MyGram: Modality-aware Graph Transformer with Global Distribution for Multi-modal Entity Alignment

Zhifei Li[1,5,6], Ziyue Qin[1,*], Xiangyu Luo[4], Xiaoju Hou[2], Yue Zhao[3,*], Miao Zhang[1], Zhifang Huang[1], Kui Xiao[1], Bing Yang[1,*]

[1]School of Computer Science, Hubei University, Wuhan 430062, China
[2]Institute of Vocational Education, Guangdong Industry Polytechnic University, Guangzhou 510300, China
[3]Shandong Police College, Ji'nan 250200, China
[4]School of Cyber Science and Technology, Hubei University, Wuhan 430062, China
[5]Hubei Key Laboratory of Big Data Intelligent Analysis and Application (Hubei University), Wuhan 430062, China
[6]Key Laboratory of Intelligent Sensing System and Security (Hubei University), Ministry of Education, Wuhan 430062, China
{zhifei1993, zhangmiao, 20160006, xiaokui}@hubu.edu.cn, 2023030010@gdip.edu.cn,
zhaoy@sdpc.edu.cn, {qinziyue, xyluo}@stu.hubu.edu.cn, yangbing@126.com

Abstract

Multi-modal entity alignment aims to identify equivalent entities between two multi-modal Knowledge graphs by integrating multi-modal data, such as images and text, to enrich the semantic representations of entities. However, existing methods may overlook the structural contextual information within each modality, making them vulnerable to interference from shallow features. To address these challenges, we propose MyGram, a **m**odality-aware **gra**ph transformer with global distribution for **m**ulti-modal entity alignment. Specifically, we develop a modality diffusion learning module to capture deep structural contextual information within modalities and enable fine-grained multi-modal fusion. In addition, we introduce a Gram Loss that acts as a regularization constraint by minimizing the volume of a 4-dimensional parallelotope formed by multi-modal features, thereby achieving global distribution consistency across modalities. We conduct experiments on five public datasets. Results show that MyGram outperforms baseline models, achieving a maximum improvement of 4.8% in Hits@1 on FBDB15K, 9.9% on FBYG15K, and 4.3% on DBP15K.

Code — https://github.com/HubuKG/MyGram

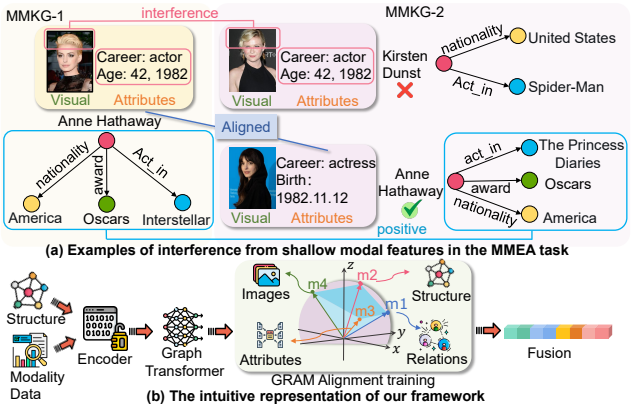

Figure 1: Illustration of modal interference and our GRAM-based alignment framework. (a) shows that when aligning the entity `Anne Hathaway`, the visual and attribute features of the entity `Kirsten Dunst` introduce interference to the task. (b) shows how Gram-based Loss imposes global constraints over cross-modal feature distributions.

Introduction

Knowledge graphs are structured graphical models for representing data (Bordes et al. 2013; Miller 1995; Xu et al. 2022). Since being introduced by Google in 2012, they have been widely adopted in applications such as intelligent question answering (Dinan et al. 2019; Xu et al. 2024; Wang et al. 2017), recommendation systems (Liu et al. 2021a; Sun et al. 2020; Zhang et al. 2016), and various domains (Yang et al. 2021; Zhu et al. 2024). A knowledge graph consists of entities and edges that denote the relationships between them, with each entity and relation associated with its attributes. At their core, they represent real-world knowledge in the form of (head entity, relation, tail entity) triples, enabling efficient data organization and management, as well as strong semantic representation and reasoning capabilities (Fan et al. 2014; Li et al. 2022).

Multi-modal knowledge graphs (MMKGs) have attracted increasing attention from researchers (Liu et al. 2019; Jian et al. 2025). By integrating modalities such as text, images, and audio, MMKGs enhance the semantic representation of entities and improve the expressive power of knowledge graphs. However, due to heterogeneous data sources and differences in construction methods, MMKGs often present inconsistent representations of the same real-world entity. Thus, integrating MMKGs from multiple sources to ensure structural and semantic completeness has become a key research challenge (Trisedya, Qi, and Zhang 2019). Among related tasks, multi-modal entity alignment (MMEA) plays a central role in knowledge fusion by identifying equivalent entities across MMKGs and establishing alignment links,

*Corresponding Authors.

thereby enabling the aggregation of cross-source knowledge (Liu et al. 2020; Chen et al. 2020).

Despite recent progress in MMEA, significant challenges remain in effectively modeling the cross-modal consistency of equivalent entities. (1) **Limitations of contrastive learning**. Most existing methods adopt intra-modal contrastive learning frameworks, aligning entities by optimizing the feature distances between positive and negative pairs. However, these approaches may overlook the distributional differences across modalities in the global feature space, which hinders cross-modal feature consistency. (2) **Interference from shallow features**. Existing methods overlook the structural contextual information within each modality, making it difficult for the model to distinguish between entities that appear similar but are essentially different. As illustrated in Figure 1(a), `Anne Hathaway` and `Kirsten Dunst` share highly similar visual and attribute features, which can introduce interference in the alignment process. However, by leveraging structural information, accurate alignment is still achievable. Therefore, it is necessary to utilize deeper and more discriminative modality-specific structural features.

