Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.069690

ARTICLE

Multi-Modal Pre-Synergistic Fusion Entity Alignment Based on Mutual Information Strategy Optimization

Huayu Li[1,2], Xinxin Chen[1,2], Lizhuang Tan[3,4,*], Konstantin I. Kostromitin[5,6], Athanasios V. Vasilakos[7] and Peiying Zhang[1,2]

[1]Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, 266580, China
[2]Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software, China University of Petroleum (East China), Qingdao, 266580, China
[3]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, 250014, China
[4]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, 250014, China
[5]Department of Physics of Nanoscale Systems, South Ural State University, Chelyabinsk, 454080, Russia
[6]Institute of Radioelectronics and Information Technologies, Ural Federal University, Yekaterinburg, 620002, Russia
[7]Department of ICT and Center for AI Research, University of Agder (UiA), Grimstad, 4879, Norway

*Corresponding Author: Lizhuang Tan. Email: tanlzh@sdas.org

Received: 28 June 2025; Accepted: 02 September 2025; Published: 23 September 2025

ABSTRACT: To address the challenge of missing modal information in entity alignment and to mitigate information loss or bias arising from modal heterogeneity during fusion, while also capturing shared information across modalities, this paper proposes a Multi-modal Pre-synergistic Entity Alignment model based on Cross-modal Mutual Information Strategy Optimization (MPSEA). The model first employs independent encoders to process multi-modal features, including text, images, and numerical values. Next, a multi-modal pre-synergistic fusion mechanism integrates graph structural and visual modal features into the textual modality as preparatory information. This pre-fusion strategy enables unified perception of heterogeneous modalities at the model's initial stage, reducing discrepancies during the fusion process. Finally, using cross-modal deep perception reinforcement learning, the model achieves adaptive multi-level feature fusion between modalities, supporting learning more effective alignment strategies. Extensive experiments on multiple public datasets show that the MPSEA method achieves gains of up to 7% in Hits@1 and 8.2% in MRR on the FBDB15K dataset, and up to 9.1% in Hits@1 and 7.7% in MRR on the FBYG15K dataset, compared to existing state-of-the-art methods. These results confirm the effectiveness of the proposed model.

KEYWORDS: Knowledge graph; multi-modal; entity alignment; feature fusion; pre-synergistic fusion

1 Introduction

In the contemporary era of information proliferation, Knowledge Graphs (KGs) serve as structured semantic knowledge bases that play a pivotal role in connecting and integrating vast amounts of heterogeneous data. Researchers extensively utilize this representation across diverse domains, including intelligent question answering [1], recommendation systems [2,3], and information extraction [4,5], among others.

However, most knowledge graphs are constructed from multiple heterogeneous data sources, resulting in substantial duplication of nodes across distinct KGs. Although these nodes represent identical real-world entities, they exhibit divergent representations, leading to structural redundancy and graph incompleteness. To address this challenge, Entity Alignment (EA) techniques have been developed. Entity Alignment (EA) techniques aim to identify corresponding entities referring to the same real-world objects across different knowledge graphs [6]. Traditional EA methodologies predominantly focus on unimodal knowledge graphs, whereas real-world entities inherently exhibit multimodal characteristics. Conventional EA approaches often neglect inter-modal interactions and complementarities, limiting comprehensive entity understanding and resulting in suboptimal alignment accuracy. Consequently, Multimodal Entity Alignment (MMEA) has emerged as an essential research direction. Multimodal knowledge learning focuses on leveraging multiple modalities to support downstream applications, such as emotion recognition [7,8] and cross-modal retrieval [9,10]. Nevertheless, in MMEA tasks, the heterogeneity and complexity of multimodal data present significant challenges for effective modal fusion. An example of an MMKG is illustrated in Fig. 1.

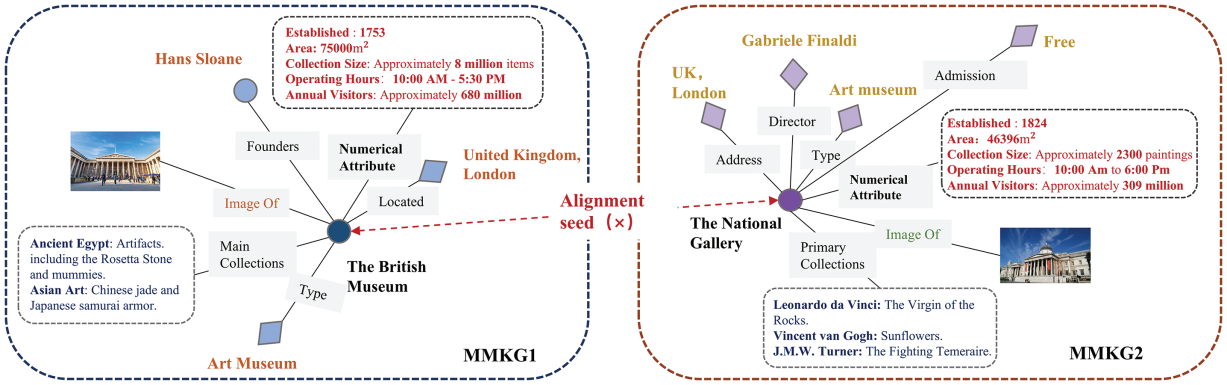

Figure 1: Multi-modal knowledge graph entity alignment

This paper proposes a Multi-modal Pre-synergistic Fusion framework based on Cross-modal Mutual Information Strategy Optimization. This framework addresses the challenges posed by inter-modal heterogeneity and aims to mitigate information loss or bias during modal fusion. The model effectively captures shared features across modalities by incorporating mutual information maximization, thereby enabling deep multi-level adaptive modal fusion. Specifically, the principal contributions of this work are as follows:

- To significantly mitigate inherent modal heterogeneity in cross-modal learning and establish tighter feature correlations, this paper innovatively proposes a Multi-modal Pre-Synergistic Fusion Mechanism. This framework proactively injects graph structural features and visual modal information as preparatory information into the cross-modal attention mechanism during the model's initial stage, thereby achieving unified cross-modal perception. Furthermore, this dual synergistic fusion strategy provides multi-perspective perception capabilities throughout the fusion process, enabling collaborative discriminative decision-making based on heterogeneous information sources. It significantly reinforces cross-modal feature correlations, ultimately achieving maximum improvements of 16% and 13.1% in Hits@1 metrics on the FBDB15K and FBYG15K datasets, respectively, compared to baseline methods.
- We design a novel reinforcement learning method based on cross-modal mutual information maximization. This approach ensures maximal information retention during modal fusion, reducing potential information loss or bias. Additionally, feedback from the reward-punishment mechanism progressively optimizes action selection and learns optimal coordination strategies.

- Extensive comparative experiments are conducted on benchmark datasets (FB15K, DB15K, and YAGO15K). The results demonstrate that the proposed MPSEA model achieves state-of-the-art performance. The Hits@1 metric showed maximum improvements of 7% and 9.1%, respectively. The MRR metric achieved maximum gains of 8.2% and 7.7%, respectively.

2 Related Work

2.1 Entity Alignment

Structure-based entity alignment approaches are primarily categorized into two major classes:

1. EA Based on Relational Triple Translation Models: These methods represent relational knowledge as triples (Head Entity, Relation, Tail Entity), exemplified by MTransE and BootEA [11]. However, a significant limitation of this approach is its inefficiency in effectively capturing information from higher-order neighboring nodes.

2. EA Utilizing Graph Neural Networks (GNNs): GNN-based EA methods [12] have garnered increasing research attention. These techniques enhance entity representations by aggregating information from neighboring nodes and edges, thereby enriching local features while simultaneously considering the global structural modeling of the KG. This category of methods—exemplified by GCN-Align [13], and RDGCN [14]—updates entity representations by aggregating information from neighboring nodes and edges. Such integration enriches local features and effectively accounts for global structural modeling of the knowledge graph.

2.2 Multi-Modal Entity Alignment

Entity alignment (EA) methods based on structural embedding primarily focus on the relational modalities of knowledge graphs (KGs). They generally encounter geometric vector space limitations [6,15]. Consequently, integrating auxiliary modal information has become essential for addressing EA challenges. EVA [16] leveraged visual similarity to establish visual anchors within a multi-modal framework and expanded training sets through iterative learning. However, these methods focus on enhancing entity semantic embeddings through various modal representations, often neglecting inter-entity interactions. MCLEA [17] improves cross-modal interactions and minimizes modal discrepancies by incorporating contrastive learning. This framework facilitates proximity between equivalent entities in vector space while ensuring separation of dissimilar entities during entity representation learning. Despite advancing modal interaction, MCLEA employs static cross-modal weighting schemes that inadequately account for cross-modal source preferences, including modal absence and attribute cardinality variations. MEAformer [18] introduces a dynamic cross-modal weight assignment mechanism, pioneering entity-level feature fusion. To further optimize data resource utilization, MEAIE [19] incorporates auxiliary numerical modalities, effectively mitigating modal incompleteness issues. While existing methods have advanced entity alignment, significant limitations persist in addressing cross-modal heterogeneity. We introduce a mutual information optimization strategy coupled with a multi-modal pre-synergistic fusion mechanism to bridge this gap. Furthermore, our approach reinforces the perception of inter-modal interdependencies, optimizing discriminative fusion of multimodal information.

3 Our Approach

This section focuses on the notational definition of the multimodal entity alignment task and the proposed modeling framework.

