SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

July 2026, Vol. 69, Iss. 7, 172105:1–172105:18
https://doi.org/10.1007/s11432-025-4721-2

Noise-augmented multi-modal entity alignment with confidence-based dynamic fusion

Xiangyu LUO[2], Yan ZHANG[1,2,3,4*], Miao ZHANG[1,3,4], Kui XIAO[1,3,4], Wenxing HUANG[1,3,4] & Zhifei LI[1,2,3,4*]

[1]*School of Computer Science, Hubei University, Wuhan 430062, China*
[2]*School of Cyber Science and Technology, Hubei University, Wuhan 430062, China*
[3]*Hubei Key Laboratory of Big Data Intelligent Analysis and Application (Hubei University), Wuhan 430062, China*
[4]*Key Laboratory of Intelligent Sensing System and Security (Hubei University), Ministry of Education, Wuhan 430062, China*

Received 26 February 2025/Revised 11 September 2025/Accepted 16 December 2025/Published online 16 June 2026

Abstract Multi-modal entity alignment seeks to match equivalent entities across different multi-modal knowledge graphs, which integrate heterogeneous multi-modal data such as images and text to enrich entity semantics. However, variations in multi-modal data quality and their inherent unreliability present significant challenges that can negatively impact alignment results. Consequently, we propose NoCo, a noise-augmented multi-modal entity alignment method with confidence-based dynamic fusion. NoCo incorporates a modality-aware noise enhancement mechanism that adaptively injects Gaussian noise into each modality, thereby improving robustness and preventing overfitting to unreliable features. Simultaneously, a confidence-based dynamic fusion framework is designed to automatically calibrate modality contributions according to data quality, effectively down-weighting noisy inputs while amplifying reliable signals. Experimental evaluations demonstrate that NoCo effectively overcomes these challenges and achieves strong performance. Compared with the state-of-the-art method, NoCo achieves a 2.8% maximum improvement on the Multi-OpenEA datasets, 4.1% maximum improvement on FB15K-DB15K, and 5.0% on FB15K-YG15K. The code of the proposed model is stored at https://github.com/HubuKG/NoCo.

Keywords multi-modal knowledge graphs, multi-modal entity alignment, dynamic fusion

Citation Luo X Y, Zhang Y, Zhang M, et al. Noise-augmented multi-modal entity alignment with confidence-based dynamic fusion. Sci China Inf Sci, 2026, 69(7): 172105, https://doi.org/10.1007/s11432-025-4721-2

1 Introduction

Knowledge graph (KG) is a structured representation method that utilizes a graph format to depict relations between entities, facilitating comprehension of deep-level relations and inherent meanings among information [1–6]. Owing to variations in construction approaches and technologies, different KGs may adopt various perspectives to describe the same real-world object [7]. Extensive research concentrates on using knowledge fusion methods to improve the completeness of KGs, thereby creating unified, comprehensive, and precise representations of entities by integrating different KGs. As a key technology in the field of knowledge fusion, entity alignment (EA) aims to match equivalent entities across different KGs, establishing alignment relationships among entities representing the same real-world object to consolidate and merge the information that they contain [8, 9].

Multi-modal knowledge graphs (MMKGs) enhance traditional static KGs by incorporating data from multiple modalities, such as images and text, each contributing distinct perspectives and enriched semantic information on entities [10–12]. The rapid development of MMKGs has enabled diverse downstream applications, including information retrieval, visual question answering [13–15], recommendation systems [16, 17], and applications across various domains [18–20]. Multi-modal entity alignment (MMEA) extends conventional entity alignment approaches by leveraging multi-modal data to identify corresponding entity pairs across different MMKGs [21, 22].

However, the complexity of multi-modal data presents significant challenges that impact the accuracy of MMEA. Firstly, data from different modalities often vary in quality and reliability due to inconsistencies in data sources and the inherent differences across modalities. Variations in data collection methods, times, and sources across MMKGs lead to discrepancies in coverage, detail, and accuracy within each modality. Consequently, a modality may offer high-quality alignment information in some instances while introducing misleading information in others. Furthermore, the inconsistency in multi-modal data quality and modality-specific characteristics introduces

* Corresponding author (email: zhangyan@hubu.edu.cn, zhifei1993@hubu.edu.cn)

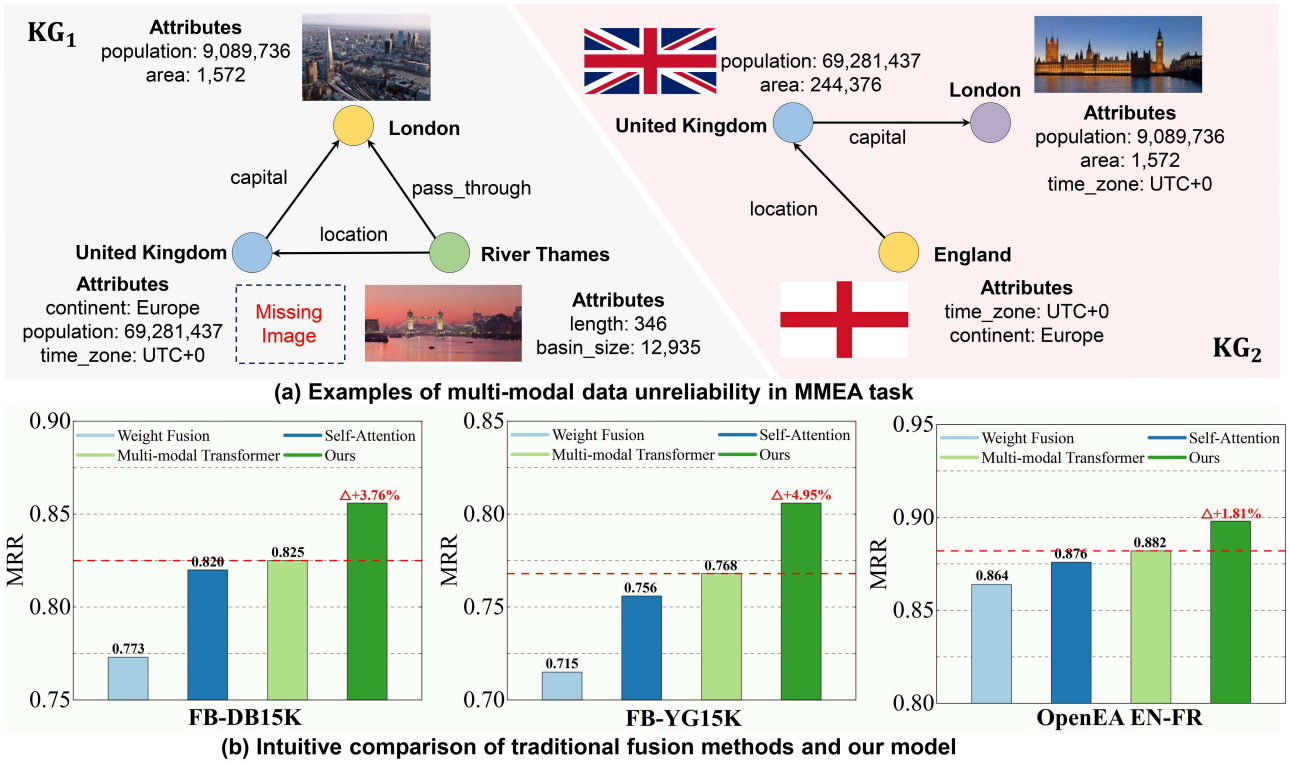

Figure 1 (Color online) Illustration of multi-modal data unreliability. Proposed NoCo compared with other fusion methods. Experiments on three datasets validate the superiority of NoCo.

irrelevant features, generating noise that disrupts the alignment process. As illustrated in Figure 1(a), the image representing the entity "*United Kingdom*" in KG1 is absent, while the visual modality is fully presented in KG2. Additionally, both KGs exhibit inconsistencies in the attribute modality. Such disparities in multi-modal information can lead to misalignment, potentially causing the "*United Kingdom*" entity in KG1 to be incorrectly matched with the "*England*" entity in KG2.

To address these challenges, we propose NoCo, a novel noise augmented multi-modal entity alignment framework with confidence-based dynamic fusion. NoCo introduces a modality-aware noise enhancement mechanism that adaptively scales Gaussian noise according to the representational characteristics of each modality. It diversifies modality spaces, improves robustness, and enables more effective performance in real-world tasks. In contrast to prior MMEA models, NoCo further incorporates a confidence-based dynamic fusion framework tailored for EA. This framework automatically adjusts modality weights during the alignment process based on data quality and uncertainty, thereby capturing inter-modality interactions more reliably. The model mitigates the adverse effects of low-quality modalities by integrating uni- and global-confidence of modalities. In addition, a relative calibration strategy constrains the generalization error bound, ensuring robustness even when multi-modal data are incomplete or noisy. As illustrated in Figure 1(b), experiments across multiple datasets validate the superiority of the proposed approach over existing fusion methods. Overall, the main contributions of this paper are summarized as follows.

- We propose NoCo, a novel fusion framework for MMEA, which employs a pre-trained encoder to process multi-modal information and integrates a modality-aware noise enhancement mechanism. This mechanism leverages the representational capacity of each modality while enhancing robustness against noisy or unreliable features.
- We design a confidence-based dynamic fusion strategy that adaptively calibrates the contribution of each modality based on its quality and uncertainty. This strategy not only mitigates the adverse effects of low-quality modalities but also achieves more balanced and reliable integration across modalities.
- We conduct extensive experiments on three separate datasets to validate the superiority of the proposed model. For the Hits@1 metric, the proposed model achieves a 1.43% improvement over the baseline on the Multi-OpenEA datasets, 3% improvement on FB15K-DB15K, and 2.18% improvement on FB15K-YG15K.

The subsequent sections are organized as follows: Section 2 presents a comprehensive discussion of the relevant literature; Section 3 elucidates the proposed model; Section 4 presents experimental results and corresponding analyses; and finally, Section 5 provides the concluding remarks.

2 Related work

This section provides a brief review of uni-modal entity alignment models and multi-modal entity alignment models.

2.1 Uni-modal entity alignment

Traditional uni-modal EA solely utilizes structural information and entity relations to identify equivalent entities in diverse KGs [23–25]. This approach typically adopts embedding techniques, matching entities by generating entity embeddings and calculating vector similarities.

