ACCURACY, COVERAGE AND FAIRNESS FOR KNOWLEDGE GRAPH EMBEDDING VIA DUAL CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing knowledge graph embedding (KGE) models achieve strong link prediction performance but lack reliable coverage and fairness guarantees. Recent conformal prediction (CP) methods for KGE provide coverage guarantees, yet they use CP as a post-hoc calibration step and do not explicitly address accuracy or fairness. We propose the first unified framework that jointly addresses accuracy, coverage, and fairness in KGE via dual label-clustering. Our key insight is that uncertainty in KGE is inherently structured: entities and relations exhibit distinct difficulty patterns that should be captured at the label level. To this end, we propose learnable low-rank factorization for entity and relation spaces, perform representation-driven dual label clustering, and integrate the resulting structure into self-adversarial training and cluster-wise conformal calibration. This produces cluster-adaptive prediction sets that maintain valid coverage while improving base KGE accuracy. We further provide a theoretical prediction set size disparity bound analysis that decomposes unfairness into intra-cluster heterogeneity, cross-cluster spread, and residual cross-group disparity, linking the dual label clustered CP to improved fairness. Empirically, our method consistently outperforms state-of-the-art baselines and achieves better accuracy, coverage effectiveness, and fairness, while acting as a general plug-in approach for diverse base KGE models.

1 INTRODUCTION

Knowledge graphs (KGs) represent structured knowledge as triples (Trouillon et al., 2016; Galkin et al., 2024), where each factual triple (h, r, t) indicates that a relation r holds between a head entity h and a tail entity t. Knowledge graph embedding (KGE) methods encode entities and relations as low-dimensional representations to capture structural knowledge such as symmetry, inversion, and composition (Bordes et al., 2013; Balazevic et al., 2019). These embeddings enable a wide range of downstream tasks (Yasunaga et al., 2022), including link prediction (Gao et al., 2020; Zhang et al., 2019), reasoning (Luo et al., 2025; Zhu et al., 2023), and question answering (Zhuo et al., 2025a;b).

Many KGE methods have been proposed, focusing on different geometric embedding spaces (Chami et al., 2020; Do Carmo, 1992; Eisenhart, 1997; Ficken, 1939) and scoring functions (Cao et al., 2021; Dettmers et al., 2018; Balazevic et al., 2019; Zhang et al., 2022). Despite these advances, most existing approaches produce ranked predictions without reliable uncertainty quantification, making it difficult to assess the confidence of predictions in high-stakes applications (Zhu et al., 2025). Recently, conformal prediction (CP) has been introduced directly to KGE, ensuring that the true answer is included in the prediction set with a given probability.

However, different groups of labels (entities or relations) in KGs exhibit highly heterogeneous characteristics, including varying frequencies, different structural roles, etc. (Ge et al., 2023; Li et al., 2024; Zhu et al., 2026). Existing conformal KGE methods ignore such heterogeneous label spaces that uncertainty is not uniformly distributed. Thus, the existing simple calibration leads to inefficient accuracy and overly conservative prediction sets, and unfairness across groups (Liu et al., 2026; Cresswell et al., 2025), as shown in Figure 1.

To address the above problems, we propose a unified KGE framework based on *dual label-clustering*. Instead of treating entities or relations as independent and homogeneous labels, we encode entities and relations into low-rank embedding spaces, enabling a representation-driven grouping of labels with similar semantic and structural properties to capture shared uncertainty characteristics. We

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

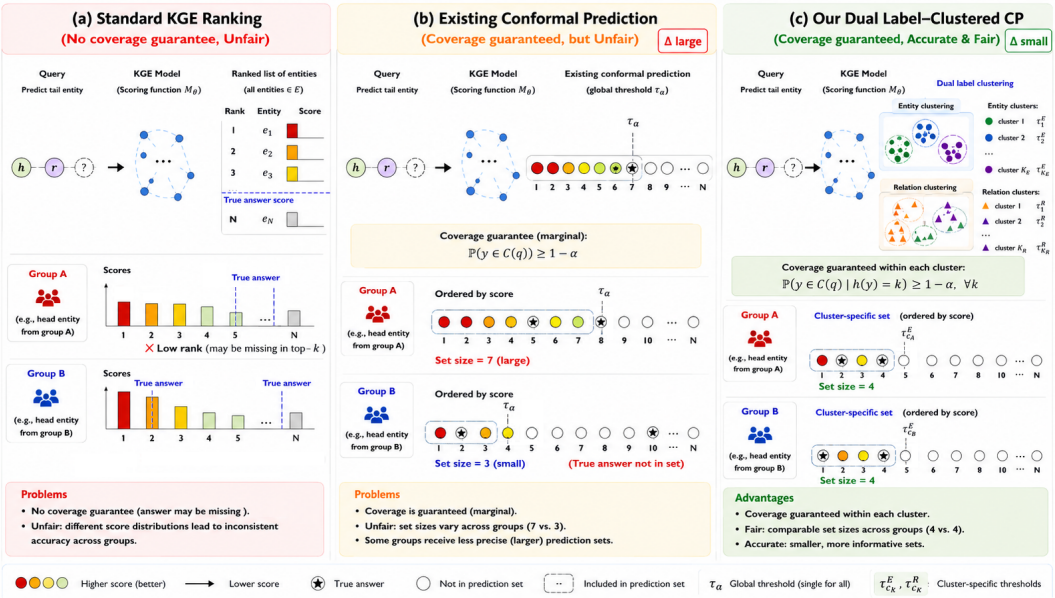


Figure 1: Comparison of KGE, CP, and ours. (a) Standard KGE produces ranked candidates without coverage guarantees and exhibits unfairness across groups, where different score distributions lead to inconsistent prediction accuracy. (b) Existing CP constructs prediction sets using a simple threshold τ_α, ensuring marginal coverage but producing unequal set sizes across groups, resulting in large set size disparity Δ. (c) Our dual label-clustered CP performs cluster-wise calibration over both entities and relations. It maintains valid coverage while producing comparable prediction set sizes across groups, resulting in smaller Δ, and improved fairness.

then integrate model training with cluster-aware self-adversarial losses to improve accuracy. Finally, we conduct conformal prediction through cluster-based calibration, resulting in reliable and tight prediction sets that adapt to variations in label difficulties.

We further provide a theoretical analysis showing that label clustering reduces prediction set size disparity,which establishes a connection between structured label clustering and improved fairness.

Our framework operates as a plug-and-play enhancement to existing KGE models. Empirically, we evaluate our method across common benchmark datasets and diverse KGE models, demonstrating consistent improvements in accuracy, coverage, and fairness compared to state-of-the-art (SOTA) baselines. In summary, our contributions are as follows:

- We propose the first unified framework that addresses accuracy, coverage, and fairness in KGE.
- We introduce a dual label-clustering method that models both entity and relation heterogeneity.
- Theoretical analysis shows that we reduce prediction set size disparity and improves fairness.
- Extensive experiments show that our method consistently improves accuracy, coverage, and fairness across diverse KGE models compared to SOTA baselines.

2 Background

2.1 Knowledge Graph Embedding

Let $\mathcal{E}$ and $\mathcal{R}$ denote the sets of entities and relations, respectively. A knowledge graph (KG) is a directed, edge-labeled graph, denoted by a set of triples:

$$\mathcal{G} = \{(h, r, t) \mid h, t \in \mathcal{V}, \ r \in \mathcal{R}\} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}, \tag{1}$$

where each triple (h, r, t) indicates that a relation $r \in \mathcal{R}$ holds between a head entity $h \in \mathcal{E}$ and a tail entity $t \in \mathcal{E}$.