3.1 Problem Definition

First, the MMKG is structured as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{N}, \mathcal{I}, \mathcal{T}\}$, where $\mathcal{E}$ denotes the set of entities, $\mathcal{R}$, $\mathcal{A}$, $\mathcal{N}$, and $\mathcal{I}$ represent the set of various modal data, including relationships, attributes, values, and images, respectively, and $\mathcal{T}$ denotes the set of relational triples, where $(h, r, t) \in \mathcal{T}$ forms a triad; h represents the head entity, t represents the tail entity, and r signifies the relation connecting the head and tail entities. Let $X^{(r)}, X^{(a)}, X^{(n)}$, and $X^{(v)}$ denote the embeddings for relationships, attributes, values, and visual representations, respectively. MMEA refers to the process of achieving a 1-to-1 mapping of corresponding entities between the source knowledge graph $\mathcal{G}_1 = (\mathcal{E}_1, \mathcal{R}_1, \mathcal{A}_1, \mathcal{N}_1, \mathcal{I}_1, \mathcal{T}_1)$ and the target knowledge graph $\mathcal{G}_2 = (\mathcal{E}_2, \mathcal{R}_2, \mathcal{A}_2, \mathcal{N}_2, \mathcal{I}_2, \mathcal{T}_2)$. That is, $L = \{(e_u, e_v) \mid e_u \in \mathcal{E}_1, e_v \in \mathcal{E}_2, e_u \equiv e_v\}$ represents the set of aligned anchor pairs in the cross-knowledge graph, where $\equiv$ denotes the equivalence of two entities in the real world.

3.2 Overview

To address the modal deficiencies identified in MMEA, this paper incorporates numerical modal embeddings alongside conventional modalities, proposing the Pre-synergistic Fusion Entity Alignment Model Optimized by Mutual Information Strategy (MPSEA). As illustrated in Fig. 2, the framework comprises four core components: 1) Multi-modal Knowledge Representation, 2) Multi-modal Feature Fusion, 3) Cross-modal Deep Perception-based Reinforcement Learning (CPLR), and 4) Joint Multi-modal Loss Optimization. The model first extracts entity-related information across five modalities: relational, attributive, visual, numerical, and structural. We then employ a multi-modal pre-synergistic fusion mechanism to dynamically weight cross-modal features, generating joint embeddings through integrated early and late fusion strategies. Subsequently, the reinforcement learning module implements a mutual information-driven reward-penalty mechanism. This is synergized with modal-adaptive contrastive learning to produce a joint alignment loss, which optimizes anchor entity representations through iterative refinement.

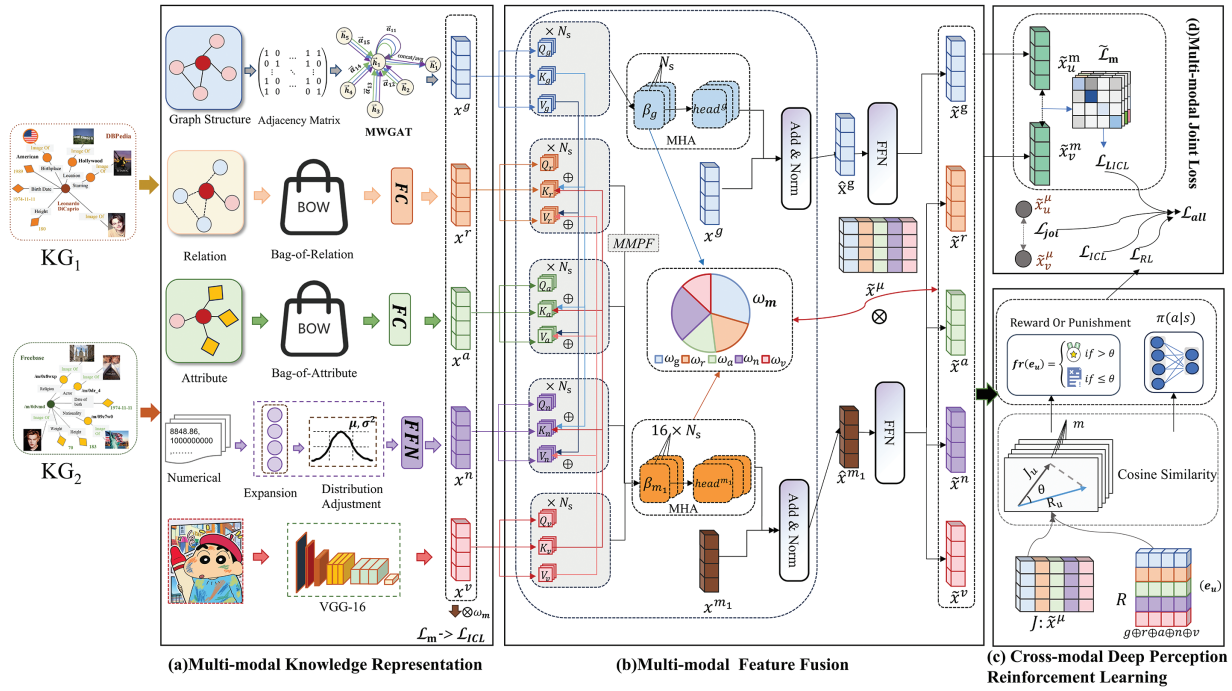

Figure 2: MPSEA modeling framework

3.3 Multi-Modal Knowledge Representatiog

Our modalities include visual, structural, relational, and attribute data, with numerical attribute values. The collective term for relations and attributes is the textual modality. This module is used to obtain the feature vectors required for early fusion of each modality x_m, where $m \in \{g, v, r, a, n\}$.

3.3.1 Graph Neighborhood Structural Embedding

To mitigate contextual gaps [20] (encompassing Entity-Attribute Cardinality Discrepancy and Modal Attribute Discrepancy) and enhance node-neighbor differentiation capability, we employ a two-headed, two-layer GAT [21] architecture. This framework incorporates dedicated linear transformation matrices $\mathbf{W}_i$ and $\mathbf{W}_j$, which separately process entity node features and neighbor node features. The output of the final GAT layer is used as the structural embedding $\mathbf{x}_i^g$. The formulation is defined in Eq. (1):

$$\tilde{e}_{ij} = \text{LeakyReLU}\left(\boldsymbol{\omega}^\top \left[W_i\mathbf{h}_i \parallel W_j\mathbf{h}_j\right]\right) \tag{1}$$

3.3.2 Attribute and Relationship Embedding

This paper integrates relational triples and attribute triples as external textual information to mitigate modal incompleteness. We employ Bag-of-Words (BoW) [22] representation to encode relational features $\mathbf{h}_i^r$ and attribute features $\mathbf{h}_i^a$ for entity e_i. These features are subsequently projected through dedicated fully connected layers to derive relation embeddings $\mathbf{x}_i^r$ and attribute embeddings $\mathbf{x}_i^a$, respectively. The formulation is as follows (see Eq. (2)):

$$\mathbf{h}_i^{m_3} = \text{bag-of-words}(m_3), \quad m_3 \in \{r, a\} \tag{2}$$

Here, r and a represent the relationship and attribute modalities, respectively. W_{m_3} is a learnable weight matrix, and $\mathbf{h}_i^{m_3} \in \mathbb{R}^{d_{m_3}}$ represents the input features of entity e_i in each modality.

3.3.3 Visual Embedding

Disambiguating homonymous entities solely through textual modalities presents significant challenges within EA tasks. Visual information (image color and texture features) provides an effective filter for eliminating false alignments. This paper employs the pre-trained VGG-16 model [23] to extract visual features $\mathbf{h}_i^v$ for entities. These features are transformed via a fully-connected layer to derive visual embeddings $\mathbf{x}_i^v$. Crucially, we generate corresponding visual representations for image-missing entities. We found that only 94.56% of the entities in the FBDB15K dataset have image features, and only 81.17% of the entities in the FBYG15K dataset have image features. Therefore, following the EVA [16] method, the missing image features are randomly initialized into 4096-dimensional feature vectors obeying a normal distribution. The formal definition is as follows (see Eq. (3)):

$$x_i^v = \text{FC}_v(W_v, h_i^v) = \text{FC}_v(W_v, \text{VGG16}(v)) \tag{3}$$

Here, v signifies the visual modality, where W_v is the learnable weight matrix, and $h_i^v \in \mathbb{R}^{d_v}$ represents the input features of entity e_i in the visual modality.

3.3.4 Numerical Embedding

We implement the Deep Numerical Embedding Framework (DNEF) to encode numerical attributes, which projects scalar values into high-dimensional vector spaces. This representation is subsequently refined

through neural networks to enhance numerical feature expressiveness. For each numerical feature $n(h^n, a_i)$, we expand it into a d-dimensional vector $\varphi(n_{(h^n,a_i)}) \in \mathbb{R}^d$.

$$\varphi(n_{(h^n,a_i)}) = n(h^n, a_i) \cdot \gamma_i + \zeta_i \tag{4}$$

Here, $\varphi(n_{(h^n,a_i)})$ denotes the numerical embedding corresponding to the numerical triple, a_i represents the attribute key embedding, $\gamma_i, \zeta_i \in \mathbb{R}^d$ are learnable scaling and shifting parameters that dynamically regulate per-dimension distribution characteristics. Next, $\varphi(n_{(h^n,a_i)})$ is fed into a feedforward neural network, resulting in the numerical embedding $x^n \in \mathbb{R}^d$.