To address the above issues, we propose **MyGram**, a modality-aware graph transformer with global distribution for multi-modal entity alignment. The model mainly includes the following two strategies: (1) **Gram-based Distribution Alignment.** As shown in Figure 1(b), we introduce Gram Loss as a regularization constraint, which minimizes the volume of a 4-dimensional parallelotope formed by modality vectors, thereby enhancing the global distribution consistency across modalities in the high-dimensional space. Compared to conventional point-wise feature alignment methods, Gram Loss promotes holistic cross-modal semantic coherence and improves the model's generalization ability. (2) **Modality Diffusion Learning**. To enrich entities' representation, we design a modality-aware graph convolutional diffusion module that captures multi-hop neighborhood information within each modality, thereby generating modality features enriched with structural context. On this basis, we further introduce a Transformer architecture that utilizes self-attention mechanisms to integrate information from various modalities, achieving deeper semantic fusion and more precise alignment. Overall, the main contributions of this paper are summarized as:

- We propose a novel modality-aware framework, MyGram, which integrates graph diffusion and Transformer architectures to obtain structurally contextualized modality-specific features for more robust and reliable entity alignment.
- We introduce a Gram-based global alignment strategy that minimizes the volume of a 4-dimensional parallelotope formed by modality embeddings, enforcing global distribution alignment and improving the semantic consistency of equivalent entities.
- We perform extensive experiments on five separate datasets to validate the superiority of our model. For the Hits@1 metric, our model achieves a maximum improvement of 4.8% on FBDB15K, 9.9% on FBYG15K, and 4.3% on DBP15K.

Related Work

This section provides a brief review of uni-modal entity alignment models and multi-modal entity alignment models.

Uni-modal Entity Alignment

Entity Alignment (EA) aims to detect semantically equivalent entities across distinct Knowledge Graphs (KGs), serving as a crucial step toward knowledge fusion (Zhang et al. 2019; Pei et al. 2019). Traditional EA methods mainly target relational triple-based KGs and can be broadly grouped into: (1) Translation-based models, e.g., TransE (Bordes et al. 2013) and TransH (Dinan et al. 2019), which embed entities and relations into vector spaces and capture relational semantics via translation. MTransE (Chen et al. 2017) adds transition matrices for cross-graph mapping, while BootEA iteratively refines alignment through graph matching and embedding learning (Sun et al. 2018). (2) GNN-based models, which employ graph neural networks to model entity structures and enhance representations (Zhao et al. 2023; Jiang et al. 2023). GCN-Align (Wang et al. 2018) applies graph convolutional networks, RDGCN (Wu et al. 2019) models dual-graph structures, and MuGNN (Cao et al. 2019) adopts multi-channel architectures for better alignment.

Uni-modal entity alignment approaches have shown effectiveness by leveraging embedding techniques to align entities. However, these methods may overlook valuable information from other modalities, such as textual descriptions and visual data, which can provide richer and more comprehensive representations of entities.

Multi-modal Entity Alignment

The emergence of MMKGs has drawn growing attention to incorporating visual and textual modalities into entity alignment (Wang et al. 2020; Xie et al. 2017, 2016). Recent studies focus on designing more effective multi-modal fusion mechanisms to improve alignment performance. MSNEA (Chen et al. 2022) introduces visually guided relation and attribute learning to merge modality-specific features into a unified semantic representation. MoAlign (Li et al. 2023b) employs hierarchical attention to capture structural, textual, and visual information. MEAformer (Chen et al. 2023) adopts a dynamic cross-modal weighting strategy that adjusts modality contributions at the instance level in real time.

Furthermore, several MMEA approaches are proposed to address real-world constraints. To alleviate data scarcity, SimDiff applies diffusion-based augmentation for more stable learning (Li et al. 2024). GSIEA (Zhang et al. 2025) reduces adverse structural discrepancies across KGs via graph structure prefix injection. PMF (Huang et al. 2024) suppresses modality-irrelevant interference by progressively freezing each modality's contribution during alignment. IBMEA (Su et al. 2024) introduces an information bottleneck to mitigate the influence of spurious cues.

Unlike prior methods, the proposed modality-aware graph convolutional diffusion Transformer captures high-order structural semantics and applies Gram-based distribution alignment to enhance cross-modal consistency. These strategies jointly improve robustness and alignment accuracy in multi-modal entity alignment tasks.

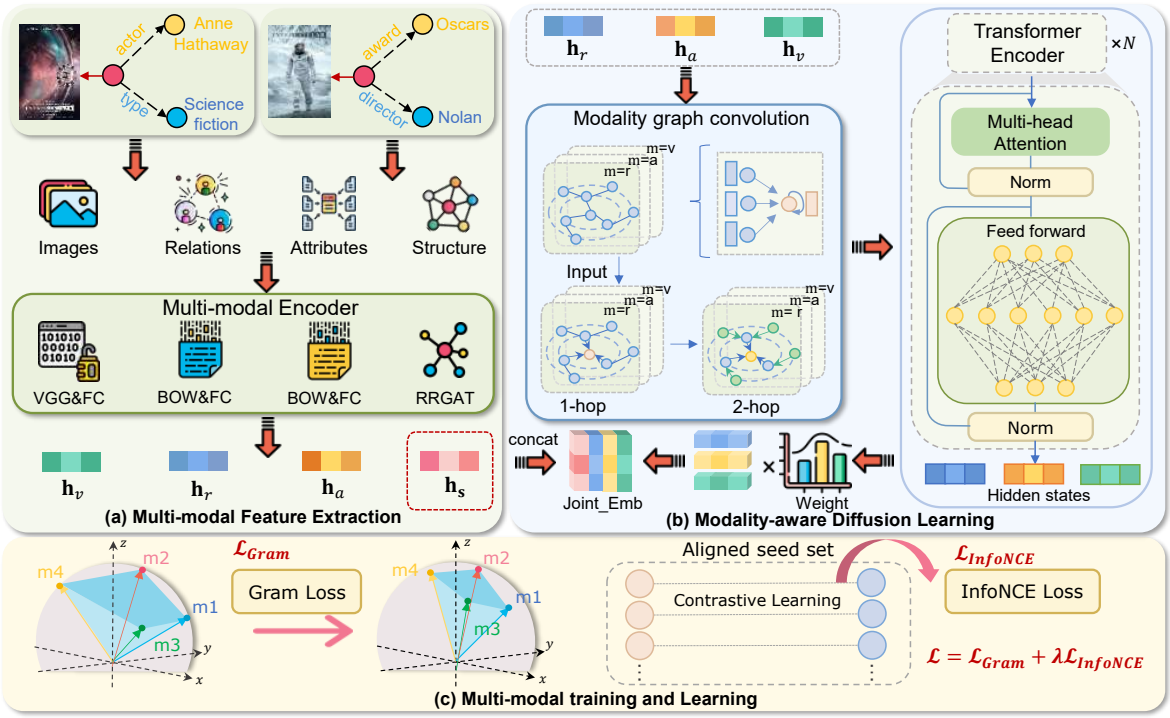

Figure 2: The overall framework of MyGram. (a) Multi-modal Feature Extraction: Extract uni-modal embeddings for each entity from different modalities; (b) Modality-aware Diffusion Learning: Enhance modality features with structural contextual information; (c) Multi-modal training and Learning: employing gram loss to establish alignment between equivalent entities.

