

Knowing When to Trust Images: Reliability-Aware Multi-modal Entity Alignment

Chenxiao Li¹ Yunhe Feng¹ Dongfang Liu² Dong Nie³ Yan Huang¹ Heng Fan¹

¹University of North Texas ²Purdue University ³Meta Inc

chenxiaoli@my.unt.edu heng.fan@unt.edu

Abstract

The visual modality, *i.e.*, images, plays a key role in multi-modal entity alignment (MMEA). Existing approaches often directly fuse the image with other modalities to align different entities. Although simple, such strategies overlook the potential noise in the images and their semantic misalignment with corresponding entities, resulting in suboptimal fusion and degraded performance. Addressing this, we propose a novel **Reliability-Aware framework for MMEA (RA-MMEA)**, which assesses visual reliability and adaptively improves unreliable visual representations for robust entity alignment. The core lies in two modules, including *dependency-aware visual reliability prediction (DA-VRP)* and *stability-regularized visual embedding generation (SR-VEG)*. The former aims to estimate the reliability of an image by leveraging multi-modal dependency within the entity, while the latter focuses on producing alternative visual representation conditioned on semantics encoded in textual modalities for multi-modal fusion. Compared to current methods, RA-MMEA enables more reliable visual representations for modality fusion, thereby improving performance. In extensive experiments, RA-MMEA achieves state-of-the-art results, verifying the importance of reliable visual modality for entity alignment and the effectiveness of RA-MMEA. *The code and results will be released.*

1 Introduction

Multi-modal knowledge graphs (MMKGs) employ heterogeneous information such as image, text, and structure to represent real-world entities. Yet, independently constructed MMKGs often contain duplicated or semantically overlapping entities, resulting in fragmented knowledge across graphs. To address this issue, multi-modal entity alignment (MMEA) (Chen et al., 2020; Liu et al., 2019) has been introduced to identify equivalent entities across different MMKGs by leveraging multi-modal information,

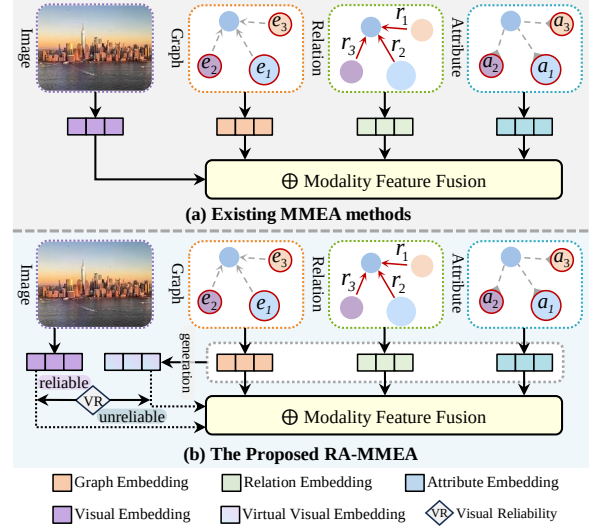


Figure 1: Comparison of existing methods directly fusing image and other modalities in (a) and RA-MMEA that first assesses the reliability of the image and then performs fusion in (b). *Best viewed in pdf for all figures.*

thus consolidating fragmented MMKGs into unified and complete knowledge resources.

In MMEA, visual modality, *i.e.*, images, offers rich appearance information and is crucial for aligning entities. Existing models typically directly fuse the image and other modalities to construct entity representation (see Fig. 1 (a)). However, images in MMKGs are *not* always reliable. They may be irrelevant to the entity or exhibit a large semantic gap from the text. Directly incorporating such noisy images into multi-modal fusion may compromise the entity representation and consequently degrade alignment. To alleviate this, a straightforward solution is to filter out unreliable images using manually annotated class conflict dictionaries (CCDs) (Shi et al., 2022). However, constructing CCDs entails significant labeling efforts, which is laborious and expensive. Besides, manually designed class conflict rules in CCDs inevitably introduce biases and struggle to generalize to real-world scenarios.

In this paper, we propose to learn visual reliability from the entity itself and adaptively handle unreliable images before multi-modal fusion, and introduce a novel *Reliability-Aware framework for MMEA (RA-MMEA)*, that assesses visual reliability and adaptively improves unreliable visual representations for robust entity alignment (see Fig. 1(b)). The key intuition is that image reliability can be inferred from its consistency with the semantic context of an entity, including graph, relation, and attribute. Based on this intuition, we design a dependency-aware visual reliability predictor (DA-VRP) in RA-MMEA, which estimates image reliability from global multi-modal dependency. Reliable images are expected to form coherent dependency with the remaining entity modalities, whereas unreliable images tend to deviate from such dependency and behave as outliers in the multi-modal representation space. DA-VRP first calculates the dependency score of the image with other modalities through dependency relation modeling, and then uses such dependency score to learn a binary reliability predictor that identifies and suppresses unreliable visual modalities. To enable end-to-end optimization with discrete binary indicators, we employ Gumbel-Softmax to predict the binary reliability in a differentiable manner.

After filtering unreliable images, the suppressed visual information needs to be compensated. To this end, we present a simple yet effective stability-regularized visual embedding generation module in RA-MMEA, termed SR-VEG, that aims to produce alternative visual representation conditioned on other modalities for multi-modal fusion. Specifically, it adaptively integrates features from other modalities and employs latent-variable modeling to generate visual representations. To preserve dependency consistency under modality removal, we further present a weight-difference penalty that aligns dependency scores computed with and without the visual modality. This regularization stabilizes the overall dependency estimation within an entity, leading to better reconstructed visual features and thereby improving entity alignment.

To the best of our knowledge, RA-MMEA is the first method that explicitly models visual reliability in MMEA. Compared with existing methods, RA-MMEA jointly estimates visual reliability and selectively reconstructs only unreliable visual representations before multi-modal fusion. This reliability-aware paradigm prevents unreliable visual information from degrading entity represen-

tations while preserving reliable visual information. Extensive experiments on multiple benchmarks demonstrate that RA-MMEA consistently achieves state-of-the-art performance, highlighting the importance of reliability-aware visual modeling for robust entity alignment.

In summary, our *contributions* are as follows: ♠ We propose RA-MMEA, the first reliability-aware framework for MMEA; ♥ We present DA-VRP that leverages global multi-modal dependency within an entity for visual reliability prediction; ♣ We introduce SR-VEG that generates virtual visual embeddings under stability-regularized consistency constraints; ♦ On multiple benchmarks, our RA-MMEA achieves state-of-the-art results.

2 Related Work

Multi-modal Entity Alignment (MMEA). Typical MMEA methods primarily focus on improving multi-modal fusion to optimize embedding representations (Zhang et al., 2024). To mitigate noise arising from modality heterogeneity, existing studies (Shi et al., 2022; Li et al., 2025a) employ annotated ontology matching as a filtering criterion to enhance embedding quality. To handle uncertain missing values in fusion, another line of work (Chen et al., 2023b; Li et al., 2023b, 2024) proposes generating pseudo-representation for the missing modality in multi-modal fusion. In contrast to these strategies that regulate input data quality or quantity, methods such as (Lin et al., 2022; Chen et al., 2022, 2023a; Li et al., 2025b) adopt contrastive learning or Transformer to precisely model inter-modal relationships, thereby alleviating the adverse effects of multi-modal information imbalance. Despite promising results, the aforementioned methods overlook the importance of addressing unreliable visual modality to improve MMEA.

