Rethinking Uncertainly Missing and Ambiguous Visual Modality in Multi-Modal Entity Alignment

Zhuo Chen[1], Lingbing Guo[1], Yin Fang[1], Yichi Zhang[1], Jiaoyan Chen[4], Jeff Z. Pan[5], Yangning Li[6], Huajun Chen[1,2], and Wen Zhang[3*]

[1] College of Computer Science, Zhejiang University, Hangzhou, China
[2] Donghai laboratory, Zhoushan, China
[3] School of Software Technology, Zhejiang University, China
`{zhuo.chen,lbguo,fangyin,zhangyichi2022,zhang.wen,huajunsir}@zju.edu.cn`
[4] The University of Manchester & University of Oxford, UK
`jiaoyan.cheni@manchester.ac.uk`
[5] School of Informatics, The University of Edinburgh, Edinburgh, UK
`https://knowledge-representation.org/j.z.pan/`
[6] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
`liyn20@mails.tsinghua.edu.cn`

Abstract. As a crucial extension of entity alignment (EA), multi-modal entity alignment (MMEA) aims to identify identical entities across disparate knowledge graphs (KGs) by exploiting associated visual information. However, existing MMEA approaches primarily concentrate on the fusion paradigm of multi-modal entity features, while neglecting the challenges presented by the pervasive phenomenon of missing and intrinsic ambiguity of visual images. In this paper, we present a further analysis of visual modality incompleteness, benchmarking latest MMEA models on our proposed dataset MMEA-UMVM, where the types of alignment KGs covering bilingual and monolingual, with standard (non-iterative) and iterative training paradigms to evaluate the model performance. Our research indicates that, in the face of modality incompleteness, models succumb to overfitting the modality noise, and exhibit performance oscillations or declines at high rates of missing modality. This proves that the inclusion of additional multi-modal data can sometimes adversely affect EA. To address these challenges, we introduce UMAEA, a robust multi-modal **e**ntity **a**lignment approach designed to tackle **u**ncertainly **m**issing and **a**mbiguous visual modalities. It consistently achieves SOTA performance across all 97 benchmark splits, significantly surpassing existing baselines with limited parameters and time consumption, while effectively alleviating the identified limitations of other models. Our code and benchmark data are available at https://github.com/zjukg/UMAEA.

Keywords: Entity Alignment · Knowledge Graph · Multi-modal Learning · Uncertainly Missing Modality.

---

* Corresponding author.
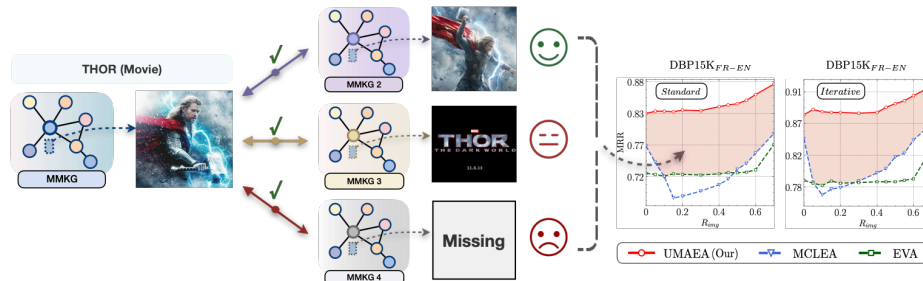
1   Introduction

Fig. 1. Phenomenon for missing and ambiguous visual modality in MMEA, where our UMAEA attains superior performance compared to MCLEA [20] and EVA [21].

Recently entity alignment (EA) has attracted wide attention as a crucial task for aggregating knowledge graphs (KGs) from diverse data sources. Multi-modal information, particularly visual images, serves as a vital supplement for entities. However, achieving visual modality completeness always proves challenging for automatically constructed KGs both on the Internet and domain-specific KGs. For instance, in the DBP15K datasets [28] for EA, only a portion of the entities have attached images (e.g., 67.58% in DBP15K$_{JA\text{-}EN}$ [21]). This incompleteness is inherent to the DBpedia KG [17], as not every entity possesses an associated image. Furthermore, the intrinsic ambiguity of visual images also impacts the alignment quality. As illustrated in Figure 1, the movie *THOR* can be represented by a snapshot of the movie (star) poster or an image of the movie title itself. While individuals familiar with the Marvel universe can effortlessly associate these patterns, machines struggle to discern significant visual feature association without the aid of external technologies like OCR and linking knowledge bases [7], posing challenges for alignment tasks. This phenomenon primarily arises from the abstraction of single-modal content, e.g., country-related images could be either national flags, landmarks or maps.

In this paper, we deliver an in-depth analysis of potential missing visual modality for MMEA. To achieve this, we propose the MMEA-UMVM dataset, which contains seven separate datasets with a total of 97 splits, each with distinct degrees of visual modality incompleteness, and benchmark several latest MMEA models. To ensure a comprehensive comparison, our dataset encompasses bilingual, monolingual, as well as normal and high-degree KG variations, with standard (non-iterative) and iterative training paradigms to evaluate the model performance. The robustness of the models against ambiguous images is discussed by comparing their performance under complete visual modality.

In our analysis, we identify two critical phenomena: (i) Models may succumb to overfitting noise during training, thereby affecting overall performance. (ii)

Models exhibit performance oscillations or even declines at high missing modality rates, indicating that sometimes the additional multi-modal data negatively impacts EA and leads to even worse results than when no visual modality information is used. These findings provide new insights for further exploration in this field. Building upon these observations, we propose our model UMAEA, which alleviates those shortcomings of other models via introducing multi-scale modality hybrid and circularly missing modality imagination. Experiments prove that our model can consistently achieve SOTA results across all benchmark splits with limited parameters and runtime, which supports our perspectives.

2   Related Work

Entity Alignment (EA) [29,10] is the task of identifying equivalent entities across multiple knowledge graphs (KGs), which can facilitate knowledge integration.

Typical Entity Alignment methods mainly rely on the relational, attribute, and surface (or literal) features of KG entity for alignment. Specifically, symbol logic-based technologies are used [13,27,24] to constrain the EA process via manually defined prior rules (e.g., logical reasoning and lexical matching). Embedding-based methods [29] eschew the ad-hoc heuristics of logic-based approaches, employing learned embedding space similarity measures for rapid alignment decisions. Among these, GNN-based EA models [18,32,23,38,9,37] emphasize local and global structural KG characteristics, primarily utilizing graph neural networks (GNNs) for neighborhood entity feature aggregation. While translation-based EA methods [43,31,39,2,12] use techniques like TransE [1] to capture the pairwise information from relational triples, positing that relations can be modeled as straightforward translations in the vector space.

Multi-modal Entity Alignment (MMEA) normally leverages visual modality as supplementary information to enhance EA, with each entity accompanied by a related image. Specifically, Chen et al. [4] propose to combine knowledge representations from different modalities, minimizing the distance between holistic embeddings of aligned entities. Liu et al. [21] use a learnable attention weighting scheme to assign varying importance to each modality. Chen et al. [5] incorporate visual features to guide relational feature learning while weighting valuable attributes for alignment. Lin et al. [20] further improve intra-modal learning with contrastive learning. Shi et al. [36] filter out mismatched images with pre-defined ontologies and an image type classifier. Chen et al. [8] dynamically predict the mutual modality weights for entity-level modality fusion and alignment.

