# COUNTERFACTUAL INTERVENTION VALIDATION AND DECISION TRANSITION LEARNING FOR GRAPH AND HYPERGRAPH NEURAL NETWORKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The decision processes of Graph Neural Networks and Hypergraph Neural Networks often rely on implicit structural propagation and distributed representations. It makes them difficult to interpret. And in multi class tasks, class imbalance and inter-class similarity can lead to ambiguous local decision boundaries. Make minority and easily confused classes difficult to learn.So we propose a counterfactual intervention validation and decision transition learning framework for GNNs and HGNNs. First, we introduce a two-stage counterfactual discovery and intervention validation method. It searches the structural space for key supports and estimates their contributions. The key supports drive a source state toward a restoration target. Node feature interventions are then applied to validate these candidate supports and to generate counterfactual samples. Second, we propose a dual target counterfactual decision transition learning method. Each source sample is paired with counterfactuals corresponding to a restoration target and a class-transition target. The model is jointly optimized using endpoint classification, decision-margin, transition-magnitude, and local decision-gradient alignment constraints.We evaluate on two public datasets using graph and hypergraph neural networks.On the TEP dataset, it improves Macro-F1 to 91.89% and Balanced Accuracy to 90.74%. And increase the restoration endpoint margins to 16.850 and transition endpoint margins to 13.954.On the Paderborn bearing dataset, the model maintains a Macro-F1 of 96.43%. Across both datasets, the Stage-1 exhibits high cross-stage intervention consistency and sparsity. The resulting reasoning chain demonstrates feature changes causally induce changes in the model predictions. Provide an intervention-based explanation of the model decisions. And the proposed enhancement strategy encourages geometrically meaningful decision directions and improves the recognition of minority and closely related classes.

## 1 INTRODUCTION

GNNs and HGNNs have been widely applied to molecular modeling, recommender systems, knowledge graphs, and industrial intelligence (Kipf & Welling, 2017; Hamilton et al., 2017; Veličković et al., 2018; Feng et al., 2019; Yadati et al., 2019; Chien et al., 2022). However, the predictions of graph-based models usually depend on multi-layer structural propagation and nonlinear representation transformations. It may make it difficult to understand why a model makes a particular prediction and how its decision would change under potential interventions. Explainable graph learning has explored gradient-based, perturbation-based, and subgraph-optimization-based methods. It analyzes model predictions by identifying important nodes, edges, or subgraphs (Ying et al., 2019; Luo et al., 2020; Schlichtkrull et al., 2021; Yuan et al., 2021; Miao et al., 2022; Yuan et al., 2023; Amara et al., 2022; Kosan et al., 2024). For example, GNNExplainer explains model decisions by searching for compact substructures. The substructures are highly relevant to a given prediction (Ying et al., 2019). Counterfactual reasoning focuses on intervening on certain nodes and examining the changes of the target state or prediction probability. If the target state or prediction probability changes after the intervention, the intervened variables are regarded as causal factors associated with the corresponding prediction. Therefore, we can analyse the causal sensitivity of different factors



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.

that make the model's decisions chang by changing the input states and observing the prediction responses. Because counterfactuals naturally have a causal explanatory logic (Pearl, 2009; Wachter et al., 2018; Karimi et al., 2021). Graph counterfactual methods extend counterfactual explanation to graph neural networks. For example, CF-GNNExplainer searches for minimal structural perturbations and generates graph counterfactual samples (Lucic et al., 2022).Proceedings of Machine Learning Research havs further investigated the validity, stability, and distributional consistency of counterfactual samples (Bajaj et al., 2021; Tan et al., 2022; Ma et al., 2022; Lin et al., 2021; Huang et al., 2023; Verma et al., 2024; Prado-Romero et al., 2024). Hence, counterfactuals naturally contain both intervention information and prediction changes. Provide a more direct perspective for understanding the decision mechanisms of graph models than static correlation-based explanations.

Existing methods mainly treat counterfactuals as post-hoc explanation tools. It's goal focuses on validating the causality of variables.But it pays limited attention to the important information from counterfactual samples. In multi class graph learning tasks, different classes often exhibit complex feature overlap and structural similarity. Meanwhile, class imbalance further biases the model toward majority-class patterns. Make minority and closely related classes more likely to be confused. As a result, relying solely on supervision from real samples is often insufficient for learning clear class boundaries, especially for classes with limited training data. For example, in anomaly detection, medical diagnosis, and industrial defect recognition, a model must not only distinguish normal from abnormal states, but also discriminate among multiple abnormal classes with similar patterns. Conventional supervised learning can learn class distributions. But it is difficult to explicitly learn the decision boundaries between classes, the key factors that drive class transitions, and the directions of change in the local decision space (Zhang et al., 2018; Verma et al., 2019; Kong et al., 2020; Han et al., 2022; You et al., 2020; 2021; Khosla et al., 2020; Pawelczyk et al., 2022). So if simply add counterfactual samples to the training set and treat them as two independent samples,it will ignore the most important relationship between them.

Therefore, we propose a counterfactual intervention validation and decision transition learning framework for GNNs and HGNNs. The proposed framework can enhance model interpretability and decision-boundary learning by exploiting the intervention directions and class-transition relationships. The framework first performs **two-stage interventional counterfactual reasoning**. In the first stage, structural interventions are conducted to identify key structural regions and quantify their contributions. The key structral regions can drive a source state back to normal. In the second stage, keep the original graph structure unchanged and interven the node-feature of the key structural regions. So that they induce the desired target transition. In this way, the method validates that the set of structural interventions is made up of causal nodes for the original prediction. And, at the same time, we can obtain reliable counterfactual samples. Subsequently, **dual-target counterfactual transition learning** is performed by constructing **counterfactual triplets**. The restoration target is used to learn the boundary between abnormal and stable states. While the transition target is used to learn the competitive boundary between different classes. The model is jointly optimized using endpoint classification, decision-boundary margin, class-transition magnitude, and **local gradient-direction consistency**. So that it learns not only the class distribution but also the local decision geometry associated with class transitions. Furthermore, we constrats the model decision direction to be aligned with the direction of actual counterfactual changes. It make the learned decision process is endowed with a more explicit geometric interpretation.

Our main contributions are summarized as follows:

- **Two-stage counterfactual reasoning.** In the first stage, structural interventions are used to search for key structural factors and quantify their contributions. The key structural factors drive changes in the source state. In the second stage, the original graph structure is restored, and independent node-feature interventions are applied to verify whether these candidate factors can still induce changes in the source phenomenon, i.e., whether they are causal with respect to the source prediction. Then select counterfactual samples with cross-intervention consistency, improving the reliability of counterfactual explanations.
- **Dual-target counterfactual boundary-gradient decision transition learning.** For each source state, both a restoration target and a class-transition target are constructed. And the original sample is paired with the corresponding counterfactual samples. Then jointly optimize endpoint classification, decision margin, finite transition magnitude, and local gradient-direction consistency. The

optimized model strengthens local class boundaries and obtains a more explicit geometric interpretation of its decision process.

- **Comprehensive experimental validation.** We validate the proposed method on two public datasets. On the TEP dataset, we employ GNN-TAM with learnable topology. And evaluate the improvement of counterfactual transition learning on complex multi-class decision boundaries and difficult-class recognition. On the Paderborn Bearing Dataset, we use a classical HGNN with a fixed physical hypergraph. And evaluate two-stage intervention counterfactual generation, structural-feature support consistency, and local decision-boundary constraints.

# 2 PRELIMINARIES

## 2.1 GRAPH AND HYPERGRAPH PREDICTION

We consider **multi-class prediction problems on GNNs and HGNNs**. The $i$-th structured sample is uniformly represented as $G_i = (X_i, S_i)$. $X_i \in \mathbb{R}^{N_i \times d}$ denotes the node feature matrix, $N_i$ is the number of nodes, and $d$ is the node feature dimension. $S_i$ represents the structural information. For a graph, $S_i = A_i$, where $A_i$ is the adjacency matrix. For a hypergraph, $S_i = H_i$, where $H_i$ is the node–hyperedge incidence matrix.

GNNs aggregate neighborhood information through message passing $h_v^{(l+1)} = \phi^{(l)}\left(h_v^{(l)}, \mathrm{AGG}\left\{h_u^{(l)} : u \in N(v)\right\}\right)$, and the layer-wise update can be written as $H^{(l+1)} = \sigma\left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$, where $A = A + I$, $D$ is the corresponding degree matrix, and $W^{(l)}$ denotes the learnable parameters (Kipf & Welling, 2017; Hamilton et al., 2017; Veličković et al., 2018) (Kipf & Welling, 2017). A hypergraph is defined as $G = (V, E)$. Each hyperedge $e \in E$ can connect multiple nodes. The relationships between nodes and hyperedges are represented by an incidence matrix $H \in \{0,1\}^{N \times E}$. And $H_{ve} = 1$ indicates that node $v$ belongs to hyperedge $e$. A typical HGNN propagation process is formulated as $X^{(l+1)} = \sigma\left(D_v^{-\frac{1}{2}}HW_eD_e^{-1}H^\top D_v^{-\frac{1}{2}}X^{(l)}W^{(l)}\right)$, where $D_v$ and $D_e$ denote the node-degree and hyperedge-degree matrices, and $W_e$ represents the hyperedge weights (Feng et al., 2019; Yadati et al., 2019; Chien et al., 2022).