In this study, we utilize the numerical triples extracted by Liu [24] for model training. The processing involves two sequential operations: first, normalizing numerical triples per entity to obtain normalized representations $\mathbf{h}^n$; second, encoding these representations through the DNEF to derive numerical embeddings $\mathbf{x}^n$.

3.4 Multi-Modal Features Fusion

3.4.1 Graph Structure Transformer

We use the encoder part of the Transformer [25] to process the graph structure, which consists of two main core components: the multi-head self-attention block (MHA) and the fully connected feedforward network (FFN).

MHA component (working with n_s parallel heads): First, for each modality $m \in \{g, v, r, a, n\}$, generate the query (Q_m), key (K_m), and value (V_m) required for the i-th self-attentive head. Specifically, for each modality m, we compute from the input embedding x^m and the corresponding weight matrices $W_q^{m(i)}$, $W_k^{m(i)}$, $W_v^{m(i)}$ as:

$$Q_m^{(i)} = x^m W_q^m(i), \quad K_m^{(i)} = x^m W_k^m(i), \quad V_m^{(i)} = x^m W_v^m(i) \tag{5}$$

When modality $m = g$, we perform the dot product computation of Q_g and K_g to get the graph structure attention weight β_g and matrix multiplication with V_g to get the i-th self-attention head $head_i^g$. Then we concatenate the n_s heads to obtain the graph structure embedding output x^g, i.e., MHA (x^g). Next, we process the output of the MHA using layer normalization(LN) and residual joining to obtain the further processing of the graph structure embedding $\hat{x}^g$.

FFN component: It consists of two fully connected layers with a ReLU activation function. Subsequently, the output of the feedforward layer is processed using layer normalization and residual connections, resulting in the graph structure embedding $\tilde{x}^g$ required for late fusion.

3.4.2 Multi-Modal Pre-Synergistic Fusion Transformer (MMPF)

We propose a Multi-modal Pre-synergistic Fusion with a Deep Perception mechanism to mitigate cross-modal heterogeneity between visual features, graph structures, and textual representations while enhancing inter-modal perception. Specifically, we embed key-value pairs $(k^{(v)}, v^{(v)})$ from visual modalities and $(k^{(g)}, v^{(g)})$ from graph structures—obtained via Multi-Head Attention (MHA)—into textual modality key-value pairs $(k^{(t)}, v^{(t)})$ (encompassing relations, attributes, and values). This integration provides multi-perspective contextual awareness during fusion, achieving hierarchical feature fusion through cross-modal self-attention. The computational workflow proceeds as follows:

Multi-modal Pre-synergistic Fusion MHA Component: We redefine the key $K_i^{(m_2)}$ and value $V_i^{(m_2)}$ of the i-th head for the textual modality $m_2 \in \{r, a, n\}$. The key $K_{m_2}^{(i)}$ and $V_{m_2}^{(i)}$ are guided by the multi-modal pre-synergistic mechanism, and through weighted concatenation with the keys and values of the graph structure modality g and the visual modality v, the new key $\hat{K}_{m_2}^{(i)}$ and $\hat{V}_{m_2}^{(i)}$ are obtained.

$$\hat{K}_{m_2}^{(i)} = K_{m_2}^{(i)} \oplus \lambda_g * K_g^{(i)} \oplus \lambda_v * K_v^{(i)}; \qquad \hat{V}_{m_2}^{(i)} = V_{m_2}^{(i)} \oplus \lambda_g * V_g^{(i)} \oplus \lambda_v * V_v^{(i)} \tag{6}$$

To unify the notation, we will use $K_{m_2}^{(i)}$ and $V_{m_2}^{(i)}$ to replace $\hat{K}_{m_2}^{(i)}$ and $\hat{V}_{m_2}^{(i)}$ in the following; λ_g and λ_v represent the weighting coefficients for the visual and graph structure modalities, respectively, and are learnable parameters.

Unlike the graph structure transformer, we perform the cross-modal self-attention mechanism in the other modalities $m_1 \in \{r, a, n, v\}$. In the i-th head, the pre-synergistic fused $K_{m_2}^i$, $V_{m_2}^i$ are used to calculate the cross-modal attention weights $\beta_{m_1,z}$ (denoted as cross-modal attention weights between m_1 and z modalities). Then, the weighted fusion with the corresponding values yields the i-th self-attention head $\text{head}_i^{m_1}$. After concatenating the n_s heads, we obtain the output $\text{MHA}(x_{m_1})$ for each modality embedding x_{m_1}. Next, we apply layer normalization (LN) and residual connections to process the output of MHA, obtaining the further processed m_1 modality embedding $\hat{x}^{m_1}$. The specific calculation process is as follows (see Eq. (7)):

$$\beta_{m_1 z} = \frac{\exp(Q_{m_1}^T K_z)/\sqrt{d_s}}{\sum_{i \in m_1} \exp(Q_{m_1}^T K_i)/\sqrt{d_s}}; \qquad \text{MHA}(x^{m_1}) = \bigoplus_{i=1}^{n_s} \text{head}_i^{m_1} \cdot W_o; \quad m_1 \in \{r, a, n, v\} \tag{7}$$

FFN component: it is divided into two parts

1. Visual Modality Processing (v): To mitigate cross-modal noise interference and enhance discriminative capability for critical information, we integrate an OutlookerAttention (OA) module within the Feed-Forward Network (FFN) layer. This module employs local sliding window aggregation to refine fine-grained feature processing. The final visual embeddings $\tilde{\mathbf{x}}^v$ are computed directly by the feed-forward network as $\tilde{\mathbf{x}}^v = \text{ReLU}\Big(\text{OA}(\hat{\mathbf{x}}^v W_1 + b_1)\Big) W_2 + b_2$.

2. Textual Modality m_2: We adopt the same processing as the graph structure Transformer.

3.4.3 Multi-Modal Fusion

To define the cross-modal weights of each modality,we weighted-average the attention weight scores of each modality and subsequently applied Softmax normalization to derive these weights.

1. The weight of modality m_1 is w_{m_1}:

$$w_{m_1} = \frac{\exp\left(\sum_{i=1}^{N_s} \beta_{m_1}^{(i)} / \sqrt{|M_1| \times N_s}\right)}{\sum_{k \in M_1} \exp\left(\sum_{i=1}^{N_s} \beta_{m_1}^{(i)} / \sqrt{|M_1| \times N_s}\right) + \exp\left(\sum_{i=1}^{N_s} \beta_g^{(i)} / \sqrt{N_s}\right)} \tag{8}$$

2. The weight of the graph structure modality g is w_g:

$$w_g = \frac{\exp\left(\sum_{i=1}^{N_s} \beta_g^{(i)} / \sqrt{N_s}\right)}{\sum_{k \in M_1} \exp\left(\sum_{i=1}^{N_s} \beta_{(m_1)}^{(i)} / \sqrt{|M_1| \times N_s}\right) + \exp\left(\sum_{i=1}^{N_s} \beta_g^{(i)} / \sqrt{N_s}\right)} \tag{9}$$

Subsequently, cross-modal weights are utilized to perform weighted concatenation of multi-modal embeddings, yielding joint embeddings x_u^{μ} for early fusion [18] and $\tilde{x}_u^{\mu}$ for late fusion [18] of the model's u-th entity.

3.5 Reinforcement Learning Based on Cross-Modal Deep Perception

Inspired by Zeng et al. [26], this paper proposes a reinforcement learning (RL) approach based on mutual information strategy optimization. Following deep multilevel fusion, we employ a custom mutual information estimator as a quality checking mechanism. This ensures maximal information preservation during modal fusion while mitigating inter-modal information loss or bias.

The model's predicted joint embedding must approximate the ground-truth joint embedding in entity alignment. We therefore define:

- Predicted embedding $J = \tilde{\mathbf{x}}_u^{\mu}$ (cross-modally weighted joint embedding)
- Ground-truth embedding $R = \bigoplus_{m \in \mathcal{M}} \tilde{\mathbf{x}}^m$ (modality-wise concatenation)

A custom cross-modal mutual information estimator is used to compute the mutual information score $MI(e_u)$ between the predicted and true embeddings of entity e_u. We compute the cosine similarity to measure the correlation between the embeddings. We then exponentiate this similarity to derive the mutual information score $MI(e_u)$.

$$MI(e_u) = \exp\left(\cos_{sim}(J_u, R_u)\right) = \exp\left(\frac{J_u \cdot R_u}{\|J_u\| \cdot \|R_u\|}\right) \tag{10}$$

To encourage the model's predicted embedding J to align more closely with the true embedding R, we integrate the mutual information score and the cosine similarity to define the reward signal $r(e_u)$ and the punishment mechanism $p(e_u)$, and ultimately form the reinforcement learning reward function $f_r(e_u)$:

$$r(e_u) = \max(\cos\_\text{sim}(J_u, R_u) - \theta, 0) + \alpha \cdot MI(e_u) \tag{11}$$

$$p(e_u) = \beta \cdot (\theta - \cos\_\text{sim}(J_u, R_u)); \quad f_r(e_u) = r(e_u) - p(e_u) \tag{12}$$

Here, θ is a custom similarity threshold, α is the weight of the mutual information score, and β is the penalty coefficient proportional to θ. When the cosine similarity between J and R exceeds the threshold θ, it indicates a high similarity between the predicted embedding and the true embedding, which is considered "correct," and the model receives a reward. On the other hand, if the similarity falls below θ, a penalty mechanism is activated, resulting in a suppressed reward for the model. We further incorporate the mutual information score, and even when the cosine similarity is low, $MI(e_u)$ can still supply extra information about the relationship between modalities. If the mutual information score between modalities is high, the model will still earn a reward, even if the cosine similarity is low. The model is guided to optimize the consistency between the predicted embedding J and the true embedding R through the balance of penalty and reward mechanisms. At the same time, the model also aims to maximize the mutual information between modalities.

