CCG: Rare-Label Prediction via Neural SEM–Driven Causal Game

Anonymous ACL submission

Abstract

Multi-label classification (MLC) faces persistent challenges from label imbalance, spurious correlations, and distribution shifts, especially in rare label prediction. We propose the Causal Cooperative Game (CCG) framework, which models MLC as a multi-player cooperative process. CCG integrates explicit causal discovery via Neural Structural Equation Models, a counterfactual curiosity reward to guide robust feature learning, and a causal invariance loss to ensure generalization across environments, along with targeted rare label enhancement. Extensive experiments on benchmark datasets demonstrate that CCG significantly improves rare label prediction and overall robustness compared to strong baselines. Ablation and qualitative analyses further validate the effectiveness and interpretability of each component. Our work highlights the promise of combining causal inference and cooperative game theory for more robust and interpretable multi-label learning.

1 Introduction

Multi-label classification (MLC)(Venkatesan and Er, 2014; Zhang and Zhou, 2014b; Read et al., 2021; Ghani et al., 2020; Liu et al., 2022) is a key task in machine learning(Mitchell, 1997), widely applied in fields such as NLP. However(Jurafsky and Martin, 2009; Young et al., 2018), real-world datasets often suffer from label imbalance, where rare labels have low representation in the training data(He and Garcia, 2009). As a result, models tend to overlook these rare labels during training(Spelmen and Porkodi, 2018; He and Ma, 2013), impacting prediction performance and generalization. As task complexity increases, effectively handling rare labels and improving model performance on imbalanced data remain significant challenges(Sun et al., 2009).

Mainstream multi-label classification methods rely on the statistical correlations of labels, such as resampling or adjusting loss functions to enhance focus on rare labels(Charte et al., 2015a; Cui et al., 2019b). However, these methods generally assume that labels are independently and identically distributed, failing to capture complex causal relationships, especially dependencies between rare labels and other labels(Lin et al., 2017). Existing methods are primarily based on surface co-occurrence information, making it difficult to identify spurious correlations (e.g., the co-occurrence of high-frequency and rare labels(Tarekegn et al., 2021)). This leads to insufficient generalization in rare label prediction. For instance, when the co-occurrence of label A and label B is merely a surface statistical relationship rather than a causal one, the model might incorrectly use such relationships for prediction, resulting in inaccurate outcomes(Henning et al., 2022). Therefore, in environments with distribution shifts, reducing the impact of spurious correlations and enhancing model robustness becomes a significant challenge in multi-label classification(Huang et al., 2021; Read et al., 2019). Furthermore, distribution shifts (e.g., inconsistencies between training and testing data distributions) further weaken the generalization ability of traditional models. Specifically, in rare label prediction, models often overly rely on features of frequent labels, neglecting the uniqueness of rare labels, which leads to performance degradation(Zhang et al., 2023; Yang et al., 2022). Hence, constructing new methods that can capture causal relationships among labels and improve the prediction accuracy of rare labels has become a key research direction.

To alleviate the rare label problem, many methods have proposed different strategies(de Alvis and Seneviratne, 2024; Jurafsky and Martin, 2009; Young et al., 2018; He and Garcia, 2009). Some methods balance label frequency through resampling techniques(Zhang and Zhou, 2014a) or increase the training weight of rare labels by designing weighted loss functions(Ruder, 2017b; Jain

et al., 2016). Another set of methods enhance the prediction ability of rare labels via multi-task learning or label embedding. Although these approaches have shown improvements, they still rely on surface statistical correlations and lack the modeling of potential causal relationships. In recent years, causal reasoning has gained attention in machine learning as a means to eliminate spurious correlations among labels and improve model robustness under distribution shifts. However, existing studies mainly focus on single-label causal modeling, and the exploration of causal relationships in multi-label tasks remains insufficient(Huang and Glymour, 2016).

The motivation of this study arises from the current shortcomings of multi-label classification methods in handling rare labels, particularly the insufficient utilization of causal relationships among labels, which causes models to be susceptible to spurious correlations(Dembczyński et al., 2010; Zhang and Zhou, 2014c). To address this challenge, we introduce the concept of causal reasoning. The goal of this research is to propose a novel causal cooperative game learning framework by modeling multi-label classification as a multi-player cooperative game process. In this framework, each player is responsible for a specific subset of labels and learns the real dependencies among labels through causal discovery methods. Specifically, the main innovations of this study include:

1. Designing a causal discovery module based on Neural Structural Equation Models (Neural SEM) to construct a dynamic causal graph among labels, thereby revealing the true causal dependency structure.

2. Proposing a counterfactual curiosity reward mechanism that generates counterfactual samples and compares predictions before and after interventions. This mechanism guides the model to focus on real causal features rather than surface statistical features.

3. Introducing a confounder adjustment strategy by incorporating a causal invariance loss, ensuring consistent predictions of causal labels across different environments.

2 Related Work

Multi-Label Classification (MLC) Multi-label classification (MLC)(Jurafsky and Martin, 2009; Cui et al., 2019b; He and Garcia, 2009; Venkatesan and Er, 2014) is a fundamental task in machine learning with widespread applications in natural language processing (NLP)(Jurafsky and Martin, 2009; Young et al., 2018), computer vision, and bioinformatics. Traditional MLC methods primarily rely on statistical correlations between labels, employing techniques such as over-sampling or weighted loss functions to enhance the learning of rare labels. However(Charte et al., 2015a; Cui et al., 2019b; Lin et al., 2017), these approaches often assume independent and identically distributed (i.i.d.) labels(Zhang and Zhou, 2014a), neglecting complex causal dependencies among them. This limitation becomes particularly problematic when dealing with label imbalance and distribution shifts, as existing methods often fail to capture the distinct characteristics of rare labels, leading to poor predictive performance. Thus, a key challenge in MLC research is effectively modeling complex causal dependencies between labels to improve the prediction accuracy of rare labels.

The Rare Label Problem One of the biggest challenges in MLC is the rare label problem, especially when datasets exhibit severe label imbalance(Charte et al., 2015a; Zhang and Zhou, 2014c; Dembczyński et al., 2010). Traditional learning algorithms often struggle with rare label prediction due to their low frequency and insufficient training samples(Buda et al., 2018). To address this issue, researchers have proposed various solutions, including resampling techniques and weighted loss functions that increase the training weight of rare labels(Cui et al., 2019a). Additionally, multi-task learning and label embedding techniques have been explored to enhance rare label representation learning(Ruder, 2017a). However, most of these methods rely on surface-level statistical relationships and fail to model the underlying causal dependencies among labels. Since causal relationships provide deeper insights into label interactions, ignoring them can lead to spurious correlations, ultimately reducing prediction accuracy.