Translation-based and graph neural network (GNN)-based models enhance KG alignment by refining the representation of entities and relations. Translation-based models, such as TransE [1], conceptualize relationships as mappings that transform entity embeddings. Enhancements such as TransH [2] incorporate relation-specific hyperplanes to reduce ambiguity, whereas MTransE [26] introduces cross-graph transition matrices to capture inter-graph relationships. BootEA [27] integrates traditional graph-matching techniques with deep learning to iteratively refine alignments. By contrast, GNN-based methods utilize graph structures to embed entities enriched with local information [28, 29]. GCN-Align [30] applies graph convolutions for entity representation, GMNN [31] leverages topic-entity graphs to capture context, RDGCN [32] employs a dual-graph approach to better model relational structures, and MuGNN [33] uses multi-channel networks to reinforce cross-graph alignments.

Uni-modal EA methods effectively perform alignment tasks using embedding techniques. However, they do not incorporate multi-modal information, such as images and text, to enrich the representation of entities.

2.2 Multi-modal entity alignment

With the advancement of MMKGs, the integration of visual and textual modalities into EA tasks has garnered significant attention [34–37]. Various multi-modal fusion techniques have been developed to leverage the rich multi-modal information within MMKGs.

MMEA [21] introduced the MKF module, which projects entity embeddings from different modalities into a shared semantic space, facilitating interaction and complementarity across modalities. ACK-MMEA [38] employs a relation-aware GNN (ConsistGNN) to integrate unified attribute information. MSNEA [39] model applies visually guided relation learning and visually adaptive attribute learning, combining feature vectors from each modality into a cohesive representation. MoAlign [40] introduces a hierarchical multi-head attention mechanism to layer multimodal features, using distinct attention heads to focus on neighboring information, textual attributes, and visual attributes. MEAformer [41] employs a dynamic cross-modal weighting mechanism to generate modality weights in real-time, allowing instance-based weight adjustments across modalities. Following feature encoding, GEEA [42] passes embeddings from each modality through an adversarial network to achieve final fusion.

Unlike previous methods, the proposed NoCo leverages modality-aware noise enhancement to improve robustness by dynamically adjusting the intensity of perturbations for different modalities. By contrast, existing regularization techniques such as adversarial training and virtual adversarial training [43] typically inject worst-case perturbations in a modality-agnostic manner to stabilize representations. Although these methods are effective in general machine learning tasks, they do not explicitly consider the heterogeneity and reliability imbalance across modalities in MMKGs. Our approach, by learning modality-specific noise scales, ensures that the injected perturbations are consistent with the distinct feature space of each modality. Furthermore, NoCo incorporates a confidence-based dynamic fusion framework to adaptively adjust the contribution of each modality in the alignment decision, thereby mitigating the adverse influence of low-quality modalities. This task-specific synergy between adaptive noise augmentation and confidence-driven fusion differentiates our work from conventional adversarial noise injection methods and provides a novel perspective for robust MMEA.

3 Methodology

This section presents a comprehensive explanation of NoCo. Figure 2 shows that NoCo comprises three main modules that form the overall framework: (1) modality-aware augment, which obtains multi-modal embeddings and applies modality-aware noise enhancement; (2) confidence-based dynamic fusion, which calculates modality confidence and optimizes the fusion process; and (3) multi-modal contrastive learning, which employs intra-modal contrastive loss for the MMEA task.

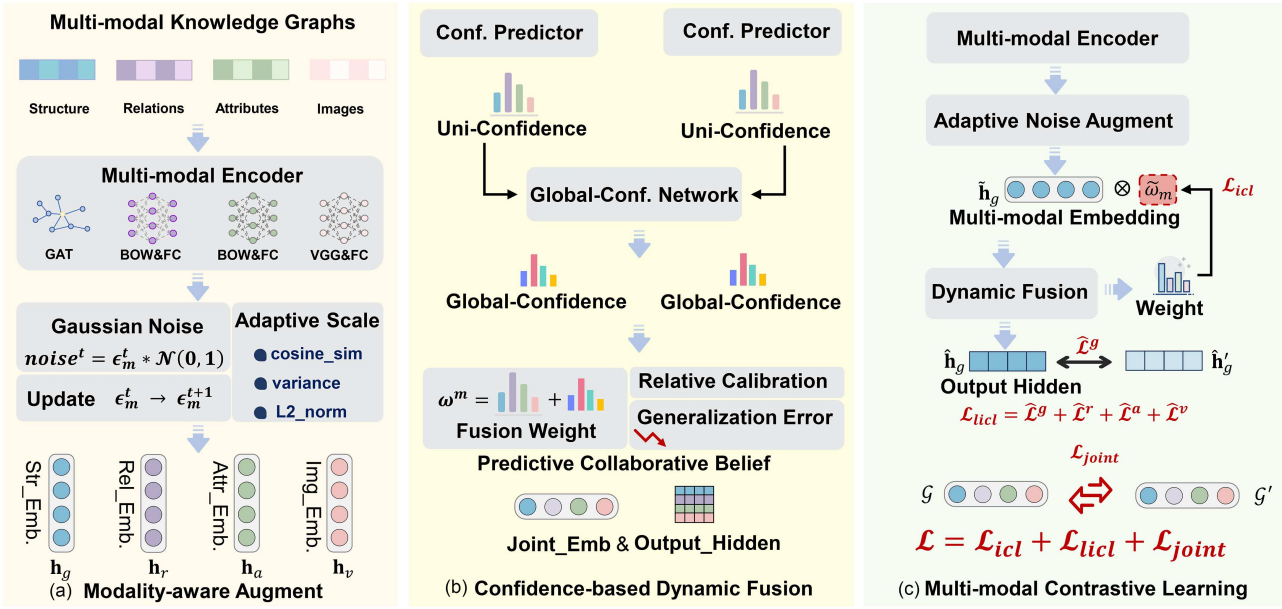

Figure 2 (Color online) Overall framework of NoCo. (a) Modality-aware augment: The multi-modal embedding is perturbed through a noise update mechanism to fully leverage the representational capacity of the modal space. (b) Confidence-based dynamic fusion: Utilize a confidence predictor and relative calibration for multi-modal fusion. (c) Multi-modal contrastive learning: Applies intra-modal and inter-modal contrastive losses to align entity representations across modalities.

3.1 Problem definition

We define a multi-modal knowledge graph as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}, \mathcal{T}\}$, where $\mathcal{E}, \mathcal{R}, \mathcal{A}$, and $\mathcal{V}$ represent the set of entities, relations, attributes, and images, respectively. $\mathcal{T} \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ represents the set of relation triple. $(h, r, t) \in \mathcal{T}$ is a relation triple, where h represents the head entity, t represents the tail entity, and r is the relation between the head and tail entities. The MMEA task aims to match corresponding entities from different KGs that describe the same concept in the real world. Give two MMKGs $\mathcal{G}_1 = \{\mathcal{E}_1, \mathcal{R}_1, \mathcal{A}_1, \mathcal{V}_1, \mathcal{T}_1\}$ and $\mathcal{G}_2 = \{\mathcal{E}_2, \mathcal{R}_2, \mathcal{A}_2, \mathcal{V}_2, \mathcal{T}_2\}$. MMEA aims to find each pair of entities $(e_1, e_2), e_1 \in \mathcal{E}_1, e_2 \in \mathcal{E}_2$, where e_1 and e_2 represent the same object in the real world. The set of aligned seeds for two MMKGs is defined as $\mathcal{H} = \{(e_1, e_2) | e_1 \in \mathcal{E}_1, e_2 \in \mathcal{E}_3, e_1 \equiv e_2\}$, where $\equiv$ represents the equivalence of the two entities. $\mathcal{H}$ provides training guidance for the MMEA task.

3.2 Modality-aware noise augment

3.2.1 *Multi-modal encoding*

In the multi-modal encoding module, we process entity features within MMKGs by handling each modality separately to capture their distinct perspectives. In this section, we will present the methods for embedding each modality of entities within the given MMKGs into a low-dimensional vector space.

(1) *Graph structure embedding.* We employed a graph attention network (GAT) for graph structure embedding in this study to capture the structural information of entities. GAT utilizes a self-attention mechanism to dynamically learn relations between node pairs, thereby allowing the model to adapt effectively to different graph topologies. Letting $\mathbf{h}_i \in \mathbb{R}^d$ be the hidden state of entity e_i, we have neighborhood aggregation characterized as

$$\mathbf{h}_i^g = ReLU\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}_g \mathbf{h}_j\right), \tag{1}$$

where $\mathcal{N}_i$ represents the first-hop neighbors of e_i, and $\mathbf{W}_g$ is the diagonal weight matrix used for the linear transformation. α_{ij} is the attention weight of entity pair (e_1, e_2), which is represented as

$$\begin{aligned} \alpha_{ij} &= \mathrm{softmax}\left(e_{ij}\right) \\ &= \frac{\exp\left(\mathrm{LeakyReLU}\left(\boldsymbol{\omega}^\top \left[\mathbf{W}_g \mathbf{h}_i || \mathbf{W}_g \mathbf{h}_j\right]\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(\mathrm{LeakyReLU}\left(\boldsymbol{\omega}^\top \left[\mathbf{W}_g \mathbf{h}_i || \mathbf{W}_g \mathbf{h}_k\right]\right)\right)}, \end{aligned} \tag{2}$$

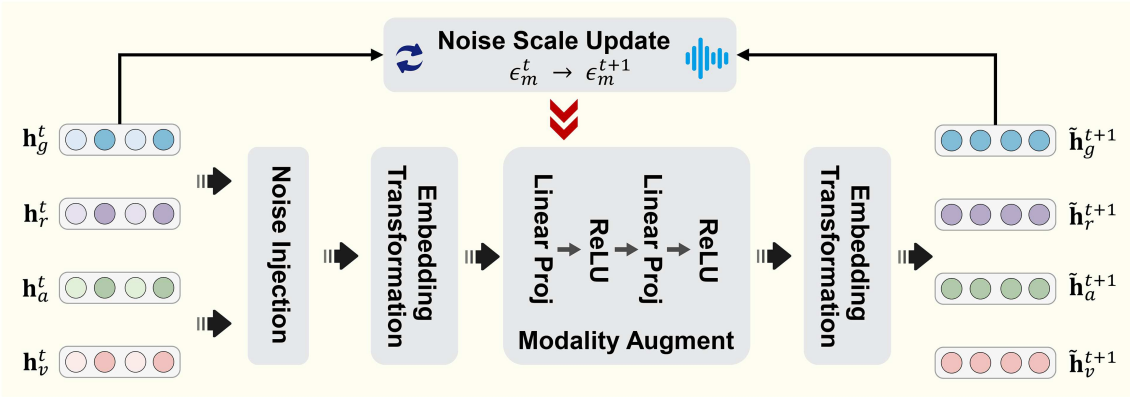

Figure 3 (Color online) Modality-aware noise augment module.

where $\boldsymbol{\omega} \in \mathbb{R}^{2d}$ is the learnable weight and $\|$ represents the concatenation operation. The GAT that we utilize has two attention heads and two layers, and we represent the output of the last layer as graph structure embedding $\mathbf{h}_g$.