Next, the policy network takes on the role of the core decision-maker, aiming to compute the action probability distribution based on the current state (i.e., the comprehensive features of the entity pair). Here, $\pi(a)$ denotes the probability of selecting action a from the action space under the current policy, where the action space can be defined as a binary discrete set $\{0, 1\}$. $a = 1$ represents a "match" action, indicating the policy network's belief that the entity pair (e_u, e_v) should be aligned, while $a = 0$ represents a "mismatch"

action, signifying the policy network's belief that the entity pair should not be aligned. W_p is the weight matrix of the policy network.

The policy network processes the concatenated joint embedding J and mutual information score $MI(e_u)$, feeding this comprehensive feature vector into a linear layer to compute the action logits, which are then transformed into a probability distribution via a Softmax function. The model selects the optimal action based on the degree of sharing between the joint embedding J and the mutual information. The higher the similarity and the stronger the mutual information, the greater the probability of action selection. This paper adopts the REINFORCE algorithm, which optimizes the alignment strategy by maximizing the reward signal $fr(e_u)$ through minimizing a loss function. The formulation is as follows (see Eq. (13)):

$$\pi(a)_{e_u} = \text{softmax}(W_p \cdot [J \oplus MI(e_u)]); \qquad \mathcal{L}_{\text{RL}} = -f_r(e_u) \cdot \log(\pi(a)_{e_u}) \tag{13}$$

The overall approach leverages mutual information and a reward-penalty feedback mechanism to guide the policy network in optimizing the performance of MMEA. This process ensures a closer alignment between the predicted embeddings and the ground-truth targets. By backpropagating this loss, the parameters of the policy network are updated, thereby increasing the selection probability of actions that yield higher rewards. Consequently, the model can dynamically adapt its alignment decisions to accommodate the complex relationships and inherent uncertainty present in multimodal data. The reinforcement learning workflow is illustrated in Fig. 3:

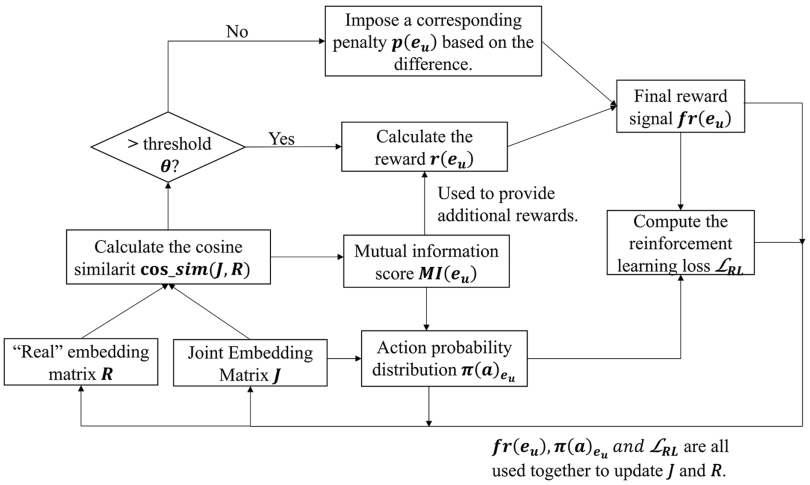

Figure 3: Cross-modal deep perception based reinforcement learning flowchart

3.6 Multi-Modal Joint Loss

In this paper, we comply with the work of Lin [17] and use a contrastive learning strategy. The pre-aligned anchor pair set L serves as the collection of positive sample anchor pairs (e_u^1, e_u^2), where each pair denotes aligned entities. We define the modality loss $\mathcal{L}_m$ for the entity pair (e_u^1, e_u^2) in modality m. Subsequently, we obtain the early fusion-based intra-modal contrastive loss $\mathcal{L}_{ICL}$ and the late fusion-based intra-modal contrastive loss $\mathcal{L}_{LICL}$. In our implementation, we replace the joint contrastive loss of early joint embeddings $p_m(e_u, e_v)$ with the contrastive loss of late joint embeddings $\tilde{p}_m(e_u, e_v)$. Here, $p_m(e_u^1, e_u^2)$ denotes the predicted alignment probability of the entity pair (e_u^1, e_u^2) in modality m. Ultimately, we sum these losses($\mathcal{L}_{joi}, \mathcal{L}_{ICL}, \mathcal{L}_{LICL}, \mathcal{L}_{RL}$) to obtain the final loss L_{all}. With the above methodological framework established, we proceed to evaluate the effectiveness and robustness of the proposed approach through

comprehensive experiments, as detailed in the following section.

$$\mathcal{L}_{ICL} = \sum_{m \in M} \mathcal{L}_m = \sum_{m \in M} -\log\left(p_m(e_u^1, e_u^2) + p_m(e_u^2, e_u^1)\right)/2 \tag{14}$$

$$\mathcal{L}_{joi} = -\log\left(\tilde{p}(e_u^1, e_u^2) + \tilde{p}(e_u^2, e_u^1)\right)/2 \tag{15}$$

The overall MPSEA mechanism is presented in Algorithm 1:

Algorithm 1: Multi-modal prefix guidance based on policy optimization.

Input: Pre-aligned anchor pairs (e_u, e_v) of $\mathcal{G}_1$ and $\mathcal{G}_2$, Modality m input features $\mathbf{h}^m$, *(e,a,n)* triples from $\mathcal{E}, \mathcal{A}, \mathcal{N}$, Adjacency matrices A_1, A_2 of $\mathcal{G}_1$ and $\mathcal{G}_2$

Output: Alignment probability p of entity pair (e_u, e_v)

Initialize the modality set $m = \{r, a, n, v, g\}$;
for each k in range($\max_{\text{epoch}}$) **do**
Embeddings for early fusion: $x_u^m \leftarrow$ for each entity e_u, pass through the multimodal knowledge representation module;
$\tilde{x}_u^g, \beta_u^g, Q_m, K_m, V_m \leftarrow$ Get graph structure embeddings for late fusion, attention weights, and preliminary query, key, and value for each modality; // Compute graph-based features and attention scores
$x_u^{m_1}, \beta_u^{m_1}$, where $m_1 \in \{r, a, n, v\} \leftarrow$ Obtain embeddings for late fusion of modality m_1 through the MMPF, attention weights for each modality; // Late fusion features for non-graph modalities
$\omega_{m_1}^g, \omega_{m_1}^u \leftarrow$ Calculate cross-modal weights; // Cross-modal weighting for feature aggregation
$x_u^\mu, \tilde{x}_u^\mu \leftarrow$ Cross-modal weighted concatenation to obtain early joint embeddings, late joint embeddings;
$f_r(e_u), \pi(a), \mathcal{L}_{\text{RL}} \leftarrow$ The reward, action probability distribution, and reinforcement learning loss are computed through the CPRL module;
$\mathcal{N}_u^{ng}, q_m(e_u, e_v) \leftarrow$ Using in-batch negative sampling strategy;
$\mathcal{L}_{\text{joi}}, \mathcal{L}_{\text{ICL}}, \mathcal{L}_{\text{LICL}}, \mathcal{L}_{\text{all}} \leftarrow$ Compute and update the loss;
Generate p;
end for
return neg_batch;

4 Experiments

To systematically evaluate the MPSEA model proposed in the previous section, this section commences with a detailed account of the experimental setup, including the data set and the evaluation metrics employed. Following this, we conduct a comparison experiment and an ablation experiment to validate the effectiveness of our proposed MPSEA model.

4.1 Experimental Setup

4.1.1 Datasets and Evaluation Metrics

This study employs the multi-modal cross-knowledge graph datasets FB15K-DB15K and FB15K-YAGO15K [24] for experimentation. We conduct three distinct dataset splits, utilizing 20%, 50%, and 80% of reference entity alignment pairs as pretraining anchors. The scale of each dataset is detailed in Table 1. We adopt Hits@n and Mean Reciprocal Rank (MRR) as performance metrics for entity alignment evaluation. The term Hits@n (n = 1, 5, 10) denotes the Top-N accuracy, which is defined as the percentage of correct entities within the top N results. The mean reciprocal rank (MRR) represents the average position of

the proper entity in the ranking. Consequently, elevated values of Hits@n and MRR are indicative of superior performance.

Table 1: Statistics of the multi-modal knowledge mapping datasets

Dataset	KG	Entity	Relation	Attribute	Image	Relation Triple	Numerical Triple	Seed
FB15K-DB15K	DB15K	10842	279	225	12837	89197	48080	12849
	FB15K	14951	1345	116	13444	592213	29395	
FB15K-YG15K	YG15K	15404	32	7	11199	122886	23532	11199
	FB15K	14951	1345	116	13444	592213	29395	

4.1.2 Implementation Details and Comparison of Models

1. Implementation details. We designed the multi-weighted GAT as a graph attention network with a 300-dimensional hidden layer and two attention heads; visual features were encoded as 4096-dimensional feature vectors. In the feature fusion stage, we use an MHA block with five attention heads. The hidden layer dimension and each modal embedding dimension are set to 300, and we performed a total of 500 epochs, with an additional 500 optional epochs. The learning rate was set to 5e-4, and we utilized the AdamW optimizer with parameters β_1 = 0.9 and β_2 = 0.999. All experiments were conducted on an Ubuntu 20.04 compute server with an NVIDIA RTXA5000 GPU, utilizing Python 3.8 and PyTorch 2.1.

