Probing Relative Interaction and Dynamic Calibration in Multi-modal Entity Alignment

Anonymous ACL submission

Abstract

Multi-modal entity alignment aims to identify equivalent entities between two different multi-modal knowledge graphs. Current methods have made significant progress by improving embedding and cross-modal fusion. However, most of them depend on using loss functions to capture the relationship between modalities or adopt a one-time strategy to directly compute modality weights using attention mechanisms, which overlooks the relative interactions between modalities at the entity level and the accuracy of modality weights, thereby hindering the generalization to diverse entities. To address this challenge, we propose RICEA, a relative interaction and calibration framework for multi-modal entity alignment, which dynamically computes weights based on the relative interaction and recalibrates the weights according to their uncertainties. Among these, we propose a novel method called ADC that utilizes attention mechanisms to perceive the uncertainty of the weight for each modality, rather than directly calculating the weight of each modality as in previous works. Across 5 datasets and 22 settings, our proposed framework significantly outperforms other baselines. Our code and data are available at https://anonymous.4open.science/r/RICEA-12D7/.

1 Introduction

Multi-modal knowledge graphs (MMKGs) organize real-world knowledge across modalities such as text and vision, have drawn massive attention in various scenarios, and supported numerous AI applications (Zhu et al., 2015; Yang et al., 2021b). Due to the increasing need for comprehensive multi-modal knowledge integration, multi-modal entity alignment (MMEA) (Chen et al., 2020; Liu et al., 2019) has emerged as a significant task in this field.

Current MMEA methods focus primarily on improving embeddings and cross-modal fusion.

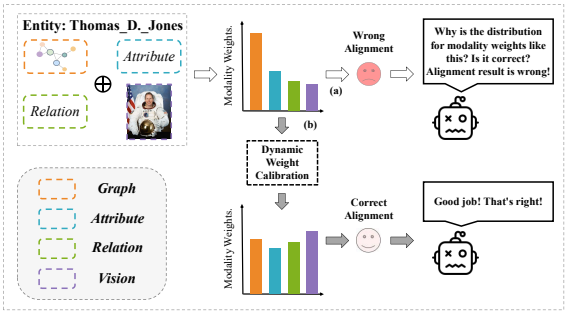

Figure 1: (a) Existing dynamic fusion methods neglect the relative interaction between modalities of each entity and the accuracy of weights, resulting in incorrect alignment results and (b) our calibration strategy dynamically adjusts the modality weights, improving the generalization.

Specifically, Chen et al. (2022) integrate visual features to guide relational feature learning, thereby fully utilizing multi-modal knowledge. Lin et al. (2022) propose two novel losses to obtain effective joint representations, ensuring that to-be-aligned entities between different KGs are semantically close with minimum gaps between modalities. Chen et al. (2023a) dynamically generate the entity-level meta weight for each modality inspired by the vanilla transformer (Vaswani, 2017). Li et al. (2025) independently obtain embeddings for each modality and select the optimal fusion strategy for the entity.

Previous works primarily utilize loss functions to model the relationship between modalities or use attention mechanisms to generate weights for each modality at one time. However, this task-oriented modeling is coarse-grained and often cannot reflect the actual importance level of modalities. They fail to fully capture the relative interactions between modalities at the entity level, leading to an over-emphasis on one modality and thus struggling to generalize to diverse real-world entities. For instance, the ablation study in (Chen et al., 2023a)

suggests adopting only the intra-modal loss but excluding inter-modal loss (Lin et al., 2022) results in the best performance. Their analysis reveals that inter-modal loss did not significantly enhance model performance and even resulted in performance degradation. As shown in Figure 1 (a), the diversity of entities in the real world limits the generalization ability of task-oriented models to a certain extent. Additionally, in (Serrano and Smith, 2019; Jain and Wallace, 2019), similar conclusions are drawn that while attention mechanisms can predict the importance of input components, in many cases, this association does not hold. Given that the importance of different modalities is relative, the importance of each modality should be dynamically adjusted according to the changing importance of other modalities. In other words, the importance of different modalities affects each other.

Therefore, to address the above issues, we propose a **R**elative **I**nteraction and **C**alibration framework for multi-modal entity alignment named RICEA. We elucidate the computation of each modality weight through the lens of relative modality interaction. Additionally, we refine modal weights by introducing an **A**ttention-Driven **D**istribution **C**alibration (ADC) mechanism. In contrast to previous efforts that directly ascertain the weight for each modality, ADC adopts a two-phase strategy to perceive the uncertainty of modality weights and adjust them accordingly. RICEA shows improved performance compared to the direct use of attention mechanisms (Chen et al., 2023a; Li et al., 2023a), particularly in low-resource scenarios, and provides new insights.

In summary, our main contributions are threefold:

- We propose a relative interaction and calibration framework for multi-modal entity alignment called RICEA. We are the first to identify the uncertainty issue of modality weights in multi-modal entity alignment and propose a calibration strategy to dynamically adjust the weights of each modality to ensure accuracy and stability.
- We propose **I**ntr**A**-modal **W**eight (IAW) and **I**nt**E**r-modal **W**eight (IEW) to enhance the relative interaction between modalities. We also propose a dynamic weight calibration for computing **C**alibration **J**oint **W**eight (CJW), which significantly improves generalization by calibrating weights with high uncertainty. To the best of our knowledge, we are the first to use attention mechanisms to perceive uncertainty in the weights for each modality.
- Extensive experiments verify the effectiveness of our proposed framework, especially its strong generalization ability in low-resource scenarios.

2 Related Work

Generally, the related work can be classified into two perspectives, i.e., typical entity alignment and multi-modal entity alignment.