LLMs in Multi-modal Entity Alignment. Large language models (LLMs) have been recently employed for MMEA due to their strong reasoning and understanding capabilities. Chen et al. (2024) employ LLMs to generate and embed attribute abstracts, facilitating attribute alignment. Ji et al. (2025) project multi-modal embeddings based on LLM-based attribute summaries to obtain the potential associations between them. Lu et al. (2025) propose LLM-guided visual alignment and attribute enhancement to achieve semantic filtering and enhancement in MMEA. Li et al. (2026) resolve dual-level noisy correspondence via a dedicated two-

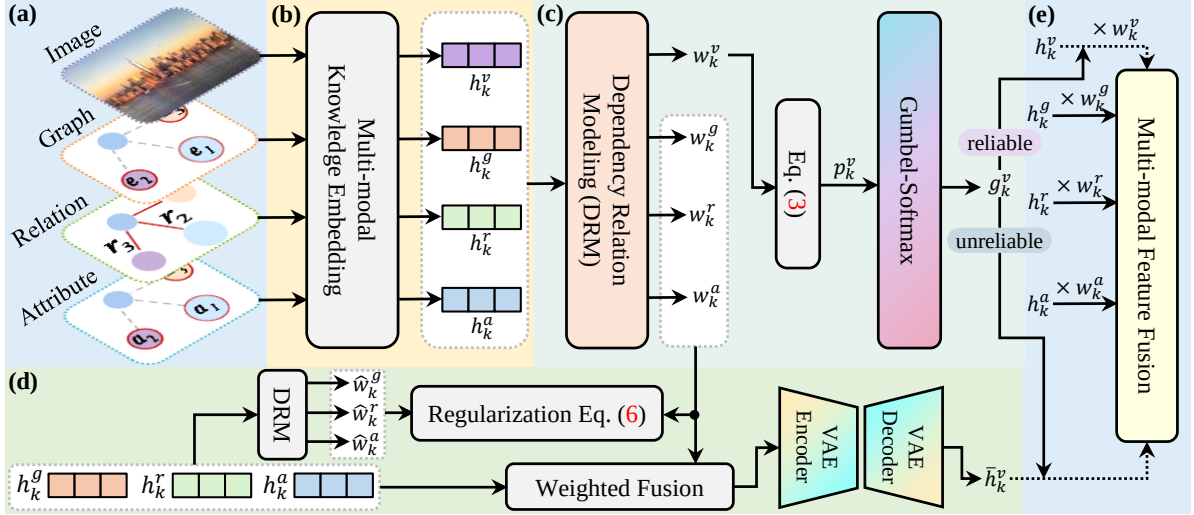


Figure 2: Illustration of the proposed RA-MMEA, which assesses visual modality reliability and adaptively improves unreliable visual representations through generated feature for MMEA. (a) Inputs with image, graph, relation, and attribute; (b) Multi-modal Embedding; (c) Dependency-Aware Visual Reliability Prediction; (d) Stability-Regularized Visual Embedding Generation; and (e) Multi-modal Fusion.

fold principle and employ MLLMs for relational reasoning. Despite excellent performance, these LLM-based models may introduce potential information leakage by relying on external world knowledge and entity names. Differently, instead of using external knowledge for MMEA, this work focuses on assessing and leveraging the reliability of visual modality based on the entity itself, aiming to improve multi-modal fusion under unreliable images.

3 Methodology

Problem Formulation. In MMEA, a multi-modal knowledge graph (MMKG) is formalized as $G = (E, R, A, T, I, P)$, where E, R, A, T , and I are the sets of entities, relations, attributes, triples, and images, and $P = \{(e, i) \mid e \in E, i \in I\}$ denotes the set of entity-image pairs. Each entity is associated to multiple attributes and $0 \sim 1$ image. Given two MMKGs $G_1 = (E_1, R_1, A_1, T_1, I_1, P_1)$ and $G_2 = (E_2, R_2, A_2, T_2, I_2, P_2)$, the set of seed alignments across them is defined as $H = \{(e_1, e_2) \mid e_1 \in E_1, e_2 \in E_2, e_1 \equiv e_2\}$, where $\equiv$ represents the equivalence of two entities. Given a set of pre-aligned entity pairs as training guidance, MMEA aims to match the counterpart entities e_1 and e_2 describing the same concepts in the real world from distinct multi-modal knowledge graphs.

Overview In this paper, we propose RA-MMEA, a reliability-aware framework for robust multi-modal entity alignment. As shown in Fig. 2, given an entity containing visual image, graph structure, rela-

tions, and attributes, we first extract their embedding (Sec. 3.1). Afterwards, DA-VRP (Sec. 3.2) is applied to predict the reliability of the image. If the image is reliable, we perform multi-modal fusion (Sec. 3.4) to obtain the entity representation; otherwise, we discard the unreliable visual embedding and apply SR-VEG (Sec. 3.3) to produce a virtual visual embedding for modality fusion (Sec. 3.4).

3.1 Multi-modal Knowledge Embedding

In this section, we describe embedding extraction for the visual image, graph structure, relations, and attributes of the entity e_k .

Visual Embedding. Given a visual image of the k -th entity e_k , we apply a pre-trained CLIP model (Radford et al., 2021), followed by fully connected (FC) layers, to extract the visual embedding h_k^v .

Graph Structure Embedding. For the graph structure of e_k , we follow (Li et al., 2025b; Chen et al., 2023a) and use graph attention network (Veličković et al., 2018) with two attention heads and two layers to extract the graph embedding h_k^g .

Relation Embedding. Following (Yang et al., 2019; Chen et al., 2023a), for the relations of e_k , we utilize bag-of-words followed by FC layers to extract the relation embedding h_k^r .

Attribute Embedding. For the attribute modality of e_k , we first employ bag-of-words and BERT (Devlin et al., 2019) to generate two attribute embeddings, respectively. These embeddings are then

combined through learnable weights and FC layers to obtain the final attribute embedding h_k^a .

3.2 Dependency-Aware Visual Reliability Prediction (DA-VRP)

In RA-MMEA, we propose DA-VRP to estimate whether the visual modality of an entity is reliable under the global multi-modal context. The key intuition is that reliable images should exhibit coherent semantic dependency with the remaining modalities of an entity, and thus have relatively high dependency score with other modalities.

Dependency Relation Modeling (DRM). Specifically, we first perform dependency relation modeling (DRM) to compute the global relevance of each modality with the remaining modalities. More concretely, drawing inspiration from the Transformer architecture (Vaswani et al., 2017), we first apply the scalable dot-product to calculate the relevance score with normalization between modalities m and j , as follows,

$$s_{mj} = \frac{\exp(\phi(h_k^m)^\top \phi(h_k^j)/d)}{\sum_{n \in M} \exp(\phi(h_k^m)^\top \phi(h_k^n)/d)} \quad (1)$$

where s_{mj} is the relevance score between modality $m \in M$ and $j \in M$, and $M = \{v, g, r, a\}$ with v, g, r, a representing the visual image, graph structure, relation, and attribute modalities. d is a scaling factor, and $\phi(\cdot)$ the linear projection.

Afterwards, we aggregate all relevance scores of each modality m with normalization to compute its dependency score within the entity e_k , as follows,

$$w_k^m = \frac{\exp(\sum_{j \in M} s_{mj})}{\sum_{n \in M} \exp(\sum_{j \in M} s_{nj})} \quad (2)$$

where w_k^m denotes the dependency score for modality m . After this, we employ the dependency score w_k^v of visual image to predict the visual reliability.

Visual Reliability Prediction. The dependency score $w_k^v \in [0, 1]$ reflects the semantic consistency between the visual modality and the remaining modalities of entity e_k , and we use it as a task-level proxy for visual reliability. This is motivated by the fact that cross-modal dependency provides a differentiable approximation of semantic consistency, a principled indicator of whether visual information agrees with the overall entity semantics. A higher dependency score indicates that the visual representation is more coherent with the overall entity semantics. In contrast, a lower score suggests that

the image may be noisy, semantically inconsistent, or provide limited visual evidence for representing the corresponding entity.

To predict the visual reliability, we first construct a probability vector based on w_k^v , as follows,

$$p_k^v = [1 - w_k^v, w_k^v] \quad (3)$$

where $p_{k,0}^v = 1 - w_k^v$ denotes the probability that the image is unreliable, and $p_{k,1}^v = w_k^v$ represents the probability that the image is reliable. To obtain the binary reliability prediction, a straightforward way is to apply the Bernoulli function on $p_{k,0}^v$ and $p_{k,1}^v$. However, this operation is non-differentiable, preventing the entire pipeline from being trained in an end-to-end manner.

