Capturing Latent Modal Association For Multimodal Entity Alignment

Yongquan Ji, Jingwei Cheng*, Fu Zhang, Chenglong Lu

[1]School of Computer Science and Engineering, Northeastern University, China,
[2]Key Laboratory of Intelligent Computing of Medical Images,
Ministry of Education, Northeastern University, China.
{chengjingwei,zhangfu}@neu.edu.cn,
{jyq0609,chenglonglu233}@gmail.com

Abstract

Multimodal entity alignment aims to identify equivalent entities in heterogeneous knowledge graphs by leveraging complementary information from multiple modalities. However, existing methods often overlook the quality of input modality embeddings during modality interaction—such as missing modality generation, modal information transfer, modality fusion—which may inadvertently amplify noise propagation while suppressing discriminative feature representations. To address these issues, we propose a novel model—CLAMEA for **c**apturing **l**atent modal **a**ssociation for **m**ultimodal **e**ntity **a**lignment. Specifically, we use a self-attention mechanism to enhance salient information while attenuating noise within individual modality embeddings. We design a dynamic modal attention flow fusion module to capture and balance latent intra- and inter-modal associations and generate fused modality embeddings. Based on both fused and available modalities, we adopt variational autoencoder (VAE) to generate high-quality embeddings for the missing modality. We use a cross-modal association extraction module to extract latent modal associations from the completed modality embeddings, further enhancing embedding quality. Experimental results on two real-world datasets demonstrate the effectiveness of our approach, which achieves an absolute 3.1% higher Hits@1 score than the sota method [1].

1 Introduction

Multimodal Knowledge Graphs (MMKGs) (Peng et al., 2023) are an extension of traditional knowledge graphs that integrate information from multiple modalities, such as text, images, and triples. MMKGs have demonstrated wide applicability in tasks like question answering (Chen et al., 2022c) and recommendation systems (Wang et al., 2019). Multimodal Entity Alignment (MMEA) aims to identify equivalent entities in MMKGs by leveraging their multimodal information.

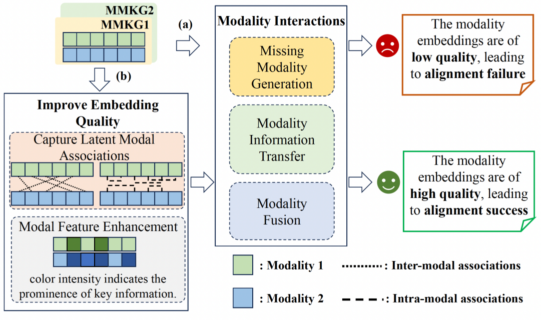
MMKG2
MMKG1
(a)
(b)
Modality Interactions
Missing Modality Generation
Modality Information Transfer
Modality Fusion
The modality embeddings are of **low quality**, leading to **alignment failure**
The modality embeddings are of **high quality**, leading to **alignment success**
Improve Embedding Quality
Capture Latent Modal Associations
Modal Feature Enhancement
color intensity indicates the prominence of key information.
: Modality 1
: Modality 2
: Inter-modal associations
: Intra-modal associations

Figure 1: (a) indicates that modality embeddings are not processed before modality interaction, resulting in low-quality embeddings and ultimately leading to alignment failure. (b) indicates that processing modality embeddings in advance (modal feature enhancement; capturing latent modal associations) improves the quality of the embeddings and leads to successful alignment.

The process of MMEA can be broadly divided into two stages: modality embedding and modality interaction. Current MMEA methods primarily focus on optimizing modality embeddings or improving modality interaction mechanisms to enhance overall model performance. Prominent examples for modality embedding encoders include TransE (Bordes et al., 2013), BERT (Devlin et al., 2019), and ResNet (He et al., 2016), which have become well-established for deriving relational, attribute, and visual embeddings, respectively.

Modality interaction refers to the integration of multimodal information, such as missing modality generation, modal information transfer and modality fusion. Specifically, EVA (Liu et al., 2021) employs weighted concatenation to fuse modality embeddings. MSNEA (Chen et al., 2022a) uses the visual modality to guide other modalities. MEAformer (Chen et al., 2023) adopts an attention-based dynamic weighting and concatenation strat-

*Corresponding author.