These approaches substantiate that visual information indeed contributes positively to EA. However, we notice that all of them are based on two ideal assumptions: (i) Entities and images have a one-to-one correspondence, meaning that a single image sufficiently encapsulates and conveys all the information about an entity. (ii) Images are always available, implying that an entity consistently possesses a corresponding image.
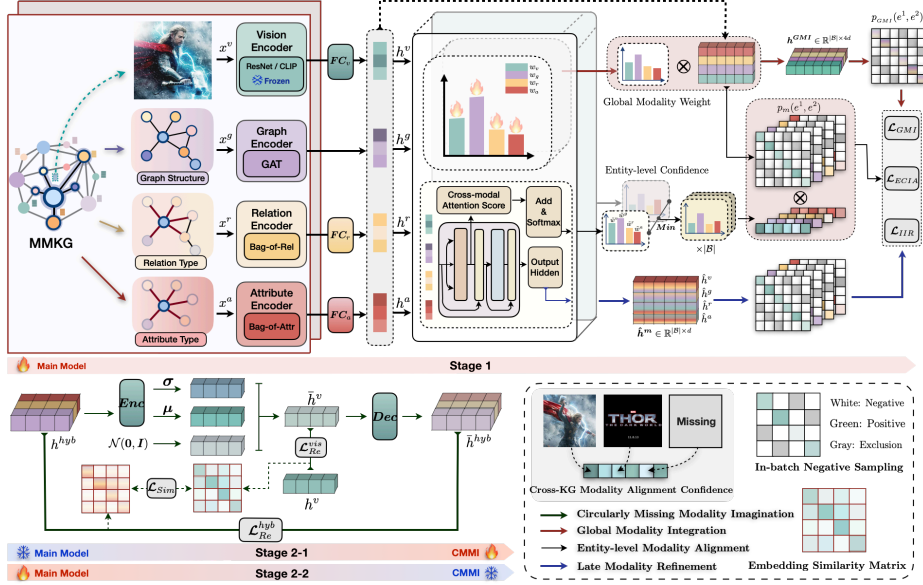
Fig. 2. The overall framework of UMAEA.

In real-world KGs, the noise is an inherent issue. Even for the standard MMEA datasets [28,22,4,21], they are hard to satisfy those two ideal conditions mentioned above. Consequently, we focus on two more pragmatic and demanding issues: (i) In MMKGs, entity images might be missing uncertainly, implying a varying degree of image absence. (ii) In MMKGs, images of the entities could be uncertainly ambiguous, suggesting that a single entity might have heterogeneous visual representations. To tackle these challenges, we present a benchmark consisting of seven datasets on which extensive experiments are conducted, and introduce our model UMAEA against these problems.

Incomplete Multi-modal Learning aims to tackle classification or reconstruction tasks, like multi-modal emotion recognition [44] and cross-modal retrieval [14], by leveraging information from available modalities when one modality is missing (e.g, a tweet may only have images or text content). In multi-modal alignment tasks, missing modality significantly impacts the performance as the symmetry of paired multi-modal data leads to noise accumulation when it is uncertain which side has modality incompleteness, further hindering model training. Prior MMEA studies [21,5,20,8] calculate mean and variance from available visual features, enabling random generation of those incomplete features using a normal distribution. In this paper, we develop an adaptive method for optimal training under the conditions with uncertainly missing or noisy visual modality, meanwhile providing a comprehensive benchmark.

3   Method

3.1   Preliminaries

We define a MMKG as a five-tuple $\mathcal{G}=\{\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}, \mathcal{T}\}$, where $\mathcal{E}, \mathcal{R}, \mathcal{A}$ and $\mathcal{V}$ denote the sets of entities, relations, attributes, and images, respectively. $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is the set of relation triples. Given two MMKGs $\mathcal{G}_1 = \{\mathcal{E}_1, \mathcal{R}_1, \mathcal{A}_1, \mathcal{V}_1, \mathcal{T}_1\}$ and $\mathcal{G}_2 = \{\mathcal{E}_2, \mathcal{R}_2, \mathcal{A}_2, \mathcal{V}_2, \mathcal{T}_2\}$, MMEA aims to discern each entity pair (e_i^1, e_i^2), $e_i^1 \in \mathcal{E}_1$, $e_i^2 \in \mathcal{E}_2$ where e_i^1 and e_i^2 correspond to an identical real-world entity e_i. For clarity, we omit the superscript symbol denoting the source KG of an entity in our context, except when explicitly required in statements or formulas. A set of pre-aligned entity pairs is provided, which is proportionally divided into a training set (i.e., seed alignments $\mathcal{S}$) and a testing set $\mathcal{S}_{te}$ based on a given seed alignment ratio (R_{sa}). We denote $\mathcal{M} = \{g, r, a, v\}$ as the set of available modalities. Commonly, in typical KG datasets for MMEA, each entity is associated with multiple attributes and 0 or 1 image, and the proportion (R_{img}) of entities containing images is uncertain (e.g., 67.58% in DBP15K$_{JA\text{-}EN}$ [21]). In this study, in order to facilitate a comprehensive evaluation, dataset MMEA-UMAM is proposed where we define R_{img} as a controlled variable for benchmarking.

3.2   Multi-modal Knowledge Embedding

Graph Structure Embedding. Let $x_i^g \in \mathbb{R}^d$ represent the randomly initialized graph embedding of entity e_i where d is the predetermined hidden dimension. We employ the Graph Attention Network (GAT) [35] with two attention heads and two layers to capture the structural information of $\mathcal{G}$, equipped with a diagonal weight matrix [40] $\boldsymbol{W}_g \in \mathbb{R}^{d \times d}$ for linear transformation. We define $h_i^g = GAT(\boldsymbol{W}_g, \boldsymbol{M}_g; x_i^g)$, where $\boldsymbol{M}_g$ denotes to the graph adjacency matrix.

Relation, Attribute, and Visual Embedding. To mitigate the information contamination arising from blending relation / attribute representations in GNN-like networks [21], we employ separate fully connected layers, parameterized by $\boldsymbol{W}_m \in \mathbb{R}^{d_m \times d}$, for embedding space harmonization via $h_i^m = FC_m(\boldsymbol{W}_m, x_i^m)$, where $m \in \{r, a, v\}$ and r, a, v, represent relation, attribute, visual modalities, respectively. Furthermore, $x_i^m \in \mathbb{R}^{d_m}$ denotes the input feature of entity e_i for the corresponding modality m. We follow Yang et al. [41] to use the bag-of-words features for relation (x^r) and attribute (x^a) representations (see Section 4.1 for details). While for the visual modality, we employ a pre-trained (frozen) visual model as the encoder (Enc_v) to obtain the visual embeddings x_i^v for each available image of the entity e_i. For entities without image data, we generate random image features using a normal distribution parameterised by the mean and standard deviation of other available images [21,5,20,8].

3.3   Multi-scale Modality Hybrid

This section describes the detailed architecture of the multi-scale modality hybrid for aligning multi-modal entities between MMKGs. The model comprises

three modality alignment modules operating at different scales, each associated with a training objective as depicted in Figure 2.

Global Modality Integration (GMI) emphasizes global alignment for each multi-modal entity pair, where the multi-modal embeddings for an entity are first concatenated and then aligned using a learnable global weight, allowing the model to adaptively learn the relative quality of each modality across two MMKGs. Let w_m be the global weight for modality m. We formulate the GMI joint embedding h_i^{GMI} for entity e_i as:

$$h_i^{GMI} = \bigoplus_{m \in \mathcal{M}} [w_m h_i^m], \qquad (1)$$

where $\bigoplus$ refers to the vector concatenation operation. To enhance model's sensitivity to feature differences between unaligned entities, we introduce a unified entity alignment contrastive learning framework, inspired by Lin et al. [20], to consolidate the training objectives of the modules. For each entity pair (e_i^1, e_i^2) in $\mathcal{S}$, we define $\mathcal{N}_i^{ng} = \{e_j^1 | \forall e_j^1 \in \mathcal{E}_1, j \neq i\} \cup \{e_j^2 | \forall e_j^2 \in \mathcal{E}_2, j \neq i\}$ as its negative entity set. To improve efficiency, we adopt the in-batch negative sampling strategy [6], restricting the sampling scope of $\mathcal{N}_i^{ng}$ to the mini-batch $\mathcal{B}$. Concretely, we define the alignment probability distribution as follows:

$$p_m(e_i^1, e_i^2) = \frac{\gamma_m(e_i^1, e_i^2)}{\gamma_m(e_i^1, e_i^2) + \sum_{e_j \in \mathcal{N}_i^{ng}} \gamma_m(e_i^1, e_j)}, \qquad (2)$$

where $\gamma_m(e_i, e_j) = \exp(h_i^{m\top} h_j^m / \tau)$ and τ represents the temperature hyperparameter. To account for the alignment direction of entity pairs in (2), we establish a bi-directional alignment objective as:

$$\mathcal{L}_m = -\mathbb{E}_{i \in \mathcal{B}} \log[p_m(e_i^1, e_i^2) + p_m(e_i^2, e_i^1)]/2, \qquad (3)$$

where m denotes a modality or an embedding type. We denote the training objective as $\mathcal{L}_{GMI}$ when the GMI join embedding is used, i.e., $\gamma_{GMI}(e_i, e_j)$ is set to $\exp(h_i^{GMI\top} h_j^{GMI} / \tau)$.

