Mixed-Curvature Multi-Modal Knowledge Graph Completion

Yuxiao Gao[1], Fuwei Zhang[1], Zhao Zhang[2*], Xiaoshuang Min[3], Fuzhen Zhuang[1,4*]

[1] Institute of Artificial In telligence, Beihang University, Beijing, China
[2] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[3] The Sixth Research Institute of China Electronics Corporation, Beijing, China
[4] Zhongguancun Laboratory, Beijing, China
{gaoyx, zhangfuwei}@buaa.edu.cn, zhangzhao2021@ict.ac.cn, minxiaoshuang@ncse.com.cn, zhuangfuzhen@buaa.edu.cn

Abstract

Multi-modal Knowledge Graph Completion (KGC), which aims to enrich knowledge graph embeddings by incorporating images and text as supplementary information alongside triplets, is an significant task in learning KGs. Existing multi-modal KGC methods mainly focus on modality-level fusion, neglecting the importance of modeling the complex structures, such as hierarchical and circular patterns. To address this, we propose a **M**ixed-**C**urvature multi-modal **K**nowledge **G**raph **C**ompletion method (MCKGC) that embeds the information into three single-curvature spaces, including hyperbolic space, hyperspherical space, and Euclidean space, and incorporates multi-modal information into a mixed space. Specifically, MCKGC consists of Modality Information Mixed-Curvature Module (MIMCM) and Progressive Fusion Module (PFM). To improve the expressive ability for different modalities, MIMCM introduces multi-modal information into three single-curvature spaces for interaction. Then, to extract useful information from different modalities and capture the complex structure from the geometric information, PFM implements a progressive fusion strategy by utilizing modality-level and space-level gates to adaptively incorporate the information from different spaces. Extensive experiments on three widely used benchmarks demonstrate the effectiveness of our method.

Introduction

In recent years, Knowledge graphs (KGs) have garnered increasing attentions and have been widely applied in many downstream tasks, such as information retrieval (Zhang et al. 2022a; Liu et al. 2018; Xiong, Callan, and Liu 2017), event forecasting (Zhang et al. 2024a, 2022b), recommendation systems (Koren, Bell, and Volinsky 2009; Yu et al. 2014; Guo et al. 2020; Chen et al. 2024b,a), common sense reasoning (Lin et al. 2019), and question answering (Huang et al. 2019). KGs are composed of large-scale structured triples of entities and relations. A triplet in KG can be defined as (h, r, t), where h, r, and t represents the subject (or head) entity, relation, and object (or tail) entity, respectively. However, existing KGs are often incomplete, leading to performance deficiencies in downstream applications. To address this issue, Knowledge Graph Completion (KGC) techniques aim to represent entities and relations in a low-dimensional space, and inference and complete missing values in incomplete triplets.

Multi-modal KGC, as an extension of traditional KGC, aims to enrich the embedding of knowledge graphs by using images and text as additional information to triplets, thereby accomplishing tasks such as link prediction in a more comprehensive manner. Specifically, multi-modal KGC projects the entities and relations into a latent space. It then uses image, text, and the inherent structural information of KG itself to learn low-dimensional vector representations for entities and relations.

Recent researches on multi-modal KGC have made significant strides, with most studies focusing on fusing different modalities. For example, LAFA (Shang et al. 2024) emphasizes selective information integration for multi-modal KG by designing a link-aware fusion module. OTKGE (Cao et al. 2022b) addresses the issue of spatial heterogeneity in multi-modal fusion by conceptualizing it as an optimal transport plan that aligns different modal embeddings into a unified space.

However, most of these methods embed the information in Euclidean space, which might face some limitations. Traditional KGC models like DistMult (Yang et al. 2014), ComplEx (Trouillon et al. 2016), and TuckER (Balažević, Allen, and Hospedales 2019)—which typically utilize Euclidean space during inference—also face limitations. Figure 1 introduces three different structure in multi-modal KGs, including chain, hierarchical, and circular structures. The hierarchical and the circular structures consist of complex connections between different entities, which increases the difficulty of modeling multi-modal KGs. Thus, relying solely on a single-space model to encapsulate the intricate structure of a multi-modal KG is a formidable task.

To address the aforementioned challenges, we propose a **M**ixed-**C**urvature multi-modal **K**nowledge **G**raph **C**ompletion method (MCKGC), which strategically embeds multi-modal information into a mixed space to enhance the completion of multi-modal knowledge graphs. MCKGC model comprises two core modules: **Modality Information Mixed-Curvature Module** (MIMCM) and **Progressive Fusion Module** (PFM). MIMCM integrates data from three distinct modalities into three single-curvature spaces—hyperbolic, hyperspherical, and Euclidean. This

*Corresponding authors.

