DTDES-KGE: Dual-Teacher Knowledge Distillation with Distinct Embedding Spaces for Knowledge Graph Embeddings

Bofan Wei, Hongyuan Xu, Yuhang Niu, Jiarui Ren, Yanlong Wen*, Xiaojie Yuan
College of Computer Science, Nankai University, Tianjin, China
{weibofan, xuhongyuan, niuyuhang, renjiarui}@dbis.nankai.edu.cn
{wenyl, yuanxj}@nankai.edu.cn

Abstract

Knowledge distillation for knowledge graph embedding (KGE) models effectively compresses KGE models by reducing their embedding dimensions. While existing methods distill knowledge from a high-dimensional teacher to a low-dimensional student, they typically rely on a single teacher embedding space, thereby overlooking valuable complementary knowledge from teachers in distinct embedding spaces. This paper introduces DTDES-KGE, a novel knowledge distillation framework that significantly enhances distillation performance by leveraging dual teachers in distinct embedding spaces. To overcome the challenge of spatial heterogeneity when integrating knowledge from dual teachers, we propose a spatial compatibility module for reconciliation. Additionally, we introduce a student-aware knowledge fusion mechanism to fuse the knowledge from dual teachers dynamically. Extensive experiments on two real-world datasets validate the effectiveness of DTDES-KGE.

1 Introduction

Knowledge graphs (KGs) represent factual knowledge with diverse structures as triples. Their efficacy in organizing complex information has driven widespread adoption in various applications such as natural language processing (Ma et al., 2025), question answering (Wu et al., 2024b), and recommendation systems (Lai et al., 2024).

Knowledge Graph Embedding (KGE) models are pivotal for leveraging KGs, as they learn embeddings for entities and relations, facilitating the integration of KGs into numerous downstream applications (Zhang et al., 2024; Mohamed et al., 2021; Choudhary et al., 2021). While diverse KGE models have emerged, where Euclidean space models prove adept at capturing chain structures and hyperbolic space models excel in representing hierarchies in KGs (Liu et al., 2024; Cao et al., 2022; Li et al., 2025), achieving state-of-the-art performance often necessitates large embedding dimensions. This subsequently leads to prohibitive storage demands and increased inference latency, which severely limits their practical deployment, particularly in resource-constrained environments.

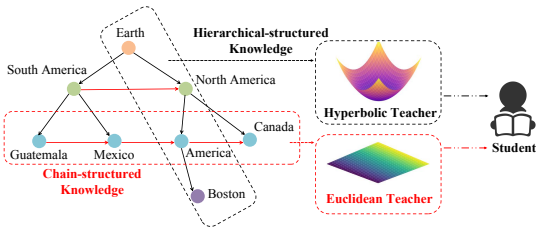

Figure 1: An example of leveraging dual teachers from distinct embedding space for knowledge distillation.

Knowledge distillation (KD) offers a promising avenue for compressing these large KGE models, where a low-dimensional student model learns from a pre-trained, high-capacity teacher model by mimicking its output and embedding structure (Zhu et al., 2022). However, existing knowledge distillation approaches for KGEs predominantly rely on teacher models trained within a single embedding space (Wang et al., 2021a; Zhu et al., 2022; Liu et al., 2023; Xu et al., 2024). This overlooks the rich, complementary knowledge that teachers from different embedding spaces, each of which excels at capturing distinct structural properties of KGs, could collectively provide. Recognizing this untapped potential, and as illustrated in Figure1, we propose to utilize dual teachers, one from Euclidean space and another from hyperbolic space, to furnish a more comprehensive knowledge base for the student model.

Despite the considerable promise of fusing knowledge from Euclidean and hyperbolic KGE models (Wang et al., 2021b), a critical challenge emerges: inter-space heterogeneity (Cao et al.,

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 4875–4887
November 4-9, 2025 ©2025 Association for Computational Linguistics

2022; Wang et al., 2024). The distinct geometric properties and scoring mechanisms inherent to these disparate spaces result in heterogeneous outputs and embedding structures. This fundamental incompatibility severely impedes effective knowledge fusion, preventing the student model from optimally learning from the combined wisdom of cross-space teachers.

This paper investigates the pivotal research question: *How can we effectively leverage dual teachers from distinct embedding spaces for KGE model distillation?* To this end, we introduce DTDES-KGE, a novel **D**ual **T**eachers with **D**istinct **E**mbedding **S**paces for **K**nowledge **G**raph **E**mbeddings distillation framework. DTDES-KGE is designed to first reconcile spatial disparities and then intelligently fuse knowledge from dual teachers. Specifically, to address the challenge of spatial heterogeneity, DTDES-KGE incorporates a spatial compatibility module. This module employs optimal distribution selection to harmonize the heterogeneous outputs from the dual teachers and utilizes a pre-distillation phase to align their disparate embedding structures, thereby fostering better cross-space knowledge transfer. Furthermore, to intelligently integrate the reconciled knowledge, DTDES-KGE employs a student-aware knowledge fusion mechanism. This mechanism dynamically assigns adaptive weights to the dual teachers throughout the distillation process. It considers prior knowledge about teacher specializations, the knowledge divergence between teachers, and the student's current grasp of different knowledge facets, ensuring an optimal blend of guidance.

We conduct comprehensive experiments on two real-world benchmark datasets. The results demonstrate that DTDES-KGE significantly enhances the performance of low-dimensional KGE models, enabling them to achieve results comparable or even superior to leading, larger KGE models and existing distillation techniques.

Our main contributions are summarized as follows: (i) We propose DTDES-KGE, a knowledge distillation framework for KGE models utilizing dual teachers with distinct embedding spaces. (ii) We design a dedicated spatial compatibility module to mitigate inter-space heterogeneity and a student-aware knowledge fusion mechanism to dynamically and judiciously integrate complementary knowledge from these diverse teachers. (iii) Experiment results show that DTDES-KGE achieves comparable or superior performance to leading approaches on two datasets.

2 Related Work

2.1 Knowledge Graph Embeddings

KGE models can be divided into three groups based on the curvature of their embedding spaces: Euclidean-space models (zero curvature), hyperbolic space models (non-zero curvature), and mixed-curvature models. For Euclidean-space KGE models, some researchers focus on treating relations as translations from head entity embedding to the tail entity embedding (Bordes et al., 2013; Sun et al., 2019; Zhang et al., 2019, 2020; Ge et al., 2023; Niu et al., 2022), whereas others focus on representing relations as linear transformations applied to the entity embeddings (Yang et al., 2015; Nickel et al., 2011; Trouillon et al., 2016). For hyperbolic KGE models, early studies explored the Poincaré ball model (Balazevic et al., 2019; Kolyvakis et al., 2019; Chami et al., 2020; Pan and Wang, 2021). Recent efforts motivated by the potential for richer semantic representation, have explored modeling within ultrahyperbolic manifold spaces or utilized the Lorentz model for relationship modeling (Xiong et al., 2022; Chen et al., 2022; Fan et al., 2024). To further enhance the performance of KGE models, recent research explores modeling in multiple spaces with varying curvatures and integrates representations from different spaces with a well-designed strategy (Cao et al., 2022; Wang et al., 2021b; Liu et al., 2024; Li et al., 2025).

