On Modality Weighting and Specificity for Multi-Modal Entity Alignment

Yu Xing[1*], Qizhuo Xie[1*], Yunhui Liu[1], Qing Gu[1], Tao Zheng[1], Bin Chong[2†], Tieke He[1†]

[1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
[2]National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing 100871, China
chongbin@pku.edu.cn, hetieke@gmail.com

Abstract

Multi-modal entity alignment aims to identify equivalent entities across different multi-modal knowledge graphs (MMKGs). While prior work has achieved notable progress through improved multi-modal encoding and cross-modal fusion techniques, two critical challenges remain unresolved. First, due to the heterogeneous and often inconsistent sources from which MMKGs are constructed, the quality and informativeness of modalities vary significantly across entities, leading to the modality weighting problem. Second, existing cross-modal fusion mechanisms predominantly emphasize modality-shared information, often at the expense of modality-specific signals that are also essential for precise alignment. To address these issues, we propose *HUMEA*, a novel framework that integrates hierarchical Mixture-of-Experts (MoE) with unimodal distillation. *HUMEA* consists of: (1) A hierarchical MoE module comprising intra-modal and inter-modal experts, which adaptively modulates modality contributions by capturing entity representations at fine-to-coarse semantic granularities. In addition, we introduce a contrastive mutual information loss to enhance expert diversity and reduce redundancy. (2) A unimodal distillation strategy that preserves modality-specific information in the fused representations through single-modality alignment and distillation, achieving a balanced integration of shared and unique modality features. Extensive experiments on two benchmark datasets, FB15K-DB15K and FB15K-YAGO15K, demonstrate state-of-the-art performance, validating the effectiveness of our approach.

Code — https://github.com/mikumifa/HUMEA

Introduction

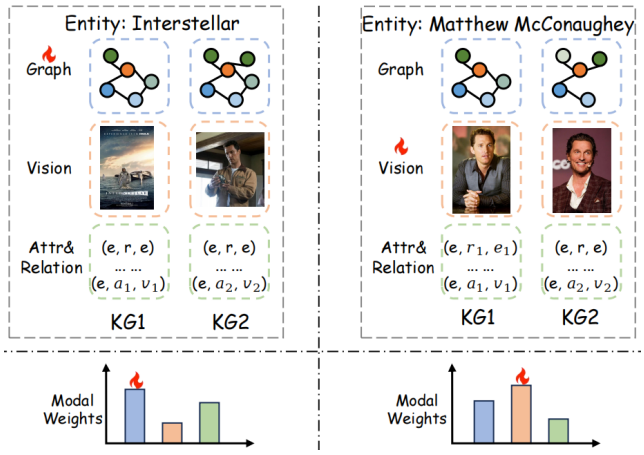
Multi-modal knowledge graphs encapsulate rich real-world semantics by integrating diverse modalities, such as structural links, visual content, relations, and attributes. Owing to their versatility and effectiveness in various downstream applications, such as recommender systems (Wu et al. 2024; Guo et al. 2024) and cross-modal retrieval (Liang et al. 2024; Qian et al. 2021), MMKGs have garnered increasing attention in recent years. However, most MMKGs are constructed independently from heterogeneous sources (e.g., DBpedia and Freebase), which often results in inconsistencies and semantic discrepancies across different graphs. Consequently, multi-modal entity alignment (MMEA) has emerged as a critical research task, aiming to identify semantically equivalent entities across disparate MMKGs.

Current MMEA methods (Liu et al. 2021; Chen et al. 2022; Huang et al. 2024; Cheng, Guo, and Zhang 2025) have achieved significant improvements by enhancing modality encoding and cross-modal fusion. Despite these efforts, two major challenges remain unresolved. The first is the dynamic modality weighting. Since multi-modal knowledge graphs are constructed from heterogeneous data sources, the quality and informativeness of each modality can vary significantly across entities during alignment. For instance, as shown in Figure 1a (left), the movie entity Interstellar exhibits a similar graph structure in both KG1 and KG2. However, the visual information differs substantially: one image presents the movie poster, whereas the other depicts the male lead actor. In this case, the graph structure provides more reliable alignment cues and should therefore be assigned a higher weight. Improperly emphasizing the visual modality may mislead the model, for example, aligning the Interstellar entity in KG2 with the Matthew McConaughey entity in KG1. In contrast, as shown in Figure 1a (right), some entities benefit more from the visual modality, which must be given greater weight to ensure correct alignment. These observations underscore the importance of a modality weighting mechanism that adaptively responds to the varying reliability of each modality across entities.

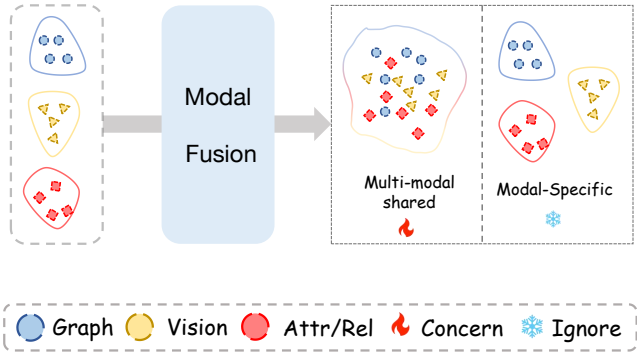
The second limitation is that most existing MMEA methods overlook modality-specific features when executing cross-modal fusion. These methods typically adopt simple fusion strategies, such as concatenation or tensor fusion, to project multi-modal data into a unified representation space. However, recent studies (Han et al. 2024; Fang et al. 2025) have demonstrated that such approaches often bias the model toward learning modality-invariant shared features, thereby overlooking modality-specific information. This oversight can lead to the loss of distinctive characteristics inherent to each modality, which are often crucial for accurate alignment.