2. Existing methods. To validate the effectiveness and sophistication of our method, some baseline models are selected for comparison in this paper. Two training strategies are employed: iterative and non-iterative training. Different baseline models are used for each of the two types of training: **1)** Non-iterative training: HMEA [27], MMEA [28], EVA [16], MCLEA [17], MEAformer [18], ACK-MMEA [20], GEEA [29], Confidence-MMEA [30], and TriFac [31]. **2)** Iterative training: EVA [16], MCLEA [17], MEAformer [18], MEAIE [19], MultiJAF [32], MSNEA [33], DFMKE [34], and UQMIEA [35].

4.2 Results and Analysis

To assess MPSEA's effectiveness, we conducted separate comparison experiments (i.e., validating MPSEA's performance against individual baseline models), ablation experiments (to validate the contribution of different components of MPSEA to its performance), and exploration of the effect of hyperparameters on MPSEA's performance.

4.2.1 Comparison Experiment

To validate the effectiveness of our method, the experiments encompassed both iterative and non-iterative settings, utilizing various training/testing alignment seed splits (i.e., 2:8, 5:5, and 8:2) to facilitate performance comparisons. The detailed results of the comparative experiments are demonstrated in Tables 2–4. Overall, MPSEA demonstrates exceptional performance across all metrics. In the non-iterative strategy, as illustrated in Table 2, our model surpasses the best baseline models, achieving improvements of 2.2%–7% in Hits@1 and 1.1%–8.2% in MRR. Table 3 indicates that the highest improvements in Hits@1 and MRR reach 9.1% and 7.7%, respectively.

Specifically, although MEAformer and GEEA propose relatively effective modality fusion methods, neither considers the information bias caused by inter-modal heterogeneity. This issue is successfully mitigated by introducing a multi-modal pre-synergistic fusion guidance module. Conversely, in contrast to

the majority of baseline models that focus solely on conventional modal information, we also incorporate numerical modalities, which effectively alleviates the problem of missing modalities.

Table 2: Performance of various alignment seeds for the FBDB15K non-iterative setting. Best values among the baseline models are underlined, the best results are highlighted in bold

Methods	FBDB15K (20%)				FBDB15K (50%)				FBDB15K (80%)			
	Hits			MRR	Hits			MRR	Hits			MRR
	@1	@5	@10		@1	@5	@10		@1	@5	@10	
HMEA	0.127	–	0.369	–	0.262	–	0.581	–	0.417	–	0.786	–
MMEA	0.265	0.451	0.541	0.357	0.416	0.621	0.703	0.512	0.590	0.804	0.868	0.685
EVA	0.199	0.356	0.448	0.283	0.334	0.522	0.589	0.422	0.484	0.625	0.696	0.563
MCLEA	0.295	0.582	0.637	0.393	0.573	0.742	0.800	0.652	0.730	0.857	0.883	0.784
MSNEA	0.114	0.296	0.388	0.175	0.288	0.499	0.590	0.388	0.518	0.744	0.779	0.613
ACK-MMEA	0.304	0.455	0.549	0.387	0.560	0.683	0.736	0.624	0.682	0.816	0.874	0.752
Meaformer	0.431	0.629	0.721	0.520	0.621	0.789	0.846	0.704	0.771	0.887	0.918	0.823
GEEA	0.343	0.565	0.661	0.450	0.651	0.788	0.852	0.723	0.787	0.886	0.918	0.833
Confidence-MMEA	0.280	0.478	0.564	0.376	0.498	0.695	0.764	0.590	0.687	0.853	0.895	0.764
TriFac	0.318	–	0.559	0.389	0.554	–	0.750	0.607	0.697	–	0.882	0.761
OURS	**0.501**	**0.701**	**0.771**	**0.602**	**0.695**	**0.837**	**0.879**	**0.776**	**0.809**	**0.912**	**0.932**	**0.847**

Table 3: Performance of various alignment seeds in the non-iterative setting of FBYG15K. Best values among the baseline models are underlined, the best results are highlighted in bold

Methods	FBYG15K (20%)				FBYG15K (50%)				FBYG15K (80%)			
	Hits			MRR	Hits			MRR	Hits			MRR
	@1	@5	@10		@1	@5	@10		@1	@5	@10	
HMEA	0.105	–	0.369	–	0.265	–	0.581	–	0.433	–	0.801	–
MMEA	0.234	0.398	0.480	0.317	0.403	0.572	0.645	0.486	0.597	0.785	0.839	0.682
EVA	0.153	0.292	0.361	0.224	0.311	0.488	0.534	0.388	0.491	0.656	0.692	0.565
MCLEA(2022)	0.254	0.401	0.484	0.332	0.543	0.702	0.759	0.616	0.653	0.794	0.835	0.715
MSNEA(2022)	0.103	0.200	0.249	0.153	0.320	0.534	0.589	0.413	0.531	0.733	0.778	0.620
ACK-MMEA	0.289	0.413	0.496	0.360	0.535	0.661	0.699	0.593	0.676	0.829	0.864	0.744
Meaformer	0.325	0.517	0.596	0.416	0.560	0.731	0.780	0.640	0.705	0.840	0.874	0.768
GEEA	0.298	0.506	0.585	0.393	0.589	0.745	0.794	0.668	0.731	0.849	0.891	0.777
Confidence-MMEA	0.261	0.422	0.503	0.343	0.485	0.672	0.738	0.572	0.672	0.825	0.871	0.745
TriFac	0.290	–	0.508	0.371	0.546	–	0.694	0.579	0.669	–	0.865	0.736
OURS	**0.416**	**0.596**	**0.672**	**0.493**	**0.639**	**0.795**	**0.832**	**0.721**	**0.769**	**0.878**	**0.919**	**0.803**

Table 4: Performance of various alignment seeds under FBDB15K, FBYG15K iterative settings. Best values among the baseline models are underlined, the best results are highlighted in bold

Seed	Methods	FBDB15K				FBYG15K			
		Hits@1	Hits@5	Hits@10	MRR	Hits@1	Hits@5	Hits@10	MRR
20%	EVA	0.231	0.422	0.488	0.318	0.188	0.312	0.403	0.260
	MCLEA(2022)	0.395	0.609	0.656	0.487	0.322	0.485	0.546	0.400
	MutiJAF	0.480	0.576	0.601	0.523	0.463	0.658	0.731	0.554
	MSNEA	0.149	0.330	0.392	0.232	0.149	0.330	0.392	0.232

(Continued)

Table 4 (continued)

Seed	Methods	FBDB15K				FBYG15K			
		Hits@1	Hits@5	Hits@10	MRR	Hits@1	Hits@5	Hits@10	MRR
	MEAIE	0.440	0.576	0.601	0.523	0.401	0.545	0.646	0.489
	DFMKE	0.338	–	0.655	0.550	0.368	–	0.643	0.510
	MEAformer	0.578	0.760	0.812	0.661	0.444	0.631	0.692	0.529
	UQMIEA	0.447	0.663	0.741	0.548	0.399	0.605	0.689	0.497
	OURS	**0.621**	**0.802**	**0.856**	**0.706**	**0.563**	**0.723**	**0.763**	**0.631**
50%	EVA	0.364	0.554	0.606	0.449	0.325	0.519	0.560	0.404
	MCLEA	0.620	0.786	0.832	0.696	0.563	0.690	0.751	0.631
	MSNEA	0.358	0.607	0.656	0.459	0.376	0.591	0.646	0.472
	MEAformer	0.690	0.834	0.871	0.755	0.612	0.766	0.808	0.682
	OURS	**0.756**	**0.887**	**0.891**	**0.801**	**0.658**	**0.819**	**0.848**	**0.709**
80%	EVA	0.491	0.661	0.711	0.573	0.493	0.654	0.695	0.572
	MCLEA	0.741	0.876	0.900	0.802	0.681	0.801	0.837	0.737
	MSNEA	0.565	0.772	0.810	0.651	0.593	0.768	0.806	0.668
	MEAformer	0.784	0.897	0.921	0.834	0.724	0.846	0.880	0.783
	OURS	**0.823**	**0.928**	**0.946**	**0.868**	**0.766**	**0.881**	**0.905**	**0.801**

In the iterative experiments, we observed a significant improvement in results compared to the non-iterative approach. This enhancement can mainly be attributed to our dynamic expansion of the training set based on the original optimization. It enables the model to learn the alignment relationships between entities better. Observing Table 4, we found that on 20% aligned seed data, compared with MutiJAF, our model's Hits@1 and MRR increased by 14.1% and 18.3%, respectively, on FBDB15K, and Hits@1 and MRR increased by 10% and 7.7%, respectively, on FBYG15K. This is because MutiJAF only considers late fusion, does not involve the correlation of early primary feature level modalities, and ignores the modal heterogeneity brought about by visual and graph structures. Although some models, such as MEAIE, MutiJAF, etc., also consider numerical modalities, none consider the deeper integration of inter-modal information and fail to utilize each modality's information fully). In contrast, while addressing inter-modal heterogeneity, our model also introduces a deep perceptual reinforcement learning module based on policy networks, which further realizes multi-level deep information interaction between modalities. Compared to MEAIE, our model achieves improvements of 18.1% and 18.3% in Hits@1, as well as 16.2% and 14.2% in MRR across the two datasets, thereby demonstrating the practical effectiveness and rationale of the module.