(2) *Relation, attribute, and visual embedding.* We initially input the relations, attributes, and visual images of entity e_i into a simple fully connected layer to obtain the relation, attribute, and visual embeddings:

$$\mathbf{h}_i^m = \mathrm{FC}_m\left(\mathbf{W}_m, \mathbf{x}_i^m\right), m \in \{r, a, v\}, \tag{3}$$

where $r, a,$ and v represent the relation, attribute, and vision modality, respectively. $\mathbf{x}_i^m \in \mathbb{R}^{d_m}$ is the feature of entity e_i for modality m. $\mathbf{W}_m \in \mathbb{R}^{d \times d}$ is the weight matrix used for linear transformation. In addition, we represent the relation and attribute modalities of entities as bag-of-words features, simplifying their processing during the embedding operation. Meanwhile, we use a pre-trained image encoder to process the visual image of an entity, obtaining the visual embedding for the corresponding image v_i of entity e_i. We select the output from the last layer of the image encoder as the image features:

$$\mathbf{x}_i^v = \mathrm{ImageEncoder}_v\left(v_i\right). \tag{4}$$

3.2.2 *Noise augment*

We introduce a modality-aware noise enhancement mechanism to enhance the robustness of the model and improve its ability to handle real-world tasks. Unlike conventional approaches that inject perturbations in a uniform manner, our method learns a separate noise scaling factor for each modality, ensuring that the injected Gaussian noise is consistent with the distinct feature distribution and reliability of different modalities. This design allows the model to adaptively regularize heterogeneous modality spaces, thereby maximizing their representational capability. In contrast to adversarial training or other general-purpose regularization techniques, which typically generate modality-agnostic perturbations, our approach exploits the intrinsic characteristics of MMKGs to achieve task-specific, customized noise augmentation tailored for the MMEA scenario. Figure 3 shows the specifics of the modality-aware noise enhancement mechanism. Specifically, for the feature $\mathbf{h}_m$ of modality m, we add Gaussian noise iteratively:

$$\mathbf{h}_m^{(t+1)} = \mathbf{W}_m\left(\mathbf{h}_m^{(t)} + \epsilon_m \times \mathcal{N}(0,1)\right) + b_m, \tag{5}$$

where ϵ_m denotes the noise scaling factor for modality m, $\mathcal{N}(0,1)$ represents standard Gaussian distribution, and t represents iteration steps. The noise scaling factor is learned and updated separately for each modality, enabling modality-specific adjustments of noise intensity rather than applying a uniform perturbation across all modalities. After noise injection, a linear transformation is applied to perform dimension mapping, aligning the features to a unified representation space. Subsequently, the noise-perturbed embedding is fed into the enhancement module for forward propagation, enabling further processing and refinement of the features:

$$\overset{\sim}{\mathbf{h}}_m^{(t+1)} = f\left(\mathbf{h}_m^{(t+1)}, \theta\right), \tag{6}$$

where $f(\cdot)$ comprises multiple linear layers combined with a ReLU activation function and is optimized using parameter θ. Through this processing, we retain the majority of the information of the original representation while introducing slight variations to distribute the learned features more broadly across the embedding space. This noise-based enhancement directly regularizes the embedding space, promoting a more uniform representation distribution. The overall process of the modality-aware noise augment module is detailed in Algorithm 1.

Algorithm 1 Modality-aware noise augment.

Input:
$\mathbf{h}_m$: the embedding of modality m; Θ: the number of iteration steps; *noise_size*: the noisy embedding size;
Output:
The augmented embedding $\widetilde{\mathbf{h}}_m$; initialize the noise scale ϵ_m for each modality;
for t in range (Θ) **do**
$noise = \epsilon_m \times \mathcal{N}(0, I) \Leftarrow$ Generate Gaussian noise for modality m;
$\mathbf{h}_m^{(t+1)} = \mathbf{W}_m \left(\mathbf{h}_m^{(t)} + noise\right) + b_m \Leftarrow$ Calculate noisy embedding for modality m;
Transform $\mathbf{h}_m^{(t+1)}$ to *noise_size* embedding;
$\widetilde{\mathbf{h}}_m^{(t+1)} = f\left(\mathbf{h}_m^{(t+1)}, \theta\right) \Leftarrow$ Process noisy embedding through the augment model;
Map noisy embedding $\widetilde{\mathbf{h}}_m^{(t+1)}$ back to input size;
$\epsilon_m^{(t)} \rightarrow \epsilon_m^{(t+1)} \Leftarrow$ Update noise scale adaptively;
Set noisy embedding $\rightarrow$ embedding for iteration.
end for

3.3 Confidence-based dynamic fusion

Owing to the variability in multi-modal data quality and inherent differences between modalities, treating each modality equally can amplify irrelevant features, thereby interfering with the performance of EA. Consequently, we aim to dynamically adjust the weight of each modality in the alignment decision. This approach ensures that more reliable modalities are assigned higher weights, while the weight of lower-quality modalities is reduced, mitigating their potential negative impact on the EA results.

Inspired by the theory proposed by Cao et al. [44], which focuses on reducing the upper bound of generalization error in multimodal fusion, we propose a confidence-based dynamic fusion framework (Figure 4). The fusion weight in a multimodal system should not only consider the individual modality but also incorporate the states of other modalities. Consequently, we calculate single modality confidence Uni_Conf_m and global modality confidence $Global\_Conf_m$ for each modality in MMEA. Uni_Conf_m is derived from the negative intra-modality covariance between fusion weights and the loss function, whereas $Global\_Conf_m$ is based on the positive inter-modality covariance between these elements. This approach ensures a balanced integration that reflects both modality-specific and cross-modality interactions. Based on the calculated single modality confidence and global modality confidence, the initial fusion weight w_m for modality m can be determined:

$$w_m = Uni\_Conf_m + Global\_Conf_m. \tag{7}$$

Given the unreliability and variability of modal data, we enhance the relative calibration strategy inspired by Cao et al. to further reduce the uncertainty of fusion weight w_m. This strategy enables the dynamic adjustment of the relative position of each modality according to the quality changes of other modalities within MMEA, ensuring a more adaptive and balanced fusion process.

We first calculate probability distribution $\mathbf{P}_m$ for augmented embedding $\widetilde{\mathbf{h}}_m$ for modality m:

$$\mathbf{P}_m = softmax\left(\widetilde{\mathbf{h}}_m\right) = \frac{\exp\left(\widetilde{\mathbf{h}}_m\right)}{\sum_{j=1}^{d} \exp\left(\widetilde{\mathbf{h}}_{m,j}\right)}, \tag{8}$$

where $\widetilde{\mathbf{h}}_{m,j}$ denotes the j-th element of the embedding vector $\widetilde{\mathbf{h}}_m$ and d represents the dimension of embedding. Subsequently, we calculate the mean deviation of the probability distribution for each modality, which serves as a measure of its uncertainty:

$$DU_m = \frac{1}{d}\sum_{i=1}^{d} |\mathbf{P}_{m,i} - \mu_m|, \tag{9}$$

where $\mathbf{P}_{m,i}$ is the i-th component of probability distribution $\mathbf{P}_m$ and μ_m represents the mean of $\mathbf{P}_m$. Thus, a uniform distribution indicates high uncertainty, as it suggests that all components have similar probabilities. Conversely, a peaked distribution implies low uncertainty, as it reflects a dominant component with higher confidence in specific features. We calculate the relative calibration factor for each modality by utilizing the mean deviation of its probability distribution:

$$RC_m = \frac{DU_m}{\sum_{k \neq m} DU_k + \varepsilon}, \tag{10}$$

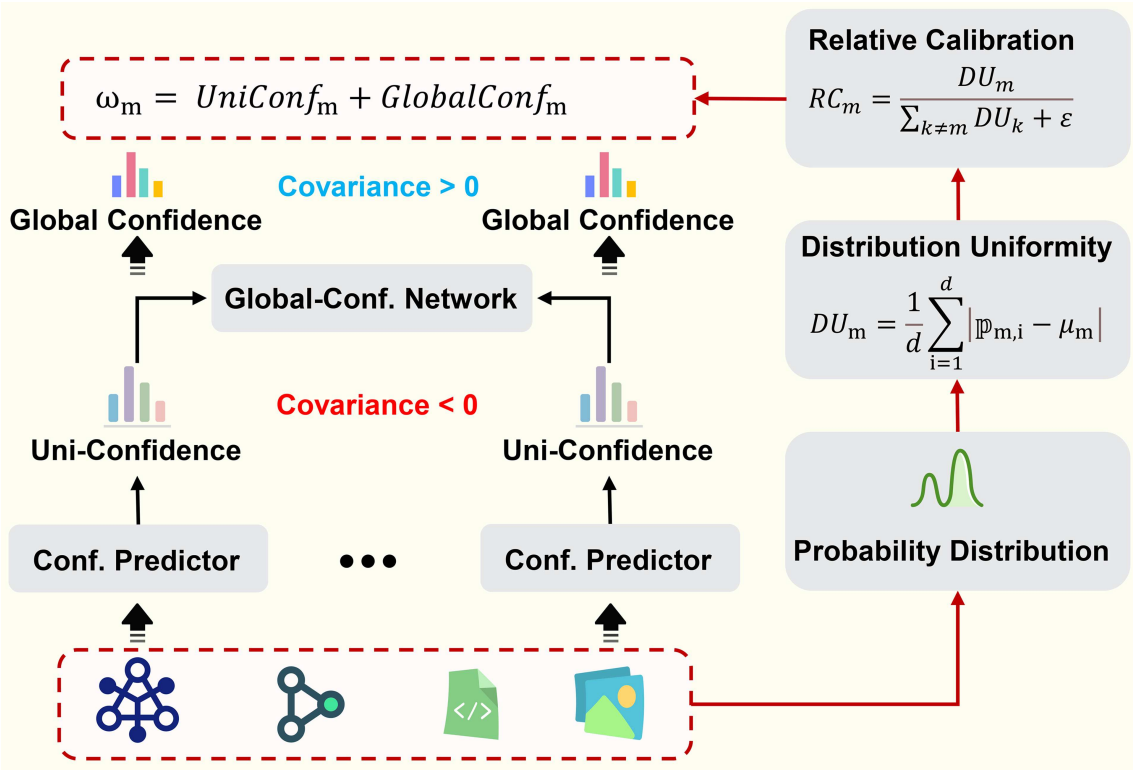

Figure 4 (Color online) Confidence-based dynamic fusion module.

which helps adjust the contribution of the modality to the fusion process based on its uncertainty. ε denotes the calibration smoothing factor, and by smoothing the numerical value of the calibration factor, it is stable when calculating the relative uncertainty. The initial fusion weights are adjusted using the relative calibration factor to yield the calibrated fusion weights:

$$\widetilde{w}_m = softmax\left(w_m \times RC_m\right). \tag{11}$$

Next, we define joint embeddings for modal fusion following the approach outlined by Chen et al. [41]:

$$\widetilde{\mathbf{h}}_{joint} = \bigoplus_{m \in \mathcal{M}} \left[\widetilde{w}_m \widetilde{\mathbf{h}}_m\right], \tag{12}$$

where $\widetilde{\mathbf{h}}_m$ represents the augmented embedding of modality m and $\bigoplus$ means the concatenation operation. The confidence-based dynamic fusion process is summarized in Algorithm 2.

