TOWARDS HIERARCHICAL INVARIANT SEMANTICS FOR INDUCTIVE KNOWLEDGE GRAPH REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Inductive Knowledge Graph Reasoning (IKGR) aims to infer missing facts for unseen entities given an observed Knowledge Graph (KG). However, most existing GNN-based methods suffer from limited generalization and poor extrapolation due to indiscriminate message passing mechanisms, which entangle task-relevant signals with spurious correlations confined to the training environment. In the inductive setting, such spurious patterns dilute task-relevant invariant semantics and significantly degrade test performance. To address this challenge, we propose **HiKGR**, a novel framework that models **H**ierarchical **i**nvariant semantics for inductive **K**nowledge **G**raph **R**easoning, thereby improving generalization and extrapolation. Specifically, we provide an analysis showing that the learned representation is prone to spurious associations without explicit invariant constraints. Then we design a multi-environment generator to simulate multiple environments with diverse structural variations, and improve extrapolation by exploring invariant semantic factors from hierarchical relation, entity, and subgraph perspectives. Further, we develop a bi-level optimization strategy to jointly encourage cross-environment invariance while preserving environment diversity. Extensive experiments on real-world datasets demonstrate that the proposed HiKGR substantially outperforms state-of-the-art methods.

1 INTRODUCTION

With the powerful representation capability, Knowledge Graphs (KGs) Cui et al. (2024) model entities and relationships as structured triples, enabling effective reasoning across various applications such as recommendation systems Wang et al. (2024a; 2019) and question answering Zhang et al. (2018); Wang et al. (2024b). Due to inherent incompleteness and potential noise during KG construction, early studies Kim et al. (2025); Li et al. (2024) primarily focus on inferring missing facts under the transductive setting, where all entities and relations are assumed to be observed during training. This assumption severely limits the ability to generalize to emerging entities. To bridge this gap, recent works Zhu et al. (2021); Liu et al. (2021) investigate Inductive Knowledge Graph Reasoning (IKGR), which aims to generalize reasoning to unseen entities.

Typically, existing IKGR methods can be roughly divided into two categories: rule-based models Xie et al. (2025); Liu et al. (2016) and message-passing-based models Zhang et al. (2024); Ding et al. (2025). Although the former facilitates inductive reasoning through rule pattern generalization, it often suffers from poor scalability and limited adaptability in complex scenarios. In contrast, the latter focuses on learning representations through local structural topology via Graph Neural Networks (GNNs) Kipf & Welling (2017), achieving desirable performance. Despite their advances, these approaches still exhibit fundamental limitations. As shown in Figure 1, message passing relies heavily on local structures and neighboring information, making it prone to capturing spurious label correlations present in the training environment. As

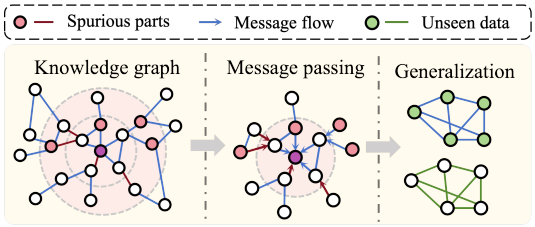


Figure 1: Illustration of most existing GNN-based IKGR methods. In message passing process, neighborhood aggregation indiscriminately captures spurious correlations tied to the training environment, leading to poor generalization to unseen data.

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.

a result, the learned representations tend to overfit to environment-specific patterns, significantly hindering their ability to generalize to unseen entities in the test environment.

Invariant Learning (IL) Arjovsky et al. (2019) offers a promising paradigm for mitigating spurious correlations by identifying invariant knowledge that remains stable across diverse environments. However, although it achieves great success in graph learning Yuan et al. (2025); Li et al. (2025), its application to IKGR remains underexplored. Unlike conventional settings, KGs exhibit richer relational dependencies and compositional semantics, making it non-trivial to directly apply existing IL principles. Moreover, although existing graph augmentation techniques Li et al. (2022); Wu et al. (2022) provide a potential way to construct multiple environments for exploring invariant knowledge in IL, naively perturbing KG structures may disrupt essential relational semantics and distort stable structural patterns, thereby causing GNNs to capture unstable signals that fail to generalize across unseen environments. Therefore, these observations naturally raise two key challenges:

1) How to construct diverse KG environments while preserving invariant knowledge?

2) How to extract invariant semantics across multiple environments for extrapolating to unseen data?

To address these challenges, we propose **HiKGR**, a unified reasoning framework that models hierarchical invariant semantics through multi-environment interaction. We provide an analysis showing that the learned representation is prone to spurious label correlations without explicit constraints, leading to degraded inference performance. Based on this insight, we jointly optimize representation sufficiency and invariance to identify task-relevant knowledge while mitigating spurious correlations. To facilitate invariant knowledge discovery, we design a multi-environment generator that constructs diverse KG environments with controlled structural perturbations. Further, considering that invariant knowledge is distributed across different structural granularities of the KG, we propose a hierarchical reasoning model aiming to capture invariant semantic factors from relation, entity and subgraph perspectives, thereby enhancing extrapolation ability. Finally, we develop a bi-level optimization strategy to jointly encourage cross-environment invariance while preserving environment diversity. The main contributions of this work are summarized as follows.

- We propose HiKGR, a novel GNN-based framework that integrates invariant learning into IKGR, providing theoretical insights for improving generalization and extrapolation.
- We design a multi-environment generator with a bi-level optimization strategy to construct diverse KG environments, enabling the model to learn cross-environment invariance.
- We develop a hierarchical reasoning mechanism to capture invariant semantic factors from relation, entity, and subgraph perspectives, facilitating knowledge inference for unseen entities.
- Extensive experiments on real-world datasets demonstrate that HiKGR consistently outperforms state-of-the-art methods on inductive knowledge graph completion tasks.

2 Related Work

Inductive knowledge graph reasoning. Inductive knowledge graph reasoning Teru et al. (2020); Mai et al. (2021); Huang et al. (2023) aims to generalize reasoning ability to unseen entities beyond the observed KG. Existing studies can be roughly grouped as follows. Rule-based methods Wang et al. (2025b); Qi et al. (2023); Du et al. (2026) facilitate inductive reasoning by extracting logical patterns. For instance, AMIE Galárraga et al. (2015) explores Horn rules from KGs, while Neural LP Yang et al. (2017) extends this idea by learning differentiable logical rules. Path-based approaches perform reasoning by exploring relational paths. For example, MINERVA Das et al. (2017) formulates path finding as a reinforcement learning problem, while NBFNet Zhu et al. (2021) generalizes path-based reasoning via differentiable dynamic programming. These methods exhibit strong inductive capability but suffer from poor scalability and high complexity in complex scenarios. Message passing-based methods Chen et al. (2022); Galkin et al. (2022); Shomer et al. (2025) perform inductive reasoning by aggregating structural information from local neighborhoods. For example, GraIL Teru et al. (2020) models relational reasoning over enclosing subgraphs, while RED-GNN Zhang & Yao (2022) and AdaProp Zhang et al. (2023b) further improve reasoning performance by refining message propagation mechanisms. Despite their advances, these approaches still suffer from limited extrapolation ability, as they tend to capture spurious label correlations induced from local structural patterns during training, hindering the extraction of task-relevant invariant semantics.

Invariant learning for inductive generalization. Invariant learning Creager et al. (2020); Arjovsky et al. (2019) has emerged as a promising paradigm for identifying stable knowledge across diverse environments, thereby improving generalization under distribution shifts. It has been successfully applied in image classification Creager et al. (2020); Kim et al. (2017) and recommendation systems Wang et al. (2022); Zhang et al. (2023a). Recently, invariant learning has been extended to graph-structured data. For instance, SIGN Yuan et al. (2025) improves robustness under distribution shifts by learning stable graph representations. VIVACE Li et al. (2025) disentangles invariant subgraphs to capture stable structural patterns. However, despite the success of IL in graph learning, its application to IKGR remains underexplored. Extending invariant learning to IKGR poses significant challenges due to the richer relational structures and compositional semantics of KGs, where structural variations may alter essential semantics and hinder the discovery of invariant knowledge.

3 Preliminary and Motivation

3.1 Preliminary

Inductive knowledge graph reasoning. We define a training KG as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ and $\mathcal{R}$ denote the sets of entities and relations, respectively, and $\mathcal{T} = \{(h, r, t)|h, t \in \mathcal{E}, r \in \mathcal{R}\}$ denotes the set of factual triplets, which is partitioned into training set $\mathcal{T}_{train}$ and validation set $\mathcal{T}_{valid}$, respectively. Besides, we define a query triplet set as $\mathcal{Q} = \{(h_q, r_q, t_q)|h_q, t_q \in \mathcal{E}', r_q \in \mathcal{R}\}$ with $\mathcal{E} \cap \mathcal{E}' = \varnothing$. We use $\mathbf{z} \in \mathbb{R}^d$ to represent any entity, $\mathbf{h} \in \mathbb{R}^d$ head entity, $\mathbf{t} \in \mathbb{R}^d$ tail entity, and $\mathbf{r}^{(l)} \in \mathbb{R}^d$ relation at each layer l, since relations exhibit layer-dependent semantics, as they capture local structural patterns at shallow layers and compositional reasoning at deeper layers. In the inductive setting, the goal of HiKGR is to train an optimal cross-environment shared predictor $F = f_{\boldsymbol{\theta}_f} \circ \Phi_{\boldsymbol{\theta}_g}$ based on $\mathcal{T}_{train}$ that can generalize to unseen triplets $\mathcal{Q}$, where $\Phi_{\boldsymbol{\theta}_g}(\cdot)$ and $f_{\boldsymbol{\theta}_f}(\cdot)$ denote a GNN encoder with L layers and linear predictor parameterized by $\boldsymbol{\theta}_g$ and $\boldsymbol{\theta}_f$, respectively.

To enhance the generalization and extrapolation of HiKGR, we aim to obtain an invariant sufficient representation given observed triplets.

Definition 1. *(Invariant sufficient representation across environments) Let $\mathbf{Z} = \Phi_{\boldsymbol{\theta}_g}(\mathcal{G})$ denote the representation learned from a KG $\mathcal{G}$. We say that $\mathbf{Z}$ is an invariant sufficient entity representation for predicting label Y across environments $\mathcal{S} = \{\mathcal{S}_k\}_{k=1}^K$ if it satisfies*

$$\mathcal{P}(Y \mid \mathcal{G}) = \mathcal{P}(Y \mid \mathbf{Z}) \text{ and } \mathcal{P}(Y \mid \mathbf{Z}, \mathcal{S}) = \mathcal{P}(Y \mid \mathbf{Z}), \tag{1}$$

where the first condition characterizes sufficiency and the second characterizes invariance.

Proposition 1. *(Optimization for sufficiency and invariance) To learn a sufficient and invariant representation $\mathbf{Z}$, we could solve the program $\max_{\boldsymbol{\theta}_g, \boldsymbol{\theta}_f} F(\boldsymbol{\theta}_g, \boldsymbol{\theta}_f)$, where*

$$F(\boldsymbol{\theta}_g, \boldsymbol{\theta}_f) = \mathcal{I}(\mathbf{Z}; Y) - \lambda \mathcal{I}(Y; \mathcal{S} \mid \mathbf{Z}), \tag{2}$$

Here, $\mathcal{I}(\mathbf{Z}; Y) = \mathcal{H}(Y) + \mathbb{E}_{\mathbf{Z},Y}[\log \mathcal{P}(Y \mid \mathbf{Z})]$ is maximized for sufficiency with $\mathcal{H}(Y)$ is constant value, while $\mathcal{I}(Y; \mathcal{S} \mid \mathbf{Z}) = \mathbb{E}_{\mathbf{Z},\mathcal{S}}[\mathrm{KL}(\mathcal{P}(Y \mid \mathbf{Z}, \mathcal{S}) \| \mathcal{P}(Y \mid \mathbf{Z}))]$ is minimized for invariance, where $\mathrm{KL}(\cdot)$ denotes KL divergence and $\mathcal{I}(\cdot)$ denotes the mutual information. Detailed derivation is given in Appendix A.1.

3.2 Motivation

We examine existing IKGR studies under a single observed environment $\mathcal{S}_k \in \mathcal{S}$, where the predictor F is trained by minimizing the training objective (i.e., binary cross-entropy loss) over the training setting. Inspired by Yao et al. (2025), given a KG $\mathcal{G}$, we consider a simplified setting, where invariant and spurious semantic factors are captured by $\Phi_{\boldsymbol{\theta}_g}(\cdot)$, and have the following finding.