4.2.2 Ablation Study

We performed the ablation experiment in two ways: 1) the impact of different modalities on entity alignment, and 2) the importance of key components in entity alignment.

1. Modal contributions validation: Five relevant modal variants were designed, including w/o structure, w/o relation, w/o attr, w/o img, and w/o number. The specific results are depicted in Table 5. We find that removing any modality results in a decrease in performance. Notably, eliminating the structural modality leads to a substantial decline in Hits@1 and MRR, with reductions of approximately 40% across both datasets. This sharp decline is due to the fact that the graph structural modality conveys contextual information about entities through their topological relationships. Removing the modality will result in the inability to

utilize the relationship information between entities, thus losing the understanding of complex structures and higher-order dependencies. Secondly, removing the relationships, the image modality performance degradation is also more significant. Compared to FBDB15K, the relational modality of FBYG15K is more significant than the image modality. This distinction may arise because, while both datasets exhibit complex relational structures, FBDB15K boasts over 90% image coverage. In contrast, only 81.17% of the entities in FBYG15K possess image features, indicating that a considerable number of entities lack visual information support. In this case, the model cannot fully leverage the image modality to mitigate the deficiencies of other modalities, resulting in its relatively weakened role in multi-modal fusion. As a result, relational modalities become an essential source used in addition to structural modalities to fill information gaps and provide key contextual and semantic associations.

Table 5: Validation of the contribution of each modality (aligned seed ratio of 20%)

Methods	FBDB15K				FBYG15K			
	Hits@1	Hits@5	Hits@10	MRR	Hits@1	Hits@5	Hits@10	MRR
w/o MMPF-img	0.408	0.623	0.674	0.513	0.333	0.505	0.599	0.413
w/o MMPF-gph	0.435	0.648	0.695	0.538	0.341	0.529	0.605	0.430
w/o MMPF	0.341	0.556	0.566	0.425	0.285	0.433	0.513	0.356
w/o OA	0.487	0.707	0.732	0.591	0.406	0.591	0.667	0.486
w/o OC	0.484	0.696	0.730	0.588	0.404	0.578	0.665	0.501
w/o CPLR	0.406	0.619	0.662	0.509	0.338	0.475	0.581	0.391
repl.GAT	0.485	0.688	0.735	0.596	0.408	0.588	0.657	0.478
MPSEA	**0.501**	**0.701**	**0.771**	**0.602**	**0.416**	**0.596**	**0.672**	**0.493**

2. Contribution verification of each component. We designed the following three variants for experimentation: **1) w/o MMPF-img:** Remove the image pre-synergistic fusion from the multi-modal pre -synergistic fusion module (MMPF); **w/o MMPF-gph:** Remove the graph structure pre-synergistic fusion from MMPF; **w/o MMPF:** Completely remove the multi-modal pre-synergistic fusion module. **2) w/o OA:** Remove the OA module from the FFN layer of the image transformer; w/o OC: Remove both the OA and convolutional dual modules from the FFN layer of the transformer. **3) w/o CPRL:** Remove the Cross-modal Deep Perception Reinforcement Learning (CPRL) module. **4) repl.GAT:** Replacing multi-weight GAT with Normal GAT. The results are illustrated in Table 6.

Table 6: Validation of the contribution of each component (alignment seeding at 20%)

Methods	FBDB15K				FBYG15K			
	Hits@1	Hits@5	Hits@10	MRR	Hits@1	Hits@5	Hits@10	MRR
w/o structure	0.078	0.163	0.238	0.101	0.057	0.105	0.146	0.074
w/o relation	0.467	0.654	0.737	0.558	0.348	0.541	0.622	0.430
w/o attr	0.471	0.669	0.761	0.575	0.394	0.567	0.644	0.476
w/o img	0.432	0.601	0.655	0.498	0.369	0.546	0.622	0.430
w/o number	0.491	0.692	0.764	0.589	0.389	0.575	0.643	0.483
MPSEA	**0.501**	**0.701**	**0.771**	**0.602**	**0.416**	**0.596**	**0.672**	**0.493**

We have the following key observations: **(a)** For the first set of variants, we find that the MMPF module effectively addresses the problem of modal heterogeneity differences during fusion. Notably, removing image pre-synergistic fusion has a more significant impact than removing graph structure pre-synergistic fusion. This is presumably due to the greater heterogeneity between image modality and text, compared to graph structure modality. The introduction of image pre-synergistic fusion facilitates the integration of heterogeneous features into other modalities in advance, thereby enabling more effective multi-modal fusion. **(b)** For the second group of variants, we observed that introducing the OA module and the convolution enhancement module in the FFN layer can further mitigate the noise interference between different modalities and enhance the model's ability to handle fine-grained features. **(c)** The removal of the CPRL module has resulted in a notable decline in model performance. It indicates that our custom cross-modal mutual information estimator can do a better job of helping the model learn information fusion and sharing, and guiding the policy network renewal through a reward and punishment mechanism to optimize the alignment policy progressively. **(d)** A slight reduction in the model is observed when the multi-weighted GAT is substituted for the regular GAT, demonstrating that the former mitigates the contextual gap issue and enhances the capacity to differentiate between various types of nodes.

4.2.3 Hyperparametric Analysis

We further investigated the influence of hyperparameters on model performance. The experimental results are illustrated in Figs. 4 and 5 (Both MRR and Hit@1 are dimensionless evaluation metrics without physical units, with values ranging from 0 to 1). Our primary focus was on the impact of the threshold θ of the reward function and the mutual information weight α, as defined in the CPRL module. The results presented in Fig. 4 indicate that as θ increases, model performance progressively improves, reaching a maximum value around 0.63, after which performance slightly declines with further increases in θ. The possible reason for this is that as θ increases, the model must attain a higher similarity in order to be rewarded, thus incentivizing the model to improve the embedding quality further. However, when θ exceeds a certain threshold (0.63), too high a threshold leads to an increase in the penalization ratio, suppressing meaningful matching pairs and making model learning difficult, thus leading to a decline in overall performance.

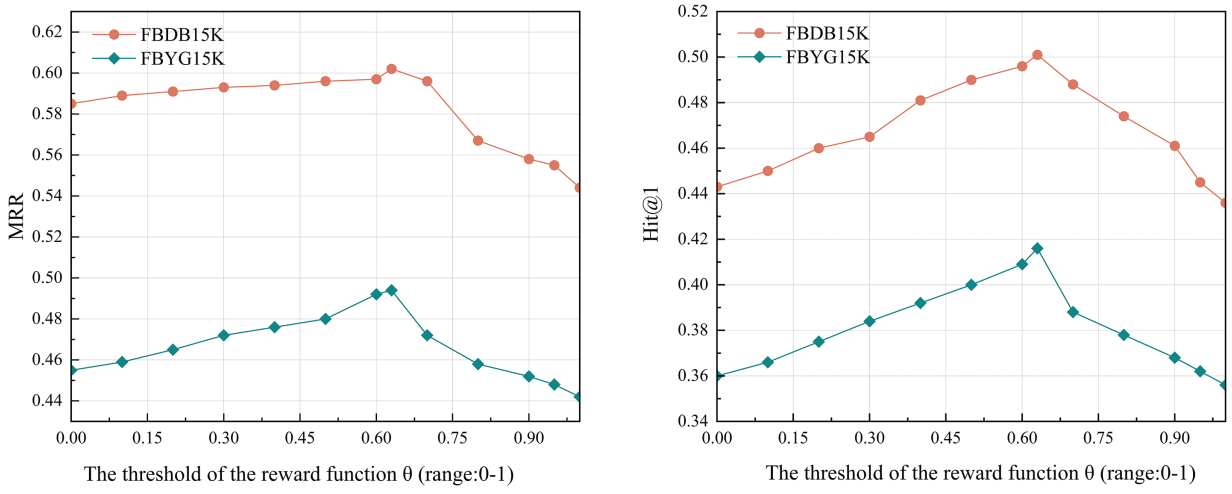

Figure 4: Impact of threshold θ on performance (non-iterative strategy, 20% alignment seed ratio)