[1]Code: https://github.com/Quanquan429/CLAMEA

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22278–22293
November 4-9, 2025 ©2025 Association for Computational Linguistics

egy to fuse modality embeddings. ACK-MMEA (Li et al., 2023) constructs attribute-consistent MMKGs by fusing and generating modality embeddings. GEEA (Guo et al., 2023) employs a Variational Autoencoder (VAE) to generate more expressive modality embeddings. PMMEA (Tang and Wang, 2024) enhances modality embeddings by incorporating positional information.

Although existing methods have achieved promising results, they often overlook the quality of input modality embeddings during modality interactions—such as missing modality generation (Guo et al., 2023), modal information transfer (Zhu et al., 2023), and modality fusion (Chen et al., 2023)—which may lead to noise amplification and the suppression of key information, ultimately affecting the entity alignment performance.

As shown in the Figure 1, the quality of the input modality embeddings significantly affects the quality of the modality embeddings obtained after interaction, which in turn impacts the final alignment performance. Moreover, studies such as IBMEA (Su et al., 2024a), SimDiff (Li et al., 2024) similarly indicate that highlighting alignment information and suppressing irrelevant information are crucial, and the quality of embeddings plays a key role in the outcomes of subsequent interactions.

To address the above issues, we propose a novel method—CLAMEA, which aims to improve the quality of modality embeddings by capturing latent modal associations. Firstly, we introduce a self-attention mechanism to highlight the key information in modal embeddings and suppress noise, thereby enhancing modal embeddings. Secondly, we design a Dynamic Modal Attention Flow Fusion module to capture and balance latent intra- and inter-modal associations, generating fused modality embeddings. Thirdly, we leverage both the fused and avaiable modalities to generate embeddings for missing modalities, thereby mitigating the negative impact of missing modalities. Finally, we further extract the latent modal associations between modalities to further optimize the modal embedding.

The main contributions of this paper are summarized as follows:

- We focus on latent modal associations, aiming to capture and extract latent intra- and inter-modality associations more deeply to improve the quality of modality embeddings.
- We propose a new model, CLAMEA, to capture latent modal associations by designing a dynamic modal attention flow mechanism, cross-modal association extraction, dynamic missing modality generation, and modal feature enhancement. In the whole process, the collaborative optimization of modal embeddings and interaction is achieved.
- We conducted extensive experiments on two real-world datasets, FB15K-DB15K and FB15K-YG15K, and the results show that our method achieves an absolute 3.1% higher Hits@1 score than the sota method, demonstrating excellent performance.

2 Related work

2.1 Entity Alignment

Traditional entity alignment aims to identify semantically equivalent entities across different knowledge graphs by comparing their attributes, names, and structural information. Details are in Appendix A.

2.2 Multimodal Entity Alignment

Multimodal entity alignment significantly improves alignment performance by integrating information from different modalities. For example, EVA (Liu et al., 2021) is the first to combine images with structure, relations, and attributes for entity alignment, and assign a learnable weight to each modality. MSNEA (Chen et al., 2022a) adopts a vision-dominant approach to guide the learning of relational and attribute modalities. MEAformer (Chen et al., 2023) dynamically predict modality weights using an attention mechanism. GEEA (Guo et al., 2023) uses a variational autoencoder (VAE) (Kingma et al., 2013) to predict modal information for entity alignment. IBMEA (Su et al., 2024a) integrates low-dimensional unimodal embeddings into high-dimensional representations. SimDiff (Li et al., 2024) enhancing the utilization of multimodal data by adding and removing noise vectors. LoginMEA (Su et al., 2024b) propose a novel local-to-global interaction network for MMEA. TMEA (Chen et al., 2024) tackling uncertain correspondences through a commonality extraction mechanism. These methods effectively fuse multimodal information to boost performance. However, most existing methods fail to consider the quality of modality embeddings before modality interaction. To address this issue, we propose a novel approach that captures latent modal associations and enhance modality embeddings, thereby improving model performance.