To address the aforementioned challenges, we propose *HUMEA*, an MMEA framework that integrates a hierar-

[*]These authors contributed equally.
[†]Bin Chong and Tieke He are the corresponding authors.



(a) Modality dynamic weights for different entity samples.



(b) The neglect of modality-specific features.

Figure 1: Two challenges in multi-modal entity alignment.

chical MoE module with a unimodal distillation strategy. *First*, inspired by the success of Mixture-of-Experts (MoE) (Cai et al. 2025) techniques in adaptively weighting expert features, we design a hierarchical MoE module, comprising two levels: intra-modal MoE and inter-modal MoE. Specifically, the former constructs a dedicated set of experts for each modality, enabling the extraction of fine-grained, modality-specific representations. To further reduce redundancy across experts, we introduce an expert disentanglement loss, formulated via contrastive mutual information, which encourages diversity and decorrelation among expert outputs. The latter treats each modality as a global expert and employs a routing network to dynamically weigh their contributions based on the input entity features. This design allows the model to adaptively emphasize higher-quality modalities, thereby addressing modality quality imbalance across heterogeneous entity instances. *Second*, to retain valuable modality-specific information that is often suppressed during fusion, we introduce a unimodal distillation strategy. This involves two components: (1) a unimodal alignment loss for each modality, which ensures that unimodal representations preserve their individual alignment capabilities; and (2) a unimodal distillation loss, which guides the learning of multi-modal features using unimodal features. This strategy facilitates the integration of both shared and specific modality semantics, thereby improving the expressiveness and alignment accuracy of the learned representations.

Extensive experiments conducted on two benchmark MMEA datasets, FB15K-DB15K and FB15K-YAGO15K, demonstrate the superior performance and effectiveness of our approach. The main contributions of our work can be summarized as follows:

- We propose a hierarchical mixture-of-experts module that enables dynamic modality weighting, allowing the model to adaptively process and fuse multi-modal information cross entity.
- A unimodal distillation method is introduced to retain the alignment capacity of each individual modality, preserving modality-specific information through unimodal alignment and distillation loss.
- Extensive experiments on two benchmark datasets, FB15K-DB15K and FB15K-YAGO15K, have achieved the state-of-the-art results, validating the effectiveness of our approach.

Related-Work

Multi-modal Entity Alignment

Multi-modal entity alignment aims to identify entities that refer to the same real-world object across different knowledge graphs. Existing approaches primarily enhance alignment performance by enhancing modality encoding and cross-modal fusion strategies. MMEA(Chen et al. 2020) proposed a simple concatenation of encoded modalities to map entities into a unified semantic space. EVA(Liu et al. 2021) introduced a learnable scalar to weight multi-modal embeddings and introduced a neighborhood consistency loss to guide alignment. MSNEA(Chen et al. 2022) leveraged visual features to guide the encoding of relational and attribute modalities. MEAformer(Chen et al. 2023a) employed a multi-modal transformer integration that performs meta-modality integration by dynamically predicting entity-level modality weights for feature aggregation. MCLEA(Lin et al. 2022) addressed modality discrepancy through contrastive learning. PCMEA(Wang et al. 2024) proposed a pseudo-label calibration mechanism with momentum-based contrastive learning to make full use of the labeled and unlabeled data. Most recently, RICEA(Li et al. 2025) tackled dynamic modality weighting by computing interaction-aware weights and recalibrating them using uncertainty estimates.

Despite these notable advancements, existing methods still face challenges in addressing dynamic modality weighting and neglecting modality-specific information. Our approach bridges these gaps by introducing the MoE framework and unimodal distillation strategy.

Mixture of Experts

The Mixture of Experts method is designed to tackle complex tasks by combining the specialized knowledge of multiple experts, which is recently widely used in AI-related fields like computer vision(Chen et al. 2023b; Wang et al.

2025; Kumar and Marttinen 2024), natural language processing(Zhang et al. 2024; Lu et al. 2025), and recommendation systems(Hou et al. 2022; Bian et al. 2023). In multimodal learning, recent approaches(Yu et al. 2025; Wu et al. 2025a,b) often treat different modalities as distinct experts and employ a routing network to assign sample-specific weights for adaptive fusion. However, to the best of our knowledge, no work has explored the performance of MoE techniques in the MMEA task. In this paper, we investigate the effectiveness of MoE in addressing the dynamic modality weighting issue in MMEA, considering both intra-modal and inter-modal perspectives for finer-grained control.

Method

Problem Definition

Multi-modal Knowledge Graph. A multi-modal knowledge graph typically consists of relational, attribute, and visual information. Formally, it can be defined as $G = \{E, R, I, A, V, T_r, T_a\}$, where E, R, I, A, and V denote the sets of entities, relations, images, attributes, and attribute values, respectively. Furthermore, $T_r = \{(e, r, e^{'})|e, e^{'} \in E, r \in R\}$ represent the set of relational triples, and $T_a = \{(e, a, v)|e \in E, a \in A, v \in V\}$ represent the set of attribute triples.

Multi-modal Entity Alignment. The goal of multi-modal entity alignment is to identify entity pairs from two knowledge graphs that refer to the same real-world concept by leveraging multi-modal information. Formally, given two multi-modal knowledge graphs $G_1 = \{E_1, R_1, I_1, A_1, V_1, T_r^1, T_a^1\}$ and $G_2 = \{E_2, R_2, I_2, A_2, V_2, T_r^2, T_a^2\}$, the alignment task aims to discover the set of equivalent entity pairs, which is denoted as $H = \{(e_1, e_2)|e_1 \equiv e_2, e_1 \in E_1, e_2 \in E_2\}$. In practice, a subset of H is typically provided as aligned seed to guide model training, and then the model is expected to align the remaining unmatched entity pairs.