A knowledge graph embedding model M_θ associates entities and relations with low-dimensional embeddings and then uses a scoring function:

$$M_\theta : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \to \mathbb{R}, \tag{2}$$

where θ denotes model parameters. The model score $M_\theta(h, r, t)$ reflects the plausibility of a triple, with the aim of scoring higher indicating more likely facts (Sun et al., 2019).

Let $\mathcal{Q}$ denote the query space and $\mathcal{Y}$ denote the label space ($\mathcal{Y} = \mathcal{E} \cup \mathcal{R}$). Each data point is $(q, y) \in \mathcal{Q} \times \mathcal{Y}$. Given a query $q \in \mathcal{Q}$, in the form of $\langle h, r, ? \rangle$, $\langle h, ?, t \rangle$ or $\langle ?, r, t \rangle$, the traditional KGE methods can output all ranked candidates for answers. The accuracy of a KGE model can be evaluated by the rank (Zhu et al., 2026) of the true answer y of the query q:

$$\mathrm{rank}(q, y) = \left| \{ y' \in \mathcal{E} \cup \mathcal{R} : M_\theta(q, y') \geq M_\theta(q, y) \} \right|. \tag{3}$$

Smaller rank values indicate a better model performance. However, rankings alone do not provide any uncertainty estimates, making it difficult to quantify prediction confidence.

2.2 Conformal Prediction for KGE

Conformal prediction can construct set-valued predictions with statistical guarantees. Specifically, a nonconformity function is defined as:

$$s : \mathcal{Q} \times \mathcal{Y} \to \mathbb{R}, \tag{4}$$

which measures how incompatible a candidate label is with a query, with higher scores indicating poorer agreement. In KGE, a natural choice is $s(q, y) = -M_\theta(q, y)$. Given a held-out calibration dataset $\mathcal{D}_{cal} = \{(q_i, y_i)\}_{i=1}^{n_{cal}}$, CP first computes calibration scores $S_i = s(q_i, y_i)$, $i = 1, \ldots, n_{cal}$. For a target coverage level $1 - \alpha \in (0, 1)$, the empirical conformal threshold $\hat{q}_\alpha$ is then given by:

$$\hat{q}_\alpha = \text{Quantile}_{\tau_\alpha}(S_1, \ldots, S_{n_{cal}}),\ \tau_\alpha = \frac{\lceil (n_{cal} + 1)(1 - \alpha) \rceil}{n_{cal}}. \tag{5}$$

The conformal prediction set for a test query q is then:

$$\mathcal{C}(q) = \{ y' \in \mathcal{Y} : s(q, y') \leq \hat{q}_\alpha \}. \tag{6}$$

Under the assumption that the training, calibration and test datasets are drawn independently and identically distributed (Vovk et al., 2005), the above constructed prediction set is guaranteed to contain the true answer y for query q with the probability of at least $1 - \alpha$, *i.e.,* $\mathbb{P}(y \in \mathcal{C}(q)) \geq 1 - \alpha$.

Thus, conformal prediction always transforms ranking-based candidates into $1 - \alpha \in (0, 1)$ coverage-guaranteed prediction sets, where the prediction set size $|\mathcal{C}(q)|$ serves as a measure of predictive uncertainty. Smaller $|\mathcal{C}(q)|$ indicates that the prediction for query q is more reliable and less uncertain, and the average prediction set size $\mathbb{E}[|\mathcal{C}|]$ over the whole query space $\mathcal{Q}$ is dictated by the accuracy and calibration of the base KGE models.

However, existing CP KGE methods (Zhu et al., 2026; 2025) ignore heterogeneous label spaces that uncertainty is not uniformly distributed. Thus, the existing simple calibration leads to inefficient accuracy and overly conservative prediction sets.

2.3 Fairness for KGE

We briefly introduce how fairness applies to KGE and review common fairness notions. In KGE-based link prediction, prediction sets are constructed over the label space $\mathcal{Y}$, and are used to support downstream decision-making. Ensuring fairness in KGE requires that these prediction sets be consistent across different sensitive groups of queries.

Specifically, let $A = [K_g]$ denote a finite set of sensitive group labels, and let $g : \mathcal{Q} \to A$ be a group assignment function that maps each query $q \in \mathcal{Q}$ to a group. Each group is defined as:

$$G_a = \{ q \in \mathcal{Q} : g(q) = a \}, \quad a \in A. \tag{7}$$

In the KGs, sensitive groups are usually applied at relation and entity level, such as entity groups of different genders, relation groups of domain-specific sensitive categories. Fairness is often defined in terms of parity across these query groups. For KGE, a notion is *group-conditional coverage*, which requires that prediction sets achieve the same coverage level for all sensitive groups:

$$\mathbb{P}(y \in \mathcal{C}(q) \mid q \in G_a) \geq 1 - \alpha, \quad \forall a \in A. \tag{8}$$

This fairness notion ensures that each query group meets the same statistical coverage guaranty. However, such above fairness requires labels of sensitive groups, and existing fair methods fails when sensitive group labels of all common KGE datasets are unknown. Thus, we use set size disparity (Liu et al., 2026) as a fairness metric for KGE.

Set size disparity as a fairness metric for KGE. As prediction set sizes serve as a measure of predictive uncertainty and confidence, disparities in set sizes across groups can therefore be the fairness metric for KGE. For any defined or undefined group $a \in A$, define its expected set size as $\mathbb{E}\left[|\mathcal{C}(q)| \mid q \in G_a\right]$. We define the set size disparity between two groups $a, b \in A$ as:

$$\Delta_{a,b} = \left|\mathbb{E}\left[|\mathcal{C}(q)| \mid q \in G_a\right] - \mathbb{E}\left[|\mathcal{C}(q)| \mid q \in G_b\right]\right|. \tag{9}$$

Then, we can quantify the model unfairness using the maximum set size disparity across all group pairs, *i.e.,* $\Delta = \max_{a,b \in A}\{\Delta_{a,b}\}$. This quantity captures differences in the predictive uncertainty and confidence across groups. Large disparities indicate that some groups receive less precise predictions while some groups receive more precise predictions, leading to more unfairness.

2.4 Discussion

In KGE, the disparities arise due to heterogeneity in entities and relations. For example, queries involving rare entities or complex relational structures often result in less accuracy and larger prediction sets, reflecting higher uncertainty. Existing CP KGE models ignore this heterogeneity, leading to inefficient accuracy and overly conservative prediction sets, and unfairness across groups. Therefore, achieving more accurate and fair KGE requires balancing coverage guarantees with controlled set size disparities with methods that explicitly model structured heterogeneity in label spaces.

3 Methodology

In this section, we present our unified framework for achieving accuracy, coverage, and fairness in knowledge graph embedding (KGE) via dual label-clustering. Our method consists of three key components: (i) dual label clustering over entities and relations with low-rank factorization, (ii) cluster-aware self-adversarial training for improving model accuracy, and (iii) cluster-wise conformal prediction for adaptive and fair uncertainty quantification.

3.1 Dual Label Clustering with Low-rank Factorization

Entities and relations in KGs exhibit highly heterogeneous uncertainty characteristics. For example, some labels are frequent or structurally simple, while others are sparse or associated with more complex relational patterns. To capture such heterogeneity, we introduce *dual label clustering with low-rank factorization*, which partitions both entity and relation spaces before conformal prediction.

Let ϕ denote the embedding layer of a base KGE model. For each entity $v \in \mathcal{E}$ and relation $r \in \mathcal{R}$, the base model maps them to embeddings:

$$e_v = \phi(v) \in \mathbb{R}^{d_\mathcal{E}}, \quad e_r = \phi(r) \in \mathbb{R}^{d_\mathcal{R}}. \tag{10}$$

The relation embedding dimension is often dependent on the entity embedding dimension (Li et al., 2022). For example, in TransE (Bordes et al., 2013), we have $d_\mathcal{R} = d_\mathcal{E}$. In the rest of this paper, we denote $d = d_\mathcal{E}$ as the entity embedding dimension.