Proposition 2. *(Spurious correlations) Let $\mathbf{Z} = \Phi_{\boldsymbol{\theta}_g}(\mathcal{G})$ denote the learned representation, which contains invariant component $\mathbf{Z}_{inv}$ satisfying Definition 1 and spurious component $\mathbf{Z}_{spu}$ that is predictive of the label but specific to the observed $\mathcal{S}_k$. Then minimizing the single-environment training objective $\mathcal{R}\left(f_{\boldsymbol{\theta}_f}(\mathbf{Z}), Y\right)$ yields an optimal predictor that depends on $\mathbf{Z}_{spu}$, denoted as*

$$f^*_{\boldsymbol{\theta}_f}(\mathbf{Z}_{inv}, \mathbf{Z}_{spu}) \not\equiv f^*_{\boldsymbol{\theta}_f}(\mathbf{Z}_{inv}). \tag{3}$$

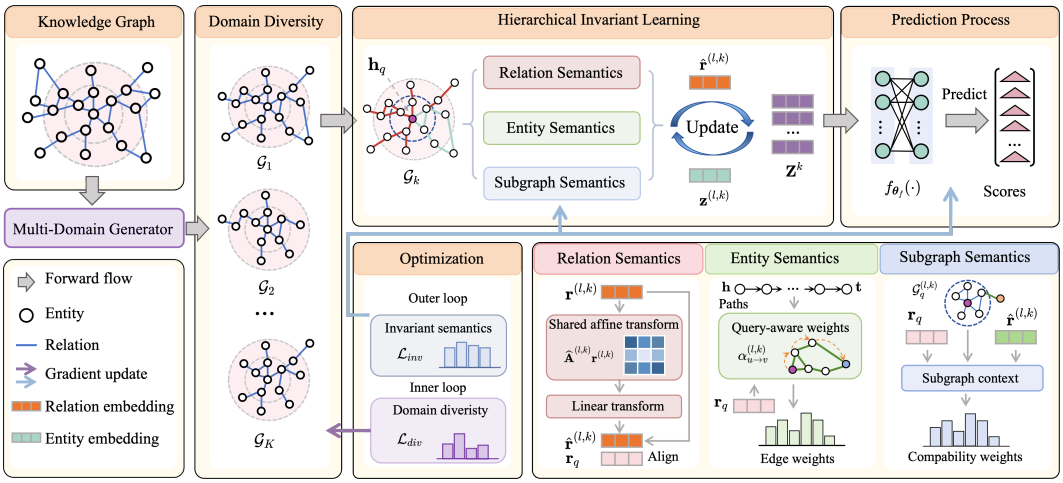


Figure 2: Overview of the proposed HiKGR for inductive knowledge graph reasoning.

Detailed analysis is provided in Appendix A.2.

Key insight. This result suggests that single-environment training favors all predictive signals, including unstable spurious correlations. Since these spurious signals are specific to $\mathcal{S}_k$, a predictor relying on them can achieve low training risk but poor extrapolation to unseen test set. In IKGR, such entanglement is further exacerbated by complex relational dependencies, severely degrading overall performance. Thus, inspired by Proposition 1, we explore hierarchical semantic invariant learning across diverse structural environments, which motivates the proposed HiKGR.

4 Proposed Method

Figure 2 illustrates the proposed HiKGR framework. The primary notations and pseudocode of HiKGR are provided in Appendix B.1 and B.5.

4.1 Multi-Environment Construction

To enable invariant learning, we construct diverse KG environments from the observed $\mathcal{G}$.

Definition 2. *(An Environment in HiKGR) An environment $\mathcal{S}_k$ contains a distribution $\mathcal{P}_{\mathcal{S}_k}(\mathcal{G}_k, Y)$ where $\mathcal{G}_k$ is generated from the observed $\mathcal{G}$ through controlled perturbation, satisfying two properties.*

- *Structural diversity: $\mathcal{S}_k$ exhibits a distinct structural distribution, providing diverse variations for identifying invariant semantics.*
- *Invariant preservation: $\mathcal{S}_k$ facilitates F in learning task-relevant semantics that remain consistent across structural distributions.*

Specifically, for each environment $\mathcal{S}_k$, we associate each triplet (h, r, t) with a specific mask $m^{(k)}_{h\to t}$, forming a mask matrix $\mathbf{M}^{(k)}$ for the observed $\mathcal{G}$. The multi-environment generator, implemented as a multi-layer perceptron $\Psi_{\boldsymbol{\theta}_k}(\cdot)$, produces a logit

$$w^{(k)}_{h\to t} = \Psi_{\boldsymbol{\theta}^k_e}\left(\left[\mathbf{h}^{(k)} \| \mathbf{r}^{(0,k)} \| \mathbf{t}^{(k)}\right]\right), \tag{4}$$

where $\cdot\|\cdot$ denotes concatenation. To enable differentiable sampling, we adopt the Gumbel-Sigmoid relaxation Maddison et al. (2016), denoted as

$$m^{(k)}_{h\to t} = \sigma\left(\left(\log u - \log(1-u) + w^{(k)}_{h\to t}\right)/\tau\right), \quad u \sim \mathcal{U}(0,1), \tag{5}$$

where τ controls the smoothness. Applying the mask matrices yields multiple KG environments $\{\mathcal{G}_k\}_{k=1}^{K}$. To support invariant preservation, we regularize mask density through the third term of Eq. (16), aiming to balance sparsification with the retention of task-relevant structural evidence. To induce structural diversity, we further relate diversity to empirical risk discrepancy as follows.

Proposition 3. *(Lower bound on environment variation) Given an observed $\mathcal{G}$, for any two generated environments $\mathcal{S}_k$ and $\mathcal{S}_{k'}$ and a predictor $F = f_{\boldsymbol{\theta}_f} \circ \Phi_{\boldsymbol{\theta}_g}$, there exists a constant $\Gamma > 0$, such that the representation variation is lower bounded by the discrepancy between the empirical risks, defined as*

$$\mathbb{E}_i \left[\left\| \mathbf{z}_i^{(k)} - \mathbf{z}_i^{(k')} \right\|_2 \right] \geq \Gamma \left| \mathcal{R}_{\mathcal{S}_k} \left(f_{\boldsymbol{\theta}_f}(\mathbf{Z}^{(k)}) \right) - \mathcal{R}_{\mathcal{S}_{k'}} \left(f_{\boldsymbol{\theta}_f}(\mathbf{Z}^{(k')}) \right) \right|, \tag{6}$$

where $\mathbb{E}_i$ averages over the same labeled query-candidate pairs in both environments, $\mathbf{Z}^{(k)} = \Phi_{\boldsymbol{\theta}_g}(\mathcal{G}_k)$ with $\mathbf{z}_i^{(k)} \in \mathbf{Z}^{(k)}$, and $\Gamma = 1/\|\boldsymbol{\theta}_f\|_2$. We omit the label Y in $\mathcal{R}_{\mathcal{S}_k}(\cdot)$ for brevity.

This motivates maximizing the empirical risk variance to promote environment diversity. Besides, the bi-level optimization strategy in Section 4.3 further encourages F to learn task-relevant semantics that are invariant across environments. Detailed derivations are provided in Appendix A.3.

4.2 Hierarchical Invariant Semantic Learning

Recalling Section 3.2, we aim to learn invariant knowledge for cross-environment extrapolation. Considering that the invariant semantics in KGs are hierarchical and entangled across multiple granularities, we explore them from relation, entity, and subgraph perspectives. Appendix B.2 provides an intuitive example demonstrating the necessity of hierarchical modeling.

Invariant relation semantics. Capturing nuanced relation semantics is crucial for inference in IKGR tasks. For example, *located_in* and *born_in* both point to location-type entities. Semantically related observed relations, such as *located_in* and *born_in*, may benefit from similar transformation patterns. Thus, we learn a transformation matrix $\widehat{\mathbf{A}}^{(l,k)} \in \mathbb{R}^{d \times d}$ in the l-th layer of $\Phi_{\boldsymbol{\theta}_g}(\cdot)$, which is shared across all relations within $\mathcal{S}_k$, enabling sharing of similar relation semantics. For each GNN layer, relation embeddings are first initialized as learnable representations, represented as $\mathbf{r}^{(l,k)}$ due to layer-dependent semantics. Then we perform a linear transformation to capture common semantics while preserving relation-specific information, defined as

$$\widehat{\mathbf{r}}^{(l,k)} = \mathbf{r}^{(l,k)} + \mathrm{MLP}_{\mathbf{W}_r^{(l,k)}} \left(\mathbf{r}^{(l,k)} \widehat{\mathbf{A}}^{(l,k)} \right), \qquad \textit{s.t.} \ \ \widehat{\mathbf{A}}^{(l,k)} = \mathrm{Norm} \left(\frac{\mathbf{A}^{(l,k)} + \mathbf{A}^{(l,k)^T}}{2} \right), \tag{7}$$

where the residual part $\mathbf{r}^{(l,k)}$ preserves relation-specific information, $\mathbf{A}^{(l,k)} \in \mathbb{R}^{d \times d}$ is a learnable parameter matrix, $\mathrm{MLP}_{\mathbf{W}_r^{(l,k)}}(\cdot)$ denotes a multi-layer perceptron with parameters $\mathbf{W}_r^{(l,k)} \in \mathbb{R}^{d \times d}$, and $\mathrm{Norm}(\cdot)$ denotes row-wise ℓ_2 normalization.

To bridge relation semantics across environments, we minimize the deviation of each transformation from the mean transformation across environments, denoted as

$$\mathcal{L}_{rel} = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{A}^{(l,k)} - \bar{\mathbf{A}}^l\|_F^2, \tag{8}$$

where L denotes the layer number of GNN encoder, and $\bar{\mathbf{A}}^l = \frac{1}{K} \sum_{k=1}^{K} \mathbf{A}^{(l,k)}$ denotes the mean transformation matrix at the l-th layer across all environments.

Invariant entity semantics. Entities encode semantics through multi-hop relational paths. However, not all paths from $h \to t$ are equally informative, as spurious and noisy paths may introduce misleading signals.

Thus, we introduce a path-aware weighting mechanism that selectively emphasizes query-consistent paths while suppressing irrelevant ones. Given a head entity of query $(h_q, r_q, ?)$ and one path $h_q \xrightarrow{r_1} \cdots \xrightarrow{r_{l-1}} u \xrightarrow{r_l} v$ in $\mathcal{S}_k$, the edge weight of sub-path (r_l, v) at the l-th layer is computed by measuring relevance between the current sub-path context and the query relation, denoted as

$$\alpha_{u \to v}^{(l,k)} = \sigma \Big(\mathbf{w}_a^\top \mathrm{ReLU} \Big(\mathbf{W}_v \mathbf{z}_v^{(l,k)} + \mathbf{W}_{r_l} \widehat{\mathbf{r}}_1^{(l,k)} + \mathbf{W}_{r_q} \widehat{\mathbf{r}}_q^{(l,k)} \Big) \Big), \tag{9}$$

where v denotes the l-th order neighboring entity of h_q, $\mathbf{W}_v, \mathbf{W}_{r_l}, \mathbf{W}_{r_q} \in \mathbb{R}^{d_a \times d}$ are learnable parameters, $\mathbf{w}_a \in \mathbb{R}^{d_a}$ is a learnable vector, and $\sigma(\cdot)$ is the Sigmoid function. By recursively composing edge weights through GNN propagation, the model implicitly adaptively assigns higher importance to invariant paths of h_q that are semantically aligned with the query.

Since explicit path alignment is computationally prohibitive, we approximate the semantics of an entity by a weight distribution induced at the final layer. Specifically, we align these distributions across environments via KL divergence, defined as

$$\mathcal{L}_{ent} = \frac{1}{K(K-1)} \sum_{k \neq k'} \mathrm{KL}\left(\mathbf{p}_v^{(L,k)} \,\|\, \mathbf{p}_v^{(L,k')}\right), \text{ with } p_v^{(L,k)}(r) = \frac{\sum_{(u,r,v)\in\mathcal{G}_k} \alpha_{u\to v}^{(L,k)}}{\sum_{r'} \sum_{(u,r',v)\in\mathcal{G}_k} \alpha_{u\to v}^{(L,k)}}. \tag{10}$$

Intuitively, $\mathbf{p}_v^{(L,k)}$ denotes the different contributions of relations to entity v under $\mathcal{S}_k$, where $p_v^{(L,k)}(r)$ denotes the importance of relation r among all incoming relations to entity v.