We note that the global adaptive weighting allows the model to capitalize on high-quality modalities while minimizing the impact of low-quality modalities, such as the redundant information within attributes / relations, and noise within images. Concurrently, it ensures the preservation of valuable information to a certain extent, ultimately contributing to the stability of the alignment process.

Entity-level Modality Alignment aims to perform instance-level modality weighting and alignment, utilizing minimum cross-KG confidence measures from seed alignments to constrain the modality alignment objectives. It allows the model to dynamically assign lower training weights to missing or ambiguous modality information, thereby reducing the risk of encoder misdirection arising from uncertainties. To achieve this, we follow Chen et al. [8] to adapt the

vanilla Transformer [34] for two types of sub-layers: the multi-head cross-modal attention (MHCA) block and the fully connected feed-forward networks (FFN).

Specifically, MHCA operates its attention function across N_h parallel heads. The i-th head is parameterized by modally shared matrices $\boldsymbol{W}_q^{(i)}$, $\boldsymbol{W}_k^{(i)}$, $\boldsymbol{W}_v^{(i)}$ $\in \mathbb{R}^{d \times d_h}$, transforming the multi-modal input h^m into modal-aware query $Q_m^{(i)}$, key $K_m^{(i)}$, and value $V_m^{(i)}$ in $\mathbb{R}^{d_h}$ ($d_h = d/N_h$):

$$Q_m^{(i)}, K_m^{(i)}, V_m^{(i)} = h^m \boldsymbol{W}_q^{(i)}, h^m \boldsymbol{W}_k^{(i)}, h^m \boldsymbol{W}_v^{(i)} \,. \tag{4}$$

MHCA generates the following output for a given feature of modality m:

$$\text{MHCA}(h^m) = \bigoplus_{i=1}^{N_h} \text{head}_i^m \cdot \boldsymbol{W}o \,, \tag{5}$$

$$\text{head}_i^m = \sum_{j \in \mathcal{M}} \beta_{mj}^{(i)} V_j^{(i)} \,, \tag{6}$$

where $\boldsymbol{W}_o \in \mathbb{R}^{d \times d}$. The attention weight ($\beta_{mj}$) between an entity's modality m and j in each head is calculated as:

$$\beta_{mj} = \frac{\exp(Q_m^\top K_j / \sqrt{d_h})}{\sum_{i \in \mathcal{M}} \exp(Q_m^\top K_i / \sqrt{d_h})} \,. \tag{7}$$

Besides, layer normalization (LN) and residual connection (RC) are incorporated to stabilize training:

$$\hat{h}^m = LayerNorm(\text{MHCA}(h^m) + h^m) \,. \tag{8}$$

The FFN consists of two linear transformation layers and a ReLU activation function with LN and RC applied afterwards:

$$\text{FFN}(\hat{h}^m) = ReLU(\hat{h}^m \boldsymbol{W}_1 + b_1) \boldsymbol{W}_2 + b_2 \,, \tag{9}$$

$$\hat{h}^m \leftarrow LayerNorm(\text{FFN}(\hat{h}^m) + \hat{h}^m) \,, \tag{10}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{d \times d_{in}}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d_{in} \times d}$. Notably, we define the entity-level confidence $\tilde{w}^m$ for each modality m as:

$$\tilde{w}^m = \frac{\exp(\sum_{j \in \mathcal{M}} \sum_{i=0}^{N_h} \beta_{mj}^{(i)} / \sqrt{|\mathcal{M}| \times N_h})}{\sum_{k \in \mathcal{M}} \exp(\sum_{j \in \mathcal{M}} \sum_{i=0}^{N_h} \beta_{kj}^{(i)} \sqrt{|\mathcal{M}| \times N_h})} \,, \tag{11}$$

which captures crucial inter-modal interface information and adaptively adjusts model's cross-KG alignment confidence for different modalities from each entity. To facilitate learning these dynamic confidences and incorporating them into the training process, we devise two distinct training objectives: $\mathcal{L}_{ECIA}$ and $\mathcal{L}_{IIR}$. The first objective is *explicit confidence-augmented intra-modal alignment* (ECIA), while the second is *implicit inter-modal refinement* (IIR), which will

be discussed in the following subsection. For the ECIA, we design the following training target which is the variation of Equation (3):

$$\mathcal{L}_{ECIA} = \sum\nolimits_{m \in \mathcal{M}} \widetilde{\mathcal{L}}_m \, , \tag{12}$$

$$\widetilde{\mathcal{L}}_m = -\mathbb{E}_{i \in \mathcal{B}} \, \log[\, \phi_m(e_i^1, e_i^2) * (p_m(e_i^1, e_i^2) + p_m(e_i^2, e_i^1))\,]/2 \, . \tag{13}$$

Considering the symmetric nature of EA and the varying quality of aligned entities and their modality features within each KG, we employ the minimum confidence value to minimize errors. For example, e_i^1 may possess high-quality image data while e_i^2 lacks image information, as illustrated in Figure 1. In such cases, using the original objective for feature alignment will inadvertently align meaningful features with random noise, thereby disrupting the encoder training process. To mitigate this issue, we define $\phi_m(e_i^1, e_i^2)$ as the minimum confidence value for entities e_i^1 and e_i^2 in modality m, calculated by $\phi_m(e_i, e_j) = Min(\tilde{w}_i^m, \tilde{w}_j^m)$.

Late Modality Refinement leverages the transformer layer outputs to further enhance the entity-level adaptive modality alignment through an *implicit inter-modal refinement* (IIR) objective, enabling the refinement of attention scores by directly aligning the output hidden states. Concretely, we define the hidden state embedding of modality m for entity e_i as $\hat{h}^m$, following Equation (10). We define:

$$\mathcal{L}_{IIR} = \sum\nolimits_{m \in \mathcal{M}} \widehat{\mathcal{L}}_m \, , \tag{14}$$

where $\widehat{\mathcal{L}}_m$ is also a variant of $\mathcal{L}_m$, as illustrated in Equation (3), with only the following modification: $\widehat{\gamma}_m(e_i, e_j) = \exp(\hat{h}_i^{m\top} \hat{h}_j^m / \tau)$.

As depicted in Figure 2, we designate the entire process so far as the first stage of our (main) model, with the training objective formulated as:

$$\mathcal{L}_1 = \mathcal{L}_{GMI} + \mathcal{L}_{ECIA} + \mathcal{L}_{IIR} \, . \tag{15}$$

3.4   Circularly Missing Modality Imagination.

Note that our primary target of the first stage is to alleviate the impact of modality noise and incompleteness on the alignment process throughout training. Conversely, the second stage draws inspiration from VAE [15,26] and CycleGAN [45], which accentuates generative modeling and unsupervised domain translation. Expanding upon these ideas, we develop our circularly missing modality imagination (CMMI) module, aiming to enable the model to proactively complete missing modality information.