Let $\mathbf{E} \in \mathbb{R}^{|\mathcal{E}| \times d}$ be the entity embedding matrix whose row corresponding to entity v is e_v, and let $\mathbf{R} \in \mathbb{R}^{|\mathcal{R}| \times d_\mathcal{R}}$ be the relation embedding matrix whose row corresponding to relation r is e_r. Directly clustering $\mathbf{E}$ and $\mathbf{R}$ may be sensitive to noise or redundancy. Instead of directly learning entity and relation embeddings $\mathbf{E}$ and $\mathbf{R}$, we therefore construct both embedding spaces through learnable low-rank factorization for the base KGE model:

$$\mathbf{E} = \mathbf{U}_E \mathbf{V}_E^\top, \quad \mathbf{R} = \mathbf{U}_R \mathbf{V}_R^\top. \tag{11}$$

Here, $\mathbf{U}_E \in \mathbb{R}^{|\mathcal{E}| \times k_E}$ is the learnable low-rank entity coordinate matrix, where the row $\mathbf{u}_v^E \in \mathbb{R}^{k_E}$ gives the compressed representation of entity v. Similarly, $\mathbf{U}_R \in \mathbb{R}^{|\mathcal{R}| \times k_R}$ is the learnable low-rank relation coordinate matrix, where the row $\mathbf{u}_r^R \in \mathbb{R}^{k_R}$ gives the compressed representation of relation r. The matrices $\mathbf{V}_E \in \mathbb{R}^{d \times k_E}$ and $\mathbf{V}_R \in \mathbb{R}^{d_\mathcal{R} \times k_R}$ are the shared affine matrices that map the low-rank coordinates back to the entity and relation embedding spaces of the base KGE model, respectively. Thus, $\mathbf{U}_E$ and $\mathbf{U}_R$ capture the main latent factors used for clustering, while $\mathbf{E}$ and $\mathbf{R}$ are used as original KGE embeddings.

This low-rank factorization replaces the direct learning of $\mathbf{E}$ and $\mathbf{R}$ with structured representations $\mathbf{U}_E\mathbf{V}_E^\top$ and $\mathbf{U}_R\mathbf{V}_R^\top$, enabling the model to disentangle global information (captured by the shared affine matrices $\mathbf{V}_E, \mathbf{V}_R$) from label-specific variability (captured by the entity and relation coordinate matrices $\mathbf{U}_E, \mathbf{U}_R$). In particular, the low-rank coordinates $\mathbf{U}_E$ and $\mathbf{U}_R$ provide a more stable and compact representation space for clustering.

3.2 Cluster-aware Self-adversarial Training

To improve prediction accuracy under structured label heterogeneity, we incorporate dual label clustering into the training objective. In existing KGE training, models are optimized to distinguish positive triples from negative ones using self-adversarial sampling weights (Sun et al., 2019). However, such training treats all negative samples uniformly and ignores the fact that both entities and relations exhibit different levels of difficulty and uncertainty across clusters.

Thus, we introduce dual cluster-aware weighting for both entity and relation negative triples. Given a positive triple (h, r, t), we construct two types of negative sample sets, *i.e.,* $\mathcal{N}_E = \{(h', r, t')\}$ and $\mathcal{N}_R = \{(h, r', t)\}$, where (h', r, t') are entity-corrupted and (h, r', t) are relation-corrupted.

Cluster-aware adversarial weighting. To incorporate dual cluster-aware weighting, we define self-adversarial sampling weights that depend on both model scores and cluster alignments. For entity-corrupted negatives (h', r, t'), we measure the similarity between the corrupted entity and the corresponding positive entity. Specifically, taking tail corruption for example, we define:

$$\mathrm{sim}_E(t', t) = \frac{\langle \mathbf{u}_{t'}^E, \mathbf{u}_t^E \rangle}{\|\mathbf{u}_{t'}^E\| \cdot \|\mathbf{u}_t^E\|}. \tag{12}$$

Thus, the tail entity-corrupted adversarial weight is:

$$w_E(h, r, t') = \frac{\exp\left[\gamma M_\theta(h, r, t') + \beta_E \mathrm{sim}_E(t', t)\right]}{\sum_{(h,r,t'') \in \mathcal{N}_E} \exp\left(\gamma M_\theta(h, r, t'') + \beta_E \mathrm{sim}_E(t'', t)\right)}. \tag{13}$$

For relation-corrupted negatives, we measure the similarity between the corrupted relation and the positive relation using $\mathrm{sim}_R(r', r) = \frac{\langle \mathbf{u}_{r'}^R, \mathbf{u}_r^R \rangle}{\|\mathbf{u}_{r'}^R\| \cdot \|\mathbf{u}_r^R\|}$. Similarly, the relation-corrupted adversarial weight is $w_R(h, r', t) = \frac{\exp\left(\gamma M_\theta(h,r',t) + \beta_R \mathrm{sim}_R(r', r)\right)}{\sum_{(h,r'',t) \in \mathcal{N}_R} \exp(\gamma M_\theta(h,r'',t) + \beta_R \mathrm{sim}_R(r'', r))}$. Here, $\gamma > 0$ is the adversarial temperature (Sun et al., 2019), while $\beta_E, \beta_R > 0$ control the influence of entity- and relation-level heterogeneity, respectively. The overall cluster-aware self-adversarial loss is defined as:

$$\mathcal{L} = -\log \sigma(M_\theta(h, r, t)) - \sum_{\mathcal{N}_E} w_E(h', r, t') \log \sigma(-M_\theta(h', r, t')) - \sum_{\mathcal{N}_R} w_R(h, r', t) \log \sigma(-M_\theta(h, r', t)). \tag{14}$$

3.3 Cluster-wise Conformal Prediction

We now extend conformal prediction to the clustered label spaces. For each entity $v \in \mathcal{E}$ and relation $r \in \mathcal{R}$, we use their learned low-rank coordinates $\mathbf{u}_v^E$ and $\mathbf{u}_r^R$ and the similarity as defined in Eq. 12 for clustering (Zhu et al., 2026), and then get two maps for entities and relations, respectively:

$$h_E : \mathcal{E} \to [K_E], \quad h_R : \mathcal{R} \to [K_R], \tag{15}$$

where $h_E(v)$ assigns entity v to one of K_E entity clusters and $h_R(r)$ assigns relation r to one of K_R relation clusters. For simplicity, we use $h(y)$ to denote the assigned cluster for an entity y or a relation y. These maps are then used for cluster-wise conformal calibration.

Instead of a simple threshold $\hat{q}_\alpha$, we compute cluster-specific thresholds. For each cluster $k \in [K_E]$ and $[K_R]$, define calibration subsets $\mathcal{D}_{cal}^{(k)} = \{(q_i, y_i) \in \mathcal{D}_{cal} : h(y_i) = k\}$. Compute cluster-specific nonconformity scores $S_i^{(k)} = s(q_i, y_i)$, where $(q_i, y_i) \in \mathcal{D}_{cal}^{(k)}$. Then we have:

$$\hat{q}_\alpha^{(k)} = \mathrm{Quantile}_{\tau_\alpha^{(k)}}(S_1^{(k)}, \ldots, S_{n_k}^{(k)}),\ \tau_\alpha^{(k)} = \frac{\lceil (|\mathcal{D}_{cal}^{(k)}| + 1)(1 - \alpha) \rceil}{|\mathcal{D}_{cal}^{(k)}|}. \tag{16}$$

Thus, our prediction set for the query q is $\mathcal{C}(q) = \{y' \in \mathcal{Y} : s(q, y') \leq \hat{q}_\alpha^{(h(y'))}\}$. which yields cluster-adaptive prediction sets that adjust to label-specific uncertainty and heterogeneity. For every cluster $k \in [K]$, we have:

$$\mathbb{P}\left(y \in \mathcal{C}(q) \mid h(y) = k\right) \geq 1 - \alpha. \tag{17}$$

Therefore, unlike existing CP KGE methods that only guarantee coverage with a simple threshold, our cluster-wise CP adapts to label-specific uncertainty and provides strong coverage guarantees conditioned on label clusters (Ding et al., 2023).