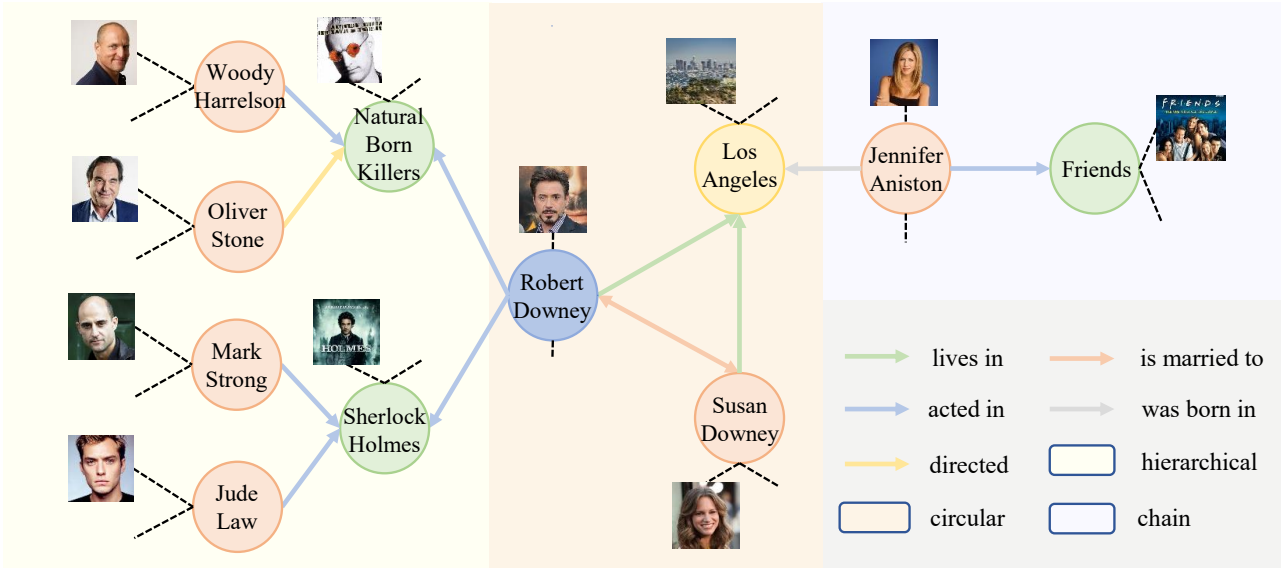


Figure 1: A simple example of a multi-modal KG, illustrating the complex structures. The yellow block illustrates the hierarchical structure, represented as a tree-like format that highlights the top-down hierarchy in professional interactions among actors. The orange block, denoting the circular structure, showcases the mutual collaboration among actors within the same film project, emphasizing the interconnectedness of their roles. Meanwhile, the purple block captures the chain structure, displaying a linear sequence of information specifically related to Robert Downey together.

module also features a learnable curvature, which adapts to the complex structure of KGs, thereby effectively capturing multi-modal information across the designated spaces and addressing various graph structures including hierarchical, chain-like, and circular forms. To promote more effective integration of multi-modal and multi-space information, PFM implements a strategic progressive fusion process. It first fuses information within individual modalities, and then it integrates information across different spaces. In particular, PFM employs both modality-level and space-level gates to adaptively incorporate the information from various modalities and spatial information. Here we summarize our contributions as follows:

- We introduce a new **M**ixed-**C**urvature multi-modal **K**nowledge **G**raph **C**ompletion method (MCKGC) that organizes information into various spaces to effectively capture the complex structures in multi-modal KGs.
- To enhance the integration of multi-modal and multi-space information, we present a progressive fusion module (PFM). PFM employs modality-level and space-level gates to adaptively calculates fusion weights, allowing the model to effectively extract relevant information across modalities and capture the complex geometric structure of multi-modal KGs.
- We conduct extensive experiments on three commonly used benchmarks. And the experimental results demonstrate the effectiveness of MCKGC.

Problem Definition

Typically, a Knowledge Graph (KG) can be represented as $\{\mathcal{E}, \mathcal{R}, \mathcal{T}\}$, where $\mathcal{E}$ and $\mathcal{R}$ denote the sets of entities (nodes) and relations (edges), respectively. And $\mathcal{T} = \{(h, r, t)|h, t \in \mathcal{E}, t \in \mathcal{R}\}$ represents the set of triplets in the KG. In multi-modal KGs, each entity can be represented as e_m ($m \in \mathcal{M} = \{\mathcal{S}, \mathcal{V}, \mathcal{T}\}$), where m includes $\mathcal{S}$, $\mathcal{V}$, and $\mathcal{T}$, which denotes structure, visual, and text modality respectively.

Multi-modal KGC is to learn multi-modal fusion embeddings for entities and embeddings for relations by projecting them into a continuous low-dimensional vector space. Based on these embeddings, the model can calculate the score of the each triplet (h, r, t). In the inference phase, the model can predict the missing entities in incomplete triplets, e.g., head entity $(?, r, t)$ and tail entity $(h, r, ?)$.

Methodology

In this section, we introduce the formal description and the details of our proposed framework **M**ixed-**C**urvature multi-modal **K**nowledge **G**raph **C**ompletion (MCKGC). Figure 2 illustrates the overall architecture of our proposed MCKGC. First, we introduce the multi-modal alignment module. Subsequently, we detail two critical components of MCKGC: the Modality Information Mixed-Curvature Module (MIMCM) and the Progressive Fusion Module (PFM). Finally, we describe the decoder and the loss function.

Multi-modal Alignment Module

To facilitate interaction of modalities in different spaces, different types of modalities need to be aligned at first. The initial image and text modality information come from the original dataset, extracted by pre-trained models VGG (Simonyan and Zisserman 2014) and BERT (Devlin 2018), re-

spectively. Next, we design a multi-modal alignment module to align different sizes of image and text information with structural information.

For each triplet in the multi-modal KG, the embedding of image information $\mathbf{e}_i$ and text information $\mathbf{e}_t$ are projected into the same dimension through a projection matrix. Subsequently, a Layer Normalization operation is conducted to stabilize training, as follows:

$$\mathbf{h}^i = \text{LayerNorm}(\mathbf{W}_i^T \mathbf{e}_i), \tag{1}$$

$$\mathbf{h}^t = \text{LayerNorm}(\mathbf{W}_t^T \mathbf{e}_t), \tag{2}$$

where $\mathbf{e}_i \in \mathbb{R}^{d_i}$ and $\mathbf{e}_t \in \mathbb{R}^{d_t}$ represent the initial image embedding and text embeddings, respectively. d_i and d_t denote the dimension size of the image embedding and text embeddings, respectively. $\mathbf{W}_i \in \mathbb{R}^{d_i \times d_n}$ and $\mathbf{W}_t \in \mathbb{R}^{d_t \times d_n}$ represent the projection matrices for image and text, respectively. d_n represents the dimension size of the aligned embeddings after projecting process. Since the structural information is learned from scratch during training, its dimension can be set to d_n, obviating the need for alignment.

Modality Information Mixed-Curvature Module

To capture the various complex structures in multi-modal KGs, we propose the Modality Information Mixed-Curvature Module (MIMCM) to obtain geometric interaction embeddings of different spaces for various modalities.