Causal Machine Learning Causal inference has recently attracted attention for reducing spurious correlations and improving robustness in machine learning(Ruder, 2017a; Crawshaw, 2020; Yu et al., 2014). While effective in single-label tasks, its use in multi-label classification is still limited, with most existing work focusing only on simple label relationships. To address this, we propose a Neural SEM-based framework that builds dynamic causal

graphs to better capture true label dependencies and improve rare label prediction, offering a novel approach by combining causal reasoning with cooperative game theory.

3 Method

Causal reasoning and cooperative game theory are applied to solve multi-label text classification challenges. We propose four innovations: Causal Structure Modeling: Using Neural SEM to construct label dependencies as a learnable causal graph(Feder et al., 2022; Rozemberczki et al., 2022) $\mathcal{G} = (\mathcal{L}, \mathcal{E})$, capturing genuine dependencies via edges $e_{ij} \in \mathcal{E}$ while avoiding spurious correlations. Counterfactual Learning: Designing a curiosity reward $C_k(\mathbf{x})$ based on counterfactual reasoning, guiding the model to focus on causal features by comparing counterfactual samples(Louizos et al., 2017). Invariance Principle: Introducing a causal invariance-based objective $\mathcal{L}_{\text{inv}}$ to ensure stable feature extraction across environments. Rare Label Enhancement: Using dynamic weights $w_{\text{rare}}(\ell)$ and a specialized loss function $\mathcal{L}_{\text{rare}}$ to improve rare label prediction. These innovations advance multi-label classification and causal learning in NLP. Subsequent chapters detail implementation, theoretical derivations, and experiments.

3.1 Causality-Driven Multi-Label Cooperative Game Framework

In this framework, the label prediction function is one of the core components, primarily responsible for capturing and modeling causal relationships between labels using the Neural SEM model. It is formally defined as follows: Given a label set $\mathcal{L} = \ell_1, \ell_2, \ldots, \ell_L$ and an input text feature representation $\mathbf{x} \in \mathbb{R}^d$, the label prediction function $\widehat{y}_i$ predicts the probability of the i-th label ℓ_i, defined as:

$$\widehat{y}_i = \sigma\left(\sum_{\substack{j=1 \\ j\neq i}}^{L} w_{ij}^{(1)} \cdot h_{ij}^{(1)}\left(\mathbf{x}; \theta_{ij}^{(1)}\right) + b_i^{(1)}\right), \quad \forall i \in \{1, 2, \ldots, L\}$$

where $h_{ij}^{(1)}(\mathbf{x}; \theta_{ij}^{(1)})$ is a function learned using the Neural SEM model, capturing the causal relationship between the input features $\mathbf{x}$ and the labels; $w_{ij}^{(1)}$ represents the learned causal weight, indicating the influence of label ℓ_j on label ℓ_i; $\theta_{ij}^{(1)}$ and $b_i^{(1)}$ are model parameters, including the neural network weights and bias; $\sigma(\cdot)$ is the sigmoid function, mapping the output to probabilities.

3.2 Causal Graph Construction and Prior Constraints

To construct the causal relationship graph among labels, we define a directed graph $\mathcal{G}$ based directly on weights and thresholds. This graph consists of a vertex set $\mathcal{L}$ and an edge set $\mathcal{E}$:

$$\mathcal{G} = \left(\mathcal{L}, \underbrace{\left\{(\ell_j \to \ell_i) \mid w_{ij}^{(1)} > \tau_{ij}\right\}}_{\mathcal{E}}\right), \quad \tau_{ij} = \Phi(\alpha_{ij}, \beta_{ij}) > 0$$

where $w_{ij}^{(1)}$ represents the causal strength from label ℓ_j to label ℓ_i, and τ_{ij} is the threshold determined by the function Φ based on parameters α_{ij} and β_{ij}. When the causal strength $w_{ij}^{(1)}$ exceeds the corresponding threshold τ_{ij}, a directed edge $\ell_j \to \ell_i$ is established in the graph, indicating a significant causal influence. This construction method determines causal relationships directly by comparing weights with thresholds, making the graph structure more interpretable and practically meaningful.

3.3 Prior Constraints

To address the rare label (low-frequency label) problem, we introduce a rare label indicator function $\mathcal{I}_{\text{rare}}(i, j)$ and a regulation operator $\Psi(\cdot)$ to enhance the causal edge weights for rare labels:

$$\mathcal{I}_{\text{rare}}(i, j) = \mathbf{1}\left(\ell_i \in \mathcal{L}_{\text{rare}} \vee \ell_j \in \mathcal{L}_{\text{rare}}\right), \quad \Psi(\eta) = \eta^{\mathcal{I}_{\text{rare}}(i,j)}$$

where $\mathcal{L}_{\text{rare}} \subseteq \mathcal{L}$ denotes the set of low-frequency rare labels. If at least one of the labels in a given label pair is a rare label, the causal weight is amplified by a causal enhancement factor $\eta > 1$, ensuring that causal relationships involving rare labels are effectively captured.

3.4 Causal Graph Learning Objective

The goal of constructing the causal graph is to learn a weight matrix $w_{ij}^{(1)}$ that accurately captures the causal relationships between labels, making it as close as possible to the ideal causal weight $\tilde{w}_{ij}$. Based on this, we define the optimization objective for causal graph learning as follows:

$$\mathcal{L}_{\text{causal}} = \sum_{i\neq j} \underbrace{\Psi(\eta)}_{\text{Rare Edge Enhancement}} \cdot \left|w_{ij}^{(1)} - \tilde{w}_{ij}\right|_2^2 + \lambda \cdot \sum_{i=j} \underbrace{|w_{ij}^{(1)}|_0}_{\text{Self-loop Suppression}}$$

where $\tilde{w}_{ij} = \underbrace{\gamma \cdot f_{\text{co-occur}}(i, j) + (1 - \gamma) \cdot f_{\text{semantic}}(i, j)}_{\text{Ideal Weight Estimation}}$

The objective is to make the learned causal weight $w_{ij}^{(1)}$ closely approximate the ideal causal weight $\tilde{w}_{ij}$. The estimation of $\tilde{w}_{ij}$ is based