4 Theoretical Analysis

We now analyze why our cluster-wise conformal prediction improves fairness in KGE. Our goal is to show that label clustering reduces the maximum set size disparity Δ by controlling three interpretable sources of unfairness: intra-cluster heterogeneity, cross-cluster spread, and residual cross-group disparity.

Let $h : \mathcal{Y} \to [K]$ denote a generic label clustering map. In our framework, h can be instantiated as h_E for entity prediction and h_R for relation prediction. For each cluster $k \in [K]$, define the corresponding label subset as $\mathcal{Y}_k = \{y \in \mathcal{Y} : h(y) = k\}$.

For any group $a \in A$, cluster $k \in [K]$, and label $y \in \mathcal{Y}$, define:

$$\mu_{k,a} = \mathbb{E}\left[|\mathcal{C}(q)| \mid h(y) = k,\ q \in G_a\right],\ r_{y,a} = \mathbb{E}\left[|\mathcal{C}(q)| \mid y,\ q \in G_a\right],\ \epsilon_{k,a} = \max_{y \in \mathcal{Y}_k} r_{y,a} - \min_{y \in \mathcal{Y}_k} r_{y,a}. \tag{18}$$

Here, $\mu_{k,a}$ denotes the expected prediction set size for group a within cluster k, $r_{y,a}$ denotes the expected prediction set size associated with label y in group a, and $\epsilon_{k,a}$ measures the within-cluster variation of prediction set sizes.

Theorem 4.1 (Set Size Disparity Bound (Liu et al., 2026)). *For any clustering map $h : \mathcal{Y} \to [K]$ and any two groups $a, b \in A$, the set size disparity satisfies:*

$$\Delta_{a,b} \leq \underbrace{\max_{k \in [K]} \epsilon_{k,a}}_{\text{(I) Intra-cluster heterogeneity}} + \underbrace{\left(\max_{k \in [K]} \mu_{k,a} - \min_{k \in [K]} \mu_{k,a}\right)}_{\text{(II) Cross-cluster spread}} + \underbrace{\left|\sum_{y \in \mathcal{Y}} \mathbb{P}(y \mid q \in G_b)\left(r_{y,a} - r_{y,b}\right)\right|}_{\text{(III) Cross-group residual}}. \tag{19}$$

Consequently, the overall unfairness satisfies:

$$\Delta \leq \max_{a,b \in A} \left\{ \max_{k \in [K]} \epsilon_{k,a} + \left(\max_{k \in [K]} \mu_{k,a} - \min_{k \in [K]} \mu_{k,a}\right) + \left|\sum_{y \in \mathcal{Y}} \mathbb{P}(y \mid q \in G_b)\left(r_{y,a} - r_{y,b}\right)\right| \right\}. \tag{20}$$

Proof. The proof details and analysis are in Appendix A. □

Theorem 4.1 decomposes the set size disparity bound into three terms. The first term, intra-cluster heterogeneity, measures how different labels are within the same cluster. Our low-rank clustering reduces this term by grouping entities and relations with similar structural and uncertainty patterns. The second term, cross-cluster spread, measures how much expected prediction set sizes vary across clusters. Cluster-wise conformal calibration controls this term by assigning cluster-specific thresholds instead of using a simple threshold. The third term captures the disparity between groups within the same label, i.e., the intra-label cross-group disparity. This term reflects differences in prediction set sizes that arise from group-specific variations in the data distribution rather than label heterogeneity. Importantly, this component cannot be reduced by label clustering. Both existing CP KGE methods and our cluster-wise approach rely on shared calibration across groups within each label or label cluster, and therefore do not explicitly correct for group-specific discrepancies at the same label. Addressing the third term would require group prior information for group-aware calibration, which

Table 1: KGE link prediction accuracy results. Results improved by ours are in **bold**. $\mathbb{P}$ and $\mathbb{Q}$ refers to elliptic and hyperbolic geometry-based score functions (Do Carmo, 1992; Li et al., 2024).

Model		TransE	Ours	DistMult	Ours	RotatE	Ours	GoldE	Ours
$M_\theta(h,r,t)$		$-\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_2$		$\langle \mathbf{h}, \mathrm{diag}(\mathbf{r}), \mathbf{t}\rangle$		$-\|\mathbf{h}\circ\mathbf{r}-\mathbf{t}\|_2$		$-\|\mathbf{h}\circ\mathbf{r}-\mathbf{t}\|_{\mathbb{P},\mathbb{Q}}$	
embedding dimension $= 32$									
WN18RR	MRR↑	6.8±0.3	**7.2**±0.3	38.9±0.4	**49.5**±0.3	29.9±0.4	**30.5**±0.3	46.8±0.3	**47.5**±0.2
	H@1↑	0.1±0.1	**0.2**±0.1	36.3±0.4	**37.0**±0.3	27.3±0.4	**29.5**±0.4	40.7±0.4	**41.2**±0.3
	H@3↑	11.2±0.5	**12.3**±0.4	39.8±0.4	**40.7**±0.4	31.9±0.4	**32.0**±0.4	49.5±0.4	**50.1**±0.3
	H@10↑	17.9±0.4	**19.4**±0.3	43.6±0.3	**45.4**±0.3	34.6±0.3	**34.8**±0.3	57.1±0.3	**57.5**±0.2
FB15K-237	MRR↑	29.1±0.3	**30.4**±0.3	28.0±0.3	**28.6**±0.3	28.1±0.3	**28.7**±0.2	32.0±0.2	**32.5**±0.2
	H@1↑	20.5±0.4	**21.6**±0.4	19.7±0.4	**20.5**±0.4	20.1±0.4	**20.5**±0.3	23.2±0.3	**23.8**±0.3
	H@3↑	31.9±0.5	**33.5**±0.5	30.6±0.5	**31.3**±0.5	30.5±0.5	**31.2**±0.4	35.0±0.4	**35.4**±0.4
	H@10↑	46.4±0.6	**48.3**±0.6	44.8±0.5	**45.4**±0.5	44.3±0.5	**44.9**±0.5	49.3±0.5	**49.9**±0.4
embedding dimension $= 128$									
FB15K-237	MRR↑	29.3±0.4	**31.0**±0.4	28.7±0.4	**28.8**±0.3	28.4±0.3	**28.7**±0.3	33.0±0.3	**33.6**±0.2
	H@1↑	20.2±0.4	**21.9**±0.4	20.3±0.4	**20.7**±0.3	20.1±0.4	**20.8**±0.3	24.1±0.4	**24.8**±0.3
	H@3↑	32.5±0.5	**35.3**±0.4	31.3±0.4	**31.3**±0.3	30.8±0.4	**31.2**±0.3	36.3±0.4	**36.7**±0.3
	H@10↑	47.3±0.6	**49.6**±0.6	45.2±0.5	**45.7**±0.5	44.8±0.5	**45.3**±0.4	51.3±0.5	**51.4**±0.4
YAGO3-10	MRR↑	30.1±0.4	**32.3**±0.4	17.2±0.4	**17.8**±0.3	39.0±0.3	**40.1**±0.3	39.1±0.3	**40.3**±0.3
	H@1↑	20.1±0.5	**22.2**±0.5	10.7±0.5	**11.3**±0.4	29.2±0.4	**30.0**±0.4	29.7±0.4	**30.5**±0.3
	H@3↑	35.3±0.6	**38.4**±0.5	18.3±0.5	**18.8**±0.4	43.7±0.4	**45.6**±0.4	43.7±0.4	**45.9**±0.4
	H@10↑	49.2±0.6	**52.3**±0.6	29.4±0.5	**30.4**±0.5	57.7±0.5	**59.1**±0.5	58.0±0.5	**59.3**±0.5
embedding dimension $= 256$									
FB15K-237	MRR↑	30.0±0.4	**31.3**±0.4	28.9±0.3	**29.4**±0.2	28.9±0.3	**29.0**±0.2	33.8±0.3	**33.9**±0.2
	H@1↑	20.8±0.5	**21.5**±0.4	20.8±0.4	**21.3**±0.3	21.0±0.4	**21.0**±0.3	24.9±0.3	**25.1**±0.2
	H@3↑	33.3±0.6	**35.3**±0.5	31.8±0.5	**32.0**±0.4	31.3±0.4	**31.4**±0.4	37.0±0.4	**37.1**±0.3
	H@10↑	48.4±0.6	**50.3**±0.6	45.5±0.5	**45.8**±0.5	45.2±0.5	**45.4**±0.5	51.8±0.5	**52.1**±0.4
embedding dimension $= 1000$									
WN18RR	MRR↑	22.1±0.4	**22.6**±0.4	43.5±0.4	**44.0**±0.3	47.4±0.4	**47.9**±0.2	51.0±0.2	**51.4**±0.1
	H@1↑	1.3±0.3	**1.7**±0.3	39.1±0.4	**39.8**±0.4	42.5±0.4	**43.6**±0.3	46.2±0.3	**46.6**±0.2
	H@3↑	40.1±0.5	**40.5**±0.4	44.9±0.5	**45.5**±0.4	49.3±0.4	**49.5**±0.3	53.0±0.3	**53.3**±0.3
	H@10↑	52.6±0.5	**53.2**±0.5	53.1±0.5	**53.5**±0.4	57.2±0.5	**57.9**±0.4	60.5±0.5	**60.8**±0.4