2.2 Knowledge Distillation for KGE Models

Knowledge distillation, introduced by Hinton (Hinton, 2015), enhances a smaller student model by transferring knowledge from a larger teacher model. MulDE (Wang et al., 2021a) first applied KD to KGE models, using four hyperbolic KGEs as teachers. Subsequent approaches include DualDE (Zhu et al., 2022), proposing a two-stage framework that models teacher-student dual influence; IterDE (Liu et al., 2023), employing an iterative method to reduce student-teacher performance gaps; and SKDE (Xu et al., 2024), exploring self-distillation for low-dimensional KGEs. However, existing KGE distillation methods predominantly use teacher models from a single embedding space. This practice transfers only one type of knowledge, thereby limiting student model performance.

3 Method

This section first outlines KGE model distillation preliminaries (§3.1). As shown in Figure 2, our framework has three key components: a spatial compatibility module to reconcile heterogeneous dual-teacher knowledge (§3.2); a student-aware mechanism to dynamically fuse this reconciled knowledge (§3.3); and a combined strategy integrating logits-based (§3.4) and feature-level distillation (§3.5) into the overall objective (§3.6).

3.1 Preliminaries

Given a set of entities E and a set of relations R, a knowledge graph $G = \{(h, r, t)|h, t \in E, r \in R\}$ is a collection of factual triples. KGE models learn entity and relation representations by training on observed positive triples T and negative triples T^-, which are generated by randomly corrupting heads or tails of positive triples.

KGE models use different scoring functions $f_r(h, t)$ to measure the plausibility of triples. Inspired by (Wu et al., 2024a), the score function of the student model is defined as:

$$f_r(e_h, e_t) = exp(\frac{e_{hr}^T e_t}{\tau_f \|e_{hr}\|_2 \|e_t\|_2}) + exp(\frac{1}{\tau_f}(\frac{e_{hr}^T e_t}{\|e_{hr}\|_2 \|e_t\|_2})^3), \quad (1)$$

where τ_f is the temperature hyperparameter. e_{hr} is the query embedding obtained by transforming the concatenation of the head and relation embeddings e_h, e_r through a linear layer.

Usually, a margin-based loss (Sun et al., 2019) is used to optimize the KGE model:

$$L_H = -\log \sigma(f_r(h, t) - \gamma) - \sum_{i=1}^{n} \frac{1}{n} \log \sigma(\gamma - f_r(h^{'}, t^{'})), \quad (2)$$

where n denotes the number of negative samples, σ denotes the sigmoid function, and γ is a hyperparameter controlling model's learning difficulty.

A widely-adopted logits-based knowledge distillation approach utilizes Kullback-Leibler (KL) divergence as a soft loss to guide the student model to mimic the teacher's output:

$$L_{KD} = \frac{1}{n} \sum \tau_s^2 KL(\sigma^{'}(P_t/\tau_s), \sigma^{'}(P_s/\tau_s)), \quad (3)$$

where P_t and P_s denote the output of the teacher model and the student model, $\sigma^{'}$ denotes the softmax function, and τ_s represents the temperature.

3.2 Spatial Compatibility Module

This section introduces our spatial compatibility module, which mitigates heterogeneous knowledge from dual teachers, thereby enabling effective cross-space knowledge fusion. To achieve this, it employs optimal distribution selection and a pre-distillation phase to reconcile their outputs and embedding structures.

Optimal Distribution Selection. To fully leverage the rich, nuanced information contained within the teacher models' logits and to facilitate their direct weighted aggregation, we first address the inherent distributional heterogeneity of their outputs. To this end, we introduce a dynamic mechanism that learns a query-specific target distribution, denoted as $\mathcal{D}^*$. By mapping the outputs of both teachers to this unified distribution, we harmonize their scores, rendering them directly comparable and enabling a more principled fusion.

Specifically, the parameters of the target distribution $\mathcal{D}^*$ are dynamically learned from key statistical properties of the two teacher distributions $(\mathcal{D}_1, \mathcal{D}_2)$. We construct a feature set comprising their individual means (μ_1, μ_2) and standard deviations (σ_1, σ_2), along with their relational statistics—the mean difference $(\mu_1 - \mu_2)$ and standard deviation ratio (σ_1/σ_2). This feature vector is then input to a linear layer, producing an output vector $\mathbf{y}$ that directly specifies the parameters of the optimal distribution $\mathcal{D}^*$:

$$\mu^* = \mathbf{y}[0], \quad (4)$$

$$\sigma^* = \text{softplus}(\mathbf{y}[1]), \quad (5)$$

where μ^* and σ^* represent the target mean and standard deviation. Finally, the raw logits from the dual teachers, $\mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^n$, are normalized and re-projected to align with $\mathcal{D}^*$, yielding the reconciled scores $\mathbf{s}_1^*, \mathbf{s}_2^*$. This element-wise transformation is formulated as:

$$\mathbf{s}_i^*[j] = \frac{\mathbf{s}_i[j] - \mu_i}{\sigma_i} \cdot \sigma^* + \mu^* \quad j = 1, \ldots, n \quad (6)$$

Pre-distillation Phase. Recent studies show that using teachers' embedding structure as hint knowledge benefits distillation (Hao et al., 2024; Li et al., 2024; Zheng et al., 2025). However, dual teacher models with heterogeneous embedding spaces pose a significant challenge to directly leveraging their embedding structures as hint knowledge. This heterogeneity implies that properties of different KGE

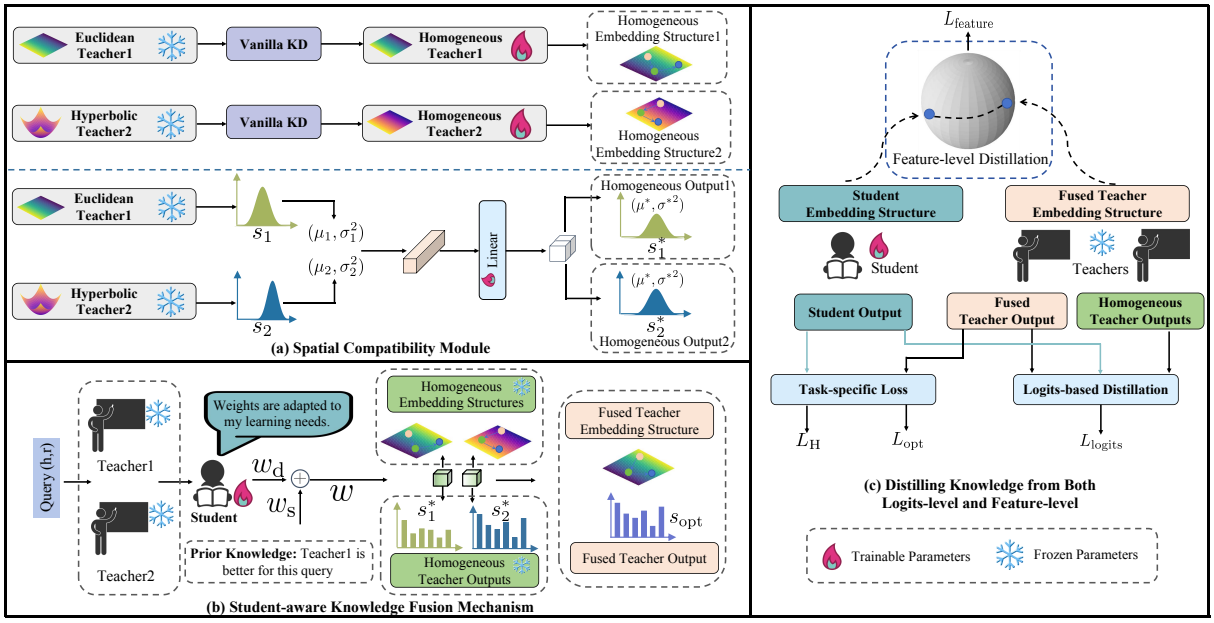

Figure 2: The overall architecture of DTDES-KGE. "Vanilla KD" indicates vanilla knowledge distillation approach utilizing KL divergence introduced by (Hinton, 2015).

models such as distance, angles and transformations are defined differently.