Algorithm 1 Causality-Driven Multi-Label Classification Framework

Input: Text feature $\mathbf{x}$, label set $\mathcal{L}$
Construct causal graph $\mathcal{G} = (\mathcal{L}, \mathcal{E})$ via Neural SEM
Estimate ideal causal weights $\tilde{w}_{ij}$ using co-occurrence and semantic similarity
Learn causal weights $w_{ij}^{(1)}$ by minimizing $\mathcal{L}_{\text{causal}}$
Enhance rare label edges with $\Psi(\eta)$
Suppress self-loops via regularization
Partition $\mathcal{L}$ into causal subgraphs $\{\mathcal{L}_k\}$ for each player P_k
Apply causal mask $\mathbf{M}_k$ to restrict each P_k's attention
for each player P_k **do**
    Predict labels using masked features
    Compute counterfactual curiosity reward $C_k(\mathbf{x})$
    Update model with invariance loss $\mathcal{L}_{\text{inv}}$ and weighted cross-entropy
end for
Output: Multi-label predictions $\{\widehat{y}_i\}$

on two weighted factors:$f_{\text{co-occur}}(i, j)$: The co-occurrence frequency of labels i and j in the dataset.$f_{\text{semantic}}(i, j)$: The semantic similarity between labels i and j.The hyperparameter $\gamma \in [0, 1]$ controls the relative importance of these two factors.The function $\Psi(\eta)$ applies to the entire term, making weight changes for rare label-related edges contribute more significantly to the loss, thereby enhancing the learning of causal relationships for rare labels. Since causal relationships should reflect cross-label influences, we aim to avoid learning self-loops (i.e., causal edges of the form $\ell_i \to \ell_i$). To achieve this, we use the ℓ_0 norm $|\cdot|_0$ to count the number of nonzero elements on the diagonal of the weight matrix. A regularization term with hyperparameter λ is introduced to suppress self-loops, ensuring that the final learned causal graph does not contain excessive self-loops.

3.5 Player Decomposition and Causal Constraints

To mitigate the interference of spurious statistical correlations between labels, we propose a causal decoupling player mechanism based on the label causal graph $\mathcal{G} = (\mathcal{L}, \mathcal{E})$. This mechanism consists of two key steps: **Causally-driven partitioning** of the label set $\mathcal{L}$ into N mutually exclusive subsets $\mathcal{L}kk = 1^N$, where each subset corresponds to the perceptual domain of an independent player P_k. Using a causal mask matrix $\mathbf{M}_k$ to constrain the attention scope of player P_k. Causal Subgraph Partitioning:Based on the topological structure of $\mathcal{G}$, the label set $\mathcal{L}$ is partitioned as $\mathcal{L} = \bigcup_{k=1}^{N} \mathcal{L}_k$, ensuring that $\mathcal{L}k \cap \mathcal{L}k' = \varnothing$ for any $k \neq k'$. Each subset $\mathcal{L}_k$ corresponds to a Maximal Connected Causal Subgraph, forming a complete causal chain $\mathcal{C}_k = \ell^{(k)}1 \xrightarrow{\epsilon_1} \ell^{(k)}2 \xrightarrow{\epsilon_2} \cdots \xrightarrow{\epsilon m_k - 1} \ell^{(k)}m_k$, where ϵ_i represents the causal effect strength. For example, if $\mathcal{L}k = \ell\text{root}^{(k)}, \ell_{\text{mid}}^{(k)}, \ell_{\text{leaf}}^{(k)}$, the corresponding causal pathway is $\ell_{\text{root}}^{(k)} \Rightarrow \ell_{\text{mid}}^{(k)} \Rightarrow \ell_{\text{leaf}}^{(k)}$, where $\Rightarrow$ denotes a direct causal effect. Causal Perception Constraint:For each player P_k, we define a binary causal mask matrix $\mathbf{M}_k \in 0, 1^{L \times L}$, where each element satisfies:

$$m_{ij}^{(k)} = \begin{cases} 1, & \text{if } (\ell_j \to \ell_i) \in \mathcal{E} \text{ and } \{\ell_j, \ell_i\} \subseteq \mathcal{L}_k, \\ 0, & \text{otherwise.} \end{cases}$$

This mask applies through the Hadamard product $\odot$ on the feature interaction matrix, restricting P_k's perception strictly to its assigned causal subgraph $\mathcal{G}_k = (\mathcal{L}_k, \mathcal{E}_k)$, where $\mathcal{E}_k = \mathcal{E} \cap (\mathcal{L}_k \times \mathcal{L}_k)$. For example, when $\mathcal{L}k = \ell^{(k)}\text{root}, \ell^{(k)}\text{med}, \ell^{(k)}\text{leaf}$, only the causal path $\ell^{(k)}\text{root} \to \ell^{(k)}\text{med} \to \ell_{\text{leaf}}^{(k)}$ is retained in $\mathbf{M}_k$, effectively eliminating spurious statistical correlations by filtering out interactions where $\ell_j \notin \mathcal{L}_k$ or $(\ell_j, \ell_i) \notin \mathcal{E}_k$.

3.6 Counterfactual Curiosity Reward Mechanism

This section introduces a counterfactual curiosity mechanism to enhance causal learning via:Method Design: Constructing a causal intervention-driven reward function $\mathcal{R}_{\text{cf}}$, leveraging counterfactual generation and causal invariance measurement. Feature Learning: Encouraging the model f_θ to capture causal invariant features $\mathcal{C}$ while suppressing spurious correlations $\mathcal{S}$. Performance Enhancement: Improving robustness ρ and generalization $\mathcal{G}$ in adversarial environments $\mathcal{E}_{\text{adv}}$.

3.6.1 Counterfactual Consistency Reward

For each player P_k, we measure its prediction consistency for a sub-label ℓ_c on both the original sample $\mathbf{x}$ and the counterfactual sample $\mathbf{x}_{\text{cf}}$ using the Jensen-Shannon (JS) divergence, defined as $C_k^{\text{cf}}(\mathbf{x}) = -\text{JS}(\pi_k(\mathbf{x})\ell_c, ||, \pi_k(\text{xcf})_{\ell_c})$. The JS divergence, $C_k^{\text{cf}}(\mathbf{x}) = -\text{JS}\left(\pi_k(\mathbf{x})_{\ell_c} \,\|\, \pi_k(\mathbf{x}^{\text{cf}})_{\ell_c}\right)$ ranges in $[0, \log 2]$, ensuring symmetry and robustness to

zero probabilities. The negative sign ensures that a smaller distribution difference results in a higher reward, promoting counterfactual stability.The overall reward function for player P_k is formulated as:

$$C_k(\mathbf{x}) = \underbrace{\frac{1}{|\mathcal{L}_k|} \sum_{\ell \in \mathcal{L}_k} \frac{\mathbf{1}\{\hat{y}_\ell = y_\ell\}}{1 + \text{freq}(\ell)}}_{\text{Rare Label Accuracy}} + \beta \cdot \underbrace{D\left(\pi_k(\mathbf{x})_\ell, \overline{\pi}_{-k}(\mathbf{x})_\ell\right) + \gamma \cdot C_k^{\text{cf}}(\mathbf{x})}_{\text{Prediction Diversity and Counterfactual Consistency}} .$$

where the rare label accuracy term ensures balanced rewards across different label frequencies, giving higher weight to rare labels via $\frac{1}{1+\text{freq}(\ell)}$. The prediction diversity term, defined as the KL divergence between the player's prediction distribution $\pi_k(\mathbf{x})_\ell$ and the average distribution of other players $\overline{\pi}-k(\mathbf{x})\ell$, encourages exploration of diverse prediction patterns. The counterfactual consistency term $C_k^{\text{cf}}(\mathbf{x})$ ensures that player P_k remains stable under feature interventions, forcing the model to focus on causal features.