is beyond the scope of this work. Thus, for KGE, our dual label clustering is directly aligned with this fairness result. Entity clustering reduces uncertainty variation across candidate answers, while relation clustering stabilizes calibration across relations with different difficulties. Together, they reduce Δ, leading to tighter prediction set sizes and improved fairness while preserving coverage.

5 Experiments

We conduct experiments to evaluate whether our method improves accuracy, coverage, and fairness. Specifically, we aim to answer the following questions: (1) Does our method improve KGE link prediction accuracy? (2) Does our method maintain valid coverage while producing tighter prediction set sizes? (3) Does our method reduce set size disparity across groups for better fairness?

5.1 Experimental Setups

Datasets. Following existing works (Sun et al., 2019; Zhu et al., 2025; 2026), we evaluated on three standard benchmarks: WN18RR (Dettmers et al., 2018), FB15k-237 (Toutanova & Chen, 2015), and YAGO3-10 (Mahdisoltani et al., 2015). Appendix B provides more details.

Base KGE models and baselines. We use a wide range of advanced base KGE models, including TransE (translation/Euclidean) (Bordes et al., 2013), DistMult (bilinear/Euclidean) (Yang et al., 2015), RotatE (rotation/2D) (Sun et al., 2019), and GoldE (rotation/hyperbolic/Euclidean/kD) (Li et al., 2024). We compare with SOTA CP KGE baselines: (1) Top-k (Zhu et al., 2025) directly uses model scores from each base KGE. (2) KGCP (Zhu et al., 2025) applies traditional conformal prediction using a global threshold. (3) CondKGCP (Zhu et al., 2026) proposes predicate-conditional conformal prediction using different types of nonconformity scores.

Implementation details: For fair comparisons (Zhu et al., 2026; Li et al., 2024), we fix the embedding dimension for experimental comparison. We use official open-source implementations for each model. We report mean reciprocal rank (MRR) and HITS at N (H@N) metrics to evaluate the link prediction accuracy. For conformal prediction coverage performance, we report empirical coverage (EC) and

average prediction set size (APSS). For fairness performance, we report the maximum set size disparity across groups, following KGCP paper Section 6.2 (Zhu et al., 2026) where each group contains similar query difficulties. We report mean±std over 5 runs. More details are provided in Appendix C.

5.2 Accuracy Results

We first evaluate whether our method improve base KGE model accuracy. Table 1 reports link prediction results across benchmark datasets. As SOTA baselines, *i.e.,* Top-k, KGCP and CondKGCP, conduct CP after training the original base KGE models, their link prediction accuracy is the same as each base KGE model.

The results show that our method consistently improves link prediction accuracy across different datasets, embedding dimensions, and base KGE models. This indicates that our method is model-agnostic and can serve as a plug-in enhancement for diverse KGE models.

The gains are particularly clear on FB15K-237 and YAGO3-10, where relation patterns are more diverse and label distributions are more heterogeneous. This supports our motivation that structured label heterogeneity should be explicitly modeled.

Since Top-k, KGCP, and CondKGCP are applied after training the base KGE models, they do not modify the base KGE models and therefore share the same link prediction accuracy as the corresponding base KGE model. In contrast, our method improves the base model itself through low-rank factorization and cluster-aware self-adversarial training. This explains why our framework improves KGE accuracy. This also indicates that low-rank factorization and cluster-aware self-adversarial weighting help the model learn more discriminative entity and relation representations.

5.3 Coverage Results

We next evaluate whether our method improves coverage performance and reduces prediction set sizes. Table 2 reports empirical coverage and average set size under different target coverage levels using the best base KGE model GoldE with embedding dimension $= 128$. The results in Table 2 demonstrate that our method achieves a better coverage performance and much better prediction set effectiveness compared to SOTA baselines.

Table 2: Coverage performance and average prediction set size comparison. The target level is $1 - \alpha = 90\%$. Smaller prediction set size is more effective.

Dataset	Method	EC ↑	APSS ↓
WN18RR	Top-k	0.90	7944
	KGCP	0.90	8057
	CondKGCP	0.90	5463
	Ours	**0.91**	**3517**
FB15k-237	Top-k	0.90	841
	KGCP	0.90	375
	CondKGCP	0.90	389
	Ours	**0.91**	**253**
YAGO3-10	Top-k	0.90	5970
	KGCP	0.90	5833
	CondKGCP	0.90	5698
	Ours	**0.90**	**4641**

First, our method consistently satisfies the target coverage requirement. This is consistent with the guarantee that cluster-wise CP adapts to label-specific uncertainty and provides better coverage performance (Ding et al., 2023). Second, our method significantly reduces the average prediction set size (APSS) compared to SOTA baselines, demonstrating that cluster-specific thresholds better adapt to label-dependent uncertainty. Sspecifically, on WN18RR, our method reduces APSS from 8057 (KGCP) and 5463 (CondKGCP) to 3517, achieving a substantial reduction of more than 30%. Overall, these results confirm that modeling structured label heterogeneity is critical for effective conformal prediction in KGE. By aligning calibration with label clusters, our method reduces unnecessary inclusion of low-confidence candidates, leading to more informative prediction sets without violating coverage guarantees.

5.4 Fairness Results

We further evaluate whether our method reduces unfairness measured by the maximum set size disparity metric Δ. Table 3 reports fairness comparison across datasets using the best base KGE model GoldE with embedding dimension $= 128$. The results in Table 3 show that our method consistently achieves the lowest set size disparity Δ across all datasets, indicating a significant improvement in fairness compared to existing baselines.