3 Problem Definition

3.1 Multimodal Knowledge Graph (MMKG)

A multimodal knowledge graph (MMKG) consists of visual, attribute, and relational modalities. It can be represented as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{I}, \mathcal{A}, \mathcal{V}, \mathcal{T}^{\mathcal{R}}, \mathcal{T}^{\mathcal{A}}, \mathcal{P})$, where $\mathcal{E}$, $\mathcal{R}$, $\mathcal{I}$, $\mathcal{A}$, $\mathcal{V}$, $\mathcal{T}^{\mathcal{R}}$, $\mathcal{T}^{\mathcal{A}}$, and $\mathcal{P}$ represent the sets of entities, relations, images, attributes, attribute values, relational triples, attribute triples, and entity-image pairs, respectively. The relational triple set $\mathcal{T}^{\mathcal{R}} = \{(h_h, r, h_t) \mid h_h, h_t \in \mathcal{E}, r \in \mathcal{R}\}$ consists of relational triples, each of which is composed of a head entity h_h, a relation r, and a tail entity h_t. The attribute triple set $\mathcal{T}^{\mathcal{A}} = \{(h, a, v) \mid h \in \mathcal{E}, a \in \mathcal{A}, v \in \mathcal{V}\}$ consists of attribute triples, where each triple contains an entity h, one of its attributes a, and the corresponding attribute value v. The entity-image pair set $\mathcal{P} = \{(h, i) \mid h \in \mathcal{E}, i \in \mathcal{I}\}$ represents the set of pairs consisting of an entity h and an image i.

3.2 Multimodal Entity Alignment (MMEA)

The task of Multimodal Entity Alignment (MMEA) aims to identify semantically equivalent entities between two multimodal knowledge graphs. Formally, given two input MMKGs $\mathcal{G}_1 = (\mathcal{E}_1, \mathcal{R}_1, \mathcal{I}_1, \mathcal{A}_1, \mathcal{V}_1, \mathcal{T}^{\mathcal{R}}{}_1, \mathcal{T}^{\mathcal{A}}{}_1, \mathcal{P}_1)$ and $\mathcal{G}_2 = (\mathcal{E}_2, \mathcal{R}_2, \mathcal{I}_2, \mathcal{A}_2, \mathcal{V}_2, \mathcal{T}^{\mathcal{R}}{}_2, \mathcal{T}^{\mathcal{A}}{}_2, \mathcal{P}_2)$, the output is a set of aligned entity pairs, defined as $\mathcal{D} = \{(h_1, h_2) \mid h_1 \in \mathcal{E}_1, h_2 \in \mathcal{E}_2, h_1 \equiv h_2\}$, where $\equiv$ indicates that the two entities are semantically equivalent in the real world.

4 Methodology

This section presents the technical details of our proposed CLAMEA model, as shown in Figure 2. The model consists of five main modules: 1) Multimodal Knowledge Encoder (MKE) which encodes information from different modalities; 2) Modal Feature Enhancement (MFE) which optimizes modality embeddings using self-attention; 3) Dynamic Modal Attention Flow Fusion (DMAFF) which captures and balances latent modal associations to generate fused modalities; 4) Dynamic Missing Modality Generation (DMMG) which combines Variational Autoencoder (VAE) and dynamic modal fusion mechanism to generate missing modality from available modalities; 5) Cross-modal Association Extraction (CMAE) which extracts associations among modalities using cross-modal attention mechanisms to enhance modality embeddings. Additionally, we incorporate contrastive learning and iterative optimization strategies.

4.1 Multimodal Knowledge Encoder

In this module, we use TransE (Bordes et al., 2013), the pre-trained visual model ResNet (He et al., 2016), and the pre-trained BERT (Devlin et al., 2019) to generate relational modality embedding $\mathbf{h}_o^r$, visual modality embedding $\mathbf{h}_o^v$, and attribute modality embedding $\mathbf{h}_o^a$ of entities, respectively. Please refer to Appendix B for details.