Benefiting from previous works (Balazevic, Allen, and Hospedales 2019; Chami et al. 2020; Cao et al. 2022a), given a point $\mathbf{x}$ in Euclidean space, we can now establish connections between hyperbolic space $\mathbb{H}$ and its tangent space $\mathcal{T}_{\mathbf{x}}\mathbb{H}$ through exponential map $\exp_{\mathbf{x}}^c : \mathcal{T}_{\mathbf{x}}\mathbb{H} \to \mathbb{H}$ and the logarithmic map $\log_{\mathbf{x}}^c : \mathbb{H} \to \mathcal{T}_{\mathbf{x}}\mathbb{H}$, where $\mathcal{T}_{\mathbf{x}}\mathbb{H}$ represents the tangent space, and $\mathbb{H}$ represents hyperbolic space. The formulas are as follows:

$$\exp_{\mathbf{x}}^c(\mathbf{v}) = \mathbf{x} \oplus_c \Big(\tanh\big(\sqrt{c}\frac{\lambda_{\mathbf{x}}\|\mathbf{v}\|}{2}\big)\frac{\mathbf{v}}{\sqrt{c}\|\mathbf{v}\|}\Big), \tag{3}$$

$$\log_{\mathbf{x}}^c(\mathbf{y}) = \frac{2}{\sqrt{c}\lambda_{\mathbf{x}}}\tanh^{-1}\big(\sqrt{c}\|-\mathbf{x} \oplus_c \mathbf{y}\|\big)\frac{-\mathbf{x} \oplus_c \mathbf{y}}{\|-\mathbf{x} \oplus_{\mathbf{c}} \mathbf{y}\|}, \tag{4}$$

where $\mathbf{x}, \mathbf{y} \in \mathbb{H}_n^{\mathbf{c}}$ and $\mathbf{v} \in \mathcal{T}\mathbb{H}_n^{\mathbf{c}}$, the tangent space $\mathcal{T}\mathbb{H}_n^{\mathbf{c}}$ is the space of tangent vectors at the point $\mathbf{x}$ on the hyperbolic space $\mathbb{H}_n^{\mathbf{c}}$, n represents the dimension of the manifold space and c represents curvature, which is positive in hyperbolic space $\mathbb{H}$ and negative in hyperspherical space $\mathbb{S}$. $\|\cdot\|$ denotes the Euclidean norm and $\oplus_{\mathbf{c}}$ represents Mobius addition:

$$\mathbf{x} \oplus_{\mathbf{c}} \mathbf{y} = \frac{\big(1 + 2c\langle\mathbf{x}, \mathbf{y}\rangle + c\|\mathbf{y}^2\|\big)\mathbf{x} + \big(1 - c\|\mathbf{x}^2\|\big)\mathbf{y}}{1 + 2c\langle\mathbf{x}, \mathbf{y}\rangle + c^2\|\mathbf{y}^2\|\|\mathbf{x}^2\|}. \tag{5}$$

Similarly, the above operations can also establish connections from tangent space to hyperspherical space.

Based on the aforementioned mapping formula, we can transform the aligned entity information of each modality into three different spaces to obtain geometric embeddings in each corresponding space. Taking the head entity of the structural modality $\mathbf{h}^s$ as an example. For Euclidean space, we obtain the embedding $\mathbf{E}_h^s = \mathbf{r}\mathbf{h}^s$. For hyperbolic space, we obtain the embedding $\mathbf{H}_h^s = \mathbf{r} \otimes_{c_1} \exp_0^{c_1}\mathbf{h}^s$. For hyperspherical space, we obtain the embedding $\mathbf{S}_h^s = \mathbf{r} \otimes_{c_2} \exp_0^{c_2}\mathbf{h}^s$, where $c_1(c_1 > 0)$, $c_2(c_2 < 0)$ represent the curvatures in hyperbolic space and spherical space, respectively. $\otimes_c$ represents the multiplication in hyperbolic space:

$$\mathbf{r} \otimes_c \mathbf{h} = \frac{1}{\sqrt{c}}\tanh\big(\mathbf{r}\tanh^{-1}(\sqrt{c}\|\mathbf{h}\|)\frac{\mathbf{h}}{\|\mathbf{h}\|}\big). \tag{6}$$

Next, to achieve interaction between modalities, we map the embeddings in the hyperbolic space and hyperspherical space back to the tangent space. For hyperbolic space, we obtain the processed embedding $\mathbf{H}'^s_h = \log_0^{c_1}\mathbf{H}_h^s$; similarly, for hyperspherical space, we obtain the processed embedding $\mathbf{S}'^s_h = \log_0^{c_2}\mathbf{S}_h^s$. Analogous operations are performed for the other two modalities to acquire embeddings across the three spaces. As a result, a total of nine embedding features are obtained. Consequently, we obtain a total of nine representations for the head entity.

Progressive Fusion Module

To better fuse information from various modalities and different spaces, we propose a progressive fusion module (PFM). First, we integrate information from different modalities into the same space to achieve modality unification. Taking hyperbolic space as an example, we input the three modality embeddings—structural embedding $\mathbf{H}'^s_h$, image embedding $\mathbf{H}'^i_h$, and text embedding $\mathbf{H}'^t_h$ obtained from MIMCM into a projection matrix to align the features of the three modalities after space interaction, as follows:

$$\mathbf{e}_{\text{H}}^s = \mathbf{H}'^s_h \mathbf{W}'_s, \tag{7}$$

$$\mathbf{e}_{\text{H}}^i = \mathbf{H}'^i_h \mathbf{W}'_i, \tag{8}$$

$$\mathbf{e}_{\text{H}}^t = \mathbf{H}'^t_h \mathbf{W}'_t, \tag{9}$$

where $\mathbf{W}'_j \in \mathbb{R}^{d_n \times d_n} (j \in \{s, i, t\})$ represents the learnable matrices.

