rMMEA: Robust Multi-Modal Entity Alignment with Missing and Noise Visual Modality

Lingbing Guo[1], Zhuo Chen[2], Yichi Zhang[2], Wenbin Guo[1], Haonan Yang[1], Zhao Li[1], Zirui Chen[1], Xin Wang[1*]

[1]Tianjin University
[2]Zhejiang University
{lbguo,wangx}@tju.edu.cn

Abstract

Recently, multi-modal embedding methods have flourished in entity alignment. As state-of-the-art approaches evolve rapidly, visual modality (i.e., images) missing emerges as a critical challenge. While visual modality typically offers the most informative signals in multi-modal entity alignment (MMEA), it is frequently unavailable for many entities. The existing methods commonly use dummy vectors to represent visual-missing embeddings, which negatively impacts both model training and inference. In this paper, we propose robust multi-modal entity alignment (rMMEA), which leverages ranking-based knowledge distillation and mutual information (MI) estimation to address missing modalities while enhancing noise robustness. Unlike conventional teacher-student distillation that requires the student to replicate teacher outputs, our rMMEA learns soft rankings from pure and complete modality sides while capturing implicit key semantics of teacher embeddings through mutual information maximization, allowing rMMEA to avoid strict point-to-point alignment. The experimental results across multiple benchmarks and settings demonstrate that rMMEA significantly outperforms the state-of-the-art anti-modality-missing methods in terms of effectiveness and efficiency.

Introduction

Multi-modal entity alignment (MMEA) has recently gained increased attention (Liu et al. 2021; Li et al. 2023; Guo et al. 2024b; Huang et al. 2024; Zhang et al. 2025b). As a crucial extension of embedding-based entity alignment (EA), MMEA encodes structural, textual, visual and various other modality information into low-dimensional vectors to identify entities that refer to the same real-world object but exist in different knowledge graphs (KGs).

Typically, an MMEA method consists of two key modules: the multi-modal encoder, responsible for encoding the input features of different modalities into their respective embeddings; and the fusion layer, which consolidates all these sub-embeddings into a joint embedding.

Unlike conventional single-modal EA (Sun et al. 2020b) that focuses on the design of structural encoders, the majority of existing MMEA methods place emphasis on the fusion of multi-modal embeddings. These MMEA methods have explored diverse fusion strategies, such as leveraging learnable weights, self-attention mechanisms, and contrastive objectives (Liu et al. 2021; Lin et al. 2022; Chen et al. 2022c, 2023; Li et al. 2023; Huang et al. 2024; Chen et al. 2024).

Nevertheless, there remain several significant challenges in MMEA that have yet to be fully addressed. For example, the visual features are frequently missing for entities within real-world KGs. Even the elaborately curated benchmarks like DBP15K contain over 20% of entities without image links (Sun, Hu, and Li 2017; Liu et al. 2021). Thus, addressing modality absence and potential noisy features represents a crucial and underexplored direction in MMEA.

To process entities without visual features, current methods often use a dummy vector (typically zero or randomly initialized) as the vision embedding (Liu et al. 2021; Chen et al. 2022c; Huang et al. 2024; Guo et al. 2024a, 2025). However, when identifying alignment entity pairs, these dummy embeddings do not provide useful information and can distort the actual similarity score between entities.

The images of entities are so crucial for identifying their counterparts that some MMEA methods even use the alignment scores of vision embeddings as supervised data (Lin et al. 2022). Take Figure 1a as an example, the absence of visual information reduces discriminability, increasing the likelihood of incorrect alignment candidates. To tackle this issue, the existing anti-modality-missing MMEA methods leverage generative models such as variational autoencoders (VAEs) (Kingma and Welling 2013) to generate visual-missing embeddings (Chen et al. 2023), which requires more parameters and multi-stage training (Figure 1b). Additionally, there are general multi-modal learning techniques (Chen et al. 2022a) employing knowledge distillation to transfer "dark knowledge" from teacher (modality-complete) to student (visual-missing). However, they may struggle with significant learning gaps, as seen in the transition from 0.36 to 0.87 in Figure 1c.

In this paper, we propose robust multi-modal entity alignment (rMMEA) for modality missing and noise resistance (Figure 1d). rMMEA leverages ranking-based distillation to teach the student the relatively rankings of entities, offering a more adaptive and softer objective that conveys essential ranking information for EA. We implement rMMEA in a self-distillation (Tian, Krishnan, and Isola 2020;

*Corresponding Author.