Invariant subgraph semantics. Given a query $(h_q, r_q, ?)$, we define query-specific subgraph $\mathcal{G}_q^{(l-1,k)}$ in $\mathcal{S}_k$ as the union of all query-relevant paths within $l-1$ hops. It provides higher-order structural context for reasoning, capturing intrinsic interactions between the query and substructures. To construct subgraph context, we aggregate relation representations within the substructure by

$$\mathbf{c}_q^{(l-1,k)} = \sigma\left(\mathbf{W}_c \sum_{(u,r,v)\in\mathcal{G}_q^{(l-1,k)}} \alpha_{u\to v}^{(l'-1,k)} \widehat{\mathbf{r}}^{(l'-1,k)}\right), \text{ with } l' \in [l-1], \tag{11}$$

where $\sigma(\cdot)$ denotes the Tanh function, and $\mathbf{W}_c \in \mathbb{R}^{d\times d}$ is a learnable projection matrix. Based on the subgraph context, we compute a subgraph-aware gate of next propagation step, aiming to measure the compatibility between local relation and the current subgraph semantics, defined as

$$\beta_{u\to v}^{(l,k)} = \sigma\Big(\mathbf{w}_g^\top \Big(\Big[\widehat{\mathbf{r}}^{(l,k)} \,\|\, \widehat{\mathbf{r}}_q^{(l,k)} \,\|\, \mathbf{c}_q^{(l,k)}\Big]\Big)\Big), \tag{12}$$

where $\sigma(\cdot)$ denotes the Sigmoid function, $\mathbf{w}_g \in \mathbb{R}^{d_g}$ is a learnable vector.

To extract invariant subgraph semantics, we align the query-specific subgraph contexts at the final propagation layer across environments via cosine similarity, denoted as

$$\mathcal{L}_{sub} = \frac{1}{K(K-1)} \sum_{k \neq k'} \left(1 - \cos\left(\mathbf{c}_q^{(L,k)}, \mathbf{c}_q^{(L,k')}\right)\right). \tag{13}$$

After extracting semantics from three perspectives, we obtain the entity representation by aggregating neighborhood messages using an additive fusion of entity and subgraph semantics without suppressing either component, denoted as

$$\mathbf{z}_v^{(l+1,k)} = \sigma\left(\sum_{u\in N_k(v)} \left(\alpha_{u\to v}^{(l,k)} + \beta_{u\to v}^{(l,k)}\right) \mathbf{m}_{u\to v}^{(l,k)}\right), \text{ with } \mathbf{m}_{u\to v}^{(l,k)} = \mathbf{W}_h\left(\mathbf{z}_u^{(l,k)} + \widehat{\mathbf{r}}_{uv}^{(l,k)}\right), \tag{14}$$

where $\sigma(\cdot)$ is the ReLU function, $N_k(v)$ is the neighbors pointing to v, $\mathbf{m}_{u\to v}^{(l,k)}$ denotes the neighboring message of (u, r_{uv}), and $\mathbf{W}_h \in \mathbb{R}^{d\times d}$ is a learnable projection matrix. After L propagation layers, the final entity representation $\mathbf{z}_v^k = \mathbf{z}_v^{(L,k)}$ is passed to $f_{\boldsymbol{\theta}_f}(\cdot)$ for downstream task.

4.3 Overall Model Optimization

Model prediction. Given query $(h_q, r_q, ?)$, we implement the knowledge graph completion by predicting the matching score of each candidate entity v using a linear predictor $f_{\boldsymbol{\theta}_f}(\cdot)$, and then we employ the binary cross-entropy loss as follows

$$\mathcal{L}_{task} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{BCELoss}(\mathcal{G}_k) = -\frac{1}{K}\sum_{k=1}^{K}\left[y \log f_{\boldsymbol{\theta}_f}(\mathbf{z}_v^{(k)}) + (1-y)\log\left(1 - f_{\boldsymbol{\theta}_f}(\mathbf{z}_v^{(k)})\right)\right], \tag{15}$$

where $\boldsymbol{\theta}_f \in \mathbb{R}^d$ is a learnable vector, and $y \in \{0, 1\}$ denotes the label of the query.

Optimization. We design a bi-level optimization strategy to jointly encourage cross-environment invariance while preserving environment diversity. Motivated by Proposition 3, in the inner loop of training, we optimize the multi-environment generator while fixing the encoder and predictor to

increase environment diversity while controlling structural perturbations. Specifically, we maximize the variance of losses and masks for environment diversity and regularize the structural size for invariant preservation, defined as

$$\mathcal{L}_{div} = -\left[\mathrm{Var}\left(\{\mathrm{BCELoss}(\mathcal{G}_k)\}_{k=1}^{K}\right) + \lambda_{var}\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{M}^{(k)} - \overline{\mathbf{M}}\|_F^2 - \lambda_{sp}\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\|\mathbf{M}^{(k)}\|_1}{|\mathcal{T}_{\text{train}}|} - \eta\right)^2\right], \tag{16}$$

where $\mathrm{Var}(\cdot)$ denotes variance, η is the target mean mask value, and $\|\mathbf{M}^{(k)}\|_1$ denotes the sum of all nonnegative mask entries associated with training triplets. The first two terms promote structural diversity, whereas the last term constrains the sparsity around the budget η.

Guided by Proposition 1, in the outer loop, we optimize the GNN encoder and the predictor while fixing the multi-environment generator, encouraging invariant preservation and enforcing semantic consistency across environments for F by minimizing

$$\mathcal{L}_{inv} = \mathcal{L}_{task} + \lambda_{rel}\mathcal{L}_{rel} + \lambda_{ent}\mathcal{L}_{ent} + \lambda_{sub}\mathcal{L}_{sub}, \tag{17}$$

where $\mathcal{L}_{rel}$, $\mathcal{L}_{ent}$, and $\mathcal{L}_{sub}$ are tractable hierarchical regularizers that promote cross-environment semantic consistency. Detailed derivation is shown in Appendix B.3.

Proposition 4. *(Generalization bound of HiKGR) Let Q be an unseen test setting and $\mathcal{F}$ be the fixed finite class of predictors. Then, under the conditional coverage assumption of Eq.* (38)*, for a bounded loss $\ell \in [0,1]$, with probability at least $1-\delta$, for $F = f_{\boldsymbol{\theta}_f} \circ \Phi_{\boldsymbol{\theta}_g} \in \mathcal{F}$, we have the following risk bound*

$$\mathcal{R}_Q\left(f_{\boldsymbol{\theta}_f}(\mathbf{Z})\right) \le \widehat{\mathcal{R}}_{\mathcal{S}}\left(f_{\boldsymbol{\theta}_f}(\mathbf{Z})\right) + \sqrt{\frac{1}{2}\mathcal{I}(Y;\mathcal{S}\mid\mathbf{Z})} + \sqrt{\frac{\log|\mathcal{F}| + \log(2/\delta)}{2|\mathcal{T}|}} + \Delta_{\mathrm{cov}}(Q,\mathcal{S}), \tag{18}$$

where $\widehat{\mathcal{R}}_{\mathcal{S}}\left(f_{\boldsymbol{\theta}_f}(\mathbf{Z})\right) = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathcal{R}}_{\mathcal{S}_k}\left(f_{\boldsymbol{\theta}_f}(\mathbf{Z}^k)\right)$ denotes the average empirical risk of training setting, $|\mathcal{T}|$ denotes the number of i.i.d. labeled query-candidate pairs for training, and $\Delta_{\mathrm{cov}}(Q,\mathcal{S})$ measures the coverage gap between the test and training distributions.

Proposition 4 shows that HiKGR can tighten the generalization bound if reducing the empirical risk $\widehat{\mathcal{R}}_{\mathcal{S}}\left(f_{\boldsymbol{\theta}_f}(\mathbf{Z})\right)$ and invariant violation $\mathcal{I}(Y;\mathcal{S}\mid\mathbf{Z})$ via $\mathcal{L}_{inv}$ in Eq. (17) and the coverage gap $\Delta_{\mathrm{cov}}(Q,\mathcal{S})$ via $\mathcal{L}_{div}$ in Eq. 16. Detailed derivation is given in Appendix A.4.

5 Experiments

In this section, we conduct experiments to evaluate the proposed HiKGR, focusing on the following research questions. **RQ1.** How does HiKGR perform compared with SOTA methods in the inductive setting? **RQ2.** How does the diversity of environments influence the effectiveness of HiKGR? **RQ3.** How do the major components affect overall performance? **RQ4.** How do different hyperparameters in HiKGR affect the performance?

5.1 Experimental Setup

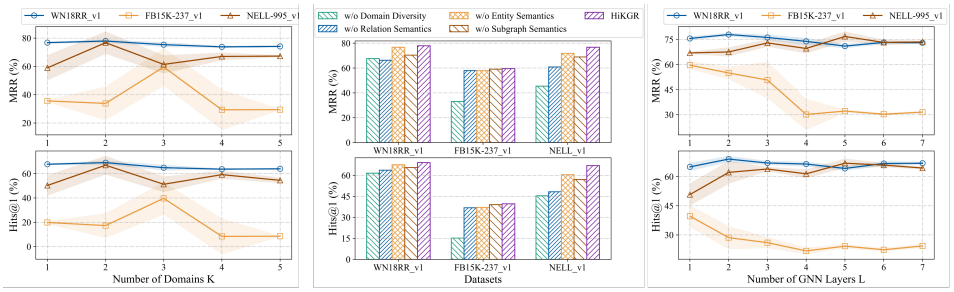
Inductive Datasets. We evaluate the extrapolation ability of HiKGR on unseen entities. We adopt three inductive benchmark datasets: *WN18RR* Dettmers et al. (2018), *FB15K-237* Toutanova & Chen (2015), and *NELL-995* Xiong et al. (2017). Following Teru et al. (2020), each dataset is split into four inductive settings with disjoint entity sets and progressively increasing graph size.

Inductive Baselines. We compare HiKGR with SOTA methods for inductive knowledge graph reasoning. We include rule-based methods *Neural LP* Yang et al. (2017) and *DRUM* Sadeghian et al. (2019), as well as message-passing-based methods *RED-GNN* Zhang & Yao (2022), *AdaProp* Zhang et al. (2023b), *Ingram* Lee et al. (2023), *KnowFormer* Liu et al. (2024), *Context-Pooling* Su et al. (2025), *HiMod* Xiong et al. (2026), and *CAKGE* Cao et al. (2026).

Evaluation metrics. Following the standard filtered setting Bordes et al. (2013), we adopt an ordinal tie-breaking policy that assigns distinct ranks to candidates with identical scores Cao et al. (2026), Xiong et al. (2026). We report Mean Reciprocal Rank (MRR) and Hits@t ($t = 1, 10$) to evaluate model performance. We run each experiment three times and report the mean performance on the test set. Detailed descriptions of experimental setup are given in Appendix C.1.

Table 1: Performance comparison (MRR, Hits@1 and Hits@10) of SOTA methods on three inductive benchmark datasets (**bold**: optimal; underline: sub-optimal).