3.4 Multi-modal contrastive learning

We introduce a multi-modal contrastive learning module that calculates both single-modal embedding and joint embedding loss to thoroughly capture intra-modal and inter-modal dynamics. This approach leverages a given pre-ranked seed pair $\mathcal{H}$ to comprehensively mine information within each modality, thereby enhancing the ability of the model to integrate and distinguish modality-specific features effectively.

According to Sun et al. [27], treating $\mathcal{H}$ as a positive sample is intuitive owing to the 1-to-1 alignment constraint, while any unaligned entity pair can be considered a negative sample. To construct informative negatives, we corrupt the pre-aligned entity pairs by including not only internally unaligned entities from the source graph KG1 but also cross-graph unaligned entities from the target graph KG2. This in-batch sampling strategy substantially enriches the diversity of negatives, as semantically or structurally similar entities are more likely to co-occur within the same batch, thus naturally providing hard negatives. Together with the subsequent contrastive learning across all modalities and their joint embeddings, the model is more likely to encounter negatives close to positives, which effectively prevents overfitting on trivial cases. For each seed alignment $\left(e_i^1, e_i^2\right)$ in $\mathcal{H}$, we define its negative set as

$$\mathcal{N}_i^{ng} = \left\{e_j^1 | e_j^1 \in \mathcal{E}_1, j \neq i\right\} \bigcup \left\{e_j^2 | e_j^2 \in \mathcal{E}_2, j \neq i\right\}. \tag{13}$$

Algorithm 2 Confidence-based dynamic fusion.

Input:
$\widetilde{\mathbf{h}}_m$: the augmented embedding of modality m; $\mathcal{M}$: the number of modalities;
Output:
The joint embedding $\hat{\mathbf{h}}_{joint}$; the fusion weight $\widetilde{w}_m$ for each modality; the hidden states $\hat{\mathbf{h}}_m$ for modality m; initialize the Uni-Conf. networks and Global-Conf. networks for modality m;
for modality m in $\mathcal{M}$ **do**
$Uni\_Conf_m \Leftarrow$ Calculate the Uni-Confidence for modality m;
$Uni\_Conf = \sum Uni\_Conf_m \Leftarrow$ Concatenate all Uni-Conf. scores as input to Global-Conf. Networks;
end for
for modality m in $\mathcal{M}$ **do**
$Global\_Conf_m \Leftarrow$ Calculate the Global-Confidence for modality m;
$w_m = Uni\_Conf_m + Global\_Conf_m \Leftarrow$ Calculate the preliminary fusion weight for modality m;
$\mathbf{P}_m = Softmax\left(\widetilde{\mathbf{h}}_m\right) = \frac{\exp\left(\widetilde{\mathbf{h}}_m\right)}{\sum_{j=1}^{d}\exp\left(\widetilde{\mathbf{h}}_{m,j}\right)} \Leftarrow$ Calculate the probability distribution for modality m;
$DU_m = \frac{1}{d}\sum_{i=1}^{d}|\mathbf{P}_{m,i} - \mu_m| \Leftarrow$ Calculate the mean deviation of $\mathbb{P}_m$ for modality m;
$RC_m = \frac{DU_m}{\sum_{k\neq m} DU_k + \varepsilon} \Leftarrow$ Get relative calibration factor of modality m;
$\widetilde{w}_m = softmax\left(w \times RC_m\right) \Leftarrow$ Obtain the calibrated fusion weight;
$\widetilde{\mathbf{h}}_{joint} = \bigoplus_{m\in\mathcal{M}}\left[\widetilde{w}_m\widetilde{\mathbf{h}}_m\right] \Leftarrow$ Generate the joint embedding;
$\hat{\mathbf{h}}_m \Leftarrow$ Process the hidden states of modality m.
end for

We apply an in-batch negative sampling strategy as outlined by Chen et al. [45], which restricts the sampling scope of $\mathcal{N}_i^{ng}$ to the mini-batch. Immediately following this step, we incorporate these negative samples into the contrastive learning module to enhance the robustness and discrimination capability of the model during training. We define the alignment probability distribution as

$$p_m\left(e_i^1, e_i^2\right) = \frac{\gamma_m\left(e_i^1, e_i^2\right)}{\gamma_m\left(e_i^1, e_i^2\right) + \sum\limits_{e_j\in\mathcal{N}_i^{ng}}\gamma_m\left(e_i^1, e_j\right)}, \tag{14}$$

where $\gamma_m\left(e_i, e_j\right) = \exp\left(s_i^{m\top}s_j^m/\tau\right)$ represents the similarity measure and τ is the temperature hyper-parameter. Notably, the distribution in the above equation is directional and asymmetric. Therefore, we define the bi-directional alignment objective for modality m as follows:

$$\mathcal{L}_m = -\log\left(p_m\left(e_i^1, e_i^2\right) + p_m\left(e_i^2, e_i^1\right)\right)/2. \tag{15}$$

Lin et al. [46] introduced intra-modality contrastive loss (ICL) to the MMEA task, which compels input embeddings to maintain the similarity of entities within the original embedding space. ICL enables the model to differentiate embeddings of the same entities across different KGs from other entity embeddings within each modality [46]. We follow their method to design ICL, and apply a late intra-modal contrastive loss (LICL) to achieve inter-modal complementarity:

$$\mathcal{L}_{icl} = \sum_{m\in\mathcal{M}} -\log\left(p_m\left(e_i^1, e_i^2\right) + p_m\left(e_i^2, e_i^1\right)\right)/2, \tag{16}$$

$$\begin{aligned}\mathcal{L}_{licl} &= \sum_{m\in\mathcal{M}}\widetilde{\mathcal{L}}_m \\ &= \sum_{m\in\mathcal{M}} -\log\left(\widetilde{p}_m\left(e_i^1, e_i^2\right) + \widetilde{p}_m\left(e_i^2, e_i^1\right)\right)/2,\end{aligned} \tag{17}$$

$$\widetilde{p}_m\left(e_i^1, e_i^2\right) = \frac{\widetilde{\gamma}_m\left(e_i^1, e_i^2\right)}{\widetilde{\gamma}_m\left(e_i^1, e_i^2\right) + \sum\limits_{e_j\in\mathcal{N}_i^{ng}}\widetilde{\gamma}_m\left(e_i^1, e_j\right)}, \tag{18}$$

where $\widetilde{\gamma}_m\left(e_i, e_j\right) = \exp\left(\widetilde{\mathbf{h}}_i^{m\top}\widetilde{\mathbf{h}}_j^m/\tau\right)$ is calculated by the last hidden state $\widetilde{\mathbf{h}}^m$. In addition, we design the loss $\mathcal{L}_{joint}$, which is computed using the joint embedding. Finally, we design the following overall loss to train our model:

$$\mathcal{L} = \mathcal{L}_{joint} + \mathcal{L}_{icl} + \mathcal{L}_{licl}. \tag{19}$$

Table 1 Statistics of the six public datasets.

Dataset	MMKGs	#Ent.	#Rel.	#Attr.	#Rel. Triple	#Attr. Triple	#Image	#Seed
FB15K-DB15K	FB15K	14951	1345	116	592213	29395	13444	12849
	DB15K	12842	279	225	89197	48080	12837	
FB15K-YG15K	FB15K	14951	1345	116	592213	29395	13444	111199
	YG15K	15404	32	7	122886	23532	11194	
Multi-OpenEA$_{\text{EN-FR-V1}}$	EN (English)	15000	267	308	47334	73121	15000	15000
	FR (French)	15000	210	404	40864	67167	15000	
Multi-OpenEA$_{\text{EN-DE-V1}}$	EN (English)	15000	215	286	47676	83755	15000	15000
	DE (German)	15000	131	194	50419	156150	15000	
Multi-OpenEA$_{\text{D-W-V1}}$	DBpeida	15000	248	342	38265	68258	15000	15000
	Wikidata	15000	169	649	42746	138246	15000	
Multi-OpenEA$_{\text{D-W-V2}}$	DBpedia	15000	167	175	73983	66813	15000	15000
	Wikidata	15000	121	457	83365	175686	15000	

4 Experiments

In this section, we conduct an extensive series of experiments to evaluate the effectiveness of the NoCo model using six subsets from two publicly available datasets. We aim to answer the following seven research queries.

- **RQ1**: How does the performance of the NoCo model compare with existing MMEA models?
- **RQ2**: How does modal information impact the performance of NoCo?
- **RQ3**: How do the principal modules affect the performance of NoCo?
- **RQ4**: How do various hyperparameter settings impact the performance of NoCo?
- **RQ5**: How does NoCo perform when applied to real-world MMEA tasks?
- **RQ6**: How does NoCo perform with low alignment seed rates?
- **RQ7**: How does the efficiency of NoCo measure up against the baseline models?