To handle this, we adopt Gumbel-Softmax (Jang et al., 2017; Maddison et al., 2017) to sample from the categorical distribution, as follows,

$$g_k^v = \text{Gumbel-Softmax}(p_k^v) \quad (4)$$

where g_k^v denotes a binary one-hot vector representing the predicted visual reliability, with $g_k^v = [0, 1]$ indicating that the image is reliable and $g_k^v = [1, 0]$ indicating that the image is unreliable. By employing the Gumbel-Softmax trick, the entire framework becomes fully end-to-end trainable, thereby improving entity alignment performance.

3.3 Stability-Regularized Visual Embedding Generation (SR-VEG)

To compensate for the loss of visual information caused by filtering the unreliable visual embedding, we propose SR-VEG to generate semantically consistent virtual visual embeddings from the remaining reliable modalities. Besides handling filtered visual modalities, it can also naturally generalize to entities that inherently lack visual information.

Specifically, we first construct a hybrid representation from the non-visual modalities, including graph structure, relation, and attribute embeddings, with the learned dependency scores from DA-VRP, as follows,

$$h_k^{\text{hyb}} = \oplus_{m \in T} [w_k^m \cdot h_k^m], \quad T = \{g, r, a\} \quad (5)$$

Here, $\oplus$ denotes concatenation. By incorporating modality-aware weights, SR-VEG preserves cross-modal dependency during visual embedding generation. However, directly removing the visual modality may perturb the original multi-modal dependency structure. Although unreliable visual

information should be suppressed, its removal can still affect dependency estimation among the remaining modalities, leading to unstable modality weights and potentially degrading the generated visual representation.

To alleviate this issue, we introduce a stability regularization term to constrain the discrepancy between dependency weights estimated before and after visual modality removal. For entities whose visual modalities are filtered, we first recompute dependency scores $\hat{w}_k^m$ ($m \in \{g, r, a\}$) with the embeddings h_k^g , h_k^r , and h_k^a using DRM in DA-VRP, and then compare them with the original dependency weights w_k^g , w_k^r , and w_k^a . We retain only positive deviations using a ReLU operation via

$$\mathcal{L}_p = \sum_{m \in T} \max(\hat{w}_k^m - w_k^m, 0) \quad (6)$$

Here, the penalty $\mathcal{L}_p$ does not directly construct the hybrid representation for generation. Instead, it acts as an auxiliary stability constraint that regularizes dependency estimation under modality perturbation. By discouraging excessive dependency shifts after visual removal, SR-VEG maintains more stable cross-modal relationships and improves the generated visual embeddings.

After obtaining h_k^{hyb} , we use a VAE-style framework (Kingma and Welling, 2014; Sohn et al., 2015) to generate the pseudo-visual embeddings. Specifically, the encoder first maps the hybrid representation into the parameters of a latent Gaussian distribution via

$$[\mu_k \oplus \log(\sigma_k)^2] = \Phi_{\text{Enc}}(h_k^{\text{hyb}}) \quad (7)$$

where $\Phi_{\text{Enc}}(\cdot)$ is the encoder implemented with MLP layers, and μ_k and σ_k denote the mean and standard deviation of the latent distribution, respectively. Then, latent variables are sampled through the reparameterization trick, as follows,

$$z_k = \mu_k + \epsilon \odot \sigma_k, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (8)$$

Afterwards, the sampled latent representation is decoded to generate pseudo-visual embeddings via

$$\bar{h}_k^v = \Phi_{\text{Dec}}(z_k) \quad (9)$$

where $\Phi_{\text{Dec}}(\cdot)$ is the decoder implemented with MLP layers, and $\bar{h}_k^v$ denotes the generated pseudo-visual embeddings. $\bar{h}_k^v$ is expected to preserve semantic consistency with the remaining modalities while compensating for the unreliable or missing visual information. The optimization objective of SR-VEG is detailed in Section 3.5.

3.4 Multi-modal Feature Fusion

In RA-MMEA, the final entity representation is obtained by fusing embeddings from different modalities in a reliability-aware manner. For each entity, when its visual modality is predicted to be reliable, we directly incorporate the original visual embedding into the entity representation. Otherwise, we replace the unreliable visual embedding with the generated pseudo-visual embedding from SR-VEG. This process can be formulated as follows,

$$h_k = w_k^g h_k^g \oplus w_k^r h_k^r \oplus w_k^a h_k^a \oplus \tilde{h}_k^v \quad (10)$$

where h_k denotes the final representation of entity e_k . $\tilde{h}_k^v = g_{k,1}^v w_k^v h_k^v + g_{k,0}^v \bar{h}_k^v$ is the reliability-aware visual representation, where $g_{k,1}^v$ and $g_{k,0}^v$ are the two elements of the visual reliability vector g_k^v , corresponding to the reliable and unreliable visual states, respectively.

3.5 Optimization and Inference

Optimization. To enhance interaction between modalities, we introduce multi-modal contrastive learning to achieve bidirectional alignment. The loss function is defined as follows,

$$\mathcal{L}_m = -\log(p_m(e_u, e_t) + p_m(e_t, e_u)/2) \quad (11)$$

where p_m denotes the alignment probability between two entities e_u and e_t . Furthermore, the intra-modal loss $\mathcal{L}_{\text{IAL}}$ and inter-modal loss $\mathcal{L}_{\text{ICL}}$ can be represented as follows,

$$\mathcal{L}_{\text{IAL}} = \sum_{m \in M} \mathcal{L}_m \quad \mathcal{L}_{\text{ICL}} = \sum_{m \in M} \hat{\mathcal{L}}_m \quad (12)$$

where $\hat{\mathcal{L}}_m$ is the post-fusion variant of $\mathcal{L}_m$. To alleviate imbalance and enforce constraints on reconstructed features, we jointly optimize the SR-VEG loss function $\mathcal{L}_{\text{SR}}$ together with the contrastive loss. Specifically, inspired by Chen et al. (2023b), we employ two reconstruction objectives, $\mathcal{L}_{\text{Re}}^{\text{vis}}$ and $\mathcal{L}_{\text{Re}}^{\text{hyb}}$, to minimize the distance between the pseudo and true embeddings, thereby emphasizing the differences between different features during the autoencoding process. In addition, minimizing the Kullback-Leibler (KL) divergence $\mathcal{L}_{\text{KL}}$ encourages the latent space to approximate a Gaussian distribution. Then, we introduce a similarity distillation loss $\mathcal{L}_{\text{Sim}}$ to quantify the discrepancy between the similarity matrix of the hybrid embeddings and that of the virtual image features. Finally, the overall loss function can be expressed as follows,

$$\mathcal{L} = \mathcal{L}_{\text{IAL}} + \mathcal{L}_{\text{ICL}} + \mathcal{L}_{\text{SR}} + \gamma \cdot \mathcal{L}_p \quad (13)$$

