment**:* Following the identification of $\mathbf{z}_T$, we perform geometric alignment on the trivial submanifold $\mathcal{M}_T$. In biased training sets, class-conditioned trivial submanifolds are spatially misaligned ($\mu_i \neq \mu_j$), creating a separation that encodes spurious correlations. To resolve this, we leverage *Theorem 1* to learn sample weights $W = \{w_i\}_{i=1}^{N^{tr}}$ and induce a new measure P_W. This enforces conditional barycenter alignment, compelling the Riemannian barycenters of all class-specific submanifolds to coincide at a unified focal point.

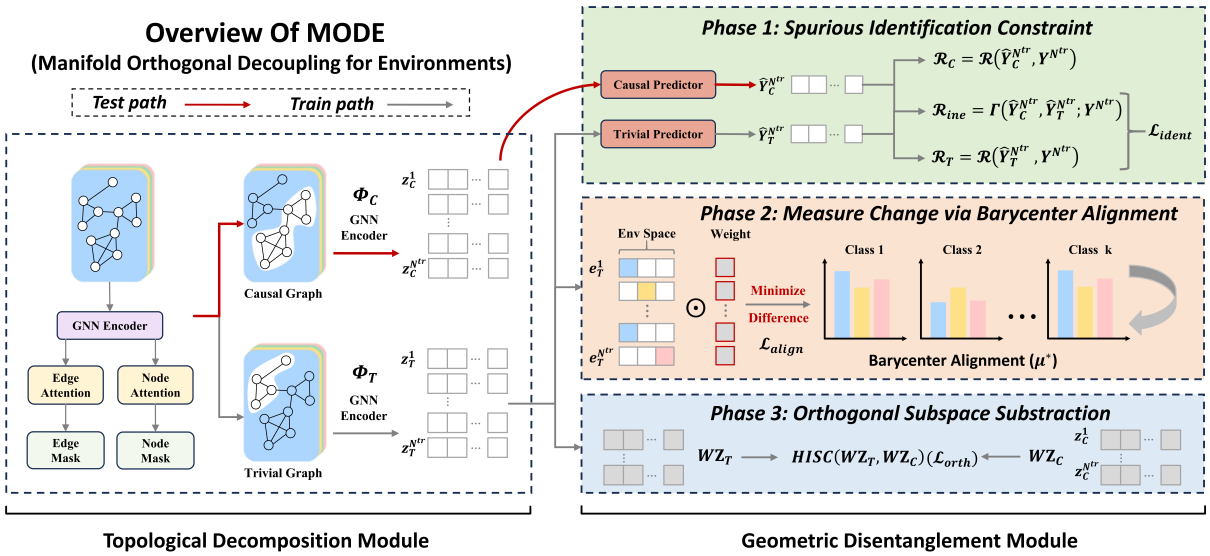


Fig. 4: The Framework of MODE.

Specifically, we compute the weighted barycenter for each class m:

$$\mu_m^{env} = \sum_{i=1}^{N^{tr}} w_i \cdot \chi\left(\mathbf{z}_T^i\right) \cdot \mathbb{1}\left(Y_i = y_m\right), \tag{14}$$

where χ projects $\mathbf{z}_T$ to a metric space for alignment. We then minimize the divergence between these class barycenters:

$$\min_{W,\chi} \mathcal{L}_{align} = \sum_{m=0}^{k} \sum_{n=m+1}^{k} \mathrm{MSE}\left(\mu_m^{env}, \mu_n^{env}\right). \tag{15}$$

Minimizing $\mathcal{L}_{align}$ forces the trivial features to collapse to a single geometric center μ^* regardless of the class label, effectively correcting the spatial misalignment.

*3) **Phase 3: Orthogonal Subspace Subtraction**:* With the trivial features barycenter aligned and the causal features remaining dynamic, a Differential Asymmetry is created. We exploit this by enforcing tangent space orthogonality to explicitly subtract the trivial subspace.

We employ the Hilbert-Schmidt Independence Criterion (HSIC) [17] to minimize the dependence between the reweighted trivial features and the causal features:

$$\min \mathcal{L}_{orth} = \mathrm{HSIC}\left(W \odot \mathbf{Z}_T, W \odot \mathbf{Z}_C\right). \tag{16}$$

Here, $W \odot \mathbf{Z}$ represents the distribution under the induced measure P_W. Since a static vector field is inherently orthogonal to a dynamic one, minimizing $\mathcal{L}_{orth}$ ensures that $\mathbf{z}_C$ contains no information from the environmental subspace, effectively achieving subtractive causal representation.