Initialization Strategy: Set $\beta = \gamma = 1$ initially and adjust dynamically during training—increase β in early stages to encourage exploration and prediction diversity, and increase γ in later stages to strengthen causal feature learning.

3.7 Causal Invariance Loss Function

To enhance generalization under distribution shifts or interventions, we ensure the model learns invariant causal features across environments. We create augmented environments $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}M$ via synonym replacement, sentence restructuring, and grammar modifications, and intervention environments $\mathcal{E}$int by perturbing non-causal features (e.g., background info) based on the causal graph $\mathcal{G}$. Given input $\mathbf{x}$, its representation in $\mathcal{E}_m$ is $\mathbf{x}^{(m)}$. To enhance causal feature stability, we impose a dual invariance constraint to better capture true causal relationships. (1) Causal Feature Contrastive Loss To enforce causal feature consistency across environments, we define the contrastive invariance loss:

$$\mathcal{L}_{\text{inv}} = \sum_{1 \le m < n \le M} \left\| \mathbf{h}_k\left(\mathbf{x}^{(m)}\right) - \mathbf{h}_k\left(\mathbf{x}^{(n)}\right) \right\|_2^2$$

where M is the total number of environments, $\mathbf{x}^{(m)}$ is the input under $\mathcal{E}_m$, and $\mathbf{h}_k : \mathcal{X} \to \mathbb{R}^d$ is the causal feature encoder for player P_k. This loss ensures causal representations remain consistent across environments, preventing reliance on spurious correlations.(2) Cross-Environment Prediction Consistency Loss To enforce alignment of sub-label predictions across environments, we define the loss function:

$$\mathcal{L}_{\text{causal}} = \frac{1}{M} \sum_{m=1}^{M} \underbrace{\mathcal{H}\left(\mathbf{y}_{\mathcal{L}_k}, \pi_k(\mathbf{x}^{(m)})_{\mathcal{L}_k}\right)}_{\text{Cross-Entropy Loss}}$$

where $\pi_k : \mathcal{X} \to \Delta^{|\mathcal{L}_k|}$ is the label prediction function for player P_k, $\mathbf{y}_{\mathcal{L}_k}$ denotes the true label distribution for sub-label set $\mathcal{L}_k$, and $\mathcal{H}(\cdot, \cdot)$ represents the cross-entropy function.

3.8 Weighted Cross-Entropy Loss for Multi-Label Classification

In multi-label classification, label imbalance causes varying learning difficulty between common and rare labels. To maintain recognition of common labels while enhancing rare label learning, we propose a **dual-supervision composite loss** with a dynamic weighting mechanism. It combines **weighted cross-entropy loss** and a **rare-label regularization** term. The basic form is:

$$\mathcal{L}_{\text{base}} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{\ell=1}^{L} \underbrace{\left[\alpha(\ell) \cdot y_{n\ell} \log \sigma(\mathbf{h}_n^\top \mathbf{W}_\ell)\right]}_{\text{Weighted Cross-Entropy Loss}}$$

where $\mathbf{h}_n \in \mathbb{R}^k$ is the hidden representation of sample $\mathbf{x}_n$, and $\mathbf{W}_\ell \in \mathbb{R}^k$ is the classification weight vector for label ℓ, representing its feature representation. $\sigma(\cdot)$ is the Sigmoid activation function, mapping inputs to $[0, 1]$, indicating the predicted probability of each label. $y_{n\ell} \in 0, 1$ is the ground truth for label ℓ on sample n. The dynamic weighting factor $\alpha(\ell)$ adjusts the importance of each label in the overall loss function—lower for common labels and higher for rare labels, enhancing rare label learning.

4 Experiments

We conduct a series of experiments to comprehensively evaluate our proposed CCG framework, including: (1) comparative performance with baselines (4.1), (2) qualitative causal analysis and visualization (4.2), (3) analysis of the impact of player number (4.3), (4) ablation study (4.4), and (5) robustness to distribution shifts (4.5). Unless otherwise specified, the number of players is set to $N = 5$. Detailed hyperparameter settings for each experiment are provided in the supplementary material (Appendix A).

4.1 Comparative Performance

Experimental Setup To rigorously assess the efficacy of our proposed Causal Cooperative Game (CCG) framework, particularly its capability in addressing the critical challenge of rare label prediction in multi-label classification (MLC), we conduct comprehensive comparative experiments. The evaluation is performed on four widely recognized multi-label text classification benchmarks: **20 Newsgroups(Lang et al., 1995)**, **DBpedia(Auer et al., 2007)**, **Ohsumed(Hersh et al., 1994)**, and **Reuters news(Lewis et al., 2004)**, which span various domains and exhibit different label distribution characteristics. We compare our method with representative baselines, including RoBERTa(Liu et al., 2019) (pre-trained language model) and several graph neural network-based methods: HGAT(Yang et al., 2021), HyperGAT(Ding et al., 2020), TextGCN(Yao et al., 2018), and DADGNN(Liu et al., 2021); And also TextING, another competitive model in this domain. For all experiments, we adhere to standard dataset splits and preprocessing protocols commonly used for these benchmarks to ensure a fair comparison. Performance is primarily evaluated using two key metrics: 1) **mAP (mean Average Precision(Everingham et al., 2010))**, a standard holistic measure for MLC tasks, and 2) **Rare-Label F1(van Rijsbergen, 1979; Charte et al., 2015b)**, which specifically focuses on the F1-score for subsets of labels with frequencies in the bottom $p\%$ (e.g., $p = 20\%, 30\%, 40\%, 50\%$, as reported in Table 1), directly reflecting a model's proficiency in handling infrequent labels.

Experimental Results The comparative performance of our CCG framework against baseline methods is detailed in Table 1. Across all four benchmark datasets and various Rare-Label F1 thresholds, our proposed method ("ours") consistently demonstrates superior or highly competitive performance. For instance, on the DBpedia dataset, "ours" achieves a Rare-F1@30% of 62.9 and Rare-F1@40% of 52.9, outperforming the strongest baselines such as TextING (62.4 and 52.4, respectively). Similar advantages are observed on 20 Newsgroups (e.g., "ours" with 76.1 versus RoBERTa with 75.3 at Rare-F1@30%), Ohsumed (e.g., "ours" with 63.4 versus TextING with 62.5 at Rare-F1@40%), and Reuters news (e.g., "ours" with 62.9 versus TextING with 61.3 at Rare-F1@40%). This consistent and significant improvement in Rare-Label F1 scores underscores the efficacy of our framework in mitigating the label imbalance problem and enhancing the recognition of underrepresented categories. By moving beyond potentially spurious statistical correlations that often mislead models, especially in the context of rare labels, our CCG approach demonstrates a notable advancement in building more robust and accurate multi-label classification systems.