While SOTA methods reduce disparity compared to Top-k, they still exhibit relatively large Δ. KGCP uses a single threshold, which fails to adapt to label-specific uncertainty and often leads to over-conservative prediction sets for difficult groups. CondKGCP partially alleviates this issue by conditioning on predicates, but it still does not fully capture label space heterogeneity, limiting its ability to balance prediction set sizes across groups. These empirical findings are consistent with our theoretical analysis in Theorem 4.1. Specifically, entity clustering reduces intra-cluster heterogeneity by grouping labels with similar uncertainty, while relation clustering reduces cross-cluster spread by stabilizing calibration across predicates with different difficulty levels. As a result, the first two components in the Theorem 4.1 disparity decomposition are effectively controlled, leading to a smaller overall Δ.

6 Related Work

KGE. Early KGE methods explore geometric operations (Ge et al., 2023; Zhang et al., 2022), including translation (Bordes et al., 2013; Wang et al., 2014; Ji et al., 2015), bilinear (Yang et al., 2015; Trouillon et al., 2016), and 2D rotations (Sun et al., 2019), to model relations. Then, many works (Li et al., 2022; Zhang et al., 2019; Gao et al., 2020; Cao et al., 2021) seek higher-dimensional rotations for better modeling capacity. Others explores rotations for different (Balazevic et al., 2019; Chami et al., 2020) or combined (Xiong et al., 2022) geometries to preserve topological structures. GoldE (Li et al., 2024) generalizes existing KGE models in terms of both higher dimensions and different geometries.

Table 3: Fairness comparison measured by maximum set size disparity Δ. Lower values indicate better fairness.

Dataset	Method	$\Delta \downarrow$
WN18RR	Top-k	10037
	KGCP	9941
	CondKGCP	8506
	Ours	**7165**
FB15k-237	Top-k	1537
	KGCP	1429
	CondKGCP	1371
	Ours	**1094**
YAGO3-10	Top-k	8962
	KGCP	8457
	CondKGCP	8170
	Ours	**7429**

Conformal prediction for KGE. Recent works introduce CP for KGE to ensure coverage guaranty. Top-k (Zhu et al., 2025) directly uses model scores from each base KGE to rank candidates. KGCP (Zhu et al., 2025) applies traditional conformal prediction using a global threshold. CondKGCP (Zhu et al., 2026) proposes predicate-conditional conformal prediction using different types of nonconformity scores.

Fairness for KGE. Many works (Cresswell et al., 2025; Ding et al., 2023; Vovk et al., 2005) explore fairness for computer vision (CV) and natural language processing (NLP). SubstantiveFair (Liu et al., 2026) shows that equalized set sizes, rather than coverage, strongly correlate with improved fairness for CV and NLP.

Despite the significant progress in the above directions, current methods exhibit several key limitations. Existing CP KGE methods, only address coverage guarantees while overlooking both accuracy and fairness, as they ignore that both entity and relation label spaces are heterogeneous. While fairness has been studied for CV and NLP, these approaches are not directly applicable to KGE, as they do not consider the unique characteristics of KGs, such as highly imbalanced label distributions and the dual role of entities and relations.

7 Conclusion

We propose a unified framework for improving accuracy, coverage, and fairness in knowledge graph embedding. Our key idea is to model uncertainty as a structured property of heterogeneous entity and relation label spaces. We introduce dual label clustering with learnable low-rank factorization, incorporate cluster-aware self-adversarial training to improve base KGE accuracy, and perform cluster-wise conformal prediction to obtain adaptive prediction sets with label-cluster conditional coverage. Theoretically, we show that label clustering reduces set size disparity by controlling intra-cluster heterogeneity and cross-cluster spread, establishing a connection between structured calibration and fairness. Empirically, our method consistently improves link prediction accuracy, maintains valid coverage, reduces average prediction set size, and achieves lower maximum set size disparity compared to SOTA baselines. Future work includes extending our framework to dynamic knowledge graphs and incorporating group-aware calibration when sensitive group information is available.

AI USE STATEMENT

We only used generative AI tools to assist with the writing of this paper. All AI-assisted text were carefully reviewed and verified by the authors. The research ideas, methodology, experimental design, analysis, and scientific claims were conducted independently by the authors. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

In this work, we have made every effort to ensure reproducibility. In Section 3, from input to output, we provide a detailed description of our method, including model structure, model training, and anomaly scores. In Appendix B, we describe all the datasets in detail. In Appendix C, we summarize all the implementation details and default hyper-parameters, including batch size, window size, learning rate, etc. Our codes will be available upon acceptance.

ETHICS AND IMPACT STATEMENT

This work aims to improve the reliability and fairness of knowledge graph embedding (KGE) systems by jointly addressing accuracy, coverage, and set size disparity. Since KGE models are increasingly used in downstream applications such as biomedical knowledge discovery, recommendation, question answering, and decision support, unreliable or over-confident predictions may lead to misleading conclusions. By incorporating conformal prediction, our method provides statistically valid prediction sets, helping users better understand model uncertainty rather than relying solely on ranked outputs.

A positive societal impact of this work is that it encourages safer deployment of KGE models in high-stakes settings. Instead of producing a single ranked answer without uncertainty information, our method returns prediction sets with coverage guarantees. This can help practitioners identify when the model is uncertain and avoid over-reliance on potentially incorrect predictions. Moreover, by reducing prediction set size disparity across groups, our framework can provide more balanced decision support for different entity or relation groups.

However, this work also has limitations and potential risks. First, fairness is evaluated through set size disparity, which captures differences in prediction set informativeness but does not fully characterize all forms of social or ethical fairness. In particular, as discussed in Theorem 4.1, label clustering does not eliminate intra-label cross-group residual disparity. Addressing this residual component may require explicit group-aware calibration or domain-specific fairness constraints.

Second, the fairness benefits of our method depend on the quality of the underlying knowledge graph and its labels. If a KG encodes historical bias, missing facts, or biased relation structures, improving prediction accuracy and calibration may still propagate or amplify biased information. Therefore, our method should not be interpreted as a complete solution to bias in knowledge graphs.

Third, in sensitive applications such as healthcare, finance, or legal reasoning, prediction sets should be used as decision-support tools rather than autonomous decision-making mechanisms. Human oversight, domain validation, and careful auditing remain necessary before deployment.

Overall, this work contributes a technical step toward more reliable and fair uncertainty quantification in KGE, but responsible deployment requires combining our method with broader data governance, fairness auditing, and domain-specific safeguards.

REFERENCES

Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. Multi-relational poincaré graph embeddings. In *Advances in Neural Information Processing Systems*, pp. 4465–4475, 2019.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 1247–1250, 2008.

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

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. Dual quaternion knowledge graph embeddings. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, pp. 6894–6902, 2021.

Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. Low-dimensional hyperbolic knowledge graph embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6901–6914, 2020.

Jesse C. Cresswell, Bhargava Kumar, Yi Sui, and Mouloud Belbahri. Conformal prediction sets can cause disparate impact. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64555–64576, 2023.

Manfredo Perdigao Do Carmo. *Riemannian geometry*. Birkhäuser, 1992.

Luther Pfahler Eisenhart. *Riemannian geometry*. Princeton University Press, 1997.

Frederick Arthur Ficken. The riemannian and affine differential geometry of product-spaces. *Annals of Mathematics*, pp. 892–913, 1939.

Mikhail Galkin, Xinyu Yuan, Hesham Mostafa, Jian Tang, and Zhaocheng Zhu. Towards foundation models for knowledge graph reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Chang Gao, Chengjie Sun, Lili Shan, Lei Lin, and Mingjiang Wang. Rotate3d: Representing relations as rotations in three-dimensional space for knowledge graph embedding. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 385–394, 2020.