Framework Overview

In this paper, we present *HUMEA*, an MMEA framework designed to address the modality weighting and specificity issues. *HUMEA* consists of three main components: (1) Multi-modal Knowledge Embedding, which extracts graph structure, visual content, relational context and attribute information with diverse encoders; (2) Hierarchical MoE, which adaptively balances modality weights from fine- to coarse-grained levels; (3) Unimodal Distillation, which preserves modality-specific information within multi-modal representations.

Multi-modal Knowledge Embedding

In this subsection, we describe the process for obtaining the initial feature representations for each modality of entity.

Graph Structure Embedding. To encode the graph structural context of each entity, we construct neighborhood-aware embeddings by aggregating information from adjacent nodes. Specifically, we choose a two-layer Graph Attention Network (GAT)(Velickovic et al. 2017) to capture local topological patterns, with the final-layer output serving as the intermediate structural embedding. The structural embedding h_i^g of entity e_i can be formally described as follows:

$$h_i^g = \text{GAT}_2(W_g, M_g, e_i^g), \tag{1}$$

where g represents graph structure modality, M_g represents the adjacency matrix, $e_i^g \in R^d$ is the randomly initialized graph embedding of entity e_i, d is the hidden dimension, and $W_g \in R_{d\times d}$ is a diagonal weight matrix.

Visual Knowledge Embedding. To extract visual features of each entity e_i, we utilize a pre-trained visual encoder ResNet-152(He et al. 2016) to encode. Specifically, we feed the image v_i of entity e_i into the pre-trained visual model and use the final layer output before the softmax layer as the image feature. The initial visual feature embedding e_i is generated as follows:

$$h_i^v = \text{ResNet}(v_i). \tag{2}$$

Relation/Attribute Knowledge Embedding. In MMKGs, relational and attribute triples also provide essential descriptive information about entities, which is essential for accurate alignment. In this paper, we adopt two approaches to embed this information. First, we convert relational and attribute triples into textual sentences and encode their semantics using a pre-trained BERT model. For example, for a given entity e_i with n attribute triples $(e_i, a_1, v_1), (e_i, a_2, v_2), ..., (e_i, a_n, a_n)$, we construct a sentence s_a in the form: "a_1 is v_1,a_2 is v_3...a_n is v_n.". Then the sentence is fed into BERT, and the final hidden layer output is used as the attribute embedding. The same procedure is applied to relational triples to obtain the semantic relation embedding. Second, we employ bag-of-words(BOW) features to encode the relations and attributes of entity e_i. The full process is detailed as follows:

$$h_i^{k_t} = \text{BERT}(s_i^k), k \in \{r, a\}, \tag{3}$$

$$h_i^{k_b} = \text{BOW}(t_i^k), k \in \{r, a\}, \tag{4}$$

where n represents modality and $\text{BERT}(\cdot)$ represents the hidden feature vector in the last layer of BERT.

Consequently, we denote the six modalities of entity e_i as h_i^m and $m \in \mathcal{M} = \{g, v, r_t, a_t, r_b, a_b\}$, and similarly hereafter.

Hierarchical Mixture of Experts

To enable the model to adapt to the dynamic modality quality of each entity, we design a hierarchical MoE module, consisting of two levels: intra-modal and inter-modal:

Intra-Modal MoE. After obtaining the raw features of each modality, we further enhance them using intra-modal MoE layers that capture entity-specific preferences within each modality. This design addresses the observation that, during alignment, different entities may focus on different aspects of the same modality. For example, when processing images, a person entity may focus more on facial details, while a city entity may focus more on landmarks. Specifically, we construct six modality-specific MoE components corresponding to the six modalities: Graph-based MoE (Graph-MoE),