We represent both graph and hypergraph prediction models as $z_i = f_\theta(X_i, S_i) \in \mathbb{R}^C$, where $C$ is the number of classes and $z_i$ denotes the output logits. The corresponding class probability is $p_\theta(c \mid X_i, S_i) = \mathrm{softmax}(z_i)_c$ and the predicted class is $\hat{y}_i = \arg\max_{c \in \mathcal{Y}} p_\theta(c \mid X_i, S_i)$ where $\mathcal{Y} = \{1, \ldots, C\}$.

## 2.2 COUNTERFACTUAL REASONING AND INTERVENTION LEARNING

**Counterfactual reasoning** is grounded in the intervention principle of causal inference. For a variable $V_j$, the intervention $do(V_j = v_j')$ replaces its original generating mechanism with the fixed value $v_j'$, and then examines how this intervention affects the outcome variables (Pearl, 2009).So if $(X, S, y) \to (X^{cf}, S^{cf}, y^{cf})$ or $P(y|X, S) \neq P(y|X^{cf}, S^{cf})$, then the intervention variable is the causal factor corresponding to the prediction result (Karimi et al., 2021).

For graph and hypergraph models $G = (X, S)$, the input contains both structural information and node attributes. Interventions can therefore be performed in two different spaces. **Structural interventions** operate on graph edges or node-hyperedge incidences in a hypergraph. They are uniformly represented as $S_i^{cf} = \mathcal{I}_S(S_i; P_i)$, where $P_i$ denotes the structural intervention variable. **Node attribute interventions** keep the original structure $S_i$ fixed and modify only a subset of node attributes $X_i^{cf} = \mathcal{I}_X(X_i; M_i, \Delta_i)$. $M_i$ specifies the nodes or feature regions allowed to be intervened on, and $\Delta_i$ denotes the corresponding attribute perturbation (Lucic et al., 2022; Bajaj et al., 2021; Tan et al., 2022; Ma et al., 2022; Verma et al., 2024).

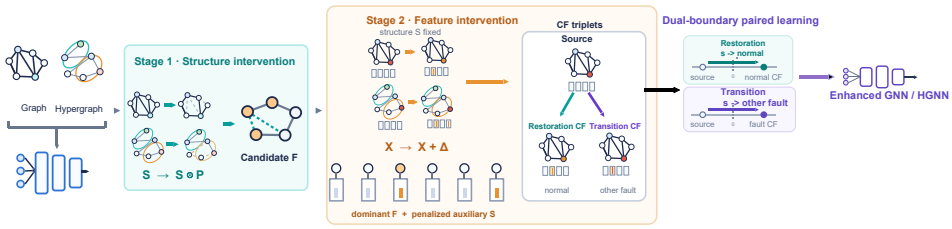



Figure 1: **Overview of the proposed framework.**

### 2.3 Gradient-based Decision Geometry

**Gradient** characterizes the local variation of a prediction function in the input space (Sundararajan et al., 2017). For a source class $s$ and a target class $t$, let $R_{t,s}(X, S)$ denote their relative decision function. Its gradient with respect to the node attributes is $G_{t,s} = \nabla_X R_{t,s}(X, S)$. Around a given point $X$, $G_{t,s}$ represents the local direction that most rapidly increases the decision advantage of target class $t$ over source class $s$. Geometrically, for the local binary decision boundary $R_{t,s}(X, S) = 0$, the gradient $\nabla_X R_{t,s}(X, S)$ can be viewed as the local normal direction of the corresponding level set. Therefore it characterizes the local decision geometry of the model in the input space. $D_{i,t} = X_{i,t}^{cf} - X_i$ describes the actual transition direction induced by an effective counterfactual intervention in the input space.

Use a first-order Taylor expansion $R_{t,s}(X + D, S) = R_{t,s}(X, S) + \nabla_X R_{t,s}(X, S)^\top D + O(\|D\|^2)$. Therefore, when the intervention magnitude is small,$\Delta R_{t,s} \approx G_{t,s}^\top D$. This relationship establishes a geometric connection between finite counterfactual transitions and local decision gradients:$D$ is aligned with $G$. And use cosine aimilarity $\cos(G, D) = \frac{G^\top D}{\|G\|_2 \|D\|_2}$ to measure it.

### 2.4 Problem Formulation

We consider multi-class classification tasks on graphs and hypergraphs. Train a base model $f_{\theta_0}$ on real training data. For a source sample $G_i = (X_i, S_i)$ that is correctly predicted by the model, its source class is $s_i$. For $s_i$, we denote its restoration target as $t_i^r \in \mathcal{Y} \mid \{s_i\}$, such as the normal state in fault diagnosis or the stable state in anomaly detection. Apart from the restoration target, all remaining non-source classes form the candidate class-transition target set $\mathcal{T}_i^c = \mathcal{Y} \mid \{s_i, t_i^r\}$. Therefore, the complete counterfactual target set for sample $i$ is $\mathcal{T}_i = \mathcal{Y} \mid \{s_i\} = \{t_i^r\} \cup \mathcal{T}_i^c$. For any target $t \in \mathcal{T}_i$, we aim to obtain a node-attribute counterfactual $X_{i,t}^{cf}$ that remains valid under the original structure $S_i$ and satisfies $\arg\max_{c \in \mathcal{Y}} p_{\theta_0}(c \mid X_{i,t}^{cf}, S_i) = t$.

We preserve the paring between each original sample and its counterfactual, represent it as $(X_i, s_i) \to (X_{i,t}^{cf}, t)$. And serves it as the basic supervision unit for the subsequent counterfactual decision transition learning.

Based on the above definitions, we focus on two related problems. First, through two-stage validation between structural interventions and node-attribute interventions, we **obtain counterfactual explanations and counterfactual samples**. Second, we further **learn the local class boundaries and decision-transition directions** of the model by exploiting the source–target correspondence between the original and counterfactual samples.

## 3 Two-Stage Counterfactual Discovery and Intervention Validation

A base prediction model is $f_{\theta_0}$ and a source sample is $G_i = (X_i, S_i)$. We first identify candidate causal supports associated with the current prediction through structural counterfactual discovery. We then restore the original structure and apply node-attribute interventions. It's goal is to verify whether these candidate supports can effectively drive the target state transition. The model parameters $\theta_0$ remain fixed throughout the counterfactual generation process.

### 3.1 RESTORATION-ORIENTED STRUCTURAL COUNTERFACTUAL DISCOVERY

Let $\mathcal{R}_i$ denote the set of structural relations in $S_i$ that are allowed to be intervened on. For a standard graph, $\mathcal{R}_i$ corresponds to intervenable edges satisfying $A_{ij} > 0$; whereas for a hypergraph, it corresponds to intervenable node-hyperedge incidences satisfying $H_{ne} > 0$ . For each intervenable relation, we introduce a continuous retention variable $P_{i,r} \in [0, 1], r \in \mathcal{R}_i$ (Ying et al., 2019; Luo et al., 2020; Schlichtkrull et al., 2021; Yuan et al., 2021). A smaller $P_{i,r}$ indicates a stronger attenuation of the corresponding node-related structural relation. The structural counterfactual is formulated as $\widetilde{S}_i = \mathcal{I}_S(S_i; P_i)$. And the actual attenuation of a structural relation is defined as $Q_{i,r} = 1 - P_{i,r}$. We aggregated the attenuation associated with node $v$ as $C_{i,v} = \sqrt{\frac{\sum_{r\in\mathcal{R}_i(v)} Q_{i,r}^2}{|\mathcal{R}_i(v)|+\epsilon}}$ to **obtain a node-level explanation from relation-level structural interventions**. $C_{i,v}$ measures **the contribution of node** $v$ to the current state of the given sample.

A structural counterfactual should not only achieve state restoration, but also keep the intervention as small as possible, concentrate the intervention on a limited number of nodes, and encourage explicit intervention states. We first define the restoration target loss as $L_{\text{tar}} = -\log p_{\theta_0}\left(t_i^r \mid X_i, \widetilde{S}_i\right)$. Further introduce the structural modification constraint $L_{\text{str}} = \sum_{r\in\mathcal{R}_i} Q_{i,r}^2$, the node-level **sparsity** constraint $L_{\text{node}} = \sum_v C_{i,v}$ and the binary regularization of the continuous mask $L_{\text{bin}} = \sum_{r\in\mathcal{R}_i} P_{i,r}(1 - P_{i,r})$.