2.1 Typical Entity Alignment

Embedding-based approaches for entity alignment (EA) can be generally divided into two categories: those that solely leverage graph structures and those that incorporate additional side information about entities (Zhang et al., 2020, 2021). Sun et al. (2018) iteratively labels likely entity alignment as training data for learning alignment-oriented KG embeddings. Wang et al. (2018) train GCNs to embed entities of each language into a unified vector space. To exploit the literal descriptions of entities expressed in different languages, Yang et al. (2019) integrate GCN-based and BERT-based modules to boost performance. Zhao et al. (2022) offer an unsupervised framework that performs entity alignment in the open world. A detailed survey for typical entity alignment can be found in (Zeng et al., 2021).

2.2 Multi-modal Entity Alignment

Beyond text and structured data, visual and auditory data, such as pictures, videos, and audio, also contain rich knowledge. Previous multi-modal entity alignment methods can be categorized into two types: **(i) Utilizing the relationships between modalities to enhance alignment.** Lin et al. (2022) utilize a multi-modal contrastive learning model to obtain effective joint representations for multi-modal entity alignment. Chen et al. (2022) employ inter-modal enhancement mechanisms to integrate visual features to guide relational feature learning to fully utilize multi-modal knowledge. Li et al. (2024) propose triplet-aware graph neural networks to aggregate multi-relational features. **(ii) Applying attention mechanisms to enhance alignment.** Chen et al. (2023a) provide a multi-modal entity alignment transformer for meta-modality hybrid, which dynamically predicts entity-level modality

weights for feature aggregation. Li et al. (2023a) propose a novel MMEA transformer, that hierarchically introduces neighbor features, multi-modal attributes, and entity types to enhance the alignment task. Fang and Yan (2024) use Transformer to obtain encoded representations of knowledge graph entities and make similar entities closer in the subspace. In addition, some methods enhance the alignment by completing the modality information. Li et al. (2023b) use attention mechanisms so that the entity embeddings can incorporate multiple images with different emphases. Chen et al. (2023b) address the issue of missing modality information to alleviate the impact of incompleteness on the alignment process.

However, all these methods overlook the relative interaction between modalities as well as the accuracy and stability of weights, resulting in overemphasis on one modality, which affects the generalization to diverse entities.

3 Methodology

3.1 Problem Formulation

Multi-modal Knowledge Graph. A multi-modal knowledge graph is formalized as $G = (E, R, A, T, I, P)$. Here, E, R, A, T, and I are the sets of entities, relations, attributes, triples, and images, respectively. $P = \{(e, i) \mid e \in E, i \in I\}$ is the set of entity-image pairs. Each entity is associated to multiple attributes and $0 \sim 1$ image.

Multi-modal Entity Alignment. Given two multi-modal knowledge graphs $G = (E, R, A, T, I, P)$ and $G^{'} = (E^{'}, R^{'}, A^{'}, T^{'}, I^{'}, P^{'})$, the set of seed alignments across two multi-modal knowledge graphs is defined as $H = \{(e, e') \mid e \in E, e' \in E', e \equiv e'\}$, where $\equiv$ represents the equivalence of two entities. A set of pre-aligned entity pairs are offered for training guidance. The task of multi-modal entity alignment targets to match the counterpart entities e and e' describing the same concepts in the real world from distinct multi-modal knowledge graphs.

3.2 Framework Overview

We propose a relative interaction and calibration framework for multi-modal entity alignment (RICEA), which comprises three major components: 1) **Multi-modal Knowledge Embedding** module extracts visual, structural, relational, and attribute features, and integrates them into holistic entity representations; 2) **Cross-Modal Interaction** **Joint Weighting** ($CIJW$) module measures the relative interaction between modalities and generates the **J**oint **Weight** (JW); 3) **Dynamic Weight Calibration** (DWC) module further dynamically adjusts the entity-level weight of each modality by calculating the uniformity of modality distribution and perceiving the uncertainty of modality weights. The framework overview is illustrated in Figure 2, and its primary components will be detailed in the following sections.

3.3 Multi-modal Knowledge Embedding

This section elaborates on how we embed the graph structure (h^g), relations (h^r), attributes (h^a), surface (h^s), and visual (h^v) modalities of entities into low-dimensional vectors.

Graph Structure Embedding. The graph attention network (GAT) is a typical neural network that deals with structured data (Veličković et al., 2018). We leverage GAT with two attention heads and two layers to capture the structural information. The structural feature embedding of the i-th entity e_i is:

$$h_i^g = GAT\left(\boldsymbol{W}_g, \boldsymbol{M}_g; x_i^g\right), \tag{1}$$

where g represents graph structure modality, $x_i^g \in \mathbb{R}^d$ represents the randomly initialized graph embedding of entity e_i, d is the predetermined hidden dimension, $\boldsymbol{W}_g \in \mathbb{R}^{d \times d}$ represents a diagonal weight matrix (Yang et al., 2015) used for linear transformation, $\boldsymbol{M}_g$ represents the graph adjacency matrix.

Relation, Attribute, and Surface Embedding. Following Yang et al. (2019) and Chen et al. (2023a), we use bag-of-words features to represent relations x^r, attributes x^a, and surfaces x^s, and employ separate fully connected layers (FC) to alleviate the information pollution caused by mixed relation/attribute representations in GNN-like networks (Liu et al., 2021). We represent the feature embedding of the m-th modality of the i-th entity e_i as:

$$h_i^m = \boldsymbol{FC}_m(x_i^m), m \in \{r, a, s\}, \tag{2}$$

where r, a, s represent the relation, attribute, and surface (a.k.a. entity name) modality, respectively.