4.2 Modal Feature Enhancement

In MMEA, the importance of each piece of information in the modality embedding varies. Certain information is essential for entity alignment, while other information has minimal or even detrimental effects (Su et al., 2024a). To address this, this module enhances the ability to capture information within the modality using a self-attention mechanism (Vaswani et al., 2017), emphasizing key information and suppressing irrelevant information, thus achieving modality embeddings enhancement. The specific enhancement method is defined as follows:

$$\mathbf{h}^m = Concat(head_1, \ldots, head_\eta)\mathbf{W}_h + \mathbf{b}_h, \quad (1)$$

$$head_i = \mathrm{Att}(\mathbf{h}_o^m \mathbf{W}_i^Q, \mathbf{h}_o^m \mathbf{W}_i^K, \mathbf{h}_o^m \mathbf{W}_i^V), \quad (2)$$

$$\mathrm{Att}(\mathbf{Q}_m, \mathbf{K}_m, \mathbf{V}_m) = softmax\left(\frac{\mathbf{Q}_m \mathbf{K}_m^T}{\sqrt{d}}\right)\mathbf{V}_m, \quad (3)$$

where $m \in \{a, r, v\}$, $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ represent the query, key, and value, respectively, softmax$(\cdot)$ is the activation function, d is the dimension of the key, η is the number of heads, $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, $\mathbf{W}_i^V$ are the projection matrices for the i-th attention head, and $\mathbf{W}_h$, $\mathbf{b}_h$ are the projection matrix and bias vector.

4.3 Dynamic Modal Attention Flow Fusion

In MMKGs, entities contain information from different modalities, often displaying complex and tightly coupled semantic associations. Existing methods mostly focus on processing modality embeddings based on intra-modal information, neglecting inter-modal associations. Furthermore, different modalities have varying importance, making it difficult for simple fusion strategies to balance the associations among modalities.

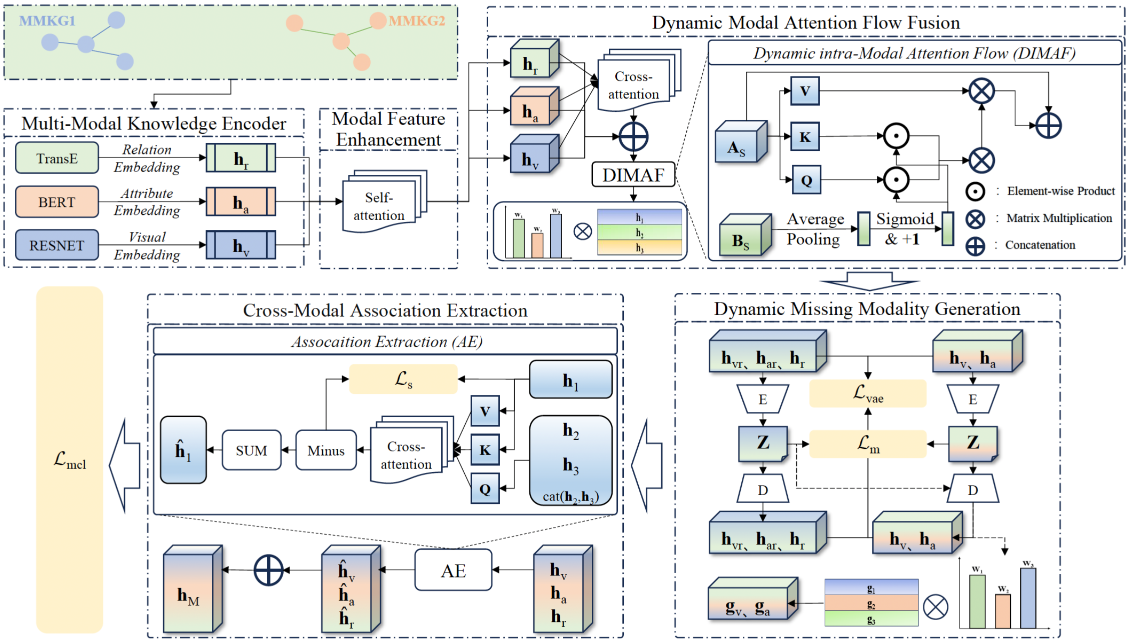

Figure 2: The overall framework of CLAMEA.