Modality-level gate. Different triplets benefit from various modalities to varying extents. To address this variability, we propose a modality-level gate that achieves dynamic fusion between modalities in Equation (10), thereby maximizing the utilization of valuable information from each modality.

$$\mathbf{e}_{\text{H}} = \sum_j \alpha_j \mathbf{e}_{\text{H}}^j, j \in \{s, i, t\}, \tag{10}$$

where α_j is the weight computed by the modality-level gate, which can be calculated as follows:

$$\begin{aligned}\alpha_j &= \text{softmax}\Big(\mathbf{e}_{\text{H}}^j \mathbf{W}''_j\Big) \\ &= \frac{\exp\big(\mathbf{e}_{\text{H}}^j \mathbf{W}''_j\big)}{\sum_{k \in \{s,i,t\}} \exp\big(\mathbf{e}_{\text{H}}^k \mathbf{W}''_k\big)},\end{aligned} \tag{11}$$

where $\mathbf{W}''_j \in \mathbb{R}^{d_n \times d_n}$ and $\mathbf{W}''_k \in \mathbb{R}^{d_n \times d_n}$ represent learnable linear transformation matrices. Similarly, we perform

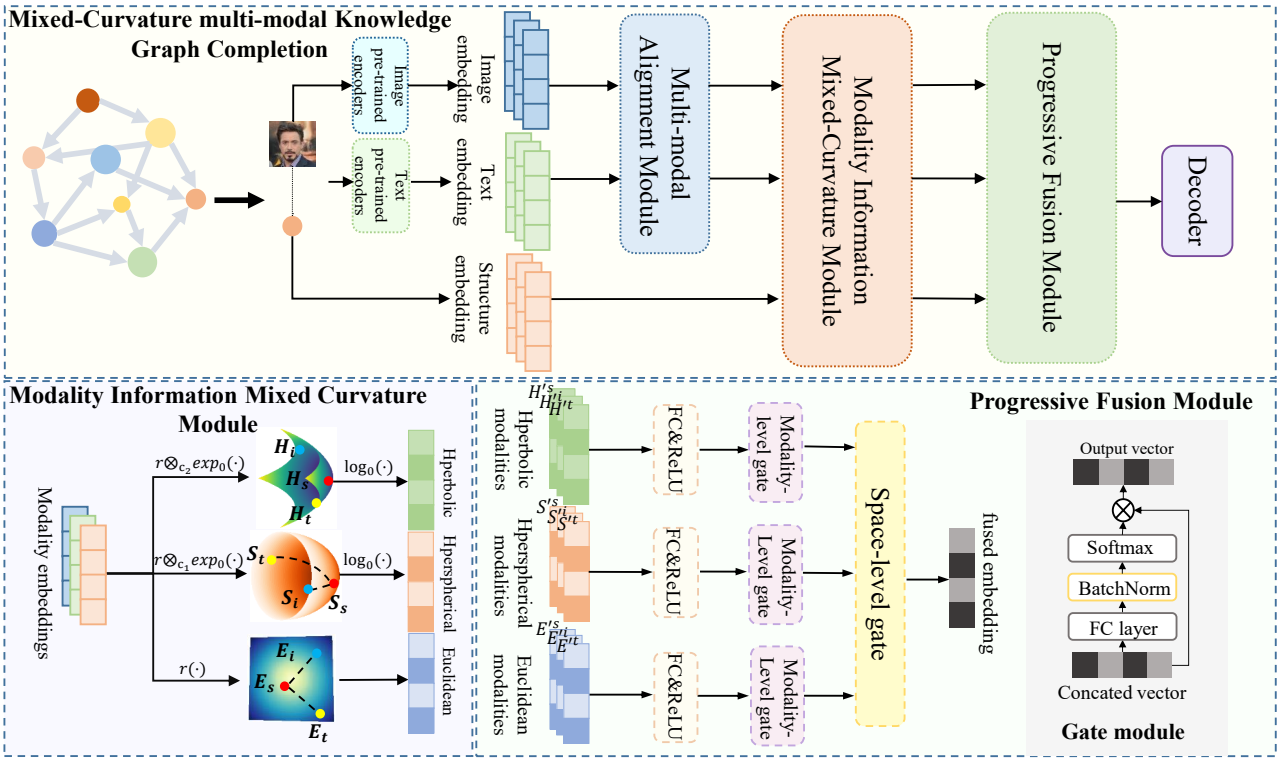


Figure 2: The overall framework of MCKGC. MCKGC consists of three modules, including Multi-modal Alignment Module (MAM), Modality Information Mixed-Curvature Module (MIMCM), and the Progressive Fusion Module (PFM). The MAM introduces the alignment progress for different modalities. The MIMCM learns the embeddings of multi-modal KG information in different curvature spaces. The PFM utilizes a modality-level gate and a space-level gate to progressively fuse the modality information across the three spaces. Notably, only the space-level gate incorporates BatchNorm in its operation.

the same fusion operation on the other two modalities. Ultimately, we obtain embeddings in the three spaces, denoted as follows: $\mathbf{e}_{\mathrm{E}}, \mathbf{e}_{\mathrm{H}}, \mathbf{e}_{\mathrm{S}}$.

Space-level gate. After individually fusing the modality information within each respective space, we enhance the integration of information from the Euclidean, hyperbolic, and hyperspherical spaces by employing a space-level gate to combine the aforementioned three embeddings. In contrast to the modality-level gating, we utilize a BatchNorm (Ioffe and Szegedy 2015) to equilibrate the information across three spaces, mitigating polarization and thus stabilizing the training process.

$$\mathbf{e} = \sum_{i \in \{\mathrm{E},\mathrm{H},\mathrm{S}\}} \beta_i \mathbf{e}_i, \tag{12}$$

where $\mathbf{e}$ is the final embedding for example entity h. β_i is the weight computed by the space-level fusion gate:

$$\begin{aligned} \beta_i &= \mathrm{softmax}\Big(\mathrm{BatchNorm}\big(\mathbf{e}_i \mathbf{W}_i'''\big)\Big) \\ &= \frac{\exp\Big(\mathrm{BatchNorm}\big(\mathbf{e}_i \mathbf{W}_i'''\big)\Big)}{\sum_{k \in \{\mathrm{E},\mathrm{H},\mathrm{S}\}} \exp\Big(\mathrm{BatchNorm}\big(\mathbf{e}_k \mathbf{W}_k'''\big)\Big)}. \end{aligned} \tag{13}$$

