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

Anonymous ACL submission

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 ***R**eliability-**A**ware 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. Code is available for anonymous review here.

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, thus consolidating fragmented MMKGs into unified and complete knowledge resources.

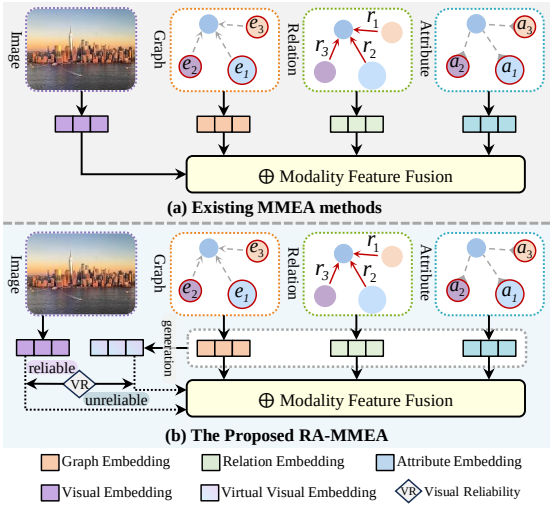


Figure 1: Comparison of existing methods directly fusing image and other modalities in (a) and RA-MMEA that first assesses the reliability of the image and then perform fusion in (b). *Best viewed in pdf for all figures*.

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 introduces biases and struggles to generalize to real-world scenarios.

In this paper, we propose to learn visual relia-

bility 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 feature 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 enables more reliable visual representations to be exploited for multi-modal fusion, leading to more robust entity alignment. To validate its effectiveness, we conduct extensive experiments on multiple benchmarks. The results show that RA-MMEA achieves state-of-the-art performance, demonstrating the importance of visual reliability in MMEA and the effectiveness of our RA-MMEA.

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-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 knowl-

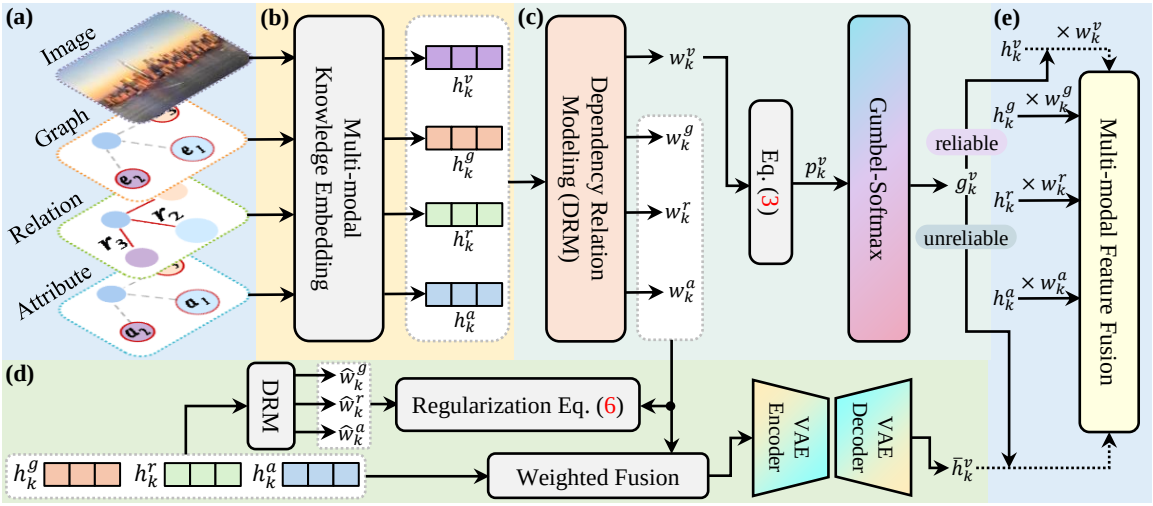


Figure 2: Overall of RA-MMEA, which assesses visual modality reliability and adaptively improves unreliable visual representations through generated feature for MMEA. (a) Inputs; (b) Multi-modal Embedding; (c) Dependency-Aware Visual Reliability Prediction; (d) Stability-Regularized Visual Embedding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

edge 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, relations, 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 VGG-16 network (Simonyan and Zisserman, 2015), 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 follow,

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

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 $w_v^k \in [0, 1]$ reflects the semantic consistency between the visual modality and remaining modalities of entity e_k, and we use this dependency score as a task-level proxy for visual reliability. A higher dependency score indicates that the visual representation is more coherent with the overall entity semantics, whereas a lower score suggests that the image may be noisy or semantically inconsistent.