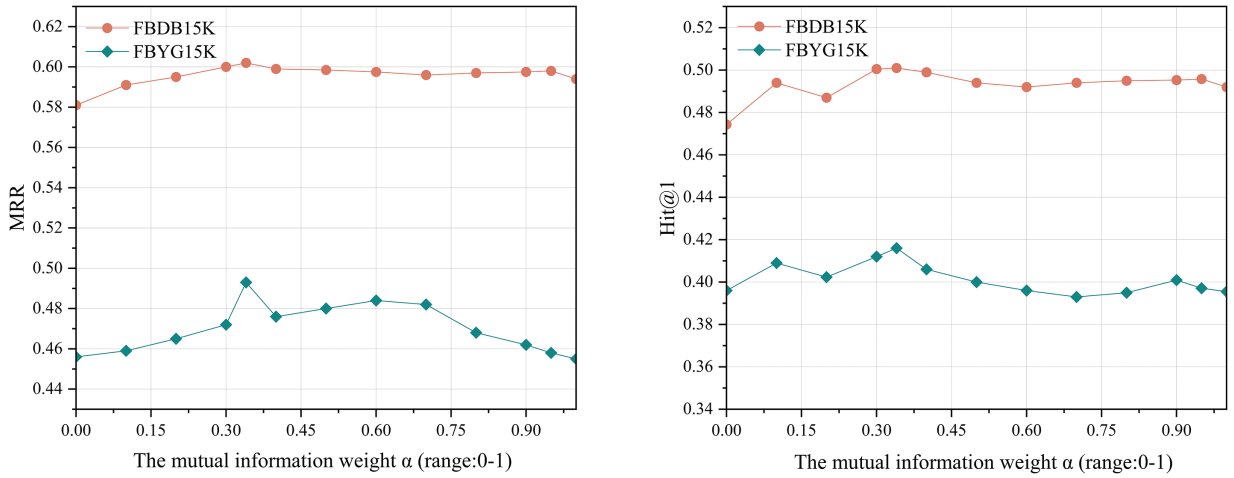

Figure 5: Impact of mutual information weights α (non-iterative strategy, 20% alignment seed ratio)

The results presented in Fig. 5 indicate that for both datasets, performance is maximized when α is approximately 0.34. α is applied to balance the contributions of similarity and cross-modal information sharing. When α is high, the model places more emphasis on inter-modal information sharing. The optimal balance appears to be achieved at α = 0.34. It is worth noting that the performance of α = 0.1 is higher than 0.2 for FBYG15K. When α = 0.1, the model may rely more on single-modal features, avoiding some of the cross-modal interference. As α increases slightly to 0.2, the contribution of mutual information is insufficient and instead introduces a degree of noise, resulting in a minor reduction in performance.

4.2.4 Low-Resource Training Data Performance

To comprehensively evaluate the performance of MPSEA in MMEA tasks, we conducted an in-depth analysis using limited training data. By setting the pre-alignment seed ratio to 0.01–0.3 and selecting MEAformer and MCLEA as baseline models, we performed experiments on the FB15K-YG15K dataset, with results illustrated in the reinforcement learning flowchart as follows in Fig. 6. It is observed that MPSEA is at the forefront even at low seed ratios, further confirming that our model is practically effective.

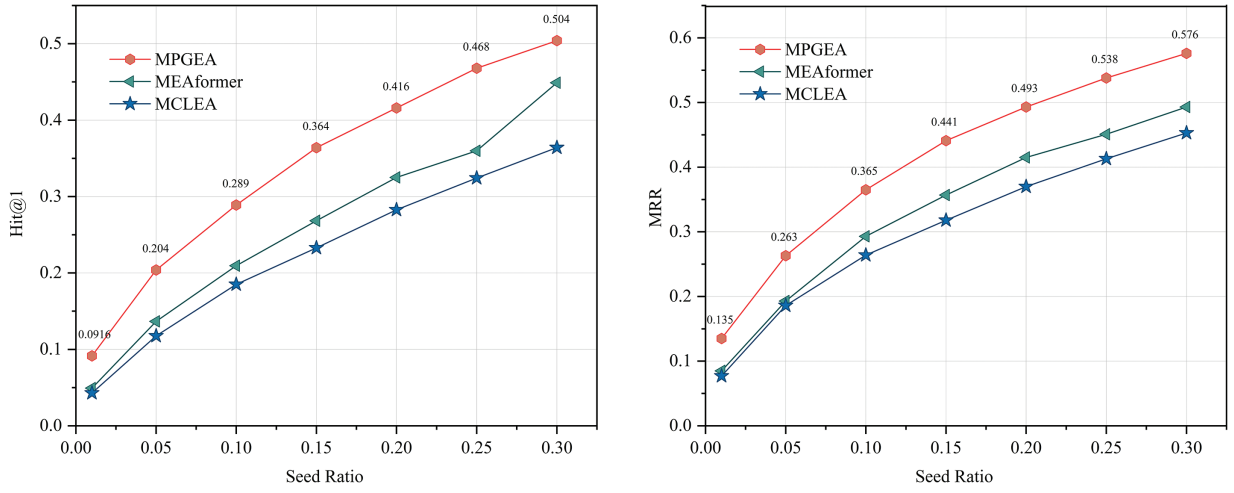

Figure 6: H@1 and MRR performance with different ratios of seed alignments ranging from 0.01 to 0.3 on FBYG15K

5 Conclusion

This paper proposes a Multi-modal Pre-synergistic Fusion Model based on Cross-modal Mutual Information Strategy Optimization (MPSEA) for the entity alignment task. The model effectively encodes multi-modal information—including numerical, graph-structural, and visual modalities—alleviating the issue of missing modality information in multi-modal entity alignment. By injecting graph structure and visual modality features as preparatory information into the cross-modal attention mechanism, MPSEA achieves multi-level, deep joint feature embeddings, which reduces information loss and bias arising from modality heterogeneity. Furthermore, a custom-designed strategy optimization network based on cross-modal mutual information is adopted to capture shared information and optimize model output via contrastive learning. The cosine similarity of embeddings ultimately determines entity alignment. Experimental results on two real-world datasets demonstrate that MPSEA achieves superior performance over existing approaches.

5.1 Discussion on Model Limitations

Although MPSEA achieves superior performance in the multimodal entity alignment task, its design and experimental results also reveal several limitations:

- **Strong Dependence on Specific Modalities:** Ablation experiments show that removing any modality leads to performance degradation. Notably, removing the graph structure modality causes a significant drop in Hits1 and MRR metrics across both datasets. This indicates a high reliance on graph structural information, fundamentally because the core of knowledge graphs lies in conveying rich contextual information through the topological relationships between entities. The structural modality captures complex connections and higher-order dependencies among entities; removing it results in the loss of understanding of these complex structures. Without the structural modality, the model cannot effectively utilize the aggregated relationships of neighboring nodes for alignment reasoning. Suppose the knowledge graph lacks high-quality structural information. In that case, the fusion of other modality features will lack a solid anchor and context, greatly reducing the efficiency and accuracy of information fusion.
- **Hyperparameter Sensitivity:** The model's performance is susceptible to hyperparameters in the reward function. For example, when the reward threshold θ exceeds 0.63, an excessively high threshold increases the penalty ratio, suppressing meaningful matches and making model training difficult, leading to overall performance decline. Similarly, if the mutual information weight α is improperly set, such as too low (e.g., 0.2), noise may be introduced, causing slight performance degradation. This indicates that careful tuning of these parameters is necessary in practical applications.
- **Computational Cost Considerations:** Core mechanisms of MPSEA, such as the multi-head self-attention (MHA) blocks, complex fusion strategies, and deep neural network components, are computationally intensive. The computational complexity of MHA grows significantly with the length of the input sequence (i.e., number of entities) and the number of modalities. This may pose challenges of higher computational overhead when handling large-scale multimodal knowledge graphs compared to simpler embedding-based alignment methods.

5.2 Future Research Directions

In future research, we plan to:

- The current study evaluates the model only on the FB15K-DB15K and FB15K-YAGO15K datasets, a relatively narrow scope for verifying the model's general effectiveness. Future work should consider

incorporating diverse datasets from different domains with varying scales and modal characteristics to validate the model's generalization ability. This will help ensure the model remains robust across a wider range of real-world scenarios.

- Explore embedding methods for richer modalities (such as audio and video) to enhance the model's applicability and accuracy further. Currently, most multimodal entity alignment (MMEA) datasets treat multimodal data as attributes of textual entities, often neglecting the intrinsic correlations between modalities, which differs from real-world knowledge graph entities that include multiple modalities such as text, images, and audio. For example, recent advances in Alzheimer's disease (AD) diagnosis demonstrate how AI processes complex, multi-level information to achieve high accuracy in disease detection and staging using a multi-stage convolutional neural network (CNN) framework on MRI data. Additionally, research on medical image denoising based on Transformer architectures shows that advanced Transformer models can effectively handle complex noise patterns in multimodal data by capturing long-range pixel dependencies through multi-head attention mechanisms and performing image reconstruction. This aligns with the OutlookerAttention mechanism in our model, which is designed to reduce cross-modal noise interference. Understanding how medical AI effectively processes diverse data types—such as MRI sequences, PET scans, and clinical data—while providing interpretable insights into fusion strategies is crucial for developing robust and trustworthy multimodal systems. These experiences offer valuable guidance for our multimodal pre-synergistic fusion approach, helping us better model the complex relationships between heterogeneous modalities to achieve more comprehensive information fusion.
- Inspired by the dual-perspective evaluation framework proposed by Wang [36], integrating dynamically learnable evaluation metrics helps better balance structural fidelity and semantic coherence in generated text, enhancing interpretability. This approach addresses the demand for explainable AI (XAI) in medical applications, where XAI aids in understanding model decisions, identifying biases, and optimizing design, especially in high-risk scenarios like Alzheimer's diagnosis. We aim to enable entity alignment model evaluation to go beyond performance metrics and reveal the reasoning process in handling heterogeneous multimodal information, thereby improving model transparency and reliability.

These efforts, we believe, will further improve the performance and generalizability of multimodal entity alignment models in complex knowledge graph scenarios.