To overcome this challenge, we introduce a pre-distillation phase. The core idea is to transform the heterogeneous embeddings into a unified, homogeneous representation. This is achieved by employing the simple but effective vanilla knowledge distillation approach utilizing KL divergence introduced by introduced by (Hinton, 2015). Specifically, each of the heterogeneous teacher models is independently distilled into a homogeneous KGE model. Through this process, the rich knowledge of each heterogeneous teacher is largely transferred to the homogeneous model, meanwhile significantly mitigating the heterogeneity of the dual teachers' embedding structures, thereby reducing student learning complexity. We select RotatE (Sun et al., 2019) as the homogeneous KGE model for the pre-distillation phase. The influence of using different homogeneous KGE models for this phase is further investigated, and experimental findings are discussed in Section 4.5.

3.3 Student-aware Knowledge Fusion Mechanism

This section details our student-aware knowledge fusion mechanism, designed to assign query-specific fusion weights to dual teachers. This approach addresses their varying proficiency across queries and adapts to student learning progress. Weights combine a static component $\mathbf{w}_\text{s}$ according to prior knowledge with a dynamic component w_d determined by teacher performance and student learning state. To simplify representation, let $\mathbf{s}_1^*$ and $\mathbf{s}_2^*$ be the outputs under optimal distribution obtained from teachers in Euclidean and hyperbolic spaces, respectively.

We first utilize Krackhardt hierarchy score $K(r)$ and estimated curvature $C(r)$ of each relation r as prior knowledge. Detailed calculations for these metrics are provided in (Chami et al., 2020). Given that Euclidean space is typically better suited for modeling chain-structured knowledge while hyperbolic space excels at modeling hierarchical-structured knowledge (Liu et al., 2024), the static weight for the hyperbolic teacher is increased for relations with strong hierarchical indicators (higher $K(r)$ and lower $C(r)$), and vice versa for the euclidean teacher.

For each relation r, a static weight $\mathbf{w}_\text{s} = [w_1^s, w_2^s]$ is computed by averaging contributions $\mathbf{c}_\text{C}(r)$ and $\mathbf{c}_\text{K}(r)$. These contributions, $\mathbf{c}_\text{C}(r), \mathbf{c}_\text{K}(r) \in \mathbb{R}^2$, are derived from metrics $C(r)$ and $K(r)$ by comparing them to thresholds C_thres and K_thres:

$$\mathbf{c}_\text{C}(r) = \begin{cases} [1 - C_\text{opt},\ C_\text{opt}], & \text{if } C(r) < C_\text{thres} \\ [C_\text{opt},\ 1 - C_\text{opt}], & \text{if } C(r) \geq C_\text{thres} \end{cases} \tag{7}$$

$$\mathbf{c}_\text{K}(r) = \begin{cases} [1 - K_\text{opt},\ K_\text{opt}], & \text{if } K(r) \geq K_\text{thres} \\ [K_\text{opt},\ 1 - K_\text{opt}], & \text{if } K(r) < K_\text{thres} \end{cases} \tag{8}$$

where $C_{\text{opt}}, K_{\text{opt}}$ are hyperparameters. The static weight w_s is then calculated as follows:

$$\mathbf{w}_{\text{s}} = \frac{\mathbf{c}_{\text{C}}(r) + \mathbf{c}_{\text{K}}(r)}{2}, \tag{9}$$

guiding fusion towards the teacher space geometrically aligned with relation r's structure.

Dynamic weights $\mathbf{w}_{\text{d}} = [w_1^d, w_2^d]$ are determined based on query-specific features extracted from the dual teachers and the student. Teacher features $\mathbf{f}_{\text{tea}}$ are extracted from output scores $\mathbf{s}_1^*$ and $\mathbf{s}_2^*$, encompassing positive sample probabilities indicating teacher confidence and KL divergences $D_{\text{KL}}(\text{softmax}(\mathbf{s}_1^*)||\text{softmax}(\mathbf{s}_2^*))$ and $D_{\text{KL}}(\text{softmax}(\mathbf{s}_2^*)||\text{softmax}(\mathbf{s}_1^*))$ measuring the degree of divergence in the dual teachers' outputs. Student features $\mathbf{f}_{\text{stu}}$ are extracted from embeddings $\mathbf{e}_h, \mathbf{e}_r$ and output scores $\mathbf{s}_{\text{stu}}$ of the student model. They comprise the student's embeddings, representing embedding structure of the student model, the positive sample probability from $\mathbf{s}_{\text{stu}}$, and KL divergences $D_{\text{KL}}(\text{softmax}(\mathbf{s}_{\text{stu}})||\text{softmax}(\mathbf{s}_i^*))$ for $i = 1, 2$, reflecting the student's learning state and distance from mastering each teacher's knowledge.

Entity embeddings $\mathbf{e}_h$ and relation embeddings $\mathbf{e}_r$ are transformed by linear layers $\mathcal{T}_e$ and $\mathcal{T}_r$ into latent representations $\mathbf{v}_h = \mathcal{T}_e(\mathbf{e}_h)$ and $\mathbf{v}_r = \mathcal{T}_r(\mathbf{e}_r)$, respectively. Similarly, features $\mathbf{f}_{\text{tea}}$ and $\mathbf{f}_{\text{stu}}$ are processed by linear layers $\mathcal{T}_{\text{tea}}$ and $\mathcal{T}_{\text{stu}}$ to yield latent representations $\mathbf{v}_{\text{tea}} = \mathcal{T}_{\text{tea}}(\mathbf{f}_{\text{tea}})$ and $\mathbf{v}_{\text{stu}} = \mathcal{T}_{\text{stu}}(\mathbf{f}_{\text{stu}})$, respectively.

These four transformed vectors are then concatenated as input $\mathbf{z} = [\mathbf{v}_h; \mathbf{v}_r; \mathbf{v}_{\text{tea}}; \mathbf{v}_{\text{stu}}] \in \mathbb{R}^{4 \cdot d_{\text{h}}}$. The dynamic weights $\mathbf{w}_{\text{dyn}}$ is calculated as follows:

$$\mathbf{w}_{\text{d}} = \sigma^{'}(W^{2 \times d_h}(GELU(W^{d_h \times 4d_h} z))). \tag{10}$$

The final fusion weights $\mathbf{w} = [w_1, w_2]$ are computed as a weighted combination of the static wieghts and the dynamic weights :

$$\mathbf{w} = \frac{p_s \mathbf{w}_{\text{s}} + \mathbf{w}_{\text{d}}}{1 + p_s}, \tag{11}$$

where p_w is a hyperparameter controlling the influence of the prior knowledge.