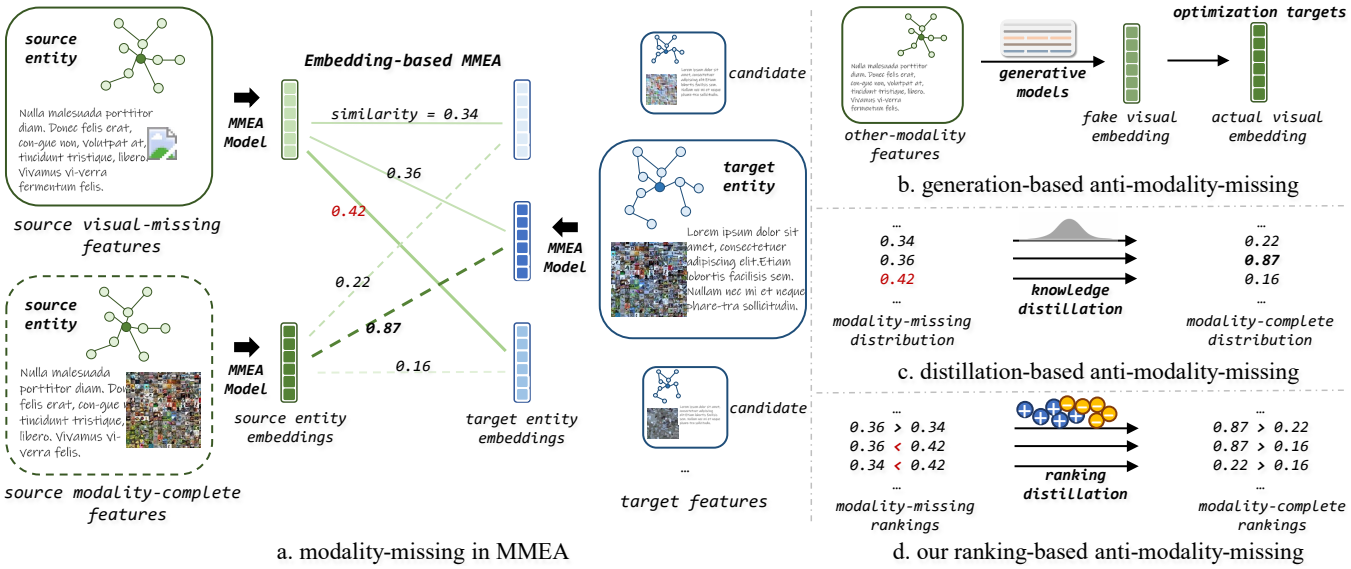


Figure 1: Illustrations of modality missing in MMEA and anti-modality-missing methods: a. MMEA models encode multi-modal features into embeddings. The source visual-missing embedding (light green) is erroneously aligned due to its low discriminability $(0.34, 0.36, 0.42)$. b. Generation-based methods use other features to synthesize visual embedding (instead of actual image), necessitating multi-stage training. c. Distillation-based approaches pursue precise knowledge transfer between alignment probability distributions, e.g., aligning 0.36 with 0.87. d. Our ranking-based distillation focuses attention on rankings and employs adaptive knowledge transfer, e.g., aligning the relationship "$<$" in "$0.36 < 0.42$" with "$>$" in "$0.87 > 0.16$".

Guo et al. 2024d) manner, allowing it to be jointly optimized with the main EA objective.

Furthermore, we also leverage mutual information (MI) estimation (Belghazi et al. 2018; van den Oord, Li, and Vinyals 2018) to distill the semantics underlying the joint embeddings from modality-complete estimation. It is also based on self-supervision without point-to-point alignment and multi-stage training. As a result, rMMEA can be easily adapted to various modality missing scenarios while handling modality noise.

We summarize our contributions as follows:

- We propose rMMEA, an efficient self-distillation method for anti-modality-missing in MMEA. rMMEA requires no external data and offers a significant speed advantage over existing anti-missing-modality methods.
- We design ranking-based distillation and MI estimation objectives for robust MMEA in a soft and adaptive manner. Experimental results demonstrate that rMMEA exhibits less performance degradation compared to baselines as more corrupted data involved.
- We conduct experiments across seven MMEA benchmarks, in which rMMEA consistently and significantly outperforms state-of-the-art anti-modality-missing methods in both effectiveness and efficiency.

Related Works

We categorize related works into two groups:

Multi-Modal Entity Alignment MMEA is a significant extension of embedding-based EA approaches (Chen et al. 2017; Guo, Sun, and Hu 2019; Sun et al. 2020a; Zeng et al. 2020; Guo et al. 2022a,b). Most MMEA methods leverage graph neural networks (GNNs) (Kipf and Welling 2017) alongside pretrained text and vision models (Mikolov et al. 2013; Devlin et al. 2019; Radford et al. 2021) to produce the embeddings of graph, attribute, and image features. Their differences rest in how to fuse the multi-modal embeddings. Specifically, EVA (Liu et al. 2021) learns the cross-modal interactions with attention-based fusion layers. MEAformer (Chen et al. 2022c) and MCLEA (Lin et al. 2022) strengthen these interactions through inter-modal and inner-modal attention mechanisms and contrastive learning, respectively. ACK-MMEA (Li et al. 2023) and AS-GEA (Luo et al. 2024) introduce multi-modal normalization and multi-modal paths, respectively.

Anti-Modality-Missing MMEA As an emerging direction in multi-modal learning and KG representation learning, modality missing in MMEA remains understudied. The closest works to ours are UMAEA (Chen et al. 2023) and GEEA (Guo et al. 2024b), which leverage generative models (e.g., VAE (Kingma and Welling 2013; Kingma et al. 2016)) to synthesize visual-missing embeddings.

In typical multi-modal learning, generation-based and distillation-based methods are two predominant approaches for anti-modality-missing. The former (including UMAEA) generates modality-missing embeddings using generative models (Tran et al. 2017; Li et al. 2020; Kim et al. 2025), such as VAEs, generative adversarial networks (GANs) (Goodfellow et al. 2014), and diffusion models (Ho, Jain, and Abbeel 2020). The latter transfers the "dark

knowledge" from modality-complete outputs and necessitates lower computational cost (Chen et al. 2022a; Wang et al. 2020; Lao et al. 2025; Sikdar, Teotia, and Sundaram 2025). Our work falls into this category and introduces a ranking-based distillation objective specifically for EA.

Methodology

In this section, we present the details of rMMEA. We start with preliminaries defining MMKG and MMEA, and then introduce multi-modal embedding learning. Finally, we discuss the proposed rMMEA in detail, illustrating the insights and implementations of ranking-based distillation and mutual information estimation.

Preliminaries

MMKG Following the previous works (Liu et al. 2021; Lin et al. 2022; Chen et al. 2022b, 2023; Guo et al. 2024b), we define a multi-modal knowledge graph (MMKG) as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{A}, \mathcal{V}\}$, where $\mathcal{E}$,$\mathcal{R}$,$\mathcal{T}$ denote the sets of entities, relations, and relational triplets, respectively, while $\mathcal{A}$ and $\mathcal{V}$ represent the sets of attributes and images (i.e., the visual modality), respectively.

MMEA MMEA extends conventional EA to multi-modal settings. Formally, given two MMKGs $\mathcal{G}^1 = \{\mathcal{E}^1, \mathcal{R}^1, \mathcal{T}^1, \mathcal{A}^1, \mathcal{V}^1\}$ and $\mathcal{G}^2 = \{\mathcal{E}^2, \mathcal{R}^2, \mathcal{T}^2, \mathcal{A}^2, \mathcal{V}^2\}$, MMEA identifies each entity pair (e_i^1, e_i^2) in respective KGs that refer to the same real-world object e_i, with a small proportion of pre-aligned entity pairs $\mathcal{S}^{tr} \in \mathcal{E}^1 \times \mathcal{E}^2$ as training data. The corresponding testing set is denoted by $\mathcal{S}^{te}$.

Embedding-based MMEA

Let $\mathcal{M} = \{g, r, a, v, \dots\}$ represent the set of all modalities, where g,r,a, and v denote graph, relation, attribute, and visual modalities, respectively. Then, we can formalize the encoding and fusion processes of MMEA as:

$$\mathbf{e}_i^m = \mathbf{E}^m(e_i, \mathbf{X}^m), m \in \mathcal{M} \quad \text{(Encoding)} \quad (1)$$

$$\mathbf{e}_i = \mathbf{F}(\{\mathbf{e}_i^m | m \in \mathcal{M}\}) \quad \text{(Fusion)} \quad (2)$$

where the boldfaced $\mathbf{e}_i^m$, $\mathbf{e}_i$ denote the modality m embedding and the joint embedding of entity e_i, respectively. $\mathbf{E}^m$ and $\mathbf{X}^m$ are the encoder and input feature of modality m, respectively. The encoder and input format can vary with different modalities. For example, graph encoding typically uses GNNs (e.g., GAT (Velickovic et al. 2018; Guo et al. 2024c)) with adjacency matrices as input, which are constructed from the relational triplet sets $\mathcal{T}^1$ and $\mathcal{T}^2$. The fusion layer $\mathbf{F}$ also has various implementations, such as fully-connection layer or self-attention. The detailed designs of the encoder and fusion layer are beyond our scope, thus we refer interest readers to the surveys (Jahanifar et al. 2023; Chen et al. 2024). In our experiments, we adopt the configurations of baselines (Chen et al. 2023) to evaluate rMMEA.