Inspired by DFAF (Gao et al., 2019), we make improvements based on its framework. DFAF is originally applied to the Visual Question Answering (VQA) task. It should be noted that after modality interactions are completed, multiple modality interactions are still retained separately, without forming a unified fused modality. This mechanism cannot directly meet the requirements of our task. To address this issue, we innovatively introduce a dynamic weight fusion strategy based on the attention mechanism to integrate the interacted modal information. We adapt this strategy and apply it for the first time to the MMEA, providing a novel and feasible solution for modality fusion.

Specifically, we introduce a dynamic fusion Modal Attention Flow (MAF) mechanism (Gao et al., 2019). MAF is further divided into Inter-Modal Attention Flow and Dynamic Intra-Modal Attention Flow. This mechanism effectively transmits information within and among modalities, capturing and balancing modality association.

We use two modalities, A and B, as an example for explanation. First, a inter-modal attention module captures the attention flow between modalities, enabling their interaction. Then, intra-modal attention modules are used to capture the attention flow within each modality, where intra-modal interactions are dynamically modulated by information from the other modality. Finally, the dynamic weight fusion module integrates the information from both inter-modal and intra-modal attention, adaptively weighting them.

4.3.1 Inter-Modal Attention Flow

We first apply a linear transformation to the modality embeddings of $\mathbf{A}$ and $\mathbf{B}$, then compute the attention flow between them to capture the inter-modal association, followed by feature aggregation to update the modality embeddings. This process enables the identification of cross-modal relationships and captures important information between modalities. The detailed computation is as follows:

$$\mathbf{A}_{update} = softmax\left(\frac{\mathbf{Q}_A\mathbf{K}_B^\top}{\sqrt{d}}\right)\mathbf{V}_B, \quad (4)$$

$$\mathbf{B}_{update} = softmax\left(\frac{\mathbf{Q}_B\mathbf{K}_A^\top}{\sqrt{d}}\right)\mathbf{V}_A, \quad (5)$$

After obtaining the updated modality embeddings, we concatenate them with the initial embeddings to ensure that the original modality information is preserved. Finally, a fully connected layer is applied to transform the concatenated embeddings, resulting in the updated modality embeddings $\mathbf{A}_s$ and $\mathbf{B}_s$, which are then fed into the Dynamic Intra-Modal Attention module.

$$\mathbf{A_s} = Linear\left([\mathbf{A}, \mathbf{A}_{update}]\right), \quad (6)$$

$$\mathbf{B_s} = Linear\left([\mathbf{B}, \mathbf{B}_{update}]\right), \quad (7)$$

4.3.2 Dynamic Intra-Modal Attention Flow

The Dynamic Intra-Modality Attention module receives modality embeddings that already contain cross-modal association. We first perform average pooling on the modality embeddings. The pooled cross-modal embeddings are then mapped to a latent space through a learnable linear transformation, followed by a Sigmoid activation function to generate gating vectors. These gating vectors are used to scale the query and key features of the target modality, thereby enhancing or suppressing specific information within the embeddings. The detailed formulation is as follows:

$$\mathbf{G}_{As \leftarrow Bs} = \sigma\left(Linear(\text{AvgPool}(\mathbf{B}_s))\right), \quad (8)$$

$$\mathbf{G}_{Bs \leftarrow As} = \sigma\left(Linear(\text{AvgPool}(\mathbf{A}_s))\right), \quad (9)$$

$$\begin{aligned} \mathbf{Q}_{\hat{A}} &= (\mathbf{1} + \mathbf{G}_{As \leftarrow Bs}) \odot \mathbf{Q}_{As}, \\ \mathbf{Q}_{\hat{B}} &= (\mathbf{1} + \mathbf{G}_{Bs \leftarrow As}) \odot \mathbf{Q}_{Bs}, \\ \mathbf{K}_{\hat{A}} &= (\mathbf{1} + \mathbf{G}_{As \leftarrow Bs}) \odot \mathbf{K}_{As}, \\ \mathbf{K}_{\hat{B}} &= (\mathbf{1} + \mathbf{G}_{Bs \leftarrow As}) \odot \mathbf{K}_{Bs}, \end{aligned} \quad (10)$$