3.4 Logits-based Distillation

The fused teacher output $\mathbf{s}_{\text{opt}}$ is a weighted sum of s_1^* and s_2^*:

$$\mathbf{s}_{\text{opt}} = w_1 \mathbf{s}_1^* + w_2 \mathbf{s}_2^* \tag{12}$$

To ensure the accuracy of the fused teacher output, we utilize formulate 2 to calculated the hard loss L_{opt} of the fused teacher output . The logits-based soft loss L_{logtis} is calculated as follows:

$$L_{logtis} = \lambda_1 L_{soft}^1 + \lambda_2 L_{soft}^2 + \lambda_{opt} L_{soft}^{opt}, \tag{13}$$

where L_{soft}^1, L_{soft}^2 and L_{soft}^{opt} are soft losses calculated with formula 3 with teacher output s_1^*, s_2^*, s_{opt} and student output s. The weights $\lambda_1, \lambda_2, \lambda_{opt}$ are dynamically adjusted based on the training epoch E. Initially, higher weights are assigned to L_{soft}^1 and L_{soft}^2 to enable the student model to acquire foundational knowledge directly from dual teachers. As training progresses, the weights gradually shift towards L_{soft}^{opt}. This encourages the model to learn the higher-quality, more complex fused knowledge.

3.5 Feature-level Distillation

We utilize InfoNCE (Gutmann and Hyvärinen, 2010) as our feature-level distillation loss function:

$$L_{feature} = \log \frac{e^{\theta(\widetilde{e_t}, e_s^+)/\tau_c}}{e^{\theta(\widetilde{e_t}, e_s^+)/\tau_c} + \sum e^{\theta(\widetilde{e_t}, e_s^-)/\tau_c}}, \tag{14}$$

where $\theta(\cdot, \cdot)$ is the cosine function and τ_c is the temperature hyperparameter, and $\widetilde{e_t}$ is the fused embedding structure from two homogeneous teachers. $\widetilde{e_t}$ is calculated from the weighted sum $\overline{e_t}$:

$$\overline{e_t} = w_1 e_t^1 + w_2 e_t^2, \tag{15}$$

where e_t^1 and e_t^2 are embeddings from dual teachers. This sum is then transformed to match the student dimension:

$$\widetilde{e_t} = W^{d \times d}(\text{GELU}(W^{d \times D}\overline{e_t} + b)), \tag{16}$$

where W, b are learnable parameters. d and D are student and teacher embedding dimensions, respectively.

3.6 Overall Objective

The framework jointly minimizes task-related hard loss and soft loss for knowledge distillation:

$$L = L_H + \lambda L_S, \tag{17}$$

where λ is the weight of soft loss. The soft loss is calculated as follows:

$$L_S = L_{logits} + L_{opt} + \lambda_s L_{feature}, \tag{18}$$

where λ_s is the weight of distilling feature knowledge.

4 Experiments

4.1 Experimental Settings

Datasets. We conducted experiments on two real-world datasets: FB15k-237 and WN18RR. The statistics of two datasets are shown in Table 1.

Datasets	#entity	#relation	#train	#valid	#test
FB15k-237	14541	237	272115	17535	20466
WN18RR	40943	11	86835	3034	3134

Table 1: Dataset statistics. #entity and #relation are the number of entities and relations.

Baselines. We evaluate DTDES-KGE by comparing it with four different types of KGE models: Euclidean, hyperbolic, knowledge distillation, and mixed-curvature. For Euclidean models, we select TransE (Bordes et al., 2013), ComplEx (Trouillon et al., 2016), RotatE (Sun et al., 2019), HAKE (Zhang et al., 2020), and CompoundE (Ge et al., 2023). For hyperbolic models, we consider MuRP (Balazevic et al., 2019), AttH (Chami et al., 2020), UltraE (Xiong et al., 2022), and LorentzKG (Fan et al., 2024). For knowledge distillation models, we evaluate MulDE(Wang et al., 2021a), DualDE (Zhu et al., 2022), IterDE (Liu et al., 2023), and SKDE (Xu et al., 2024). For mixed-curvature models, we choose GIE(Cao et al., 2022), M^2GNN (Wang et al., 2021b), UniGE (Liu et al., 2024), and MRME (Li et al., 2025) as baseline models.

Metrics. KGE models are evaluated on link prediction. For each test triple (h, r, t), candidates are generated by replacing its head or tail with other entities. The correct triple's rank ($rank_t$) among these candidates is used to calculate Mean Reciprocal Rank (MRR) and $Hits@k$ ($k \in 1, 3, 10$) in the filtered setting (Sun et al., 2019).

4.2 Implementation Details

DTDES-KGE, implemented in PyTorch(Paszke et al., 2019), was optimized via grid search for learning rate, γ, and τ_f. The optimal hyperparameters are 0.005, 9, 0.6 for FB15k-237 and 0.005, 9, 1.0 for WN18RR. For negative sampling, we employ a query-aware sampling strategy, the details of which are provided in Appendix A. The teacher models we utilized include HAKE, RotatE, and LorentzKG. Appendices B and C detail hyperparameter selection and teacher model training. Experiments were accelerated on a single NVIDIA RTX 4090 GPU. Our code can be found here.

4.3 Main Result

Table 2 compares the performance of DTDES-KGE with various baseline methods. These results form the basis for the following discussion.

Q1. Can dual teachers from distinct embedding space enhance knowledge distillation? Yes. DTDES-KGE demonstrates enhanced knowledge distillation performance by leveraging dual teachers from distinct embedding spaces, outperforming existing KGE baselines on both datasets. Specifically, on FB15k-237, DTDES-KGE achieved an MRR of 40.5, surpassing the second-best baseline (38.4) by 5.5%. On WN18RR, its MRR of 51.5 exceeded the next leading model (50.2) by 2.6%. Notably, this strong performance was achieved with a low embedding dimension (32). These findings confirm that dual teachers from distinct spaces transfer more comprehensive knowledge to the student model, thereby enhancing distillation efficacy.

4.4 Ablation Studies

Table 3 shows ablation results for the DTDES-KGE. In this table, "-LKD" indicates without using outputs of dual teachers as logits-based knowledge, "-FKD" indicates without using embedding structures of dual teachers as feature-level knowledge, "-Fusion" indicates without fusion the outputs and embedding structures of dual teachers, and "ST" indicates using only a single teacher during distillation. The other variants represent the full framework with a specific component removed: "-SCM" (without the **S**patial **C**ompatibility **M**odule), "-ODS" (without the **O**ptimal **D**istribution **S**election), "-PP" (without the **P**re-distillation **P**hase), and "-SKFM" (without the **S**tudent-aware **K**nowledge **F**usion **M**echanism). Guided by the ablation results, we answer the following questions:

Q2. Can distilling both outputs and embedding structures from dual teachers benefit KGE model distillation? Yes, our ablation studies confirm this. Removing logits-based distillation (-LKD) led to a severe degradation in MRR. Furthermore, eliminating feature-level distillation (-FKD) or relying solely on a single teacher (ST) also resulted in a distinct performance reduction. These findings underscore that complementary guidance from dual teachers, incorporating both output logits and embedding structures, is crucial for effective KGE model distillation.