4.1 Experimental settings

4.1.1 *Datasets*

We selected two publicly available MMEA datasets to evaluate the effectiveness of our proposed method: MMKG [10] and Multi-OpenEA [47]. We enumerate in Table 1 the statistics of all the datasets used in the experiments. Among these, MMKG first extracts subsets from Freebase, YAGO, and DBpedia to construct a cross-KG dataset. We select FB15K-DB15K and FB15K-YG15K from MMKG, utilizing three data splits containing 20%, 50%, and 80% of reference EA pairs as pre-aligned seeds for training, respectively. Multi-OpenEA is supplemented with entity images obtained from Google search queries. From this dataset, we select two bilingual subsets, EN-FR-V1 and EN-DE-V1, and two monolingual subsets D-W-V1 and D-W-V2, adopting 30% of the pre-aligned seed pairs for training.

4.1.2 *Evaluation metrics*

We use cosine similarity to assess the probability of EA and rank the matching candidates for each entity to be aligned. We use metrics such as Hits@N (N = 1, 5, 10) and mean reciprocal rank (MRR). Specifically, Hits@N measures whether the correct item appears within the top N-ranked results, indicating retrieval performance. MRR captures the quality of the ranking by calculating the inverse of the rank at which the first correct answer appears, thereby assessing how well the model ranks the correct entity within its candidate list. In the following tables of experimental results, the best results are marked in bold, while the second-best results are underlined.

4.1.3 *Baseline models*

We compare the proposed NoCo with a range of leading MMEA models to demonstrate its effectiveness and advancements. These baseline models are categorized into three distinct groups based on the fusion method employed.

(1) *Static fusion method*.

• PoE [10] (ESWC'19): PoE processes the data of each modality through multiple independent "expert models" and finally fuses the respective outputs to generate decisions.

• MMEA [21] (KSEM'20): MMEA introduces the MKF module, which projects entity embeddings from different modalities into a shared semantic space, facilitating interaction and complementarity across modalities.

(2) *Dynamic fusion method.*

• EVA [48] (AAAI'21): EVA encodes image, relation, structure, and attribute features independently using different feature extractors and then fuses these features with learnable weights in an intermediate layer.

• ACK-MMEA [38] (WWW'23): ACK-MMEA utilizes a relationship-aware GNN to fuse the consistent attribute information.

• MSNEA [39] (KDD'22): MSNEA applies visually guided relation learning and visually adaptive attribute learning, combining feature vectors from each modality into a cohesive representation.

• MCLEA [46] (COLING'22): MCLEA introduces intra-modality contrastive loss and inter-modality alignment loss into MMEA.

• TRIFAC [49] (NEUNET'24): TRIFAC decomposes the original MMKG by employing embedding refinement through a two-stage MMKG decomposition process.

• GEEA [42] (ICLR'24): GEEA explores EEA from the generative models perspective and introduces a generative EEA framework with the proposed mutual variational autoencoder as the generative model.

(3) *Attention-based method.*

• MoAlign [40] (EMNLP'23): MoAlign uses a hierarchical modifiable multi-head attention transformer to model entity representations from both structural and semantic aspects.

• MEAformer [41] (ACM MM'23): MEAformer proposes a meta-modal hybrid MMEA transformer to dynamically predict the cross-cor-relation coefficient between modalities.

• SNAG [50] (COLING'25): SNAG incorporates a multi-head cross-modal attention mechanism within the Transformer framework and fuses the features of each modality at the entity level using a dynamic weighting mechanism.

• DESAlign [51] (ICDE'24): DESAlign tackles the issue of over-smoothing that arises from semantic inconsistency and leverages existing modalities to insert missing semantics.

• SGMEA [52] (COLING'25): SGMEA introduces structure-guided modality enhancement to leverage graph structure for enriching visual and attribute embeddings, thereby improving MMEA accuracy.

4.1.4 *Implementation details*

The hidden layer dimensions for all networks are standardized to 300. We set the total number of training epochs to 500, with an additional 500 epochs for an optional iterative training strategy. Training strategies include cosine warm-up scheduling, early stopping, and gradient accumulation to enhance performance and stability during training. In the multi-modal embedding module, we select VGG-16 [53] as the image encoder and set d_v to 4096. The Bag-of-Words representation for relation and attribute modalities follows the method of Yang et al. [54], where each modality feature is represented by a fixed-length vector of 1000. The choice of VGG-16 and Bag-of-Words for visual and textual representation is consistent with most prior MMEA work, ensuring fair comparison with existing baselines and allowing us to isolate the contribution of our proposed modules from the representational power of stronger encoders. For MSNEA, following the setting in [41], we eliminate the use of raw attribute values for input consistency across baselines and extend MSNEA with iterative training ability. Additionally, we apply an alignment editing method to mitigate error accumulation, as outlined in [27].

4.2 Performance comparison (RQ1)

As illustrated in Tables 2–4, the results of the MMEA task demonstrate that NoCo surpasses the current state-of-the-art models across most evaluation metrics. Notably, the proposed model outperformed recent approaches, including GEEA and DESAlign. Additionally, as a model designed to utilize noisy data, our model demonstrated superiority over SNAG. Specifically, for the Hits@1 metric, our model achieved a 1.43% improvement over the baseline on the Multi-OpenEA datasets, 3% improvement on FB15K-DB15K, and 2.18% improvement on FB15K-YG15K.

NoCo employed a modality-aware noise update mechanism for multi-modal data augmentation, enhancing the representation capacity of the modal space and improving the robustness of the model. Simultaneously, it incorporates a confidence-based dynamic fusion framework to adjust the weight of each modality in the alignment decision, effectively capturing the interaction between modalities. Additionally, a relative calibration strategy is utilized to reduce the upper bound of the model generalization error, ensuring more reliable performance. The superior performance of NoCo compared with other multi-modal models highlights promising research directions for addressing multi-modal data quality issues and modality unreliability.

Table 2 Performance comparison of different MMEA models on FB15K-DB15K dataset.

FB15K-DB15K												
Model	$R_{seed}=20\%$				$R_{seed}=50\%$				$R_{seed}=80\%$			
	MRR	Hits			MRR	Hits			MRR	Hits		
		@1	@5	@10		@1	@5	@10		@1	@5	@10
Static-based												
PoE (ESWC'19)	0.170	0.126	0.164	0.251	0.533	0.464	0.601	0.658	0.721	0.666	0.783	0.820
MMEA (KSEM'20)	0.357	0.265	0.451	0.541	0.512	0.416	0.621	0.703	0.685	0.590	0.804	0.868
Dynamic-based												
EVA (AAAI'21)	0.283	0.199	0.356	0.448	0.422	0.334	0.522	0.589	0.563	0.484	0.625	0.696
ACK-MMEA (WWW'23)	0.387	0.304	0.455	0.549	0.624	0.560	0.683	0.736	0.752	0.682	0.816	0.874
MSNEA (KDD'22)	0.175	0.114	0.203	0.296	0.388	0.288	0.499	0.590	0.613	0.518	0.744	0.779
MCLEA (COLING'22)	0.393	0.295	0.496	0.582	0.652	0.573	0.742	0.800	0.784	0.730	0.857	0.883
TRIFAC (NEUNET'24)	0.389	0.318	0.487	0.559	0.607	0.554	0.623	0.750	0.761	0.697	0.851	0.882
GEEA (ICLR'24)	0.450	0.343	0.565	0.661	0.723	0.651	0.788	0.852	0.836	0.787	0.886	0.918
Attention-based												
MoAlign (EMNLP'23)	0.409	0.318	0.478	0.564	0.634	0.576	0.696	0.749	0.773	0.699	0.854	0.882
MEAformer (ACM MM'23)	0.534	0.434	0.634	0.728	0.704	0.625	0.790	0.847	0.825	0.773	0.887	0.918
SNAG (COLING'25)	0.495	0.389	0.642	0.702	0.742	0.669	0.825	0.871	0.848	0.802	0.893	0.928
DESAlign (ICDE'24)	0.539	0.441	0.652	0.740	0.728	0.656	0.822	0.853	0.850	0.805	0.898	0.926
SGMEA (COLING'25)	0.540	0.450	0.649	**0.752**	0.737	0.677	0.824	0.869	0.826	0.801	0.895	0.922
NoCo (Ours)	**0.544**	**0.459**	**0.665**	0.739	**0.753**	**0.689**	**0.831**	**0.875**	**0.856**	**0.819**	**0.912**	**0.931**

Table 3 Performance comparison of different MMEA models on FB15K-YG15K dataset.

FB15K-YG15K												
Model	$R_{seed}=20\%$				$R_{seed}=50\%$				$R_{seed}=80\%$			
	MRR	Hits			MRR	Hits			MRR	Hits		
		@1	@5	@10		@1	@5	@10		@1	@5	@10
Static-based												
PoE (ESWC'19)	0.154	0.113	0.196	0.229	0.414	0.347	0.475	0.536	0.635	0.573	0.711	0.746
MMEA (KSEM'20)	0.317	0.234	0.398	0.480	0.486	0.403	0.572	0.645	0.682	0.597	0.785	0.839
Dynamic-based												
EVA (AAAI'21)	0.224	0.153	0.292	0.361	0.388	0.311	0.488	0.534	0.565	0.491	0.656	0.692
ACK-MMEA (WWW'23)	0.360	0.289	0.413	0.496	0.593	0.535	0.661	0.699	0.744	0.676	0.829	0.864
MSNEA (KDD'22)	0.153	0.103	0.200	0.249	0.413	0.320	0.534	0.589	0.620	0.531	0.733	0.778
MCLEA (COLING'22)	0.332	0.254	0.401	0.484	0.616	0.543	0.702	0.759	0.715	0.653	0.794	0.835
TRIFAC (NEUNET'24)	0.371	0.290	0.442	0.508	0.579	0.546	0.606	0.694	0.736	0.669	0.787	0.895
GEEA (ICLR'24)	0.393	0.298	0.506	0.585	0.668	0.589	0.744	0.794	0.780	0.732	0.849	0.890
Attention-based												
MoAlign (EMNLP'23)	0.378	0.296	0.482	0.525	0.617	0.550	0.658	0.713	0.769	0.689	0.851	0.884
MEAformer (ACM MM'23)	0.416	0.325	0.517	0.598	0.640	0.560	0.731	0.780	0.768	0.705	0.840	0.874
SNAG (COLING'25)	0.405	0.309	0.524	0.596	0.658	0.578	0.750	0.804	0.804	0.747	0.861	0.902
DESAlign (ICDE'24)	0.425	**0.342**	0.530	0.601	0.680	0.612	0.748	0.799	0.790	0.734	0.866	0.887
SGMEA (COLING'25)	0.431	**0.335**	**0.521**	**0.593**	0.668	0.607	0.752	0.809	0.802	0.745	0.860	0.898
NoCo (Ours)	**0.446**	0.339	**0.533**	**0.614**	**0.692**	**0.619**	**0.765**	**0.816**	**0.811**	**0.762**	**0.874**	**0.905**

4.3 Modality effect (RQ2)

We conducted ablation studies on different modal features involved in NoCo to comprehensively explore the influence of various factors. Table 5 lists the impact of distinct modality information in NoCo on the MMEA task. Specifically, we consider four variants of the NoCo model for the ablation study. (1) **w/o structure**: a variant that excludes the graph structure modality. (2) **w/o relation**: a variant that omits the entity relation modality. (3) **w/o attribute**: a variant that removes the entity attributes. (4) **w/o image**: a variant that excludes the vision modality.