Metrics	Methods	WN18RR				FB15K-237				NELL-995			
		v1	v2	v3	v4	v1	v2	v3	v4	v1	v2	v3	v4
MRR	Neural LP	0.649	0.635	0.361	0.628	0.325	0.389	0.400	0.396	0.610	0.361	0.367	0.261
	DRUM	0.666	0.646	0.380	0.627	0.333	0.395	0.402	0.410	0.628	0.365	0.375	0.273
	RED-GNN	0.701	0.690	0.427	0.651	0.369	0.447	0.445	0.442	0.637	0.419	0.430	0.363
	AdaProp	0.733	0.715	0.474	0.662	0.310	0.451	0.471	0.454	0.644	0.451	0.429	0.366
	Ingram	0.277	0.236	0.230	0.118	0.293	0.274	0.233	0.214	0.697	0.358	0.308	0.221
	KnowFormer	0.752	0.709	0.467	0.646	0.466	0.432	**0.494**	**0.480**	**0.827**	0.465	0.428	0.378
	Context-Pooling	0.708	0.694	0.425	0.648	0.383	0.463	0.447	0.427	0.634	0.430	0.426	0.382
	HiMod	0.773	0.743	0.497	0.734	0.456	0.453	0.458	0.458	0.732	0.428	0.438	0.380
	CAKGE	0.713	0.693	0.412	0.634	0.465	0.461	0.452	0.451	0.745	0.501	0.465	0.349
	HiKGR	**0.779**	**0.764**	**0.497**	**0.736**	**0.600**	**0.463**	0.479	0.473	0.768	**0.505**	**0.471**	**0.497**
Hits@1	Neural LP	0.592	0.575	0.304	0.583	0.243	0.286	0.309	0.289	0.500	0.249	0.267	0.137
	DRUM	0.613	0.595	0.330	0.586	0.247	0.284	0.308	0.309	0.500	0.271	0.262	0.163
	RED-GNN	0.653	0.633	0.368	0.606	0.302	**0.381**	0.351	0.340	0.525	0.319	0.345	0.259
	AdaProp	0.657	0.642	0.396	0.587	0.226	0.352	0.350	0.353	0.522	0.348	0.341	0.247
	Ingram	0.130	0.112	0.116	0.041	0.167	0.163	0.140	0.114	0.575	0.253	0.199	0.124
	KnowFormer	**0.695**	0.656	**0.406**	0.609	0.378	0.333	**0.400**	**0.383**	**0.730**	0.357	0.372	0.267
	Context-Pooling	0.660	0.637	0.365	0.603	0.316	0.368	0.358	0.330	0.517	0.329	0.324	0.278
	HiMod	0.676	0.650	0.362	0.639	0.312	0.307	0.367	0.347	0.631	0.326	0.349	0.264
	CAKGE	0.662	0.646	0.375	0.614	0.326	0.372	0.353	0.351	0.659	0.375	0.342	0.284
	HiKGR	**0.691**	**0.662**	0.367	**0.640**	**0.403**	0.359	0.369	0.369	0.669	**0.399**	**0.374**	**0.304**
Hits@10	Neural LP	0.772	0.749	0.476	0.706	0.468	0.586	0.571	0.593	0.871	0.564	0.576	0.539
	DRUM	0.777	0.747	0.477	0.720	0.474	0.595	0.571	0.593	0.873	0.540	0.577	0.531
	RED-GNN	0.799	0.780	0.599	0.721	0.483	0.629	0.603	0.621	0.866	0.601	0.594	0.556
	AdaProp	0.866	0.836	0.626	0.755	0.551	0.649	0.639	0.638	0.886	0.652	0.618	0.607
	Ingram	0.606	0.480	0.466	0.259	0.493	0.482	0.408	0.397	0.865	0.596	0.509	0.440
	KnowFormer	0.819	0.817	0.571	0.727	0.606	0.603	0.639	0.653	0.889	0.657	0.647	0.598
	Context-Pooling	0.791	0.783	0.522	0.721	0.487	0.629	0.592	0.598	0.804	0.593	0.624	0.546
	HiMod	0.929	0.923	0.780	0.952	0.669	0.650	0.612	0.640	0.888	0.646	0.609	0.633
	CAKGE	0.808	0.757	0.504	0.721	0.604	**0.697**	**0.641**	0.652	0.835	0.674	0.635	0.543
	HiKGR	**0.973**	**0.983**	**0.784**	**0.953**	**0.853**	0.651	**0.644**	**0.664**	**0.896**	**0.709**	**0.654**	**0.778**



(a) Performance of environment diversity (b) Performance of designed mechanisms (c) Performance of different GNN layers

Figure 3: Overall performance under (a) varying environment diversity K, (b) environment construction and hierarchical invariant semantics, and (c) different GNN layer numbers L.

5.2 Experimental Results

RQ1: Performance comparison. To evaluate the extrapolation ability of HiKGR on unseen entities, we compare it with SOTA methods, shown in Table 1. Overall, HiKGR achieves average improvements of 0.61%, 1.78%, and 3.79% on MRR, Hits@1, and Hits@10 over the second-best methods across all datasets, respectively, demonstrating its superior generalization ability. Compared to rule-based methods, HiKGR performs better by modeling complex relational interactions beyond predefined logical patterns. It also consistently outperforms message-passing-based methods, owing to its ability to mitigate spurious correlations and capture hierarchical invariant semantics. We observe larger gains on Hits@10, suggesting that HiKGR effectively improves the plausible coverage by capturing invariant semantic patterns across environments. We further provide performance comparison in the transductive setting, shown in Table 7 of Appendix C.2.

RQ2: Impact of environment diversity. To evaluate the impact of environment diversity, we vary the number of environments K on three datasets, shown in Figure 3 (a). We observe that the number of environments substantially affects the reasoning performance. For example, HiKGR shows substantial gains when increasing K to $[2, 3]$. This indicates that introducing moderate environment diversity helps expose structural variation, enabling the model to identify invariant semantics and reduce reliance on environment-specific correlations. However, the performance tends to decline when K becomes too large. This is because excessive environment variation may introduce noisy and redundant perturbations, which weaken the consistency of invariant patterns and hinder stable learning. Notably, the performance exhibits a more gradual improvement as K increases in NELL-995_v1. This is because the characteristic of sparsity and longer-path dependencies in NELL-995_v1 require additional environments to provide useful structural signals for capturing invariance.

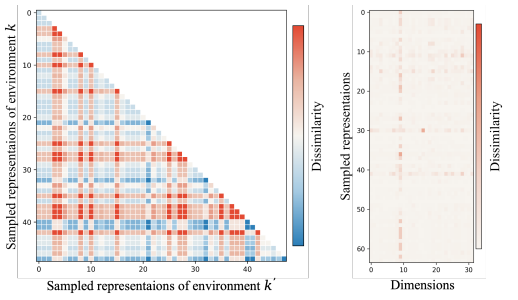
We further compare HiKGR with general graph augmentation methods for constructing multi-environments, including DropEdge Rong et al. (2019), GraphSAINT Zeng et al. (2019), and GDA4Rec Wang et al. (2025a). As shown in Table 2, HiKGR consistently outperforms these baselines, demonstrating that the proposed multi-environment construction is more effective than generic augmentations in generating informative environment variations for invariant semantic learning.

Table 2: Performance comparison between HiKGR and other graph augmentation methods on WN18RR_v1.

Methods	MRR	Hits@1	Hits@10
DropEdge	0.715	0.620	0.913
GraphSAINT	0.695	0.602	0.911
GDA4Rec	0.738	0.644	0.969
HiKGR	**0.779**	**0.691**	**0.973**

RQ3: Analysis of environment diversity and hierarchical semantics. To examine the role of environment construction and hierarchical invariant semantics in extrapolation, we conduct a rigorous ablation by removing both the corresponding objective functions and their associated modeling mechanisms, shown in Figure 3 (b). We observe a clear performance degradation without multi-environment construction, indicating its importance in introducing diverse structural variations for invariant learning. Besides, the removal of any invariant semantic level consistently leads to a performance drop, indicating that each level contributes essential invariant information, where relation level ensures stable semantic transformations, entity level filters query-inconsistent paths, and subgraph level further preserves contextually relevant structures. These components jointly enable robust extrapolation to unseen data, demonstrating the effectiveness of hierarchical invariant semantic learning.

We further provide a qualitative visualization of environment variations and learned representations on WN18RR_v1 in Figure 4. The heatmaps illustrate that the difference in graph structures across two environments, while the learned representations from distinct environments appear more aligned. This contrast indicates that HiKGR effectively captures invariant semantics under the moderately diverse environments, and this observation is further supported by Table 8. More visualizations are provided in Figure 6 of Appendix C.2.



(a) Visualization of environment diversity (b) Visualization of representation difference

Figure 4: Visualization of (a) interpolated heatmap of KGs across environments and (b) learned representations across environments in WN18RR_v1.

RQ4: Parameter sensitivity. We analyze the sensitivity of the number of GNN layers L on three datasets. As shown in Figure 3 (c), we observe that HiKGR achieves optimal performance at moderate depths ($L \in [1, 3]$) on most datasets, while NELL-995_v1 benefits from slightly deeper architectures due to its sparse structure and longer dependency paths. This is because NELL-995_v1 exhibits a sparser graph structure with longer dependency paths, requiring deeper message passing to capture useful invariant information. In contrast, moderate depths are sufficient for WN18RR_v1 and FB15K-237_v1, while performance degrades as L increases due to the introduction of noisy neighbors and spurious paths that hinder invariant semantic learning. Additional hyperparameter analysis is provided in Appendix C.2.

6 Conclusion

In this paper, we propose a novel reasoning framework that models hierarchical invariant semantics across hierarchical relation, entity and subgraph semantic levels to address the challenge of generalization and extrapolation for the IKGR tasks. First, we design a multi-environment generator to construct diverse environments, providing a foundation for invariant semantic learning. Second, we capture invariant semantic factors across multiple granularities, enabling the learning of invariant knowledge while suppressing environment-specific variations. Finally, a bi-level optimization strategy is introduced to jointly enforce cross-environment invariance and environment diversity, thereby improving generalization ability. Extensive experiments on real-world datasets demonstrate the effectiveness of HiKGR in generalizing to unseen entities under the inductive setting. This work advances invariant semantic modeling for inductive knowledge graph reasoning, providing valuable insights for future research in this area.

AI USE STATEMENT

In this work, we did not use generative AI tools for tasks with required disclosure; all required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for polishing and improving the writing quality of the manuscript, and for assisting with code development. We have reviewed all AI-assisted work. The AI-polished text was manually checked by 3 authors against the original drafts to ensure that the intended meaning, technical accuracy, and tone were preserved. AI-assisted code was verified and tested for correctness by 3 authors; we confirmed that all implemented methods match the descriptions in this paper and that the experimental results are reproducible. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Table 6 of Appendix C lists the key hyperparameters of HiKGR, and the complete experimental setup is described in Section 5.1 and Appendix C.1. Complete proofs and assumptions for our theoretical results are provided in Appendix A. A supplementary description of the method is provided in Appendix B. Code is available at `https://anonymous.4open.science/r/HiKGR`.

REFERENCES

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, 2019. URL http://arxiv.org/abs/1907.02893.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*, pp. 2787–2795, 2013.

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. Cakge: Context-aware adaptive learning for dynamic knowledge graph embeddings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. URL 10.1109/TPAMI.2026.3655896.

Mingyang Chen, Wen Zhang, Yushan Zhu, Hongting Zhou, Zonggang Yuan, Changliang Xu, and Huajun Chen. Meta-knowledge transfer for inductive knowledge graph embedding. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 927–937, 2022.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *Proceedings of the International Conference on Machine Learning*, pp. 2189–2200, 2020.

Yuanning Cui, Zequn Sun, and Wei Hu. A prompt-based knowledge graph foundation model for universal in-context reasoning. *Advances in Neural Information Processing Systems*, pp. 7095–7124, 2024.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. Go for a walk and arrive at the answer: reasoning over paths in knowledge bases using reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Syg-YfWCW.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1811–1818, 2018.

Ling Ding, Lei Huang, Zhizhi Yu, Di Jin, and Dongxiao He. Towards global-topology relation graph for inductive knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11581–11589, 2025.

Enjun Du, Siyi Liu, and Yongqi Zhang. Graphoracle: efficient fully-inductive knowledge graph reasoning via relation-dependency graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 19055–19063, 2026.

Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian M Suchanek. Fast rule mining in ontological knowledge bases with amie. *The VLDB Journal*, pp. 707–730, 2015.

Michael Galkin, Zhaocheng Zhu, Hongyu Ren, and Jian Tang. Inductive logical query answering in knowledge graphs. *Advances in Neural Information Processing Systems*, pp. 15230–15243, 2022.

Mikhail Galkin, Xinyu Yuan, Hesham Mostafa, Jian Tang, and Zhaocheng Zhu. Towards foundation models for knowledge graph reasoning. In *Proceedings of the International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jVEoydFOl9.

Xingyue Huang, Miguel Romero, Ismail Ceylan, and Pablo Barceló. A theory of link prediction via relational weisfeiler-leman on knowledge graphs. *Advances in Neural Information Processing Systems*, pp. 19714–19748, 2023.

Li Jin, Zezhong Ding, and Xike Xie. Duetgraph: Coarse-to-fine knowledge graph reasoning with dual-pathway global-local fusion. pp. 46077–46108, 2025.

Junsik Kim, Jinwook Park, and Kangil Kim. Rscf: Relation-semantics consistent filter for entity embedding of knowledge graph. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 12320–12336, 2025.

Seungryong Kim, Rui Cai, Kihong Park, Sunok Kim, and Kwanghoon Sohn. Modality-invariant image classification based on modality uniqueness and dictionary learning. *IEEE Transactions on Image Processing*, pp. 884–899, 2017.

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

Jaejun Lee, Chanyoung Chung, and Joyce Jiyoung Whang. Ingram: Inductive knowledge graph embedding via relation graphs. In *Proceedings of the International Conference on Machine Learning*, pp. 18796–18809, 2023.