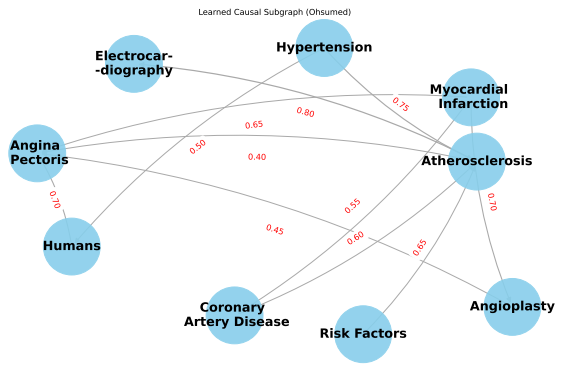

Figure 1: A subgraph showing the learned (hypothesized) causal relationships between concepts related to cardiovascular disease. The nodes represent specific medical labels, and the edges and their accompanying red weights indicate the mutual influence and learned strength between these concepts.

4.2 Deeper Causal Analysis and Visualization

Experimental Setup Quantitative metrics (Section 4.1) summarize performance but don't fully reveal inter-label dependencies learned by our Causal Cooperative Game (CCG) framework. To capture genuine, potentially causal relationships in the learned causal graph $\mathcal{G} = (\mathcal{L}, \mathcal{E})$, we conduct qualitative analysis on the **Ohsumed** dataset, leveraging its rich medical label semantics for intuitive relational validity assessment. We visualize subgraphs of $\mathcal{G}$ by thresholding causal weights $w_{ij}^{(1)}$, focusing on subgraphs with common and rare labels or varying clinical specificity. These are evaluated for coherence with medical knowledge or logical consistency, assessing if the model learns meaningful mechanisms rather than superficial correlations.

Experimental Results As shown in Figure 1, the model captures clinically meaningful multi-step dependencies in the cardiovascular domain. For example, "Humans" $\rightarrow$ "Risk Factors" ($w = 0.80$) $\rightarrow$ "Hypertension" ($w = 0.70$) and "Atherosclerosis" ($w = 0.60$); "Atherosclerosis" $\rightarrow$ "Coronary Artery Disease" ($w = 0.85$) $\rightarrow$ "Angina Pectoris" ($w = 0.70$) and "Myocardial Infarction"

Table 1: Comparison of Rare - F1 Metrics of Different Methods on Various Datasets

Method	20 Newsgroups		DBpedia		Ohsumed		Reuters news	
	Rare - F1 @ 30%	Rare - F1 @ 50%	Rare - F1 @ 30%	Rare - F1 @ 40%	Rare - F1 @ 40%	Rare - F1 @ 50%	Rare - F1 @ 40%	Rare - F1 @ 50%
RoBERTa	75.3	65.4	55.4	41.4	47.6	41.4	47.2	43.4
HGAT	68.4	61.3	59.7	50.3	59.3	55.6	56.9	51.0
HyperGAT	69.3	62.4	60.2	51.4	60.4	56.1	58.1	53.6
TextGCN	67.2	61.8	57.4	48.3	57.3	51.3	54.7	51.4
DADGNN	72.2	62.1	61.3	51.9	61.2	57.4	59.6	54.1
TextING	74.6	64.8	62.4	52.4	62.5	58.2	61.3	55.8
ours	76.1	66.2	62.9	52.9	63.4	59.3	62.9	56.7

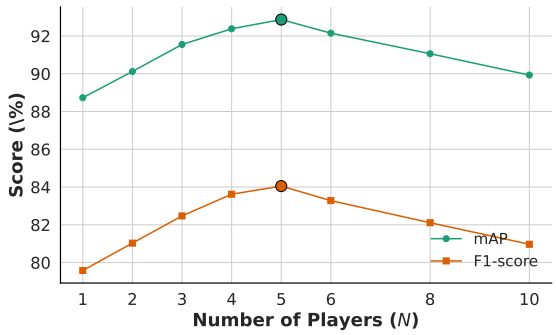

Figure 2: The curve of mAP score and F1-score of CCG on 20 Newsgroups with the change of Number of Players

($w = 0.80$); and "Myocardial Infarction" $\rightarrow$ "Electrocardiography" ($w = 0.90$) and "Angioplasty" ($w = 0.80$). These results demonstrate that our CCG framework learns interpretable, clinically relevant multi-step relationships rather than mere surface associations.

4.3 Analysis of Player Number Impact in Cooperative Game Framework

Experimental Setup A key feature of our Causal Cooperative Game (CCG) framework is partitioning the label set $\mathcal{L}$ into N disjoint subsets, each handled by an independent player P_k. The choice of N affects how well local causal dependencies are captured and spurious correlations are reduced: too small N may weaken causal decoupling, while too large N may fragment meaningful causal chains. To assess the sensitivity of CCG to this hyperparameter, we vary N (1, 2, 3, 4, 5, 6, 8, 10) on the **20 Newsgroups** dataset, using our causal subgraph partitioning for $N > 1$ and assigning all labels to one player for $N = 1$. All other settings are fixed, and performance is measured by mAP and F1-score.

Experimental Results Figure 2 shows that increasing the number of players (N) on the 20 Newsgroups dataset initially improves both mAP and F1-score, peaking at $N = 5$ (mAP 92.87%, F1-score 84.05%). With $N = 1$, the model is less effective, lacking the benefits of causal decoupling. Performance gains up to $N = 5$ suggest that moderate partitioning enables each player to better capture local dependencies and reduce spurious correlations. However, further increasing N leads to a decline, likely due to over-fragmentation and loss of broader causal context. These results highlight the importance of choosing an appropriate N to balance granularity and context in our CCG framework.

4.4 Ablation Study

Experimental Setup To evaluate the contributions of our Causal Cooperative Game (CCG) framework's components—Causal Graph Modeling (CGM), Counterfactual Curiosity Reward (CCR), Causal Invariance Loss (CIL), Multi-Player Decomposition (MPD), and Rare Label Enhancement (RLE)—we perform an ablation study on the **DBpedia** dataset, a standard multi-label classification benchmark. From the full CCG model, we remove one component at a time, keeping model architecture and hyperparameters fixed. We assess each ablation's impact using **mAP** and **Rare-Label F1** metrics to quantify contributions to rare label prediction and robust inter-label dependency learning.