The results indicate that, although the extent of performance reduction varies, the removal of any modality feature results in a notable decrease in overall performance. This confirms that the inclusion of multi-modal information plays a pivotal role in improving the effectiveness of EA tasks. Each modality contributes unique and complementary

Table 4 Performance comparison of different MMEA models on Multi-OpenEA datasets.

Model	Mulit-OpenEA							
	EN-FR-V1		EN-DE-V1		D-W-V1		D-W-V2	
	MRR	Hits@1	MRR	Hits@1	MRR	Hits@1	MRR	Hits@1
MSNEA	0.734	0.692	0.804	0.753	0.826	0.800	0.873	0.838
EVA	0.836	0.785	0.945	0.922	0.891	0.858	0.922	0.890
MCLEA	0.864	0.819	0.957	0.939	0.908	0.881	0.949	0.928
MEAformer	0.882	0.836	0.971	0.954	0.933	0.909	0.962	0.944
UMAEA	0.891	0.847	0.970	0.955	0.930	0.905	0.967	0.948
NoCo (Ours)	**0.898**	**0.859**	**0.978**	**0.964**	**0.943**	**0.934**	**0.973**	**0.960**

Table 5 Ablation study on the different modalities of NoCo.

Model	FB15K-DB15K				FB15K-YG15K			
	MRR	Hits			MRR	Hits		
		@1	@5	@10		@1	@5	@10
w/o structure	0.569	0.677	0.804	0.882	0.524	0.571	0.707	0.735
w/o relation	0.838	0.799	0.886	0.921	0.792	0.732	0.830	0.890
w/o attribute	0.834	0.795	0.881	0.920	0.789	0.737	0.826	0.891
w/o image	0.822	0.788	0.878	0.916	0.770	0.731	0.815	0.884
NoCo	**0.856**	**0.819**	**0.912**	**0.931**	**0.811**	**0.762**	**0.874**	**0.905**

information that, when integrated, enhances the ability of the model to align entities accurately.

Among the different modalities, the exclusion of the graph structure modality leads to the most significant drop in performance. This finding underscores the critical importance of structural data in MMEA, highlighting its role as a foundational component for accurate representation and alignment across KGs. The reliance on graph structure data reinforces the understanding that rich connectivity and relational context are vital for distinguishing and aligning entities effectively in complex multi-modal scenarios.

4.4 Key components (RQ3)

We constructed three variants for ablation studies to evaluate the importance of specific components in the NoCo model: (1) **w/o NA**, a version that omits the modality-aware noise enhancement mechanism, (2) **w/o CDF**, a version that removes the confidence-based dynamic fusion framework, (3) **w UN**, a variant where the same level of noise is applied uniformly across all modalities. These modifications are designed to isolate and assess the individual contributions of these key components to the overall performance of the model. Figure 5 presents the impact of these variations on the model.

The results indicate that omitting the noise enhancement module leads to a decrease across various performance metrics, emphasizing its role in enhancing robustness by perturbing the embedding space and diversifying learned features. We further introduce the model variant **w UN** to validate the effectiveness of the modality-aware noise. Experimental results demonstrate that applying uniform noise across all modalities leads to improved performance compared with the variant without noise augmentation, but it still falls noticeably short of the full model equipped with modality-aware noise. This indicates that uniform noise alone cannot sufficiently capture the heterogeneous characteristics of different modalities, whereas modality-aware noise can better align with modality-specific properties and thus provides more significant performance gains. The performance decline is even more significant when the confidence-based dynamic fusion framework is excluded, affirming the value of this method in ensuring the adaptive integration of modality weights. This framework enhances the robustness of the model by dynamically balancing the contribution of each modality using confidence calculations and relative calibration strategies, which optimize alignment outcomes.

Furthermore, the observed drop in performance when the noise enhancement mechanism is removed highlights its essential role in bolstering the MMEA process. This mechanism allows NoCo to leverage modal space representation more effectively, helping to counteract the impact of unreliable or noisy data and resulting in a more uniform and stable embedding distribution. Overall, these findings underscore the importance of both the modality-aware noise enhancement and the confidence-based dynamic fusion framework in achieving optimal performance in MMEA.

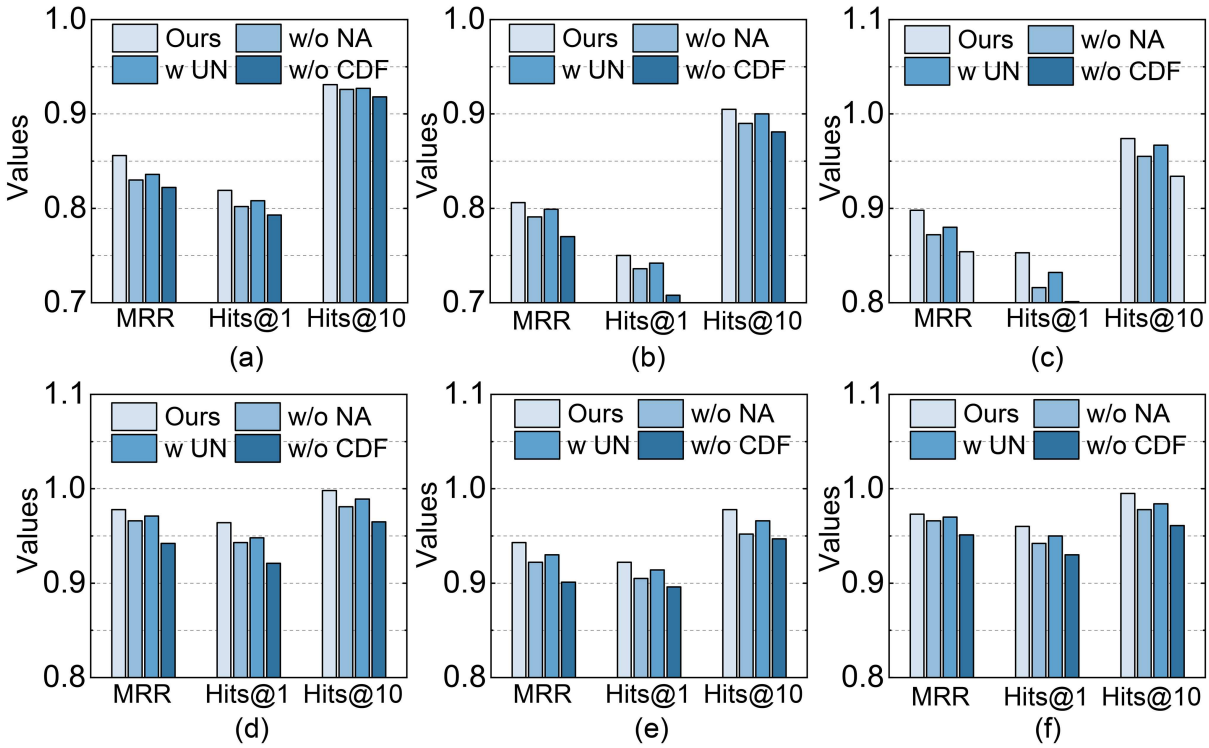

Figure 5 (Color online) Ablation study on key components of proposed NoCo. (a) FB15K-DB15K; (b) FB15K-YG15K; (c) OpenEA EN-FR-V1; (d) OpenEA EN-DE-V1; (e) OpenEA D-W-V1; (f) OpenEA-D-W-V2.

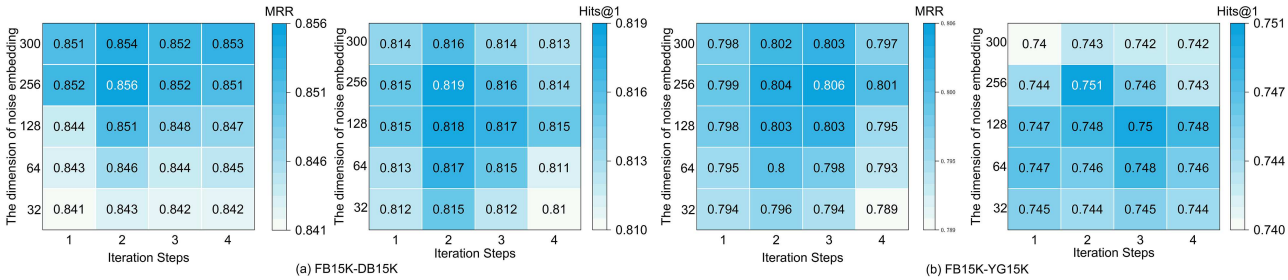

Figure 6 (Color online) Hyper-parameter analysis is performed on the dimension of noise embedding and iteration steps.

4.5 Hyper-parameter sensitivity (RQ4)

In this subsection, we explore the impact of hyper-parameters on model performance from two perspectives. First, we analyze the influence of the noise embedding dimension and the number of iterative update steps on the model performance. Figure 6 shows the effects of these variations. Second, we examine the impact of the calibration smoothing factor on performance, with the corresponding results presented in Figure 7.

(1) *Noise dimension and iteration steps.* The results indicate that setting the number of iteration steps either excessively high or low can lead to a degradation in model performance. This finding suggests that achieving optimal performance requires careful tuning to strike the right balance in the number of iteration steps. Additionally, the analysis shows that increasing the dimension of the noise embedding generally enhances the expressiveness and performance of the model. However, while larger embedding dimensions improve the capacity of the model to represent complex patterns, this comes with a trade-off: higher computational and storage requirements.