Methodology

In this section, we present our MyGram framework, illustrated in Figure 2. Our model consists of three main modules: (1) Multi-modal Feature Extraction: extracting uni-modal embeddings from each entity; (2) Modality-aware Diffusion Learning: obtaining modality features with structural contextual information; (3) Multi-modal training and Learning: employing gram loss to establish alignment between equivalent entities.

Preliminaries

Multi-modal entity alignment seeks to identify entities from different knowledge graphs that refer to the same real-world object. Specifically, we consider two multi-modal knowledge graphs, denoted as $\mathcal{G}_1 = \{\mathcal{E}_1, \mathcal{R}_1, \mathcal{A}_1, \mathcal{V}_1, \mathcal{T}_1\}$ and $\mathcal{G}_2 = \{\mathcal{E}_2, \mathcal{R}_2, \mathcal{A}_2, \mathcal{V}_2, \mathcal{T}_2\}$, with $\mathcal{E}, \mathcal{R}, \mathcal{A}$, and $\mathcal{V}$ representing the set of entities, relations, attributes, and images, respectively. $\mathcal{T}$ signifies the set of triplets. MMEA aims to identify equivalent entity pairs $\mathcal{S} = \{(e_1, e_2)|e_1 \in \mathcal{E}_1, e_2 \in \mathcal{E}_2, e_1 \equiv e_2\}$, During training, the model is given a set of pre-aligned entity pairs $\mathcal{S}$. In the evaluation stage, for each given entity $e_1 \in \mathcal{E}_1$, the model is expected to retrieve its corresponding equivalent entity $e_2 \in \mathcal{E}_2$ from the full candidate set.

Multi-modal Feature Extraction

To extract features from different modalities, we extract entity features from MMKGs by processing each modality independently, thereby preserving its unique semantic information.

Structure To capture the relationships between entities and their neighbors as well as the varying importance of the neighbors, we employ a relational reflection graph attention network (RRGAT) to aggregate the entity neighbors that retain the structural information of the relationships (Mao et al. 2020). Let $x_g \in \mathbb{R}^d$ represents the initialized entity feature, the feature of neighborhood aggregation is:

$$\mathbf{h}_g = RRGAT\left(\omega, \mathbf{M}_g, x_g\right), \tag{1}$$

where ω is a learnable vector, $\mathbf{M}_s$ represents a relational transformation matrix that reflects the relationship.

Relation, Attribute and Visual We first project the input features of relation r, attribute a, and visual v modalities into a shared feature space, and obtain the modality embedding representation through linear transformation:

$$\mathbf{h}_m = \mathbf{W}_m x_m + b_m, m \in \{r, a, v\}, \tag{2}$$

where x_m is the initial feature of entity for modality m. $\mathbf{W}_m \in \mathbb{R}^{d \times d_m}$ and $\mathbf{b}_m$ denote the weight matrix and learnable parameter for the modality m, respectively. For attributes and relations, we represent them using bag-of-words features. For images, we preprocess its visual image using a pretrained image encoder and take the output from the encoder's final layer as the image feature:

$$x_v = ImageEncoder\left(v\right). \tag{3}$$

Modality-aware Diffusion Learning

To address the challenge of interference from shallow features and ensure effective multi-modal fusion, we propose a novel **m**odality-aware **g**raph convolutional **d**iffusion module (MGD). While traditional methods may neglect the modality information of neighboring entities, our approach captures high-order neighbor features across modalities to obtain modality-specific representations enriched with structural context. This is followed by a Transformer-based self-attention mechanism that enables deeper semantic alignment and more comprehensive multi-modal fusion (Vaswani et al. 2017).

For the relation, attribute, and visual embeddings obtained in the previous section, we apply MGD to each modality individually. For each modality, we first construct a shared adjacency matrix with self-loops:

$$\hat{A} = D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}}, \tag{4}$$

where A is the original adjacency matrix, I is the identity matrix (used to introduce self-loops), and D is the degree matrix computed by $D_{ii} = \sum_j (A + I)_{ij}$, where D_{ii} represents the degree of node i.

After constructing the adjacency matrix $\hat{A}$, we process the input features of each modality as follows:

$$\mathbf{H}_m^{(l)} = \beta \cdot \hat{A}\mathbf{H}_m^{(l-1)} + \alpha \cdot \mathbf{H}_m^{(0)} \ \ (l = 1, 2, \cdots, k), \tag{5}$$

where $\mathbf{H}_{m,i}^{(0)}$ is the result of applying dropout to $\mathbf{h}_i^m$, β denotes the neighborhood propagation coefficient, α represents the residual retention coefficient, and $H_{m,i}^{(l)}$ is the result of the l-th propagation. The final output is obtained after k-th propagation as follows, where γ is a stabilization factor utilized for preventing gradient explosion:

$$\mathbf{H}_m = Dropout\left(\frac{1}{\gamma}\mathbf{H}_m^{(k)}\right), \gamma = \beta^k + \alpha\sum_{c=0}^{k-1}\beta^c. \tag{6}$$

After obtaining modality features enriched with structural context, we further introduce the transformer self-attention mechanism to achieve multi-modal interactive fusion. Specifically, the transformer can effectively capture long-range dependencies across different modalities. Through self-attention, it dynamically adjusts the weights of modality features, thereby facilitating comprehensive interaction and fusion among modalities.