Setting	Models	Venue	FB15K-DB15K			FB15K-YG15K		
			Hits@1↑	Hits@10↑	MRR↑	Hits@1↑	Hits@10↑	MRR↑
Non-Iterative	MMEA (Chen et al., 2020)	KSEM'20	0.265	0.541	0.357	0.234	0.480	0.317
	EVA (Liu et al., 2021)	AAAI'21	0.199	0.448	0.283	0.153	0.361	0.224
	MSNEA (Chen et al., 2022)	KDD'22	0.114	0.296	0.175	0.103	0.249	0.153
	MCLEA (Lin et al., 2022)	COLING'22	0.295	0.582	0.393	0.254	0.484	0.332
	MoAlign (Li et al., 2023a)	EMNLP'23	0.318	0.564	0.409	0.296	0.525	0.378
	MEAformer (Chen et al., 2023a)	ACM MM'23	0.417	0.715	0.518	0.327	0.595	0.417
	PMF (Huang et al., 2024)	ACL'24	0.539	-	0.620	0.459	-	0.539
	RICEA (Li et al., 2025b)	ACL'25	0.471	0.720	0.557	0.411	0.658	0.497
	SGMEA (Cheng et al., 2025)	COLING'25	<u>0.543</u>	<u>0.777</u>	<u>0.625</u>	<u>0.587</u>	<u>0.826</u>	<u>0.670</u>
	RULE (Li et al., 2026)	ICLR'26	0.444	-	0.538	0.389	-	0.470
Iterative	HUMEA (Xing et al., 2026a)	AAAI'26	0.512	0.764	0.598	0.439	0.699	0.528
	RA-MMEA (Ours)	-	0.849	0.948	0.879	0.787	0.928	0.835
	EVA (Liu et al., 2021)	AAAI'21	0.231	0.488	0.318	0.188	0.403	0.260
	MSNEA (Chen et al., 2022)	KDD'22	0.149	0.392	0.232	0.138	0.346	0.210
	MCLEA (Lin et al., 2022)	COLING'22	0.395	0.656	0.487	0.322	0.546	0.400
	MEAformer (Chen et al., 2023a)	ACM MM'23	0.578	0.812	0.661	0.444	0.692	0.529
	PMF (Huang et al., 2024)	ACL'24	0.624	-	0.702	0.543	-	0.620
	IBMEA (Su et al., 2024)	ACM MM'24	0.631	0.813	0.697	0.521	0.708	0.584
	PCMEA (Wang et al., 2024)	AAAI'24	0.676	0.887	0.728	0.590	0.835	0.646
	RICEA (Li et al., 2025b)	ACL'25	0.567	0.804	0.652	0.516	0.733	0.593
	SGMEA (Cheng et al., 2025)	COLING'25	0.661	0.847	0.729	0.750	0.901	0.805
	EIEA (Wang et al., 2025)	KDD'25	0.769	0.869	0.804	0.830	0.915	0.860
	DMEA (Xing et al., 2026b)	WWW'26	<u>0.804</u>	<u>0.896</u>	<u>0.849</u>	<u>0.898</u>	<u>0.957</u>	<u>0.934</u>
	RA-MMEA (Ours)	-	0.915	0.976	0.937	0.899	0.964	0.936

Table 1: Results on two cross-KG datasets using 20% of the reference entity alignments for training. The best results are shown in **bold** and the second best results are underlined.

where γ controls the strength of $\mathcal{L}_p$, and $\mathcal{L}_{SR} = \mathcal{L}_{KL} + \mathcal{L}_{Re}^{vis} + \mathcal{L}_{Re}^{hyb} + \mathcal{L}_{Sim}$.

Inference. We employ the cosine similarity (Sim) to quantify the alignment probability. The similarity matrix between the source entity set E and the target entity set E' is denoted as $\text{Sim}(E, E')$.

4 Experiments

Datasets. We validate the effectiveness and robustness of RA-MMEA on two cross-KG datasets, FB15K-DB15K/YG15K (Liu et al., 2019), with 20%, 50%, and 80% reference entity alignments as seed pairs, and three bilingual DBP15K datasets (ZH/JA/FR-EN) (Sun et al., 2017) under the unsupervised setting. Appendix A depicts the statistics of multi-modal datasets. The details of the evaluation metrics are shown in Appendix B.

Implementation Details. RA-MMEA is implemented in PyTorch on a machine with NVIDIA RTX A6000 GPUs. RA-MMEA is trained for 500 epochs, and optionally further trained by 500 additional epochs with iterative training. We apply AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with batch size of 3,500. The scaling factor d is set to 17. The parameter γ that controls $\mathcal{L}_p$ is set to 15.

Compared Methods. To rigorously evaluate our framework’s effectiveness, we compare it with 16 methods that focus on multi-modal fusion optimization as in Sec. 4.1. Notably, several recent MMEA methods, such as TMEA (Chen et al.,

2024), LGEA (Lu et al., 2025), and CLAEA (Ji et al., 2025), introduce external knowledge for MMEA. Since our goal is to design a reliability-aware framework that detects and mitigates unreliable visual modalities by exploiting inherent multi-modal dependency, rather than enhancing inputs with additional external information, we exclude these information-enhancement methods to ensure a fair comparison. For RULE (Li et al., 2026), we adopt the variant without MLLM during testing.

4.1 State-of-the-art Comparison

Tab. 1 shows the non-iterative and iterative results on the cross-KG datasets with 20% seed alignments. RA-MMEA consistently outperforms all baselines on both datasets and under both training settings across all metrics. Specifically, on FB15K-DB15K, RA-MMEA improves Hits@1 by 11.1%–30.6% and MRR by 8.8%–25.4% over the second-best baselines. On FB15K-YG15K, it achieves gains of 0.1%–20.0% in Hits@1 and 0.2%–16.5% in MRR. Notably, the improvements are particularly substantial under the non-iterative setting, demonstrating RA-MMEA’s effectiveness even with limited seed alignments. Additional results with 50% and 80% seed alignments are provided in Appendix C, while results on the bilingual DBP15K datasets are reported in Appendix D. Overall, these results demonstrate that RA-MMEA consistently achieves state-of-the-art performance across datasets, seed ratios, and training settings.

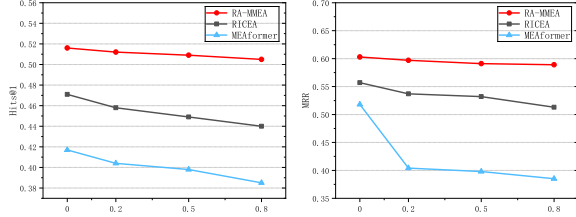


Figure 3: Performance of strong baselines across visual noise levels on FB15K-DB15K.

4.2 Robustness Evaluation

Visual information in real-world scenarios often contains varying degrees of noise, making it important for MMEA models to maintain stable alignment performance as visual information becomes increasingly unreliable. To evaluate this robustness, we construct artificial Entity-Image mismatches on FB15K-DB15K by randomly selecting 20%, 50%, and 80% of entities and shuffling their associated images while preventing them from being reassigned to their original entities, thereby simulating progressively increasing visual noise.

As shown in Fig. 3, the performance of all methods gradually decreases as the noise ratio increases, indicating that incorrect Entity-Image associations make MMEA increasingly challenging. Nevertheless, RA-MMEA consistently achieves the best performance across all noise ratios and exhibits substantially smaller performance degradation than MEAformer and RICEA. These results demonstrate that DA-VRP effectively mitigates interference caused by incorrect Entity-Image associations, enabling RA-MMEA to preserve robust multi-modal representations under increasing visual noise and confirming its robustness in challenging noisy MMEA scenarios.

4.3 Visual Noise Detection

The above results demonstrate that RA-MMEA remains robust under progressively increasing visual noise. We further investigate whether this advantage stems from the ability of DA-VRP to identify unreliable visual information. Specifically, we treat the artificially injected Entity-Image mismatches as known unreliable samples and measure the proportion of these samples correctly identified as unreliable by DA-VRP.

As shown in Tab. 2, DA-VRP achieves detection rates of 75.24%, 74.87%, and 75.08% under visual noise ratios of 20%, 50%, and 80%, respectively, remaining stable at approximately 75%

Noise Ratio	Ratio=20%	Ratio=50%	Ratio=80%
Detection Rate (%)	75.24	74.87	75.08

Table 2: Detection rate of DA-VRP under varying visual noise ratios.

	DA-VRP	SR-VEG	Hits@1↑	Hits@10↑	MRR↑
①	✗	✗	0.682	0.860	0.733
②	✓	✗	0.798	0.886	0.855
③	✗	✓	0.701	0.873	0.752
④	✓	✓	0.849	0.948	0.879

Table 3: Ablation studies on DA-VRP and SR-VEG.

		Hits@1↑	Hits@10↑	MRR↑
①	DA-VRP w/ BS	0.503	0.762	0.590
②	DA-VRP w/ GS	0.849	0.948	0.879

Table 4: Ablation studies on Gumbel-Softmax (GS) and Bernoulli Sampling (BS) in DA-VRP.