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

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

where $p_{k,0}^v = 1 - w_v^k$ 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, inspired by Xue et al. (2020), we employ 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

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

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

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

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

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

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 unreliable and reliable 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,

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

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, $\mathcal{L}_{\text{Re}}^{vis}$ and $\mathcal{L}_{\text{Re}}^{hyb}$, 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

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

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

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 $\text{Sim}\langle E, E' \rangle$.

4 Experiments

Datasets. We validate the effectiveness of RA-MMEA on two cross-KG benchmarks, including

Seeds	Models	FB15K-DB15K			FB15K-YG15K		
		Hits@1↑	Hits@10↑	MRR↑	Hits@1↑	Hits@10↑	MRR↑
20%	EVA (Liu et al., 2021)	0.199	0.448	0.283	0.153	0.361	0.224
	MSNEA (Chen et al., 2022)	0.114	0.296	0.175	0.103	0.249	0.153
	MCLEA (Lin et al., 2022)	0.295	0.582	0.393	0.254	0.484	0.332
	MoAlign (Li et al., 2023a)	0.318	0.564	0.409	0.296	0.525	0.378
	MEAformer (Chen et al., 2023a)	0.417	0.715	0.518	0.327	0.595	0.417
	$\mathrm{T_{RI}F_{AC}}$ (Li et al., 2024)	0.318	0.559	0.389	0.290	0.508	0.371
	RICEA (Li et al., 2025b)	0.471	0.720	0.557	0.411	0.658	0.497
	RULE (Li et al., 2026)	0.444	-	0.538	0.389	-	0.470
	RA-MMEA(Ours)	**0.516**	**0.767**	**0.603**	**0.430**	**0.685**	**0.517**
	Gain over Prior Best (%)	4.5	4.7	4.6	1.9	2.7	2.0
50%	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
	$\mathrm{T_{RI}F_{AC}}$ (Li et al., 2024)	0.554	0.750	0.607	0.546	0.694	0.579
	RICEA (Li et al., 2025b)	0.648	0.852	0.721	0.617	0.811	0.687
	RA-MMEA(Ours)	**0.702**	**0.876**	**0.764**	**0.639**	**0.838**	**0.709**
	Gain over Prior Best (%)	5.4	2.4	4.3	2.2	2.7	2.2
80%	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
	$\mathrm{T_{RI}F_{AC}}$ (Li et al., 2024)	0.697	0.882	0.761	0.669	0.865	0.736
	RICEA (Li et al., 2025b)	0.776	0.916	0.829	0.734	0.892	0.792
	RA-MMEA(Ours)	**0.806**	**0.928**	**0.851**	**0.761**	**0.921**	**0.818**
	Gain over Prior Best (%)	3.0	1.2	2.2	2.7	2.9	2.6

Table 1: Non-iterative results on two cross-KG datasets are presented. The variable X% indicates the percentage of reference entity alignments used for training. The best results are shown in **bold**.

FB15K-DB15K/YG15K (Liu et al., 2019), with 20%, 50%, and 80% reference entity alignments. Appendix A depicts the statistics of multi-modal datasets. The details of the evaluation metrics are shown in Appendix B. Additionally, we conduct more experiments on three bilingual DBP15K (ZH/JA/FR-EN) (Sun et al., 2017), using 30% reference entity alignments as seed pairs in Section C.

Implementation Details. RA-MMEA is implemented in PyTorch on a machine with NVIDIA RTX 4090 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 10.

Compared Methods. To rigorously evaluate the effectiveness of our framework, we compare it with methods that use the same types of modality inputs and 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), CLAEA (Ji et al., 2025), and DMEA (Xing et al., 2026), leverage LLM techniques and 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 LLM-based 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

The non-iterative and iterative results on the cross-KG datasets are shown in Tab. 1 and Tab. 2, respectively. Our framework consistently outperforms all baselines across 6 experimental settings on both datasets under all metrics. Under the 20% seed ratio, RA-MMEA achieves notable performance gains. For FB15K-DB15K, Hits@1 improves by 4.5%–6.2% and MRR by 2.2%–4.8%. On FB15K-YG15K, Hits@1 rises by 1.9%–4.3% and MRR by 2.0%–3.6%. At 50% seed ratios, it consistently outperforms all baselines, yielding 4.0%–5.4% and 3.2%–4.3% gains in Hits@1 and MRR on FB15K-DB15K and 2.0%–2.2% and 1.7%–2.2% improvements on FB15K-YG15K. Even at 80% seeds, where performance saturates, it still achieves 3.0%–3.8% and 2.2%–2.4% gains on FB15K-DB15K and 1.1%–2.7% and 1.5%–2.6% on FB15K-YG15K. Overall, these results demonstrate the effectiveness of our framework for im-