C. Learning Strategy

The optimization of MODE is divided into two alternating stages to ensure stable manifold reconstruction:

Stage 1 (Manifold Reconstruction): We fix the decomposition encoders and optimize the weights W to align the barycenters of the currently identified trivial features.

$$\min_{W,\chi} \mathcal{L}_{align}. \tag{17}$$

Stage 2 (Trivial Submanifold Subtraction): We fix the weights W and update the Topological Decomposition Module and classifiers. This stage simultaneously identifies the spurious shortcuts (via $\mathcal{L}_{ident}$) and subtracts them via orthogonality (via $\mathcal{L}_{orth}$) to extract robust causal features.

$$\min_{f,\Phi,g} \mathcal{L}_{total} = \mathcal{L}_{pred} + \lambda_1 \mathcal{L}_{ident} + \lambda_2 \mathcal{L}_{orth}. \tag{18}$$

This two-stage mechanism works in synergy to ensure that the model can accurately identify and subtract the trivial features.

V. EXPERIMENT

A. Datasets

To evaluate the effectiveness of **MODE** and its robustness against controllable biases, we conduct experiments on both real-world and synthetic benchmarks. For real-world evaluation, we utilize four representative datasets from **TU-Dataset** [18]: **MUTAG** and **PROTEINS** for bioinformatics, and **IMDB-BINARY** and **IMDB-MULTI** for social network analysis.

To investigate model performance under distribution shifts, we employ the **Spurious-Motif** dataset [9] with the bias degree fixed at $b = 0.9$ (denoted as **SPMotif-0.9**). This setup ensures that 90% of samples within each class exhibit specific trivial features, creating a strong interdependency between causal and non-causal components. Additionally, we incorporate the **CRCG** dataset [19], a synthetic benchmark with diverse subgraph categories, to assess stability under varying spurious correlation probabilities. We use b for SPmotif and P for CRCG to denote the bias strength.

B. Experimental Setup

All models use GeLU activation [20] and global add pooling, with hyper-parameters λ_1 and λ_2 selected via grid search in $[0.1, 1.0]$ (step size 0.1). Experiments are conducted on an NVIDIA GeForce RTX 3090 GPU (24GB). For TUDataset, we employ a 3-layer GCN encoder [1] with hidden dimensions of 128 or 256, trained using Adam (learning rate 0.001) for 200 epochs with batch sizes of 64 or 128. For synthetic benchmarks, we use a 3-layer LEGNN encoder [21] with a 64-dimensional hidden layer, batch size 32, and learning rate 0.001. Following [19], [22], we report mean accuracy (ACC) with standard deviation, using 10-fold cross-validation on TUDataset and five independent runs with fixed seeds on synthetic datasets.

C. Baselines

We compare **MODE** against the following baselines:

ERM [23]: Standard Empirical Risk Minimization, which trains the model without specific interventions for distribution shifts.

DIR [9]: A causal learning method that identifies invariant features by minimizing prediction variance across different interventional distributions.

CIGA [12]: An approach that maximizes mutual information to extract subgraphs preserving class-invariant information, effectively filtering out spurious noise.

RCGRL [13]: A method that constructs instrumental variables to remove confounding factors, extracting features causally related to the labels.

DISC [14]: A framework combining counterfactual sample generation with Generalized Cross Entropy (GCE) loss [24] to mitigate the effects of severe coupling between causal and trivial features.

D. Results

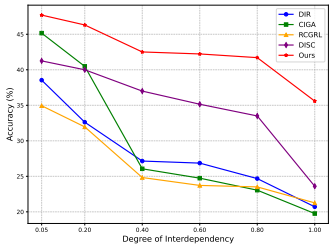


Fig. 5: The performances of different models as the degree of interdependency gradually increases.

1) Results on Synthetic Datasets.: To evaluate model robustness, we systematically regulate the degree of interdependency within the CRCG datasets. This dependency is parameterized by the bias ratio P, where a value approaching 1 indicates a progressively stronger interdependency. The quantitative results are summarized in Table I and Figure 5, from which we derive the following in-depth conclusions:

MODE demonstrates superior performance in high-interdependency scenarios. As presented in Table I, MODE consistently outperforms baseline methods across varying bias levels. This performance advantage becomes particularly evident in extreme settings. Specifically, on the SPMotif-0.9, MODE achieves an accuracy of 63.41%, exceeding the second-best method (DISC, 53.68%) by approximately 10%. Furthermore, on CRCG with $P = 100\%$, while traditional invariant learning methods such as CIGA and DIR experience severe performance degradation comparable to random guessing, MODE retains a relatively stable accuracy of 35.59%.