Once obtaining the joint embedding $\mathbf{e}_i^1$ of entity e_i^1 in the source KG $\mathcal{G}^1$, we are capable of seeking its optimal aligned entity $e_i^{2,*}$ in the target KG $\mathcal{G}^2$ through:

$$\mathbf{p}(\mathcal{E}^2 | e_i^1) = \{\mathbf{S}(\mathbf{e}_i^1, \mathbf{e}_j^2)\}_{e_j^2 \in \mathcal{E}^2} \quad (3)$$

$$e_i^{2,*} = \arg\max(\mathbf{p}(\mathcal{E}^2 | e_i)) \quad (4)$$

where $\mathbf{p}(\mathcal{E}^2|e_i)$ is the unnormalized alignment probability distribution over $\mathcal{E}^2$, and $\mathbf{S}(\cdot,\cdot)$ is the score/similarity function used to estimate the alignment score between an arbitrary pair of entity embeddings $(\mathbf{e}_i^1, \mathbf{e}_j^2)$.

Robust MMEA for Missing Visual Modality

From the analysis of the existing works (Chen et al. 2023; Huang et al. 2024), most MMEA benchmarks exhibit varying degrees of visual missing. To mitigate the negative impact, rMMEA incorporates the ranking-based knowledge distillation and mutual-information-based estimation objectives into MMEA, both requiring neither extra training data nor additional training stages.

Ranking-based Knowledge Distillation Knowledge distillation is a well-known and effective approach in multi-modal learning to address the issue of modality missing. In the specific case of MMEA, supposed our goal is to find the aligned entities of e_i^1 from $\mathcal{E}^2$, and the alignment probability distribution under modality-complete modeling is $\mathbf{p}_t(\mathcal{E}^2|e_i^1)$ while that under visual-missing modeling is $\mathbf{p}_s(\mathcal{E}^2|e_i^1)$, then we can optimize a Kullback-Leibler (KL) divergence (Kullback and Leibler 1951) loss to distill the "dark knowledge" from $\mathbf{p}_t(\mathcal{E}^2|e_i^1)$ into $\mathbf{p}_s(\mathcal{E}^2|e_i^1)$:

$$\mathcal{L}_{kd} = D_{\mathrm{KL}}(\mathbf{p}_s(\mathcal{E}^2|e_i^1), \tilde{\mathbf{p}}_t(\mathcal{E}^2|e_i^1)) \quad (5)$$

where $\mathbf{p}_s(\mathcal{E}^2|e_i^1)$ represents the output of a student model in knowledge distillation, while the gradient computation for teacher's output $\tilde{\mathbf{p}}_t(\mathcal{E}^2|e_i^1))$ is stopped to avoid multi-stage training and prevent reverse learning (Tian, Krishnan, and Isola 2020; Guo et al. 2024d). This self-distillation loss compels the student to mimic the teacher and can be jointly optimized with the main EA objective (Pei et al. 2019; Guo et al. 2024b). Furthermore, the paired examples in Equation (5) are constructed from the training set without external data augmentation, making it widely adopted in multi-modal learning area (Chen et al. 2022a; Wang et al. 2020; Lao et al. 2025; Sikdar, Teotia, and Sundaram 2025).

Considering the number of candidates for each source entity, aligning the two full-size distributions can be challenging. In this case, negative sampling is helpful. We alternatively use a small subset of negative entities $\mathcal{N}_i^2$ sampled from $\mathcal{E}^2$ to approximate the full set estimation, which will rewrite Equation (5) as:

$$\mathcal{L}_{kd} = D_{\mathrm{KL}}(\mathbf{p}_s(\mathcal{N}_i^2|e_i^1), \tilde{\mathbf{p}}_t(\mathcal{N}_i^2|e_i^1)) \quad (6)$$

$$\mathcal{N}_i^2 = \{e_i^2\} \cup \{e_j^2 | e_j^2 \in \mathcal{E}^2, e_j^2 \neq e_i^2\} \quad (7)$$

However, as discussed in the introduction section and Figure 1, the visual-missing embeddings may reduce discriminability in the student distribution $\mathbf{p}_s(\mathcal{N}_i^2|e_i^1)$. Directly compensating for the numerical gap is difficult, and we prioritize target entity rankings over magnitude.

Building on this insight, we propose ranking-based distillation. Let $r(e_j^2|e_i^1)$ be the alignment ranking of e_j^2 w.r.t. e_i^1, then it can be computed by the following equation:

$$r(e_j^2|e_i^1) = 1 + \sum_{e_k \in \mathcal{N}_i^2} \mathbb{I}(\mathbf{S}(\mathbf{e}_i^1, \mathbf{e}_k^2) - \mathbf{S}(\mathbf{e}_i^1, \mathbf{e}_j^2)) \quad (8)$$

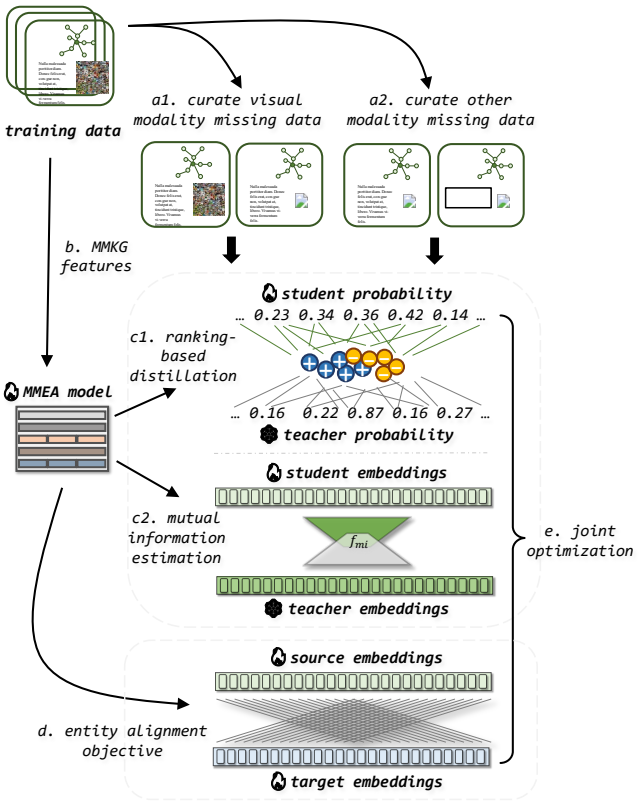


Figure 2: An overview of rMMEA: a. We first construct the supervised data from the original training set. b. Then, we leverage an MMEA model to produce entity embeddings. c. We compute the ranking-based distillation and mutual information estimation losses using these embeddings with examples from first step for supervision. d. We also employ conventional entity alignment objective for MMEA learning. e. Finally, we minimize all losses jointly.