Xiou Ge, Yun-Cheng Wang, Bin Wang, and C.-C. Jay Kuo. Compounding geometric operations for knowledge graph completion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6947–6965, 2023.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 687–696, 2015.

Rui Li, Jianan Zhao, Chaozhuo Li, Di He, Yiqi Wang, Yuming Liu, Hao Sun, Senzhang Wang, Weiwei Deng, Yanming Shen, Xing Xie, and Qi Zhang. House: Knowledge graph embedding with householder parameterization. In *International Conference on Machine Learning*, pp. 13209–13224, 2022.

Rui Li, Chaozhuo Li, Yanming Shen, Zeyu Zhang, and Xu Chen. Generalizing knowledge graph embedding with universal orthogonal parameterization. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Pengqi Liu, Zijun Yu, Mouloud Belbahri, Arthur Charpentier, Masoud Asgharian, and Jesse C. Cresswell. Beyond procedure: Substantive fairness in conformal prediction. *CoRR*, abs/2602.16794, 2026.

Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Yuan-Fang Li, Chen Gong, and Shirui Pan. Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek. Yago3: A knowledge base from multilingual wikipedias. In *7th Biennial Conference on Innovative Data Systems Research*, 2015.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, pp. 39–41, 1995.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations*, 2019.

Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and Their Compositionality*, pp. 57–66, 2015.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *Proceedings of the 33nd International Conference on Machine Learning*, pp. 2071–2080, 2016.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer New York, 2005.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 1112–1119, 2014.

Bo Xiong, Shichao Zhu, Mojtaba Nayyeri, Chengjin Xu, Shirui Pan, Chuan Zhou, and Steffen Staab. Ultrahyperbolic knowledge graph embeddings. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2130–2139, 2022.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations*, 2015.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. In *Advances in Neural Information Processing Systems*, pp. 37309–37323, 2022.

Qianjin Zhang, Ronggui Wang, Juan Yang, and Lixia Xue. Knowledge graph embedding by reflection transformation. *Knowl. Based Syst.*, 238:107861, 2022.

Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. Quaternion knowledge graph embeddings. In *Advances in Neural Information Processing Systems*, pp. 2731–2741, 2019.

Yuqicheng Zhu, Nico Potyka, Jiarong Pan, Bo Xiong, Yunjie He, Evgeny Kharlamov, and Steffen Staab. Conformalized answer set prediction for knowledge graph embedding. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *NAACL 2025*, pp. 731–750. Association for Computational Linguistics, 2025.

Yuqicheng Zhu, Daniel Hernández, Yuan He, Zifeng Ding, Bo Xiong, Evgeny Kharlamov, and Steffen Staab. Predicate-conditional conformalized answer sets for knowledge graph embeddings. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *ACL 2025*, Findings of ACL, pp. 4145–4167. Association for Computational Linguistics, 2026.

Zhaocheng Zhu, Xinyu Yuan, Michael Galkin, Louis-Pascal A. C. Xhonneux, Ming Zhang, Maxime Gazeau, and Jian Tang. A*net: A scalable path-based reasoning approach for knowledge graphs. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Xingrui Zhuo, Shirui Pan, Jiapu Wang, Gongqing Wu, Zan Zhang, Rui Li, Zizhong Wei, and Xindong Wu. Progressive prefix-memory tuning for complex logical query answering on knowledge graphs. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pp. 3716–3724. ijcai.org, 2025a.

Xingrui Zhuo, Jiapu Wang, Gongqing Wu, Shirui Pan, and Xindong Wu. Effective instruction parsing plugin for complex logical query answering on knowledge graphs. In Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (eds.), *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, pp. 4780–4792. ACM, 2025b.

A Theoretical Proofs and Analysis

Our goal is to show that label clustering reduces the maximum set size disparity Δ by controlling three interpretable sources of unfairness: intra-cluster heterogeneity, cross-cluster spread, and residual cross-group disparity.

$h : \mathcal{Y} \to [K]$ denote a generic label clustering map. In our framework, h can be instantiated as h_E for entity prediction and h_R for relation prediction. For each cluster $k \in [K]$, define the corresponding label subset as $\mathcal{Y}_k = \{y \in \mathcal{Y} : h(y) = k\}$.

For any group $a \in A$, cluster $k \in [K]$, and label $y \in \mathcal{Y}$, define:

$$\mu_{k,a} = \mathbb{E}\left[|\mathcal{C}(q)| \mid h(y) = k,\ q \in G_a\right],\ r_{y,a} = \mathbb{E}\left[|\mathcal{C}(q)| \mid y,\ q \in G_a\right],\ \epsilon_{k,a} = \max_{y \in \mathcal{Y}_k} r_{y,a} - \min_{y \in \mathcal{Y}_k} r_{y,a}. \tag{21}$$

Here, $\mu_{k,a}$ denotes the expected prediction set size for group a within cluster k, $r_{y,a}$ denotes the expected prediction set size associated with label y in group a, and $\epsilon_{k,a}$ measures the within-cluster variation of prediction set sizes.

Theorem A.1 (Set Size Disparity Bound (Liu et al., 2026)). *For any clustering map $h : \mathcal{Y} \to [K]$ and any two groups $a, b \in A$, the set size disparity satisfies:*

$$\Delta_{a,b} \leq \underbrace{\max_{k \in [K]} \epsilon_{k,a}}_{\text{(I) Intra-cluster heterogeneity}} + \underbrace{\left(\max_{k \in [K]} \mu_{k,a} - \min_{k \in [K]} \mu_{k,a}\right)}_{\text{(II) Cross-cluster spread}} + \underbrace{\left|\sum_{y \in \mathcal{Y}} \mathbb{P}(y \mid q \in G_b)\left(r_{y,a} - r_{y,b}\right)\right|}_{\text{(III) Cross-group residual}}. \tag{22}$$

Consequently, the overall unfairness satisfies:

$$\Delta \leq \max_{a,b \in A} \left\{ \max_{k \in [K]} \epsilon_{k,a} + \left(\max_{k \in [K]} \mu_{k,a} - \min_{k \in [K]} \mu_{k,a}\right) + \left|\sum_{y \in \mathcal{Y}} \mathbb{P}(y \mid q \in G_b)\left(r_{y,a} - r_{y,b}\right)\right| \right\}. \tag{23}$$

Proof. By the law of total expectation, the expected prediction set size for group a can be written as:

$$\mathbb{E}\left[|\mathcal{C}(q)| \mid q \in G_a\right] = \sum_{y \in \mathcal{Y}} \mathbb{P}(y \mid q \in G_a) r_{y,a}. \tag{24}$$

Thus, for two groups $a, b \in A$, we have:

$$\Delta_{a,b} = \left|\sum_{y \in \mathcal{Y}} \mathbb{P}(y \mid q \in G_a) r_{y,a} - \sum_{y \in \mathcal{Y}} \mathbb{P}(y \mid q \in G_b) r_{y,b}\right|. \tag{25}$$

Adding and subtracting $\sum_{y \in \mathcal{Y}} \mathbb{P}(y \mid q \in G_b) r_{y,a}$ gives:

$$\Delta_{a,b} \leq T_1 + T_2, \tag{26}$$

where

$$T_1 = \left|\sum_{y \in \mathcal{Y}} \left(\mathbb{P}(y \mid q \in G_a) - \mathbb{P}(y \mid q \in G_b)\right) r_{y,a}\right|, \tag{27}$$

and

$$T_2 = \left| \sum_{y \in \mathcal{Y}} \mathbb{P}(y \mid q \in G_b) \left(r_{y,a} - r_{y,b} \right) \right|. \tag{28}$$