First, we apply cross-modal attention to each set of modality features $\mathbf{H}_m$ to obtain the attention weights:

$$head_m^i = \beta_m^{(i)} V_m^{(i)}, \quad \beta_m = \mathrm{softmax}\left(\frac{Q_m^\top K_m}{\sqrt{d_h}}\right), \tag{7}$$

where $\mathbf{Q}_m$, $\mathbf{K}_m$, and $\mathbf{V}_m^{(i)}$ denote the query, key, and value matrices for the i-th attention head, and d_h is the dimension of each head. Then we employ multi-head cross-attention (*MA*) to enhance the ability to capture diverse information across modalities:

$$MA(\mathbf{H}_m) = \left[head_m^1 \oplus, \cdots, \oplus head_m^{N_h}\right]\mathbf{W}_o, \tag{8}$$

where $\oplus$ denotes the concatenation operation, N_h is the number of attention heads, and $\mathbf{W}_o$ is the output projection matrix.

After processing, residual links, normalization, and feed forward are applied, ultimately resulting in the hidden state $\widetilde{\mathbf{H}}_m$. Finally, we define the cross-modal weights for each modality to achieve multi-modal fusion:

$$\omega_m = \frac{exp\left(\sum_{j\in M}\sum_{i=0}^{N_h}\beta_{m,j}^{(i)}/\sqrt{|M| \times N_h}\right)}{\sum_{k\in M} exp\left(\sum_{k\in M}\sum_{i=0}^{N_h}\beta_{m,k}^{(i)}/\sqrt{|M| \times N_h}\right)}. \tag{9}$$

Next, we define the joint embedding for modality fusion:

$$\mathbf{H}_o = \mathbf{H}_g \oplus_{m\in M} [\omega_m \mathbf{H}_m]. \tag{10}$$

Multi-modal training and Learning

Most existing methods adopt contrastive learning during training. They impose distance constraints between the features of positive and negative entity pairs to promote accurate entity alignment. However, such methods may ignore the geometric relationships among vectors in high-dimensional space, making it difficult to fully capture deeper semantic consistency across modalities.

To address the above issue, we draw inspiration from (Cicchetti et al. 2025), who propose using the volume formed by multiple modality vectors in high-dimensional space as a geometric indicator of vector relationships. This volume offers a more intuitive way to reflect the consistency of multi-modal features in high-dimensional space. We apply this design to multi-modal entity alignment, where introducing high-dimensional volume as a regularization constraint helps enforce semantic consistency between different modality features of the same entity from a geometric perspective. A smaller volume implies that the embeddings lie in a more compact subspace, thereby indicating stronger semantic coherence across modalities.

During the training process, we first compute the similarity matrix between the structural feature of the source entity and the visual feature of the target entity in each entity pair. The multi-modal features are represented by the hidden states $\widetilde{\mathbf{H}}_m$ obtained in the previous section:

$$\mathrm{sim} = \frac{\langle \widetilde{\mathbf{H}}_g^s, \widetilde{\mathbf{H}}_v^t \rangle}{\|\widetilde{\mathbf{H}}_g^t\| \cdot \|\widetilde{\mathbf{H}}_v^t\|}. \tag{11}$$

Based on the computed similarity matrix, we select the top-K most similar candidate entities for each source entity. Then, we built a 4-dimensional parallelotope using the structural feature of the source entity and the visual, attribute, and relation features of the target entity in each entity pair.

$$\mathcal{M} = \left[\widetilde{\mathbf{H}}_g^s, \widetilde{\mathbf{H}}_v^t, \widetilde{\mathbf{H}}_a^t, \widetilde{\mathbf{H}}_r^t\right] \in \mathbb{R}^{d_h \times 4}, \tag{12}$$

where $\mathcal{M}$ represents the multi-modal matrix we construct. Then, the Gram matrix $G \in \mathbb{R}^{4\times 4}$ is defined:

$$G = \mathcal{M}^\top \mathcal{M} = \begin{bmatrix} \langle \widetilde{\mathbf{H}}_g^s, \widetilde{\mathbf{H}}_g^s \rangle & \cdots & \langle \widetilde{\mathbf{H}}_g^t, \widetilde{\mathbf{H}}_r^t \rangle \\ \vdots & \ddots & \vdots \\ \langle \widetilde{\mathbf{H}}_r^t, \widetilde{\mathbf{H}}_g^s \rangle & \cdots & \langle \widetilde{\mathbf{H}}_r^t, \widetilde{\mathbf{H}}_r^t \rangle \end{bmatrix}. \tag{13}$$

According to (Gantmacher 1959), if $\widetilde{\mathbf{H}}_g, \ldots, \widetilde{\mathbf{H}}_r$ are vectors in $\mathbb{R}^{d_h}$ spanning a 4-dimensional parallelotope, then the square of the volume of this shape is given by the determinant of the Gram matrix $G \in \mathbb{R}^{4\times4}$, known as the Gramian. From this, we obtain the volume of the 4-dimensional parallelotope as:

$$Vol = \sqrt{|det\,(G\,)| + \epsilon}\,. \tag{14}$$

To ensure that the correct positive match is included in the top-k candidate entities and to further locate its position within the top-k, we define a binary mask:

$$mask^{(i,k)} = \begin{cases} 1 & if\ topk\_idx^{(i,k)} = target^{(i)} \\ 0 & otherwise \end{cases}, \tag{15}$$

where $topk\_idx^{(i,k)}$ denotes the global index of the k-th most similar neighbor of sample i. The mask is used to identify the correct position p of the positive match, which is then used to extract its corresponding log-probability from the softmax distribution in the sparse contrastive loss:

$$\mathcal{L}_{Gram} = -\frac{1}{M}\sum_{m=1}^{M} log\frac{exp\left(-Vol^{(m,p)}/\tau\right)}{\sum\limits_{k=1}^{K} exp\left(-Vol^{(m,k)}/\tau\right)}, \tag{16}$$

where $Vol^{(i,k)}$ denotes the volume spanned by the structural embedding of the source entity and the multi-modal embedding of the target entity, and τ is the temperature coefficient. We further introduce a contrastive alignment loss, which aims to maximize the similarity of true aligned entity pairs and separate them from negative samples:

$$\mathcal{L}_{InfoNCE} = \sum_{(e_i,e_j)\in\mathcal{S}} -log\frac{exp\,(sim\,(e_i,e_j)\,/\mathcal{T})}{\sum\limits_{e_k\in\mathcal{N}_i^{ng}} exp\,(sim\,(e_i,e_k)\,/\mathcal{T})}, \tag{17}$$

where sim(e_i, e_j) represents the similarity (e.g., cosine similarity) between the source entity e_i and the target entity e_j, and $\mathcal{T}$ is the temperature coefficient. For each aligned entity pair (e_i, e_j), we treat it as a positive sample, while sampling multiple negative entities from the candidate set $\mathcal{N}_i$ to construct a local contrastive learning objective. The final loss is formulated as a weighted combination of the two components, λ is used to adjust the weight of gram loss:

$$\mathcal{L}_{total} = \mathcal{L}_{InfoNCE} + \lambda\mathcal{L}_{Gram}. \tag{18}$$

Experiments

To evaluate the effectiveness of the proposed MyGram model, we conduct a thorough experimental study on six subsets from two benchmark datasets. This evaluation is designed to investigate the following five research questions.