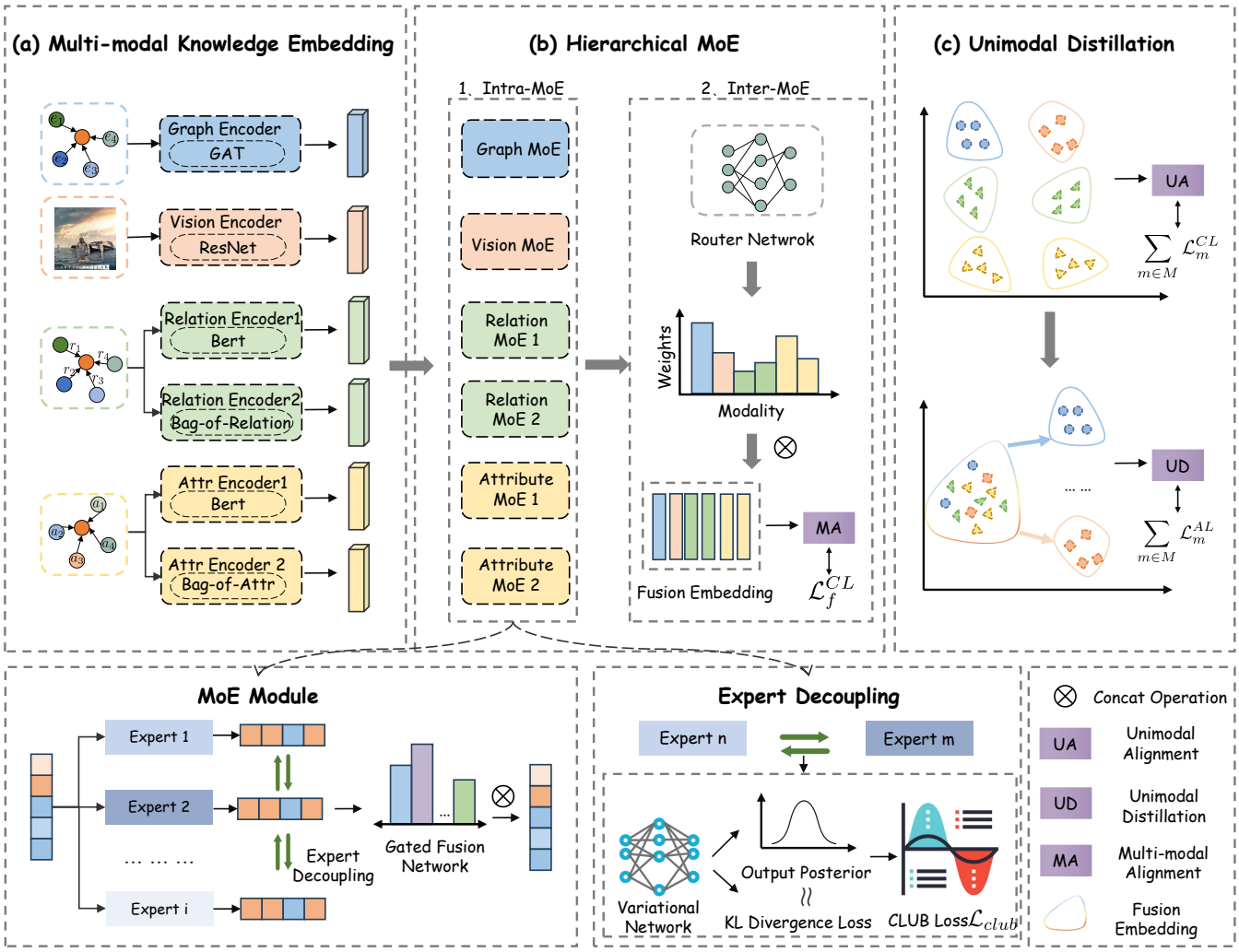


Figure 2: The overview of our framework, which first incorporates a heterogeneous multi-modal knowledge embedding. Second, a hierarchical MoE module is designed, comprising intra-modal and inter-modal MoE, where the intra-modal MoE includes an expert mutual information loss to reduce redundancy. Finally, we introduce a unimodal distillation strategy, which integrates single-modal alignment loss and unidirectional distillation loss to retain modality-specific information.

Image-based MoE (Img-MoE), Text-based MoE for relations/attributes (Text-MoE), and Bag-of-Words-based MoE for relations/attributes (BoW-MoE).

Each MoE consists of multiple expert networks, where each expert is implemented as a feedforward network (FFN) and learns to capture complementary intra-modal patterns. The outputs from each expert are aggregated through a learned gating mechanism, which assigns dynamic weights to each expert based on the entity input. This design allows the model to learn dynamic modality embeddings for different entities during alignment. Each intra-modal MoE shares the same configuration. Formally:

$$\mathrm{FFN}(h_i^m) = (\mathrm{ReLU}(h_i^m W_1 + b_1))W_2 + b_2, \tag{5}$$

$$h_i^{m'} = \sum_{k=1}^{K} g_k \cdot \mathrm{FFN}_k(h_i^m), \tag{6}$$

$$\mathbf{g} = \mathrm{Softmax}(h_i^m \cdot w_3 + b_3), \tag{7}$$

where k represents the number of experts, g_k is the weight coefficient of the k-th expert from the routing vector $\mathbf{g}$, and $h_i^{m'}$ denotes the output of intra-modal MoE.

In addition, we introduce an expert information disentanglement loss to reduce redundancy. Specifically, we aim to minimize the mutual information between the outputs of different experts, encouraging each expert to learn distinct intra-modal patterns. To achieve this, we adopt the Contrastive Log-ratio Upper Bound (CLUB) (Cheng et al. 2020) method, which provides a tractable variational upper bound on mutual information. The formulation of CLUB can be denoted as:

$$\begin{aligned} I_{\mathrm{CLUB}}(x;y) :=& \mathbb{E}_{p(x;y)}[\log p(y|x)] \\ & - \mathbb{E}_{p(x)}\mathbb{E}_{p(y)}[\log p(y|x)], \end{aligned} \tag{8}$$

Here, x and y are two variables, which in our context can be

replaced by the output vectors of expert. However, since the true conditional probability distribution $p(y|x)$ is generally intractable, we employ a variational approximation network to estimate it. Denote the output of the k-th expert for e_i and modality m as $h^i_{k,m}$. Therefore, for the K intra-modal experts of modality m, the the CLUB-based disentanglement loss is defined as:

$$\mathcal{L}_{CLUB} = \frac{1}{K^2}\sum_{m\in\mathcal{M}}\sum_{i\in\mathcal{B}}\sum_{a=1}^{K}\sum_{b\neq i}^{K}(\log Q_{\theta,m}(h^i_{b,m} \mid h^i_{a,m}) - \sum_{i'\in\mathcal{B}\setminus\{i\}}\log Q_{\theta,m}(h^{i'}_{b,m} \mid h^i_{a,m})), \tag{9}$$

where $\mathcal{B}$ is a batch of entities, and $Q_{\theta,m}(y|x)$ represents a variational approximation of truth posterior distribution $p(y|x)$, parameterized by a neural network θ for modality m. The term i' refers to another entity sampled from the same batch $\mathcal{B}$.