To reach our goal, we develop a variational multi-modal autoencoder framework, allowing the hidden layer output between the encoder MLP_{Enc} and decoder MLP_{Dec} (parameterized by $\boldsymbol{W}_{Enc} \in \mathbb{R}^{3d \times 2d}$ and $\boldsymbol{W}_{Dec} \in \mathbb{R}^{d \times 3d}$, respectively) to act as an imagined pseudo-visual feature $\bar{h}_i^v$, using reparameterization

strategy [15] with tri-modal hybrid feature $h_i^{hyb} = [h_i^r \oplus h_i^a \oplus h_i^g]$ as the input:

$$[\,\mu_i \oplus \log(\sigma_i)^2\,] = MLP_{Enc}(h_i^{hyb})\,, \tag{16}$$

$$\bar{h}_i^v = z \odot \sigma_i + \mu_i\,, \ \ z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})\,, \tag{17}$$

$$\bar{h}_i^{hyb} = MLP_{Dec}(\bar{h}_i^v)\,. \tag{18}$$

Concretely, two reconstruction objectives $\mathcal{L}_{Re}^{vis}$ and $\mathcal{L}_{Re}^{hyb}$ are utilized to minimize $|h_i^{hyb} - \bar{h}_i^{hyb}|$ and $|h_i^v - \bar{h}_i^v|$, where h_i^v represents the real image feature. Besides, we adhere to the standard VAE algorithm [15] to regularize the latent space by encouraging it to be similar to a Gaussian distribution through minimizing the Kullback–Leibler (KL) divergence:

$$\mathcal{L}_{KL} = \mathbb{E}_{i \in \bar{\mathcal{B}}}\left((\mu_i)^2 + (\sigma_i)^2 - \log(\sigma_i)^2 - 1\right)/2\,, \tag{19}$$

where $\bar{\mathcal{B}}$ refers to those entities with complete images within a mini-batch.

Furthermore, we exploit the internal embedding similarity matrix obtained from the hybrid embeddings h^{hyb}, and distill this information into the virtual image feature similarity matrix based on $\bar{h}^v$:

$$\mathcal{L}_{Sim} = \mathbb{E}_{i \in \bar{\mathcal{B}}}\, D_{KL}(p_{hyb}(e_i^1, e_i^2)||\bar{p}_v(e_i^1, e_i^2))\,, \tag{20}$$

where p_{hyb} and $\bar{p}_v$ all follow Equation (2) with $\gamma_{hyb}(e_i, e_j) = \exp(h_i^{hyb\top} h_j^{hyb}/\tau)$ and $\bar{\gamma}_v(e_i, e_j) = \exp(\bar{h}_i^{v\top} \bar{h}_j^v/\tau)$. This strategy not only curbs the overfitting of visible visual modalities in the autoencoding process, but also emphasizes the differences between distinct characteristics. Crucially, the knowledge mapping of original tri-modal hybrid features to the visual space is maximally preserved, thereby mitigating modal collapse when most of the visual content is missing and the noise is involved. The final loss in stage two is formulated as:

$$\mathcal{L}_2 = \mathcal{L}_{KL} + \mathcal{L}_{Re}^{vis} + \mathcal{L}_{Re}^{hyb} + \mathcal{L}_{Sim}\,. \tag{21}$$

3.5   Training Details

Pipeline. As previously mentioned, the training process consists of two stages. In the first stage, the primary model components are trained independently, while in the second stage, the CMMI module is additionally incorporated. The training objective $\mathcal{L}$ is defined as follows:

$$Stage\ 1 : \mathcal{L} \leftarrow \mathcal{L}_1\,, \tag{22}$$

$$Stage\ 2\text{-}1/2\text{-}2 : \mathcal{L} \leftarrow \mathcal{L}_1 + \mathcal{L}_2\,, \tag{23}$$

where the second stage is further divided into two sub-stages. Concretely, in order to stabilize model training and avoid knowledge forgetting caused by the cold-start of module insertion [42], as shown in Figure 2, the models from stage 1 (i.e., main model) are frozen to facilitate CMMI training when entering stage 2-1. While in stage 2-2, the CMMI is frozen and the main model undergoes further refinement to establish the entire pipeline. This process is easy to implement, just by switching the range of learnable parameters during model training.

Entity Representation. During evaluation, we replace the original random vectors with the generated μ_i for those entities without images. While in the second training stage, we employ the pseudo-visual embedding $\bar{h}_i^v$ (rather than μ_i) as a substitute as we observe that actively introducing noise during training could introduce randomness and uncertainty into the reconstruction process, which has been demonstrated to be beneficial in learning sophisticated distributions and enhances the model's robustness [16]. Furthermore, we select h_i^{GMI}, as formulated in Equation (1), for the final multi-modal entity representation.

4   Experiment

4.1   Experiment Setup

To guarantee a fair assessment, we use a total of seven MMEA datasets derived from three major categories (bilingual, monolingual, and high-degree), with two representative pre-trained visual encoders (ResNet-152 [11] and CLIP [25]), and evaluated the performance of four models under two distinct settings (standard (non-iterative) and iterative). In this research, we intentionally set aside the surface modality (literal information) to focus on understanding the effects of absent visual modality on model performance.

Datasets. DBP15K [28] contains three datasets ($R_{sa} = 0.3$) built from the multilingual versions of DBpedia, including DBP15K$_{ZH\text{-}EN}$, DBP15K$_{JA\text{-}EN}$ and DBP15K$_{FR\text{-}EN}$. We adopt their multi-model variants [21] with entity-matched images attached. Besides, four Multi-OpenEA datasets ($R_{sa} = 0.2$) [19] are used, which are the multi-modal variants of the OpenEA benchmarks [33] with entity images achieved by searching the entity names through the Google search engine. We include two bilingual datasets { EN-FR-15K, EN-DE-15K } and two monolingual datasets { D-W-15K-V1, D-W-15K-V2 }, where V1 and V2 denote two versions with distinct average relation degrees. To create our **MMEA-UMVM** (uncertainly missing visual modality) datasets, we perform random image dropping on MMEA datasets. Specifically, we randomly discard entity images to achieve varying degrees of visual modality missing, ranging from 0.05 to the maximum R_{img} of the raw datasets with a step of 0.05 or 0.1. Finally, we get a total number of 97 data split. See appendix [7] for more details.

Iterative Training. Following Lin et al. [20], we adopt a probation technique for iterative training. The probation can be viewed as a buffering mechanism, which maintains a temporary cache to store cross-graph mutual nearest entity pairs from the testing set. Concretely, every K_e (where $K_e = 5$) epochs, we propose cross-KG entity pairs that are mutual nearest neighbors in the vector space and add them to a candidate list $\mathcal{N}^{cd}$. Furthermore, an entity pair in $\mathcal{N}^{cd}$ will be added into the training set if it remains a mutual nearest neighbour for K_s ($= 10$) consecutive rounds.

---

[7] The appendix is attached with the arXiv version of this paper.

Baselines. Six prominent EA algorithms proposed in recent years are selected as our baseline comparisons, excluding the surface information for a parallel evaluation. We further collect 3 latest MMEA methods as the strong baselines, including EVA [21], MSNEA [5], and MCLEA [20]. Particularly, we reproduce them with their original pipelines unchanged in our benchmark.

Table 1. Non-iterative results of four models with "w/o CMMI" setting indicating the absence of the stage-2. The best results within the baselines are marked with underline, and we highlight our results with **bold** when we achieve SOTA.