In the formula, $\odot$ denotes element-wise multiplication. The scaling factor takes the form of **1 + G**, which not only preserves the original embedding distribution (with a base value of 1) but also enhances the embeddings through the cross-modal gating **G**. The key aspect of this mechanism is that the gating **G** is derived from the conditional information of heterogeneous modalities. This establishes semantic dependencies across modalities.

$$\mathbf{h}^A = Linear(\mathbf{A}_s + \hat{\mathbf{A}}_{update}), \quad (11)$$

$$\mathbf{h}^B = Linear(\mathbf{B}_s + \hat{\mathbf{B}}_{update}), \quad (12)$$

$$\hat{\mathbf{A}}_{update} = softmax\left(\frac{\mathbf{Q}_{\hat{A}}\mathbf{K}_{\hat{A}}}{\sqrt{d}}\right)\mathbf{V}_{As}, \quad (13)$$

$$\hat{\mathbf{B}}_{update} = softmax\left(\frac{\mathbf{Q}_{\hat{B}}\mathbf{K}_{\hat{B}}}{\sqrt{d}}\right)\mathbf{V}_{Bs}, \quad (14)$$

Finally, the fused embedding is obtained through a dynamic weighted summation (Dsum) of each modality-specific unimodal embedding $\mathbf{h}^m$ with its corresponding dynamic weight w_m, as follows:

$$\mathbf{h}^{AB} = \sum_{m \in \{A,B\}} w_m \cdot \mathbf{h}^m, \quad (15)$$

Detailed explanation of w_m, please refer to Appendix C.

4.4 Dynamic Missing Modality Generation

In MMKGS, entities often suffer from missing information in the visual or attribute modalities. Such missing data leads to the incompleteness of MMKGs, which in turn affects entity alignment. Generally, the relational modality of MMKGs is complete, as determined by its characteristics. Therefore, this module focuses on the generation of visual and attribute modalities (Li et al., 2023).

In this module, inspired by (Chen et al., 2022b, 2024), we leverage the existing modality information to generate pseudo-embeddings for the missing modalities. Specifically, we utilize the fused modality embeddings $\mathbf{h}^{ar}$ and $\mathbf{h}^{vr}$ produced by the Dynamic Modal Attention Flow Fusion module, along with the available modality embeddings, to generate pseudo-embeddings for the missing modalities through Variational Autoencoders (VAEs) (Kingma et al., 2013). We used multiple VAEs to learn latent representations. For each VAE, $\mathbf{h}^x$ is the target modality embedding, $\mathbf{z}^x$ is the latent representation of $\mathbf{h}^x$, $q_\phi(\mathbf{z}^x|\mathbf{h}^x)$ is the probabilistic encoder, and $p_\theta(\mathbf{h}^x|\mathbf{z}^x)$ is the probabilistic decoder. The loss function of the VAE is defined as follows:

$$\begin{aligned} \mathcal{L}_{(\theta,\varphi;\mathbf{h}^x)} &= \mathbb{E}_{q_\varphi(\mathbf{z}^x|\mathbf{h}^x)}\left[\log p_\theta(\mathbf{h}^x \mid \mathbf{z}^x)\right] \\ &\quad - D_{\text{KL}}\left[q_\varphi(\mathbf{z}^x \mid \mathbf{h}^x) \parallel p(\mathbf{z}^x)\right], \end{aligned} \quad (16)$$

where $D_{\text{KL}}(\cdot)$ is the KL divergence. The total VAE loss is computed as:

$$\begin{aligned} \mathcal{L}_{vae} &= \mathcal{L}_{(\theta,\varphi;\mathbf{h}^a)} + \mathcal{L}_{(\theta,\varphi;\mathbf{h}^v)} + \mathcal{L}_{(\theta,\varphi;\mathbf{h}^r)} \\ &\quad + \mathcal{L}_{(\theta,\varphi;\mathbf{h}^{vr})} + \mathcal{L}_{(\theta,\varphi;\mathbf{h}^{ar})}, \end{aligned} \quad (17)$$

We minimize the distance between latent modality representations to improve the quality of missing modality generation:

$$\begin{aligned} \mathcal{L}_m = &\text{MSE}(\mathbf{z}^a, \mathbf{z}^{vr}) + \text{MSE}(\mathbf{z}^v, \mathbf{z}^{ar}) + \\ &\text{MSE}(\mathbf{z}^a, \mathbf{z}^r) + \text{MSE}(\mathbf{z}^v, \mathbf{z}^r), \end{aligned} \quad (18)$$

where $\text{MSE}(\cdot)$ denotes mean squared error. Through this operation, $\mathbf{z}^{vr}$, $\mathbf{z}^r$ is used to approximate $\mathbf{z}^a$, and $\mathbf{z}^{ar}$, $\mathbf{z}^r$ is used to approximate $\mathbf{z}^v$.