Visual Embedding. We employ the pre-trained vision model as the encoder (Enc_v) to get the visual embeddings x_i^v for each available image of the entity e_i. Following Chen et al. (2020); Liu et al. (2021); Lin et al. (2022), we utilize the VGG-16 (Simonyan and Zisserman, 2015) on FB15K-DB15K/YAGO15K and the ResNet-152 (He et al.,

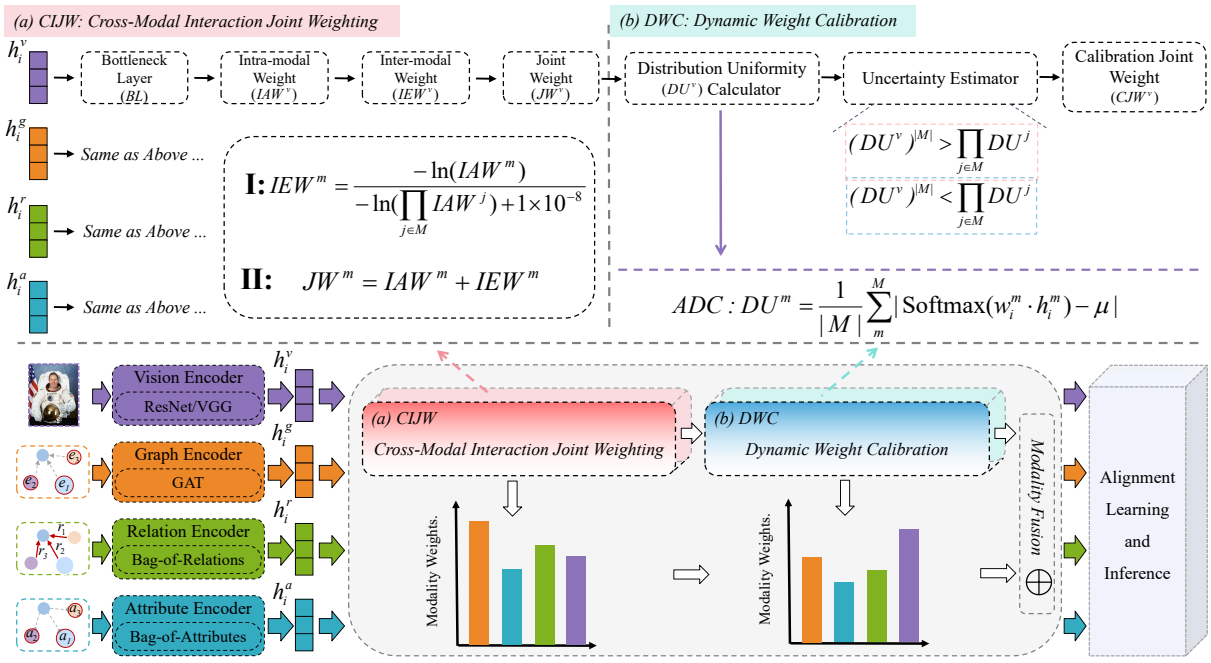

Figure 2: The overall framework of RICEA (down) and the implementation details of (a) Cross-Modal Interaction Joint Weighting ($CIJW$) and (b) Dynamic Weight Calibration (DWC).

2016) on DBP15K. To alleviate information pollution, we also employ a fully connected layer. Details of the settings will be provided in Section 4.1. For entities lacking image data, we create random image features by utilizing a normal distribution, which is defined by the mean and standard deviation of the existing images (Chen et al., 2022, 2023a; Lin et al., 2022; Liu et al., 2021). Image embedding is calculated as:

$$x_i^v = Enc_v(v_i), \quad h_i^v = FC_v(x_i^v). \tag{3}$$

3.4 Cross-Modal Interaction Joint Weighting

The **C**ross-Modal **I**nteraction **J**oint **W**eighting ($CIJW$) module aims to enhance the relative interaction between modalities and combines the **IntrA**-modal **W**eight (IAW) and **IntEr**-modal **W**eight (IEW) of the current modality to generate the **J**oint **W**eight (JW). By considering the $IAWs$ of all modalities, we can more comprehensively evaluate the importance of the current modality in the i-th entity.

Specifically, to ensure high precision, stability, and efficient processing of input features, we designed a Bottleneck Layer (BL), which is a fully connected module consisting of an up-projection layer, a nonlinear mapping, and a down-projection layer. The up-projection layer expands the input features to a higher dimensional space to increase the feature expression ability, and then compresses it back to the original dimension through the down-projection layer after nonlinear mapping, thereby improving the efficiency of feature processing. Finally, the dimension d is reduced to a single dimension to generate IAC, which is regarded as the reliability of unimodality.

$$IAW^m = BL(h_i^m). \tag{4}$$

To consider the importance of all modalities and smooth out small numerical differences to enhance computational stability, we take the natural logarithm of the IAW for the m-th modality and divide it by the natural logarithm of the product of the $IAWs$ across all modalities M. The IEW of the m-th modality can be calculated as:

$$IEW^m = \frac{-\ln(IAW^m)}{-\ln\left(\prod_{j \in M} IAW^j\right) + 1 \times 10^{-8}}, \tag{5}$$

where M denotes the set of available modalities. To prevent zero division errors or numerical instability in mathematical operations, we add a small constant of $1e-8$ to ensure smooth transitions, avoid abrupt changes under boundary conditions, maintain the value positive for logarithm calculations, and preserve the stability of the computational process, thus preventing numerical overflow

or underflow. To improve collaborative interaction between modalities, we define JW as a linear combination of IAW and IEW and use it as the temporary weight for the m-th modality. JW is defined as:

$$JW^m = IAW^m + IEW^m, \tag{6}$$

where IAW^m, IEW^m, and JW^m represent the intra-modal weight, inter-modal weight, and joint weight of the m-th modality respectively.

3.5 Dynamic Weight Calibration

Inspired by Cao et al. (2024), we proposed the **D**ynamic **W**eight **C**alibration (DWC) to ensure that the output weights are more accurate and reliable. DWC aims to use the weights obtained by the attention mechanisms to calculate distribution uniformity to perceive the uncertainty of modality and use it as a calibration standard to dynamically adjust the JWs. It is worth noting that the uncertainty of the modality means that some modalities may have uncertain missingness and ambiguity, which is also observed by Chen et al. (2023b). The experimental results in Section 4.3 show that our method improves the performance, surpassing the methods that directly use the attention mechanisms (Chen et al., 2023a; Li et al., 2023a).