where $\mathbb{I}$ is an indicator function returning 1 if $(\mathbf{S}(\mathbf{e}_i^1, \mathbf{e}_k^2) - \mathbf{S}(\mathbf{e}_i^1, \mathbf{e}_j^2)) > 0$ holds and 0 otherwise.

Then, we can employ a self-distillation loss similar to Equation (5) to minimize the ranking discrepancies between modality-complete output and visual-missing output:

$$\mathcal{L}_{rd} = \sum_{e_j^2 \in \mathcal{N}_i^2 \cup \{e_i^2\}} |r_s(e_j^2|e_i^1) - r_t(e_j^2|e_i^1)| \tag{9}$$

A key limitation of above formulation is the discontinuity of the indicator function $\mathbb{I}$, which is incompatible with gradient-based optimization. To address this, we parameterize $\mathbb{I}$ using a special activation function σ and reformulate Equation (8) as:

$$r(e_j^2|e_i^1) = 1 + \sum_{e_k \in \mathcal{N}_i^2} \sigma((\mathbf{S}(\mathbf{e}_i^1, \mathbf{e}_k^2) - \mathbf{S}(\mathbf{e}_i^1, \mathbf{e}_j^2))) \tag{10}$$

$$\sigma(x) = \frac{1}{1 + \beta e^{-\min(\alpha x, x)}} \tag{11}$$

where the leading 1 in Equation (10) maintains correct ranking positions. α and β are two hyper-parameters controlling the temperature of the activation.

When $\alpha < 1$ and $\beta = 1$, σ resembles a composition of LeakyReLU (Xu et al. 2015) and Sigmoid activations, excepted that the $\max$ is replaced with a $\min$. This design precisely meets our motivation: setting $\alpha > 1$ magnifies the negative alignment scores to approximate $\mathbb{I}$ via near-zero Sigmoid outputs, while positive scores remain unaffected.

MI-based Semantics Estimation In conventional multi-modal learning, the performance typically depends on the predicted probability distribution (Chen et al. 2022a; Wang et al. 2020). Differently, the effectiveness of MMEA relies more heavily on entity embedding quality (Equation (3)). Thus, it is also crucial to align the visual-missing embeddings with modality-complete counterparts in MMEA.

As discussed in previous sections, rMMEA avoids explicit point-to-point alignment, concentrating instead on the more essential information for EA. Directly minimizing L1/L2 distance between visual-missing and modality-complete embeddings would easily create a significant learning gap, as it forces precise dimensional matching.

Drawing inspiration from domain adaption area (Ganin and Lempitsky 2015; Shen et al. 2018; Ben-David et al. 2010; Courty et al. 2017), we propose mutual information estimation as alternative approach. Unlike distance-based metrics that demand exact dimensional correspondence, mutual information captures implicit embedding correlations and is also in line with our insight in designating ranking-based distillation. Following the established methodologies (van den Oord, Li, and Vinyals 2018; Tian, Krishnan, and Isola 2020), we first leverage a neural function f_{mi} to estimate the mutual information density:

$$f_{\text{mi}}(\mathbf{e}_s, \mathbf{e}_t) = \exp(\mathbf{e}_{i,s}^T \mathbf{W}_m \hat{\mathbf{e}}_{i,t} + \mathbf{b}_m) \tag{12}$$

where $\mathbf{W}_m$ and $\mathbf{b}_m$ are learnable weight matrix and bias, respectively. We then employ contrastive learning to estimate and maximize the mutual information:

$$\begin{aligned}\mathcal{L}_{\text{mi}} = &- \log\left(f_{\text{mi}}(\mathbf{e}_{i,s}, \mathbf{e}_{i,t})\right) \\ &+ \mathop{\mathbb{E}}_{e_{j,t} \in \mathcal{N}_{i,t}} \log\left(f_{\text{mi}}(\mathbf{e}_{i,s}, \mathbf{e}_{j,t})\right).\end{aligned} \tag{13}$$

where $\mathcal{N}_{i,t}$ denotes the sampled negative set for teacher embedding $\mathbf{e}_{i,t}$. The proof of how this contrastive objective approximates true mutual information have been discussed in (Belghazi et al. 2018; van den Oord, Li, and Vinyals 2018; Tian, Krishnan, and Isola 2020; Guo et al. 2024d). It is worth noting that, mutual information estimation also demonstrates superior noise resistance compared to distance-based approaches.

Robust MMEA for Missing and Noise Modalities

The ranking-based distillation and mutual-information estimation in rMMEA are applicable not only to visual-missing scenarios but also to other cases of modality missing or noise. We believe incorporating such data during training can further enhance the robustness of MMEA. Notably, the required training examples can be conveniently sampled from the existing training set.

r_{vm}	Model	DBP15K$_{\text{ZH-EN}}$			DBP15K$_{\text{JA-EN}}$			DBP15K$_{\text{FR-EN}}$		
		H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
0.4	EVA (Liu et al. 2021)	.625	.876	.717	.624	.881	.716	.634	.900	.728
	MSNEA (Chen et al. 2022b)	.520	.786	.611	.480	.744	.569	.478	.772	.574
	MCLEA (Lin et al. 2022)	.661	.896	.744	.686	.898	.761	.675	.901	.757
	UMAEA (Chen et al. 2023)	.750	.933	.805	.775	**.963**	.845	.792	**.970**	.859
	GEEA (Guo et al. 2024b)	.753	.931	.801	.759	.955	.841	.784	.960	.843
	rMMEA	**.766**	**.948**	**.833**	**.799**	.960	**.853**	**.797**	.965	**.864**
0.95	EVA (Liu et al. 2021)	.623	.878	.715	.615	.877	.708	.624	.895	.720
	MSNEA (Chen et al. 2022b)	.413	.722	.517	.313	.643	.425	.297	.690	.427
	MCLEA (Lin et al. 2022)	.638	.905	.732	.599	.897	.706	.634	.930	.741
	UMAEA (Chen et al. 2023)	.720	.938	.800	.725	.949	.807	.752	**.970**	.830
	GEEA (Guo et al. 2024b)	.721	.931	.799	.723	.944	.803	.750	.961	.828
	rMMEA	**.759**	**.939**	**.826**	**.760**	**.950**	**.828**	**.786**	.964	**.854**