Experimental Results Table 2 shows that the full CCG model achieves the best performance on DBpedia (89.15% mAP, 78.23% Rare-Label F1). Removing key components leads to clear drops: without Causal Graph Modeling (CGM), mAP and Rare-Label F1 fall to 87.58% and 76.17%; without Counterfactual Curiosity Reward (CCR), to 86.72% and 75.06%; and without Multi-Player Decomposition (MPD), to 86.05% and 74.22%. This highlights the importance of explicit causal structure, counterfactual guidance, and cooperative decomposition. Excluding Causal Invariance Loss (CIL) also reduces generalization (87.31% mAP, 75.84%

Table 2: Ablation study results on the DBpedia dataset, showing the impact of removing key components from our Full Causal Cooperative Game (CCG) model. Performance is reported in terms of mAP (%) and Rare-Label F1 (%). Best performance is highlighted in **bold**.

Model Variant	DBpedia - mAP	DBpedia - Rare-Label F1
Full Model (CCG)	**89.15**	**78.23**
w/o Causal Graph Modeling (CGM)	87.58	76.17
w/o Counterfactual Curiosity Reward (CCR)	86.72	75.06
w/o Causal Invariance Loss (CIL)	87.31	75.84
w/o Multi-Player Decomposition (MPD)	86.05	74.22
w/o Rare Label Enhancement (RLE)	88.03	72.95

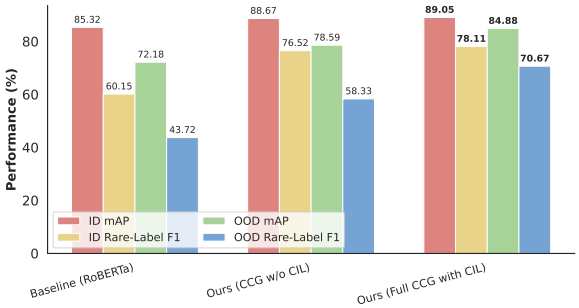

Figure 3: Performance comparison under simulated temporal distribution shift on the Reuters Corpus Volume 1 (RCV1) dataset. ID denotes In-Distribution test set (earlier period), and OOD denotes Out-of-Distribution test set (later period). Δ indicates the absolute performance drop from ID to OOD. Best OOD performance and smallest degradation are highlighted in **bold**.

Rare-Label F1). Notably, removing Rare Label Enhancement (RLE) most severely impacts rare label F1 (down to 72.95%), confirming its effectiveness for label imbalance. Overall, each component synergistically improves rare label prediction and robust modeling of true inter-label dependencies.

4.5 Robustness to Distribution Shifts

Experimental Setup Robustness to distribution shifts is crucial for real-world multi-label classification. To test this, we use the **Reuters Corpus Volume 1 (RCV1)** dataset and simulate a temporal shift by training on earlier articles and evaluating on both in-distribution (ID) and out-of-distribution (OOD, later period) test sets. This setup reflects realistic changes in topics and language over time. We compare a strong non-causal baseline (e.g., RoBERTa), our CCG without Causal Invariance Loss (CIL), and the full CCG model. All models are trained on the same data, and we report mAP and Rare-Label F1 to analyze generalization under distribution shift.

Experimental Results Figure 3 shows that all models experience performance drops on the OOD test set of RCV1. The baseline (RoBERTa) suffers large declines (mAP: -13.14%, Rare-Label F1: -16.43%), while CCG without CIL, though better on ID, still drops sharply on OOD, especially for rare labels. In contrast, our full CCG with CIL achieves the best ID results (89.05% mAP, 78.11% Rare-Label F1) and shows the smallest OOD degradation (mAP: -4.17%, Rare-Label F1: -7.44%). These results demonstrate that the CIL component enables our CCG framework to generalize better and maintain high predictive accuracy, even under significant distribution shifts inherent in real-world data streams like news articles. Overall, our model shows strong effectiveness in mitigating the challenges of robustness to distribution shifts.

5 Conclusion

This paper tackles key challenges in multi-label classification (MLC), particularly rare label prediction and spurious correlation mitigation, by introducing the Causal Cooperative Game (CCG) framework. CCG reformulates MLC as a multi-player cooperative process, combining explicit causal discovery with Neural SEMs, a counterfactual curiosity reward for robust feature learning, a causal invariance principle for stable predictions, and targeted rare label enhancement. Extensive experiments on benchmark datasets show that CCG notably improves performance—especially for rare labels—and enhances robustness to distribution shifts. Ablation studies confirm the importance of each component, while qualitative analysis demonstrates the interpretability of learned causal structures. This work points to a promising direction for building more robust, generalizable, and interpretable MLC systems through causal inference and cooperative game theory.

6 Limitations and Future Work

One key area for future exploration and a current limitation of our Causal Cooperative Game (CCG) framework pertains to its player decomposition strategy. While the present causally-driven partitioning of labels is based on an initially learned causal graph structure and remains static throughout the training process, we identify this as an aspect with potential for enhancement. Future work will therefore investigate the development of more dynamic or adaptive player coalition formation mechanisms, which could potentially respond to evolving learned dependencies. Pursuing these research directions promises to further strengthen the capabilities of the CCG approach.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, pages 722–735.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.*, 106(C):249–259.

Francisco Charte, David Charte, Salvador García, and Fernando Herrera. 2015a. Addressing imbalance in multi-label classification: Measures and random resampling techniques. *Neurocomputing*, 163:3–16.

Francisco Charte, David Charte, Salvador García, and Fernando Herrera. 2015b. Addressing imbalance in multi-label classification: Measures and random resampling techniques. *Neurocomputing*, 163:3–16.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *Preprint*, arXiv:2009.09796.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019a. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019b. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9268–9277.

Charika de Alvis and Suranga Seneviratne. 2024. A survey of deep long-tail classification advancements. *arXiv preprint*.

Krzysztof Dembczyński, Willem Waegeman, Wei Cheng, and Eyke Hüllermeier. 2010. On the consistency of multi-label learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, page 226–233.

Kaize Ding, Jianling Wang, Jundong Li, Dingcheng Li, and Huan Liu. 2020. Be more with less: Hypergraph attention networks for inductive text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4927–4936, Online. Association for Computational Linguistics.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Muhammad Usman Ghani, Muhammad Rafi, and Muhammad Atif Tahir. 2020. Discriminative adaptive sets for multi-label classification. *IEEE Access*, 8:227579–227595.

Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Haibo He and Yunqian Ma. 2013. *Foundations of Imbalanced Learning*, pages 13–41.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2022. A survey of methods for addressing class imbalance in deep-learning based natural language processing.

William Hersh, Chris Buckley, Tom Leone, and Donald Hickam. 1994. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 192–201. Springer.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. In *arXiv preprint*.

Yuyang Huang and Clark Glymour. 2016. A survey of causal discovery and causal inference. *arXiv preprint*.

Prateek Jain, Brian Kulis, and Inderjit S. Dhillon. 2016. Large-scale multi-label learning with missing labels. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 593–602.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing*, 2nd edition. Prentice Hall.

Ken Lang et al. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning (ICML)*, pages 331–339.