Inter-modal MoE. After enhancing representations with fine-grained intra-modal MoE, we further design an inter-modal MoE to dynamically adjust the contribution of each modality at a coarse-grained level. In this design, Unlike conventional MoE architectures where each expert is a separate neural network, we treat each modality as a distinct expert and design a router network that outputs a refined set of weights conditioned on the input, which enables more specialized and powerful capabilities. The formal representation is as follows:

$$h_c = \bigoplus_{m\in\mathcal{M}} h_i^{m'}, \tag{10}$$

$$\begin{aligned} W &= [w_g, w_v, w_{r_t}, w_{a_t}, w_{r_b}, w_{a_b}] \\ &= g(h_c;\theta) = \sigma(\alpha_i g_i(h_c)), \end{aligned} \tag{11}$$

where $\oplus$ denotes the concatenation operation, g is the routing network with parameters θ, and softmax is applied to normalize the weights. The symbol σ denotes a normalization operation, and $\alpha_i g_i$ represents multiple linear layers. Then, w_m denotes the dynamic learned modality weights.

After obtaining the weights for each modality, we integrate the multi-modal features into a fusion embedding h^f_i through a weighted concatenation operation:

$$h^f_i = \text{Concat}[w_m \cdot h_i^{m'}]. \tag{12}$$

Then we employ a multi-modal contrastive loss as the primary optimization objective of our framework. This loss is designed to reduce the distance between aligned entity pairs while increasing the distance between unaligned pairs. In our setting, the pre-aligned seed set naturally serves as positive samples, while all other non-aligned pairs are treated as negatives. Formally, for the i-th entity $e^1_i \in E_1$ in a mini-batch $\mathcal{B}$, the positive set is defined as $\mathcal{P}^i_2 = \{e^2_i | e^2_i \in E_2\}$, where (e^1_i, e^2_i) is an aligned pair, E_1 and E_2 denote the sets of entities in graph G_1 and G_2, respectively. The negative set includes two parts, intra-graph negatives from the graph $G_1 : \mathcal{N}^1_j = \{e^1_j | \forall e^1_j \in E_1, j \neq i\}$, and cross-graph negatives from the graph $G_2 : \mathcal{N}^2_j = \{e^2_j | \forall e^2_j \in E_2, j \neq i\}$. Overall, we define the alignment probability distribution $q_f(e^1_i, e^2_i)$ for each positive pair (e^1_i, e^2_i) as:

$$q_f(e^1_i, e^2_i) = \frac{\delta(h^{1,f}_i, h^{2,f}_i)}{\delta(h^{1,f}_i, h^{2,f}_i) + \sum_{e^1_j\in\mathcal{N}^1_j}\delta(h^{1,f}_i, h^{1,f}_j) + \sum_{e^2_j\in\mathcal{N}^2_j}\delta(h^{1,f}_i, h^{2,f}_j)}, \tag{13}$$

where $\delta(u, v) = \exp(u^T \cdot v/\mathcal{T})$, and $\mathcal{T}$ is a temperature coefficient. Notably, the alignment distribution is directional and asymmetric for each input. Therefore, the distribution for another direction is defined as $q_f(e^i_2, e^i_2)$ analogously. The overall contrastive loss is then defined as:

$$\mathcal{L}^{CL}_f = -\mathbb{E}_{i\in\mathcal{B}}\log\left[\frac{1}{2}\left(q_f(e^i_1, e^i_2) + q_f(e^i_2, e^i_1)\right)\right]. \tag{14}$$

Unimodal Distillation

Different modalities encode complementary aspects of an entity from diverse perspectives. However, as previously discussed, the fused multi-modal representation may suppress or overlook modality-specific details during training. To address this issue and preserve modality-specific information, we draw inspiration from knowledge distillation (Hinton, Vinyals, and Dean 2015) and propose a unimodal distillation strategy. This strategy allows each individual modality to guide the learning of the fused representation, thereby retaining its unique semantic contributions. Specifically, we first introduce a unimodal alignment loss for each modality m to ensure that unimodal representations retain effective alignment capability. The unimodal alignment loss for a given positive pair (e^1_i, e^2_i) is defined as:

$$\mathcal{L}^{CL}_m = \mathcal{L}^{CL}_m(h^{1,m}_i, h^{2,m}_i). \tag{15}$$

Then, we employ a unidirectional knowledge distillation mechanism to transfer knowledge from unimodal embeddings to the fused multi-modal representation. This mechanism encourages the fused representation to absorb modality-specific information, resulting in a more fine-grained and informative representation. Concretely, we minimize the discrepancy between the unimodal and fused representations by introducing an alignment loss. For aligned sample pairs (e^i_1, e^i_2), the distillation loss is defined as:

$$\begin{aligned} \mathcal{L}^{AL}_m = -\sum_m \mathbb{E}_{i\in\mathcal{B}}[&D_{KL}(Q^h_m(e^1_i, e^2_i)\|Q^h_f(e^1_i, e^2_i)) \\ &+ D_{KL}(Q^h_m(e^2_i, e^1_i)\|Q^h_f(e^2_i, e^2_i))], \end{aligned} \tag{16}$$

where $D_{KL}(\cdot)$ represents the Kullback-Leibler Divergence, $Q^h_f(e^1_i, e^2_i) = h^{1,f}_i \otimes h^{2,f}_i$ and h^f_i is the fusion embedding.