Q3. Do both reconciling the heterogeneity of dual teachers and employing a student-aware

	FB15k-237					WN18RR				
Model	MRR	H@1	H@3	H@10	Dim	MRR	H@1	H@3	H@10	Dim
TransE	28.0	17.7	32.1	48.0	200	22.7	3.5	38.6	50.6	180
ComplEx	25.7	16.5	29.3	44.3	200	43.2	39.6	45.2	50.0	230
RotatE	30.1	21.0	33.1	48.5	1024	47.3	43.2	48.8	55.3	1000
HAKE	34.6	25.0	38.1	54.2	1000	49.7	45.2	51.6	58.2	1000
CompoundE	35.7	26.4	39.3	54.5	300	49.1	45.0	50.8	57.6	300
MuRP	33.5	24.3	36.7	51.8	200	48.1	44.0	49.5	56.6	200
AttH	34.8	25.2	38.4	54.0	500	48.6	44.3	49.9	57.3	500
UltraE	36.8	27.6	40.0	56.3	400	50.1	45.0	51.5	59.2	400
LorentzKG	38.4	28.7	42.2	**57.9**	32	50.2	45.6	52.3	58.9	32
MulDE	32.3	23.7	32.7	47.7	32	45.4	41.1	47.8	55.7	32
DualDE	34.6	24.0	34.5	51.9	32	46.0	41.9	48.1	56.3	32
IterDE	37.4	27.5	39.3	53.8	32	47.3	42.4	49.8	57.1	32
SKDE	35.3	26.0	39.0	54.0	200	48.0	44.0	49.0	55.0	200
GIE	36.2	27.1	40.1	55.2	200	49.1	45.2	50.5	57.5	200
M^2GNN	36.2	27.5	39.8	56.5	200	48.5	44.4	49.8	57.2	200
UniGE	34.3	25.7	37.5	52.3	32	49.1	44.7	51.2	56.3	32
MRME	35.9	28.6	38.3	52.4	32	49.8	46.8	51.9	56.2	32
DTDES-KGE	**40.5**	**31.8**	**44.1**	**57.9**	32	**51.5**	**47.2**	**53.4**	**59.5**	32

Table 2: Experimental results comparing KGE models with different embedding dimensions on link prediction tasks over the FB15k-237 and WN18RR datasets.

knowledge fusion strategy improve distillation performance? Yes, both are crucial. Firstly, merely using dual teachers independently without fusion (-Fusion) significantly degraded performance, consistent with previous studies (Cao et al., 2022), thus underscoring the necessity of knowledge integration. Secondly, effective fusion demands heterogeneity reconciliation. Removing the spatial compatibility module (-SCM) caused a substantial performance drop, even with fusion attempts, indicating naive fusion is ineffective. This was further validated by performance reductions upon removing SCM's components: optimal distribution selection (-ODS) or the pre-distillation phase (-PP). Finally, the student-aware knowledge fusion mechanism (-SKFM) is also vital; its removal led to significant performance degradation, confirming its efficacy in adaptive knowledge combination. Details are in Appendix D.

Dataset	Setting	MRR	Hit@1	Hit@3	Hit@10
FB15k-237	DTDES-KGE	**40.5**	**31.8**	**44.1**	**57.9**
	-LKD	29.7	22.0	32.4	45.2
	-FKD	39.8	31.0	43.2	56.9
	ST	38.8	30.2	42.0	55.6
	-Fusion	38.9	30.1	42.4	56.4
	-SCM	38.7	29.7	42.4	56.2
	-ODS	39.1	30.0	42.5	56.2
	-PP	40.0	31.1	43.2	57.1
	-SKFM	39.4	30.5	42.8	56.3
WN18RR	DTDES-KGE	**51.5**	**47.2**	**53.4**	**59.5**
	-LKD	36.8	35.1	37.6	39.4
	-FKD	50.6	46.2	52.6	58.8
	ST	48.1	43.1	50.7	57.1
	-Fusion	49.5	46.0	51.1	56.8
	-SCM	49.7	45.8	51.0	57.1
	-ODS	49.9	45.7	51.7	57.6
	-PP	50.8	46.5	52.9	59.0
	-SKFM	50.3	45.8	50.6	57.2

Table 3: Ablation Results of DTDES-KGE

4.5 Further Exploration

Q4. Is optimal distribution selection a superior choice for reconciling output heterogeneity? Yes. Table 4 compares various distribution selection strategies. "Proj-to" maps one teacher model's output distribution onto another teacher's. 'Stand Norm" and "Min-Max Scaling" respectively standardize and apply Min-Max scaling to dual teachers' output distributions. "Softmax" converts scores to a probability distribution using the softmax function. Fixed distribution selection strategies fail to account for query-dependent variations in teacher distributions, thereby reducing distillation performance. In contrast, our proposed strategy dynamically adapts to these query-specific teacher distribution characteristics, yielding superior performance.

Dataset	Setting	MRR	Hit@1	Hit@3	Hit@10
FB15k-237	Ours	**40.5**	**31.8**	**44.1**	**57.9**
	Proj-to-1st	37.2	28.1	41.0	55.4
	Proj-to-2nd	38.1	29.0	41.6	56.0
	Stand Norm	37.4	28.7	40.7	54.6
	Min-Max Scaling	25.4	17.6	27.4	40.9
	Softmax	37.1	28.3	40.6	54.6
WN18RR	Ours	**51.5**	**47.2**	**53.4**	**59.5**
	Proj-to-1st	42.3	38.1	44.4	49.7
	Proj-to-2nd	44.0	38.5	47.4	53.2
	Stand Norm	41.7	36.4	44.8	50.5
	Min-Max Scaling	28.9	23.6	33.1	36.8
	Softmax	44.3	40.6	46.0	51.1

Table 4: Analyzing the impact of alternative algorithmic choices for optimal distribution selection.

Q5. Are homogeneous teacher models more effective for feature-level distillation? Yes. Figure 3 explores different settings for the embedding space and dimension of homogeneous teacher models during feature-level distillation. Experimental results indicate that employing the Euclidean space model RotatE as a homogeneous teacher yields superior performance compared to the hyperbolic space model AttH. The rationale behind this observation is that aligning the embedding space of homogeneous teacher models with the student's embedding space facilitates distillation effectiveness through reduced spatial mismatch. Regarding dimension, while sufficient capacity is necessary, excessively large dimensions offer limited gain. Conversely, insufficient dimensions degrade performance. As shown in Table 5, a noteworthy finding is that distilling both logits and features from the homogeneous teacher model leads to performance degradation compared to relying solely on features. We attribute this phenomenon to the pre-distillation phase. Although the transfer of logits from original teachers largely reduces embedding heterogeneity, this transfer is incomplete, thereby incurring knowledge loss.

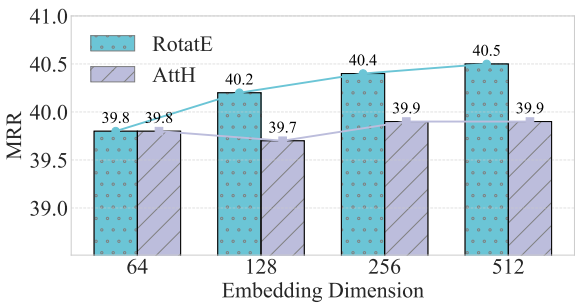

Figure 3: Impact of different homogeneous teacher model selected in euclidean and hyperbolic spaces on feature-level distillation.