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

Haoyang Li, Xin Wang, Xueling Zhu, Weigao Wen, and Wenwu Zhu. Disentangling invariant subgraph via variance contrastive estimation under distribution shifts. In *Proceedings of the International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/li25cv.html.

Rui Li, Chaozhuo Li, Yanming Shen, Zeyu Zhang, and Xu Chen. Generalizing knowledge graph embedding with universal orthogonal parameterization. In *Proceedings of the International Conference on Machine Learning*, pp. 28040–28059, 2024.

Junnan Liu, Qianren Mao, Weifeng Jiang, and Jianxin Li. Knowformer: Revisiting transformers for knowledge graph reasoning. In *Proceedings of the International Conference on Machine Learning*, pp. 31669–31690, 2024.

Qiao Liu, Liuyi Jiang, Minghao Han, Yao Liu, and Zhiguang Qin. Hierarchical random walk inference in knowledge graphs. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 445–454, 2016.

Shuwen Liu, Bernardo Grau, Ian Horrocks, and Egor Kostylev. Indigo: gnn-based inductive knowledge graph completion using pair-wise encoding. *Advances in Neural Information Processing Systems*, pp. 2034–2045, 2021.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: a continuous relaxation of discrete random variables. In *Proceedings of the International Conference on Learning Representations*, 2016. URL https://openreview.net/forum?id=S1jE5L5gl.

Sijie Mai, Shuangjia Zheng, Yuedong Yang, and Haifeng Hu. Communicative message passing for inductive relation reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 4294–4302, 2021.

Kunxun Qi, Jianfeng Du, and Hai Wan. Learning from both structural and textual knowledge for inductive knowledge graph completion. *Advances in Neural Information Processing Systems*, pp. 26546–26562, 2023.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: towards deep graph convolutional networks on node classification. In *Proceedings of the International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hkx1qkrKPr.

Ali Sadeghian, Mohammadreza Armandpour, Patrick Ding, and Daisy Zhe Wang. Drum: End-to-end differentiable rule mining on knowledge graphs. *Advances in Neural Information Processing Systems*, pp. 15321–15331, 2019.

Harry Shomer, Jay Revolinsky, and Jiliang Tang. Towards better benchmark datasets for inductive knowledge graph completion. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5777–5788, 2025.

Zhixiang Su, Di Wang, and Chunyan Miao. Context pooling: Query-specific graph pooling for generic inductive link prediction in knowledge graphs. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 2680–2689, 2025.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. In *Proceedings of the International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkgEQnRqYQ.

Komal K. Teru, Etienne G. Denis, and William L. Hamilton. Inductive relation prediction by subgraph reasoning. In *Proceedings of the International Conference on Machine Learning*, pp. 9448–9457, 2020.

Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pp. 57–66, 2015.

Shuyao Wang, Yongduo Sui, Chao Wang, and Hui Xiong. Unleashing the power of knowledge graph for recommendation via invariant learning. In *Proceedings of the ACM Web Conference*, pp. 3745–3755, 2024a.

Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 950–958, 2019.

Yansong Wang, Qihui Lin, Junjie Huang, and Tao Jia. Generative data augmentation in graph contrastive learning for recommendation. In Meeyoung Cha, Chanyoung Park, Noseong Park, Carl Yang, Senjuti Basu Roy, Jessie Li, Jaap Kamps, Kijung Shin, Bryan Hooi, and Lifang He (eds.), *Proceedings of the ACM International Conference on Information and Knowledge Management*, pp. 3155–3165, 2025a.

Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 19206–19214, 2024b.

Zhe Wang, Suxue Ma, Kewen Wang, and Zhiqiang Zhuang. Rule-guided graph neural networks for explainable knowledge graph reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12784–12791, 2025b.

Zimu Wang, Yue He, Jiashuo Liu, Wenchao Zou, Philip S. Yu, and Peng Cui. Invariant preference learning for general debiasing in recommendation. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1969–1978, 2022.

Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: an invariance perspective. In *Proceedings of the International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=FQOC5u-1egI.

Zhiwen Xie, Zhuo Zhao, Jinjin Ma, Guangyou Zhou, and Jimmy Xiangji Huang. Adarpt: an adaptive rule pattern transfer model for fully inductive knowledge graph reasoning. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1207–1217, 2025.

Chenyi Xiong, Jing Hu, Yue Zhao, Miao Zhang, Kui Xiao, Zhifang Huang, Dunhui Yu, Yan Zhang, and Zhifei Li. Himod: Hierarchical modeling with graph perturbation for enhanced inductive knowledge graph reasoning. *ACM Transactions on Knowledge Discovery from Data*, pp. 1–24, 2026.

Wenhan Xiong, Thien Hoang, and William Yang Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 564–573, 2017.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1412.6575.

Fan Yang, Zhilin Yang, and William W Cohen. Differentiable learning of logical rules for knowledge base reasoning. *Advances in Neural Information Processing Systems*, pp. 2319–2328, 2017.

Tianjun Yao, Haoxuan Li, Yongqiang Chen, Tongliang Liu, Le Song, Eric Xing, and Zhiqiang Shen. Pruning spurious subgraphs for graph out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 2025. URL https://doi.org/10.48550/arXiv.2506.05957.

Ruiwen Yuan, Yongqiang Tang, and Wensheng Zhang. A structure-aware invariant learning framework for node-level graph ood generalization. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1879–1890, 2025.

Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor K. Prasanna. Graphsaint: graph sampling based inductive learning method. In *Proceedings of the International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJe8pkHFwS.

An Zhang, Jingnan Zheng, Xiang Wang, Yancheng Yuan, and Tat-Seng Chua. Invariant collaborative filtering to popularity distribution shift. In *Proceedings of the ACM Web Conference*, pp. 1240–1251, 2023a.

Qinggang Zhang, Keyu Duan, Junnan Dong, Pai Zheng, and Xiao Huang. Logical reasoning with relation network for inductive knowledge graph completion. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4268–4277, 2024.

Yongqi Zhang and Quanming Yao. Knowledge graph reasoning with relational digraph. In *Proceedings of the ACM Web Conference*, pp. 912–924, 2022.

Yongqi Zhang, Zhanke Zhou, Quanming Yao, Xiaowen Chu, and Bo Han. Adaprop: learning adaptive propagation for graph neural network based knowledge graph reasoning. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3446–3457, 2023b.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander Smola, and Le Song. Variational reasoning for question answering with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. URL https://doi.org/10.1609/aaai.v32i1.12057.

Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal Xhonneux, and Jian Tang. Neural bellman-ford networks: a general graph neural network framework for link prediction. *Advances in Neural Information Processing Systems*, pp. 29476–29490, 2021.

A Details of Proofs

This section provides detailed proofs of the propositions proposed in the main text.

A.1 Proof of Proposition 1

Proof. For a sufficient representation, we can minimize the expected negative log likelihood of $\mathcal{P}(Y \mid \mathbf{Z})$ in Definition 1. Since $\mathcal{H}(Y)$ is constant with respect to $\boldsymbol{\theta}_g$ and $\boldsymbol{\theta}_f$, it holds that

$$\min_{\boldsymbol{\theta}_g, \boldsymbol{\theta}_f} -\mathbb{E}_{\mathbf{Z},Y}\left[\log \mathcal{P}(Y \mid \mathbf{Z})\right] \Leftrightarrow \max_{\boldsymbol{\theta}_g, \boldsymbol{\theta}_f} \mathcal{H}(Y) + \mathbb{E}_{\mathbf{Z},Y}\left[\log \mathcal{P}(Y \mid \mathbf{Z})\right], \text{ i.e., } \max_{\boldsymbol{\theta}_g, \boldsymbol{\theta}_f} \mathcal{I}(\mathbf{Z}; Y), \tag{19}$$

which aligns with the first term of Proposition 1.

For an invariant representation, since

$$\mathcal{I}(Y; S \mid Z) = \mathbb{E}_{Z,S}\left[\mathrm{KL}\big(\mathcal{P}(Y \mid Z, S) \| \mathcal{P}(Y \mid Z)\big)\right], \tag{20}$$

and the KL divergence is non-negative, $\mathcal{I}(Y; S \mid Z) = 0$ if and only if $\mathcal{P}(Y \mid Z, S) = \mathcal{P}(Y \mid Z)$ almost surely. Therefore, minimizing $\mathcal{I}(Y; S \mid Z)$ promotes the invariance condition in Definition 1.

Given a control coefficient λ, it aligns with the second term of Proposition 1. Therefore, we arrive at the optimization objective for invariant and sufficient representation in Eq. (2). □

A.2 Proof of Proposition 2

Proof. We consider the binary case where $Y \in \{0, 1\}$ and $\mathbf{Z}_{spu} \in \{0, 1\}$. Let the learned representation be $\mathbf{Z} = (\mathbf{Z}_{inv}, \mathbf{Z}_{spu})$.

The predictor outputs a logit score

$$s = f_{\boldsymbol{\theta}_f}(\mathbf{Z}_{inv}, \mathbf{Z}_{spu}), \tag{21}$$

and the prediction probability is given by

$$\widehat{Y} = \sigma(s), \tag{22}$$

where $\sigma(\cdot)$ is the Sigmoid function. Under the binary cross-entropy loss, the derivative with respect to the logit satisfies

$$\frac{\partial \mathcal{R}(s, y)}{\partial s} = \sigma(s) - y. \tag{23}$$

At the optimum, the prediction satisfies

$$\sigma\left(f^*_{\boldsymbol{\theta}_f}(\mathbf{Z}_{inv}, \mathbf{Z}_{spu})\right) = \mathcal{P}_{\mathcal{S}_k}(Y = 1 \mid \mathbf{Z}_{inv}, \mathbf{Z}_{spu}). \tag{24}$$

Since the sigmoid function is invertible, the optimal logit is

$$f^*_{\boldsymbol{\theta}_f}(\mathbf{Z}_{inv}, \mathbf{Z}_{spu}) = \log \frac{\mathcal{P}_{\mathcal{S}_k}(Y = 1 \mid \mathbf{Z}_{inv}, \mathbf{Z}_{spu})}{\mathcal{P}_{\mathcal{S}_k}(Y = 0 \mid \mathbf{Z}_{inv}, \mathbf{Z}_{spu})}. \tag{25}$$

By Bayes' rule, we have

$$\mathcal{P}_{\mathcal{S}_k}(Y = 1 \mid \mathbf{Z}_{inv}, \mathbf{Z}_{spu}) = \frac{\mathcal{P}_{\mathcal{S}_k}(\mathbf{Z}_{spu} \mid Y = 1, \mathbf{Z}_{inv})\mathcal{P}(Y = 1 \mid \mathbf{Z}_{inv})}{\sum_{y \in \{0,1\}} \mathcal{P}_{\mathcal{S}_k}(\mathbf{Z}_{spu} \mid Y = y, \mathbf{Z}_{inv})\mathcal{P}(Y = y \mid \mathbf{Z}_{inv})}. \tag{26}$$

Let

$$\pi(\mathbf{Z}_{inv}) = \mathcal{P}(Y = 1 \mid \mathbf{Z}_{inv}), \tag{27}$$

and since $\mathbf{Z}_{spu}$ is predictive of Y, we assume the spurious component satisfies

$$\mathcal{P}_{\mathcal{S}_k}(\mathbf{Z}_{spu} = Y \mid Y, \mathbf{Z}_{inv}) = q_k, \qquad q_k \neq \frac{1}{2}. \tag{28}$$

Then for $\mathbf{Z}_{spu} = 1$, we have

$$\mathcal{P}_{\mathcal{S}_k}(Y \mid \mathbf{Z}_{inv}, \mathbf{Z}_{spu} = 1) = \frac{q_k \pi(\mathbf{Z}_{inv})}{q_k \pi(\mathbf{Z}_{inv}) + (1 - q_k)(1 - \pi(\mathbf{Z}_{inv}))}. \tag{29}$$

For $\mathbf{Z}_{spu} = 0$, we have

$$\mathcal{P}_{\mathcal{S}_k}(Y = 1 \mid \mathbf{Z}_{inv}, \mathbf{Z}_{spu} = 0) = \frac{(1 - q_k)\pi(\mathbf{Z}_{inv})}{(1 - q_k)\pi(\mathbf{Z}_{inv}) + q_k(1 - \pi(\mathbf{Z}_{inv}))}. \tag{30}$$