Optimization Objective

The overall loss of the *HUMEA* is as follows:

$$\mathcal{L} = \mathcal{L}^{CL}_f + \lambda_1\sum_{m\in\mathcal{M}}\mathcal{L}^{CL}_m + \mathcal{L}_{club} + \lambda_2\sum_{m\in\mathcal{M}}\mathcal{L}^{AL}_m, \tag{17}$$

where λ_1 and λ_2 are adjustable hyperparameters.

Seeds	Model	FB15K-DB15K				FB15K-YAGO15K			
		Hits@1	Hits@5	Hits@10	MRR	Hits@1	Hits@5	Hits@10	MRR
20%	MMEA	0.2648	0.4513	0.5411	0.3570	0.2339	0.3976	0.4800	0.3170
	EVA	0.1990	-	0.4480	0.2830	0.1530	-	0.3610	0.2240
	MSNEA	0.1140	-	0.2960	0.1750	0.1030	-	0.2490	0.1530
	MCLEA	0.2950	-	0.5820	0.3930	0.2540	-	0.4840	0.3320
	MEAformer	0.4170	-	0.7150	0.5180	0.3270	-	0.5950	0.4170
	PCMEA(*)	0.3379	0.5552	0.6417	0.4410	0.3064	0.4995	0.5812	0.3990
	RICEA	0.4710	-	0.7200	0.5570	0.4110	-	0.6580	0.4970
	Ours	**0.5118**	**0.6997**	**0.7643**	**0.5980**	**0.4393**	**0.6278**	**0.6989**	**0.5280**
50%	MMEA	0.4165	0.6210	0.7035	0.5120	0.4026	0.5723	0.6451	0.4860
	EVA	0.3340	-	0.5890	0.4220	0.3110	-	0.5340	0.3880
	MSNEA	0.2880	-	0.5900	0.3880	0.3200	-	0.5890	0.4130
	MCLEA	0.5550	-	0.7840	0.6370	0.5010	-	0.7050	0.5740
	MEAformer	0.6190	-	0.8430	0.6980	0.5600	-	0.7780	0.6390
	PCMEA(*)	0.5487	0.7547	0.8112	0.6420	0.5148	0.7071	0.7697	0.6050
	RICEA	0.6480	-	0.8520	0.7210	0.6170	-	0.8110	0.6870
	Ours	**0.6949**	**0.8434**	**0.8803**	**0.7630**	**0.6491**	**0.8032**	**0.8474**	**0.7200**
80%	MMEA	0.5903	0.8041	0.8687	0.6850	0.5976	0.7849	0.8389	0.6820
	EVA	0.4840	-	0.6960	0.5630	0.4910	-	0.6920	0.5650
	MSNEA	0.5180	-	0.7790	0.6130	0.5310	-	0.7780	0.6200
	MCLEA	0.7350	-	0.8900	0.7900	0.6670	-	0.8240	0.7220
	MEAformer	0.7650	-	0.9160	0.8200	0.7030	-	0.8730	0.7660
	PCMEA(*)	0.7152	0.8601	0.9037	0.7820	0.6804	0.8297	0.8748	0.7490
	RICEA	0.7760	-	0.9160	0.8290	0.7340	-	0.8920	0.7920
	Ours	**0.8094**	**0.9146**	**0.9353**	**0.8560**	**0.7737**	**0.8904**	**0.9196**	**0.8270**

Table 1: Non-iterative results on FB15K-DB15K and FB15K-YAGO15K with different proportions of entity alignment seeds. The best results are highlighted in bold. The "-" denotes that the results are not available and "*" denotes reproduced results.

Experiments

Experimental Settings

Datasets. To verify the effectiveness of our proposed framework in real-world scenarios, we conduct experiments on two publicly available cross-KG datasets FB15K-DB15K and FB15K-YAGO15K. In addition, following established protocols, we use three different proportions of alignment seed datasets for training, specifically 20%, 50%, and 80%, with the remaining data used as the test set for performance validation.

Baselines. The optimal baseline model for the MMEA task is the RICEA (Li et al. 2025), which also addresses the dynamic modality weighting issue. Therefore, we select RICEA as our main baseline. In addition, we choose 7 other state-of-the-art multi-modal alignment methods, which are: MMEA (Chen et al. 2020), and EVA (Liu et al. 2021), MSNEA (Chen et al. 2022), MCLEA (Lin et al. 2022), MEAFormer (Chen et al. 2023a) and PCMEA(Wang et al. 2024). Since PCMEA does not report results under the non-iterative training setting, we report it utilizing the official open-source code. The results of all other baselines are taken directly from the RICEA paper.

Implementation Details. The hidden layer dimensions for all networks are standardized to 300. The total number of epochs is set to $1,000$, with an optional iterative training strategy applied for an additional 500 epochs. Our model is implemented based on Pytorch, an open-source deep learning framework. All experiments were conducted on a server with two GPUs (NVIDIA Tesla V100). To fully leverage the large amount of unlabeled data, and in line with previous works (Wang et al. 2024; Li et al. 2025), we also adopt an iterative training strategy, in which newly pseudo-labeled entity pairs are progressively incorporated into the training set. For fair comparison and practical relevance, we first report results under the standard (non-iterative) training setup, followed by results from the iterative training regime.

Results

We compare our method against several state-of-the-art and representative multi-modal entity alignment baselines on the FB15K-DB15K and FB15K-YAGO15K datasets. The performance of all methods under the non-iterative training setting is summarized in Table 1. Our model consistently achieves significant improvements across both datasets. For instance, on FB15K-DB15K, our model improves the Hits@1 metric by 4.00%, 4.70%, and 3.34% under different training seed settings, respectively, compared to the baseline RICEA. Similarly, on FB15K-YAGO15K, our method achieves relative improvements of 2.83%, 3.21%, and 3.97% under the same training ratios. These results underscore the effectiveness of our framework in addressing modality weighting across entities and preserving modality-specific information. Furthermore, we evaluate all methods under the iterative training setting, where pseudo-labeled seed pairs are progressively added during training. The results are reported in the code repository.