Dataset	Setting	MRR	Hit@1	Hit@3	Hit@10
FB15k-237	only feature	**40.5**	**31.8**	**44.1**	**57.9**
	logits+feature	40.0	31.1	43.2	57.1
WN18RR	only feature	**51.5**	**47.2**	**53.4**	**59.5**
	logits+feature	50.8	46.5	52.9	59.0

Table 5: Impact of different settings for knowledge source on knowledge distillation.

Q6. Does DTDES-KGE Increase Training and Inference Time?

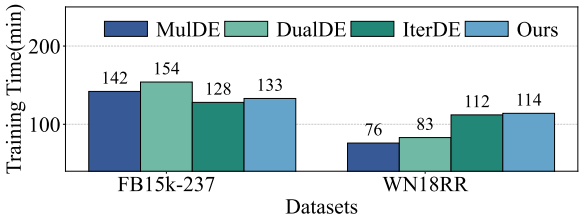

Figure 4: Traing Time of different KGE model distillation approach.

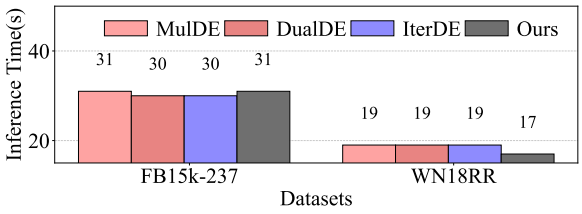

Figure 5: Inference Time of different KGE model distillation approach.

Figure 4 illustrates the training times of various KGE distillation models across two datasets. The results indicate that the training time of DTDES-KGE remains within expected limits, with the addition of auxiliary tasks not causing any notable increase in the convergence time. On the FB15K-237 dataset, DTDES-KGE exhibits a more efficient training time compared to both MulDE and DualDE. Although on the WN18RR dataset, the training time of DTDES-KGE is somewhat longer than the three baseline models, the increase is not substantial, remaining within an acceptable range. Figure 5 presents the inference times of the different KGE distillation methods on the test set. Due to the reduced embedding dimensions of the student models and the relatively simple design of the scoring functions, all four models demonstrate high inference performance. Each result is based on the average of three separate trials.

5 Conclusion

This paper introduced DTDES-KGE, a novel knowledge distillation framework leveraging dual teachers from distinct embedding spaces to enhance KGE distillation performance. It addresses teacher heterogeneity in output scores and embedding structures via a spatial compatibility module and dynamically fuses knowledge using a student-aware mechanism. Extensive experiments on two real-world datasets validated DTDES-KGE's effectiveness in improving student performance while compressing KGE models.

6 Limitations

We acknowledge that the scope of this study is primarily limited to utilizing translation-based KGE models as teachers. This may restrict the breadth of the findings. Future research should explore a wider variety of teacher architectures and evaluate the method's performance in these broader settings. When the content of a knowledge graph involves privacy risks, federated learning can be employed to address this issue.

7 Acknowledgments

We thank anonymous reviewers for their valuable comments. We also thank Jiaqi Ye, Ciyi Liu, Xin Li for their insightful suggestions and kind help. This research is supported by the National Natural Science Foundation of China (No.62372252,72342017).

References

Ivana Balazevic, Carl Allen, and Timothy Hospedales. 2019. Multi-relational poincaré graph embeddings. *NIPS*, 32.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *NIPS*, 26.

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. 2022. Geometry interaction knowledge graph embeddings. In *AAAI*, volume 36, pages 5521–5529.

Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. 2020. Low-dimensional hyperbolic knowledge graph embeddings. *ACL*.

Weize Chen, Xu Han, Yankai Lin, Hexu Zhao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2022. Fully hyperbolic neural networks. In *ACL*, pages 5672–5686.

Shivani Choudhary, Tarun Luthra, Ashima Mittal, and Rajat Singh. 2021. A survey of knowledge graph embedding and their applications. *arXiv preprint arXiv:2107.07842*.

Xiran Fan, Minghua Xu, Huiyuan Chen, Yuzhong Chen, Mahashweta Das, and Hao Yang. 2024. Enhancing hyperbolic knowledge graph embeddings via lorentz transformations. In *Findings of ACL*, pages 4575–4589.

Xiou Ge, Yun Cheng Wang, Bin Wang, and C-C Jay Kuo. 2023. Compounding geometric operations for knowledge graph completion. In *ACL*, pages 6947–6965.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR.

Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. 2024. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. *NIPS*, 36.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Prodromos Kolyvakis, Alexandros Kalousis, and Dimitris Kiritsis. 2019. Hyperkg: Hyperbolic knowledge graph embeddings for knowledge base completion. *arXiv preprint arXiv:1908.04895*.

Kai-Huang Lai, Zhe-Rui Yang, Pei-Yuan Lai, Chang-Dong Wang, Mohsen Guizani, and Min Chen. 2024. Knowledge-aware explainable reciprocal recommendation. In *AAAI*, volume 38, pages 8636–8644.

Linyu Li, Zhi Jin, Xuan Zhang, Haoran Duan, Jishu Wang, Zhengwei Tao, Haiyan Zhao, and Xiaofeng Zhu. 2025. Multi-view riemannian manifolds fusion enhancement for knowledge graph completion. *TKDE*.

Zheng Li, Xiang Li, Lingfeng Yang, Renjie Song, Jian Yang, and Zhigeng Pan. 2024. Dual teachers for self-knowledge distillation. *Pattern Recognition*, 151:110422.

Jiajun Liu, Peng Wang, Ziyu Shang, and Chenxiao Wu. 2023. Iterde: an iterative knowledge distillation framework for knowledge graph embeddings. In *AAAI*, volume 37, pages 4488–4496.

Yuhan Liu, Zelin Cao, Xing Gao, Ji Zhang, and Rui Yan. 2024. Bridging the space gap: Unifying geometry knowledge graph embedding with optimal transport. In *WWW*, pages 2128–2137.

Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, Huaren Qu, Cehao Yang, Jiaxin Mao, and Jian Guo. 2025. Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation. *ICLR*.

Sameh K Mohamed, Aayah Nounu, and Vít Nováček. 2021. Biological applications of knowledge graph embedding models. *Briefings in bioinformatics*, 22(2):1679–1693.

Maximilian Nickel, Volker Tresp, Hans-Peter Kriegel, et al. 2011. A three-way model for collective learning on multi-relational data. In *ICML*, volume 11, pages 3104482–3104584.

Guanglin Niu, Bo Li, Yongfei Zhang, and Shiliang Pu. 2022. Cake: A scalable commonsense-aware framework for multi-view knowledge graph completion. In *ACL*, pages 2867–2877.

Zhe Pan and Peng Wang. 2021. Hyperbolic hierarchy-aware knowledge graph embedding for link prediction. In *Findings of EMNLP 2021*, pages 2941–2948.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NIPS*.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *ICLR*.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *ICML*, pages 2071–2080.

Jiapu Wang, Zheng Cui, Boyue Wang, Shirui Pan, Junbin Gao, Baocai Yin, and Wen Gao. 2024. Ime: Integrating multi-curvature shared and specific embedding for temporal knowledge graph completion. In *WWW*, pages 1954–1962.

Kai Wang, Yu Liu, Qian Ma, and Quan Z Sheng. 2021a. Mulde: Multi-teacher knowledge distillation for low-dimensional knowledge graph embeddings. In *WWW*, pages 1716–1726.