Therefore the Stage-1 objective is

$$L_{\text{S1}} = \lambda_{\text{tar}} L_{\text{tar}} + \lambda_{\text{str}} L_{\text{str}} + \lambda_{\text{node}} L_{\text{node}} + \lambda_{\text{bin}} L_{\text{bin}}. \tag{1}$$

So the optimal structural intervention is obtained by $P_i^* = \arg\min_{P_i} L_{\text{S1}}$. Based on the resulting node contributions, we extract the causal candidate set as $F_i = \left\{v \mid C_{i,v} \geq \tau\right\}$, where $\tau$ is the contribution threshold; alternatively, a Top-$K$ rule can be used. $F_i$ denotes the candidate causal supports.

### 3.2 NODE-ATTRIBUTE INTERVENTION VALIDATION AND COUNTERFACTUAL GENERATION

We first restore the original structure $S_i$, keep it fixed, and search for a node-attribute counterfactual for each $t \in \mathcal{T}_i$. This stage contains both the restoration transition $s_i \to t_i^r$ and the class transition $s_i \to t, t \in \mathcal{T}_i^c$. Characterize **the reachable decision-boundary relationships** between the current state and other states. The candidate set $F_i$ is used as the primary intervention set, while an auxiliary set $A_i \subseteq V_i | F_i$ is constructed from the remaining nodes. Let $B_i^F$ and $B_i^A$ denote the corresponding node masks. For a target class $t$, the node-attribute counterfactual is defined as $X_{i,t}^{cf} = \Pi_{\mathcal{B}_X}\left[X_i + \left(B_i^F + B_i^A\right) \odot \Delta_{i,t}\right]$, where $\Delta_{i,t}$ is the node-attribute perturbation to be optimized. $\Pi_{\mathcal{B}_X}$ projects the counterfactual into the feasible feature range. It is determined by the real training data or task-specific constraints. Variables that are not allowed to be intervened on satisfy $\Delta_{i,t,n} = 0, n \notin F_i \cup A_i$.

To ensure that the counterfactual not only flips the prediction but also maintains a stable discriminative advantage for the target class, we define the target-class decision margin $M_t(X, S) = z_t(X, S) - \max_{c\neq t} z_c(X, S)$. So the basic node-attribute counterfactual objective is

$$\begin{aligned} L_{\text{cf}}^X = \lambda_{\text{cf}}\, \text{CE}\left(f_{\theta_0}(X_{i,t}^{cf}, S_i), t\right) \\ + \lambda_m\left[\kappa - M_t(X_{i,t}^{cf}, S_i)\right]_+ + \lambda_\delta \|\widetilde{\Delta}_{i,t}\|_2^2 + \lambda_{\text{pl}} L_{\text{pl}}(X_{i,t}^{cf}), \end{aligned} \tag{2}$$

where $\kappa$ is the target margin, $\widetilde{\Delta}_{i,t}$ denotes the perturbation standardized according to the scale of the real training data, and $L_{\text{pl}}$ represents data-plausibility constraints associated with statistical relationships, physical boundaries, or other task-specific requirements (Wachter et al., 2018; Mothilal et al., 2020; Tan et al., 2022; Ma et al., 2022).

We introduce group-sparsity constraint and let $\lambda_A > \lambda_F$, to encourage $F_i$ to preferentially account for the target transition, so that interventions on auxiliary nodes incur a higher cost,

$$L_{\text{spa}} = \lambda_F \sum_{v\in F_i} \|\widetilde{\Delta}_{i,t}^{(v)}\|_2 + \lambda_A \sum_{v\in A_i} \|\widetilde{\Delta}_{i,t}^{(v)}\|_2 + \lambda_1 \|\widetilde{\Delta}_{i,t}\|_1. \tag{3}$$

The final optimization problem is $\Delta_{i,t}^* = \arg\min_{\Delta_{i,t}} \Big( L_{\text{cf}}^X + L_{\text{spa}} \Big)$.

We adopt a hierarchical searching strategy. We first set $B_i^A = \mathbf{0}$ and allow only nodes in $F_i$ to change. If a counterfactual satisfying the strict constraints cannot be obtained, the auxiliary set $A_i$ is then activated. Once a feasible solution is found, greedy pruning and re-optimization are applied to the modified nodes. Further remove unnecessary interventions.

We verify whether the final target transition truly depends on $F_i$. For the same optimized perturbation $\Delta_{i,t}^*$ we construct $X_{i,t}^F = X_i + B_i^F \odot \Delta_{i,t}^*, X_{i,t}^A = X_i + B_i^A \odot \Delta_{i,t}^*, X_{i,t}^{F+A} = X_i + \big(B_i^F + B_i^A\big) \odot \Delta_{i,t}^*$. According to the three independent forward evaluations, counterfactual candidates are categorized into three types:

$$\text{Type}_{i,t} = \begin{cases} \text{F-only}, & \hat{y}(X_{i,t}^F) = t, \\ \text{A-dominant}, & \hat{y}(X_{i,t}^F) \neq t,\ \hat{y}(X_{i,t}^A) = t, \\ \text{Cooperative}, & \hat{y}(X_{i,t}^F) \neq t,\ \hat{y}(X_{i,t}^A) \neq t,\ \hat{y}(X_{i,t}^{F+A}) = t. \end{cases} \tag{4}$$

Finally, a causally valid counterfactual is required to satisfy target prediction, decision margin, actual participation of primary candidate supports, and exclude the A-dominant case. We define **the validity indicator**:

$$\text{Valid}_{i,t} = \mathbb{I}\big[\hat{y}(X_{i,t}^{cf}) = t\big]\, \mathbb{I}\big[M_t(X_{i,t}^{cf}, S_i) \geq \kappa\big]\, \mathbb{I}\big[\|\Delta_{i,t}^F\|_2 > 0\big]\, \mathbb{I}\big[\text{Type}_{i,t} \neq \text{A-dominant}\big]. \tag{5}$$

Only F-only and Cooperative counterfactuals are kept. The whole set of valid counterfactual samples is collected as $\Big\{ (X_i, S_i, s_i, X_{i,t}^{cf}, t) \,\Big|\, t \in \mathcal{T}_i,\ \text{Valid}_{i,t} = 1 \Big\}$. We further compute individual counterfactual effect $\text{ICE}_{i,t} = f_{\theta_0}(X_i, S_i) - f_{\theta_0}(X_{i,t}^{cf}, S_i)$ (Pearl, 2009). Finally the two-stage pipeline outputs sample-wise explanations $\mathcal{E}_i = \big\{ C_{i,v},\ F_i,\ \text{Type}_{i,t},\ \text{ICE}_{i,t} \big\}$ together with the valid counterfactual set $D_{\text{CF}}^{\text{valid}}$.

## 4 Counterfactual Decision-Transition Augmented Learning

We exploit the endpoint, boundary, finite-transition, and local-direction information contained in $(X_i, s_i) \to (X_{i,t}^{cf}, t)$, to locally refine the local decision geometry of the base model.

### 4.1 Dual-Target Counterfactual Triplet Construction

For source sample $i$, it has both a valid restoration counterfactual and at least one valid class-transition counterfactual.We define valid class-transition target set as $\mathcal{V}_i^c = \big\{ t \in \mathcal{T}_i^c \,\big|\, \text{Valid}_{i,t} = 1 \big\}$.Denote the counterfactual for restoration target as $X_{i,r}^{cf} = X_{i,t_i^r}^{cf}$. If multiple valid class-transition targets exist, we select the target achieving highest confidence under the base model:$t_i^c = \arg\max_{t \in \mathcal{V}_i^c} p_{\theta_0}\big(t \mid X_{i,t}^{cf}, S_i\big)$. And denote its corresponding counterfactual as $X_{i,c}^{cf} = X_{i,t_i^c}^{cf}$. We then construct **the dual-target counterfactual triplet** $\mathcal{Q}_i = \big(X_i,\ X_{i,r}^{cf},\ X_{i,c}^{cf}\big)$. They correspond respectively to the two transitions $s_i \to t_i^r, s_i \to t_i^c$.

### 4.2 Dual-Branch Decision Transition Learning

The augmented model is initialized directly from the base model $\theta \leftarrow \theta_0$. Real training data remains the major supervision source for classification. For a real-data training batch $\mathcal{B}_R$, we define real-data classification loss with class weights $L_{\text{real}} = \frac{1}{|\mathcal{B}_R|} \sum_{j \in \mathcal{B}_R} w_{y_j}\, \text{CE}\big(f_\theta(X_j, S_j), y_j\big)$, where class weight $w_{y_j}$ is computed purely from real training statistics.