RQ1: How does the performance of MyGram compare to other MMEA models?

RQ2: How does modal information impact the performance of MyGram?

RQ3: How do the principal modules affect the performance of MyGram?

RQ4: How does MyGram perform with a low alignment pair rate?

RQ5: How does MyGram perform when applied to real-world MMEA tasks?

Experimental Settings

Datasets Statistics: To assess the effectiveness of our proposed method, we adopt two widely used types of MMEA datasets: cross-knowledge graph datasets and bilingual datasets. For the cross-KG setting, we select FB15K-DB15K and FB15K-YAGO15K (Liu et al. 2019). For the bilingual setting, we employ the DBP15K dataset (Sun, Hu, and Li 2017). Following previous studies, we utilize 20%, 50%, and 80% of the alignment pairs as training seeds for cross-KG datasets, and 30% for bilingual datasets. For entities lacking associated images, we assign randomly initialized vectors for the visual modality, consistent with the settings adopted in prior works (Liu et al. 2021b).

Evaluation Metrics: Evaluation is conducted using standard metrics, including Hits@N (where N = 1, 10) and Mean Reciprocal Rank (MRR). Hits@N indicates the proportion of correct entities ranked in the top N, while MRR (Mean Reciprocal Rank) represents the average reciprocal rank of the correct entities. Higher values of Hits@N and MRR indicate better performance.

Baseline Models: To verify the effectiveness of the proposed method, we perform a comprehensive comparison against a set of representative and competitive MMEA models: PoE (Liu et al. 2019), MMEA (Chen et al. 2020), MSNEA (Chen et al. 2022), MCLEA (Lin et al. 2022), ACK-MMEA (Li et al. 2023a), MoAlign (Li et al. 2023b), MEAformer (Chen et al. 2023), GEEA (Guo et al. 2024), SimDiff (Li et al. 2024), DESAlign (Wang et al. 2024), PMF (Huang et al. 2024), IBMEA (Su et al. 2024), SNAG (Chen et al. 2025), GSIEA (Zhang et al. 2025).

Parameter Settings: We standardize the hidden layer size across all network components to 300 dimensions. The training process is conducted for a total of 1000 epochs, with the learning rate initialized at 5e-3. For the multi-modal embedding module, we adopt VGG-16 as the image feature extractor, setting the visual embedding $\mathbf{d}_v$ to 4096. Regarding the transformer component within our model, the intermediate layer dimension is configured to 400, and the number of self-attention heads is set to 5.

Performance Comparison (RQ1)

Tables 1 and 2 show the experimental results of the multi-modal entity alignment task on two monolingual datasets FBDB15K, FBYG15K, as well as the bilingual dataset DBP15K, under an iterative setting. As anticipated, the proposed MyGram model demonstrates superior performance in all metrics across the nine benchmark experiments when compared to existing state-of-the-art models. Specifically, compared to the second-best performing models, MyGram exhibits average improvements of 4.8%, 9.9%, and 4.3% in the Hit@1 metric on the FBDB15K, FBYG15K, and DBP15K datasets, respectively.

Model	FB15K-DB15K									FB15K-YG15K								
	$R_{Seed} = 20\%$			$R_{Seed} = 50\%$			$R_{Seed} = 80\%$			$R_{Seed} = 20\%$			$R_{Seed} = 50\%$			$R_{Seed} = 80\%$		
	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10
PoE	.170	.126	.251	.533	.464	.658	.721	.666	.820	.154	.113	.229	.414	.347	.536	.635	.573	.746
MMEA	.357	.265	.541	.512	.416	.703	.685	.590	.868	.317	.234	.480	.486	.403	.645	.682	.597	.839
MSNEA	.175	.114	.296	.388	.288	.590	.613	.518	.779	.153	.103	.249	.413	.320	.589	.620	.531	.778
MCLEA	.393	.295	.582	.652	.573	.800	.784	.730	.883	.332	.254	.484	.616	.543	.759	.715	.653	.835
ACK-MMEA	.387	.304	.549	.624	.560	.736	.752	.682	.874	.360	.289	.496	.593	.535	.699	.744	.676	.864
MoAlign	.409	.318	.564	.634	.576	.749	.773	.699	.882	.378	.296	.525	.617	.550	.713	.769	.689	.884
MEAformer	.534	.434	.728	.704	.625	.847	.825	.773	.918	.416	.325	.598	.640	.560	.780	.768	.705	.874
GEEA	.450	.343	.661	.723	.651	.852	.836	.787	.918	.393	.298	.585	.668	.589	.794	.780	.732	.890
SimDiff	.678	.615	.820	.786	.731	.880	.865	.829	.929	.595	.530	.736	.716	.659	.820	.791	.743	.886
DESAlign	.586	.497	.750	.728	.656	.853	.850	.805	.926	.495	.410	.660	.642	.573	.763	.782	.728	.877
PMF	.627	.547	.776	.786	.724	.878	.861	.823	.923	.546	.466	.701	.706	.644	.815	.806	.756	.892
IBMEA	.697	.631	.813	.793	.742	.880	.859	.821	.922	.584	.521	.708	.714	.655	.821	.800	.751	.890
SNAG	.495	.389	.702	.742	.669	.875	.848	.802	.928	.405	.309	.596	.658	.578	.804	.804	.747	.902
GSIEA	.547	.458	.715	.736	.669	.855	.844	.801	.923	.468	.380	.637	.676	.604	.806	.789	.735	.892
MyGram	**.739**	**.679**	**.849**	**.813**	**.764**	**.907**	**.879**	**.842**	**.948**	**.693**	**.629**	**.820**	**.771**	**.715**	**.884**	**.836**	**.783**	**.938**

Table 1: Performance comparison of different MMEA models on FBDB15K and FBYG15K. The optimal results are highlighted in **bold**, while the second-best results are underlined.