Acknowledgement: We sincerely acknowledge the financial support from the National Natural Science Foundation of China and the Natural Science Foundation of Shandong Province.

Funding Statement: This work is partially supported by the National Natural Science Foundation of China under Grants 62471493 and 62402257 (for conceptualization and investigation), partially supported by the Natural Science Foundation of Shandong Province, China under Grants ZR2023LZH017, ZR2024MF066, and 2023QF025 (for formal analysis and validation), partially supported by the Open Foundation of Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Qilu University of Technology (Shandong Academy of Sciences) under Grant 2023ZD010 (for methodology and model design), and partially supported by the Russian Science Foundation (RSF) Project under Grant 22-71-10095-P (for validation and results verification).

Author Contributions: Conceptualization: Huayu Li, Xinxin Chen, Lizhuang Tan; Methodology: Huayu Li, Xinxin Chen; Validation: Huayu Li, Xinxin Chen, Konstantin I. Kostromitin, Peiying Zhang; Formal analysis: Huayu Li, Xinxin Chen, Athanasios V. Vasilakos, Peiying Zhang; Investigation: Huayu Li, Xinxin Chen; Writing—original draft: Huayu Li, Xinxin Chen, Lizhuang Tan, Peiying Zhang; Writing—review and editing: Huayu Li, Xinxin Chen, Lizhuang Tan; Supervision: Lizhuang Tan, Konstantin I. Kostromitin, Athanasios V. Vasilakos, Peiying Zhang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the results of this study are openly available at https://github.com/lzxlin/MCLEA (accessed on 01 September 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Cao Y, Li X, Liu H, Dai W, Chen S, Wang B, et al. Pay more attention to relation exploration for knowledge base question answering. In: Findings of the Association for Computational Linguistics: ACL 2023; 2023 Jul 9–14; Toronto, ON, Canada: Association for Computational Linguistics. p. 2119–36.
2. Cao X, Shi Y, Wang J, Chen E. Cross-modal knowledge graph contrastive learning for machine learning method recommendation. In: Proceedings of the 30th ACM International Conference on Multimedia (MM '22); 2022 Oct 10–14; Lisbon, Portugal: Association for Computing Machinery. p. 3694–702.
3. Cao Y, Wang X, He X, Hu Z, Chua TS. Unifying knowledge graph learning and recommendation: towards a better understanding of user preferences. In: Proceedings of the 2019 World Wide Web Conference (WWW '19); 2019 May 13–17; San Francisco, CA, USA: Association for Computing Machinery. p. 151–61.
4. Chen X, Zhang N, Li L, Yao Y, Deng S, Tan C, et al. Hybrid Transformer with multi-level fusion for multi-modal knowledge graph completion. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22); 2022 Jun 11–12; Madrid, Spain: ACM. p. 904–15.
5. Han X, Liu Z, Sun M. Neural knowledge acquisition via mutual attention between knowledge graph and text. In: Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press; 2018. p. 32–9.
6. Sun Z, Zhang Q, Hu W, Wang C, Chen M, Akrami F, et al. A benchmarking study of embedding-based entity alignment for knowledge graphs. Proc VLDB Endow. 2020;13(12):2326–40. doi:10.14778/3407790.3407828.
7. Zhang J, Yin Z, Chen P, Nichele S. Emotion recognition using multi-modal data and machine learning techniques: a tutorial and review. Inf Fusion. 2020;59(1):103–26. doi:10.1016/j.inffus.2020.01.011.
8. Jiang S, Sun B, Wang L, Bai Y, Li K, Fu Y. Skeleton aware multi-modal sign language recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 19–25; Nashville, TN, USA. p. 3413–23.
9. Zhen L, Hu P, Wang X, Peng D. Deep supervised cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA, USA. p. 10394–403.
10. Chun S, Oh SJ, De Rezende RS, Kalantidis Y, Larlus D. Probabilistic embeddings for cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 8415–24.
11. Sun Z, Hu W, Zhang Q, Qu Y. Bootstrapping Entity Alignment with Knowledge Graph Embedding. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18); 2018 Jul 13–19; Stockholm, Sweden. p. 4396–402.
12. Sanchez-Lengeling B, Reif E, Pearce A, Wiltschko AB. A gentle introduction to graph neural networks. Distill. 2021;6(9):e33.
13. Wang Z, Lv Q, Lan X, Zhang Y. Cross-lingual knowledge graph alignment via graph convolutional networks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018); 2018 Oct 31–Nov 4; Brussels, Belgium: Association for Computational Linguistics. p. 349–57.
14. Wu Y, Liu X, Feng Y, Wang Z, Yan R, Zhao D. Relation-aware entity alignment for heterogeneous knowledge Graphs. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19); 2019 Aug 10–16; Macao, China. p. 5278–84.
15. Gao Y, Liu X, Wu J, Zhou A, Li T. ClusterEA: scalable entity alignment with stochastic training and normalized mini-batch similarities. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22); 2022 Aug 14–18; Washington, DC, USA: ACM. p. 421–31.

16. Liu F, Chen M, Roth D, Collier N. Visual pivoting for (Unsupervised) entity alignment. Proc AAAI Conf Artif Intell. 2021;35(5):4257–66. doi:10.1609/aaai.v35i5.16550.
17. Lin Z, Zhang Z, Wang M, Shi Y, Wu X, Zheng Y. Multi-modal contrastive representation learning for entity alignment. In: Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022); 2022 Oct 12–17; Gyeongju, Republic of Korea: International Committee on Computational Linguistics. p. 2572–84.
18. Chen Z, Chen J, Zhang W, Guo L, Fang Y, Huang Y, et al. MEAformer: multi-modal entity alignment transformer for meta modality hybrid. In: Proceedings of the 31st ACM International Conference on Multimedia (MM '23). Association for Computing Machinery; 2023 Oct 29–Nov 3; Ottawa, ON, Canada: Association for Computing Machinery. p. 3317–27.
19. Yuan S, Lu Z, Li Q, Gu J. A Multi-modal entity alignment method with inter-modal enhancement. Big Data Cogn Comput. 2023;7(2):77. doi:10.3390/bdcc7020077.
20. Li Q, Guo S, Luo Y, Ji C, Wang L, Sheng J, et al. Attribute-consistent Knowledge Graph Representation Learning for Multi-modal Entity Alignment. In: Proceedings of the ACM Web Conference 2023 (WWW '23); 2023 Apr 30–May 4; Austin, TX, USA: ACM. p. 2499–508.
21. Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv:1710.10903. 2017.
22. Yang H-W, Zou Y, Shi P, Lu W, Lin J, Sun X. 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); 2019 Nov; Hong Kong, China. Vol. 1. p. 4430–40.
23. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014.
24. Liu Y, Li H, Garcia-Duran A, Niepert M, Onoro-Rubio D, Rosenblum DS. Multi-modal knowledge graphs. In: Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019). Portorož, Slovenia: Springer; 2019. p. 459–74.
25. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS); 2017. p. 5998–6008.
26. Zeng W, Zhao X, Tang J, Wang X, Zhang W, Xiong H, et al. Reinforcement learning-based collective entity alignment with adaptive features. ACM Trans Inf Syst. 2021;39(3):1–31. doi:10.1145/3446428.
27. Guo H, Tang J, Zeng W, Zhao X, Liu L. Multi-modal entity alignment in hyperbolic Space. Neurocomputing. 2021;461(4):598–607. doi:10.1016/j.neucom.2021.03.132.
28. Chen L, Li Z, Wang Y, Xu T, Wang Z, Chen E. MMEA: entity alignment for multi-modal knowledge graph. In: Proceedings of the 13th International Conference on Knowledge Science, Engineering and Management (KSEM 2020); 2020 Aug 18–20; Hangzhou, China: Springer. p. 134–47.
29. Guo L, Chen Z, Chen J, Fang Y, Zhang W. Revisit and outstrip entity alignment: a perspective of generative models. In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024). Vienna, Austria; 2024.
30. Zhang X, Chen T, Wang H. A novel method for boosting knowledge representation learning in entity alignment through triple confidence. Mathematics. 2024;12(8):1214. doi:10.3390/math12081214.
31. Li Q, Li J, Wu J, Peng X, Ji C, Peng H, et al. Triplet-aware graph neural networks for factorized multi-modal knowledge graph entity alignment. Neural Netw. 2024;171(1):106479. doi:10.1016/j.neunet.2024.106479.
32. Cheng B, Zhu J, Guo M. MultiJAF: multi-modal joint entity alignment framework for multi-modal knowledge graph. Neurocomputing. 2022;500(4):581–91. doi:10.1016/j.neucom.2022.05.058.
33. Chen L, Li Z, Xu T, Wu H, Wang Z, Yuan NJ, et al. Multi-modal siamese network for entity alignment. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22); 2022 Aug 14–18; Washington, DC, USA: ACM. p. 118–26.
34. Zhu J, Huang C, De Meo P. DFMKE: a dual fusion multi-modal knowledge graph embedding framework for entity alignment. Inform Fus. 2023;90(1):111–9. doi:10.1016/j.inffus.2022.09.012.

35. Hama K, Matsubara T. Multi-modal entity alignment using uncertainty quantification for modality importance. IEEE Access. 2023;11:28479–89. doi:10.1109/access.2023.3259987.
36. Wang H, Wang L, Lepage Y. Dual-perspective evaluation of knowledge graphs for graph-to-text generation. Comput Mater Contin. 2025;84:305–24.