Table 1: MMEA results on DBP15K w.r.t. different *visual missing ratio* r_{vm}. The best and second-best results are boldfaced and underlined, respectively. H@1, H@10, and MRR indicate Hits@1, Hits@10, and mean reciprocal rank, respectively.

r_{vm}	Model	OpenEA$_{\text{EN-FR}}$			OpenEA$_{\text{EN-DE}}$			OpenEA$_{\text{D-W-V1}}$			OpenEA$_{\text{D-W-V2}}$		
		H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
0.4	EVA	.547	.830	.647	.734	.921	.800	.595	.811	.673	.788	.954	.848
	MSNEA	.360	.560	.427	.412	.622	.484	.432	.601	.490	.545	.781	.626
	MCLEA	.597	.852	.688	.745	.906	.803	.655	.848	.726	.800	.948	.855
	UMAEA	.665	.914	.753	.804	.957	.860	.724	.908	.791	.859	.987	.905
	GEEA	.660	.911	.749	.805	.957	.860	.722	.901	.785	.858	.984	.900
	rMMEA	**.687**	**.921**	**.771**	**.819**	**.956**	**.869**	**.736**	**.913**	**.801**	**.881**	**.989**	**.921**
0.95	EVA	.528	.833	.634	.717	.917	.787	.570	.801	.653	.775	.952	.839
	MSNEA	.200	.431	.278	.242	.486	.323	.238	.452	.310	.397	.690	.497
	MCLEA	.545	.852	.653	.723	.918	.791	.585	.834	.675	.771	.965	.842
	UMAEA	.605	.898	.708	.757	.942	.823	.647	.881	.733	.840	.984	.890
	GEEA	.600	.894	.704	.758	.940	.824	.639	.875	.727	.840	.981	.887
	rMMEA	**.628**	**.900**	**.726**	**.779**	**.946**	**.839**	**.677**	**.889**	**.756**	**.862**	**.987**	**.909**

Table 2: MMEA results on OpenEA w.r.t. different *visual missing ratio* r_{vm}.

Figure 2 illustrates the overall workflow of rMMEA. Since most entities retain complete features for non-visual modalities, we can first sample a small proportion of entities as modality-complete entities. Then, we can generate simulated missing/noisy data by either random initialization or zero-padding of specific modality features. After obtaining the embeddings and alignment scores from the basic MMEA model, we can estimate and optimize rMMEA on diverse modality-missing/noise scenarios.

We also present an algorithm of training rMMEA in the Appendix. Briefly, we first initialize all parameters of rMMEA, then employ mini-batch training similar to conventional MMEA methods. For each iteration, we compute the joint embeddings, alignment scores, and main EA prediction loss $\mathcal{L}_{main}$ as usual. Meanwhile, we also curate the modality-missing and modality-noise data, based on which we compute the ranking-based knowledge distillation loss $\mathcal{L}_{rd}$ (Equation (9)) and mutual information estimation loss $\mathcal{L}_{mi}$ (Equation (13)). We minimize three losses jointly until the performance on the validation set converges.

Experiment

In this section, we conduct experiments to validate the effectiveness of rMMEA and compare it with state-of-the-art MMEA methods.

Settings

We implement the backbone MMEA model following existing anti-modality-missing works. The hidden-size and batch-size are set to 300 and $3,500$, respectively. The maximal number of training epochs is set to 500, we use AdamW optimizer (Kingma and Ba 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for optimization. The basic learning rate is set to 0.003 and adjusted based on different datasets. We do not employ iterative training for all methods to ensure a fair comparison. Due to space constraints, detailed hyper-parameter settings are moved to the Appendix.

We primarily compare our method with the generation-based MMEA methods UMAEA (Chen et al. 2023) and GEEA (Guo et al. 2024b). UMAEA is the current state-of-the-art in anti-visual-missing MMEA, while GEEA is also

Model	$\text{DBP15K}_{\text{ZH-EN}}$			$\text{OpenEA}_{\text{EN-FR}}$		
	H@1	H@10	MRR	H@1	H@10	MRR
rMMEA	**.759**	**.939**	**.826**	**.628**	**.900**	**.726**
- *w/o* $\mathcal{L}_{rd}$	.723	.932	.803	.601	.889	.704
- *w/o* $\mathcal{L}_{mi}$	.736	.935	.812	.612	.893	.710
- *w/o OMMD*	.757	.939	.824	.627	**.900**	**.726**
- *w/ KLD*	.724	.937	.817	.607	.895	.707

Table 3: Ablation studies on $\text{DBP15K}_{\text{ZH-EN}}$ and $\text{OpenEA}_{\text{EN-FR}}$, with $r_{\text{vm}} = 0.95$. *w/o* $\mathcal{L}_{rd}$, *w/o* $\mathcal{L}_{mi}$, and *w/o OMMD* refer to the proposed method without ranking-based distillation, mutual information estimation, and other modality-missing data, respectively. *w/ KLD* refers to the method with conventional KL divergence distillation instead of ranking-based distillation.

based on generative models and capable of generating missing visual embeddings, but it is not tailored for visual missing setting. To demonstrate the effectiveness of the anti-modality-missing methods, we also include several conventional MMEA methods as baselines for comparison, such as EVA (Liu et al. 2021), MSNEA (Chen et al. 2022b), and MCLEA (Lin et al. 2022).