Method	FB15K-DB15K			
	Hits@1	Hits@5	Hits@10	MRR
Ours	**0.5118**	**0.6997**	**0.7643**	**0.5980**
w/o g	0.4131	0.5090	0.5643	0.4640
w/o v	0.4803	0.6747	0.7391	0.5680
w/o r_b	0.4736	0.6560	0.7191	0.5560
w/o r_t	0.4841	0.6748	0.7362	0.5690
w/o a_b	0.4731	0.6646	0.7291	0.5590
w/o a_t	0.4844	0.6750	0.7383	0.5690
w/o $\mathcal{L}_m^{CL}$	0.4840	0.6722	0.7346	0.5680
w/o $\mathcal{L}_m^{AL}$	0.4864	0.6753	0.7389	0.5710
w/o $\mathcal{L}_{CLUB}$	0.5033	0.6892	0.7576	0.5918
w/o $HMoE$	0.4778	0.6680	0.7304	0.5620

Table 2: Ablation Study of the key components in *HUMEA* on FB15K-DB15K dataset with 20% pre-aligned seed set.

Ablation Study

To assess the effectiveness of each component in our framework, we conducted a series of ablation studies, with results presented in Table 2. Several key observations can be seen: (1) Among all modalities, the graph structure proves to be the most critical, underscoring its importance in capturing rich semantic knowledge. Additionally, the bag-of-words features for relations and attributes consistently outperform their corresponding textual features, suggesting that aligned entities rely more heavily on relation and attribute types rather than their semantic formulations. (2) The removal of unimodal alignment leads to a substantial performance drop, emphasizing the necessity of preserving modality-specific signals essential for accurate alignment. Moreover, eliminating the unimodal distillation loss results in a further decline in performance, reinforcing the importance of transferring knowledge from unimodal representations to the fused multi-modal embedding. (3) Lastly, the inclusion of the hierarchical Mixture-of-Experts design yields clear performance gains. This confirms that dynamically adjusting modality weights based on their instance-specific quality is an effective strategy for improving alignment performance.

Hyperparameter Analysis

We conduct a hyperparameter analysis on the FB15K-DB15K dataset to evaluate the sensitivity of our model to key configuration choices. The main hyperparameters considered include: the number of intra-MoE experts k, temperature parameter $\mathcal{T}$, adjustable coefficient λ_1, and λ_2. The results are presented in Figure 3. Several observations can be obtained: (1) The number of experts should not be too small, once a sufficient number is reached, the performance tends to stabilize. (2) Both λ_1 and λ_2 exhibit relatively stable behavior with only minor fluctuations. (3) Notably, the temperature $\mathcal{T}$ has the most significant impact on performance and needs to be maintained around 0.1. Otherwise, it may lead to a catastrophic imbalance in the routing behavior.

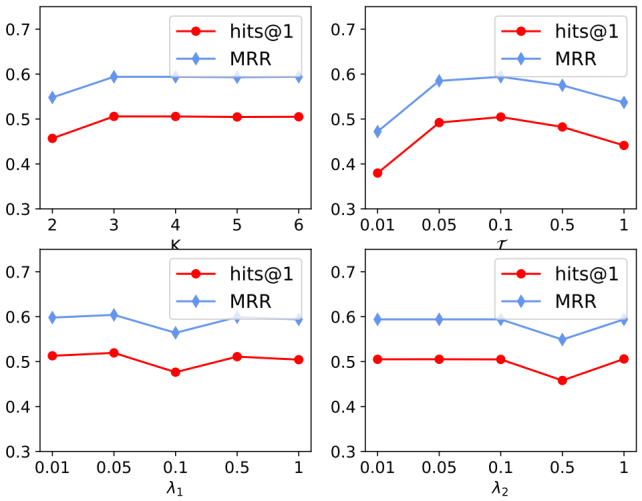


Figure 3: Hyperparameter analysis on the number of expert k, the temperature parameter $\mathcal{T}$, coefficient λ_1 and λ_2.

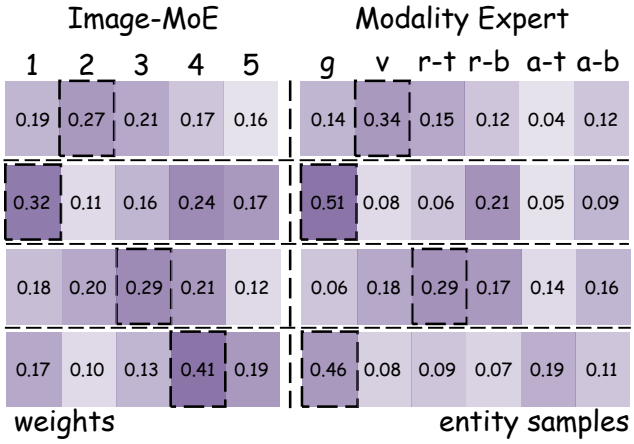


Figure 4: Visualization of modality weight for different entity samples. The number of image-MoE experts is set to 5.

Dynamic Weight Visualization

To further illustrate the practical effectiveness of *HUMEA*, we select four representative entity pairs from the FB15K-DB15K dataset and visualize both the img-moe expert importance and inter-modal modality weights, as shown in Figure 4. Darker color intensity indicates higher contribution. The results highlight *HUMEA*'s capability to adjust feature importance based on different entity instances.