Datasets	Models	MRR	Hit@1	Hit@10
$DBP15K_{ZH-EN}$	MSNEA	.684	.601	.830
	MCLEA	.788	.715	.923
	MEAformer	.835	.771	.951
	DESAlign	.865	.810	.957
	GSIEA	.855	.786	.952
	MyGram	**.876**	**.833**	**.960**
$DBP15K_{JA-EN}$	MSNEA	.617	.535	.775
	MCLEA	.785	.715	.909
	MEAformer	.834	.764	.959
	DESAlign	.869	.811	.963
	GSIEA	.852	.787	.962
	MyGram	**.879**	**.836**	**.964**
$DBP15K_{FR-EN}$	MSNEA	.630	.543	.801
	MCLEA	.782	.711	.909
	MEAformer	.841	.772	.962
	DESAlign	.885	.826	.972
	GSIEA	.865	.796	.968
	MyGram	**.908**	**.869**	**.979**

Table 2: Performance comparison of different MMEA models on DBP15K.

The experimental results demonstrate that our proposed MyGram model not only significantly improves the accuracy of multi-modal entity alignment but also enhances its generalization capability across a variety of datasets. Notably, MyGram outperforms the state-of-the-art methods SimDiff and IBMEA. MyGram proposes a modality graph convolutional diffusion method to introduce contextual semantic information for multi-modal features. In addition, it utilizes a specially designed Gram alignment constraint loss that promotes inter-modal semantic structure consistency to improve model robustness. The superior performance of MyGram compared to other multi-modal models suggests promising research directions for utilizing the neighborhood context of multi-modal features in the MMEA task.

Model	FB15K-DB15K			FB15K-YG15K		
	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10
w/o relation	.842	.822	.934	.811	.761	.922
w/o attributes	.859	.834	.940	.818	.768	.925
w/o image	.851	.829	.942	.824	.772	.927
MyGram	**.879**	**.842**	**.948**	**.836**	**.783**	**.938**

Table 3: Ablation study for modalities with 80% seed.

Modality Effect (RQ2)

In order to comprehensively analyze the effects of various factors on the model performance, we design ablation experiments for the modalities. Specifically, we consider three variants with missing modalities: w/o Relation, w/o Attribute, and w/o Image, which denote the model with relation, attribute, and visual modality of the entity removed, respectively. The ablation results on FBDB15K and FBYG15K datasets are presented in Table 3.

Experimental results on both datasets show that the model achieves the best performance when all modal features are used for entity representation. Removing any single modality leads to performance degradation to varying degrees. In particular, excluding the relational modality results in the most significant performance drop across both datasets, indicating the important role of relational information in mul-

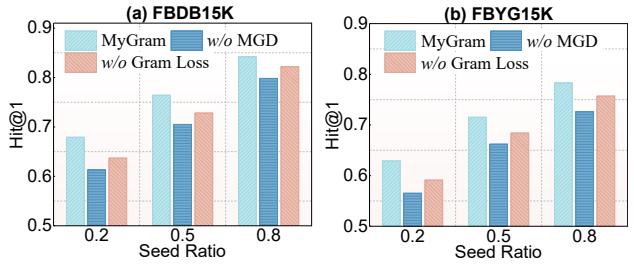

Figure 3: Ablation study on key components of proposed MyGram on (a) FBDB15K and (b) FBYG15K.

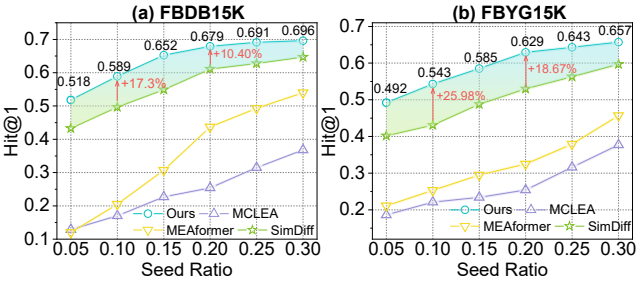

Figure 4: Low resource performance comparison on (a) FBDB15K and (b) FBYG15K.

timodal entity alignment. Moreover, the results of other variants further validate the effectiveness of the model in leveraging multi-modal information for entity alignment.

Key Components (RQ3)

Analogously, we design two variants for the key components ablation study of the proposed model: (1) w/o MGD: a variant that removes the modality-aware graph convolutional diffusion module; (2) w/o Gram: a variant with Gram-based loss removed. Figure 3 presents the impact of these variations on the model.

For the key components of the model, the model performance is significantly degraded in the absence of the modal map convolutional diffusion module. In addition, removing the Gram-based contrast loss also affects the performance. These results validate the effectiveness of the modal graph convolutional diffusion approach and the strategy for promoting multi-modal semantic coherence studied in this paper. These components enhance semantic propagation by taking into account the context of each modal neighborhood and strengthen the semantic consistency between the different modalities of an entity, demonstrating the effective role of modal semantics for MMEA tasks.

Low-Resource Study (RQ4)

To further evaluate the performance of MyGram in low-resource scenarios, we conduct experiments to assess its stability under conditions with an extremely low ratio of aligned seed pairs. In MMEA tasks, pre-aligned entity pairs are usually used to guide the training of the model, which can be viewed as a form of supervised learning. However, in real-world application scenarios, the availability of aligned

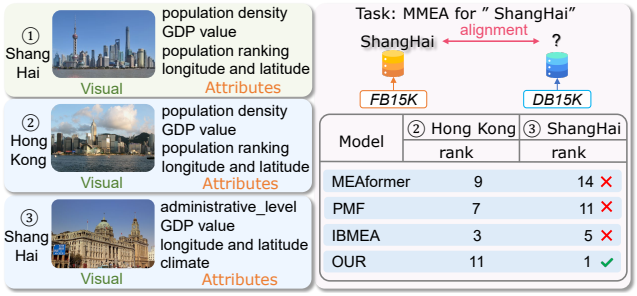

Figure 5: A case example of multi-modal entity alignment task for entity `Shang Hai`.

seed pairs is rare or almost non-existent. Therefore, conducting low-resource training experiments can effectively analyze the robustness and generalization ability of the model in a weakly supervised or unsupervised environment. As illustrated in Figure 4, we compare the performance metrics of MyGram, MEAformer, and SimDiff across varying seed proportions, ranging from 5% to 30%. It can be clearly observed that as the alignment seed rate decreases, all the metrics of each model show performance degradation. However, our proposed MyGram still maintains the performance advantage over the other three models during the decrease in the resource ratio, demonstrating the significant effectiveness of our approach in low-resource scenarios.