Datasets

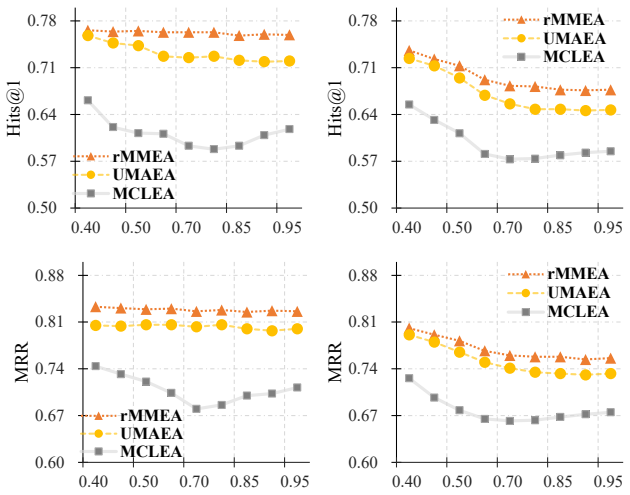
We use the multi-modal versions of DBP15K (Sun, Hu, and Li 2017) and OpenEA (Sun et al. 2020b) as benchmarks. DBP15K consists of three cross-lingual datasets: $\text{DBP15K}_{\text{ZH-EN}}$, $\text{DBP15K}_{\text{JA-EN}}$, and $\text{DBP15K}_{\text{FR-EN}}$, all sampled from DBPedia (Auer et al. 2007). The multi-modal version DBP15K was curated by (Liu et al. 2021). OpenEA is another popular EA benchmark, and its multi-modal version was curated by (Chen et al. 2023), containing two cross-lingual datasets, $\text{OpenEA}_{\text{EN-FR}}$ and $\text{OpenEA}_{\text{EN-DE}}$, as well as two cross-KG datasets (DBPedia to WikiData (Vrandečić and Krötzsch 2014)) $\text{OpenEA}_{\text{D-W-V1}}$ and $\text{OpenEA}_{\text{D-W-V2}}$. The statistics of all datasets can be found in the Appendix.

We conduct the visual missing experiments following UMAEA (Chen et al. 2023), with each dataset having four main settings for the *visual missing ratio*, i.e., $r_{\text{vm}} \in [0.4, 0.6, 0.8, 0.95]$, indicating the proportions of entities without image features in the KGs.

Main Results

Table 1 and Table 2 present the MMEA results under visual missing setting on the DBP15K and OpenEA datasets, respectively. It is evident that the anti-visual-missing methods UMAEA, GEEA, and rMMEA achieve significantly better performance compared to the conventional MMEA methods EVA, MSNEA, and MCLEA across all datasets and settings. Our rMMEA stands out as the leading approach, surpassing the current state-of-the-art UMAEA by substantial margins in most metrics. The Hits@1 and MRR results of rMMEA are even double or triple those of the conventional method MSNEA on $\text{OpenEA}_{\text{EN-FR}}$ and $\text{OpenEA}_{\text{D-W-V1}}$.

Notably, all methods exhibit better results under lower visual missing ratios. Some conventional methods, such as



a. $\text{DBP15K}_{\text{ZH-EN}}$ b. $\text{OpenEA}_{\text{D-W-V1}}$

Figure 3: The Hits@1 and MRR results on $\text{DBP15K}_{\text{ZH-EN}}$ and $\text{OpenEA}_{\text{D-W-V1}}$ w.r.t. various visual missing ratios.

EVA and MCLEA, also demonstrate a considerable anti-modality-missing capability, although there remains a substantial gap between them and rMMEA. Additionally, we observe that the Hits@10 results of the generation-based method UMAEA are often superior to its other metrics, possibly due to that the generated visual embedding is better in providing outline of the vision information.

In summary, our rMMEA consistently and significantly outperforms all baseline methods by a large margin across various metrics and datasets, demonstrating its superiority in anti-visual-missing. More detailed results are provided in the Appendix and reinforce the consistent findings.

Ablation Study

We perform ablation studies on the $\text{DBP15K}_{\text{ZH-EN}}$ and $\text{OpenEA}_{\text{EN-FR}}$ datasets to assess the effectiveness of each module in rMMEA. We design several variants for comparison: *w/o* $\mathcal{L}_{rd}$, *w/o* $\mathcal{L}_{mi}$, and *w/o OMMD* represent the proposed method without ranking-based distillation, mutual information estimation, and other modality-missing data, respectively. Additionally, *w/ KLD* denotes the method with a conventional KL divergence distillation objective instead of our ranking-based distillation.

Table 3 shows the ablation study results, where the proposed rMMEA achieves the best performance on both datasets across all metrics, indicating that the removal of any module leads to a performance decline. Particularly, replacing the ranking-based distillation with a standard KLD version (*w/ KLD*) performs slightly better than removing it (*w/o* $\mathcal{L}_{rd}$). Removing mutual information estimation also leads to a significant performance drop, while training rMMEA with other modality-missing data only marginally improves performance under visual-missing setting.

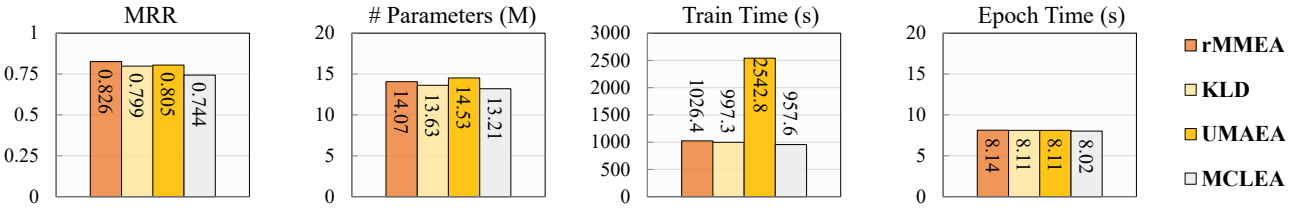
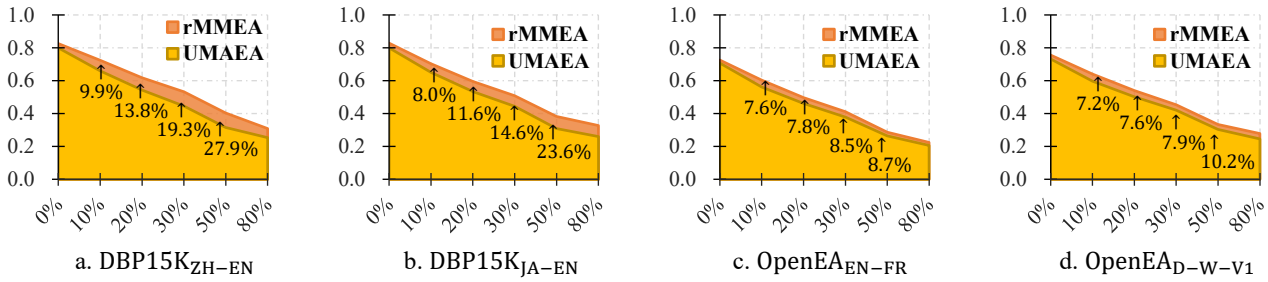


Figure 4: A comprehensive comparison of different methods on $\text{DBP15K}_{\text{ZH-EN}}$ using a single H100 GPU.



a. $\text{DBP15K}_{\text{ZH-EN}}$ b. $\text{DBP15K}_{\text{JA-EN}}$ c. $\text{OpenEA}_{\text{EN-FR}}$ d. $\text{OpenEA}_{\text{D-W-V1}}$

Figure 5: The MRR results of rMMEA and UMAEA (second-best method) w.r.t. different chaotic ratios on four datasets.