Shen Wang, Xiaokai Wei, Cicero Nogueira Nogueira dos Santos, Zhiguo Wang, Ramesh Nallapati, Andrew Arnold, Bing Xiang, Philip S Yu, and Isabel F Cruz. 2021b. Mixed-curvature multi-relational graph neural network for knowledge graph completion. In *WWW*, pages 1761–1771.

Yihong Wu, Le Zhang, Fengran Mo, Tianyu Zhu, Weizhi Ma, and Jian-Yun Nie. 2024a. Unifying graph convolution and contrastive learning in collaborative filtering. In *KDD*, pages 3425–3436.

Yike Wu, Yi Huang, Nan Hu, Yuncheng Hua, Guilin Qi, Jiaoyan Chen, and Jeff Z Pan. 2024b. Cotkr: Chain-of-thought enhanced knowledge rewriting for complex knowledge graph question answering. *EMNLP*.

Bo Xiong, Shichao Zhu, Mojtaba Nayyeri, Chengjin Xu, Shirui Pan, Chuan Zhou, and Steffen Staab. 2022. Ultrahyperbolic knowledge graph embeddings. In *SIGKDD*, pages 2130–2139.

Haotian Xu, Yuhua Wang, and Jiahui Fan. 2024. Self-knowledge distillation for knowledge graph embedding. In *CoLing*, pages 14595–14605.

Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*.

Jin-Cheng Zhang, Azlan Mohd Zain, Kai-Qing Zhou, Xi Chen, and Ren-Min Zhang. 2024. A review of recommender systems based on knowledge graph embedding. *ESWA*, page 123876.

Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. 2019. Quaternion knowledge graph embeddings. *NIPS*, 32.

Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang. 2020. Learning hierarchy-aware knowledge graph embeddings for link prediction. In *AAAI*, volume 34, pages 3065–3072.

Xu Zheng, Yunhao Luo, Pengyuan Zhou, and Lin Wang. 2025. Distilling efficient vision transformers from cnns for semantic segmentation. *Pattern Recognition*, 158:111029.

Yushan Zhu, Wen Zhang, Mingyang Chen, Hui Chen, Xu Cheng, Wei Zhang, and Huajun Chen. 2022. Dualde: Dually distilling knowledge graph embedding for faster and cheaper reasoning. In *WSDM*, pages 1516–1524.

A Query-aware Negative Sampling Strategy

Given a query $(head, relation, ?)$, each teacher model i selects the top-100 highest-scoring entities from the entity set as candidates, excluding those present in the training set. We denote each selected set as C_i. Since a well-trained teacher typically captures semantic relationships between entities correctly, entities in C_i predominantly exhibit semantic proximity to the ground truth. Through learning the score distribution among these entities, the student model can achieve more efficient and effective acquisition of semantic relationships between entities. The candidate set from both teacher models are then merged, and duplicates are removed, resulting in C_{qas}.

The Query-aware Sampling Size n_{qas} is treated as a hyperparameter. During the negative sampling process, n_{qas} entities are randomly selected from C_{qas} to serve as Query-aware Sampling entities. Any additional negative entities are then selected

through random negative sampling, ensuring no overlap with the previously sampled entities.

B The selection of hyperparameters

The main hyperparameter settings for DTDES-KGE on the FB15k-237 and wn18rr datasets are summarized in Table 6. For key parameters, we performed a grid search over the specified ranges to find the optimal values.

Hyperparameter	Dataset	
	FB15k-237	wn18rr
γ	search:(6,9,12)	search:(6,9,12)
batch size	3000	3000
negative sample size n	500	500
learning rate	search:(0.005,0.02)	search:(0.005,0.02)
τ_f	search:(0.5,1.0,1.5)	search:(0.5,1.0,1.5)
τ_s	0.6	1.0
p_w	0.1	0.1
τ_c	1.0	1.0
$\lambda, \lambda_{opt}, \lambda_s$	[1,1,0.001]	[1,1,0.001]

Table 6: The selection of hyperparameters used in DTDES-KGE

C Settings For Training Teacher model

For the FB15k-237 dataset, we selected RotatE(Sun et al., 2019) (dimension: 1024) and LorentzKG(Fan et al., 2024) (dimension: 64) as teacher models. Similarly, for the WN18RR dataset, we employed HAKE(Zhang et al., 2020) (dimension: 1000) and LorentzKG (dimension: 100) as teacher models. The optimal hyperparameters are obtained from their respective original publications. RotatE, LorentzKG, and HAKE were selected as teacher models due to their demonstrated strong performance on FB15k-237 and wn18rr(Li et al., 2025; Ge et al., 2023), enabling them to provide rich and reliable teacher signals. Given that the source code for all teacher models we employed is publicly available, this greatly facilitates the reproducibility of our work by subsequent researchers. We train the teacher model using the official code from the original paper.

D Extended Results and Discussion

Q7. Can the student-aware weight assignment mechanism achieve superior performance? Table 7 presents experimental results evaluating the impact of various knowledge fusion strategies on DTDES-KGE. We compare different approaches: "equal weight" assigns equal weights to the two teachers; w/o L_{opt} denotes without using L_{opt} ;"w/o auxiliary" learns query-specific teacher weights without incorporating auxiliary information. Further ablations investigate the contribution of specific information sources: "w/o prior", "w/o tea_Infor", and "w/o stu_Infor" denote scenarios where prior knowledge, teacher model information, and student model information are excluded, respectively. The results demonstrate that prior knowledge and the information from both teacher and student models are valuable resources for determining query-specific fusion weights. Notably, student model information is particularly crucial, as it enables dynamic adjustment of these weights based on the student's current learning state.

Dataset	Setting	MRR	Hit@1	Hit@3	Hit@10
FB15k-237	DTDES-KGE	**40.5**	**31.8**	**44.1**	**57.9**
	equal weight	36.9	27.8	40.4	54.8
	w/o opt	39.3	30.4	42.8	56.0
	w/o L_{opt}	40.3	31.7	43.6	57.7
	w/o auxiliary	39.1	30.3	42.6	56.7
	w/o prior	39.8	31.0	43.4	57.3
	w/o tea_Infor	39.8	31.0	43.2	57.2
	w/o stu_Infor	38.9	30.4	42.1	56.3
WN18RR	DTDES-KGE	**51.5**	**47.2**	**53.4**	**59.5**
	equal weight	49.3	45.2	50.6	57.2
	w/o opt	50.2	45.7	52.1	58.1
	w/o L_{opt}	51.2	46.5	52.5	58.9
	w/o auxiliary	49.1	44.9	50.3	57.9
	w/o prior	50.3	46.0	52.2	58.6
	w/o tea_Infor	50.1	45.8	51.9	58.1
	w/o stu_Infor	49.2	44.4	50.5	57.8

Table 7: Impact of different knowledge fusion strategies of DTDES-KGE

Q8. Can better results be achieved by distilling directly from a mixed-curvature model? Is distilling from teacher models in the same space superior to distilling from teacher models in different spaces? No. "All Euclidean" and "All Hyperbolic" respectively represent the use of dual teachers that are both from Euclidean space and dual teachers that are both from hyperbolic space. Distillation is less effective when using teachers from a single embedding space due to their limited complementary knowledge compared to teachers from distinct spaces. Similarly, direct distillation from state-of-the-art mixed-curvature models proved less effective than distilling from teachers from different embedding spaces. This suggests that mixed-curvature models provide a complex, pre-fused signal that is difficult for students to learn, whereas distinct, specialized knowledge from single-curvature models allows the student to learn and integrate this knowledge more effectively.