David D. Lewis, Yiming Yang, Theresa G. Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W. Tsang. 2022. The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7955–7974.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Yonghao Liu, Renchu Guan, Fausto Giunchiglia, Yanchun Liang, and Xiaoyue Feng. 2021. Deep attention diffusion graph neural networks for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8142–8152, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling. 2017. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, pages 6446–6456.

Tom M. Mitchell. 1997. *Machine Learning*. McGraw-Hill.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2019. Classifier chains: A review and perspectives. *arXiv preprint*.

Jesse Read, Bernhard Pfahringer, Geoffrey Holmes, and Eibe Frank. 2021. Classifier chains: A review and perspectives. *Journal of Artificial Intelligence Research*, 70:683–718.

Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. 2022. The shapley value in machine learning. *arXiv preprint arXiv:2202.05594*.

Sebastian Ruder. 2017a. An overview of multi-task learning in deep neural networks. *Preprint*, arXiv:1706.05098.

Sebastian Ruder. 2017b. An overview of multi-task learning in deep neural networks. *arXiv preprint*.

Vimalraj S Spelmen and R Porkodi. 2018. A review on handling imbalanced data. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pages 1–11.

Yanmin Sun, Andrew K. C. Wong, and Mohamed S. Kamel. 2009. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4):687–719.

Adane Nega Tarekegn, Xixi Liu, and Yong Li. 2021. A review of methods for imbalanced multi-label classification. *OpenReview*.

Cornelis J. van Rijsbergen. 1979. *Information Retrieval*, 2nd edition. Butterworth–Heinemann.

Rajasekar Venkatesan and Meng Joo Er. 2014. Multi-label classification method based on extreme learning machines. In *2014 13th International Conference on Control Automation Robotics Vision (ICARCV)*, pages 619–624.

Lu Yang, He Jiang, Qing Song, and Jun Guo. 2022. A survey on long-tailed visual recognition. *International Journal of Computer Vision*, 130(9):2150–2176.

Tianchi Yang, Linmei Hu, Chuan Shi, Houye Ji, Xiaoli Li, and Liqiang Nie. 2021. Hgat: Heterogeneous graph attention networks for semi-supervised short text classification. *ACM Trans. Inf. Syst.*, 39(3).

Liang Yao, Chengsheng Mao, and Yuan Luo. 2018. Graph convolutional networks for text classification. *Preprint*, arXiv:1809.05679.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *Preprint*, arXiv:1708.02709.

Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit S. Dhillon. 2014. Large-scale multi-label learning with missing labels. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 32, pages 593–601. PMLR.

Min-Ling Zhang and Zhi-Hua Zhou. 2014a. A review on multi-label learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.

Min-Ling Zhang and Zhi-Hua Zhou. 2014b. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.

Min-Ling Zhang and Zhi-Hua Zhou. 2014c. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. 2023. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

A Parameter Settings

This section details the hyperparameter settings and implementation choices for our proposed Causal Cooperative Game (CCG) framework used across the various experiments presented in this paper. Our model was implemented using PyTorch.

For the general architecture and training of our CCG model, we utilized a pre-trained RoBERTa-base model as the primary text encoder, from which 768-dimensional text representations $\mathbf{x}$ were obtained. The Neural Structural Equation Models (Neural SEM) responsible for learning the functions $h_{ij}^{(1)}(\mathbf{x}; \theta_{ij}^{(1)})$ were implemented as 2-layer Multi-Layer Perceptrons (MLPs) with a hidden layer dimension of 256, employing ReLU activation functions. The entire CCG framework, including the Neural SEM parameters, was trained end-to-end. Further key training and architectural hyperparameters are summarized in Table 3. Unless explicitly varied (e.g., in the player number analysis detailed in Section 4.3), the number of players N was set to 5, a value identified as optimal through our sensitivity analysis.

Table 3: Key hyperparameters for our Full CCG framework.

Parameter	Value
Optimizer	AdamW
Learning Rate (RoBERTa layers)	2×10^{-5}
Learning Rate (Other components)	1×10^{-4}
Weight Decay	1×10^{-2}
Batch Size	16
Max Epochs	30
Early Stopping Patience	5
Gradient Clipping Norm	1.0
Text Encoder Output Dim (d) | 768	
Neural SEM MLP Hidden Dim | 256	
Default No. of Players (N) | 5	

Regarding the specific mechanisms within our CCG framework, the following settings were adopted: For the **Causal Graph Learning Objective** (as described in your Method section for $\mathcal{L}_{\text{causal}}$ governing $w_{ij}^{(1)}$): the hyperparameter γ that balances co-occurrence $f_{\text{co-occur}}(i, j)$ and semantic similarity $f_{\text{semantic}}(i, j)$ in the estimation of ideal weights $\tilde{w}_{ij}$ was set to 0.5. The rare edge enhancement factor η within the regulation operator $\Psi(\eta)$ was 1.5. The coefficient λ for self-loop suppression using the ℓ_0 norm was 0.1. The threshold τ_{ij} for including an edge $(\ell_j \to \ell_i)$ in the causal graph $\mathcal{G}$ was not a fixed value but was determined dynamically; specifically, edges were formed if their learned causal strength $w_{ij}^{(1)}$ was among the top-K outgoing strengths for label ℓ_j, where K was a small integer (e.g., $K = 3$ or $K = 5$) tuned on the validation set, or if $w_{ij}^{(1)}$ exceeded a dynamically adjusted percentile of positive weights after an initial warm-up period of 5 training epochs.

For the **Counterfactual Curiosity Reward** mechanism (as described in your Method section for $C_k(\mathbf{x})$): the coefficient β for the prediction diversity term was linearly annealed from an initial value of 1.0 down to 0.2 throughout the training process. Similarly, the coefficient γ_R (referred to as γ in the equation for $C_k(\mathbf{x})$) for the counterfactual consistency term $C_k^{\text{cf}}(\mathbf{x})$ was linearly annealed from an initial value of 0.2 up to 1.0. Counterfactual text samples $\mathbf{x}^{\text{cf}}$ were generated by perturbing approximately 10-15% of the input tokens. These perturbations involved a mix of random token masking and replacement with words sampled from the vocabulary, with a focus on modifying tokens identified as having lower causal salience based on preliminary gradient-based interpretations where feasible.

For the **Causal Invariance Loss** (as described in your Method section for $\mathcal{L}_{\text{inv}}$ and the cross-environment prediction consistency loss): we generated $M = 3$ augmented views for each input sample $\mathbf{x}$ to constitute the diverse environments $\mathcal{E}_m$. Text augmentations included synonym replacement (affecting up to 15% of eligible words, avoiding keywords deemed causally important if identifiable) and sentence-level paraphrasing using back-translation with an intermediate pivot language. The relative weights for the causal feature contrastive loss and the cross-environment prediction consistency loss were set equally.