Avoiding excessively high dimensions that would disproportionately increase resource consumption without a significant gain in performance is crucial to balancing these considerations. After evaluating these factors, we opted to set the embedding dimension to 256. This configuration provides a satisfactory compromise, optimizing the expressiveness and alignment accuracy of the model while maintaining computational and storage efficiency. This balanced approach ensures that the model remains practical for real-world applications, where resource constraints and processing efficiency are critical.

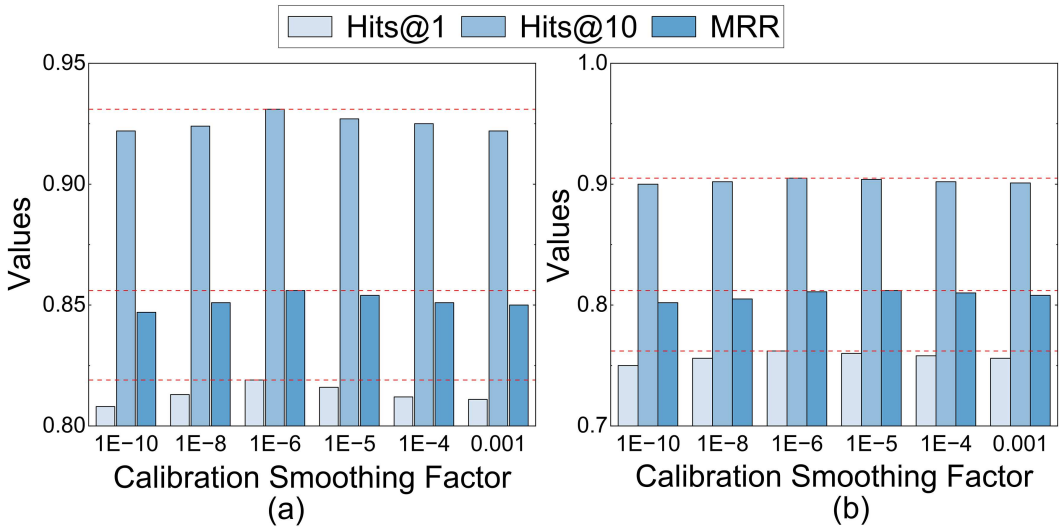

Figure 7 (Color online) Hyper-parameter analysis is performed on calibration smoothing factor. (a) FB15K-DB15K; (b) FB15K-YG15K.

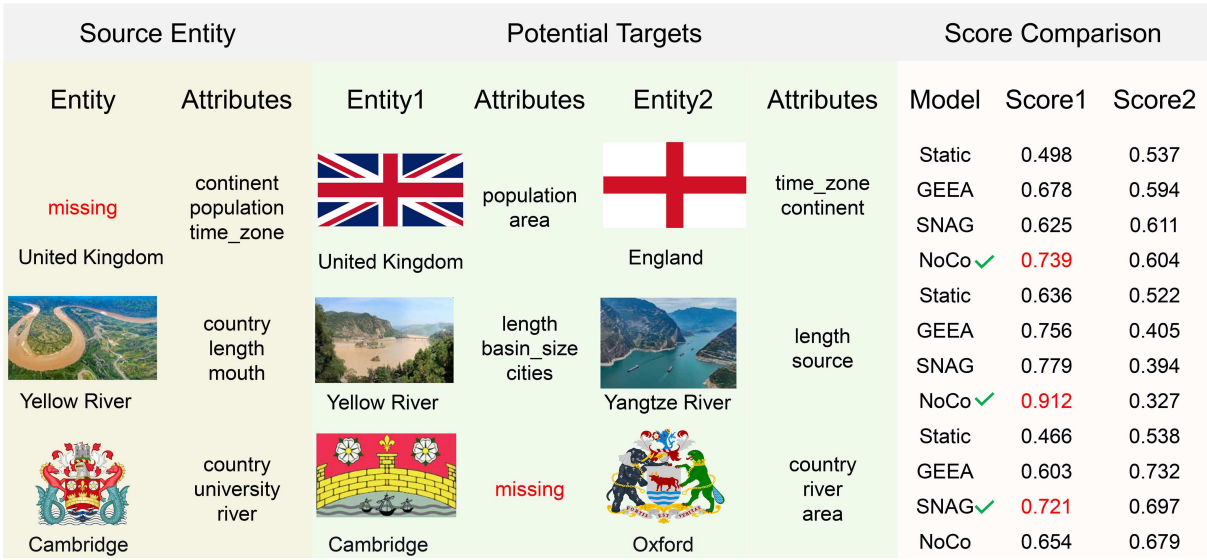

Figure 8 (Color online) Case example: MMEA tasks and score comparison with different models.

(2) *Calibration smoothing factor.* The calibration smoothing factor ε plays a critical role in regulating the smoothness of modal weights within the relative calibration strategy, ensuring that the dynamic adjustment of each modal weight remains stable during the measurement of uncertainty. This factor allows for fine-tuning the sensitivity of the model to variations in modality reliability. The degree to which uncertainty impacts the final weight assignment of different modalities can be effectively controlled by adjusting the calibration smoothing factor.

Figure 7 shows that setting the calibration smoothing factor excessively high or low results in a decline in model performance. An overly small factor makes the model excessively sensitive to minor variations, potentially causing instability, while a factor that is excessively large can overly smooth the weight adjustments, limiting the adaptability of the model. To balance sensitivity and stability, we opt for an intermediate value for the parameter, which ensures optimal performance by providing sufficient responsiveness while maintaining robust weight distribution.

4.6 Case study (RQ5)

To confirm the practicality of NoCo when applied to the real-world MMEA tasks, we conduct a case study, as illustrated in Figure 8. We illustrate the images and attributes of source entities alongside potential targets, together with a comparison of alignment scores produced by models with different fusion strategies. The results clearly demonstrate that our proposed NoCo achieves the best performance in aligning *United Kingdom* and *Yellow River*, respectively. The static fusion method treats all modalities equally without accounting for their differences.

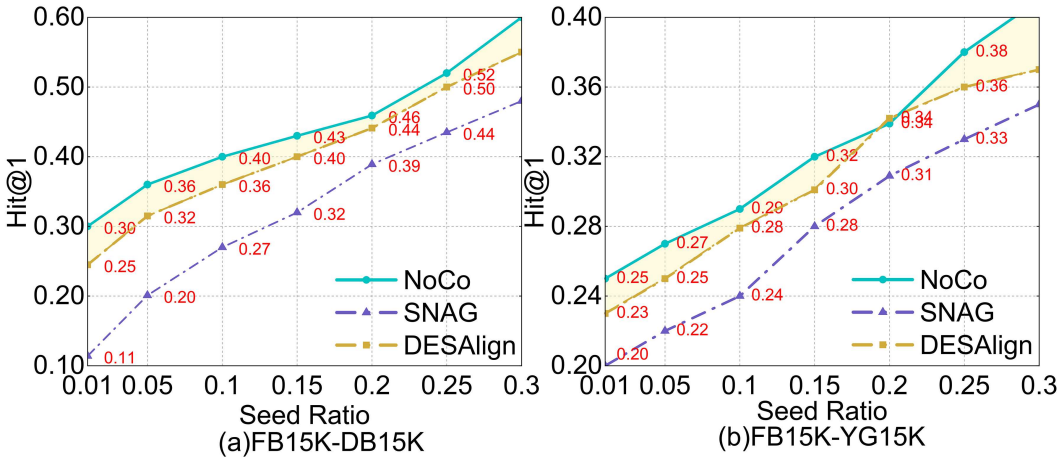

Figure 9 (Color online) Hits@1 performance with low seed ratios.

Table 6 Efficiency comparison of NoCo against the selected baselines.

Dataset	Metric	SNAG	GEEA	MEAformer	NoCo
FB15K-DB15K	Time (s./epoch)	3.5	4.1	3.1	3.0
FB15K-YG15K	Time (s./epoch)	3.1	3.2	2.7	2.5
OpenEA EN-FR-V1	Time (s./epoch)	3.7	3.8	3.6	3.6
OpenEA EN-DE-V1	Time (s./epoch)	3.8	3.6	3.7	3.5

Although GEEA and SNAG dynamically adjust modality weights, their prediction scores remain lower than those of NoCo. This shows that our confidence-based dynamic fusion effectively reduces the influence of low-quality modalities and validates the effectiveness of the framework. However, in the case of aligning *Cambridge*, NoCo was confused and incorrectly matched it with *Oxford*. This indicates that mitigating the influence of low-quality modalities and irrelevant features remains a major challenge in MMEA tasks.

4.7 Low resource study (RQ6)

To thoroughly assess the efficiency of NoCo in practical MMEA tasks, we examined its stability under conditions with very few pre-aligned seed pairs. Pre-aligned entity seed pairs are those already aligned and used in prior experiments to guide the training of the model, enhancing alignment accuracy and efficiency. In real-world EA tasks, the availability of pre-aligned seed pairs is often minimal or nonexistent, testing the generalization ability of the model and its capacity to learn from unlabeled data. This evaluation helps determine the performance of the model when supervised information is limited.

We varied the ratio of pre-aligned seed pairs from 0.01 to 0.3 and conducted experiments on the FB15K-DB15K and FB15K-YG15K datasets. The results, shown in Figure 9, demonstrate that our proposed NoCo consistently achieves leading performance, even with minimal seed pairs, and maintains this advantage as the seed ratio increases. Notably, on the FB15K-YG15K dataset, the performance gap between NoCo and other methods widened as the ratio improved.

4.8 Efficiency of NoCo (RQ7)

In addition to evaluating alignment accuracy, we selected several baseline models and measured the training time per epoch for each, performing a comprehensive comparative analysis. All evaluations were performed on an RTX 4090 GPU with 24 GB of memory, as detailed in Table 6, which underscores the efficiency of different models.

SNAG, which incorporates a Gaussian modality noise masking module and multi-head cross-modality attention, requires considerable training time. This extended processing time is attributed to the complexity of its multi-head attention mechanism, which demands significant computational resources to handle intricate cross-modality interactions and fusion. GEEA, featuring a generative entity alignment (EEA) framework with a mutual variational autoencoder, requires the longest training time among all the models tested owing to the heavy computational demands of its variational inference processes.

By contrast, NoCo, which integrates a confidence-based dynamic fusion approach, demonstrates a shorter training time compared to MEAformer and exhibits superior efficiency and performance overall. This efficiency is achieved

Table 7 Performance comparison with CLIP-based module.