Case Study (RQ5)

To evaluate the effectiveness of MyGram, we conducted a case study on the entity `Shang Hai` from the FB15K-DB15K dataset, as shown in Figure 4. In this case, the entities `Shang Hai` and `Hong Kong` share similar modality features and are easily affected by shallow feature interference. The structure of the MMEA task involves predicting the missing equivalent entity in a pair `(Shang Hai, ?)`. When provided with pre-aligned seed pairs as supervised examples to guide the training process, MyGram identifies the potential target entity of `Shang Hai` in DB15K and ranks the candidate entities accordingly. As shown in the prediction results, we observe that MEAformer and PMF assign a lower rank to the correct entity. In contrast, MyGram accurately identifies the correct match, demonstrating its superior ability to capture deeper information in the MMEA task.

Conclusion

In this paper, we propose MyGram for multi-modal entity alignment (MMEA). Our method studies the impact of shallow modality features on alignment and introduces Modality-aware Diffusion Learning to obtain modality representations enriched with structural context, effectively mitigating shallow feature interference. Moreover, MyGram incorporates a Gram-based loss to regularize cross-modal feature distributions, thus promoting global semantic consistency. Experiments show that MyGram consistently outperforms state-of-the-art methods on MMEA. In future work, we plan to explore integrating large language models (LLMs) to further enhance MMEA.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62207011, 62407013, 62377009, 62101179), the Natural Science Foundation of Hubei Province of China (No. 2025AFB653), the Natural Science Foundation of Shandong Province of China (No. ZR2024QF257), the Science and Technology Support Plan for Youth Innovation of Colleges and Universities of Shandong Province of China (No. 2023KJ370), the Open Fund of Hubei Key Laboratory of Big Data Intelligent Analysis and Application, Hubei University (No. 2024BDIAA05), and the Open Fund of Key Laboratory of Intelligent Sensing System and Security of Hubei University, Ministry of Education (No. KLISSS202410).

References

Bordes, A.; Usunier, N.; García-Durán, A.; Weston, J.; and Yakhnenko, O. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, 2787–2795.

Cao, Y.; Liu, Z.; Li, C.; Liu, Z.; Li, J.; and Chua, T. 2019. Multi-Channel Graph Neural Network for Entity Alignment. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 1452–1461.

Chen, L.; Li, Z.; Wang, Y.; Xu, T.; Wang, Z.; and Chen, E. 2020. MMEA: Entity Alignment for Multi-modal Knowledge Graph. In *Proceedings of the 13th International Conference on Knowledge Science, Engineering and Management*, volume 12274, 134–147.

Chen, L.; Li, Z.; Xu, T.; Wu, H.; Wang, Z.; Yuan, N. J.; and Chen, E. 2022. Multi-modal Siamese Network for Entity Alignment. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 118–126.

Chen, M.; Tian, Y.; Yang, M.; and Zaniolo, C. 2017. Multilingual Knowledge Graph Embeddings for Cross-lingual Knowledge Alignment. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 1511–1517.

Chen, Z.; Chen, J.; Zhang, W.; Guo, L.; Fang, Y.; Huang, Y.; Zhang, Y.; Geng, Y.; Pan, J. Z.; Song, W.; and Chen, H. 2023. MEAformer: Multi-modal Entity Alignment Transformer for Meta Modality Hybrid. In *Proceedings of the ACM International Conference on Multimedia*, 3317–3327.

Chen, Z.; Fang, Y.; Zhang, Y.; Guo, L.; Chen, J.; Pan, J. Z.; Chen, H.; and Zhang, W. 2025. Noise-powered Multi-modal Knowledge Graph Representation Framework. In *Proceedings of the 31st International Conference on Computational Linguistics*, 141–155.

Cicchetti, G.; Grassucci, E.; Sigillo, L.; and Comminiello, D. 2025. Gramian Multimodal Representation Learning and Alignment. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *Proceedings of The 7th International Conference on Learning Representations*.

Fan, M.; Zhou, Q.; Chang, E.; and Zheng, T. F. 2014. Transition-based Knowledge Graph Embedding with Relational Mapping Properties. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation*, 328–337.

Gantmacher, F. R. 1959. Matrix theory. *Chelsea, New York*, 21: 6.

Guo, L.; Chen, Z.; Chen, J.; and Chen, H. 2024. Revisit and Outstrip Entity Alignment: A Perspective of Generative Models. In *Proceedings of the twelfth International Conference on Learning Representations*.

Huang, Y.; Zhang, X.; Zhang, R.; Chen, J.; and Kim, J. 2024. Progressively Modality Freezing for Multi-Modal Entity Alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 3477–3489.

Jian, Y.; Luo, X.; Li, Z.; Zhang, M.; Zhang, Y.; Xiao, K.; and Hou, X. 2025. APKGC: Noise-enhanced Multi-Modal Knowledge Graph Completion with Attention Penalty. In *Proceedings of the AAAI Conference on Artificial Intelligence.*, 15005–15013.

Jiang, X.; Zhu, R.; Ji, P.; and Li, S. 2023. Co-Embedding of Nodes and Edges With Graph Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7075–7086.

Li, Q.; Guo, S.; Luo, Y.; Ji, C.; Wang, L.; Sheng, J.; and Li, J. 2023a. Attribute-Consistent Knowledge Graph Representation Learning for Multi-Modal Entity Alignment. In *Proceedings of the ACM Web Conference*, 2499–2508.

Li, Q.; Ji, C.; Guo, S.; Liang, Z.; Wang, L.; and Li, J. 2023b. Multi-Modal Knowledge Graph Transformer Framework for Multi-Modal Entity Alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 987–999.