		Hits@1↑	Hits@10↑	MRR↑
①	SR-VEG w/o $\mathcal{L}_p$	0.833	0.941	0.874
②	SR-VEG w/ $\mathcal{L}_p$	0.849	0.948	0.879

Table 5: Ablation studies on $\mathcal{L}_p$ for SR-VEG.

		Hits@1↑	Hits@10↑	MRR↑
①	generation w/ $\hat{w}_k^m$	0.830	0.938	0.867
②	generation w/ w_k^m	0.849	0.948	0.879

Table 6: Ablation studies on using $\hat{w}_k^m$ and using w_k^m for hybrid representation in SR-VEG.

across different noise levels. These results indicate that DA-VRP can effectively identify unreliable visual information based on cross-modal semantic inconsistency, while its detection capability remains largely insensitive to changes in the proportion of injected noise. Since existing MMEA datasets do not provide ground-truth reliability annotations for the original Entity-Image pairs, we evaluate only the artificially injected visual noise with known labels. Nevertheless, the consistent detection performance provides empirical evidence that cross-modal dependency serves as an effective task-level proxy for visual reliability. Further qualitative evidence is provided in Appendix I, where the predicted reliability is consistent with the semantic correspondence between entities and their associated images.

4.4 Ablation Study

To further analyze RA-MMEA, we conduct ablation studies on FB15K-DB15K with 20% seeds. Our final configuration is highlighted in cyan.

		Hits@1↑	Hits@10↑	MRR↑
❶	$\gamma = 1$	0.834	0.942	0.874
❷	$\gamma = 5$	0.836	0.943	0.875
❸	$\gamma = 10$	0.845	0.946	0.877
❹	$\gamma = 15$	0.849	0.948	0.879
❺	$\gamma = 20$	0.840	0.945	0.879

Table 7: Ablation studies on the parameter γ for $\mathcal{L}_p$.

Impacts of DA-VRP and SR-VEG. The core of RA-MMEA lies in two key modules, including DA-VRP and SR-VEG. To analyze their impacts, we conduct ablation studies in Tab. 3. As shown, without DA-VRP and SR-VEG, the Hits@1 score is 0.682 (❶). When using DA-VRP alone, the Hits@1 score is largely improved to 0.798 with 11.6% gains (❷); when applying SR-VEG only, the Hits@1 score is boosted to 0.701 with moderate improvements (❸). Using both modules together yields the best performance of 0.849 Hits@1 score with 16.7% gains (❹), confirming their strong complementarity for improving MMEA.

Gumbel-Softmax v.s. Bernoulli Sampling. In our DA-VRP, we adopt Gumbel-Softmax to predict the visual reliability, enabling end-to-end training of RA-MMEA. To analyze its effectiveness, we conduct ablation studies by comparing Gumbel-Softmax and Bernoulli sampling for visual reliability prediction. The results are reported in Tab. 4. As shown, we can observe that, when using Bernoulli sampling for reliability prediction, the Hits@1 score is 0.503 (❶). In contrast, when applying Gumbel-Softmax, the Hits@1 score is improved to 0.849 with 34.6% gains (❷). This confirms that differentiable reliability estimation is essential for learning accurate modality-filtering behavior.

Impact of $\mathcal{L}_p$. In SR-VEG, we design a stability regularization term $\mathcal{L}_p$, which aims to constrain dependency shifts after removing the visual modality. To analyze the impact of $\mathcal{L}_p$, we conduct the ablation studies in Tab. 5. As shown, without $\mathcal{L}_p$, the model suffers from degraded dependency estimation, leading to suboptimal pseudo-visual embeddings and performance. Incorporating $\mathcal{L}_p$ consistently improves the Hits@1 score from 0.833 (❶) to 0.849 (❷) with 1.6% gains, demonstrating that stabilizing cross-modal dependency is critical for reliable virtual visual embedding generation.

Analysis of w_k^m for Hybrid Representation. In SR-VEG, we apply the original dependency scores w_k^m ($m \in \{g, r, a\}$) to construct the hybrid representation in Eq. (5). To validate its effective-

Models	Hits@1↑	Hits@10↑	MRR↑
MEAformer (Chen et al., 2023a)	0.417	0.715	0.518
MEAformer (+)	0.506	0.758	0.594
Improvement (%)	8.9	4.3	7.6
RICEA (Li et al., 2025b)	0.471	0.720	0.557
RICEA (+)	0.501	0.755	0.589
Improvement (%)	3.0	3.5	3.2

Table 8: Plug-and-play evaluation of DA-VRP and SR-VEG across representative MMEA frameworks. "+" denotes their variants.

ness, we compare hybrid representation construction with w_k^m and $\hat{w}_k^m$ in Tab. 6. As shown, using the original dependency scores w_k^m is more stable (❷ v.s. ❶), as they are computed with full multi-modal information, whereas re-estimating weights after removing the visual modality relies on less information and therefore becomes unstable.

Impact of γ . To analyze the impact of γ for $\mathcal{L}_p$, we conduct ablation studies in Tab. 7 by changing the value of γ . As shown in Tab. 7, we can see that, when $\gamma = 15$, we show the best performance (❹). Therefore, we empirically set γ to 15.

4.5 Plug-and-Play Capability

We integrate DA-VRP and SR-VEG into two MMEA frameworks with similar processing pipelines, MEAformer and RICEA, to evaluate the generalizability and plug-and-play capability of the proposed modules. As shown in Tab. 8, incorporating DA-VRP and SR-VEG consistently improves both frameworks across all evaluation metrics. Specifically, MEAformer achieves improvements of 8.9% and 7.6% in Hits@1 and MRR, respectively, while RICEA obtains corresponding gains of 3.0% and 3.2%. These consistent improvements demonstrate that DA-VRP and SR-VEG are not restricted to a specific MMEA architecture but can be flexibly integrated into existing frameworks to enhance their ability to effectively handle diverse unreliable visual information.

5 Conclusion

In this paper, we propose RA-MMEA, a reliability-aware framework for robust multi-modal entity alignment. RA-MMEA estimates visual reliability through cross-modal dependency and generates alternative representations for unreliable images. Extensive experiments demonstrate its effectiveness and robustness, highlighting the importance of reliability-aware visual modeling in MMEA.

Limitations

Although RA-MMEA demonstrates strong performance across multiple benchmarks, existing MMEA datasets typically provide only limited and relatively static visual information for each entity, which may not fully capture the diversity of entity-related images in real-world scenarios. In practice, the same entity may be associated with images exhibiting different visual content across different contexts, making it challenging to evaluate MMEA models under more diverse visual conditions.

Exploring more diverse visual settings could further broaden the applicability of reliability-aware MMEA. It would also provide additional insights into visual reliability across different entity types and scenarios. Such investigations could further enrich our understanding of reliability-aware visual modeling in MMEA. These directions may also offer further insights into how visual information contributes to multi-modal entity alignment. Future work could develop MMEA benchmarks with richer and more diverse visual information, enabling a more comprehensive evaluation of the robustness and generalizability of MMEA methods in real-world scenarios.