This empirically validates our theory regarding Topological Collapse. When $P \to 100\%$, the trivial and causal submanifolds perfectly overlap in the observed distribution, making statistical independence constraints (used by CIGA/DIR) ill-posed. MODE succeeds because it does not rely solely on statistical independence; instead, the Barycenter Alignment mechanism geometrically reconstructs the measure space, allowing for the effective subtraction of spurious features even when they are statistically inseparable in the original space.

MODE exhibits the strongest resistance to manifold collapse. Figure 5 illustrates the accuracy trajectory as the degree of interdependency increases from 0.05 to 1.0. Initially, MODE achieves the highest accuracy , outperforming CIGA. Furthermore, as interdependency intensifies, the performance degradation slope of MODE (Red line) is significantly flatter compared to that of CIGA (Green) and DIR (Blue). While DISC (Purple) demonstrates a degree of robustness, its performance declines more precipitously than MODE in extreme settings where $P > 0.8$.

The flat decay curve confirms the effectiveness of the **Geometric Disentanglement Module**. By enforcing tangent orthogonality, MODE ensures that the learned causal representation $\mathbf{z}_C$ remains invariant to the shifts in the trivial submanifold $\mathcal{M}_T$, providing a geometric shield against the intensifying bias.

2) Results on Real-World Datasets.: Real-world datasets present a more complex challenge: the spurious correlations are not injected via simple rules but are topologically entangled with the graph structure. The results on TUDataset benchmarks are reported in Table I.

MODE achieves superior OOD generalization on complex topologies. As reported in Table I, MODE consistently surpasses all baseline models across the four real-world datasets. Notably, on the MUTAG task, MODE achieves an accuracy of 90.96%, which exceeds the nearest competitor DISC by a margin of 1.57%. Furthermore, MODE establishes a substantial lead on PROTEINS and IMDB-BINARY, exemplified by an improvement of approximately 7% over DIR on the PROTEINS dataset.

This success highlights the capability of our **Topological Decomposition Module**. Unlike synthetic biases which are often node-attribute based, real-world biases (e.g., functional groups in molecules or community structures in social networks) require precise structural separation. The superior

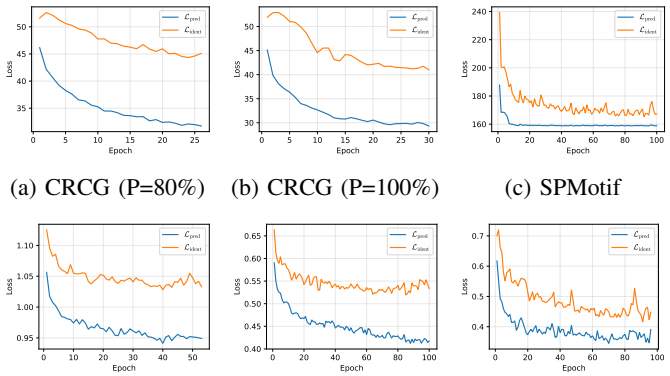
TABLE I: Performance comparison on synthetic and real-world datasets.

Dataset	CRCG				
	P=20%	P=40%	P=60%	P=80%	P=100%
ERM	28.86±1.17	25.94±1.63	24.43±1.40	23.15±1.10	22.62±1.79
DIR	32.62±1.29	27.15±1.38	26.86±0.87	24.68±0.94	20.71±1.17
CIGA	40.48±1.08	26.06±0.86	24.74±1.04	23.05±1.28	19.76±0.95
RCGRL	31.96±1.17	24.83±0.69	23.72±0.75	23.51±0.62	21.26±0.53
DISC	40.00±0.98	37.00±0.69	35.15±1.35	33.50±1.08	23.60±0.64
Ours	**46.31±1.91**	**42.51±2.91**	**42.23±2.89**	**41.71±2.61**	**35.59±1.92**
Dataset	MUTAG	PROTEINS	IMDB-BINARY	IMDB-MULTI	SPMotif-0.9
ERM	85.23±9.81	75.48±3.31	73.90±2.60	50.47±3.40	33.61±1.02
DIR	79.30±11.97	68.48±8.52	70.10±5.15	47.67±6.53	39.20±1.94
CIGA	76.05±7.67	70.25±4.60	71.00±3.83	50.40±1.76	36.01±0.80
RCGRL	78.30±9.32	70.00±3.16	72.50±4.20	49.40±2.82	41.58±1.12
DISC	89.39±8.21	68.73±7.91	72.90±4.31	50.13±3.48	53.68±8.50
Ours	**90.96±3.55**	**77.01±3.82**	**74.70±5.01**	**51.07±2.93**	**63.41±7.17**

performance indicates that MODE effectively projects the complex graph topology onto the correct latent submanifolds, isolating the invariant semantic structure from environmental noise.