DWC contains four sub-layers: 1) In **Entity-level Modality Weighted Attention**, we use the prediction weights obtained by attention mechanisms to assess the importance of each modality at the entity level, which provides a basis for the next sub-layer; 2) In **Attention-Driven Distribution Calibration**, we use the prediction weights to calculate the uniformity of the modality distribution and evaluate the uncertainty of its prediction weights. We then calibrate weights with high uncertainty and recalculate new joint weights; 3) In **Modality Fusion**, we fuse all weighted modalities. 4) In **Alignment Learning and Inference**, we use the cosine similarity to measure the alignment probability.

Entity-level Modality Weighted Attention. Inspired by Chen et al. (2023a), we introduce attention mechanisms to predict entity-level modality weights to avoid over-emphasizing one modality, instead of using the same approach as $CIJW$. Specifically, the multi-head cross-modal attention (MHCA) block performs the attention function in parallel over N_h heads where the i-th head is parameterized by modally shared weight matrix $\boldsymbol{W}_q^{(i)}, \boldsymbol{W}_k^{(i)}, \boldsymbol{W}_v^{(i)} \in \mathbb{R}^{d \times d_h}$ to project the multimodal input h^m into modal-aware query $Q_m^{(i)} \in \mathbb{R}^{d_h}$, key $K_m^{(i)} \in \mathbb{R}^{d_h}$, and value $V_m^{(i)} \in \mathbb{R}^{d_h}$.

$$Q_m^{(i)}, K_m^{(i)}, V_m^{(i)} = h^m \boldsymbol{W}_q^{(i)}, h^m \boldsymbol{W}_k^{(i)}, h^m \boldsymbol{W}_v^{(i)}. \tag{7}$$

For the feature of modality m, its output is:

$$\text{MHCA}(h_i^m) = \left[\text{head}_1^m \oplus \cdots \oplus \text{head}_{N_h}^m\right] \boldsymbol{W}_o, \tag{8}$$

$$\text{head}_i^m = \sum_{j \in M} \beta_{mj}^{(i)} V_j^{(i)}. \tag{9}$$

The attention weight β_{mj} between entity's modality m and j in each head is formulated as:

$$\beta_{mj} = \frac{\exp(Q_m^\top K_j / \sqrt{d_h})}{\sum_{n \in M} \exp(Q_m^\top K_n / \sqrt{d_h})}, \tag{10}$$

where $d_h = d/N_h$. Besides, layer normalization is used to stabilize the training:

$$\hat{h}_i^m = \textit{LayerNorm}(\text{MHCA}(h_i^m) + h_i^m). \tag{11}$$

Then the fully connected feed-forward network (FFN) consists of two linear transformation layers and a ReLU as the activation function:

$$\text{FFN}(\hat{h}_i^m) = \textit{ReLU}(\hat{h}_i^m \boldsymbol{W}_1 + b_1)\boldsymbol{W}_2 + b_2, \tag{12}$$

$$\hat{h}_i^m \leftarrow \textit{LayerNorm}(\text{FFN}(\hat{h}_i^m) + \hat{h}_i^m), \tag{13}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{d \times d_{in}}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d_{in} \times d}$. The weight of the m-th modality (w^m) is defined as:

$$w^m = \frac{\exp(\sum_{j \in M} \sum_{i=0}^{N_h} \beta_{mj}^{(i)} \sqrt{|M| \times N_h})}{\sum_{k \in M} \exp(\sum_{j \in M} \sum_{i=0}^{N_h} \beta_{kj}^{(i)} \sqrt{|M| \times N_h})}. \tag{14}$$

Attention-Driven Distribution Calibration (ADC). A uniform distribution typically suggests high uncertainty, whereas a peaked distribution implies low uncertainty in predictions (Huang et al., 2021). We apply the aforementioned attention mechanism to determine the prediction weights for each modality, treating them as prediction probabilities of the classification. This classification method serves as an implicit label for categorizing based on modality uncertainty. Subsequently, we compute the mean μ of these probability distributions. Hence, we define the **D**istribution **U**niformity (DU) of m-th modality as:

$$DU^m = \frac{1}{|M|}\sum_{m}^{M} |\text{Softmax}(w_i^m \cdot h_i^m) - \mu|, \quad (15)$$

where $|M|$ is the number of available modalities, w_i^m is the weight of the m-th modality of the i-th entity.

Taking into account the dynamic environment, the uncertainty of different modalities should be relative. In other words, the uncertainty of each modality should adjust dynamically in response to changes in the uncertainty of other modalities. The **R**elative **V**ariance (RV) of m-th modality is defined as:

$$RV^m = \frac{(DU^m)^{|M|}}{\prod_{j \in M} DU^j}. \quad (16)$$

Specifically, we assume that the quality of the modality with $RV^m < 1$ has a larger uncertainty, and it tends to produce relatively unreliable predictions, so the weight of the current modality has potential risks in terms of accuracy. Therefore, we reduce the contribution of such modalities by multiplying their predicted JW by RV^m ($RV^m < 1$). On the contrary, the quality of the modality with $RV^m > 1$ is considered to have a smaller uncertainty, so the contribution of these modalities can be retained to reduce the consumption of computing resources.

$$k^m = \begin{cases} RV^m & \text{if } (DU^m)^{|M|} < \prod_{j \in M} DU^j, \\ 1 & \text{otherwise.} \end{cases} \quad (17)$$

The **C**alibration **J**oint **W**eight (CJW) can be calculated as follows and used as the final weight of the m-th modality:

$$CJW^m = JW^m \cdot k^m. \quad (18)$$

Modality Fusion. We use the CJW of each modality as its fusion weight and assign it to each entity e_i.

$$h_i = \bigoplus_{m \in M} \left[CJW_i^m \cdot h_i^m\right], \quad (19)$$

where h_i is defined as the fusion embedding.