Ratio of Visual Missing Data

We conduct experiments to analyze how the ratio of visual missing data affects the performance of rMMEA. The results on the $\text{DBP15K}_{\text{ZH-EN}}$ and $\text{OpenEA}_{\text{D-W-V1}}$ datasets are illustrated in Figure 3, from which we can obverse that rMMEA consistently achieves the best Hits@1 and MRR results across different missing ratios. It is worth noting that, the relative performance gap between rMMEA and UMAEA widens as the proportion of visual-missing entities in KGs increases, highlighting the superior anti-visual-missing capability of our method.

Interestingly, the performance of MCLEA shows improvement in a reverse manner when more than 70% of entities lack visual features. This phenomenon may be attributed to that the model adjusts its attention away from visual embeddings, and suggests that the conventional MMEA methods may also have their own mechanisms for addressing visual-missing scenarios.

Effectiveness and Efficiency

We conduct further experiments to comprehensively compare rMMEA with anti-modality-missing and conventional MMEA methods.

Figure 4 presents a comparison of different methods on the $\text{DBP15K}_{\text{ZH-EN}}$ dataset, in terms of MRR, number of parameters, total training time, and per epoch time. It is clear that rMMEA and UMAEA exhibit the best performance. While the training time per epoch is similar across all methods, the total time for UMAEA is almost double that of others due to its multi-stage training strategy. Additionally, UMAEA employs more parameters because of the additional generative models it leverages. Comparing with UMAEA, the proposed rMMEA is undoubtedly a superior choice in anti-modality-missing scenarios.

MMEA with Chaotic Data

We design experiments to proportionally remove or replace other modality data with noise to evaluate rMMEA in a more challenging setting. Specifically, for each modality except visual and basic adjacency information, we first sample a constant proportion of entities. Subsequently, we replace the features of these entities in each modality with either zero or a random vector with equal probability for testing, and term this task *MMEA with chaotic data*.

Figure 5 illustrates the MRR results on four datasets ($r_{vm} = 0.95$) with chaotic ratios ranging from 0% to 80%. Evidently, our rMMEA outperforms UMAEA (the second-best method in previous experiments) across all four datasets and various chaotic ratios. Remarkably, although the performance of both methods decreases as more missing and noisy data is introduced, the performance curve of rMMEA is comparatively smoother. For example, as the chaotic ratio increases from 10% to 80%, the performance improvement of rMMEA compared to UMAEA also rises from 9.9% to 27.9%. Consequently, rMMEA demonstrates significantly better robustness in multi-modality-missing settings.

Conclusion

In this paper, we introduce rMMEA, a robust MMEA method designed to handle modality missing and noise scenarios. Our method surpasses state-of-the-art baselines on seven datasets across various metrics and under different visual-missing ratios. Additionally, it exhibits notable advantages over existing generation-based methods in terms of efficiency and robustness. In future, we plan to explore the integration of anti-modality-missing techniques with large language models (LLMs) (Zhang et al. 2024, 2025a), which also presents a potential limitation of our current method.

Acknowledgments

We would like to thank all anonymous reviewers for their insightful and invaluable comments. This work is founded by National Natural Science Foundation of China (62472311) and Key Research and Development Program of Ningxia Hui Autonomous Region (2023BEG02067).

References

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. G. 2007. DBpedia: A nucleus for a web of open data. In *ISWC*.

Belghazi, M. I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Hjelm, R. D.; and Courville, A. C. 2018. Mutual Information Neural Estimation. In *ICML*.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Mach. Learn.*, 79: 151–175.

Chen, C.; Dou, Q.; Jin, Y.; Liu, Q.; and Heng, P. 2022a. Learning With Privileged Multimodal Knowledge for Unimodal Segmentation. *IEEE Trans. Medical Imaging*, 41(3): 621–632.

Chen, L.; Li, Z.; Xu, T.; Wu, H.; Wang, Z.; Yuan, N. J.; and Chen, E. 2022b. Multi-modal Siamese Network for Entity Alignment. In *KDD*, 118–126.

Chen, M.; Tian, Y.; Yang, M.; and Zaniolo, C. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *IJCAI*.

Chen, Z.; Chen, J.; Zhang, W.; Guo, L.; Fang, Y.; Huang, Y.; Geng, Y.; Pan, J. Z.; Song, W.; and Chen, H. 2022c. MEAformer: Multi-modal Entity Alignment Transformer for Meta Modality Hybrid. *arXiv preprint arXiv:2212.14454*.

Chen, Z.; Guo, L.; Fang, Y.; Zhang, Y.; Chen, J.; Pan, J. Z.; Li, Y.; Chen, H.; and Zhang, W. 2023. Rethinking Uncertainly Missing and Ambiguous Visual Modality in Multi-Modal Entity Alignment. In *ISWC*, volume 14265 of *Lecture Notes in Computer Science*, 121–139. Springer.

Chen, Z.; Zhang, Y.; Fang, Y.; Geng, Y.; Guo, L.; Chen, X.; Li, Q.; Zhang, W.; Chen, J.; Zhu, Y.; Li, J.; Liu, X.; Pan, J. Z.; Zhang, N.; and Chen, H. 2024. Knowledge Graphs Meet Multi-Modal Learning: A Comprehensive Survey. *CoRR*, abs/2402.05391.

Courty, N.; Flamary, R.; Habrard, A.; and Rakotomamonjy, A. 2017. Joint distribution optimal transportation for domain adaptation. In *NeurIPS*, 3730–3739.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, 4171–4186. Association for Computational Linguistics.

Ganin, Y.; and Lempitsky, V. S. 2015. Unsupervised Domain Adaptation by Backpropagation. In *ICML*, 1180–1189.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks. arXiv:1406.2661.

Guo, L.; Bo, Z.; Chen, Z.; Zhang, Y.; Chen, J.; Lan, Y.; Sun, M.; Zhang, Z.; Luo, Y.; Li, Q.; Zhang, Q.; Zhang, W.; and Chen, H. 2024a. MKGL: Mastery of a Three-Word Language. In *NeurIPS*.

Guo, L.; Chen, Z.; Chen, J.; Fang, Y.; Zhang, W.; and Chen, H. 2024b. Revisit and Outstrip Entity Alignment: A Perspective of Generative Models. In *ICLR*. OpenReview.net.