Seeds	Models	FB15K-DB15K			FB15K-YG15K		
		Hits@1↑	Hits@10↑	MRR↑	Hits@1↑	Hits@10↑	MRR↑
20%	EVA (Liu et al., 2021)	0.231	0.488	0.318	0.188	0.403	0.260
	MSNEA (Chen et al., 2022)	0.149	0.392	0.232	0.138	0.346	0.210
	MCLEA (Lin et al., 2022)	0.395	0.656	0.487	0.322	0.546	0.400
	MEAformer (Chen et al., 2023a)	0.578	0.812	0.661	0.444	0.692	0.529
	RICEA (Li et al., 2025b)	0.567	0.804	0.652	0.516	0.733	0.593
	RA-MMEA(Ours)	**0.640**	**0.838**	**0.709**	**0.559**	**0.763**	**0.629**
	Gain over Prior Best (%)	6.2	2.6	4.8	4.3	3.0	3.6
50%	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
	RICEA (Li et al., 2025b)	0.692	0.869	0.757	0.658	0.827	0.720
	RA-MMEA(Ours)	**0.732**	**0.893**	**0.789**	**0.678**	**0.850**	**0.737**
	Gain over Prior Best (%)	4.0	2.2	3.2	2.0	2.3	1.7
80%	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
	RICEA (Li et al., 2025b)	0.787	0.919	0.838	0.752	0.899	0.804
	RA-MMEA(Ours)	**0.825**	**0.929**	**0.862**	**0.763**	**0.914**	**0.819**
	Gain over Prior Best (%)	3.8	0.8	2.4	1.1	1.5	1.5

Table 2: Iterative results on two cross-KG datasets are presented. The variable X% indicates the percentage of reference entity alignments used for training. The best results are shown in **bold**.

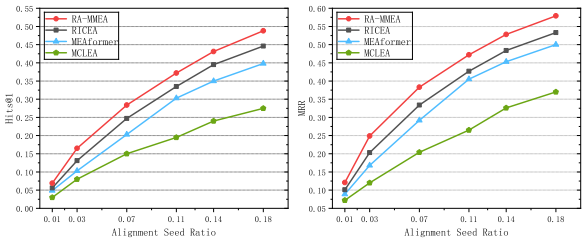


Figure 3: Performance of strong baselines with fewer seed alignments on FB15K-DB15K, including our reproduced results for MCLEA, MEAformer and RICEA.

proving multi-modal entity alignment.

4.2 Low-Resource Comparison

To discuss the reliance of our framework on alignment labels during training, we assess RA-MMEA on FB15K-DB15K with fewer seed alignments. Following Li et al. (2025b), we choose seed alignment ratio R_{sa} = {0.01, 0.03, 0.07, 0.11, 0.14, 0.18} in FB15K-DB15K.

Fig. 3 shows the results and comparison to other methods. As shown, we can see that, RA-MMEA outperforms other methods across all seed alignment ratios. Besides, as the seed alignment ratio increases, the performance of our framework improves more significantly. Specifically, when $R_{sa} = 0.01$, RA-MMEA outperforms all recent strong baselines in Hits@1 and MRR. When $R_{sa} = 0.18$, RA-MMEA even surpasses the performance of RICEA on $R_{sa} = 0.20$. This phenomenon not only verifies the robustness of RA-MMEA in handling unreliable visual data, but also demonstrates its potential in low-resource scenarios.

	DA-VRP	SR-VEG	Hits@1↑	Hits@10↑	MRR↑
❶	✗	✗	0.563	0.801	0.648
❷	✓	✗	0.609	0.812	0.685
❸	✗	✓	0.571	0.802	0.653
❹	✓	✓	0.639	0.838	0.709

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

		Hits@1↑	Hits@10↑	MRR↑
❶	DA-VRP w/ BS	0.578	0.804	0.658
❷	DA-VRP w/ GS	0.639	0.838	0.709

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

		Hits@1↑	Hits@10↑	MRR↑
❶	SR-VEG w/o $\mathcal{L}_p$	0.627	0.835	0.700
❷	SR-VEG w/ $\mathcal{L}_p$	0.639	0.838	0.709

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

		Hits@1↑	Hits@10↑	MRR↑
❶	generation w/ $\hat{w}_k^m$	0.625	0.830	0.698
❷	generation w/ w_k^m	0.639	0.838	0.709

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

4.3 Ablation Study

To further analyze RA-MMEA, we conduct ablation studies on FB15K-YG15K with 50% seeds. Our final configuration is highlighted in cyan.