References

- Liyi Chen, Zhi Li, Yijun Wang, Tong Xu, Zhefeng Wang, and Enhong Chen. 2020. MMEA: entity alignment for multi-modal knowledge graph. In *Knowledge Science, Engineering and Management: 13th International Conference, KSEM 2020, Hangzhou, China, August 28–30, 2020, Proceedings, Part I* 13, pages 134–147. Springer.
- Liyi Chen, Zhi Li, Tong Xu, Han Wu, Zhefeng Wang, Nicholas Jing Yuan, and Enhong Chen. 2022. Multi-modal siamese network for entity alignment. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 118–126.
- Liyi Chen, Ying Sun, Shengzhe Zhang, Yuyang Ye, Wei Wu, and Hui Xiong. 2024. Tackling uncertain correspondences for multi-modal entity alignment. *Advances in Neural Information Processing Systems*, 37:119386–119410.
- Zhuo Chen, Jiaoyan Chen, Wen Zhang, Lingbing Guo, Yin Fang, Yufeng Huang, Yichi Zhang, Yuxia Geng, Jeff Z Pan, Wenting Song, et al. 2023a. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3317–3327.
- Zhuo Chen, Lingbing Guo, Yin Fang, Yichi Zhang, Jiaoyan Chen, Jeff Z Pan, Yangning Li, Huajun Chen, and Wen Zhang. 2023b. Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In *International Semantic Web Conference*, pages 121–139. Springer.
- Jingwei Cheng, Mingxiao Guo, and Fu Zhang. 2025. Sgmea: Structure-guided multimodal entity alignment. In *Proceedings of the 31st international conference on computational linguistics*, pages 7851–7861.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Yani Huang, Xuefeng Zhang, Richong Zhang, Junfan Chen, and Jaein Kim. 2024. Progressively modality freezing for multi-modal entity alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3477–3489.
- Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations (ICLR 2017)*. OpenReview. net.
- Yongquan Ji, Jingwei Cheng, Fu Zhang, and Chenglong Lu. 2025. Capturing latent modal association for multimodal entity alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22278–22293.
- Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations*.
- Chenxiao Li, Jingwei Cheng, Qiang Tong, and Fu Zhang. 2025a. Exploring the impacts of feature fusion strategy in multi-modal entity alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7809–7818.
- Chenxiao Li, Jingwei Cheng, Qiang Tong, Fu Zhang, and Cairui Wang. 2025b. Probing relative interaction and dynamic calibration in multi-modal entity alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28504–28516.
- Haobin Li, Yijie Lin, Peng Hu, Mouxing Yang, and Xi Peng. 2026. Learning with dual-level noisy correspondence for multi-modal entity alignment. In *The Fourteenth International Conference on Learning Representations*.

- Qian Li, Cheng Ji, Shu Guo, Zhaoji Liang, Lihong Wang, and Jianxin Li. 2023a. Multi-modal knowledge graph transformer framework for multi-modal entity alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 987–999.
- Qian Li, Jianxin Li, Jia Wu, Xutan Peng, Cheng Ji, Hao Peng, Lihong Wang, and S Yu Philip. 2024. Triplet-aware graph neural networks for factorized multi-modal knowledge graph entity alignment. *Neural Networks*, page 106479.
- Yangning Li, Jiaoyan Chen, Yinghui Li, Yuejia Xiang, Xi Chen, and Hai-Tao Zheng. 2023b. Vision, deduction and alignment: An empirical study on multi-modal knowledge graph alignment. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multi-modal contrastive representation learning for entity alignment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2572–2584.
- Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual pivoting for (unsupervised) entity alignment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4257–4266.
- Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019. Mmkg: multi-modal knowledge graphs. In *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16*, pages 459–474. Springer.
- Chenglong Lu, Chenxiao Li, Jingwei Cheng, Yongquan Ji, Guoqing Chen, and Fu Zhang. 2025. Breaking the noise barrier: Llm-guided semantic filtering and enhancement for multi-modal entity alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33141–33155.
- Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmlR.
- Yinghui Shi, Meng Wang, Ziheng Zhang, Zhenxi Lin, and Yefeng Zheng. 2022. Probing the impacts of visual context in multimodal entity alignment. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 255–270. Springer.
- Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.
- Taoyu Su, Jiawei Sheng, Shicheng Wang, Xinghua Zhang, Hongbo Xu, and Tingwen Liu. 2024. Ibmea: Exploring variational information bottleneck for multi-modal entity alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4436–4445.
- Zejun Sun, Wei Hu, and Chengkai Li. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *The Semantic Web-ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part I 16*, pages 628–644. Springer.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.
- Luyao Wang, Pengnian Qi, Xigang Bao, Chunlai Zhou, and Biao Qin. 2024. Pseudo-label calibration semi-supervised multi-modal entity alignment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 9116–9124.
- Luyao Wang, Chunlai Zhou, and Biao Qin. 2025. Explicit-implicit entity alignment method in multi-modal knowledge graphs. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 2996–3007.
- Yu Xing, Qizhuo Xie, Yunhui Liu, Qing Gu, Tao Zheng, Bin Chong, and Tieke He. 2026a. On modality weighting and specificity for multi-modal entity alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 27144–27152.
- Yu Xing, Qizhuo Xie, You Lv, Ziyang Zhou, Qianzi Hou, Qing Gu, Bin Chong, and Tieke He. 2026b. Enhancing multi-modal entity alignment via multi-grained decision fusion. In *Proceedings of the ACM Web Conference 2026*, pages 4047–4056.
- Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, and Xu Sun. 2019. Aligning cross-lingual entities with multi-aspect information. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4431–4441.
- Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, et al. 2024. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*.

A Appendix: Datasets Statistics

Tab. 9 shows the statistics of datasets, including the number of entities (Ent.), relations (Rel.), attributes (Attr.), number of relation triples (Rel Tr.) and attribute triples (Attr Tr.), number of images (Image), and number of the reference entity pairs (EA pairs). Each entity is associated to multiple attributes and $0 \sim 1$ image.

B Appendix: Evaluation Metrics

We adopt Hits@n and MRR as evaluation metrics for all methods. Hits@n measures the proportion of correct entities ranked within the top n according to similarity scores, while MRR represents the mean reciprocal rank of the correct entities. Higher values of Hits@n and MRR indicate superior performance of the evaluated method.

C Appendix: More Results on Cross-KG Datasets

Tabs. 10 and 11 report the results on the cross-KG datasets using 50% and 80% of the reference entity alignments as seed pairs, respectively. RA-MMEA consistently achieves the best performance across all settings. Specifically, with 50% seed alignments, RA-MMEA improves Hits@1 by 9.2%–20.0% and MRR by 7.2%–16.2% over the second-best baselines. With 80% seed alignments, it further achieves gains of 6.0%–13.6% in Hits@1 and 4.4%–11.1% in MRR.

As the proportion of seed alignments increases, the performance of most methods improves due to the stronger alignment supervision. Nevertheless, RA-MMEA maintains a clear advantage over existing methods under both non-iterative and iterative settings. Notably, the improvements remain substantial even with 80% seed alignments, where

competing methods already achieve strong alignment performance. This indicates that the benefit of reliability-aware visual modeling is not limited to low-supervision scenarios but remains effective when abundant alignment supervision is available.

Together with the results using 20% seed alignments in Tab. 1, these results demonstrate that RA-MMEA consistently benefits from reliable visual information across different levels of alignment supervision, further confirming the effectiveness and generalizability of the proposed framework on cross-KG datasets.

D Appendix: Unsupervised Training on Bilingual Datasets

To validate the performance of our framework on the bilingual domain, we also conducted unsupervised experiments on the bilingual datasets DBP15K (Sun et al., 2017). Unsupervised training was initially introduced by Liu et al. (2021) to mitigate dependence on gold labels. Following Lin et al. (2022), the vision encoders Enc_v are configured as ResNet-152 (He et al., 2016) with a vision feature dimension of $d_v = 2048$. We utilize pre-trained 300-d GloVe vectors along with character bigrams for surface representation after applying machine translations for entity names.

Tab. 12 shows that our framework consistently outperforms all strong baselines on both Hits@1 and MRR, achieving improvements of 0.4%–3.3% and 0.2%–2.8%, respectively, across two settings on three bilingual datasets. Notably, the Hits@10 score reaches 1.0 on the DBP15K_{JA-EN} and DBP15K_{FR-EN}.