We unify the formulation for two branches by introducing branch indicator $b \in \{r, c\}$: $b = r$ for **restoration branch** and $b = c$ for **class-transition branch**. The corresponding target class is $t_i^b = \begin{cases} t_i^r & b = r \\ t_i^c & b = c \end{cases}$, and counterfactual sample are written as $X_{i,b}^{cf} = \begin{cases} X_{i,r}^{cf} & b = r \\ X_{i,c}^{cf} & b = c \end{cases}$. We use

the class labels of the source and counterfactual samples to preserve the class semantics at both endpoints $L^b_{\text{cls}} = \text{CE}\big(f_\theta(X_i, S_i), s_i\big) + \text{CE}\big(f_\theta(X^{cf}_{i,b}, S_i), t^b_i\big)$ (Khosla et al., 2020).

To ensure that each class remains sufficiently separated from competing classes, for an arbitrary class $c$, we define the multi-**class decision margin** as $M_c(X, S) = z_c(X, S) - \max_{k \neq c} z_k(X, S)$. So for branch $b$, we constrain the **source and target endpoints** through $L^b_{\text{mar}} = \big[m - M_{s_i}(X_i, S_i)\big]_+ + \big[m - M_{t^b_i}(X^{cf}_{i,b}, S_i)\big]_+$, where $m > 0$ is predefined minimal margin.

To measure decision change along direction $s_i \to t^b_i$, we define source-target relative decision function $R_{t,s}(X, S) = z_t(X, S) - z_s(X, S)$. The decision change in branch $b$ is $\Delta R^b_i = R_{t^b_i, s_i}\big(X^{cf}_{i,b}, S_i\big) - R_{t^b_i, s_i}\big(X_i, S_i\big)$. We encourage the target over source advantage to increase by at least $m_\Delta$:$L^b_\Delta = \big[m_\Delta - \Delta R^b_i\big]_+$. This explicitly exploits the pairing between the two to constrain the magnitude of the decision transition.

Let $D^b_i = X^{cf}_{i,b} - X_i$ denote the actual input transition direction from source to counterfactual target. Following Section 2.3, the gradient of relative decision function evaluated on source sample is $G^b_i = \nabla_X R_{t^b_i, s_i}(X_i, S_i)$ (Sundararajan et al., 2017), It represents the local direction around the source sample. In this direction, the decision advantage of the target class over the source class increases.So we enforce alignment between model gradient $G^b_i$ and empirically validated counterfactual direction $D^b_i$ via cosine-based consistency loss,$L^b_{\text{GS}} = 1 - \frac{\langle G^b_i, D^b_i \rangle}{\|G^b_i\|_2 \, \|D^b_i\|_2 + \epsilon}$. The counterfactual direction $D^b_i$ is treated as a fixed supervision signal during augmented training. And gradients are not propagated through it.

So the counterfactual decision-transition loss for each semantic branch is formulated as

$$L_b = \lambda_{\text{cls}} L^b_{\text{cls}} + \lambda_{\text{mar}} L^b_{\text{mar}} + \lambda_\Delta L^b_\Delta + \lambda_{\text{GS}} L^b_{\text{GS}}, \quad b \in \{r, c\}. \tag{6}$$

Finally, the overall augmentation objective is

$$L = L_{\text{real}} + \lambda_r L_r + \lambda_c L_c. \tag{7}$$

$L_{\text{real}}$ preserves global classification performance learned from real data. $L_r$ regularizes local decision boundaries between source states and restoration targets. $L_c$ shapes competitive decision boundaries between source samples and other reachable classes.

# 5 Experiments

## 5.1 Experimental Setup

**Datasets and backbones.** We evaluate the proposed framework on two public benchmarks with different graph structures. TEP contains one normal condition and 28 process-fault classes (**?**). We use GNN-TAM (Kovalenko et al., 2024) to learn a directed graph among 52 process variables. Paderborn contains bearing signals collected under multiple operating conditions (**?**). We formulate a three-class task consisting of Healthy, Outer-race Fault, and Inner-race Fault. And apply a two-layer HGNN (Feng et al., 2019) to a fixed physical hypergraph with seven nodes and five hyperedges. The train/validation/test sizes are 1904/336/560 for TEP and 1623/348/348 for Paderborn.

**Base Models.**On TEP, we use GNN-TAM as the backbone. It jointly learns a directed sensor graph and a fault-classification representation (Kovalenko et al., 2024). On Paderborn, we use a two-layer HGNN for graph-level classification on the fixed physical hypergraph (Feng et al., 2019).

**Baseline.** We compare counterfactual generation with **RCExplainer**(Tan et al., 2022), **CF**$^2$(Tan et al., 2022), and **CLEAR**(Ma et al., 2022). For decision-transition learning, we compare against **Mixup**(Zhang et al., 2018), **FLAG**(Kong et al., 2020), **Naive CF Augmentation**, **Endpoint CE only**, and the corresponding ablations without **finite-transition** or **gradient-alignment constraints**.

On TEP, we use one GNN-TAM graph module with two graph-convolution layers (hidden dimension 1024), trained with Adam (learning rate $10^{-3}$, batch size 256). Stage 1 initializes the retention mask at 0.98 and uses 500 acquisition and 1000 refinement steps (learning rate $5 \times 10^{-2}$, $\tau = 0.05$). Triplet learning uses a batch size of 16, endpoint/transition margins of 0.5/1.0, and margin/transition/gradient weights of 0.5/0.25/0.05 (Adam, $10^{-4}$). On Paderborn, a two-layer HGNN

(hidden dimension 64, GELU, dropout 0.2) is trained with AdamW ($10^{-3}$). Stage 1 uses at most 800 steps (0.05, $\tau = 0.05$). Triplet learning updates the final HGNN block and classifier with a batch size of 14 and learning rate $5 \times 10^{-6}$; restoration/class-transition weights are 0.1/0.3, and margin/transition/gradient weights are 0.5/0.25/0.02.

**Evaluation Metrics.** The main evaluation considers counterfactual validity, support sparsity, cross-intervention validity, Macro-F1, Balanced Accuracy, and source-to-target decision margins. Complete hyperparameters, metric definitions, extended audits, and additional ablations are reported in the Appendix(Amara et al., 2022; Kosan et al., 2024).

### 5.2 Results and Analysis

We evaluate the proposed framework from three perspectives: (1) the validity and sparsity of the generated counterfactuals; (2) whether the structural supports identified in Stage 1 provide effective intervention regions for Stage 2; and (3) whether counterfactual decision-transition learning improves classification performance and local decision boundaries. Additionally, Figures 8 and 9 illustrate the **complete counterfactual reasoning process**. They demonstrates the interpretability of our method and shows that feature changes within $F_i$ causally induce changes in the model predictions.

**Counterfactual Quality.** Table 1 compares the generation quality of different counterfactual methods on TEP. CLEAR obtains the highest target flip rate, whereas our method attains a substantially smaller average support ratio 0.279 with flip rate of 78.57%. It indicates that our method localizes concentrates successful interventions on on a sparser structural support involving fewer nodes. This matches more practical expectations, that the key cause points really leading to the result are few. The expanded-source evaluation in the appendix further examines the sensitivity of this comparison to sample selection.

Table 1: Counterfactual generation quality on TEP.

| Method | Target Flip Rate | Avg. Size | CIV | Runtime |
|---|---|---|---|---|
| RCExplainer | 7.14% | 0.923 | 7.14% | 0.098 |
| CF² | 0.00% | – | 0.00% | 0.107 |
| CLEAR | 90.00% | 0.962 | 86.43% | 0.203 |
| Ours | 78.57% | 0.279 | 75.00% | 0.985 |

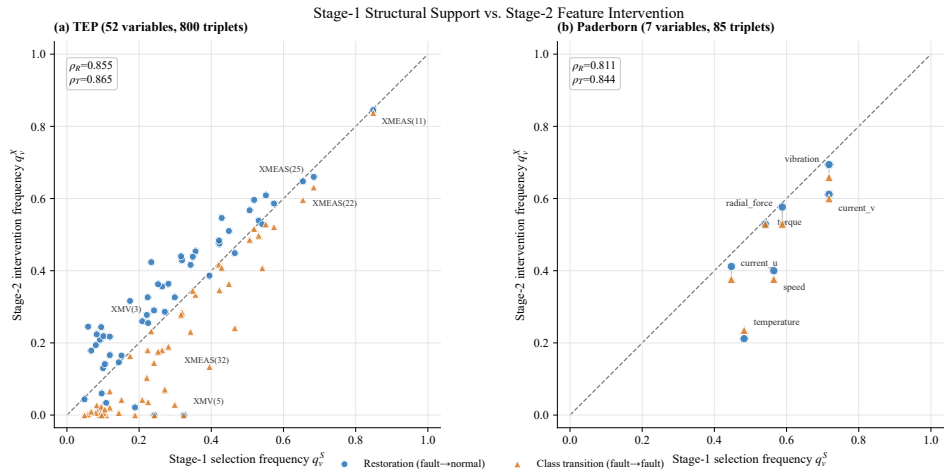



Figure 2: **Cross-intervention support consistency.** Stage-1 structural selection frequency versus Stage-2 feature-intervention frequency on TEP and Paderborn.