Guo, L.; Chen, Z.; Chen, J.; Zhang, Q.; and Chen, H. 2024c. DET: A Dual-Encoding Transformer for Relational Graph Embedding. In *LREC/COLING*, 4685–4696. ELRA and ICCL.

Guo, L.; Chen, Z.; Chen, J.; Zhang, Y.; Sun, Z.; Bo, Z.; Fang, Y.; Liu, X.; Chen, H.; and Zhang, W. 2024d. Distributed representations of entities in open-world knowledge graphs. *Knowl. Based Syst.*, 290: 111582.

Guo, L.; Han, Y.; Zhang, Q.; and Chen, H. 2022a. Deep Reinforcement Learning for Entity Alignment. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of ACL*, 2754–2765.

Guo, L.; Sun, Z.; and Hu, W. 2019. Learning to Exploit Long-term Relational Dependencies in Knowledge Graphs. In *ICML*.

Guo, L.; Zhang, Q.; Sun, Z.; Chen, M.; Hu, W.; and Chen, H. 2022b. Understanding and Improving Knowledge Graph Embedding for Entity Alignment. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *ICML*, volume 162, 8145–8156.

Guo, L.; Zhang, Y.; Bo, Z.; Chen, Z.; Sun, M.; Zhang, Z.; Zhang, W.; and Chen, H. 2025. K-ON: Stacking Knowledge on the Head Layer of Large Language Model. In *AAAI*, 11745–11753. AAAI Press.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851.

Huang, Y.; Zhang, X.; Zhang, R.; Chen, J.; and Kim, J. 2024. Progressively Modality Freezing for Multi-Modal Entity Alignment. In *ACL (1)*, 3477–3489. Association for Computational Linguistics.

Jahanifar, M.; Raza, M.; Xu, K.; Vuong, T. T. L.; Jewsbury, R.; Shephard, A.; Zamanitajeddin, N.; Kwak, J. T.; Raza, S. E. A.; Minhas, F.; and Rajpoot, N. M. 2023. Domain Generalization in Computational Pathology: Survey and Guidelines. *CoRR*, abs/2310.19656.

Kim, J.; Kang, H.; Kim, S.; Kim, K.; and Park, C. 2025. Disentangling and Generating Modalities for Recommendation in Missing Modality Scenarios. In *SIGIR*, 1820–1829. ACM.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for stochastic optimization. In *ICLR*.

Kingma, D. P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; and Welling, M. 2016. Improved variational inference with inverse autoregressive flow. *NeurIPS*, 29.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86.

Lao, M.; Li, Z.; Guo, Y.; Zhang, X.; Cai, S.; Ding, Z.; and Li, H. 2025. Boosting Discriminability for Robust Multimodal Entity Linking with Visual Modality Missing. In *SIGIR*, 989–999. ACM.

Li, L.; Du, B.; Wang, Y.; Qin, L.; and Tan, H. 2020. Estimation of missing values in heterogeneous traffic data: Application of multimodal deep learning model. *Knowl. Based Syst.*, 194: 105592.

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

Lin, Z.; Zhang, Z.; Wang, M.; Shi, Y.; Wu, X.; and Zheng, Y. 2022. Multi-modal Contrastive Representation Learning for Entity Alignment. In *COLING*, 2572–2584.

Liu, F.; Chen, M.; Roth, D.; and Collier, N. 2021. Visual Pivoting for (Unsupervised) Entity Alignment. In *AAAI*, 4257–4266.

Luo, Y.; Chen, Z.; Guo, L.; Li, Q.; Zeng, W.; Cai, Z.; and Li, J. 2024. ASGEA: Exploiting Logic Rules from Align-Subgraphs for Entity Alignment. *CoRR*, abs/2402.11000.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR Workshop*.

Pei, S.; Yu, L.; Hoehndorf, R.; and Zhang, X. 2019. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference. In *WWW*, 3130–3136.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Shen, J.; Qu, Y.; Zhang, W.; and Yu, Y. 2018. Wasserstein Distance Guided Representation Learning for Domain Adaptation. In *AAAI*, 4058–4065.

Sikdar, A.; Teotia, J.; and Sundaram, S. 2025. OGP-Net: Optical Guidance Meets Pixel-Level Contrastive Distillation for Robust Multi-Modal and Missing Modality Segmentation. In *AAAI*, 6922–6930. AAAI Press.

Sun, Z.; Hu, W.; and Li, C. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *ISWC*.

Sun, Z.; Wang, C.; Hu, W.; Chen, M.; Dai, J.; Zhang, W.; and Qu, Y. 2020a. Knowledge Graph Alignment Network with Gated Multi-hop Neighborhood Aggregation. In *AAAI*.

Sun, Z.; Zhang, Q.; Hu, W.; Wang, C.; Chen, M.; Akrami, F.; and Li, C. 2020b. A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs. *CoRR*, abs/2003.07743.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Representation Distillation. In *ICLR*.

Tran, L.; Liu, X.; Zhou, J.; and Jin, R. 2017. Missing Modalities Imputation via Cascaded Residual Autoencoder. In *CVPR*, 4971–4980. IEEE Computer Society.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR*, abs/1807.03748.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *ICLR*.

Vrandečić, D.; and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57.

Wang, Q.; Zhan, L.; Thompson, P. M.; and Zhou, J. 2020. Multimodal Learning with Incomplete Modalities by Knowledge Distillation. In *KDD*, 1828–1838. ACM.

Xu, B.; Wang, N.; Chen, T.; and Li, M. 2015. Empirical Evaluation of Rectified Activations in Convolutional Network. *CoRR*, abs/1505.00853.

Zeng, W.; Zhao, X.; Tang, J.; and Lin, X. 2020. Collective Entity Alignment via Adaptive Features. In *ICDE*, 1870–1873.

Zhang, Y.; Chen, Z.; Guo, L.; Xu, Y.; Chen, S.; Sun, M.; Hu, B.; Zhang, Z.; Liang, L.; Zhang, W.; and Chen, H. 2025a. Have We Designed Generalizable Structural Knowledge Promptings? Systematic Evaluation and Rethinking. In *ACL (1)*, 2210–2226. Association for Computational Linguistics.

Zhang, Y.; Chen, Z.; Guo, L.; Xu, Y.; Zhang, M.; Zhang, W.; and Chen, H. 2025b. Abstractive Visual Understanding of Multi-modal Structured Knowledge: A New Perspective for MLLM Evaluation. *CoRR*, abs/2506.01293.

Zhang, Y.; Chen, Z.; Guo, L.; Xu, Y.; Zhang, W.; and Chen, H. 2024. Making Large Language Models Perform Better in Knowledge Graph Completion. In *ACM Multimedia*, 233–242. ACM.