E. Additional Analysis

1) Analysis of Suboptimal Priority Constraint: To validate the inequality constraint $\mathcal{R}_T > \mathcal{R}_C$ mandating higher predictive power for causal features, we monitor the training loss dynamics (Fig. 6). Across CRCG, IMDB-MULTI, PROTEINS, and MUTAG, $\mathcal{L}_{pred}$ descends faster and remains consistently lower than $\mathcal{L}_{ident}$. On SPMotif-0.9, the narrower margin reflects the extreme interdependency ($b = 0.9$) where trivial features are highly predictive, yet the inequality is strictly maintained. These trajectories confirm that the constraint successfully restricts trivial predictors to suboptimal paths, enabling the model to distinguish true causes from shortcuts and avoid overfitting to spurious correlations even in high-bias scenarios.

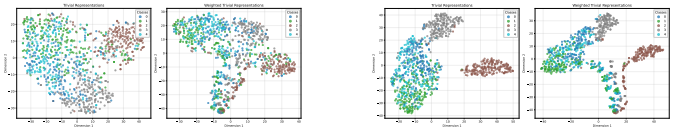


(a) CRCG (P=80%) (b) CRCG (P=100%) (c) SPMotif

(d) IMDB-MULTI (e) PROTEINS (f) MUTAG

Fig. 6: The value of $\mathcal{L}_{pred}$ and $\mathcal{L}_{ident}$ in training stage.

2) Visual Verification of Barycenter Alignment: To empirically verify our Barycenter Alignment theory, we visualize the distribution of trivial features $\mathbf{z}_T$ on the CRCG dataset under both strong ($P = 80\%$) and extreme ($P = 100\%$) interdependency settings using t-SNE [25], as illustrated in Figure 7. In the raw representation space (Left Column), the trivial features exhibit distinct clustering patterns corresponding to class labels, confirming that the environmental noise is spatially misaligned and inherently carries spurious predictive power. Notably, this spatial separation is intensified at the extreme setting of $P = 100\%$, where the class boundaries become significantly sharper and more distinct compared to the $P = 80\%$ setting. Conversely, after applying the learned weights W (Right Column), a dramatic geometric transformation is observed: the distributions of different classes significantly overlap and collapse towards a unified central region. This visual evidence confirms that MODE successfully enforces conditional barycenter alignment, effectively freezing the environmental variation to prepare for subtraction.



(a) CRCG (P=80%) (b) CRCG (P=100%)

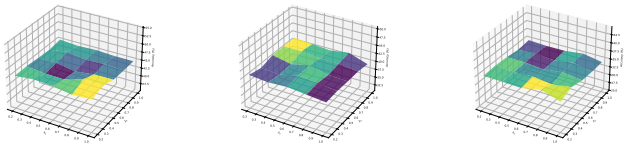
Fig. 7: Visualization of the barycenter alignment effect on trivial representations

3) Ablation Study: We validate the geometric components of MODE by comparing the full framework against variants without the Spurious Identification Constraint ($w/o\ \mathcal{L}_{ident}$) and Orthogonal Subspace Subtraction ($w/o\ \mathcal{L}_{orth}$). As shown in Table II, performance declines across PROTEINS, IMDB-M, and CRCG when either component is removed, underscoring their synergy. Specifically, the drop in $w/o\ \mathcal{L}_{ident}$ indicates that failing to capture shortcuts prevents accurate localization of the trivial submanifold $\mathcal{M}_T$. Conversely, $w/o\ \mathcal{L}_{orth}$ reveals that without tangent orthogonality, spurious information leaks into final representations even if bias is identified. By leveraging differential asymmetry, $\mathcal{L}_{orth}$ ensures statistical independence between subspaces. These results demonstrate that identification and subtraction constitute an essential closed-loop geometric filter for robust representation learning.