Decoder

Multi-modal KGC methods typically require a base KG embedding model as a decoder for link prediction. After deriving the final embeddings of head entities and tail entities from the aforementioned modules, we employ ComplEx (Trouillon et al. 2016) as a decoder. Specifically, based on the above operations, we obtain the unified representation of the entity. For a given triplet $(h, r, t) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, $\mathbf{e}_h$, $\mathbf{r}$, and $\mathbf{e}_t$ are the embeddings of head entity, relation, and tail entity, respectively. Then, we calculate the score by using a multilinear dot product, as follows:

$$\begin{aligned} \phi(h, r, t) &\triangleq \mathrm{Re}\left(\langle \mathbf{e}_h, \mathbf{r}, \bar{\mathbf{e}}_t \rangle\right) \\ &\triangleq \mathrm{Re}\left(\sum_{\ell} [\mathbf{e}_h]_\ell [\mathbf{r}]_\ell [\bar{\mathbf{e}}_t]_\ell\right), \end{aligned} \tag{14}$$

where $\mathrm{Re}(\cdot)$ refers to the real part of a vector. $\bar{\mathbf{e}}_t$ denotes the conjugate of $\mathbf{e}_t$. $[\cdot]_\ell$ represents the ℓ-th entry of a vector. To optimize the parameters, we minimize the cross-entropy loss function to train the model. The loss function is defined as follows:

$$\mathcal{L} = \sum_{(h,r,t) \in \Omega \cap \Omega'} \log\big(1 + \exp(-Y\phi(h, r, t))\big), \tag{15}$$

Datasets	Entities	Relations	Training triplets	Validation triplets	Test triplets
DB15K	12842	279	79222	9902	9904
MKG-W	15000	169	34196	4276	4274
MKG-Y	15000	28	21310	2665	2663

Table 1: Statistics of the datasets.All text and image features are provided by the original datasets.

where Ω and Ω' represent the set of observed triplets and the set of unobserved triplets, respectively. Here, $\Omega' = \mathcal{E} \times \mathcal{R} \times \mathcal{E} - \Omega$, and $Y \in \{1, -1\}$ denotes the label of the triplet (h, r, t) .

Experiments

Experimental Setup

Datasets We evaluate our model on three publicly available multi-modal KGC benchmarks: MKG-W (Xu et al. 2022), MKG-Y (Xu et al. 2022), and DB15K (Liu et al. 2019). DB15K, derived from DBpedia (Lehmann et al. 2015), and MKG-W and MKG-Y, subsets of Wikidata (Vrandečić and Krötzsch 2014). All feature rich image and text data from official releases. Dataset details are provided in Table 1.

Evaluation Protocol Building upon prior research, we utilize link prediction task (Bordes et al. 2013) to assess our model using the specified datasets. Our evaluation approach is based on the scoring function described in our methodology. We rank all entities within the sequences that lack either head entities or tail entities. We employ two rank-based metrics (Sun et al. 2019): Mean Reciprocal Rank (MRR), which measures the average of the reciprocal ranks assigned to the correct entities, and Hits@k, where k is in the set 1,3,10, which measures the proportion of correct triplets in the top k predicted triplets. We adhere to the standard evaluation protocol (Bordes et al. 2013), filtering out all true triplets in the KG during the evaluation, to ensure that low rank predictions for these triplets do not lead to penalties.

Baselines To comprehensively evaluate the effectiveness, we compare our model against various state-of-the-art baselines. The unimodal models, which focus solely on structural information, include TransE (Bordes et al. 2013), DistMult (Yang et al. 2014), ComplEx (Trouillon et al. 2016), RotatE (Sun et al. 2019), and GC-OTE (Tang et al. 2019). The multimodal models, integrating text and image data with structural information, include IKRL (Xie et al. 2016), TBKGC (Mousselly-Sergieh et al. 2018), TransAE (Wang et al. 2019), MMKRL (Lu et al. 2022), RSME (Wang et al. 2021), VBKGC (Zhang and Zhang 2022), OTKGE (Cao et al. 2022b), IMF (Li et al. 2023), QEB (Wang et al. 2023), VISTA (Lee et al. 2023), AdaMF (Zhang et al. 2024c), and MyGO (Zhang et al. 2024b).

Implementation Details Our model is implemented using PyTorch (Paszke et al. 2019). All the experiments are conducted on a RTX 3090 GPU. For all MKG datasets, the optimal hyperparameters are determined via grid search. The learning rate is chosen from {0.5, 0.1, 0.05, 0.01, 0.005}, and the regularization parameter is selected from {0.1, 0.05, 0.01}. The batch size is chosen from {256, 512, 1024}, and the dimension size is set at 256. We use Adagrad (Duchi, Hazan, and Singer 2011) as our optimizer and apply N3 regularization (Chami et al. 2020) to constrain the model parameters. Owing to the limitation of training time, our experimental results are the average of five random repetitions.

Main Results

Table 2 shows the link prediction results of MCKGC as well as all baseline models on DB15K, MKG-W, and MKG-Y datasets. From this table, we have following findings: 1) In general, MCKGC outperforms all baselines across all datasets. Notably, it achieves a 5.4% improvement in the MRR metric compared to the state-of-the-art baseline model MyGO (Zhang et al. 2024b) on DB15k, demonstrating the superiority of our proposed model. 2) Unlike unimodal models such as TransE (Bordes et al. 2013), which only learn structural information, our model effectively incorporates multi-modal auxiliary information to create more comprehensive embeddings, resulting in improved performance. This demonstrates that multi-modal information positively contributes to representation learning for KGC. 3) In addition to the improvements over unimodal models, our model shows significant enhancements when compared to other multi-modal models across the DB15K, MKG-W, and MKG-Y datasets. This underscores the effectiveness of incorporating multi-modal information from a mixed-curvature perspective.

Ablation Study

To verify the effectiveness of each component in MCKGC, we conduct an ablation study focusing on three perspectives: 1) modality information, 2) curvature information, and 3) model design. In the ablation study, we conduct link prediction experiments on the DB15K (Liu et al. 2019) dataset, with the results presented in Table 3. For the modality information, we directly remove the information of images (w/o Image) and texts (w/o Text), respectively. For evaluating the curvature information, we conduct experiments by sequentially removing the hyperbolic vector space (w/o H), hyperspherical vector space (w/o S), and Euclidean space (w/o E). Finally, we investigate the effectiveness of the gated mechanism and Modality Information Mixed-Curvature Module(MIMCM). To be specific, for the model without the gated mechanism (w/o Gate), we use a mean operation as a substitute. And for the model without MIMCM (w/o MIMCM), we directly feed multi-modal information into the progressive fusion module instead of mapping it to multiple spaces for interaction before fusion. From this table, we have the following observations: 1) At the modality level, our complete model surpasses all models with missing modal information, demonstrating that both image and text information are beneficial. 2) From the perspective of curvature information, the results further demonstrate that missing any spatial dimension leads to performance losses. Only through the comprehensive utilization of information