Q9. Would employing more than two teachers further enhance performance? Not necessarily.

Dataset	Setting	MRR	Hit@1	Hit@3	Hit@10
FB15k-237	DTDES-KGE	**40.5**	**31.8**	**44.1**	**57.9**
	All Euclidean	34.3	26.2	38.7	55.0
	All Hyperbolic	39.2	30.3	41.8	55.8
	Distill GIE	33.1	24.2	37.4	51.3
	Distill MRME	29.7	21.7	32.3	45.7
WN18RR	DTDES-KGE	**51.5**	**47.2**	**53.4**	**59.5**
	All Euclidean	48.8	44.1	50.9	57.2
	All Hyperbolic	48.6	44.2	51.2	57.4
	Distill GIE	46.4	42.9	48.5	54.0
	Distill MRME	43.6	40.2	44.9	49.8

Table 8: Impact of using different teacher models on distillation results.

Dataset	**Model**	MRR	H@1	H@3	H@10
FB15k-237	Dual Teachers	**40.5**	**31.8**	**44.1**	**57.9**
	Triple Teachers	40.7	**31.8**	43.2	57.6
WN18RR	Dual Teachers	**51.5**	47.2	**53.4**	**59.5**
	Triple Teachers	51.3	**47.4**	53.1	59.0

Table 9: Performance comparison between our proposed dual-teacher framework and a triple-teacher configuration. The results indicate that simply increasing the number of teachers does not guarantee improved performance.

While leveraging multiple teachers is beneficial, our investigation reveals that the diversity of knowledge is more critical than the sheer quantity of teachers. To validate this, we conducted additional experiments comparing our proposed dual-teacher setup against a triple-teacher configuration, with the results presented in Table 9.

The results show that increasing the number of teachers to three yields no significant performance gains; on WN18RR, it even leads to a slight degradation. This finding supports our hypothesis that the primary benefit of multi-teacher distillation stems from fusing complementary knowledge. Our dual-teacher framework is designed to integrate knowledge from two geometrically distinct spaces (Euclidean and hyperbolic), which already captures a rich set of diverse structural information. Adding a third teacher, particularly one with similar geometric properties, likely introduces information redundancy and offers diminishing returns. Furthermore, an increased number of teachers complicates the optimization of our student-aware fusion mechanism, making it more challenging to assign optimal weights.

Therefore, our dual-teacher approach represents a well-balanced trade-off between knowledge diversity and model complexity, achieving state-of-the-art performance without incurring the additional costs and optimization challenges of a larger teacher ensemble.

Dataset	**Model**	MRR	H@1	H@3	H@10
FB15k-237	HAKE	0.948	0.918	0.976	0.989
	RotatE	0.945	0.915	0.974	0.987
	AttH	0.961	0.932	0.981	0.989
	LorentzKG	0.977	0.943	0.980	0.989
	DTDES-KGE	**0.984**	**0.955**	**0.983**	**0.991**
WN18RR	HAKE	0.829	0.754	0.867	0.985
	RotatE	0.814	0.754	0.822	0.957
	AttH	0.856	0.821	0.894	0.983
	LorentzKG	0.887	0.843	0.900	0.986
	DTDES-KGE	**0.901**	**0.868**	**0.912**	**0.991**

Table 10: Performance comparison on the relation prediction task. DTDES-KGE consistently outperforms all strong baselines on both datasets.

Q10. Is the effectiveness of DTDES-KGE generalizable to other tasks beyond link prediction? Yes. To validate the generalizability of our framework, we conducted additional experiments on the crucial task of *relation prediction*. We compare DTDES-KGE against several strong baseline models that are also capable of performing this task. The results, presented in Table 10, demonstrate the superior performance of our method.

On both the FB15k-237 and WN18RR datasets, DTDES-KGE significantly outperforms all selected baselines across all metrics, including MRR and Hits@1. For instance, on WN18RR, our method achieves an MRR of 0.901, surpassing the next-best baseline, LorentzKG, by a notable margin. These strong results on a distinct task provide robust evidence that the benefits of our dual-teacher distillation framework are not confined to link prediction, but are indeed generalizable.

Q11. Why specifically choose Euclidean and hyperbolic spaces for the dual teachers? Our choice is motivated by the principle that knowledge diversity, driven by fundamental geometric differences, is key to effective distillation. The distinction between zero-curvature (Euclidean) and non-zero curvature (hyperbolic) spaces represents a primary and functionally significant taxonomy in KGE . This pairing allows us to fuse knowledge from two highly complementary knowledge. While more granular model classifications exist within each curvature family (e.g., RotatE and HAKE in Euclidean space), we hypothesize that fusing models from within the same family offers limited complementary knowledge compared to fusing across different curvature families.

To empirically validate this, we conducted experiments with various teacher pairings, with the results presented in Table 11. The findings strongly

Teacher Combination	Geometry Family	MRR	H@1	H@3	H@10
RotatE + LorentzKG	Euc. + Hyper.	**40.5**	**31.8**	**44.1**	**57.9**
HAKE + Lorentz	Euc. + Hyper.	40.5	31.6	43.8	57.4
RotatE + Poincaré	Euc. + Hyper.	39.9	31.0	42.7	56.9
HAKE + Poincaré	Euc. + Hyper.	40.1	30.9	43.4	57.1
RotatE + HAKE	All Euclidean	34.3	26.2	38.7	55.0
MuRP + LorentzKG	All Hyperbolic	39.2	30.3	41.8	55.8

Table 11: Performance of DTDES-KGE with different teacher pairings on FB15k-237. Pairings that bridge Euclidean (Euc.) and Hyperbolic (Hyper.) geometries consistently outperform those within a single geometry family.

support our rationale. As shown, any pairing of a Euclidean-family model (RotatE, HAKE) with a hyperbolic-family model (LorentzKG, MuRP) consistently and significantly outperforms pairings where both teachers reside in the same geometric space (e.g., "All Euclidean"). This confirms that the primary benefit of our framework stems from bridging disparate geometric curvatures, rather than merely combining different architectures.

E Analysis of Scientific Artifacts Utilized in the Research Process

We evaluated various Knowledge Graph Embedding models using the extensively adopted FB15k-237 and WN18RR datasets. FB15k-237 is licensed under CC BY 4.0, while WN18RR operates under Apache 2.0 License Revision 97a0bb6b. Our implementation utilized the PyTorch framework, which is distributed under a BSD-style license. All conclusions drawn from existing literature are comprehensively attributed through rigorous citations. The utilization of these datasets and framework adheres strictly to their respective licensing terms. In developing our code, we referenced the official open-source implementations of RotatE, HAKE, and LorentzKG on GitHub. Notably, RotatE and HAKE's repositories are licensed under MIT, whereas LorentzKG's repository lacks an explicit licensing declaration.

F Statement

All references utilized in this paper were rigorously sourced from peer-reviewed academic journals and conferences, accessed through established scholarly databases. The code referenced was obtained from previously published works made publicly available by their original authors. The datasets employed in this research are well-established in the field and have been validated to ensure they do not compromise the ethical standards protecting animal or human subjects. The AI assistant's involvement in this paper was restricted to proofreading for spelling and grammatical accuracy.