E Appendix: Low-Resource Comparison

To investigate the reliance of our framework on alignment labels during training, we evaluate RA-

Dataset	KG	Ent.	Rel.	Attr.	Rel. Tr.	Attr. Tr.	Image	EA pairs
FB15K-DB15K	FB15K	14,951	1,345	116	592,213	29,395	14,836	12,846
	DB15K	12,842	279	225	89,197	48,080	12,812	
FB15K-YG15K	FB15K	14,951	1,345	116	592,213	29,395	14,836	11,199
	YG15K	15,404	32	7	122,886	23,532	11,172	
DBP15K _{ZH-EN}	ZH	19,388	1,701	8,111	70,414	248,035	15,912	15,000
	EN	19,572	1,323	7,173	95,142	343,218	14,125	
DBP15K _{JA-EN}	JA	19,814	1,299	5,882	77,214	248,991	12,739	15,000
	EN	19,780	1,153	6,066	93,484	320,616	13,741	
DBP15K _{FR-EN}	FR	19,661	903	4,547	105,998	273,825	14,174	15,000
	EN	19,993	1,208	6,422	115,722	351,094	13,858	

Table 9: Statistics of MMEA datasets, with EA pairs representing the reference entity alignments.

Setting	Models	FB15K-DB15K			FB15K-YG15K		
		Hits@1↑	Hits@10↑	MRR↑	Hits@1↑	Hits@10↑	MRR↑
Non-iterative	MMEA (Chen et al., 2020)	0.417	0.703	0.512	0.403	0.645	0.486
	EVA (Liu et al., 2021)	0.334	0.589	0.422	0.311	0.534	0.388
	MSNEA (Chen et al., 2022)	0.288	0.590	0.388	0.320	0.589	0.413
	MCLEA (Lin et al., 2022)	0.555	0.784	0.637	0.501	0.705	0.574
	MoAlign (Li et al., 2023a)	0.576	0.749	0.634	0.550	0.713	0.617
	MEAformer (Chen et al., 2023a)	0.619	0.843	0.698	0.560	0.778	0.639
	RICEA (Li et al., 2025b)	0.648	0.852	0.721	0.617	0.811	0.687
	SGMEA (Cheng et al., 2025)	0.716	0.882	0.775	0.780	0.924	0.832
	HUMEA (Xing et al., 2026a)	0.695	0.880	0.763	0.649	0.847	0.720
	RA-MMEA (Ours)	0.916	0.978	0.937	0.888	0.969	0.917
Iterative	EVA (Liu et al., 2021)	0.364	0.606	0.449	0.325	0.560	0.404
	MSNEA (Chen et al., 2022)	0.358	0.656	0.459	0.376	0.646	0.472
	MCLEA (Lin et al., 2022)	0.620	0.832	0.696	0.563	0.751	0.631
	MEAformer (Chen et al., 2023a)	0.690	0.871	0.755	0.612	0.808	0.682
	IBMEA (Su et al., 2024)	0.742	0.880	0.793	0.655	0.821	0.714
	PCMEA (Wang et al., 2024)	0.738	0.915	0.781	0.670	0.886	0.721
	RICEA (Li et al., 2025b)	0.692	0.869	0.757	0.658	0.827	0.720
	SGMEA (Cheng et al., 2025)	<u>0.752</u>	0.894	<u>0.802</u>	<u>0.827</u>	<u>0.938</u>	<u>0.868</u>
	RA-MMEA (Ours)	0.932	0.984	0.950	0.919	0.977	0.940

Table 10: Results on two cross-KG datasets using 50% of the reference entity alignments for training. The best results are shown in **bold** and the second best results are underlined.

Setting	Models	FB15K-DB15K			FB15K-YG15K		
		Hits@1↑	Hits@10↑	MRR↑	Hits@1↑	Hits@10↑	MRR↑
Non-iterative	MMEA (Chen et al., 2020)	0.590	0.869	0.685	0.598	0.839	0.682
	EVA (Liu et al., 2021)	0.484	0.696	0.563	0.491	0.692	0.565
	MSNEA (Chen et al., 2022)	0.518	0.779	0.613	0.531	0.778	0.620
	MCLEA (Lin et al., 2022)	0.735	0.890	0.790	0.667	0.824	0.722
	MoAlign (Li et al., 2023a)	0.699	0.882	0.773	0.689	0.884	0.769
	MEAformer (Chen et al., 2023a)	0.765	0.916	0.820	0.703	0.873	0.766
	RICEA (Li et al., 2025b)	0.776	0.916	0.829	0.734	0.892	0.792
	SGMEA (Cheng et al., 2025)	<u>0.815</u>	0.931	0.828	<u>0.857</u>	0.951	<u>0.894</u>
	HUMEA (Xing et al., 2026a)	0.809	0.935	0.856	0.774	0.920	0.827
	RA-MMEA (Ours)	0.951	0.990	0.967	0.939	0.987	0.957
Iterative	EVA (Liu et al., 2021)	0.491	0.711	0.573	0.493	0.695	0.572
	MSNEA (Chen et al., 2022)	0.565	0.810	0.651	0.593	0.806	0.668
	MCLEA (Lin et al., 2022)	0.741	0.900	0.802	0.681	0.837	0.737
	MEAformer (Chen et al., 2023a)	0.784	0.921	0.834	0.724	0.880	0.783
	IBMEA (Su et al., 2024)	0.821	0.922	0.859	0.751	0.890	0.800
	PCMEA (Wang et al., 2024)	0.820	<u>0.964</u>	0.858	0.756	0.942	0.802
	RICEA (Li et al., 2025b)	0.787	0.919	0.838	0.752	0.899	0.804
	SGMEA (Cheng et al., 2025)	<u>0.828</u>	0.921	0.861	<u>0.882</u>	<u>0.967</u>	<u>0.915</u>
	RA-MMEA (Ours)	0.958	0.991	0.970	0.942	0.989	0.959

Table 11: Results on two cross-KG datasets using 80% of the reference entity alignments for training. The best results are shown in **bold** and the second best results are underlined.

Models	DBP15K _{ZH-EN}			DBP15K _{JA-EN}			DBP15K _{FR-EN}		
	Hits@1↑	Hits@10↑	MRR↑	Hits@1↑	Hits@10↑	MRR↑	Hits@1↑	Hits@10↑	MRR↑
<i>Non-iterative</i>									
EVA (Liu et al., 2021)	0.891	0.961	0.917	0.941	0.986	0.958	0.970	0.996	0.982
MSNEA (Chen et al., 2022)	0.859	0.936	0.887	0.921	0.970	0.939	0.954	0.989	0.968
MCLEA (Lin et al., 2022)	0.860	0.950	0.893	0.914	0.975	0.938	0.953	0.990	0.967
MEAformer (Chen et al., 2023a)	0.909	0.974	0.933	0.950	0.990	0.965	0.972	0.997	0.983
RA-MMEA (Ours)	0.942	0.992	0.961	0.971	0.998	0.982	0.991	0.999	0.995
Gain over Prior Best (%)	3.3	1.8	2.8	2.1	0.8	1.7	1.9	0.2	1.2
<i>Iterative</i>									
EVA (Liu et al., 2021)	0.948	0.992	0.964	0.977	0.997	0.985	0.990	1.000	0.995
MSNEA (Chen et al., 2022)	0.871	0.943	0.898	0.927	0.976	0.945	0.959	0.993	0.979
MCLEA (Lin et al., 2022)	0.942	0.994	0.963	0.974	0.999	0.984	0.986	1.000	0.992
MEAformer (Chen et al., 2023a)	0.964	0.998	0.975	0.985	1.000	0.991	0.992	1.000	0.996
RA-MMEA (Ours)	0.969	0.999	0.978	0.989	1.000	0.993	0.998	1.000	0.999
Gain over Prior Best (%)	0.5	0.1	0.3	0.4	0.0	0.2	0.6	0.0	0.3

Table 12: Unsupervised results on bilingual datasets are presented.

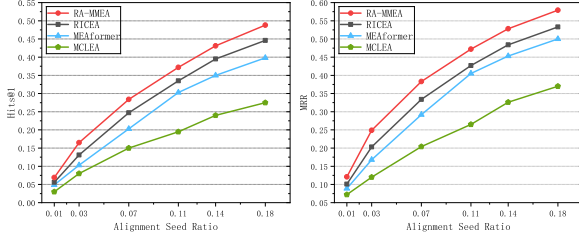


Figure 4: Performance of strong baselines with fewer seed alignments on FB15K-DB15K.