Model	DB15K				MKG-W				MKG-Y			
	MRR	**Hits@1**	**Hits@3**	**Hits@10**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**
Uni-modal												
TransE	24.86	12.78	31.48	47.07	29.19	21.06	33.20	44.23	30.73	23.45	35.18	43.37
DistMult	23.03	14.78	26.28	39.59	20.99	15.93	22.28	30.86	28.71	22.26	27.80	35.95
ComplEx	27.48	18.37	31.57	45.37	24.93	19.09	26.69	36.73	25.04	19.33	32.12	40.93
RotatE	29.28	17.87	36.12	49.66	33.67	26.80	36.68	46.73	34.95	29.10	38.35	45.30
GC-OTE	31.85	22.11	36.52	51.18	33.92	26.55	35.96	46.05	32.95	26.77	36.44	44.08
Multi-modal												
IKRL	26.82	14.09	34.93	49.09	32.36	26.11	34.75	44.07	33.22	30.37	34.28	38.26
TBKGC	28.40	15.61	37.03	49.86	31.48	25.31	33.98	43.24	33.99	30.47	35.27	40.07
TransAE	28.09	21.25	31.17	41.17	30.00	21.23	34.91	44.72	28.10	25.31	29.10	33.03
MMKRL	26.81	13.85	35.07	49.39	30.10	22.16	34.90	44.69	36.81	31.66	39.79	45.31
RSME	29.76	24.15	32.12	40.29	29.23	23.36	31.97	40.43	34.44	31.78	36.07	39.09
VBKGC	30.61	19.75	37.18	49.44	30.61	24.91	33.01	40.88	37.04	33.76	38.75	42.30
OTKGE	23.86	14.85	25.89	34.83	34.36	28.85	36.25	44.88	35.51	31.97	37.18	41.38
IMF	32.25	24.20	36.06	48.19	32.58	27.77	36.06	45.44	35.79	32.95	37.10	40.60
QEB	28.18	14.82	36.67	51.55	32.38	25.47	35.06	45.32	34.37	29.49	37.00	42.30
VISTA	30.42	22.49	33.56	45.94	32.91	26.12	35.38	45.61	30.45	24.87	32.40	41.50
AdaMF	32.51	21.31	39.67	51.68	34.27	27.21	37.86	47.21	38.06	33.49	40.44	**45.48**
MyGO	37.72	30.08	41.26	52.21	36.10	29.78	38.54	**47.75**	-	-	-	-
MCKGC	**39.79** +5.4%	**31.92** +6.1%	**43.80** +6.2%	**54.66** +4.7%	**36.88** +2.2%	**31.32** +5.2%	**38.92** +1.0%	47.43 -	**38.92** +2.3%	**35.49** +5.1%	**40.57** +0.3%	45.21 -

Table 2: The primary results on the DB15K, MKG-W, and MKG-Y datasets are as follows. The best results are highlighted in **bold**, and the second-best results are marked with an underline.

Setting		MRR	Hits@1	Hits@3	Hits@10
MCKGC		**39.79**	**31.92**	**43.80**	**54.66**
Modality Information	w/o Image	39.44	31.52	43.66	54.55
	w/o Text	39.00	30.90	43.14	54.29
Curvature Information	w/o H	38.72	30.71	42.93	53.95
	w/o S	38.45	30.23	42.61	54.36
	w/o E	38.76	30.73	43.00	53.98
Model Design	w/o Gate	38.60	30.57	42.56	54.03
	w/o MIMCM	38.48	30.54	42.65	52.65

Table 3: Ablation studies on the DB15K dataset. The best results are highlighted in **bold**.

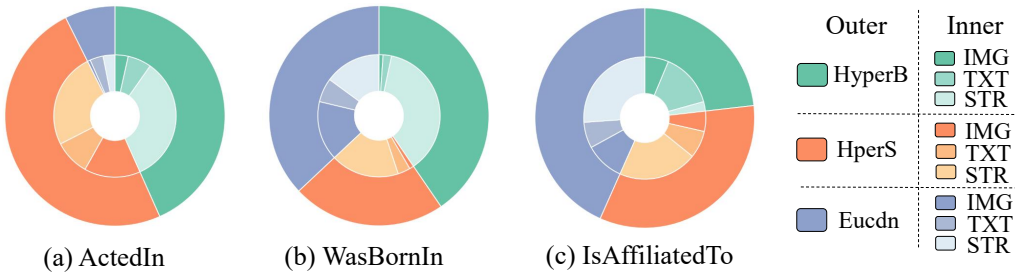
across all three spaces, Euclidean, hyperbolic, and hyperspherical, can the model fully capture the complex structure of KGs. 3) Regarding model design, the results affirm that both MIMCM and the gated mechanism are essential, and only the complete model achieves optimal performance. 4) In summary, the experimental results of ablation studies demonstrate the importance of each module in MCKGC.

Case Study

To demonstrate the significance of our designed progressive fusion module (PFM), we visualized the weight distribution for different types of relationship within the module in the MKG-Y (Xu et al. 2022) dataset. In PFM, we employ a gradual integration approach to better leverage information from different spaces and modalities. Initially, we merge information from different modalities within the same space. Once a unified spatial representation is acquired, we then integrate representations from various spaces. For this purpose, we introduced two gating mechanisms: modality-level gate and space-level gate. These gates automatically allocate the suitable weights for each modality and space in the fusing process, adaptively merging different types of information to achieve the optimal representation. We particularly highlight the distribution of weights for modal information and the modal distribution of spatial information in these two levels of gating.

As shown in Figure 3, the distribution of weights across different spaces and modalities varies among relations. For example, the relation "ActedIn" tends to form hierarchical (c.f. Figure 1) or circular structures (actors co-star a film) in KG, which is exactly hyperbolic space and hyperspherical space goot at handling. The relation "WasBornIn", due to its inherent chain-like and hierarchical nature, fits well into both hyperbolic (good at hierarchical structures) and Euclidean spaces (good at chain-like structures). Similarly, the relation "IsAffiliatedTo", involving affiliations, predominantly utilizes Euclidean space, which is suitable for chain-like structures.