Impacts of DA-VRP and SR-VEG. The core of

		Hits@1↑	Hits@10↑	MRR↑
❶	$\gamma = 1$	0.631	0.830	0.703
❷	$\gamma = 5$	0.632	0.834	0.704
❸	$\gamma = 10$	0.639	0.838	0.709
❹	$\gamma = 15$	0.633	0.833	0.704

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

Models	Train. Time	Learnable Paras.	GPU Mem.
MEAformer (Chen et al., 2023a)*	16min	10.8M	17G
RICEA (Li et al., 2025b)*	16min	11.1M	17G
RA-MMEA(ours)	17min	14.7M	22G

Table 8: Comparison of complexity on FB15K-DB15K. * indicates our reproduced results.

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 0.563 (❶). When using DA-VRP alone, the Hits@1 score is largely improved to 0.609 with 4.6% gains (❷); when applying SR-VEG only, the Hits@1 score is boosted to 0.571 with moderate improvements (❸). Using both modules together yields the best performance of 0.639 Hits@1 score with +7.6% 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 0.578 (❶). In contrast, when applying Gumbel-Softmax, the Hits@1 score is improved to 0.639 with 6.1% 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 0.627 (❶) to 0.639 (❷) with 1.2% 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 effectiveness, 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 (❶ *v.s.* ❷), 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 $\gamma = 10$, we show the best performance (❸). Therefore, we empirically set γ to 10.

4.4 Model Complexity Analysis

To further evaluate our framework, we compare the complexity of RA-MMEA with two state-of-the-art methods on FB15K-DB15K. As shown in Tab. 8, the training time, learnable parameters, and GPU memory of our RA-MMEA are 17 minutes, 14.7M, and 22G, 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 and Tab. 2), showing an effective balance between accuracy and model complexity.

Due to space limitation, we provide more analyses and experimental results in the Appendices.

5 Conclusion

In this paper, we propose RA-MMEA, a reliability-aware framework for robust multi-modal entity alignment. RA-MMEA explicitly estimates visual reliability and mitigates the impact of unreliable images. The core of RA-MMEA lies in two modules, including dependency-aware visual reliability prediction (DA-VRP) and stability-regularized visual embedding generation (SR-VEG). Specifically, DA-VRP aims to assess the reliability of the visual image using global dependency information within the entity, while SR-VEG focuses on generating alternative visual representations to compensate for the unreliable images. Extensive experiments demonstrate the effectiveness and robustness of RA-MMEA, highlighting the importance of reliability-aware visual modality in MMEA.

Limitations

Although RA-MMEA demonstrates strong performance across diverse benchmarks, one limitation is that its current visual reliability modeling adopts a binary filtering strategy. Specifically, unreliable visual embeddings are suppressed before multi-modal fusion, which effectively reduces the negative impact of noisy images but may also discard potentially useful visual cues, especially when an image is only partially unreliable or contains both noisy and informative regions. Future work could explore more fine-grained visual reliability modeling, such as region-level reliability estimation, and leverage useful partial visual information as auxiliary signals for visual embedding generation. This may help the model better distinguish harmful visual noise from useful visual evidence and further improve the completeness of multi-modal representations.

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, and 1 others. 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.

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.

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*.

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.

K Simonyan and A Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

Zequn 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*.

Yu Xing, Qizhuo Xie, You Lv, Ziyang Zhou, Qianzi Hou, Qing Gu, Bin Chong, and Tieke He. 2026. Enhancing multi-modal entity alignment via multi-grained decision fusion. In *Proceedings of the ACM Web Conference 2026*, pages 4047–4056.

Lanqing Xue, Xiaopeng Li, and Nevin L Zhang. 2020. Not all attention is needed: Gated attention network for sequence data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6550–6557.

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, and 1 others. 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: 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. 10 shows that our framework consistently outperforms all strong baselines on both Hits@1 and MRR, achieving improvements of 0.4%–3.3% and 0.3%–2.8%, respectively, across two settings on three bilingual datasets. Notably, the Hits@10 score reaches 1.0 on the $\text{DBP15K}_{JA\text{-}EN}$ and $\text{DBP15K}_{FR\text{-}EN}$.

D 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-YG15K with 50% 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. 11 reports the results. RA-MMEA-V, RA-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. 11, DA-VRP continues to provide clear benefits when applied to textual modalities, outperforming all other baselines except RICEA (Li et al., 2025b) 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.