**Cross-stage intervention consistency.** Figure 2 compares how frequently each node is selected in Stage 1 and modified in Stage 2. For each node $v$, we compute its Stage-1 selection frequency $q_v^S$ and

Stage-2 intervention frequency $q_v^X$ which are defined in Appendix. The Spearman correlations are 0.855/0.865 on TEP and 0.811/0.844 on Paderborn for the restoration and class-transition branches. Together with the $F$-only and size-matched Random-$F$ ablations in Table 2, this result suggests that the underlying drivers in the model predictions identified in Stage 1 are actually related to interventions. And also calculated the drivers' contribution scores shown in Figure 3. This provides an intervention-backed, sample-level explanation of the model's prediction.

**Effect of the Stage-1 support.** Table 2 tests whether the discovered support $F_i$ provides a meaningful search region on TEP. When interventions are restricted to $F_i$, the average changed ratio decreases to 17.0%, while the target flip rate remains at 93.18%. It indicates that the final perturbations remain largely concentrated on the $F_i$, while auxiliary nodes are mainly used for supplementing. Random-$F$ achieves a target flip rate of only $84.09\% \pm 1.61\%$. Indicate that $F_i$ is not merely an attribution-based importance set, but an interventionally relevant and more compact explanation of the model prediction(Qu et al., 2025).

Table 2: Ablation of Stage-2 intervention strategies.

| Dataset | Strategy | Validity (%) | Dual-target Success(%) | Avg. Size | $F$-energy (%) |
|---|---|---|---|---|---|
| TEP | All nodes | 97.73 | 95.45 | 0.480 | 39.34 |
| | Random-F | $84.09 \pm 1.61$ | $68.18 \pm 3.21$ | $0.143 \pm 0.006$ | $36.14 \pm 2.07$ |
| | $F$-only | 93.18 | 86.36 | 0.170 | 100.00 |
| | **F+A** (Ours) | **97.73** | **95.45** | 0.197 | 98.83 |
| Paderborn | All nodes | 100 | 100 | 0.680 | 57.32 |
| | Random-F | 100.00 | 100.00 | $0.473 \pm 0.010$ | $58.01 \pm 3.51$ |
| | $F$-only | 100.00 | 100.00 | 0.468 | 100.00 |
| | **F+A** (Ours) | 100.00 | 100.00 | **0.468** | 100.00 |

**Baseline and ablation of Counterfactual Decision Transition Learning**. Table 6 in the APPENDIX reports the downstream results. On TEP, our method improves Macro-F1 from 90.15% to 91.89% and BA from 90.10% to 90.74%. The restoration and class-transition margins increase from 9.360/10.357 to 16.850/13.954. Compared with standard augmentation, these improvements show that counterfactual triplets provide supervision not only for class labels but also for source-to-target decision transitions. Paderborn already achieves a Macro-F1 of 96.43%; our method preserves this performance but produces only minor margin changes. These results indicate that the finite-transition and local-direction constraints further help the model.

**Decision-Transition Geometry.** Figure 7 in the APPENDIX visualizes the decision trajectories for the restoration-transition and class-transition directions. The shaded regions denote 95% confidence intervals obtained by resampling the source samples. Overall, the full model achieves a larger average target-class advantage near the counterfactual endpoint as $\alpha \to 1$.

**Per-class Analysis.** We further compare the per-class F1 changes between the base model and the full model across 28 fault classes. As shown in Figure 5, the proposed method improves the F1 scores of most fault classes.

## 6 Conclusion

We presented a counterfactual intervention validation and decision-transition learning framework for GNNs and HGNNs. It identifies sparse structural supports and validating them through feature interventions on the restored graph. So it produces explanations that are both compact and interventionally grounded. And it organizes validated restoration and class-transition counterfactuals into dual-target triplets and aligns model gradients with actual intervention directions. So it converts counterfactual explanations into structured supervision for refining fault-to-normal and inter-class decision boundaries. Experiments show that the identified supports enable more effective interventions than size-matched random supports. The method improves local decision boundaries and difficult-class recognition. Future work will investigate more efficient counterfactual reasoning and enhancement strategies for hypergraphs.

## Reproducibility Statement

To facilitate reproducibility, we provide the complete algorithms, implementation details, dataset preprocessing and split protocols, hyperparameter configurations, evaluation definitions, and additional experimental results in the Appendix and supplementary material. The supplementary package contains anonymized source code for the two-stage counterfactual procedure and counterfactual decision-transition learning, together with configuration files and scripts for reproducing the reported experiments, tables, and figures. All experiments use publicly available TEP and Paderborn datasets, and the random seeds used for each experiment are explicitly documented.

### AI Use Statement

Generative AI tools were used to assist with writing refinement, English language editing, literature retrieval and discovery, and drafting parts of the manuscript. The use of AI tools was limited to assisting the presentation and organization of the manuscript. All methodological decisions, experimental implementations, numerical results, and scientific claims were developed, verified, and reviewed by the authors.

## References

Kenza Amara, Zhitao Ying, Zitao Zhang, Zhichao Han, Yang Zhao, Yinan Shan, Ulrik Brandes, Sebastian Schemm, and Ce Zhang. GraphFramEx: Towards systematic evaluation of explainability methods for graph neural networks. In *Proceedings of the First Learning on Graphs Conference*, volume 198 of *Proceedings of Machine Learning Research*, pp. 44:1–44:23. PMLR, 2022.

Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are AllSet: A multiset function framework for hypergraph neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hpBTIv2uy_E.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3558–3565, 2019. doi: 10.1609/aaai.v33i01.33013558.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1024–1034, 2017.

Xiaotian Han, Zhimeng Jiang, Ninghao Liu, and Xia Hu. G-Mixup: Graph data augmentation for graph classification. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8230–8248. PMLR, 2022.

Zexi Huang, Mert Kosan, Sourav Medya, Sayan Ranu, and Ambuj K. Singh. Global counterfactual explainer for graph neural networks. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining*, pp. 141–149. ACM, 2023. doi: 10.1145/3539597.3570376.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 353–362. ACM, 2021. doi: 10.1145/3442188.3445899.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Kezhi Kong, Guohao Li, Mucong Ding, Zuxuan Wu, Chen Zhu, Bernard Ghanem, Gavin Taylor, and Tom Goldstein. FLAG: Adversarial data augmentation for graph neural networks. *arXiv preprint arXiv:2010.09891*, 2020.

Mert Kosan, Samidha Verma, Burouj Armgaan, Khushbu Pahwa, Ambuj K. Singh, Sourav Medya, and Sayan Ranu. GNNX-BENCH: Unravelling the utility of perturbation-based GNN explainers through in-depth benchmarking. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VJvbOSXRUq.

Aleksandr Kovalenko, Vitaliy Pozdnyakov, and Ilya Makarov. Graph neural networks with trainable adjacency matrices for fault diagnosis on multivariate sensor data. *IEEE Access*, 12:152860–152872, 2024. doi: 10.1109/ACCESS.2024.3481331.

Wanyu Lin, Hao Lan, and Baochun Li. Generative causal explanations for graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6666–6679. PMLR, 2021.

Ana Lucic, Maartje A. Ter Hoeve, Gabriele Tolomei, Maarten De Rijke, and Fabrizio Silvestri. CF-GNNExplainer: Counterfactual explanations for graph neural networks. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4499–4511. PMLR, 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, 2020.

Jing Ma, Ruocheng Guo, Saumitra Mishra, Aidong Zhang, and Jundong Li. CLEAR: Generative counterfactual explanations on graphs. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1878.

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*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15524–15543. PMLR, 2022.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617. ACM, 2020. doi: 10.1145/3351095.3372850.

Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4574–4594. PMLR, 2022.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2 edition, 2009. ISBN 9780521895606.

Mario Alfonso Prado-Romero, Bardh Prenkaj, Giovanni Stilo, and Fosca Giannotti. A survey on graph counterfactual explanations: Definitions, methods, evaluation, and research challenges. *ACM Computing Surveys*, 56(7):171:1–171:37, 2024. doi: 10.1145/3618105.

Zhan Qu, Daniel Gomm, and Michael Färber. CoDy: Counterfactual explainers for dynamic graphs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50762–50785. PMLR, 2025.

Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. Interpreting graph neural networks for NLP with differentiable edge masking. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=WznmQa42ZAx.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017.

Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. In *Proceedings of the ACM Web Conference 2022*, pp. 1018–1027. ACM, 2022. doi: 10.1145/3485447.3511948.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Samidha Verma, Burouj Armgaan, Sourav Medya, and Sayan Ranu. InduCE: Inductive counterfactual explanations for graph neural networks. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=RZPN8cgqST.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6438–6447. PMLR, 2019.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31 (2):841–887, 2018.

Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. HyperGCN: A new method for training graph convolutional networks on hypergraphs. In *Advances in Neural Information Processing Systems*, volume 32, pp. 1509–1520, 2019.

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, pp. 9240–9251, 2019.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5812–5823, 2020.

Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. Graph contrastive learning automated. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12121–12132. PMLR, 2021.

Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12241–12252. PMLR, 2021.

Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5782–5799, 2023. doi: 10.1109/TPAMI.2022.3204236.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

# A Appendix

## A.1 Supplementary Experimental Details

Table **??** summarizes the dataset statistics and graph configurations.

Table 3: Dataset statistics and graph configurations.

| Dataset | Classes | Train / Val / Test | Nodes | Structure | Backbone |
|---|---|---|---|---|---|
| TEP | 29 | 1904 / 336 / 560 | 52 | Learned graph | GNN-TAM |
| Paderborn | 3 | 1623 / 348 / 348 samples | 7 | 5 hyperedges | HGNN |

### A.2 Evaluation Metrics

For a source sample $i$, let $s_i$ denote the original class and $t_i$ the desired counterfactual target.

The counterfactual validity is defined as

$$\text{Validity} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\Big[\hat{y}(X_i^{cf}) = t_i\Big].$$

The average node-support ratio is

$$\text{AvgSupport} = \frac{1}{N_{\text{cf}}}\sum_{i=1}^{N_{\text{cf}}}\frac{|M_i|}{|V_i|},$$

where $M_i$ is the set of nodes with nonzero final intervention and $V_i$ is the complete node set.

Cross-intervention validity measures whether a generated counterfactual remains valid under the subsequent intervention procedure:

$$\text{CIV} = \frac{\#\{\text{counterfactuals valid under both stages}\}}{\#\{\text{source samples}\}}.$$

For the Stage-2 support concentration, we define

$$\text{F-Energy} = \frac{\sum_{v\in F_i}\|\Delta_i^{(v)}\|_2^2}{\sum_{v\in V_i}\|\Delta_i^{(v)}\|_2^2 + \epsilon}.$$

For a paired restoration and transition triplet, the branch retention rates are

$$R_{\text{normal}} = \frac{1}{N}\sum_i \mathbb{I}[\hat{y}(X_{i,r}^{cf}) = t_i^r],$$

$$R_{\text{fault}} = \frac{1}{N}\sum_i \mathbb{I}[\hat{y}(X_{i,c}^{cf}) = t_i^c].$$

The joint retention rate requires both branches to satisfy their target labels:

$$R_{\text{joint}} = \frac{1}{N}\sum_i \mathbb{I}\Big[\hat{y}(X_{i,r}^{cf}) = t_i^r \wedge \hat{y}(X_{i,c}^{cf}) = t_i^c\Big].$$

For a target class $t$ and source class $s$, the relative decision margin at the counterfactual endpoint is

$$M_{t,s}(X^{cf}) = z_t(X^{cf}) - z_s(X^{cf}),$$

where $z_c(\cdot)$ denotes the logit of class $c$.

The finite-transition satisfaction rate is

$$\mathrm{FTS} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[M_{t_i,s_i}(X_i^{cf}) - M_{t_i,s_i}(X_i) \geq m_\Delta\right].$$

The local gradient-direction agreement is measured by

$$\mathrm{Cos} = \frac{\langle \nabla_X M_{t_i,s_i}(X_i), D_i \rangle}{\|\nabla_X M_{t_i,s_i}(X_i)\|_2 \, \|D_i\|_2 + \epsilon},$$

where $D_i = X_i^{cf} - X_i$ is the validated counterfactual direction.

Finally, the consistency between Stage 1 structural selection and Stage 2 intervention is measured using Spearman correlation:

$$\rho_S = \mathrm{Spearman}\left(\{q_v^S\}_{v\in V}, \{q_v^X\}_{v\in V}\right),$$

where

$$q_v^S = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[v \in F_i], \qquad q_v^X = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\|\Delta_i^{(v)}\|_2 > 0\right].$$

### A.3 Class-wise Structural Contribution Analysis

For each class $c$, the class-wise mean contribution of node $v$ is defined as

$$\bar{C}_{c,v} = \frac{1}{N_c}\sum_{i:y_i=c} C_{i,v},$$

where $N_c$ denotes the number of complete counterfactual triplets available for class $c$.

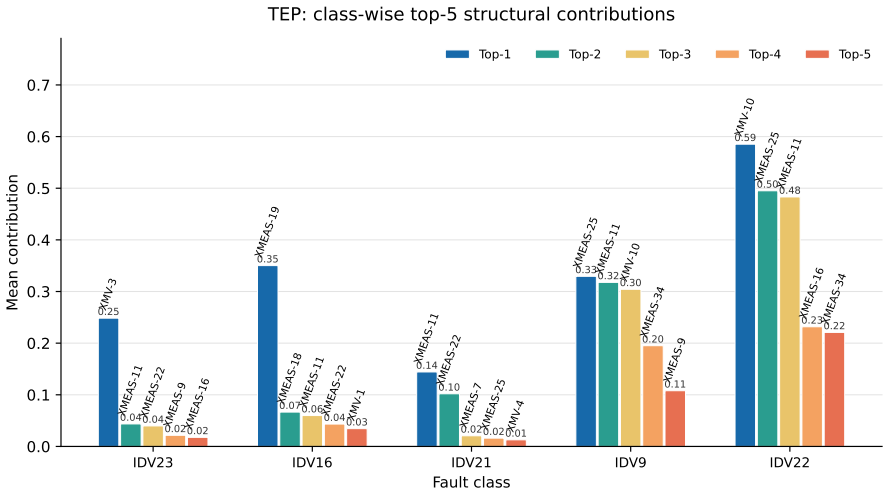



Figure 3: Class-wise top-5 structural contribution scores on TEP. The horizontal axis denotes the fault class and the vertical axis denotes the mean Stage-1 contribution. Each class contains its five highest-ranked nodes. The selected classes include IDV23, IDV16, IDV21, IDV9, and IDV22.

On TEP, the dominant contribution nodes vary across fault classes. For example, XMV(3) is the highest-contribution node for IDV23, XMEAS(19) dominates IDV16, and XMV(10) is the top node for IDV22. This class dependence indicates that Stage 1 does not assign a uniform importance

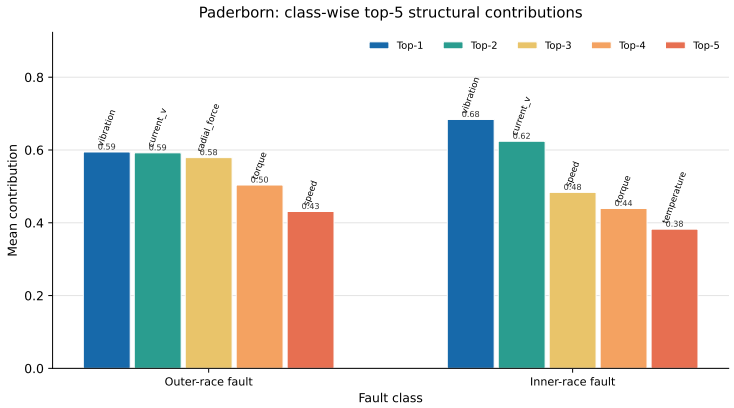



Figure 4: Class-wise top-5 structural contribution scores on Paderborn.

pattern to all faults. On Paderborn, vibration and current-$v$ are the most consistently important nodes for both outer-race and inner-race faults, which is consistent with their physical relevance to bearing condition monitoring. These results support the interpretation of $F_i$ as model-level causal candidate nodes rather than confirmed physical root causes.

Figures 3 and 4 provide a class-specific analysis of the structural contribution scores. On TEP, the dominant contribution nodes vary across fault classes, indicating that different faults rely on different structural supports. For example, XMV(3) is the highest-contribution node for IDV23, with a mean contribution of 0.248. XMEAS(19) dominates IDV16 with a mean contribution of 0.350, while XMV(10) is the top node for IDV22 with a mean contribution of 0.585. These class-dependent patterns suggest that counterfactual method does not assign a uniform importance pattern to all fault classes, but instead identifies sample- and class-specific candidate supports. On Paderborn, vibration and current-$v$ are the most consistently important nodes for both outer-race and inner-race faults. Their mean contribution scores are approximately 0.59–0.68 across the two fault classes, which is consistent with the physical relevance of vibration and electrical response in bearing condition monitoring. Overall, these results support the interpretation of $F_i$ as model-level causal candidate nodes that provide more specific and interpretable structural support for each fault prediction.