	Models	$R_{img}=0.05$			$R_{img}=0.2$			$R_{img}=0.4$			$R_{img}=0.6$		
		H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
DBP15K$_{ZH-EN}$	MSNEA [5]	.413	.722	.517	.411	.725	.518	.446	.743	.546	.520	.786	.611
	EVA [21]	.623	.878	.715	.624	.878	.716	.623	.875	.714	.625	.876	.717
	MCLEA [20]	.638	.905	.732	.588	.865	.686	.611	.874	.704	.661	.896	.744
	w/o CMMI	.703	.934	.787	.710	.937	.793	.721	.939	.801	.753	.949	.825
	UMAEA	**.720**	**.938**	**.800**	**.727**	**.941**	**.806**	**.727**	**.941**	**.806**	**.758**	**.951**	**.829**
	Improve ↑	8.2%	3.3%	.068	10.3%	6.3%	.090	10.4%	6.6%	.092	9.7%	5.5%	.085
DBP15K$_{JA-EN}$	MSNEA [5]	.313	.643	.425	.311	.644	.422	.369	.678	.472	.480	.744	.569
	EVA [21]	.615	.877	.708	.616	.877	.710	.616	.878	.711	.624	.881	.716
	MCLEA [20]	.599	.897	.706	.579	.846	.675	.613	.867	.703	.686	.898	.761
	w/o CMMI	.708	.943	.794	.712	.947	.798	.730	.950	.810	.772	.962	.843
	UMAEA	**.725**	**.949**	**.807**	**.726**	**.949**	**.808**	**.732**	**.952**	**.813**	**.775**	**.963**	**.845**
	Improve ↑	11.0%	5.2%	.099	11.0%	7.2%	.098	11.6%	7.4%	.102	8.9%	6.5%	.084
DBP15K$_{FR-EN}$	MSNEA [5]	.297	.690	.427	.304	.690	.428	.360	.710	.474	.478	.772	.574
	EVA [21]	.624	.895	.720	.624	.895	.720	.626	.898	.721	.634	.900	.728
	MCLEA [20]	.634	.930	.741	.582	.863	.682	.601	.879	.702	.675	.901	.757
	w/o CMMI	.727	.956	.813	.733	.960	.817	.746	.961	.828	.790	.968	.857
	UMAEA	**.752**	**.970**	**.830**	**.755**	**.960**	**.832**	**.763**	**.962**	**.838**	**.792**	**.970**	**.859**
	Improve ↑	11.8%	4.0%	.089	13.1%	6.7%	.112	13.7%	6.4%	.117	11.7%	6.9%	.102
OpenEA$_{EN-FR}$	MSNEA [5]	.200	.431	.278	.213	.439	.290	.260	.477	.334	.360	.560	.427
	EVA [21]	.528	.833	.634	.533	.835	.638	.539	.835	.642	.547	.830	.647
	MCLEA [20]	.545	.852	.653	.547	.852	.655	.531	.839	.637	.597	.852	.688
	w/o CMMI	.587	.893	.695	.590	.893	.697	.614	.900	.715	.664	.912	.753
	UMAEA	**.605**	**.898**	**.708**	**.604**	**.896**	**.708**	**.618**	.899	**.718**	**.665**	**.914**	**.753**
	Improve ↑	6.0%	4.6%	.055	5.7%	4.4%	.053	7.9%	6.1%	.076	6.8%	6.2%	.065
OpenEA$_{EN-DE}$	MSNEA [5]	.242	.486	.323	.253	.495	.333	.309	.542	.387	.412	.622	.484
	EVA [21]	.717	.917	.787	.718	.918	.788	.721	.920	.791	.734	.921	.800
	MCLEA [20]	.723	.918	.791	.721	.915	.789	.697	.907	.771	.745	.906	.803
	w/o CMMI	.752	.938	.818	.757	.941	.822	.771	.946	.833	.804	.954	.858
	UMAEA	**.757**	**.942**	**.823**	**.759**	**.943**	**.824**	**.774**	**.947**	**.835**	**.804**	**.957**	**.860**
	Improve ↑	3.4%	2.4%	.032	3.8%	2.5%	.035	5.3%	2.7%	.044	5.9%	3.6%	.057
OpenEA$_{D-W-V1}$	MSNEA [5]	.238	.452	.31	.254	.465	.326	.318	.514	.385	.432	.601	.490
	EVA [21]	.570	.801	.653	.575	.806	.658	.567	.797	.650	.595	.811	.673
	MCLEA [20]	.585	.834	.675	.574	.824	.663	.581	.813	.665	.655	.848	.726
	w/o CMMI	.640	.879	.727	.644	.882	.730	.667	.891	.749	.722	**.908**	.790
	UMAEA	**.647**	**.881**	**.733**	**.649**	**.882**	**.735**	**.669**	**.892**	**.750**	**.724**	**.908**	**.791**
	Improve ↑	6.2%	4.7%	.058	7.4%	5.8%	.072	8.8%	7.9%	.085	6.9%	6.0%	.065
OpenEA$_{D-W-V2}$	MSNEA [5]	.397	.690	.497	.405	.695	.503	.454	.727	.546	.545	.781	.626
	EVA [21]	.775	.952	.839	.767	.947	.832	.773	.950	.837	.788	.954	.848
	MCLEA [20]	.771	.965	.842	.753	.957	.827	.757	.935	.822	.800	.948	.855
	w/o CMMI	.828	.983	.883	.829	.982	.885	.844	.984	.896	.857	.986	.905
	UMAEA	**.840**	**.984**	**.890**	**.832**	**.982**	**.887**	**.844**	**.984**	**.896**	**.859**	**.987**	**.905**
	Improve ↑	6.5%	1.9%	.048	6.5%	2.5%	.055	7.1%	3.4%	.059	5.9%	3.3%	.050

Implementation Details. To ensure fairness, we consistently reproduce or implement all methods with the following settings: (i) The hidden layer dimensions d for all networks are unified into 300. The total epochs for baselines are set to 500 with an optional iterative training strategy applied for another 500 epochs, following [20]. Training strategies including cosine warm-up schedule (15% steps

for LR warm-up), early stopping, and gradient accumulation are adopted. The AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) is used, with a fixed batch size of 3500. (ii) To demonstrate model stability, following [4,20], the vision encoders Enc_v are set to ResNet-152 [11] on DBP15K where the vision feature dimension d_v is 2048, and set to CLIP [25] on Multi-OpenEA with $d_v = 512$. (iii) An alignment editing method is employed to reduce the error accumulation [30]. (iv) Following Yang et al. [41], Bag-of-Words (BoW) is selected for encoding relations (x^r) and attributes (x^a) as fixed-length (i.e., $d_r = d_a = 1000$) vectors. Specially, we firstly sort relations/attributes across KGs by frequencies in descending order. At rank d_r/d_a, we truncated or padded the list to discard the long-tail relations/attributes and obtain fixed-length all-zero vectors x^r and x^a. For entity e_i: if it includes any of the top-k attributes, the corresponding position in x_i^a is set to 1; if a relation of e_i is among the top-k, the corresponding position in x_i^r is incremented by 1.

In our UMAEA model, τ is set to 0.1 which determines how much attention the contrast loss pays to difficult negative samples. Besides, the head number N_h in MHCA is set to 1, and the training epochs are set to {250, 50, 100} for stage 1, 2-1, 2-2, respectively. Despite potential performance variations resulting from parameter searching, our focus remained on achieving broad applicability rather than fine-tuning for specific datasets. During iterative training, the pipeline is repeated; but the expansion of the training set occurs exclusively in stage 1. For MSNEA, we eliminate the attribute values for input consistency, and extend MSNEA with iterative training capability. All experiments are conducted on RTX 3090Ti GPUs.