Alignment Learning and Inference. We use the cosine similarity (Sim) to measure the alignment probability. The similarity matrix of source entity set E and target entity set $E^{'}$ can be denoted as $\text{Sim}\langle E, E^{'}\rangle$.

4 Experiments

4.1 Experiment Setup

Datasets. To verify the effectiveness of our proposed framework in practical applications, we conducted experiments using two cross-KG datasets FB15K-DB15K/YAGO15K (Liu et al., 2019) and three bilingual datasets ZH/JA/FR-EN versions of DBP15K (Sun et al., 2017). Appendix A depicts the statistics of multi-modal datasets. Following conventions, we used 30% reference entity alignments as pre-aligned entity pairs (seeds alignments) for DBP15K. For FB15K-DB15K/YAGO15K, we used 20%, 50%, and 80% reference entity alignments. The details of the evaluation metrics are given in Appendix B.

Implementation Details. For a fair comparison, we kept all the settings of Chen et al. (2023a). The hidden layer dimensions *d* for all networks are standardized to 300. The total number of epochs is set to 500, with an optional iterative training strategy applied for an additional 500 epochs. The AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) is used, with a fixed batch size of 3500.

We adopt the approach of Mao et al. (2021), utilizing pre-trained 300-d GloVe vectors along with character bigrams for surface representation after applying machine translations for entity names. The vision encoders Enc_v are configured as ResNet-152 (He et al., 2016) for DBP15K, following EVA/MCLEA, with a vision feature dimension of $d_v = 2048$. For FB15K-DB15K/YAGO15K, the encoders are set to VGG-16 (Simonyan and Zisserman, 2015), with $d_v = 4096$. Specifically, for entity-level modality weighted attention, the intermediate dimension d_{in} is set to 400. γ is set to 0.1, and the head number N_h in MHCA is set to 1.

4.2 Comparative Methods

To comprehensively verify the effectiveness of our framework, we selected 20 prominent EA algorithms proposed in recent years as benchmarks and validated them on 5 real-world datasets and 22 settings.

The recent multi-modal entity alignment method LODEME (Li et al., 2023b) and UMAEA (Chen et al., 2023b) primarily enhances the accuracy of entity alignment by completing missing modalities, thereby increasing the available information. Our research aims to identify the relative interaction between modalities and adjust incorrect modal weights. This enhancement aims to improve gener-

Seeds	Models	FB15K-DB15K			FB15K-YAGO15K		
		Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
20%	MMEA (Chen et al., 2020)	0.265	0.541	0.357	0.234	0.480	0.317
	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
	TriFac (Li et al., 2024)	0.318	0.559	0.389	0.290	0.508	0.371
	RICEA(Ours)	**0.471**	**0.720**	**0.557**	**0.411**	**0.658**	**0.497**
	Improv. best%	5.4	0.5	3.9	8.4	6.3	8.0
50%	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
	TriFac (Li et al., 2024)	0.554	0.750	0.607	0.546	0.694	0.579
	RICEA(Ours)	**0.648**	**0.852**	**0.721**	**0.617**	**0.811**	**0.687**
	Improv. best%	2.9	0.9	2.3	5.7	3.3	4.8
80%	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
	TriFac (Li et al., 2024)	0.697	0.882	0.761	0.669	0.865	0.736
	RICEA(Ours)	**0.776**	**0.916**	**0.829**	**0.734**	**0.892**	**0.792**
	Improv. best%	1.1	0.0	0.9	3.1	0.8	2.3

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**.

alization capabilities, enabling adaptation to variations in modal quality in real-world scenarios. For a fair comparison, we did not include them.

4.3 Overall Results

The results of the monolingual datasets are shown in Table 1 (non-iterative) and Appendix C (iterative), and the results of the bilingual datasets are shown in Appendix D. Our framework outperforms the baselines on almost all datasets under all metrics. Especially on FB15K-DB15K/YAGO15K with 20% and 50% data splits (non-iterative), Hits@1 increased by 5.4%, 2.9%, 8.4% and 5.7%, and MRR increases by 3.9%, 2.3%, 8.0%, and 4.8%, respectively. This phenomenon confirms the strong generalization ability of our framework when learning with fewer samples, addressing the aforementioned issues.

4.4 Ablation Studies and Analysis

We conducted a series of experiments to thoroughly investigate the effectiveness of RICEA and whether it addresses the issues we identified. The experiments were designed to address two primary questions:

- *Does our RICEA framework have better generalization ability than its counterparts?*

  In Section 4.4.1, we conducted comparative experiments on 2 datasets with 12 data splits in low-resource scenarios. Our framework demonstrated the capability to use a small number of samples as training data and achieve better performance than baselines, verifying RICEA's strong generalization ability.

- *Do our proposed components really work?*

  In Section 4.4.2, we performed an ablation study to verify the effectiveness of each com-

ponent of our framework. In addition, we visually demonstrate the effectiveness of $CIJW$ module and DWC module, as well as the importance of each modality in multi-modal entity alignment.

4.4.1 Low Resource

To discuss that our framework has better generalization than similar baselines, we conducted comparative experiments in low-resource scenarios.

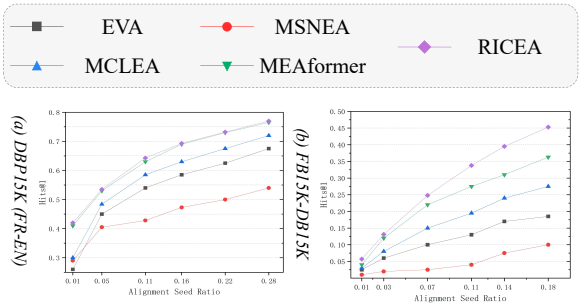

Figure 3: Generalization and low Resource. Model's Hits@1 performance with fewer seed alignments on (a) DBP15K$_{FR\text{-}EN}$ and (b) FB15K-DB15K.