### A.4 Per-class Analysis

We further compare the per-class F1 changes between the base model and the full model across 28 fault classes. As shown in Figure 5, the proposed method improves the F1 scores of most fault classes. In particular, the F1 score of IDV23 increases from 70.52% to 81.37%, corresponding to a gain of 10.85 percentage points, while its recall improves from 55.72% to 77.19%. The F1 scores of IDV16 and IDV21 increase by 4.06 and 5.05 percentage points, respectively. In contrast, IDV9 and IDV22 decrease by 3.11 and 1.78 percentage points. These results indicate that the benefits of counterfactual feedback are mainly reflected in several difficult classes and in balanced evaluation metrics.

### A.5 Dual-Branch Training Ablation

As shown in Table 4, the dual-branch model on TEP achieves the highest joint retention rate (95.21%), exceeding the restoration-only and transition-only variants (94.42% and 60.54%, respectively). It also obtains the best Macro-F1 (90.75%), indicating that the restoration and alternative-fault branches provide complementary supervision for learning two local decision boundaries. Thus, the main advantage of the dual branch is its more balanced joint behavior. On Paderborn, all variants preserve the same classification performance and nearly saturated retention rates, indicating that the dual branch is applicable but provides no clear additional gain on this already well-separated task.

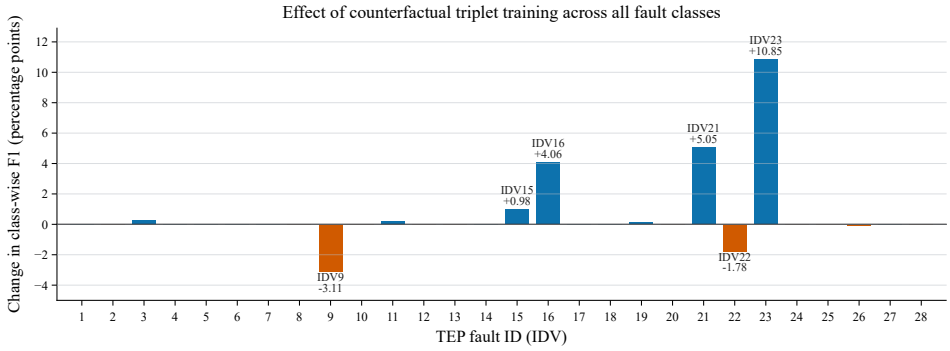



Figure 5: Per-class F1 changes on TEP. Changes in class-wise F1 after counterfactual decision-transition learning across the 28 fault classes.

Table 4: Ablation of the dual-target counterfactual training branches.

| Variant | Macro-F1 (%) | BA (%) | Normal (%) | Fault (%) | Joint (%) |
|---|---|---|---|---|---|
| | | | **TEP** | | |
| Restoration only | $90.73 \pm 0.05$ | $90.71 \pm 0.11$ | $97.83 \pm 1.16$ | $96.58 \pm 0.51$ | $94.42 \pm 0.62$ |
| Transition only | $90.72 \pm 0.03$ | $90.53 \pm 0.05$ | $61.38 \pm 0.33$ | $99.17 \pm 0.47$ | $60.54 \pm 0.19$ |
| Dual branch | $\mathbf{90.75 \pm 0.04}$ | $90.62 \pm 0.15$ | $96.71 \pm 0.92$ | $98.67 \pm 0.52$ | $\mathbf{95.21 \pm 1.46}$ |
| | | | **Paderborn** | | |
| Restoration only | $96.43 \pm 0.00$ | $96.91 \pm 0.00$ | 100.00 | $99.22 \pm 0.68$ | $99.22 \pm 0.68$ |
| Transition only | $96.43 \pm 0.00$ | $96.91 \pm 0.00$ | 100.00 | $98.82 \pm 0.00$ | $98.82 \pm 0.00$ |
| Dual branch | $96.43 \pm 0.00$ | $96.91 \pm 0.00$ | 100.00 | $98.82 \pm 0.00$ | $98.82 \pm 0.00$ |

Table 5: Sensitivity to the Stage-1 sparsity penalty on TEP.

| $\ell_1$ weight | Validity (%) | Support ratio | CIV (%) |
|---|---|---|---|
| 0.05 | 75.54 | 0.462 | 70.50 |
| 0.10 | 71.94 | 0.389 | 66.19 |
| 0.20 | 63.31 | 0.322 | 58.99 |
| 0.30 | 64.75 | 0.298 | 59.71 |
| 0.50 | 51.80 | 0.245 | 49.64 |

### A.6 Validity–Sparsity Trade-off

Table 5 and Figure 6 show the effect of the Stage-1 sparsity penalty on counterfactual validity and intervention support. Increasing the $\ell_1$ weight generally reduces the average support ratio, from 0.462 at 0.05 to 0.245 at 0.50, but also decreases the validity and CIV. The default weight of 0.30 provides a relatively balanced solution, achieving 64.75% validity, 59.71% CIV, and a support ratio of 0.298. The slight non-monotonic change between 0.20 and 0.30 is attributed to the stochastic optimization process. In particular, the 0.20 setting is dominated by the 0.30 setting, which achieves both higher validity and a smaller support ratio. These results show that there's a trade-off between sparse structure explanations and successfully generating counterfactuals.

### A.7 Decision-Transition Geometry

To examine whether the enhanced model learns the decision geometry associated with the generated counterfactuals, we interpolate between each source sample $X_i$ and its validated counterfactual endpoint $X_{i,t}^{cf}$ in Figure 7b.Here, $\alpha = 0$ corresponds to the original source sample and $\alpha = 1$ corresponds to the counterfactual endpoint. Figure **??** presents the average trajectories for the fault-to-normal restoration branch and the fault-to-fault class-transition branch.

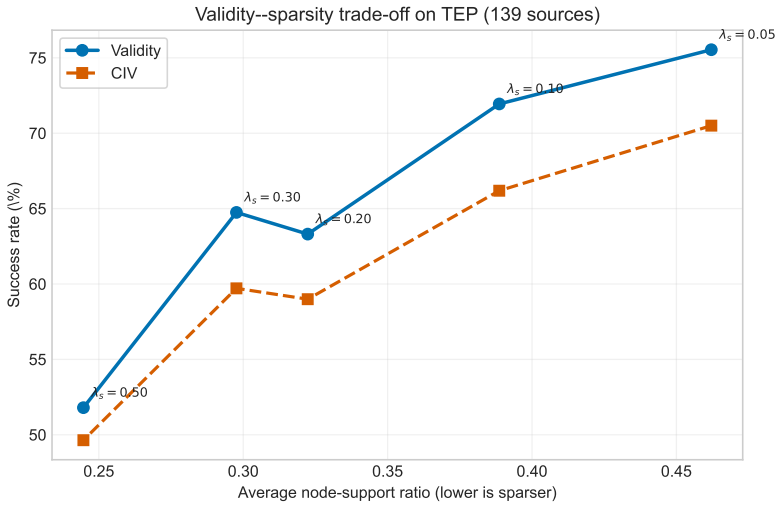
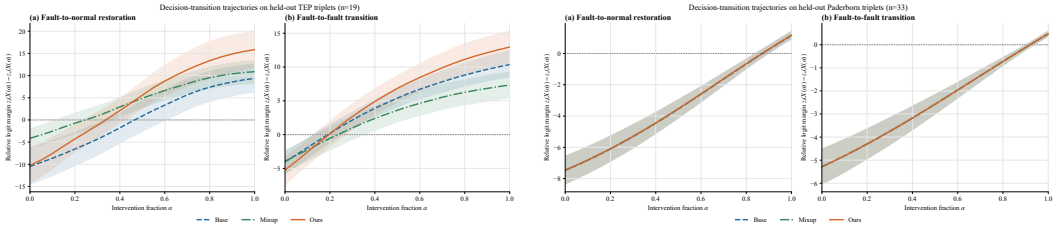



Figure 6: Validity–sparsity trade-off under different Stage-1 penalties.



(a) TEP (b) Paderborn

Figure 7: Counterfactual decision-transition trajectories. Relative source-to-target logit margins along the counterfactual intervention path for (a) TEP and (b) Paderborn. The shaded regions denote 95% confidence intervals.

### A.8 Counterfactual Decision Transition Learning

We compare the proposed counterfactual decision-transition learning method with conventional augmentation and naive counterfactual augmentation in Table 6. We further remove individual components from the paired training objective to examine the contributions of endpoint classification, finite-transition supervision, and gradient-direction alignment.