Since $q_k \neq \frac{1}{2}$ and $0 < \pi(\mathbf{Z}_{inv}) < 1$, we have

$$\mathcal{P}_{\mathcal{S}_k}(Y = 1 \mid \mathbf{Z}_{inv}, \mathbf{Z}_{spu} = 1) \neq \mathcal{P}_{\mathcal{S}_k}(Y = 1 \mid \mathbf{Z}_{inv}, \mathbf{Z}_{spu} = 0). \tag{31}$$

Therefore,

$$\mathcal{P}_{\mathcal{S}_k}(Y \mid \mathbf{Z}_{inv}, \mathbf{Z}_{spu}) \neq \mathcal{P}_{\mathcal{S}_k}(Y \mid \mathbf{Z}_{inv}), \tag{32}$$

which implies

$$f^*_{\boldsymbol{\theta}_f}(\mathbf{Z}_{inv}, \mathbf{Z}_{spu}) \not\equiv f^*_{\boldsymbol{\theta}_f}(\mathbf{Z}_{inv}). \tag{33}$$

This completes the proof. □

A.3 Proof of Proposition 3

Proof. Let $\ell(s, y)$ denote the binary cross-entropy loss for logit s and label y. According to the Mean Value Theorem, for any two variables s_1 and s_2, there exists $s' \in (s_1, s_2)$ such that

$$\ell(s_1, y) - \ell(s_2, y) = \frac{\partial \ell(s', y)}{\partial s}(s_1 - s_2), \quad \exists s' \in (s_1, s_2). \tag{34}$$

Taking absolute values and noting that $|\sigma(s') - y| \leq 1$, we obtain

$$|\ell(s_1, y) - \ell(s_2, y)| \leq |s_1 - s_2|. \tag{35}$$

Substituting $s_1 = f_{\boldsymbol{\theta}_f}(\mathbf{z}_i^{(k)})$ and $s_2 = f_{\boldsymbol{\theta}_f}(\mathbf{z}_i^{(k')})$, we have

$$\begin{aligned}
\left|\mathcal{R}_{\mathcal{S}_k}\left(f_{\boldsymbol{\theta}_f}(\mathbf{Z}^{(k)})\right) - \mathcal{R}_{\mathcal{S}_{k'}}\left(f_{\boldsymbol{\theta}_f}(\mathbf{Z}^{(k')})\right)\right| &= \left|\mathbb{E}_i\left[\ell\left(f_{\boldsymbol{\theta}_f}(\mathbf{z}_i^{(k)}), y_i\right) - \ell\left(f_{\boldsymbol{\theta}_f}(\mathbf{z}_i^{(k')}), y_i\right)\right]\right| \\
&\leq \mathbb{E}_i\left[\left|\ell\left(f_{\boldsymbol{\theta}_f}(\mathbf{z}_i^{(k)}), y_i\right) - \ell\left(f_{\boldsymbol{\theta}_f}(\mathbf{z}_i^{(k')}), y_i\right)\right|\right] \\
&\leq \mathbb{E}_i\left[\left|f_{\boldsymbol{\theta}_f}(\mathbf{z}_i^{(k)}) - f_{\boldsymbol{\theta}_f}(\mathbf{z}_i^{(k')})\right|\right] \\
&= \mathbb{E}_i\left[\left|\boldsymbol{\theta}_f^\top\left(\mathbf{z}_i^{(k)} - \mathbf{z}_i^{(k')}\right)\right|\right] \\
&\overset{(1)}{\leq} \|\boldsymbol{\theta}_f\|_2 \mathbb{E}_i\left[\left\|\mathbf{z}_i^{(k)} - \mathbf{z}_i^{(k')}\right\|_2\right].
\end{aligned} \tag{36}$$

where $\mathbb{E}_i$ averages over the same labeled query-candidate pairs in both environments, $\mathbf{Z}^{(k)} = \Phi_{\boldsymbol{\theta}_g}(\mathcal{G}_k)$ and $\mathbf{Z}^{(k')} = \Phi_{\boldsymbol{\theta}_g}(\mathcal{G}_{k'})$ with $\mathbf{z}_i^{(k)} \in \mathbf{Z}^{(k)}$ and $\mathbf{z}_i^{(k')} \in \mathbf{Z}^{(k')}$. (1) follows from Cauchy-Schwarz Inequality, and $\Gamma = \frac{1}{\|\boldsymbol{\theta}_f\|_2}$.

This completes the proof. □

A.4 Proof of Proposition 4

Proof. Let $\mathcal{P}_{\mathcal{S}} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{P}_{\mathcal{S}_k}$ denote the mixture distribution over the generated environments. For a random environment variable $\mathcal{S}$, defined as

$$\mathcal{P}(Y \mid \mathbf{Z}, \mathcal{S}) = \sum_{k=1}^{K} \mathcal{P}(\mathcal{S} = \mathcal{S}_k \mid \mathbf{Z}) \mathcal{P}(Y \mid \mathbf{Z}, \mathcal{S} = \mathcal{S}_k). \tag{37}$$

For the learned representation induced by F, we assume that the conditional label shift in the unseen test environment is covered by the generated environments, defined as

$$\mathbb{E}_{\mathbf{Z}\sim\mathcal{P}_Q(\mathbf{Z})}\left[\mathrm{KL}\left(\mathcal{P}(Y \mid \mathbf{Z}, Q) \,\|\, \mathcal{P}(Y \mid \mathbf{Z}, \mathcal{S})\right)\right] \leq \mathcal{I}(Y; \mathcal{S} \mid \mathbf{Z}). \tag{38}$$

For theoretical analysis, let $\ell \in [0, 1]$ be a bounded surrogate loss. The training and test risks are defined as

$$\begin{aligned} \mathcal{R}_{\mathcal{S}}(f_{\boldsymbol{\theta}_f}(\mathbf{Z})) &= \mathbb{E}_{\mathbf{Z}\sim\mathcal{P}_{\mathcal{S}}(\mathbf{Z})}\mathbb{E}_{Y\sim\mathcal{P}(Y|\mathbf{Z},\mathcal{S})}\left[\ell(f_{\boldsymbol{\theta}_f}(\mathbf{Z}), Y)\right], \\ \mathcal{R}_{Q}(f_{\boldsymbol{\theta}_f}(\mathbf{Z})) &= \mathbb{E}_{\mathbf{Z}\sim\mathcal{P}_{Q}(\mathbf{Z})}\mathbb{E}_{Y\sim\mathcal{P}(Y|\mathbf{Z},Q)}\left[\ell(f_{\boldsymbol{\theta}_f}(\mathbf{Z}), Y)\right]. \end{aligned} \tag{39}$$

For simplicity, we denote $\mathcal{R}_{\mathcal{S}}(f_{\boldsymbol{\theta}_f}(\mathbf{Z}))$ as $\mathcal{R}_{\mathcal{S}}$ with bounded $[0, 1]$, and $g(\mathbf{Z}) = \mathbb{E}_{Y\sim\mathcal{P}(Y|\mathbf{Z},\mathcal{S})}\left[\ell(f_{\boldsymbol{\theta}_f}(\mathbf{Z}), Y)\right]$ as $g(\mathbf{Z})$.

We introduce the invariant oracle risk on the test distribution as

$$\mathcal{R}^{\mathrm{inv}}_{Q}(f_{\boldsymbol{\theta}_f}(\mathbf{Z})) = \mathbb{E}_{\mathbf{Z}\sim\mathcal{P}_{Q}(\mathbf{Z})}\mathbb{E}_{Y\sim\mathcal{P}(Y|\mathbf{Z},\mathcal{S})}\left[\ell(f_{\boldsymbol{\theta}_f}(\mathbf{Z}), Y)\right]. \tag{40}$$

By the triangle inequality, we have

$$|\mathcal{R}_Q - \mathcal{R}_{\mathcal{S}}| \leq \underbrace{\left|\mathcal{R}_Q - \mathcal{R}^{\mathrm{inv}}_Q\right|}_{(a)} + \underbrace{\left|\mathcal{R}^{\mathrm{inv}}_Q - \mathcal{R}_{\mathcal{S}}\right|}_{(b)}. \tag{41}$$

For term (a), Jensen's inequality and the variational characterization of total variation distance yield

$$(a) \leq \mathbb{E}_{\mathbf{Z}\sim\mathcal{P}_Q(\mathbf{Z})}\left[\mathrm{TV}\left(\mathcal{P}(Y \mid \mathbf{Z}, Q), \mathcal{P}(Y \mid \mathbf{Z}, \mathcal{S})\right)\right]. \tag{42}$$

By Pinsker's inequality, Jensen's inequality, and Eq. (38), we have

$$\begin{aligned} (a) &\leq \mathbb{E}_{\mathbf{Z}\sim\mathcal{P}_Q(\mathbf{Z})}\sqrt{\frac{1}{2}\mathrm{KL}\left(\mathcal{P}(Y \mid \mathbf{Z}, Q) \,\|\, \mathcal{P}(Y \mid \mathbf{Z}, \mathcal{S})\right)} \\ &\leq \sqrt{\frac{1}{2}\mathbb{E}_{\mathbf{Z}\sim\mathcal{P}_Q(\mathbf{Z})}\left[\mathrm{KL}\left(\mathcal{P}(Y \mid \mathbf{Z}, Q) \,\|\, \mathcal{P}(Y \mid \mathbf{Z}, \mathcal{S})\right)\right]} \\ &\leq \sqrt{\frac{1}{2}\mathcal{I}(Y; \mathcal{S} \mid \mathbf{Z})}. \end{aligned} \tag{43}$$

For term (b), by the variational characterization of total variation distance, we have

$$\begin{aligned} (b) &= \left|\mathbb{E}_{\mathbf{Z}\sim\mathcal{P}_Q(\mathbf{Z})}[g(\mathbf{Z})] - \mathbb{E}_{\mathbf{Z}\sim\mathcal{P}_{\mathcal{S}}(\mathbf{Z})}[g(\mathbf{Z})]\right| \\ &\leq \mathrm{TV}\left(\mathcal{P}_Q(\mathbf{Z}), \mathcal{P}_{\mathcal{S}}(\mathbf{Z})\right) \\ &= \Delta_{\mathrm{cov}}(Q, \mathcal{S}). \end{aligned} \tag{44}$$

Combining Eqs. (41), (43), and (44), we obtain

$$|\mathcal{R}_Q - \mathcal{R}_{\mathcal{S}}| \leq \sqrt{\frac{1}{2}\mathcal{I}(Y; \mathcal{S} \mid \mathbf{Z})} + \Delta_{\mathrm{cov}}(Q, \mathcal{S}). \tag{45}$$

Assume that $|\mathcal{T}|$ query-label samples are independently drawn from $\mathcal{P}_{\mathcal{S}}$ and that $\mathcal{F}$ is finite. By Hoeffding's inequality and the union bound, with probability at least $1 - \delta$, we derive

$$\left|\mathcal{R}_{\mathcal{S}} - \widehat{\mathcal{R}}_{\mathcal{S}}\right| \leq \sqrt{\frac{\log|\mathcal{F}| + \log(2/\delta)}{2|\mathcal{T}|}}. \tag{46}$$

Finally, combining Eqs. (45) and (46), we have

$$\mathcal{R}_Q \leq \widehat{\mathcal{R}}_{\mathcal{S}} + \sqrt{\tfrac{1}{2}\mathcal{I}(Y; \mathcal{S} \mid \mathbf{Z})} + \sqrt{\tfrac{\log|\mathcal{F}| + \log(2/\delta)}{2|\mathcal{T}|}} + \Delta_{\mathrm{cov}}(Q, \mathcal{S}). \tag{47}$$

This completes the proof. □

B Details of Method

This section provides explanations of primary notation, toy examples derivation of optimization, and pseudocode in Section 4.

B.1 Primary Notations

Table 3 lists the definitions of primary notations in HiKGR.

Table 3: Summary of primary notation of this paper.