The term T_2 is exactly the cross-group residual term in Theorem 4.1. We now bound T_1. Grouping labels by clusters gives:

$$T_1 = \left| \sum_{k=1}^{K} \sum_{y \in \mathcal{Y}_k} \left(\mathbb{P}(y \mid q \in G_a) - \mathbb{P}(y \mid q \in G_b) \right) r_{y,a} \right|. \tag{29}$$

For each $y \in \mathcal{Y}_k$, decompose $r_{y,a}$ around the cluster-level quantity $\mu_{k,a}$:

$$r_{y,a} = \mu_{k,a} + \left(r_{y,a} - \mu_{k,a} \right). \tag{30}$$

Substituting this decomposition into T_1 yields:

$$T_1 \leq T_{1,1} + T_{1,2}, \tag{31}$$

where

$$T_{1,1} = \left| \sum_{k=1}^{K} \sum_{y \in \mathcal{Y}_k} \left(\mathbb{P}(y \mid q \in G_a) - \mathbb{P}(y \mid q \in G_b) \right) \left(r_{y,a} - \mu_{k,a} \right) \right|, \tag{32}$$

and

$$T_{1,2} = \left| \sum_{k=1}^{K} \mu_{k,a} \sum_{y \in \mathcal{Y}_k} \left(\mathbb{P}(y \mid q \in G_a) - \mathbb{P}(y \mid q \in G_b) \right) \right|. \tag{33}$$

Since $\mu_{k,a}$ lies between the minimum and maximum of $r_{y,a}$ within cluster k, we have:

$$|r_{y,a} - \mu_{k,a}| \leq \epsilon_{k,a}. \tag{34}$$

Therefore, the within-cluster contribution is controlled by:

$$T_{1,1} \leq \max_{k \in [K]} \epsilon_{k,a}. \tag{35}$$

Next, since each $\mu_{k,a}$ is bounded between $\min_{k \in [K]} \mu_{k,a}$ and $\max_{k \in [K]} \mu_{k,a}$, the cluster-level contribution is bounded by the range of cluster-level expected set sizes:

$$T_{1,2} \leq \max_{k \in [K]} \mu_{k,a} - \min_{k \in [K]} \mu_{k,a}. \tag{36}$$

Combining the bounds for $T_{1,1}$, $T_{1,2}$, and T_2, we obtain:

$$\Delta_{a,b} \leq \max_{k \in [K]} \epsilon_{k,a} + \left(\max_{k \in [K]} \mu_{k,a} - \min_{k \in [K]} \mu_{k,a} \right) + \left| \sum_{y \in \mathcal{Y}} \mathbb{P}(y \mid q \in G_b) \left(r_{y,a} - r_{y,b} \right) \right|. \tag{37}$$

This proves the first result. Taking the maximum over all group pairs $a, b \in A$ gives the bound on Δ.

□

B DATASETS

Table 4 summarizes the detailed statistics of three datasets that are standard benchmarks for the knowledge graph embedding and link prediction task.

WN18RR (Dettmers et al., 2018) and FB15k-237 (Toutanova & Chen, 2015) datasets are subsets of WN18 (Bordes et al., 2013) and FB15k (Bordes et al., 2013), respectively, with inverse relations removed to avoid test leakage. WN18 is extracted from WordNet (Miller, 1995), a database with lexical relations between words. FB15k is extracted from Freebase (Bollacker et al., 2008), a large-scale knowledge graph with general knowledge facts. YAGO3-10 is a subset of YAGO3 (Mahdisoltani et al., 2015), containing the most common 37 relations and all entities with more than 10 linked relations to ensure sufficient knowledge contents. Most of the triples in YAGO3-10 are general knowledge facts of people, such as citizenship, gender, profession and marital status.

Table 4: Statistics of five standard benchmarks.

Dataset	#Entity	#Relation	#Training	#Validation	#Test
WN18RR	40,943	11	86,835	3,034	3,134
FB15k-237	14,541	237	272,115	17,535	20,466
YAGO3-10	123,182	37	1,079,040	5,000	5,000

C IMPLEMENTATION DETAILS

All experiments are conducted in Python 3.8 and PyTorch 1.8.0, executed on the Ubuntu 24.04.1 LTS server equipped with AMD EPYC 7642 48-Core CPU @ 2.30GHz, 128 GB of RAM, and NVIDIA RTX 4090 GPUs (24 GB), running CUDA 12.9 and driver version 575.64.03.

Evaluation: We follow the filtered ranking protocol (Bordes et al., 2013; Li et al., 2024; 2022) for all KGE methods during accuracy evaluation. For a test triple (h, r, t), we will rank it against all negative triplets (h, r, t') or (h', r, t) that do not appear in the knowledge graph.

The mean reciprocal rank (MRR) and HITS at N (H@N) metrics are defined as:

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}, \quad \text{Hits@K} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\text{rank}_i \leq K), \tag{38}$$

where rank_i is the rank position of the correct one and $\mathbb{I}$ is the indicator function. Larger MRR and Hits@K indicate better model accuracy.

For CP coverage performance, the empirical coverage is defined as:

$$\text{Empirical Coverage} = \frac{1}{|\mathcal{D}_{test}|}\sum_{(q_i,y_i)\in\mathcal{D}_{test}}\mathbf{1}\{y_i \in \mathcal{C}(q_i)\}. \tag{39}$$

The average prediction set size is defined as:

$$\text{Average Set Size} = \frac{1}{|\mathcal{D}_{test}|}\sum_{(q_i,y_i)\in\mathcal{D}_{test}}|\mathcal{C}(q_i)|. \tag{40}$$

Smaller Size indicates better CP coverage performance.

For fairness performance, the maximum set size disparity across groups is defined as:

$$\Delta = \max_{a,b\in A}\left\{\left|\mathbb{E}\left[|\mathcal{C}(q)| \mid q \in G_a\right] - \mathbb{E}\left[|\mathcal{C}(q)| \mid q \in G_b\right]\right|\right\}. \tag{41}$$

Here, G_a and G_b are test query groups as shown in KGCP paper Section 6.2 (Zhu et al., 2026) where each group contains similar query difficulties. Lower Δ indicates better fairness.

Hyperparameters: We fine-tune the hyperparameters on the validation datasets. The tuning of specific hyperparameters (*e.g.,* batch sizes, learning rates, etc.) follows the original papers of each base KGE model. For example, the hyperparameters of GoldE (Li et al., 2024) are tuned by the random search (Bergstra & Bengio, 2012), including batch size b, self-adversarial negative sampling temperature γ, β_E, β_R, loss margin α, learning rate lr, dimension d, low-rank k_E, k_R that are smaller than dimension d, number of elliptic component spaces $m_{\mathbb{P}}$, and number of hyperbolic component spaces $m_{\mathbb{Q}}$. The hyperparameter search spaces are shown in Table 5. The nonconformity functions are chosen from NegScore, Softmax and Minmax as described in KGCP (Zhu et al., 2025).

Table 5: Hyperparameter search space.

Hyperparameter	Search Space	Type
b	$\{500, 800, 1000, 1500, 2000\}$	Choice
γ, β_E, β_R	$[0.5, 2.0]$	Range
α	$\{6, 8, 10, 12, 16, 20, 24, 28\}$	Choice
lr	$\{0.0001, 0.0005, 0.001, 0.003\}$	Choice
d	$\{32, 128, 256, 1000\}$	Choice
k_E, k_R	$\{28, 30, 120, 240, 800, 960\}$	Choice
$m_{\mathbb{P}}, m_{\mathbb{Q}}$	$\{2, 4, 8, 10, 15, 20, 25, 30\}$	Choice