Table 6: Classification performance and counterfactual endpoint margins. R/T denotes restoration/transition margins at the counterfactual endpoint.

| | TEP | | | Paderborn | |
|---|---|---|---|---|---|
| Method | Macro-F1 (%) | BA (%) | R/T Margin | Macro-F1 (%) | R/T Margin |
| Base | 90.15 | 90.10 | 9.360 / 10.357 | 96.43 | 1.149 / 0.473 |
| Mixup | 90.70 | 90.10 | 10.886 / 7.337 | 96.43 | 1.163 / 0.473 |
| FLAG | 90.54 | 90.69 | 8.188 / 7.638 | 96.43 | 1.150 / 0.468 |
| Naive CF Aug. | 90.73 | 90.47 | 16.192 / 12.682 | 96.43 | 1.171 / 0.475 |
| Endpoint CE only | 90.71 | 90.43 | 16.166 / 12.846 | 96.43 | 1.167 / 0.474 |
| w/o Finite Trans. | 91.25 | 90.44 | 15.873 / 12.939 | 96.43 | 1.170 / 0.475 |
| w/o Grad. Align. | 91.25 | 90.46 | 15.993 / 13.043 | 96.43 | 1.170 / 0.475 |
| **Ours** | 91.89 | 90.74 | 16.850 / 13.954 | 96.43 | 1.170 / 0.475 |

### A.9 Representative Counterfactual Reasoning Cases

We show a sample case from each dataset to present the full counterfactual reasoning process at the sample level. These cases clearly show the cause set and contribution in the first stage, the feature intervention in the second stage, and the prediction changes caused in both the recovery branch and the category conversion branch.

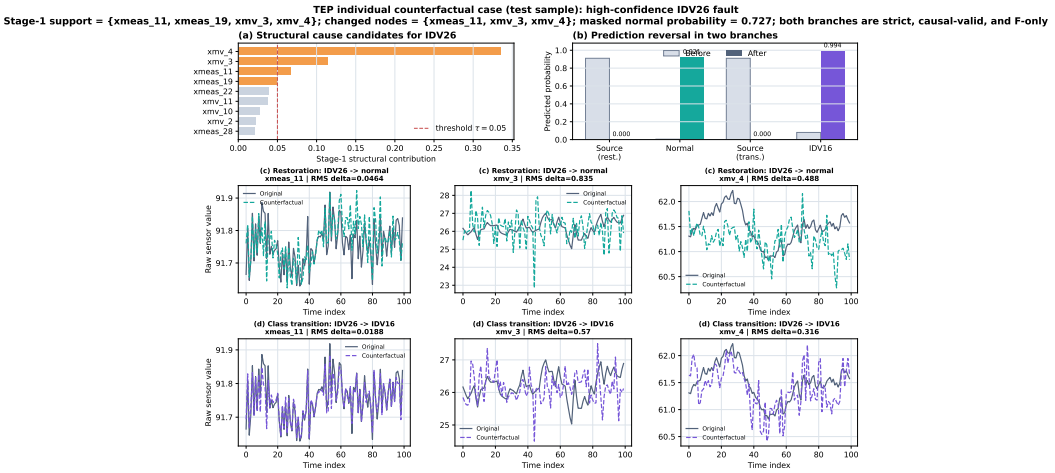



Figure 8: Representative counterfactual reasoning case IDV26 test sample on TEP. Panel (a) shows the Stage-1 structural contribution scores and the selected support $F_i$. Panel (b) reports the source and target probabilities before and after intervention. Panels (c) and (d) compare the original and counterfactual sensor trajectories for the restoration and class-transition branches, respectively.

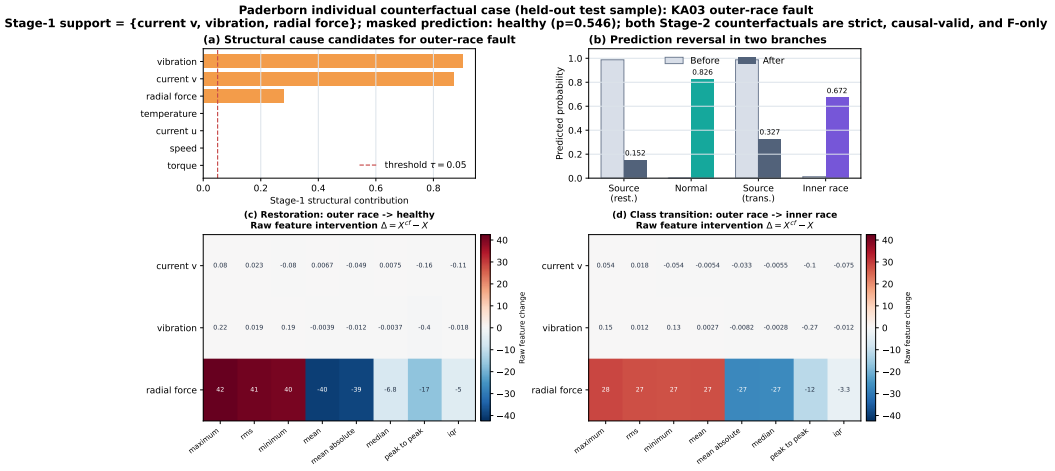



Figure 9: Representative counterfactual reasoning case KA03 outer-race fault sample on Paderborn. Panel (a) shows the node contribution scores obtained from the structural intervention. Panel (b) reports prediction changes in the restoration and class-transition branches. Panels (c) and (d) visualize the feature changes applied to the selected hypergraph nodes.

As shown in Figure 8, the original sample is classified as IDV26 with a probability of 0.910. Stage 1 identifies $F_i = \{\mathrm{XMEAS}(11), \mathrm{XMEAS}(19), \mathrm{XMV}(3), \mathrm{XMV}(4)\}$ as the set of causal factors. Stage 2 modifies nodes within this support: $\mathrm{XMEAS}(11)$, $\mathrm{XMV}(3)$, and $\mathrm{XMV}(4)$. In the restoration branch, the normal-class probability increases from 0.004 to 0.921, while the IDV26 probability decreases from 0.910 to below $10^{-3}$. In the class-transition branch, the IDV16 probability increases from 0.080 to 0.994, while the original-class probability again falls below $10^{-3}$. Both

counterfactuals only used nodes from $F_i$. So, this case provides a sparse, traceable explanation for the prediction shift: $F_i$ are the main causal nodes for IDV26.

Figure 9 presents a KA03 outer-race fault sample whose original source-class probability is 0.987. Stage 1 selects current-$v$, vibration, and radial force as $F_i$. These nodes participate in the electrical, mechanical, bearing-response, electromechanical, and load-related hyperedges. In the restoration branch, interventions on these three nodes increase the healthy-class probability from 0.0005 to 0.826 and reduce the outer-race probability from 0.987 to 0.152. In the class-transition branch, the inner-race probability increases from 0.012 to 0.672, while the outer-race probability decreases to 0.327. The feature-change heatmaps further specify the intervention direction and magnitude for each selected node. So, counterfactual reasoning can explain the causal relationship between sensor states and faults through the sensor variables and feature changes responsible for the transitions.

### A.10 Two-Stage Counterfactual Discovery

**Algorithm 1** Two-Stage Counterfactual Discovery and Intervention Validation

**Require:** Frozen model $f_{\theta_0}$, sample $G_i = (X_i, S_i)$, source $s_i$, restoration target $t_i^r$, transition set $\mathcal{T}_i^c$, threshold $\tau$

Construct intervenable relations $\mathcal{R}_i$ and initialize retention mask $P_i$
$P_i^* \leftarrow \arg\min_{P_i} \mathcal{L}_{\mathrm{S1}}$ ▷ Structural intervention
$Q_{i,r} \leftarrow 1 - P_{i,r}^*$ and compute node contributions $C_{i,v}$
$F_i \leftarrow \{v : C_{i,v} \geq \tau\}$ ▷ Causal support discovery
Restore $S_i$ and construct $A_i \subseteq V_i | F_i$
$\mathcal{D}_i^{\mathrm{valid}} \leftarrow \varnothing$
**for** $t \in \{t_i^r\} \cup \mathcal{T}_i^c$ **do**
    Optimize $\Delta_{i,t}$ over $F_i$ ▷ $F$-only search
    **if** no strict counterfactual is found **then**
        Enable $A_i$ and optimize over $F_i \cup A_i$ with $\lambda_A > \lambda_F$ ▷ $F{+}A$ search
    **end if**
    Prune redundant nodes and re-optimize $\Delta_{i,t}^*$
    Evaluate $X_{i,t}^F$, $X_{i,t}^A$, and $X_{i,t}^{F+A}$
    Determine $\mathrm{Type}_{i,t}$ ▷ Post-hoc validation
    **if** $\mathrm{Valid}_{i,t} = 1$ **then**
        $\mathcal{D}_i^{\mathrm{valid}} \leftarrow \mathcal{D}_i^{\mathrm{valid}} \cup \{(X_i, S_i, s_i, X_{i,t}^{cf}, t)\}$ ▷ Validated CF
    **end if**
**end for**
$\mathcal{E}_i \leftarrow \{C_{i,v}, F_i, \mathrm{Type}_{i,t}, \mathrm{ICE}_{i,t}\}$
**return** $\mathcal{E}_i, \mathcal{D}_i^{\mathrm{valid}}$