4.2   Overall Results

Uncertainly Missing Modality. Our primary experiment focuses on the model performances with varying missing modality proportions R_{img}. In Table 1, we select four representative proportions: $R_{img} \in \{0.05, 0.2, 0.4, 0.6\} \times 100\%$ to simulate the degree of uncertainly missing modality that may exist in real-world scenarios, and evaluate the robustness of different models. Our UMAEA demonstrates stable improvement on the DBP15K datasets across different R_{img} values in comparison to the top-performing benchmark model: 10.3% ($R_{img} = 0.05$), 11.6% ($R_{img} = 0.2$), 11.9% ($R_{img} = 0.4$), and 10.3% ($R_{img} = 0.6$). We note that it exhibits the most significant improvement when the R_{img} lies between 20% and 40%. For the Multi-OpenEA datasets, our average improvement is: 5.5% ($R_{img} = 0.05$), 5.9% ($R_{img} = 0.2$), 7.3% ($R_{img} = 0.4$), and 6.4% ($R_{img} = 0.6$). Although the improvement is slightly lower than in DBP15K, the overall advantage range remains consistent, aligning with our motivation. Besides, Figure 3 visualizes performance variation curves for three models. The overall performance trend fits the conclusions drawn in Table 1, showing that our method outperforms the baseline in terms of significant performance gap, regardless of whether iterative or non-iterative learning is employed.

Additionally, we notice a phenomenon that existing models exhibit performance oscillations (EVA) or even declines (MCLEA) at higher modality missing

rates. This kind of adverse effect peaks within a particular R_{img}^1 range and gradually recovers and gains benefits as R_{img} rises to a certain level R_{img}^2. In other words, when $0 \leq R_{img} \leq R_{img}^2$, the additional multi-modal data negatively impacts EA. This observation seems counterintuitive since providing more information leads to side effects, but it is also logical. Introducing images for half of the entities means that the remaining half may become noise, which calls for a necessary trade-off. Under the standard (non-iterative) setting, MCLEA's R_{img}^2 averages 63.6%, which is 57.14% for MSNEA and 46.43% for EVA across seven datasets. Our method, augmented with the CMMI module, reaches 20.71% for R_{img}^2. Even without CMMI, the R_{img}^2 of UMAEA remains at 34.29%. This implies that our method can gain benefits with fewer visual modality data in entity. Meanwhile, UMAEA exhibits less oscillation and greater robustness than other methods, as further evidenced by the entity distribution analysis in Section 4.3.
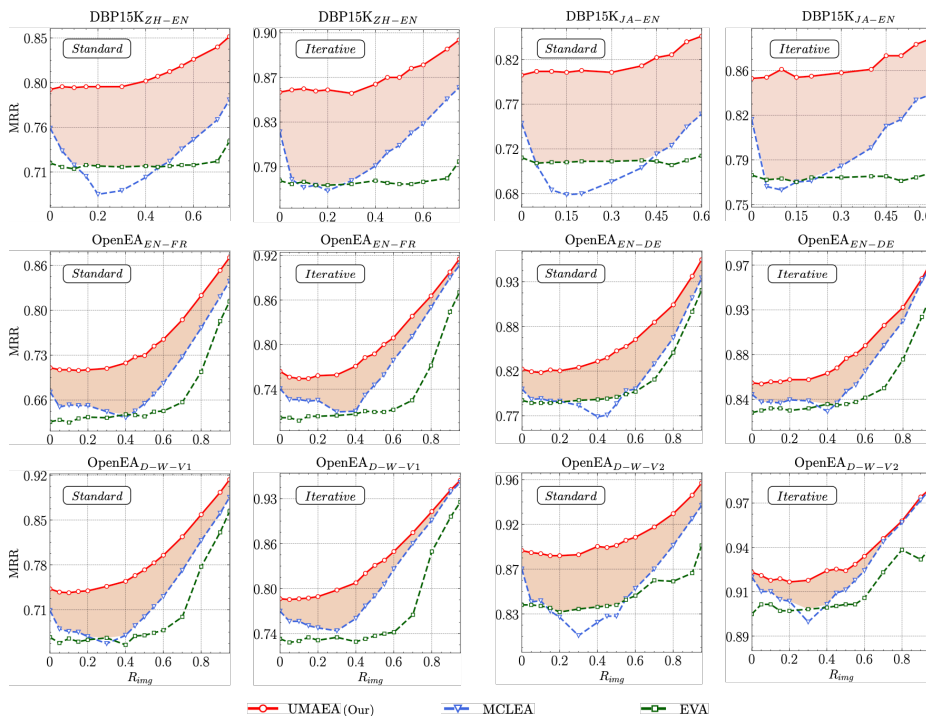

Fig. 3. The overall standard (non-iterative) and iterative model performance under the setting of uncertainly missing modality with $R_{img} \in \{0.2, 0.4, 0.6\}$. The performance of DBP15K$_{FR\text{-}EN}$ are shown in Figure 1.

We observe that our performance improvement on Multi-OpenEA is less pronounced compared to the DBP15K dataset. This may be due to the higher

image feature quality of CLIP compared to ResNet-152, which in turn diminishes the relative benefit of our model in addressing feature ambiguity. Additionally, as the appendix shows, these datasets have fewer relation and attribute types, allowing for better feature training with comparable data sizes (with a fixed 1000-word bag size, long tail effects are minimized) which partially compensates for missing image modalities. This finding can also explain why, as seen in Figure 3, our model's performance improvement decreases as R_{img} increases, and our enhancement in the dense graph (D-W-V2) is slightly less pronounced than in the sparse graph (D-W-V1) which has richer graph structure information.

Complete Modality. We also evaluate our model on the standard multi-modal DBP15K [21] dataset, achieving satisfactory results with or without the visual modality (w/o IMG), as shown in Table 2. It is noteworthy that the DBP15K dataset only has part of the entities with images attached (e.g., 78.29% in DBP15K$_{ZH\text{-}EN}$, 70.32% in DBP15K$_{FR\text{-}EN}$, and 67.58% in DBP15K$_{JA\text{-}EN}$), which is inherent to the DBPedia database. To further showcase our method's adaptability, in Table 3, we evaluate it on the standard Multi-OpenEA dataset with 100% image data attached, demonstrating that our method can be superior in the (MM)EA task against the potentially ambiguous modality information.

Table 2. Non-iterative (Non-iter.) and iterative (Iter.) results on three multi-modal DPB15K [28] datasets, where " * " refers to involving the visual information for EA.

	Models	DBP15K$_{ZH-EN}$			DBP15K$_{JA-EN}$			DBP15K$_{FR-EN}$		
		H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
Non-iter.	AlignEA [30]	.472	.792	.581	.448	.789	.563	.481	.824	.599
	KECG [18]	.478	.835	.598	.490	.844	.610	.486	.851	.610
	MUGNN [3]	.494	.844	.611	.501	.857	.621	.495	.870	.621
	AliNet [32]	.539	.826	.628	.549	.831	.645	.552	.852	.657
	MSNEA* [5]	.609	.831	.685	.541	.776	.620	.557	.820	.643
	EVA* [21]	.683	.906	.762	.669	.904	.752	.686	.928	.771
	MCLEA* [20]	.726	.922	.796	.719	.915	.789	.719	.918	.792
	UMAEA*	**.800**	**.962**	**.860**	**.801**	**.967**	**.862**	**.818**	**.973**	**.877**
	w/o IMG	**.718**	**.930**	**.797**	**.723**	**.941**	**.803**	**.748**	**.956**	**.826**
Iter.	BootEA [30]	.629	.847	.703	.622	.854	.701	.653	.874	.731
	NAEA [46]	.650	.867	.720	.641	.873	.718	.673	.894	.752
	MSNEA* [5]	.648	.881	.728	.557	.804	.643	.583	.848	.672
	EVA* [21]	.750	.912	.810	.741	.921	.807	.765	.944	.831
	MCLEA* [20]	.811	.957	.865	.805	.958	.863	.808	.963	.867
	UMAEA*	**.856**	**.974**	**.900**	**.857**	**.980**	**.904**	**.873**	**.988**	**.917**
	w/o IMG	**.793**	**.952**	**.852**	**.794**	**.960**	**.857**	**.820**	**.976**	**.880**

4.3   Details Analysis

Component Analysis. We further analyze the impact of each training objective on our model's performance in Figure 4, where the absence of any objective results in varying performance degradation. As mentioned in Section 3.3, IIR serves as an enhancement for ECIA, and its influence is comparatively less significant than that of $\mathcal{L}_{GMI}$ and $\mathcal{L}_{ECIA}$. The CMMI module's influence is detailed

Table 3. Non-iterative (Non-iter.) and iterative (Iter.) results on four standard Multi-OpenEA [19] datasets with $R_{img} = 1.0$.