Further exploration of the modal weight distribution across different spaces reveals a general trend of structural information (s) $>$ image information (i) $>$ textual information (t) (i.e., $s > i > t$), consistent with previous research findings (Zhang et al. 2024c; Wang et al. 2021). This indicates that structural information typically has the highest weight, followed by imagery and textual information. More-



(a) ActedIn (b) WasBornIn (c) IsAffiliatedTo

Figure 3: Visualization of the fusion weights on the MKG-Y dataset. We selected three typical relations and present the fusion weights of two gates in PFM. The outer layer represents the fusion weights for three geometric spaces: hyperbolic space, hyperspherical space, and Euclidean space. The inner layer indicates the fusion weights for three modalities in each space: image information, textual information, and structural information.

over, the variation in modal weights across different spaces underscores the capability of our progressive fusion module to effectively utilize diverse types of information, adaptively generating weights for different data modalities.

Related Works

Unimodal Knowledge Graph Completion

Knowledge Graph Completion (KGC) methods are typically embedding-based, where the entities and relations of a KG are embedded into a low-dimensional vector space. Current works can be categorized into models that learn in Euclidean spaces and those in non-Euclidean spaces.

Euclidean Embedding KGC has been extensively studied in Euclidean space with methods like the translation-based TransE (Bordes et al. 2013), which posits $h + r \approx t$ for head (h), tail (t), and relation (r) embeddings. Variants such as TransR (Lin et al. 2015), TransH (Wang et al. 2014), and TransD (Ji et al. 2015) build on this. Euclidean bilinear models like RESCAL (Nickel et al. 2011) and DistMult (Yang et al. 2014) have also been developed. Extensions into complex space include ComplEx (Trouillon et al. 2016), which uses latent semantics for plausibility, and RotatE (Sun et al. 2019), which models relations as rotations from head to tail. WeightE (Zhang et al. 2023) uses bilevel optimization to differentially weight entities and relations, addressing long-tail imbalance and enhancing KGE. These methods, focusing solely on the Euclidean distance between entity embeddings, fail to capture the complex cyclic or hierarchical structures.

Non-Euclidean Embedding To address the limitations of Euclidean models, recent studies have explored non-Euclidean spaces. For example, MURP (Balazevic, Allen, and Hospedales 2019) uses a translation-based model in hyperbolic space to capture graph hierarchies. However, since MURP struggles with certain logical properties of relations, ATTH (Chami et al. 2020) introduces a hyperbolic equidistant scheme to better represent hierarchical structures and logical relation patterns. GC-OTE (Tang et al. 2019) extends RotatE (Sun et al. 2019) into higher-dimensional spaces, utilizing orthogonal transform embedding to effectively model relational patterns.

Multi-modal Knowledge Graph Completion

Multi-modal KGC models enhance traditional KGC by incorporating multi-modal information (Wang et al. 2023) to improve graph embeddings. Current methods focus on making better multi-modal integration, such as IKRL (Xie et al. 2016), TBKGC (Mousselly-Sergieh et al. 2018). OTKGE (Cao et al. 2022b) models the multi-modal fusion process as a transport plan to align different modality embeddings into a unified space. MoSE (Zhao et al. 2022) represents each modality separately before integration for inference, while RSME (Wang et al. 2021) selectively utilizes image information and disregards irrelevant data. IMF (Li et al. 2023) uses an interactive multi-modal fusion framework for integrating diverse modality information. Additionally, Xu et al. tried to enhance negative sampling on multi-modal KGC. VBKGC (Zhang and Zhang 2022) employs a twin negative sampling strategy to align different embedding information. AdaMF (Zhang et al. 2024c) and MMKRL (Lu et al. 2022) both use adversarial training but focus differently: AdaMF balances modality information, while MMKRL enhances model robustness. MyGO (Zhang et al. 2024b) concentrates on processing fine-grained modal information, employing fine-grained contrastive learning to enhance entity representations.

Despite existing methods using Euclidean decoders, which may not fully capture the complexities of multi-modal KGs. This paper proposes embedding multi-modal information into different curvature spaces to improve the representation learning of entities and relations in multi-modal KGs.

Conclusion

In this paper, we introduces a Mixed-Curvature multi-modal Knowledge Graph Completion (MCKGC) method to address the challenge of modeling complex structures in multi-modal KGs. Specifically, we propose two key modules: Modality Information Mixed-Curvature Module (MIMCM) and Progressive Fusion Module (PFM). MIMCM facilitates interactions among multi-modal information within three distinct spaces. while PFM employs a progressive strategy with modality- and space-level gates for adaptive information fusion. Extensive experiments on three public benchmarks demonstrate the effectiveness of our model.

Acknowledgments

The research work is supported by the National Key Research and Development Program of China under Grant Nos. 2021ZD0113602, the National Natural Science Foundation of China under Grant No. 62176014 and No. 62206266, the Fundamental Research Funds for the Central Universities.

References

Balazevic, I.; Allen, C.; and Hospedales, T. 2019. Multi-relational poincaré graph embeddings. *Advances in Neural Information Processing Systems*, 32.

Balažević, I.; Allen, C.; and Hospedales, T. M. 2019. Tucker: Tensor factorization for knowledge graph completion. *arXiv preprint arXiv:1901.09590*.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Cao, Z.; Xu, Q.; Yang, Z.; Cao, X.; and Huang, Q. 2022a. Geometry interaction knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 5521–5529.

Cao, Z.; Xu, Q.; Yang, Z.; He, Y.; Cao, X.; and Huang, Q. 2022b. Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in Neural Information Processing Systems*, 35: 39090–39102.

Chami, I.; Wolf, A.; Juan, D.-C.; Sala, F.; Ravi, S.; and Ré, C. 2020. Low-dimensional hyperbolic knowledge graph embeddings. *arXiv preprint arXiv:2005.00545*.

Chen, W.; Wu, Y.; Zhang, Z.; Zhuang, F.; He, Z.; Xie, R.; and Xia, F. 2024a. FairGap: Fairness-aware Recommendation via Generating Counterfactual Graph. *ACM Transactions on Information Systems*, 42(4): 1–25.