Model	FB15K-DB15K				FB15K-YG15K			
	MRR	Hits			MRR	Hits		
		@1	@5	@10		@1	@5	@10
NoCo (VGG+BOW)	0.856	0.819	0.912	0.931	0.811	0.762	0.874	0.905
NoCo (CLIP)	**0.860**	**0.823**	**0.918**	**0.931**	**0.819**	**0.774**	**0.883**	**0.916**

by dynamically adjusting modality weights based on confidence, optimizing resource allocation, and maintaining robust performance without sacrificing speed.

4.9 Analysis of feature encoders

To rigorously assess the robustness and adaptability of our method with SOTA multi-modal representations, we conducted a supplementary study utilizing CLIP, extending our original data process framework. We extracted features with intrinsic cross-modal alignment and superior semantic coherence by leveraging the image and text encoders of CLIP. Empirical results across the FB15K-DB15K and FB15K-YG15K datasets in Table 7 indicate that the CLIP-enhanced model consistently outperformed the original model, highlighting the efficacy of CLIP in capturing complex cross-modal semantics.

Although CLIP features exhibit clear advantages in semantic representation capability and cross-modal consistency, their reliance on large-scale pretrained models also incurs higher computational and storage costs. In future work, we will further explore the integration of lightweight multi-modal pretrained models with our proposed method to enhance its practicality and deployment efficiency while maintaining competitive performance. Meanwhile, we plan to validate the generalization ability of our approach in more real-world, large-scale MMEA scenarios, and to investigate deeper integration with more advanced vision-language pretrained models, to further promote the development of cross-modal EA research.

5 Conclusion

This study proposed NoCo, a noise-augmented multi-modal entity alignment method with confidence-based dynamic fusion. NoCo investigated the impact of multi-modal data quality unreliability on MMEA tasks and introduced a confidence-based dynamic fusion framework to adjust the weight of each modality dynamically in the alignment decision based on data quality and uncertainty. This approach effectively mitigated the negative effects of low-quality modalities on the results. Additionally, we proposed a modality-aware noise enhancement method to fully leverage the representational capacity of the modal space, thereby improving the robustness of the model. Experimental results on two public datasets demonstrate the effectiveness of our method. In future work, we aim to further explore the MMEA task based on NoCo. We believe that addressing modal heterogeneity represents a promising research direction.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62207011, 62377009, 62301213, 62407013), Natural Science Foundation of Hubei Province of China (Grant No. 2025AFB653), and Open Fund of Key Laboratory of Intelligent Sensing System and Security of Hubei University, Ministry of Education (Grant No. KLISSS202410).

References

1 Bordes A, Usunier N, García-Durán A, et al. Translating embeddings for modeling multi-relational data. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems, 2013. 2787–2795
2 Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014. 1112–1119
3 Fan M, Zhou Q, Chang E, et al. Transition-based knowledge graph embedding with relational mapping properties. In: Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation, 2014. 328–337
4 Bi Z, Chen J, Jiang Y, et al. CodeKGC: code language model for generative knowledge graph construction. ACM Trans Asian Low-Resour Lang Inf Process, 2024, 23: 1–16
5 Liu J, Ke W, Wang P, et al. Towards continual knowledge graph embedding via incremental distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 8759–8768
6 Xu F, Li M H, Huang Q, et al. Knowledge graph-driven graph neural network-based model for rumor detection (in Chinese). Sci Sin Inform, 2023, 53: 663–681
7 Bai Y D, Chen S, Xing Z C, et al. ArgusDroid: detecting Android malware variants by mining permission-API knowledge graph. Sci China Inf Sci, 2023, 66: 192101

8 Mao X, Wang W, Xu H, et al. Relational reflection entity alignment. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management, 2020. 1095–1104

9 Zhang R, Su Y, Trisedya B D, et al. AutoAlign: fully automatic and effective knowledge graph alignment enabled by large language models. IEEE Trans Knowl Data Eng, 2024, 36: 2357–2371

10 Liu Y, Li H, García-Durán A, et al. MMKG: multi-modal knowledge graphs. In: Proceedings of the 16th International Conference on Semantic Web, 2019. 459–474

11 Liu Z, Cao Y, Pan L, et al. Exploring and evaluating attributes, values, and structures for entity alignment. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020. 6355–6364

12 Trisedya B D, Qi J, Zhang R. Entity alignment between knowledge graphs using attribute embeddings. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, 2019. 297–304

13 Ding Y, Yu J, Liu B, et al. Mukea: multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 5079–5088

14 Wang P, Wu Q, Shen C, et al. Explicit knowledge-based reasoning for visual question answering. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017. 1290–1296

15 Xu N, Gao Y, Liu A A, et al. Multi-modal validation and domain interaction learning for knowledge-based visual question answering. IEEE Trans Knowl Data Eng, 2024, 36: 6628–6640

16 Sun R, Cao X, Zhao Y, et al. Multi-modal knowledge graphs for recommender systems. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management, 2020. 1405–1414

17 Zhang F, Yuan N J, Lian D, et al. Collaborative knowledge base embedding for recommender systems. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. 353–362

18 Yang S, Zhang R, Erfani S M, et al. Unimf: a unified framework to incorporate multimodal knowledge bases into end-to-end task-oriented dialogue systems. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021. 3978–3984

19 Zhu Y, Kiros R, Zemel R S, et al. Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 19–27

20 Zhu X, Li Z, Wang X, et al. Multi-modal knowledge graph construction and application: a survey. IEEE Trans Knowl Data Eng, 2024, 36: 715–735

21 Chen L, Li Z, Wang Y, et al. MMEA: entity alignment for multi-modal knowledge graph. In: Proceedings of the 13th International Conference on Knowledge Science, Engineering and Management, 2020. 134–147

22 He W, Li Z, Lu D, et al. Multimodal dialogue systems via capturing context-aware dependencies of semantic elements. In: Proceedings of the 28th ACM International Conference on Multimedia. 2755–2764

23 Zhang Q, Sun Z, Hu W, et al. Multi-view knowledge graph embedding for entity alignment. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019. 5429–5435

24 Pei S, Yu L, Hoehndorf R, et al. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference. In: Proceedings of the World Wide Web Conference, 2019. 3130–3136

25 Luo M Q, Zhang C X, Peng C, et al. Knowledge graph completion based on parsing graph embedding and a weighted graph convolutional network (in Chinese). Sci Sin Inform, 2022, 52: 2037–2057

26 Chen M, Tian Y, Yang M, et al. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017. 1511–1517

27 Sun Z, Hu W, Zhang Q, et al. Bootstrapping entity alignment with knowledge graph embedding. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018. 4396–4402

28 Zhao Y, Zhou H, Zhang A, et al. Connecting embeddings based on multiplex relational graph attention networks for knowledge graph entity typing. IEEE Trans Knowl Data Eng, 2023, 35: 4608–4620

29 Jiang X, Zhu R, Ji P, et al. Co-embedding of nodes and edges with graph neural networks. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 7075–7086

30 Wang Z, Lv Q, Lan X, et al. Cross-lingual knowledge graph alignment via graph convolutional networks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018. 349–357

31 Xu K, Wang L, Yu M, et al. Cross-lingual knowledge graph alignment via graph matching neural network. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019. 3156–3161

32 Wu Y, Liu X, Feng Y, et al. Relation-aware entity alignment for heterogeneous knowledge graphs. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019. 5278–5284

33 Cao Y, Liu Z, Li C, et al. Multi-channel graph neural network for entity alignment. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019. 1452–1461

34 Wang Y, Huang W, Sun F, et al. Deep multimodal fusion by channel exchanging. In: Proceedings of the 33rd Conference on Neural Information Processing Systems, 2020. 4835-4845

35 Xie R, Liu Z, Luan H, et al. Image-embodied knowledge representation learning. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017. 3140–3146

36 Xie R, Liu Z, Jia J, et al. Representation learning of knowledge graphs with entity descriptions. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016. 2659–2665

37 Pezeshkpour P, Chen L, Singh S. Embedding multimodal relational data for knowledge base completion. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018. 3208–3218

38 Li Q, Guo S, Luo Y, et al. Attribute-consistent knowledge graph representation learning for multi-modal entity alignment. In: Proceedings

of the ACM Web Conference, 2023. 2499–2508
39 Chen L, Li Z, Xu T, et al. Multi-modal siamese network for entity alignment. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022. 118–126
40 Li Q, Ji C, Guo S, et al. Multi-modal knowledge graph transformer framework for multi-modal entity alignment. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023. 987–999
41 Chen Z, Chen J, Zhang W, et al. Meaformer: multi-modal entity alignment transformer for meta modality hybrid. In: Proceedings of the 31st ACM International Conference on Multimedia, 2023. 3317–3327
42 Guo L, Chen Z, Chen J, et al. Revisit and outstrip entity alignment: a perspective of generative models. In: Proceedings of the 12th International Conference on Learning Representations, 2024
43 Miyato T, Maeda S I, Koyama M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans Pattern Anal Mach Intell, 2019, 41: 1979–1993
44 Cao B, Xia Y, Ding Y, et al. Predictive dynamic fusion. In: Proceedings of the Forty-first International Conference on Machine Learning, 2024
45 Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning, 2020. 1597–1607
46 Lin Z, Zhang Z, Wang M, et al. Multi-modal contrastive representation learning for entity alignment. In: Proceedings of the 29th International Conference on Computational Linguistics, 2022. 2572–2584
47 Sun Z, Zhang Q, Hu W, et al. A benchmarking study of embedding-based entity alignment for knowledge graphs. Proc VLDB Endow, 2020, 13: 2326–2340
48 Liu F, Chen M, Roth D, et al. Visual pivoting for (unsupervised) entity alignment. In: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, 2021. 4257–4266
49 Li Q, Li J, Wu J, et al. Triplet-aware graph neural networks for factorized multi-modal knowledge graph entity alignment. Neural Netws, 2024, 179: 106479
50 Chen Z, Fang Y, Zhang Y, et al. Noise-powered multi-modal knowledge graph representation framework. In: Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, 2025. 141–155
51 Wang Y, Sun H, Wang J, et al. Towards semantic consistency: Dirichlet energy driven robust multi-modal entity alignment. In: Proceedings of the 40th IEEE International Conference on Data Engineering, 2024. 3559–3572
52 Cheng J, Guo M, Zhang F. SGMEA: structure-guided multimodal entity alignment. In: Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, 2025. 7851–7861
53 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations, 2015
54 Yang H, Zou Y, Shi P, et al. 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, 2019. 4430–4440