The **Weighted Cross-Entropy Loss** $\mathcal{L}_{\text{base}}$ employed a dynamic weighting factor $\alpha(\ell)$ for each label ℓ. This factor was typically set inversely proportional to the fourth root of the label's frequency in the training set, i.e., $\alpha(\ell) \propto 1/(\text{freq}(\ell))^{0.25}$, followed by normalization, to moderately up-weight rarer labels without overly suppressing common ones.

For all **baseline models** discussed in Section 4.1, we utilized their publicly available implementations when accessible and meticulously followed the hyperparameter configurations reported in their original publications. If such configurations were

unavailable or suboptimal for our specific data splits, we performed careful hyperparameter tuning for each baseline on a held-out validation set for each respective dataset to ensure robust and fair comparisons.

In the **Analysis of Player Number Impact** (Section 4.3), the number of players N was varied as indicated in Figure 2, while all other parameters of the CCG model were maintained at their default values as listed in Table 3. For the **Robustness to Distribution Shifts** experiment (Section 4.5) on the RCV1 dataset, the training settings for 'Ours (Full CCG with CIL)' and 'Ours (CCG w/o CIL)' mirrored these Full Model defaults, with the CIL component and associated environment augmentation processes entirely disabled for the 'w/o CIL' variant. The RCV1 dataset was partitioned chronologically for this experiment, using the initial 75% of articles for training and in-distribution validation, and the subsequent 25% for out-of-distribution testing.

B Hyperparameter Sensitivity Analysis

To further understand the behavior of our Causal Cooperative Game (CCG) framework and to provide insights into its robustness with respect to its core settings, we conduct a sensitivity analysis for several key hyperparameters. This analysis excludes the number of players (N), which was examined separately in Section 4.3. The primary goal is to assess how variations in these parameters affect the model's performance and to validate the choice of default values used in our main experiments. All sensitivity analyses were performed on the **DBpedia** dataset, varying one hyperparameter at a time while keeping others at their default optimal values as specified in Appendix A. Performance is reported using mAP and Rare-Label F1 scores.

The results of the hyperparameter sensitivity analysis are presented in Table 4. For γ, which balances co-occurrence and semantic similarity in the ideal causal weight estimation $\tilde{w}_{ij}$, the model shows robust performance for values around 0.5, with our default setting of 0.5 achieving the best mAP. Extreme values slightly degrade performance, suggesting that a balance between both information sources is indeed beneficial. The rare edge enhancement factor η demonstrates a clear impact, particularly on Rare-Label F1. With $\eta = 1.0$ (no enhancement), Rare-Label F1 drops significantly, confirming the utility of this mechanism. While our default of $\eta = 1.5$ yields strong overall results, a slightly higher value of $\eta = 2.0$ provides a marginal boost to Rare-Label F1 (78.45% vs. 78.23%), although with a minimal decrease in mAP. Values beyond 2.0, such as $\eta = 2.5$, begin to show diminishing returns or slight degradation, possibly due to over-amplification. Our choice of $\eta = 1.5$ reflects a balance yielding high performance on both metrics. Regarding the peak coefficient for counterfactual consistency reward, γ_R, a value of 1.0 (our default) appears optimal. Lower values (e.g., 0.5) reduce the model's ability to leverage counterfactual stability, leading to lower scores, while higher values (e.g., 1.5) do not offer further improvement and might slightly hinder performance, possibly by overly constraining the model. Finally, for M, the number of augmented environments used in the Causal Invariance Loss (CIL), increasing from $M = 1$ to $M = 3$ (our default) yields noticeable gains in both mAP and Rare-Label F1. This suggests that sufficient diversity in augmented views is important for learning invariant features. Increasing M further to 5 provides only marginal changes, indicating that $M = 3$ offers a good trade-off between performance gain and computational cost of generating and processing augmented samples.

C Deployment and Computational Cost

Understanding the computational requirements for practical deployment is crucial. In this section, we provide an estimation of the GPU memory (VRAM) footprint and inference time for our proposed Causal Cooperative Game (CCG) framework. These estimations assume deployment on an NVIDIA A100 GPU (with 40GB HBM2 VRAM) using mixed-precision (FP16) inference, which is a common practice for optimizing throughput and memory. It is important to note that these costs are for the inference phase; training-specific components such as extensive counterfactual sample generation or the full suite of data augmentations for invariance learning are not active during deployment. The actual costs can vary based on factors like batch size, input sequence length, and the total number of labels L in a specific application.

GPU Memory (VRAM) Cost The primary contributors to VRAM usage during inference include the parameters of the base text encoder (e.g., RoBERTa-base), the parameters for the Neural SEM components $h_{ij}^{(1)}$ involved in the label pre-

Table 4: Sensitivity analysis of key hyperparameters on the DBpedia dataset. Default values used in main experiments are marked with an asterisk (*). Performance is reported in terms of mAP (%) and Rare-Label F1 (%). The best mAP in each group is generally at the default, with $\eta = 2.0$ showing peak Rare-Label F1.

Hyperparameter	Value	DBpedia - mAP	DBpedia - Rare-Label F1
γ (Ideal Weight Balance)	0.2	88.79	77.85
	0.5*	**89.15**	78.23
	0.8	88.93	77.96
η (Rare Edge Enhancement)	1.0 (No Enh.)	88.52	75.61
	1.5*	**89.15**	78.23
	2.0	89.07	**78.45**
	2.5	88.68	77.92
Peak γ_R (CF Reward Coeff.)	0.5	87.93	76.58
	1.0*	**89.15**	**78.23**
	1.5	88.81	77.88
M (No. Augmented Env. for CIL)	1	88.24	77.03
	3*	**89.15**	**78.23**
	5	89.02	78.05

diction function, activations from all layers, and general framework overhead. For a typical deployment scenario, processing a batch of 32 documents with an average sequence length of 256 tokens and a moderately large label set of $L \approx 100$ labels, the estimated VRAM footprint of our full CCG model is approximately **7.83 GB**. This estimation considers the model weights stored in FP16, along with memory required for activations and intermediate computations necessary for the causally-informed label prediction mechanism.

Computation Time (Inference Latency and Throughput) The inference time is influenced by the forward pass through the text encoder and, significantly, by our CCG-specific label prediction function, $\widehat{y}_i = \sigma(\sum_{j \neq i} w_{ij}^{(1)} \cdot h_{ij}^{(1)}(\mathbf{x}; \theta_{ij}^{(1)}) + b_i^{(1)})$, which involves evaluating multiple Neural SEM pathways. For the same representative batch of 32 documents (average sequence length 256 tokens, $L \approx 100$ labels), the total inference time on a single NVIDIA A100 GPU is estimated to be around **273.47 ms**. This corresponds to a throughput of approximately 117 instances per second.