Following the method of Chen et al. (2023a), in FB15K-DB15K, we chose the seed alignment ratio $R_{sa} = \{0.01, 0.03, 0.07, 0.11, 0.14, 0.18\}$, in DBP15K$_{FR\text{-}EN}$, we chose the seed alignment ratio $R_{sa} = \{0.01, 0.05, 0.11, 0.16, 0.22, 0.28\}$. As shown in Figure 3, as the seed alignment ratio increases, the performance of our framework improves more significantly. When $R_{sa} = \{0.11, 0.14\}$, our framework surpasses the performance of MEAformer (Chen et al., 2023a) on $R_{sa} = \{0.14, 0.18\}$ respectively. Furthermore, when $R_{sa} = 0.01$, our framework outperforms all baselines. This not only verifies the strong generalization ability of our framework but also offers new insights into research on low-resource communities.

4.4.2 Component Analysis

We developed different variants of RICEA to evaluate the optimal combination and explore the impact of various modalities on multi-modal entity alignment. The results are shown in Figure 4. We found that performance drops most significantly when the visual modality is removed, indicating that the visual modality is more important than the other two modalities. This result also verifies the conclusion of Chen et al. (2023a). When the DWC module is removed, performance drops significantly, indicating that our proposed dynamic weight calibration module effectively calibrates erroneous modality weights. Nevertheless, our framework still outperforms methods that explore the impact of the relationships between modalities on alignment results (Chen et al., 2022; Lin et al., 2022).

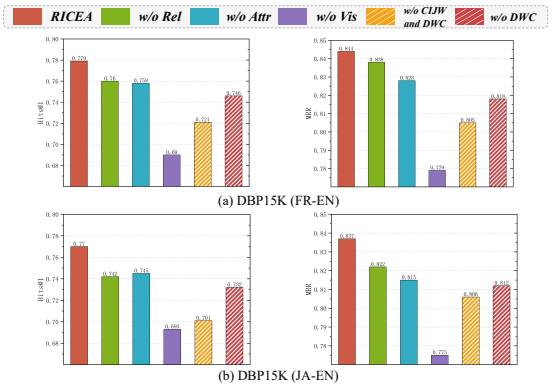

Figure 4: Component analysis for RICEA on (a) DBP15K$_{FR\text{-}EN}$ and (b) DBP15K$_{JA\text{-}EN}$.

Interestingly, we found that performance degradation when removing the proposed components is much greater than when removing modalities. This shows that our framework is almost unaffected by other information reduction, further verifying the framework's strong generalization capabilities. Moreover, we performed experiments to show that BL surpasses the attention mechanism in evaluating relative interaction, with more details provided in Appendix E.

5 Conclusion

In this work, we present a new framework called RICEA for multi-modal entity alignment. RICEA explains the weight calculation process for each modality from the perspectives of relative interaction, encouraging the development of low-resource communities. Additionally, our weight calculation is more precise and meticulous than all previous MMEA methods. We are the first to propose Dynamic Weight Calibration to further improve the framework's generalization to new data in the real world. Our research shows that RICEA can outperform all the recent methods across 5 datasets and 22 settings without increasing computational costs, especially in low-resource scenarios.

Limitations

In the efficiency analysis, we use the number of learnable parameters for evaluation. In the four settings on DBP15K, the number of learnable parameters of MEAformer is approximately 13.7 M to 14.2 M, while our framework is about 13.9 M to 14.5 M. The very small increase in the number of parameters is understandable because our framework indirectly uses attention mechanisms. In future work, we will find a better calibration standard than the attention mechanisms, so that the number of learnable parameters in our framework will be even smaller. We believe that the confidence obtained using a single-modal classifier will be more effective than the attention mechanisms and may outperform the attention mechanism in computing the uniformity of single-modal distribution. However, similar methods have not yet appeared in multi-modal knowledge graphs, making this our future research focus.

References

Weishan Cai, Wenjun Ma, Jieyu Zhan, and Yuncheng Jiang. 2022. Entity alignment with reliable path reasoning and relation-aware heterogeneous graph transformer. In *IJCAI*, pages 1930–1937.

Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. 2024. Predictive dynamic fusion. In *Forty-first International Conference on Machine Learning*.

Yixin Cao, Zhiyuan Liu, Chengjiang Li, Juanzi Li, and Tat-Seng Chua. 2019. Multi-channel graph neural network for entity alignment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1452–1461.

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.

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.

Jianyong Fang and Xuefeng Yan. 2024. Mdsea: Knowledge graph entity alignment based on multimodal data supervision. *Applied Sciences*, 14(9):3648.

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.

Rui Huang, Andrew Geng, and Yixuan Li. 2021. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. 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 3543–3556.

Chengjiang Li, Yixin Cao, Lei Hou, Jiaxin Shi, Juanzi Li, and Tat-Seng Chua. 2019. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. Association for Computational Linguistics.

Chenxiao Li, Jingwei Cheng, Qiang Tong, and Fu Zhang. 2025. 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.

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.

Zhiyuan Liu, Yixin Cao, Liangming Pan, Juanzi Li, and Tat-Seng Chua. 2020. Exploring and evaluating attributes, values, and structures for entity alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6355–6364.

Xinnian Mao, Wenting Wang, Yuanbin Wu, and Man Lan. 2021. From alignment to assignment: Frustratingly simple unsupervised entity alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2853.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

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.

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.

Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping entity alignment with knowledge graph embedding. In *IJCAI*, volume 18.

Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2020. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 222–229.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. 2018. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 349–357.

Tianxing Wu, Chaoyu Gao, Lin Li, and Yuxiang Wang. 2022. Leveraging multi-modal information for cross-lingual entity matching across knowledge graphs. *Applied Sciences*, 12(19):10107.

Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan, and Dongyan Zhao. 2019. Relation-aware entity alignment for heterogeneous knowledge graphs. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization.

Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*.

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.