Chen, W.; Yuan, M.; Zhang, Z.; Xie, R.; Zhuang, F.; Wang, D.; and Liu, R. 2024b. FairDgcl: Fairness-aware Recommendation with Dynamic Graph Contrastive Learning. *arXiv preprint arXiv:2410.17555*.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duchi, J.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Guo, Q.; Zhuang, F.; Qin, C.; Zhu, H.; Xie, X.; Xiong, H.; and He, Q. 2020. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8): 3549–3568.

Huang, X.; Zhang, J.; Li, D.; and Li, P. 2019. Knowledge graph embedding based question answering. In *Proceedings of the twelfth ACM international conference on web search and data mining*, 105–113.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456. pmlr.

Ji, G.; He, S.; Xu, L.; Liu, K.; and Zhao, J. 2015. 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 (volume 1: Long papers)*, 687–696.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8): 30–37.

Lee, J.; Chung, C.; Lee, H.; Jo, S.; and Whang, J. 2023. VISTA: Visual-Textual Knowledge Graph Representation Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 7314–7328.

Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P. N.; Hellmann, S.; Morsey, M.; Van Kleef, P.; Auer, S.; et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2): 167–195.

Li, X.; Zhao, X.; Xu, J.; Zhang, Y.; and Xing, C. 2023. IMF: interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023*, 2572–2580.

Lin, B. Y.; Chen, X.; Chen, J.; and Ren, X. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint arXiv:1909.02151*.

Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; and Zhu, X. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Liu, Y.; Li, H.; Garcia-Duran, A.; Niepert, M.; Onoro-Rubio, D.; and Rosenblum, D. S. 2019. MMKG: multi-modal knowledge graphs. In *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16*, 459–474. Springer.

Liu, Z.; Xiong, C.; Sun, M.; and Liu, Z. 2018. Entity-Duet Neural Ranking: Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2395–2405.

Lu, X.; Wang, L.; Jiang, Z.; He, S.; and Liu, S. 2022. MMKRL: A robust embedding approach for multi-modal knowledge graph representation learning. *Applied Intelligence*, 1–18.

Mousselly-Sergieh, H.; Botschen, T.; Gurevych, I.; and Roth, S. 2018. A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 225–234.

Nickel, M.; Tresp, V.; Kriegel, H.-P.; et al. 2011. A three-way model for collective learning on multi-relational data. In *Icml*, volume 11, 3104482–3104584.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Shang, B.; Zhao, Y.; Liu, J.; and Wang, D. 2024. LAFA: Multimodal Knowledge Graph Completion with Link Aware Fusion and Aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8957–8965.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, Z.; Deng, Z.-H.; Nie, J.-Y.; and Tang, J. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*.

Tang, Y.; Huang, J.; Wang, G.; He, X.; and Zhou, B. 2019. Orthogonal relation transforms with graph context modeling for knowledge graph embedding. *arXiv preprint arXiv:1911.04910*.

Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; and Bouchard, G. 2016. Complex embeddings for simple link prediction. In *International conference on machine learning*, 2071–2080. PMLR.

Vrandečić, D.; and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10): 78–85.

Wang, M.; Wang, S.; Yang, H.; Zhang, Z.; Chen, X.; and Qi, G. 2021. Is visual context really helpful for knowledge graph? A representation learning perspective. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2735–2743.

Wang, X.; Meng, B.; Chen, H.; Meng, Y.; Lv, K.; and Zhu, W. 2023. TIVA-KG: A multimodal knowledge graph with text, image, video and audio. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2391–2399.

Wang, Z.; Li, L.; Li, Q.; and Zeng, D. 2019. Multimodal data enhanced representation learning for knowledge graphs. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28.

Xie, R.; Liu, Z.; Luan, H.; and Sun, M. 2016. Image-embodied knowledge representation learning. *arXiv preprint arXiv:1609.07028*.

Xiong, C.; Callan, J.; and Liu, T.-Y. 2017. Word-entity duet representations for document ranking. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, 763–772.

Xu, D.; Xu, T.; Wu, S.; Zhou, J.; and Chen, E. 2022. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *Proceedings of the 30th ACM international conference on multimedia*, 3857–3866.

Yang, B.; Yih, W.-t.; He, X.; Gao, J.; and Deng, L. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Yu, X.; Ren, X.; Sun, Y.; Gu, Q.; Sturt, B.; Khandelwal, U.; Norick, B.; and Han, J. 2014. Personalized entity recommendation: A heterogeneous information network approach. In *Proceedings of the 7th ACM international conference on Web search and data mining*, 283–292.

Zhang, F.; Zhang, Z.; Ao, X.; Gao, D.; Zhuang, F.; Wei, Y.; and He, Q. 2022a. Mind the gap: Cross-lingual information retrieval with hierarchical knowledge enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 4345–4353.

Zhang, F.; Zhang, Z.; Ao, X.; Zhuang, F.; Xu, Y.; and He, Q. 2022b. Along the time: timeline-traced embedding for temporal knowledge graph completion. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2529–2538.

Zhang, F.; Zhang, Z.; Zhuang, F.; Zhao, Y.; Wang, D.; and Zheng, H. 2024a. Temporal Knowledge Graph Reasoning With Dynamic Memory Enhancement. *IEEE Transactions on Knowledge and Data Engineering*.

Zhang, Y.; Chen, Z.; Guo, L.; Xu, Y.; Hu, B.; Liu, Z.; Chen, H.; and Zhang, W. 2024b. MyGO: Discrete Modality Information as Fine-Grained Tokens for Multi-modal Knowledge Graph Completion. *arXiv preprint arXiv:2404.09468*.

Zhang, Y.; Chen, Z.; Liang, L.; Chen, H.; and Zhang, W. 2024c. Unleashing the Power of Imbalanced Modality Information for Multi-modal Knowledge Graph Completion. *arXiv preprint arXiv:2402.15444*.

Zhang, Y.; and Zhang, W. 2022. Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling. *arXiv preprint arXiv:2209.07084*.

Zhang, Z.; Guan, Z.; Zhang, F.; Zhuang, F.; An, Z.; Wang, F.; and Xu, Y. 2023. Weighted knowledge graph embedding. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, 867–877.

Zhao, Y.; Cai, X.; Wu, Y.; Zhang, H.; Zhang, Y.; Zhao, G.; and Jiang, N. 2022. Mose: Modality split and ensemble for multimodal knowledge graph completion. *arXiv preprint arXiv:2210.08821*.