	Models	OpenEA$_{EN-FR}$			OpenEA$_{EN-DE}$			OpenEA$_{D-W-V1}$			OpenEA$_{D-W-V2}$		
		H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
Non-iter.	MSNEA* [5]	.692	.813	.734	.753	.895	.804	.800	.874	.826	.838	.940	.873
	EVA* [21]	.785	.932	.836	.922	.983	.945	.858	.946	.891	.890	.981	.922
	MCLEA* [20]	.819	.943	.864	.939	.988	.957	.881	.955	.908	.928	.983	.949
	UMAEA*	**.848**	**.966**	**.891**	**.956**	**.994**	**.971**	**.904**	**.971**	**.930**	**.948**	**.996**	**.967**
Iter.	MSNEA* [5]	.699	.823	.742	.788	.917	.835	.809	.885	.836	.862	.954	.894
	EVA* [21]	.849	.974	.896	.956	.985	.968	.915	.986	.942	.925	.996	.951
	MCLEA* [20]	.888	.979	.924	.969	.993	.979	.944	.989	.963	.969	.997	.982
	UMAEA*	**.895**	**.987**	**.931**	**.974**	**998**	**.984**	**.945**	**.994**	**.965**	**.973**	**.999**	**.984**
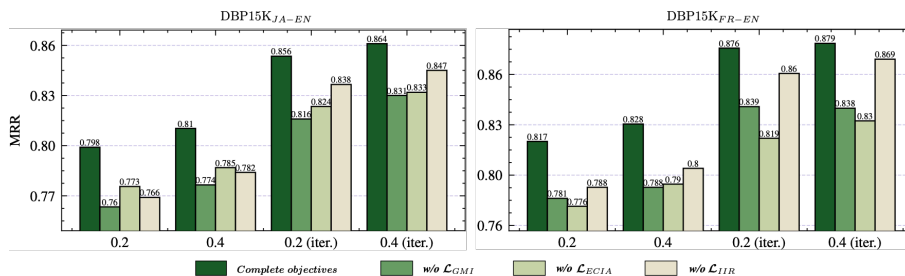
Fig. 4. The component analysis of UMAEA (w/o CMMI), where the scales on the horizontal axis represent $R_{img} \in \{0.2, 0.4\}$ and "iter." represents the model performance on iterative setting.

in Table 1, where it becomes more significant when R_{img} is low. CMMI's primary function is to mitigate noise in the missing modalities, facilitating efficient learning at high noise levels and minimizing the noise to existing information.

Efficiency Analysis. Concurrently, we briefly compare the relationship between model parameter size, training time, and performance. Our model improves the performance with only a minor increase in parameters and time consumption. This indicates that in many cases, our method can directly substitute these models with minimal additional overhead. While there is potential for enhancing UMAEA's efficiency, we view this as a direction for future research.

Table 4. Efficiency Analysis. Non-iterative model performance on three datasets with $R_{img} = 0.4$, where "Para." refers to the number of learnable parameters and "Time" refers to the total time required for model to reach the optimal performance.

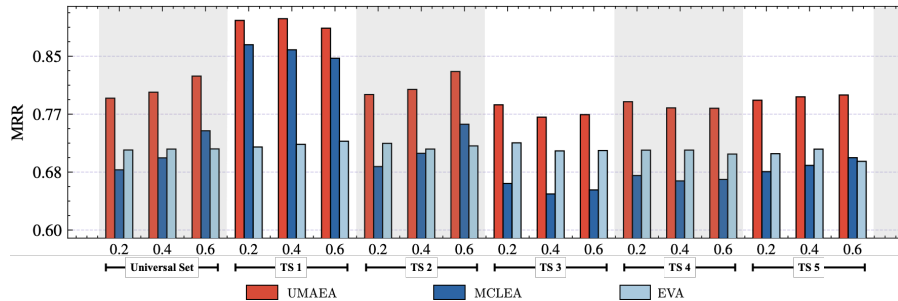| Models	DBP15K$_{JA-EN}$			DBP15K$_{FR-EN}$			OpenEA$_{EN-FR}$		
	Para. (M)	Time (Min)	MRR	Para. (M)	Time (Min)	MRR	Para. (M)	Time (Min)	MRR
EVA* [21]	13.27	30.9	.711	13.29	30.8	.721	9.81	17.8	.642
MCLEA* [20]	13.22	15.3	.703	13.24	15.7	.702	9.75	19.5	.637
w/o CMMI	13.82	30.2	.810	13.83	28.8	.828	10.35	17.9	.715
UMAEA	14.72	33.4	.813	14.74	32.7	.838	11.26	23.1	.718

Fig. 5. EA prediction distribution analysis on DBP15K$_{ZH-EN}$ (non-iterative), with $R_{img} \in \{0.2, 0.4, 0.6\}$. "TS" denotes the testing set, where: TS 1 (both entities in an alignment pair have images); TS 2 (at least one entity in an alignment pair has images); TS 3 (only one entity in an alignment pair has images); TS 4 (at least one entity in an alignment pair loss images); TS 5 (neither entity in an alignment pair has images).

Entity Distribution Analysis. To further evaluate the robustness of our method, we analyze the model's prediction performance under different distributions of entity's visual modality. Concretely, we compare five testing sets under $R_{img} \in \{0.2, 0.4, 0.6\}$ with details presented in Figure 5, where we exclude the CMMI module during the comparison. We observe that EVA's performance is generally stable but underperforms when visual modality is complete (TS 1), suggesting its overfitting to modality noise in the training stage. In contrast, MCLEA exhibits more extreme performance fluctuations, performing worse than EVA does when there's incomplete visual information within the entity pairs (TS 2, 3, 4, 5). Our superior performance reflects the intuition that the optimal performance occurs in TS 1, with tolerable fluctuations in other scenarios.

5   Conclusion

In this work, we discussed the challenges and limitations of existing MMEA methods in dealing with modality incompleteness and visual ambiguity. Our analysis revealed that certain models overfit to modality noise and suffer from oscillating or declining performance at high modality missing rates, emphasizing the need for a more robust approach. Thus, we introduced UMAEA which introduces multi-scale modality hybrid and circularly missing modality imagination to tackle this problem, performing well across all benchmarks. There remain opportunities for future research, such as evaluating our techniques for the incompleteness of other modalities (e.g., attribute), and investigating effective techniques to utilize more detailed visual contents for MMEA.