Jinzhu Yang, Ding Wang, Wei Zhou, Wanhui Qian, Xin Wang, Jizhong Han, and Songlin Hu. 2021a. Entity and relation matching consensus for entity alignment. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2331–2341.

Shiquan Yang, Rui Zhang, Sarah M Erfani, and Jey Han Lau. 2021b. Unimf: A unified framework to incorporate multimodal knowledge bases intoend-to-end task-oriented dialogue systems. In *IJCAI*, pages 3978–3984.

Kaisheng Zeng, Chengjiang Li, Lei Hou, Juanzi Li, and Ling Feng. 2021. A comprehensive survey of entity alignment for knowledge graphs. *AI Open*, 2:1–13.

Fu Zhang, Jianwei Li, and Jingwei Cheng. 2023. Improving entity alignment via attribute and external knowledge filtering. *Applied Intelligence*, 53(6):6671–6681.

Yuxin Zhang, Bohan Li, Han Gao, Ye Ji, Han Yang, Meng Wang, and Weitong Chen. 2021. Fine-grained evaluation of knowledge graph embedding model

in knowledge enhancement downstream tasks. *Big Data Research*, 25:100218.

Ziheng Zhang, Hualuo Liu, Jiaoyan Chen, Xi Chen, Bo Liu, Yuejia Xiang, and Yefeng Zheng. 2020. An industry evaluation of embedding-based entity alignment. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 179–189.

Xiang Zhao, Weixin Zeng, Jiuyang Tang, Xinyi Li, Minnan Luo, and Qinghua Zheng. 2022. Toward entity alignment in the open world: an unsupervised approach with confidence modeling. *Data Science and Engineering*, 7(1):16–29.

Qiannan Zhu, Xiaofei Zhou, Jia Wu, Jianlong Tan, and Li Guo. 2019. Neighborhood-aware attentional representation for multilingual knowledge graphs. In *ijcai*, pages 1943–1949.

Yao Zhu, Hongzhi Liu, Zhonghai Wu, and Yingpeng Du. 2021. Relation-aware neighborhood matching model for entity alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4749–4756.

Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. 2015. Building a large-scale multimodal knowledge base system for answering visual queries. *arXiv preprint arXiv:1507.05670*.

A Appendix: Datasets Statistics

Table 2 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 employ Hits@n and MRR as metrics to evaluate all the methods. Hits@n means the rate correct entities rank in the top n according to similarity computing. MRR denotes the mean reciprocal rank of correct entities. The higher values of Hits@n and MRR explain the better performance of the method.

C Appendix: Iterative Results on Two Cross-KG Datasets

Table 3 shows the iterative results on two cross-KG datasets. On the FB15K-DB15K with a 20% data split, Hits@1 dropped by 1.1%. We attribute this to substantial noise in the attribute modality. According to the statistics of datasets in Appendix A, DB15K-FB15K contains significantly more attribute information than FB15K-YAGO15K. This observation also supports the conclusions of Zhang et al. (2023) and Shi et al. (2022). On the FB15K-DB15K with 50% and 80% data splits, Hits@10 is also affected by this interference. Even so, our framework can still obtain correct alignment results from it. However, on the FB15K-YAGO15K dataset with less available information, our framework once again demonstrates strong generalization capabilities, especially in the 20% data split setting, where Hits@1 and MRR increase by 7.2% and 6.4% respectively.

D Appendix: Non-iterative and Iterative Results on Three Bilingual Datasets

Table 5 shows the non-iterative and iterative results on three bilingual datasets. The slight increase on the DBP15K is primarily due to the cross-lingual

Dataset	KG	Ent.	Rel.	Attr.	Rel. Tr.	Attr. Tr.	Image	EA pairs
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	
FB15K-DB15K	FB15K	14,951	1,345	116	592,213	29,395	13,444	12,846
	DB15K	12,842	279	225	89,197	48,080	12,837	
FB15K-YAGO15K	FB15K	14,951	1,345	116	592,213	29,395	13,444	11,199
	YAGO15K	15,404	32	7	122,886	23,532	11,194	

Table 2: Statistics of multi-modal datasets, with EA pairs representing the reference entity alignments.

Seeds	Models	FB15K-DB15K			FB15K-YAGO15K		
		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(Ours)	0.567	0.804	0.652	**0.516**	**0.733**	**0.593**
	Improv. best%	-1.1	-0.8	-0.9	7.2	4.1	6.4
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(Ours)	**0.692**	0.869	**0.757**	**0.658**	**0.827**	**0.720**
	Improv. best%	0.2	-0.2	0.2	4.6	1.9	3.8
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(Ours)	**0.787**	0.919	**0.838**	**0.752**	**0.899**	**0.804**
	Improv. best%	0.3	-0.2	0.4	2.8	1.9	2.1

Table 3: Iterative results on two cross-KG datasets are presented.

nature of the data, which tends to be more sparse and unbalanced. This makes it challenging to ensure alignment and consistency between different languages. In some settings, our framework still maintains the SOTA performance (1.000). Experimental results show that surface information (e.g., entity names) still plays a very positive role in entity alignment. However, we recommend that future studies use more visual information and discard entity names, considering the name bias.

E Appendix: Is Attention Mechanism Better than Bottleneck Layer?

In Section 3.4, we employ the Bottleneck Layer (BL) to calculate the Intra-modal Weight (IAW). An important question arises: does the attention mechanism outperform BL in computing IAW? Our experimental results demonstrate otherwise. When we replaced BL with the attention mechanism on the $\text{DBP15K}_{FR\text{-}EN}$, Hits@1 achieved only 0.744, indicating a 3.5% decrease compared to using BL. MRR achieved only 0.816, indicating a 2.8% decrease compared to using BL. Additionally, the training time was tripled compared to BL.

F Appendix: Statistics of Learnable Parameters

In Table 4, we present the number of learnable parameters for the baselines on DBP15K. While MEAformer demonstrates improved performance compared to MCLEA, it increases the number of learnable parameters by 0.5 M. Compared to MEAformer, our framework not only enhances performance but also increases the number of learnable parameters by only 0.2 M, representing a reduction of 0.2 M compared to MSNEA.