4) Hyper-parameter Sensitivity Analysis: We systematically evaluate the sensitivity of MODE regarding the loss

TABLE II: Ablation Study on real-world datasets and synthetic datasets

Dataset	PROTEINS	IMDB-M	CRCG
$w/o\ \mathcal{L}_{ident}$	74.80±3.99	50.87±3.85	33.58±1.10
$w/o\ \mathcal{L}_{orth}$	74.30±3.71	50.60±3.76	35.11±1.16
Full	**77.01±3.82**	**51.07±2.93**	**36.05±1.16**



(a) CRCG (40%) (b) CRCG (80%) (c) CRCG (100%)

Fig. 8: Hyper-parameter sensitivity of λ_1 and λ_2 on CRCG datasets.

coefficients λ_1 and λ_2. Figure 8 illustrates the 3D accuracy surfaces derived from the CRCG dataset across three distinct bias levels. A general observation across these settings is that the performance surfaces exhibit minimal fluctuation, thereby demonstrating the robustness of the framework to hyper-parameter variations within the interval $[0.1, 1.0]$. Notably, under the extreme interdependency condition of $P = 100\%$ shown in Figure 8c, the model sustains competitive accuracy, provided that λ_1 and λ_2 are not simultaneously minimized. This empirical evidence suggests that imposing sufficient weight on the geometric constraints effectively enables the model to mitigate manifold collapse.

VI. CONCLUSION

This work revisits graph out-of-distribution (OOD) generalization from a manifold perspective, showing that manifold collapse under strong interdependency makes traditional invariance constraints ill-posed. We propose MODE (Manifold Orthogonal Decoupling for Environments), a "reconstruct-then-subtract" approach that uses Radon-Nikodym derivatives to align submanifold barycenters and enforce geometric orthogonality between causal and environmental features. Experiments demonstrate improved generalization and effective disentanglement. However, several limitations remain. Under extreme environmental dominance, numerical instability may arise, motivating the exploration of stronger topological constraints.

REFERENCES

[1] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[2] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[3] L. Duan, X. Chen, W. Liu, D. Liu, K. Yue, and A. Li, "Structural entropy based graph structure learning for node classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 8372–8379.

[4] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *International Conference on Learning Representations*, 2019.

[5] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. USA: Cambridge University Press, 2014.

[6] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation (rex)," in *International conference on machine learning*. PMLR, 2021, pp. 5815–5826.

[7] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, "Deep stable learning for out-of-distribution generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 5372–5382.

[8] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," 2020.

[9] Y. Wu, X. Wang, A. Zhang, X. He, and T.-S. Chua, "Discovering invariant rationales for graph neural networks," in *International Conference on Learning Representations*, 2022.

[10] T. Jia, H. Li, C. Yang, T. Tao, and C. Shi, "Graph invariant learning with subgraph co-mixup for out-of-distribution generalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 8562–8570.

[11] H. L. Royden and P. Fitzpatrick, *Real analysis*. Macmillan New York, 1988, vol. 32.

[12] Y. Chen, Y. Zhang, Y. Bian, H. Yang, K. Ma, B. Xie, T. Liu, B. Han, and J. Cheng, "Learning causally invariant representations for out-of-distribution generalization on graphs," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[13] H. Gao, J. Li, W. Qiang, L. Si, B. Xu, C. Zheng, and F. Sun, "Robust causal graph representation learning against confounding effects," ser. AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

[14] S. Fan, X. Wang, Y. Mo, C. Shi, and J. Tang, "Debiasing graph neural networks via learning disentangled causal substructure," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 934–24 946, 2022.

[15] S. Zhang, H. Huang, J. Liu, and H. Li, "Spelling error correction with soft-masked bert," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 882–890.

[16] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65 6, pp. 386–408, 1958.

[17] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *International conference on algorithmic learning theory*. Springer, 2005, pp. 63–77.

[18] C. Morris, N. M. Kriege, F. Bause, K. Kersting, P. Mutzel, and M. Neumann, "Tudataset: A collection of benchmark datasets for learning with graphs," *arXiv preprint arXiv:2007.08663*, 2020.

[19] H. Gao, C. Yao, J. Li, L. Si, Y. Jin, F. Wu, C. Zheng, and H. Liu, "Rethinking causal relationships learning in graph neural networks," ser. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024.

[20] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2023. [Online]. Available: https://arxiv.org/abs/1606.08415

[21] E. Ranjan, S. Sanyal, and P. Talukdar, "Asap: Adaptive structure aware pooling for learning hierarchical graph representations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 5470–5477.

[22] F. Errica, M. Podda, D. Bacciu, A. Micheli *et al.*, "A fair comparison of graph neural networks for graph classification," in *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*, 2020.

[23] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. New York, NY, USA: Cambridge University Press, 2014.

[24] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 8792–8802.

[25] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.