Notations	Explanation	Notations	Explanation	Notations	Explanation
$\mathbf{z}$	Any entity embedding	$\mathbf{h}$	Head entity embedding	$\mathbf{t}$	Tail entity embedding
$\mathbf{r}^{(l,k)}$	Relational embedding	$\widehat{\mathbf{r}}^{(l,k)}$	Refined relational embedding	$\{\mathcal{S}_k\}_{k=1}^{K}$	environment set
$\{\mathcal{G}_k\}_{k=1}^{K}$	Knowledge graph set	$\{\mathbf{M}^{(k)}\}_{k=1}^{K}$	Mask set	$\Psi_{\boldsymbol{\theta}_k}(\cdot)$	Multi-environment generator
$\Phi_{\boldsymbol{\theta}_g}(\cdot)$	Cross-environment GNN encoder	$f_{\boldsymbol{\theta}_f}(\cdot)$	Linear predictor	$\widehat{\mathbf{A}}^{(l,k)}$	Specific transformation matrix
$\alpha_{u\to v}^{(l,k)}$	Edge weight of $u \to v$	$\beta_{u\to v}^{(l,k)}$	Compatibility weight of $u \to v$	$\mathbf{c}_q^{(l,k)}$	Subgraph context embedding

B.2 Toy Example of Hierarchical Invariant Semantic Learning

To illustrate the necessity of hierarchical invariant semantic learning, we consider a toy example within a single KG. As shown in Figure 5, given a query $(h_q, r_q, ?) = (\textit{Alice}, \textit{located_in}, ?)$, the goal is to infer the correct location *Paris*.

Invariant Relation Semantics. At the relation level, HiKGR learns that semantically similar relations tend to induce similar transformation patterns. For example, relations such as *come_from* and *lives_in* both capture location-related semantics, and thus guide the query relation *located_in* through similar transformations toward location-related entities. By aligning relation transformation patterns across environments, HiKGR enforces consistency among semantically similar relations, thereby capturing invariant relation semantics.

Invariant Entity Semantics. At the entity level, HiKGR encourages query-consistent paths to promote prediction while suppressing noisy or spurious paths. For example, compared with the direct path *Alice* $\xrightarrow{\textit{located_in}}$ *Paris*, the path *Alice* $\xrightarrow{\textit{friend_of}}$ *Bob* $\xrightarrow{\textit{work_at}}$ *CompanyX* $\xrightarrow{\textit{located_in}}$ *Paris* contributes less to the query prediction, since it is semantically inconsistent with the query and introduces spurious correlations that may not hold under test distribution. HiKGR addresses this by learning stable path patterns via aligning them across environments, and adaptively assigning higher importance to query-consistent paths while suppressing noisy or misleading ones.

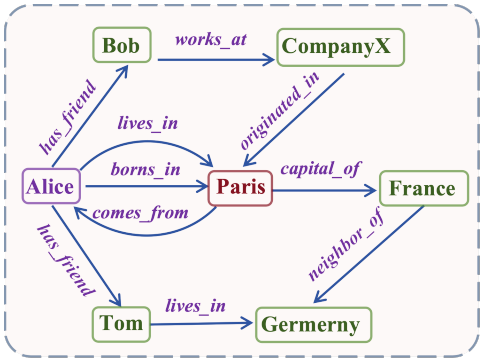


Figure 5: An example of KG.

Subgraph invariant semantics. At the subgraph level, the model progressively expands the query-centric subgraph during message passing. However, not all structural expansions are beneficial for reasoning. For example, a path such as *Paris* $\xrightarrow{\textit{capital_of}}$ *France* $\xrightarrow{\textit{neighbor_of}}$ *Germany* introduces additional entities that are structurally valid but irrelevant to the query. These inconsistencies can be resolved by considering the compatibility between the query and the evolving subgraph context across environments.

These three levels correspond to different sources of spurious correlations in KGs. Relation level captures stable relation transformation patterns across environments, entity level provides evidence for query-consistent paths, and subgraph level further suppresses structurally valid but contextually incompatible reasoning patterns, which necessitates hierarchical modeling to promote extrapolation.

B.3 Derivation of Optimization

Derivation of Eq. (17). According to Proposition 1, optimizing Eq. (2) leads to learning an invariant sufficient representation. We derive tractable objectives for the two mutual information terms accordingly.

For the first term, we introduce a variational lower bound, denoted as

$$\max_{\boldsymbol{\theta}_g,\boldsymbol{\theta}_f} \mathcal{I}(\mathbf{Z};Y) \overset{(1)}{\geq} \mathbb{E}_{\mathbf{Z},Y}\left[\log \mathcal{P}(Y \mid \mathbf{Z})\right] \overset{(2)}{\geq} \mathbb{E}_{\mathbf{Z},Y}\left[\log \widehat{\mathcal{P}}(Y \mid \mathbf{Z})\right], \tag{48}$$

where (1) follows from the definition of mutual information, and (2) introduces a variational distribution $\widehat{\mathcal{P}}(Y \mid \mathbf{Z})$ implemented by linear predictor $f_{\boldsymbol{\theta}_f}(\cdot)$ to approximate the intractable conditional distribution $\mathcal{P}(Y \mid \mathbf{Z})$. In practice, we optimize this lower bound via the binary cross-entropy loss $\mathcal{L}_{task}$ in Eq. (15).

Directly estimating the conditional mutual information $\mathcal{I}(Y;S \mid Z)$ is intractable in our setting. Therefore, instead of explicitly optimizing it, we introduce three hierarchical cross-environment consistency regularizers: $\mathcal{L}_{rel}$, $\mathcal{L}_{ent}$, and $\mathcal{L}_{sub}$.

Specifically, $\mathcal{L}_{rel}$ aligns relation transformation patterns, $\mathcal{L}_{ent}$ aligns path-induced entity semantics, and $\mathcal{L}_{sub}$ aligns query-specific subgraph contexts across the generated environments. These objectives provide tractable surrogates that promote cross-environment semantic consistency. They are not claimed to be exact estimators or upper bounds of $\mathcal{I}(Y;S \mid Z)$.

Combining the task objective with these hierarchical regularizers yields Eq. (17).

Derivation of Eq. (16). The discrete nature of KGs makes it difficult to directly estimate the differences between them. Thus, we assess the environment variation between two graphs by comparing their representations. Proposition 3 demonstrates that the variation in empirical risk serves as a lower bound for environment disparity. Based on this, we promote structural diversity by maximizing the variance of empirical and mask matrices across environments. Combining these two objectives leads to the formulation in Eq. (16).

B.4 Complexity Analysis

Time Complexity. Let $|\mathcal{E}|$, $|\mathcal{R}|$, $|\mathcal{T}|$, and $|\mathcal{Q}|$ denote the numbers of entities, relations, training triplets, and queries respectively. For each environment, the time complexity of HiKGR can be divided into two parts: (1) For the multi-environment generator, the environment generation costs $\mathcal{O}(|\mathcal{T}|d^2)$. (2) For hierarchical invariant semantic learning, the relation semantic level modeling costs $\mathcal{O}(L|\mathcal{T}|d^2)$, modeling the entity and subgraph semantic levels costs $\mathcal{O}(L|\mathcal{T}|dd_a)$, and the message aggregation process costs $\mathcal{O}(L|\mathcal{T}|d^2)$. In our implementation, the number of GNN layers L is relatively small, so the overall computational complexity is not significantly increased. Thus, the overall complexity is $\mathcal{O}(|\mathcal{T}|d^2 + |\mathcal{T}|dd_a)$.

Space Complexity. For each environment, the main memory cost comes from storing entity and relation representations and intermediate weights, denoted as $\mathcal{O}(|\mathcal{E}|d + |\mathcal{R}|d + |\mathcal{Q}|d + |\mathcal{T}|)$.

B.5 Pseudocode of HiKGR

Algorithm 1 details the pseudocode of the proposed HiKGR.

C Details of Experiments

This section provides basic information about datasets, SOTA methods, parameter settings, and additional experimental results.

C.1 Experimental Setup

Inductive Datasets. Table 4 details the information about the inductive benchmark datasets.

Inductive Methods.

Algorithm 1 Algorithm of the proposed HiKGR

Require: Training KG $\mathcal{G}$, query set $\mathcal{Q}$, GNN encoder $\Phi_{\boldsymbol{\theta}_g}(\cdot)$, linear predictor $f_{\boldsymbol{\theta}_f}(\cdot)$, multi-environment generator $\Psi_{\boldsymbol{\theta}_k}(\cdot)$

Ensure: Learned parameters $\boldsymbol{\theta}_g$, $\boldsymbol{\theta}_f$, $\{\boldsymbol{\theta}_e^k\}_{k=1}^K$

```
Initialize model parameters;
while not converged do
    # Inner loop: optimize multi-environment generator
    Generate K graphs {G_k ← Ψ_{θ_e^k}(G)}_{k=1}^K in Eq. (5);
    Update {θ_e^k}_{k=1}^K by minimizing L_div with θ_g, θ_f fixed in Eq. (16);
    # Outer loop: optimize GNN encoder and linear predictor
    for k = 1 to K do
        Encode relation semantics via transformation within each environment in Eq. (7);
        Compute sub-path weight and subgraph-level gate in Eqs. (9), (12);
        Perform message passing to obtain Z^k in Eq. (14);
    end for
    Update θ_g, θ_f by minimizing L_inv with fixed {θ_e^k}_{k=1}^K in Eq. (17);
end while
return θ_g, θ_f.
```

Table 4: Statistics of inductive benchmark datasets with four versions, where # Train and # Test denote the training and test sets, respectively. # Ent, # Rel and # Tri denote the numbers of entity, relation and triplets, respectively.

Datasets		WN18RR			FB15K-237			NELL-995		
		# Ent	# Rel	# Tri	# Ent	# Rel	# Tri	# Ent	# Rel	# Tri
v1	# Train	2,746	9	6,678	1,594	180	5,226	3,103	14	5,540
	# Test	922	8	1,991	1,093	142	2,404	225	14	1,034
v2	# Train	6,954	10	18,968	2,608	200	12,085	2,564	88	10,109
	# Test	2,757	10	4,863	1,660	172	5,092	2,086	79	5,521
v3	# Train	12,078	11	32,150	3,668	215	22,394	4,647	142	20,117
	# Test	5,084	11	7,470	2,501	183	9,137	3,566	122	9,688
v4	# Train	3,861	9	9,842	4,707	219	33,916	2,092	76	9,289
	# Test	7,084	9	15,157	3,051	200	14,554	2,795	61	8,520

- *Neural LP* Yang et al. (2017) learns interpretable logical rules over relational paths for inductive reasoning.
- *DRUM* Sadeghian et al. (2019) performs neural rule learning and tensor factorization via recurrent modeling of multi-hop relational paths.
- *RED-GNN* Zhang & Yao (2022) performs relational and feature reasoning via message passing over relational subgraph extraction.
- *AdaProp* Zhang et al. (2023b) improves reasoning by adaptively controlling message propagation weights in extracted subgraphs.
- *Ingram* Lee et al. (2023) constructs a relation graph by capturing affinity weights between relations.
- *KnowFormer* Liu et al. (2024) revisits the transformer and structure-aware attention mechanism for knowledge graph reasoning.
- *Context-Pooling* Su et al. (2025) enhances reasoning by identifying logically relevant neighbors via adaptive message aggregation over subgraph contexts.
- *HiMod* Xiong et al. (2026) improves inductive reasoning via hierarchical relational modeling with graph perturbation-enhanced message passing.
- *CAKGE* Cao et al. (2026) enhances semantic modeling via expressive knowledge graph embeddings with adaptive message aggregation.

We follow the same data split protocol described in Teru et al. (2020) to implement the compared baselines. Implementation details of these methods are given as follows.

- *Neural LP* Yang et al. (2017): https://github.com/fanyangxyz/Neural-LP
- *DRUM* Sadeghian et al. (2019): https://github.com/alisadeghian/DRUM
- *RED-GNN* Zhang & Yao (2022): https://github.com/LARS-research/RED-GNN
- *AdaProp* Zhang et al. (2023b): https://github.com/LARS-research/AdaProp
- *Ingram* Lee et al. (2023): https://github.com/bdi-lab/InGram
- *KnowFormer* Liu et al. (2024): https://github.com/jnanliu/KnowFormer
- *Context-Pooling* Su et al. (2025): http://github.com/ZhixiangSu/Context-Pooling
- *HiMod* Xiong et al. (2026): https://github.com/HubuKG/HiMod
- *CAKGE* Cao et al. (2026): https://github.com/Lion-ZS/CAKGE

Transductive Datasets. We conduct experiments on three transductive benchmark datasets: *WN18RR* Dettmers et al. (2018), *FB15K-237* Toutanova & Chen (2015), and *NELL-995* Xiong et al. (2017). Table 5 details the information about the transductive benchmark datasets.

Table 5: Statistics of transductive benchmark datasets.