Models	**Paras.**
EVA (Liu et al., 2021)	13.3 M
MSNEA (Chen et al., 2022)	14.1 M
MCLEA (Lin et al., 2022)	13.2 M
MEAformer (Chen et al., 2023a)	13.7 M
RICEA (Ours)	13.9 M

Table 4: Statistics of learnable parameters on DBP15K, using a non-iterative method without (w/o) surface form (SF).

This highlights the potential of our framework for few-sample data training and generalization, as well as its advantage in lightweight computing. Moreover, reducing the consumption of computing resources will be a focus of our future research.

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
w/o SF and Non-iterative									
AlignEA (Sun et al., 2018)	0.472	0.792	0.581	0.448	0.789	0.563	0.481	0.824	0.599
KECG (Li et al., 2019)	0.478	0.835	0.598	0.490	0.844	0.610	0.486	0.851	0.610
MUGNN (Cao et al., 2019)	0.494	0.844	0.611	0.501	0.857	0.621	0.495	0.870	0.621
AliNet (Sun et al., 2020)	0.539	0.826	0.628	0.549	0.831	0.645	0.552	0.852	0.657
EVA (Liu et al., 2021)	0.680	0.910	0.762	0.673	0.908	0.757	0.683	0.923	0.767
MSNEA (Chen et al., 2022)	0.601	0.830	0.684	0.535	0.775	0.617	0.543	0.801	0.630
MCLEA (Lin et al., 2022)	0.715	0.923	0.788	0.715	0.909	0.785	0.711	0.909	0.782
MEAformer (Chen et al., 2023a)	0.771	0.951	0.835	0.764	**0.959**	0.834	0.770	0.961	0.841
MDSEA (Fang and Yan, 2024)	0.768	0.904	0.814	0.769	0.946	0.832	0.765	0.947	0.834
RICEA(Ours)	**0.774**	**0.954**	**0.840**	**0.770**	0.953	**0.837**	**0.779**	**0.961**	**0.844**
Improv. best%	0.3	0.3	0.5	0.1	-0.6	0.3	0.9	0.0	0.3
w/ SF and Non-iterative									
RDGCN (Wu et al., 2019)	0.708	0.846	-	0.767	0.895	-	0.886	0.957	-
AttrGNN (Liu et al., 2020)	0.777	0.920	0.829	0.763	0.909	0.816	0.942	0.987	0.959
RNM (Zhu et al., 2021)	0.840	0.919	0.870	0.872	0.944	0.899	0.938	0.981	0.954
CLEM (Wu et al., 2022)	0.854	0.935	0.879	0.885	0.958	0.904	0.936	0.977	0.952
RPR-RHGT (Cai et al., 2022)	0.693	-	0.754	0.886	-	0.912	0.889	-	0.919
ERMC (Yang et al., 2021a)	0.903	0.946	0.899	0.942	0.944	0.925	0.962	0.982	0.973
EVA (Liu et al., 2021)	0.929	0.986	0.951	0.964	0.997	0.976	0.990	0.999	0.994
MSNEA (Chen et al., 2022)	0.887	0.961	0.913	0.938	0.983	0.955	0.969	0.997	0.980
MCLEA (Lin et al., 2022)	0.926	0.983	0.946	0.961	0.994	0.973	0.987	0.999	0.992
MEAformer (Chen et al., 2023a)	0.948	0.993	0.965	0.977	**0.999**	0.986	0.991	1.000	0.995
RICEA(Ours)	**0.950**	**0.993**	**0.967**	**0.978**	0.998	**0.988**	**0.991**	**1.000**	**0.995**
Improv. best%	0.2	0.0	0.2	0.1	-0.1	0.2	0.0	0.0	0.0
w/o SF and Iterative									
BootEA (Sun et al., 2018)	0.629	0.847	0.703	0.622	0.854	0.701	0.653	0.874	0.731
NAEA (Zhu et al., 2019)	0.650	0.867	0.720	0.641	0.873	0.718	0.673	0.894	0.752
EVA (Liu et al., 2021)	0.746	0.910	0.807	0.741	0.918	0.805	0.767	0.939	0.831
MSNEA (Chen et al., 2022)	0.643	0.865	0.719	0.572	0.832	0.660	0.584	0.841	0.671
MCLEA (Lin et al., 2022)	0.811	0.954	0.865	0.806	0.953	0.861	0.811	0.954	0.865
MEAformer (Chen et al., 2023a)	0.847	0.970	0.892	0.842	0.974	0.892	0.845	0.976	0.894
RICEA(Ours)	**0.858**	**0.971**	**0.896**	**0.843**	**0.976**	**0.896**	**0.847**	**0.979**	**0.898**
Improv. best%	1.1	0.1	0.4	0.1	0.2	0.4	0.2	0.3	0.4
w/ SF and Iterative									
EVA (Liu et al., 2021)	0.956	0.993	0.969	0.979	0.998	0.987	0.995	0.999	0.997
MSNEA (Chen et al., 2022)	0.896	0.969	0.922	0.942	0.986	0.958	0.971	0.998	0.982
MCLEA (Lin et al., 2022)	0.964	0.996	0.977	0.986	0.999	0.992	0.995	1.000	0.997
MEAformer (Chen et al., 2023a)	0.973	0.998	0.983	0.991	1.000	**0.995**	0.996	1.000	**0.998**
RICEA(Ours)	**0.977**	**0.998**	**0.989**	**0.996**	**1.000**	0.992	**0.997**	**1.000**	0.997
Improv. best%	0.4	0.0	0.6	0.5	0.0	-0.3	0.1	0.0	-0.1

Table 5: Non-iterative and iterative results on three bilingual datasets, with (w/) and without (w/o) surface forms (SF) are presented.