References

1. Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NIPS. pp. 2787–2795 (2013)
2. Cai, W., Ma, W., Zhan, J., Jiang, Y.: Entity alignment with reliable path reasoning and relation-aware heterogeneous graph transformer. In: IJCAI. pp. 1930–1937. ijcai.org (2022)
3. Cao, Y., Liu, Z., Li, C., Li, J., Chua, T.: Multi-channel graph neural network for entity alignment. In: ACL (1). pp. 1452–1461. Association for Computational Linguistics (2019)
4. Chen, L., Li, Z., Wang, Y., Xu, T., Wang, Z., Chen, E.: MMEA: entity alignment for multi-modal knowledge graph. In: KSEM (1). Lecture Notes in Computer Science, vol. 12274, pp. 134–147. Springer (2020)
5. Chen, L., Li, Z., Xu, T., Wu, H., Wang, Z., Yuan, N.J., Chen, E.: Multi-modal siamese network for entity alignment. In: KDD. pp. 118–126. ACM (2022)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: ICML. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (2020)
7. Chen, Z., Chen, J., Geng, Y., Pan, J.Z., Yuan, Z., Chen, H.: Zero-shot visual question answering using knowledge graph. In: ISWC. Lecture Notes in Computer Science, vol. 12922, pp. 146–162. Springer (2021)
8. Chen, Z., Chen, J., Zhang, W., Guo, L., Fang, Y., Huang, Y., Zhang, Y., Geng, Y., Pan, J.Z., Song, W., Chen, H.: Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In: ACM Multimedia. ACM (2023)
9. Gao, Y., Liu, X., Wu, J., Li, T., Wang, P., Chen, L.: Clusterea: Scalable entity alignment with stochastic training and normalized mini-batch similarities. In: KDD. pp. 421–431. ACM (2022)
10. Guo, L., Chen, Z., Chen, J., Chen, H.: Revisit and outstrip entity alignment: A perspective of generative models. CoRR **abs/2305.14651** (2023)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778. IEEE Computer Society (2016)
12. Huang, J., Sun, Z., Chen, Q., Xu, X., Ren, W., Hu, W.: Deep active alignment of knowledge graph entities and schemata. CoRR **abs/2304.04389** (2023)
13. Jiménez-Ruiz, E., Grau, B.C.: Logmap: Logic-based and scalable ontology matching. In: ISWC (1). Lecture Notes in Computer Science, vol. 7031, pp. 273–288. Springer (2011)
14. Jing, M., Li, J., Zhu, L., Lu, K., Yang, Y., Huang, Z.: Incomplete cross-modal retrieval with dual-aligned variational autoencoders. In: ACM Multimedia. pp. 3283–3291. ACM (2020)
15. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
16. Lee, H., Nam, T., Yang, E., Hwang, S.J.: Meta dropout: Learning to perturb latent features for generalization. In: ICLR. OpenReview.net (2020)
17. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web **6**(2), 167–195 (2015)
18. Li, C., Cao, Y., Hou, L., Shi, J., Li, J., Chua, T.: Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. In: EMNLP/IJCNLP (1). pp. 2723–2732. Association for Computational Linguistics (2019)

19. Li, Y., Chen, J., Li, Y., Xiang, Y., Chen, X., Zheng, H.: Vision, deduction and alignment: An empirical study on multi-modal knowledge graph alignment. CoRR **abs/2302.08774** (2023)
20. Lin, Z., Zhang, Z., Wang, M., Shi, Y., Wu, X., Zheng, Y.: Multi-modal contrastive representation learning for entity alignment. In: COLING. pp. 2572–2584. International Committee on Computational Linguistics (2022)
21. Liu, F., Chen, M., Roth, D., Collier, N.: Visual pivoting for (unsupervised) entity alignment. In: AAAI. pp. 4257–4266. AAAI Press (2021)
22. Liu, Y., Li, H., García-Durán, A., Niepert, M., Oñoro-Rubio, D., Rosenblum, D.S.: MMKG: multi-modal knowledge graphs. In: ESWC. Lecture Notes in Computer Science, vol. 11503, pp. 459–474. Springer (2019)
23. Liu, Z., Cao, Y., Pan, L., Li, J., Chua, T.: Exploring and evaluating attributes, values, and structures for entity alignment. In: EMNLP (1). pp. 6355–6364. Association for Computational Linguistics (2020)
24. Qi, Z., Zhang, Z., Chen, J., Chen, X., Xiang, Y., Zhang, N., Zheng, Y.: Unsupervised knowledge graph alignment by probabilistic reasoning and semantic embedding. In: IJCAI. pp. 2019–2025 (2021)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
26. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: NIPS. pp. 3483–3491 (2015)
27. Suchanek, F.M., Abiteboul, S., Senellart, P.: PARIS: probabilistic alignment of relations, instances, and schema. Proc. VLDB Endow. **5**(3), 157–168 (2011)
28. Sun, Z., Hu, W., Li, C.: Cross-lingual entity alignment via joint attribute-preserving embedding. In: ISWC (1). Lecture Notes in Computer Science, vol. 10587, pp. 628–644. Springer (2017)
29. Sun, Z., Hu, W., Wang, C., Wang, Y., Qu, Y.: Revisiting embedding-based entity alignment: A robust and adaptive method. IEEE Transactions on Knowledge and Data Engineering pp. 1–14 (2022). https://doi.org/10.1109/TKDE.2022.3200981
30. Sun, Z., Hu, W., Zhang, Q., Qu, Y.: Bootstrapping entity alignment with knowledge graph embedding. In: IJCAI. pp. 4396–4402. ijcai.org (2018)
31. Sun, Z., Huang, J., Hu, W., Chen, M., Guo, L., Qu, Y.: Transedge: Translating relation-contextualized embeddings for knowledge graphs. In: ISWC (1). Lecture Notes in Computer Science, vol. 11778, pp. 612–629. Springer (2019)
32. Sun, Z., Wang, C., Hu, W., Chen, M., Dai, J., Zhang, W., Qu, Y.: Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In: AAAI. pp. 222–229. AAAI Press (2020)
33. Sun, Z., Zhang, Q., Hu, W., Wang, C., Chen, M., Akrami, F., Li, C.: A benchmarking study of embedding-based entity alignment for knowledge graphs. Proc. VLDB Endow. **13**(11), 2326–2340 (2020)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS. pp. 5998–6008 (2017)
35. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: ICLR (Poster). OpenReview.net (2018)
36. Wang, M., Shi, Y., Yang, H., Zhang, Z., Lin, Z., Zheng, Y.: Probing the impacts of visual context in multimodal entity alignment. Data Sci. Eng. **8**(2), 124–134 (2023)
37. Wang, Y., Cui, Y., Liu, W., Sun, Z., Jiang, Y., Han, K., Hu, W.: Facing changes: Continual entity alignment for growing knowledge graphs. In: ISWC. Lecture Notes in Computer Science, vol. 13489, pp. 196–213. Springer (2022)

38. Wu, Y., Liu, X., Feng, Y., Wang, Z., Zhao, D.: Neighborhood matching network for entity alignment. In: ACL. pp. 6477–6487. Association for Computational Linguistics (2020)
39. Xin, K., Sun, Z., Hua, W., Hu, W., Zhou, X.: Informed multi-context entity alignment. In: WSDM. pp. 1197–1205. ACM (2022)
40. Yang, B., Yih, W., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. In: ICLR (Poster) (2015)
41. Yang, H., Zou, Y., Shi, P., Lu, W., Lin, J., Sun, X.: Aligning cross-lingual entities with multi-aspect information. In: EMNLP/IJCNLP (1). pp. 4430–4440. Association for Computational Linguistics (2019)
42. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Li, C., Xu, Y., Chen, H., Tian, J., Qi, Q., Zhang, J., Huang, F.: mplug-owl: Modularization empowers large language models with multimodality. CoRR **abs/2304.14178** (2023)
43. Zhang, Q., Sun, Z., Hu, W., Chen, M., Guo, L., Qu, Y.: Multi-view knowledge graph embedding for entity alignment. In: IJCAI. pp. 5429–5435. ijcai.org (2019)
44. Zhao, J., Li, R., Jin, Q.: Missing modality imagination network for emotion recognition with uncertain missing modalities. In: ACL/IJCNLP (1). pp. 2608–2618. Association for Computational Linguistics (2021)
45. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2242–2251. IEEE Computer Society (2017)
46. Zhu, Q., Zhou, X., Wu, J., Tan, J., Guo, L.: Neighborhood-aware attentional representation for multilingual knowledge graphs. In: IJCAI. pp. 1943–1949. ijcai.org (2019)