Conclusion

This paper tackles modality weighting and specificity issues in multi-modal entity alignment by introducing *HUMEA*. It integrates multi-modal knowledge encoders, a hierarchical mixture-of-expert module, and a unimodal distillation strategy. Experimental and empirical analysis validate the superiority and universality of *HUMEA*.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (7240041225, 62306137).

References

Bian, S.; Pan, X.; Zhao, W. X.; Wang, J.; Wang, C.; and Wen, J.-R. 2023. Multi-modal mixture of experts represetation learning for sequential recommendation. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, 110–119.

Cai, W.; Jiang, J.; Wang, F.; Tang, J.; Kim, S.; and Huang, J. 2025. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*.

Chen, L.; Li, Z.; Wang, Y.; Xu, T.; Wang, Z.; and Chen, E. 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*, 134–147. Springer.

Chen, L.; Li, Z.; Xu, T.; Wu, H.; Wang, Z.; Yuan, N. J.; and Chen, E. 2022. Multi-modal siamese network for entity alignment. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 118–126.

Chen, Z.; Chen, J.; Zhang, W.; Guo, L.; Fang, Y.; Huang, Y.; Zhang, Y.; Geng, Y.; Pan, J. Z.; Song, W.; et al. 2023a. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *Proceedings of the 31st ACM international conference on multimedia*, 3317–3327.

Chen, Z.; Shen, Y.; Ding, M.; Chen, Z.; Zhao, H.; Learned-Miller, E. G.; and Gan, C. 2023b. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11828–11837.

Cheng, J.; Guo, M.; and Zhang, F. 2025. SGMEA: Structure-Guided Multimodal Entity Alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, 7851–7861.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, 1779–1788. PMLR.

Fang, Y.; Huang, W.; Wan, G.; Su, K.; and Ye, M. 2025. EMOE: Modality-Specific Enhanced Dynamic Emotion Experts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14314–14324.

Guo, Z.; Li, J.; Li, G.; Wang, C.; Shi, S.; and Ruan, B. 2024. Lgmrec: Local and global graph learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8454–8462.

Han, X.; Nguyen, H.; Harris, C.; Ho, N.; and Saria, S. 2024. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. *Advances in Neural Information Processing Systems*, 37: 67850–67900.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hou, Y.; Mu, S.; Zhao, W. X.; Li, Y.; Ding, B.; and Wen, J.-R. 2022. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 585–593.

Huang, Y.; Zhang, X.; Zhang, R.; Chen, J.; and Kim, J. 2024. Progressively modality freezing for multi-modal entity alignment. *arXiv preprint arXiv:2407.16168*.

Kumar, Y.; and Marttinen, P. 2024. Improving medical multi-modal contrastive learning with expert annotations. In *European Conference on Computer Vision*, 468–486. Springer.

Li, C.; Cheng, J.; Tong, Q.; Zhang, F.; and Wang, C. 2025. Probing Relative Interaction and Dynamic Calibration in Multi-modal Entity Alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*. To appear.

Liang, M.; Du, J.; Liang, Z.; Xing, Y.; Huang, W.; and Xue, Z. 2024. Self-supervised multi-modal knowledge graph contrastive hashing for cross-modal search. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 13744–13753.

Lin, Z.; Zhang, Z.; Wang, M.; Shi, Y.; Wu, X.; and Zheng, Y. 2022. Multi-modal Contrastive Representation Learning for Entity Alignment. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 2572–2584. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Liu, F.; Chen, M.; Roth, D.; and Collier, N. 2021. Visual pivoting for (unsupervised) entity alignment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 4257–4266.

Lu, Y.; Zhang, S.; Cheng, D.; Liang, G.; Xing, Y.; Wang, N.; and Zhang, Y. 2025. Training Consistent Mixture-of-Experts-Based Prompt Generator for Continual Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19152–19160.

Qian, S.; Xue, D.; Zhang, H.; Fang, Q.; and Xu, C. 2021. Dual adversarial graph neural networks for multi-label cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2440–2448.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *stat*, 1050(20): 10–48550.

Wang, L.; Qi, P.; Bao, X.; Zhou, C.; and Qin, B. 2024. Pseudo-label calibration semi-supervised multi-modal entity

alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 9116–9124.

Wang, Y.; Liu, Y.; Zheng, A.; and Zhang, P. 2025. Decoupled feature-based mixture of experts for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8141–8149.

Wu, C.; Shuai, Z.; Tang, Z.; Wang, L.; and Shen, L. 2025a. Dynamic modeling of patients, modalities and tasks via multi-modal multi-task mixture of experts. In *The Thirteenth International Conference on Learning Representations*.

Wu, Q.; Ke, Z.; Zhou, Y.; Sun, X.; and Ji, R. 2025b. Routing experts: Learning to route dynamic experts in existing multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*.

Wu, W.; Wang, C.; Shen, D.; Qin, C.; Chen, L.; and Xiong, H. 2024. Afdgcf: Adaptive feature de-correlation graph collaborative filtering for recommendations. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, 1242–1252.

Yu, Y.; Cao, C.; Zhang, Y.; Lv, Q.; Min, L.; and Zhang, Y. 2025. Building a multi-modal spatiotemporal expert for zero-shot action recognition with clip. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9689–9697.

Zhang, Y.; Chen, Z.; Guo, L.; Xu, Y.; Hu, B.; Liu, Z.; Zhang, W.; and Chen, H. 2024. Multiple heads are better than one: Mixture of modality knowledge experts for entity representation learning. *arXiv preprint arXiv:2405.16869*.