Models	Hits@1↑	Hits@10↑	MRR↑
RA-MMEA-R	0.708	0.824	0.745
RA-MMEA-A	0.732	0.845	0.811
RA-MMEA-V	0.798	0.886	0.855

Table 13: Performance of applying our reliability prediction method to textual modalities.

MMEA on FB15K-DB15K with fewer seed alignments. Following [Chen et al. \(2023a\)](#) and [Li et al. \(2025b\)](#), we set the seed alignment ratios to $R_{sa} = \{0.01, 0.03, 0.07, 0.11, 0.14, 0.18\}$. For a more intuitive comparison of performance trends, we adopt the same backbone as their methods.

Fig. 4 presents the comparison with other methods. RA-MMEA consistently outperforms all baselines across all seed alignment ratios, with increasingly pronounced advantages as more seed alignments become available. Notably, even when $R_{sa} = 0.01$, RA-MMEA achieves the best Hits@1 and MRR among all compared methods. When $R_{sa} = 0.18$, RA-MMEA even surpasses RICEA with $R_{sa} = 0.20$. These results not only demonstrate the robustness of RA-MMEA to unreliable visual data but also highlight its potential in low-resource scenarios.

F Appendix: Applying DA-VRP to Textual Modalities

To further evaluate DA-VRP’s ability to filter unreliable information beyond the visual modality, we additionally examine its performance on the relational and attribute modalities on FB15K-DB15K with 20% seeds. In this section, we focus solely on identifying and removing unreliable textual modalities, without involving modality reconstruction. Please note that, we do not filter the structural modality, as it provides the essential backbone for modeling cross-modal dependency. Removing structural information would compromise the stability of the alignment framework.

Tab. 13 reports the results. RA-MMEA-V, RA-

	Hits@1↑	Hits@10↑	MRR↑
① Concatenation	0.831	0.941	0.873
② Weighted-Concatenation	0.835	0.944	0.876
③ Ours	0.849	0.948	0.879

Table 14: Comparison of strategies for handling unreliable visual representations.

MMEA-R, and RA-MMEA-A respectively represent the variants applying reliability prediction on the visual image as proposed in this work, the relation modality, and the attribute modality. Please note that, for all these three variants, we do not perform modality reconstruction. As shown in Tab. 13, DA-VRP continues to provide clear benefits when applied to textual modalities, outperforming all other baselines in Tab. 1. However, textual modalities often contain complementary relational cues, and aggressive filtering may inadvertently discard useful information. In future work, we will develop finer-grained criteria to better separate complementary from redundant cross-modal relationships, enabling more selective modality filtering.

G Appendix: Visual Replacement Strategy

Although unreliable visual representations may still contain useful entity-related information, retaining them may also introduce additional noise. We therefore compare two alternative strategies that preserve the original unreliable visual representations: Concatenation and Weighted-Concatenation. Specifically, Concatenation combines the unreliable visual representation with the generated virtual visual representation, while Weighted-Concatenation reduces its contribution according to the estimated reliability.

As shown in Tab. 14, Weighted-Concatenation slightly outperforms Concatenation (② v.s. ①), indicating that suppressing unreliable visual information can alleviate its negative impact. Nevertheless, our strategy (③) achieves the best performance across all metrics, improving Hits@1 by 1.4 and 1.8 percentage points over Weighted-Concatenation and Concatenation, respectively. These results suggest that retaining unreliable visual representations may still introduce misleading visual semantics into multi-modal fusion and consequently affect entity representations during the alignment process, especially when visual information is highly unreliable. In contrast, directly replacing them with virtual visual representations generated from the

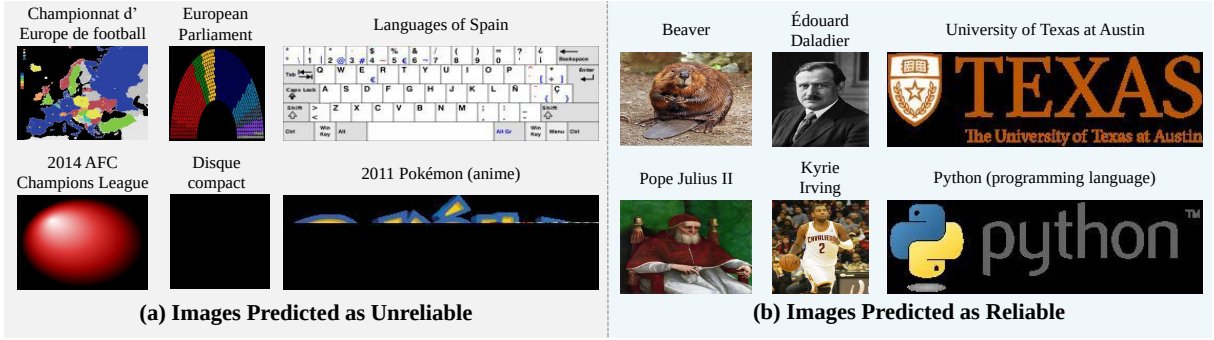


Figure 5: Case study of visual reliability on DBP15K, showing representative images predicted as unreliable (left) and reliable (right) by DA-VRP based on their cross-modal dependency with the corresponding entity semantics. The entity associated with each image is shown above the image.

Models	Train. Time	Learnable Paras.	GPU Mem.
MEAformer (Chen et al., 2023a)*	22min	10M	13.6G
RICEA (Li et al., 2025b)*	22min	10M	13.2G
RA-MMEA (Ours)	24min	13M	16.0G

Table 15: Comparison of complexity on FB15K-DB15K. * indicates our reproduced results.

remaining reliable modalities provides more semantically consistent and informative visual representations for accurate entity alignment, validating the visual replacement strategy in RA-MMEA.

H Appendix: Model Complexity Analysis

To further evaluate our framework, we compare the complexity of RA-MMEA with two representative lightweight baselines on FB15K-DB15K. As shown in Tab. 15, the training time, learnable parameters, and GPU memory of our RA-MMEA are 24 minutes, 13M, and 16G, which are comparable to the complexities of MEAformer and RICEA. However, compared to these two methods, RA-MMEA achieves superior alignment performance on different metrics (see Tab. 1, Appendix C and Appendix D), showing an effective balance between accuracy and model complexity.

I Appendix: Case Study

To qualitatively examine whether DA-VRP can distinguish unreliable from reliable visual information, we present cases from DBP15K in Fig. 5. The left side shows images predicted as unreliable by DA-VRP. For *European Parliament*, the image is an abstract seating diagram and provides limited visual semantics for representing the entity. The image associated with *Compact disc* is nearly blank, while *Championnat d'Europe de football* is represented

by a general map of Europe rather than the football competition itself. Similar cases include *2014 AFC Champions League*, *2011 Pokémon (anime)*, and *Languages of Spain*, whose images provide ambiguous, incomplete, or weakly related visual cues. These cases show that unreliable images may be irrelevant or capture only limited or indirect entity semantics.

In contrast, the reliable cases on the right exhibit entity-image semantic consistency. *Édouard Daladier* is depicted by a portrait, *Beaver* by an image of the corresponding animal, and *University of Texas at Austin* by its recognizable institutional identity. Likewise, the images of *Python (programming language)*, *Pope Julius II*, and *Kyrie Irving* contain distinctive visual information that directly corresponds to their entity semantics. Compared with the unreliable cases, these images provide more explicit and discriminative visual evidence for identifying the entities. The comparison shows that unreliable images may provide uninformative or weakly related semantics, potentially introducing misleading evidence, whereas reliable images exhibit stronger entity-image semantic consistency.

Notably, these examples cover diverse entity types, suggesting that the predicted reliability is not restricted to entity or visual content types. Despite their different visual characteristics, reliable images exhibit clearer semantic correspondence, while unreliable images provide more ambiguous or incomplete visual evidence. The qualitative results show that DA-VRP captures differences in visual reliability through cross-modal dependency. Together with the quantitative results in Sec. 4.3, these cases further support cross-modal dependency as an effective task-level proxy for visual reliability.