Dataset	# Rel	# Ent	# Tri	
			# Train	# Test
WN18RR	11	40,493	86,835	3,134
FB5K-237	237	14,541	272,115	20,466
NELL-995	200	74,536	149,678	2,818

Transductive Baselines. We compare HiKGR with SOTA methods for transductive knowledge graph reasoning. We include triple-based methods *TransE* Bordes et al. (2013), *DistMult* Yang et al. (2014), and *RotatE* Sun et al. (2019), rule-based methods *DRUM* Sadeghian et al. (2019), message-passing-based methods *RED-GNN* Zhang & Yao (2022), *AdaProp* Zhang et al. (2023b), *KnowFormer* Liu et al. (2024), *DuetGraph* Jin et al. (2025), as well as pre-training model *Ultra* Galkin et al. (2023).

Implementation Details. Based on a workstation with an AMD EPYC 7642 CPU, a GeForce RTX 4090 GPU (24GB), and 109GB RAM, the proposed HiKGR is implemented using Python 3.8 and PyTorch 2.1.0. The multi-environment generator $\Psi_{\boldsymbol{\theta}_k}(\cdot)$ is implemented as a two-layer MLP with hidden dimensions $\{3d, d, 1\}$, where $ReLU(\cdot)$ is used as the activation function. The GNN encoder $\Phi_{\boldsymbol{\theta}_g}(\cdot)$ consists of L message passing layers. The learnable parameters include $\mathbf{W}_u, \mathbf{W}_{r_2}, \mathbf{W}_{r_q} \in \mathbb{R}^{d_a \times d}$, $\mathbf{w}_a \in \mathbb{R}^{d_a}$, $\mathbf{W}_c, \mathbf{W}_h \in \mathbb{R}^{d \times d}$, and $\mathbf{w}_g \in \mathbb{R}^{d_g}$. The linear predictor $f_{\boldsymbol{\theta}_f}(\cdot)$ is implemented as a fully connected layer with dimension $\{d, 1\}$. We use the Adam optimizer and tune the learning rate lr in $\{1e-5, 1e-2\}$. Table 6 lists the key parameters of HiKGR, where n_batch denotes the batch size, $epoch$ denotes the training epochs and dp is the dropout of GNN layers. For bi-level optimization strategy, we adopt an alternating optimization scheme, where the multi-environments generator and the GNN encoder and linear predictor are updated in turn.

Table 6: Parameter setting of HiKGR

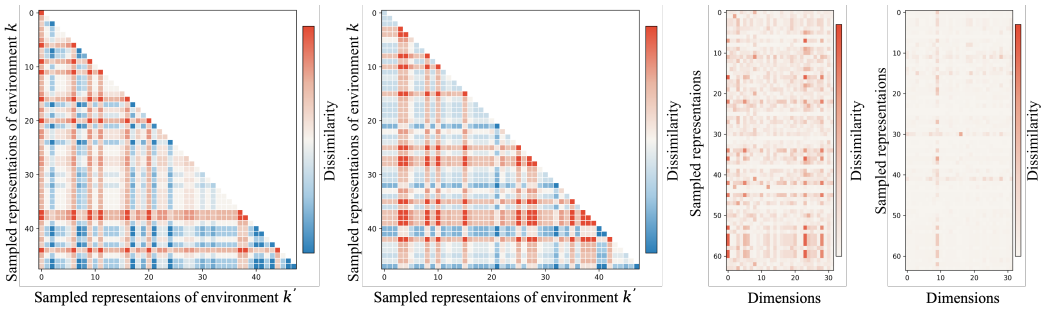
Dataset	lr	d	d_a	L	n_batch	dp	K	$epoch$	Dataset	lr	d	d_a	L	n_batch	dp	K	$epoch$
WN18RR_v1	0.0066	32	4	2	30	0.4770	2	30	WN18RR_v2	0.0036	32	3	2	30	0.3537	2	30
WN18RR_v3	0.0055	16	3	2	50	0.2335	2	30	WN18RR_v4	0.0010	32	3	1	50	0.3071	2	30
FB15K-237_v1	0.0009	32	3	1	100	0.3037	3	30	FB15K-237_v2	0.0067	48	5	5	20	0.3265	3	50
FB15K-237_v3	0.0065	32	5	7	20	0.4632	2	30	FB15K-237_v4	0.0030	16	3	3	50	0.4793	2	30
NELL-995_v1	0.0056	48	5	5	30	0.3068	2	80	NELL-995_v2	0.0060	32	5	3	30	0.4336	2	30
NELL-995_v3	0.0034	48	5	5	30	0.4336	2	30	NELL-995_v4	0.00012	16	5	1	20	0.2511	2	30

C.2 Experimental Results

RQ1: Performance comparison. We further evaluate the performance of HiKGR and the compared baselines in the transductive setting, as shown in Table 7. Overall, HiKGR achieves substantial improvements on all three datasets across most metrics. Notably, DuetGraph performs better on WN18RR and NELL-995 on some metrics, which can be attributed to its fine-tuning ability on the test set, allowing it to capture knowledge from a coarsely pretrained model. These results further validate the effectiveness of HiKGR.

Table 7: Performance comparison (MRR, Hits@1 and Hits@10) of SOTA methods on three transductive benchmark datasets (**bold**: optimal; underline: sub-optimal).

Methods	WN18RR			FB15K-237			NELL-995		
	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10
TransE	0.222	0.014	0.528	0.330	0.232	0.526	0.507	0.424	0.648
DistMult	0.455	0.410	0.544	0.358	0.264	0.550	0.510	0.438	0.636
RotatE	0.477	0.428	0.571	0.337	0.266	0.551	0.508	0.448	0.608
DRUM	0.486	0.425	0.586	0.343	0.255	0.516	0.532	0.460	0.662
RED-GNN	0.533	0.485	0.624	0.415	0.321	0.599	0.543	0.476	0.651
AdaProp	0.562	0.499	0.671	0.417	0.331	0.585	0.554	0.493	0.655
KnowFormer	0.579	0.528	0.687	0.430	0.343	0.608	0.566	0.502	0.675
DuetGraph	0.593	0.542	**0.699**	0.453	0.361	0.624	**0.590**	0.521	**0.712**
Ultra	0.480	0.414	0.614	0.368	0.272	0.564	0.509	0.441	0.660
HiKGR	**0.620**	**0.568**	0.698	**0.488**	**0.379**	**0.637**	0.586	**0.523**	0.702



(a) Visualization of domain diversity w/o and with $\mathcal{L}_{div}$ (b) Visualization of representation difference w/o and with $\mathcal{L}_{rel}$, $\mathcal{L}_{ent}$ and $\mathcal{L}_{sub}$

Figure 6: Visualization of (a) interpolated heatmaps of KGs across environments w/o and with $\mathcal{L}_{var}$ and (b) representations across environments w/o and with $\mathcal{L}_{rel}$, $\mathcal{L}_{ent}$ and $\mathcal{L}_{sub}$ in WN18RR_v1.

RQ2: Impact of environment diversity. We visualize the difference of structural variations and representation alignment across two environments, both with and without the proposed constraints. As shown in Figure 6 (a), incorporating $\mathcal{L}_{div}$ encourages greater environment diversity. In contrast, Figure 6 (b) shows that applying $\mathcal{L}_{rel}$, $\mathcal{L}_{ent}$, and $\mathcal{L}_{sub}$ leads to more consistent representations across environments. These observations are consistent with the performance trends in Table 8, suggesting that the proposed bi-level strategy effectively balances environment diversity and cross-environment invariance.

RQ3: Analysis of environment diversity and hierarchical semantics. As shown in Table 8, we evaluate the impact of the proposed objective functions on HiKGR. Removing $\mathcal{L}_{div}$ leads to significant performance degradation, as it prevents the model from effectively capturing invariant patterns from diverse structural environments. Removing any of $\mathcal{L}_{rel}$, $\mathcal{L}_{ent}$, or $\mathcal{L}_{sub}$, either individually or jointly, consistently degrades performance. This is because each level constrains invariant semantics at a different granularity, and jointly plays a unique role in extracting invariant patterns for extrapolation in the IKGR task.

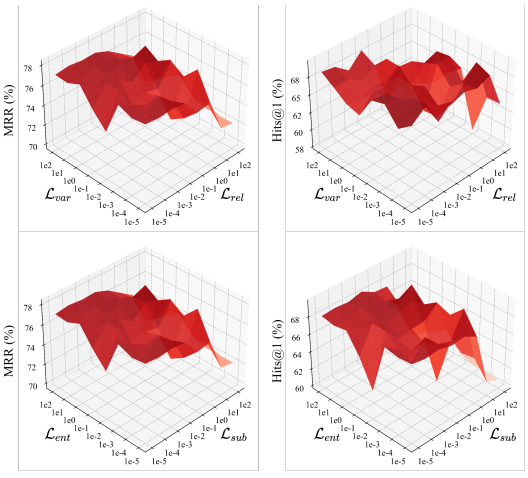


Figure 7: Parameter sensitivity of λ_{var}, λ_{rel}, λ_{ent} and λ_{sub} on WN18RR_v1 with MRR and Hits@1.

RQ4: Parameter Sensitivity. As shown in Figure 7, we analyze the parameter sensitivity of the proposed HiKGR on WN18RR_v1 with respect to the hyperparameters λ_{var}, λ_{rel}, λ_{ent} and λ_{sub}. We observe that the performance remains stable when λ_{var} and λ_{rel} are under $1e-1$, while it stays a lower level when they take large value. In comparison, the influence of λ_{ent} and λ_{sub} is gentler than λ_{var} and λ_{rel}, with

Table 8: Ablation study on three datasets, where w/o $\mathcal{L}_{rel/ent/sub}$ denotes w/o $\mathcal{L}_{rel}$, $\mathcal{L}_{ent}$ and $\mathcal{L}_{sub}$.

Datasets	WN18RR_v1			FB15K-237_v1			NELL-995_v1		
Metrics	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10
w/o $\mathcal{L}_{div}$	0.759	0.658	0.971	0.571	0.381	0.806	0.623	0.438	0.883
w/o $\mathcal{L}_{rel}$	0.773	0.681	0.973	0.589	0.393	0.853	0.669	0.508	0.886
w/o $\mathcal{L}_{ent}$	0.771	0.684	0.972	0.593	0.395	0.853	0.659	0.491	0.894
w/o $\mathcal{L}_{sub}$	0.773	0.684	0.972	0.590	0.394	0.844	0.550	0.388	0.878
w/o $\mathcal{L}_{rel/ent/sub}$	0.755	0.656	0.969	0.478	0.327	0.751	0.527	0.368	0.788
HiKGR	**0.779**	**0.691**	**0.973**	**0.600**	**0.403**	**0.853**	**0.768**	**0.669**	**0.896**

stable performance across a wider range in $[1e-4, 1]$. Overall, the satisfactory performance can be obtained when $\lambda_{var}, \lambda_{rel}$ are small and $\lambda_{ent}, \lambda_{sub}$ are moderately scaled.

Runtime Analysis. In Table 9, we report the training and inference runtime of various IKGR methods on three benchmark datasets. Despite introducing multi-environment construction and hierarchical invariant modeling, HiKGR maintains comparable training efficiency and is even faster than methods such as Neural LP, DRUM, and Context-Pooling. In terms of inference, HiKGR outperforms most baselines and remains competitive with RED-GNN and AdaProp, as it only requires standard forward propagation without additional cross-environment optimization. Although the training time is relatively higher on NELL-995_v1 due to increased structural complexity, HiKGR still achieves a favorable trade-off between performance and efficiency.

Table 9: Training and inference time comparison on different datasets.

Datasets	Time	Neural LP	DRUM	RED-GNN	AdaProp	Context-Pooling	HiMod	CAKGE	**HiKGR**
WN18RR_v1	Training Time	6.9200	6.8235	1.1518	0.7707	11.9807	1.5876	7.4828	6.3745
	Inference Time	2.2567	0.4389	0.9187	1.1208	8.7037	1.1303	5.2988	1.0613
FB15K-237_v1	Training Time	37.5167	49.0825	1.5903	1.1106	14.3430	15.3720	6.1323	1.7848
	Inference Time	8.8440	13.5499	0.9923	0.6710	11.2479	11.1572	4.2307	0.4600
NELL-995_v1	Training Time	6.5424	8.2377	3.2358	2.6463	16.5915	12.6488	4.5497	15.998
	Inference Time	0.8282	0.9281	1.6329	1.6755	11.7596	10.5114	2.3043	1.4883

D LIMITATION OF HIKGR

While HiKGR captures hierarchical invariant semantics for KGs, HiKGR can be further extended to more challenging settings, such as scenarios with completely disjoint entities and relations between training and test KGs. Moreover, although the model learns substantial invariant information across environments, some spurious patterns may still persist due to their strong correlation with target labels. Developing more effective mechanisms for modeling hierarchical invariant semantics remains an important direction for future work.