Li, R.; Di, S.; Chen, L.; and Zhou, X. 2024. SimDiff: Simple Denoising Probabilistic Latent Diffusion Model for Data Augmentation on Multi-modal Knowledge Graph. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1631–1642.

Li, Z.; Liu, H.; Zhang, Z.; Liu, T.; and Xiong, N. N. 2022. Learning Knowledge Graph Embedding With Heterogeneous Relation Attention Networks. *IEEE Trans. Neural Networks Learn. Syst.*, 33(8): 3961–3973.

Lin, Z.; Zhang, Z.; Wang, M.; Shi, Y.; Wu, X.; and Zheng, Y. 2022. Multi-modal Contrastive Representation Learning for Entity Alignment. In *Proceedings of the 29th International Conference on Computational Linguistics*, 2572–2584.

Liu, D.; Lian, J.; Liu, Z.; Wang, X.; Sun, G.; and Xie, X. 2021a. Reinforced Anchor Knowledge Graph Generation for News Recommendation Reasoning. In *Proceedings of The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1055–1065.

Liu, F.; Chen, M.; Roth, D.; and Collier, N. 2021b. Visual Pivoting for (Unsupervised) Entity Alignment. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, 4257–4266.

Liu, Y.; Li, H.; García-Durán, A.; Niepert, M.; Oñoro-Rubio, D.; and Rosenblum, D. S. 2019. MMKG: Multi-modal Knowledge Graphs. In *Proceedings of the 16th International Conference on Semantic Web*, volume 11503, 459–474.

Liu, Z.; Cao, Y.; Pan, L.; Li, J.; and Chua, T. 2020. Exploring and Evaluating Attributes, Values, and Structures for Entity Alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 6355–6364.

Mao, X.; Wang, W.; Xu, H.; Wu, Y.; and Lan, M. 2020. Relational Reflection Entity Alignment. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, 1095–1104.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the Acm*, 38(11): 39–41.

Pei, S.; Yu, L.; Hoehndorf, R.; and Zhang, X. 2019. Semi-Supervised Entity Alignment via Knowledge Graph Embedding with Awareness of Degree Difference. In *Proceedings of the world wide web conference*, 3130–3136.

Su, T.; Sheng, J.; Wang, S.; Zhang, X.; Xu, H.; and Liu, T. 2024. IBMEA: Exploring Variational Information Bottleneck for Multi-modal Entity Alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4436–4445.

Sun, R.; Cao, X.; Zhao, Y.; Wan, J.; Zhou, K.; Zhang, F.; Wang, Z.; and Zheng, K. 2020. Multi-modal Knowledge Graphs for Recommender Systems. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, 1405–1414.

Sun, Z.; Hu, W.; and Li, C. 2017. Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding. In *Proceedings of the 16th International Semantic Web Conference*, volume 10587, 628–644.

Sun, Z.; Hu, W.; Zhang, Q.; and Qu, Y. 2018. Bootstrapping Entity Alignment with Knowledge Graph Embedding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 4396–4402.

Trisedya, B. D.; Qi, J.; and Zhang, R. 2019. Entity Alignment between Knowledge Graphs Using Attribute Embeddings. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 297–304.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Wang, P.; Wu, Q.; Shen, C.; Dick, A. R.; and van den Hengel, A. 2017. Explicit Knowledge-based Reasoning for Visual Question Answering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 1290–1296.

Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; and Huang, J. 2020. Deep Multimodal Fusion by Channel Exchanging. In *Proceedings of the 33th Conference on Neural Information Processing Systems*, 4835–4845.

Wang, Y.; Sun, H.; Wang, J.; Wang, J.; Tang, W.; Qi, Q.; Sun, S.; and Liao, J. 2024. Towards Semantic Consistency: Dirichlet Energy Driven Robust Multi-Modal Entity Alignment. In *Proceedings of the 40th IEEE International Conference on Data Engineering*, 3559–3572.

Wang, Z.; Lv, Q.; Lan, X.; and Zhang, Y. 2018. Cross-lingual Knowledge Graph Alignment via Graph Convolutional Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 349–357.

Wu, Y.; Liu, X.; Feng, Y.; Wang, Z.; Yan, R.; and Zhao, D. 2019. Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 5278–5284.

Xie, R.; Liu, Z.; Jia, J.; Luan, H.; and Sun, M. 2016. Representation Learning of Knowledge Graphs with Entity Descriptions. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2659–2665.

Xie, R.; Liu, Z.; Luan, H.; and Sun, M. 2017. Image-embodied Knowledge Representation Learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 3140–3146.

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.

Xu, N.; Gao, Y.; Liu, A.; Tian, H.; and Zhang, Y. 2024. Multi-Modal Validation and Domain Interaction Learning for Knowledge-Based Visual Question Answering. *IEEE Transactions on Knowledge and Data Engineering*, 36(11): 6628–6640.

Yang, S.; Zhang, R.; Erfani, S. M.; and Lau, J. H. 2021. UniMF: A Unified Framework to Incorporate Multimodal Knowledge Bases into End-to-End Task-Oriented Dialogue Systems. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 3978–3984.

Zhang, F.; Yuan, N. J.; Lian, D.; Xie, X.; and Ma, W. 2016. Collaborative Knowledge Base Embedding for Recommender Systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 353–362.

Zhang, Q.; Sun, Z.; Hu, W.; Chen, M.; Guo, L.; and Qu, Y. 2019. Multi-view Knowledge Graph Embedding for Entity Alignment. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 5429–5435.

Zhang, Y.; Luo, X.; Hu, J.; Zhang, M.; Xiao, K.; and Li, Z. 2025. Graph structure prefix injection transformer for multi-modal entity alignment. *Inf. Process. Manag.*, 62(3): 104048.

Zhao, Y.; Zhou, H.; Zhang, A.; Xie, R.; Li, Q.; and Zhuang, F. 2023. Connecting Embeddings Based on Multiplex Relational Graph Attention Networks for Knowledge Graph Entity Typing. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 4608–4620.

Zhu, X.; Li, Z.; Wang, X.; Jiang, X.; Sun, P.; Wang, X.; Xiao, Y.; and Yuan, N. J. 2024. Multi-Modal Knowledge Graph Construction and Application: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(2): 715–735.