E Appendix: Multi-modal Gap Analysis

During the experiments, we observe an intriguing phenomenon: under the $R_{sa} = \{0.2\}$ setting

Dataset	KG	Ent.	Rel.	Attr.	Rel. Tr.	Attr. Tr.	Image	EA pairs
FB15K-DB15K	FB15K DB15K	14,951 12,842	1,345 279	116 225	592,213 89,197	29,395 48,080	13,444 12,837	12,846
FB15K-YG15K	FB15K YG15K	14,951 15,404	1,345 32	116 7	592,213 122,886	29,395 23,532	13,444 11,194	11,199
DBP15K_{ZH-EN}	ZH EN	19,388 19,572	1,701 1,323	8,111 7,173	70,414 95,142	248,035 343,218	15,912 14,125	15,000
DBP15K_{JA-EN}	JA EN	19,814 19,780	1,299 1,153	5,882 6,066	77,214 93,484	248,991 320,616	12,739 13,741	15,000
DBP15K_{FR-EN}	FR EN	19,661 19,993	903 1,208	4,547 6,422	105,998 115,722	273,825 351,094	14,174 13,858	15,000

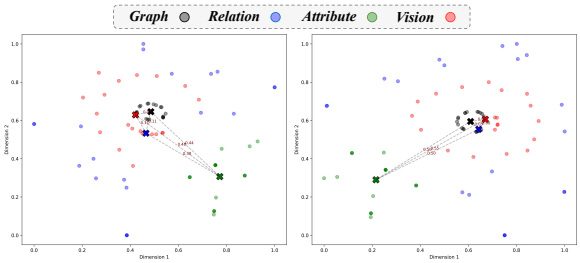
Table 9: Statistics of MMEA datasets, with EA pairs representing the reference entity alignments.

Models	DBP15K$_{ZH\text{-}EN}$			DBP15K$_{JA\text{-}EN}$			DBP15K$_{FR\text{-}EN}$		
	Hits@1↑	Hits@10↑	MRR↑	Hits@1↑	Hits@10↑	MRR↑	Hits@1↑	Hits@10↑	MRR↑
Non-iterative									
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
Iterative									
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 10: Unsupervised results on bilingual datasets are presented.

Models	Hits@1↑	Hits@10↑	MRR↑
RA-MMEA-R	0.526	0.778	0.615
RA-MMEA-A	0.563	0.802	0.646
RA-MMEA-V	0.609	0.812	0.685

Table 11: Performance of applying our reliability prediction method to textual modalities.



(a) FB15K-DB15K (b) FB15K-YG15K

Figure 4: T-SNE visualization of multi-modal embeddings.

on the FB15K-YG15K, removing the weighted fusion strategy unexpectedly improved performance, yielding Hits@1 and MRR reaching 44.0% and 52.6%, respectively. We hypothesize that the multi-modal gap, referring to the disparity between different modalities, is substantially larger in FB15K-YG15K than in FB15K-DB15K. Consequently, the weighted fusion process may amplify noise together with useful signals. To validate this hypothesis, we visualized the multi-modal embeddings. As shown in Fig. 4, the visualization supports this interpretation.

Specifically, we randomly selected 20 entities from each dataset and applied T-SNE to map their multi-modal embeddings into a lower-dimensional space. Because our goal is to cluster embeddings of the same modality, and the spatial layout in t-SNE visualizations does not reliably correspond to true geometric distances, we compute Euclidean distance to quantify the separation between multi-modal embeddings. Under this formulation, the dispersion of embeddings within a single modality reflects the stability of that modality, while the distance between the centroids of different modalities reflects the degree of separation across modalities. This analysis allows us to evaluate multi-modal relationships from the perspectives of complementarity and redundancy.

As reported by Shi et al. (2022) and Li et al. (2025a), the structural modality plays a central role in MMEA. Accordingly, we designate the structural modality as the anchor modality in our analysis. The results reveal that the distance associated with the attribute modality in FB15K-YG15K (0.50) is notably larger than that in FB15K-DB15K (0.44). This finding is further supported by the observation that methods specifically designed to enhance attribute modality, such as LGEA (Lu et al., 2025), achieve substantially greater performance gains on FB15K-YG15K compared to FB15K-DB15K. This is also the reason why we use BERT to enhance attribute embeddings. In essence, weighted fusion increases the model's sensitivity to differences across modalities. When the amount of data is limited, the model becomes more vulnerable to noise and unstable modalities. In contrast, when data is sufficiently abundant, weighted fusion strengthens robust modalities while suppressing weaker ones, ultimately leading to improved overall performance.