Inspired by (Zhang et al., 2019), we fuse pseudo-embeddings generated from different modalities and use them as the missing modality embeddings. For example, for the attribute modality we feed the latent representations $\mathbf{z}^{vr}$ and $\mathbf{z}^r$ into the decoder $p_\theta(\mathbf{h}^a|\mathbf{z}^a)$ to generate pseudo-embeddings $\mathbf{h}_1^a$ and $\mathbf{h}_2^a$, respectively. Subsequently, a dynamic weighted summation (Dsum) of the two pseudo-embeddings is performed to generate the missing attribute modality embedding:

$$\mathbf{g}^a = w_1 \cdot \mathbf{h}_1^a + w_2 \cdot \mathbf{h}_2^a, \tag{19}$$

4.5 Cross-modal Association Extraction

In MMEA, due to the heterogeneity between modalities, extracting latent associations between different modalities often presents challenges. At the same time, we observe that modality associations exist not only between individual modalities, but also between the fused modalities and individual modalities. On the basis of Multi-Modal Commonality Enhancement module of TMEA (Chen et al., 2024), we propose a Cross-modal Association Extraction module. In particular, we introduce the extraction of associations between fused modalities and individual modalities. That is, the proposed module not only focuses on associations among individual modalities but also considers those between fused and individual modalities, enabling a more comprehensive capture of latent cross-modal associations. We conduct experiments to validate the effectiveness of our improvements and demonstrate that latent associations also exist between fused and individual modalities.

For example, when performing Association Extracting (AE) between the attribute modality (*a*), the visual modality (*v*), and the relational modality (*r*), we not only consider associations between individual modalities but also the associations between the fused modality and individual modalities. Specifically, we concatenate and fuse the visual and relational modalities to generate a fused modality $\mathbf{h}^{vr}$, and then extract the latent associations between $\mathbf{h}^a$ and $\mathbf{h}^{vr}$. The specific definition of modality association extraction is as follows:

$$\begin{aligned}\hat{\mathbf{h}}^r &= w_r\mathbf{h}^r + w_v MH_{rv}(\mathbf{h}^v, \mathbf{h}^r) + w_a MH_{ra}(\mathbf{h}^a, \mathbf{h}^r)\\ &\quad + w_{av} MH_{rav}(\mathbf{h}^{av}, \mathbf{h}^r),\end{aligned} \tag{20}$$

$$\begin{aligned}\hat{\mathbf{h}}^a &= w_a\mathbf{h}^a + w_v MH_{av}(\mathbf{h}^v, \mathbf{h}^a) + w_r MH_{ar}(\mathbf{h}^r, \mathbf{h}^a)\\ &\quad + w_{rv} MH_{arv}(\mathbf{h}^{rv}, \mathbf{h}^a),\end{aligned} \tag{21}$$

$$\begin{aligned}\hat{\mathbf{h}}^v &= w_v\mathbf{h}^v + w_r MH_{vr}(\mathbf{h}^r, \mathbf{h}^v) + w_a MH_{va}(\mathbf{h}^a, \mathbf{h}^v)\\ &\quad + w_{ra} MH_{vra}(\mathbf{h}^{ra}, \mathbf{h}^v),\end{aligned} \tag{22}$$

where $MH_{*\#}(\mathbf{h}^*, \mathbf{h}^\#)$ represents the multi-head attention output between modality $\mathbf{h}^*$ and $\mathbf{h}^\#$ for extracting modality associations. The weights w_* are learnable attention fusion weights, $\mathbf{h}^{xy}$ is the concatenated fusion of the other two modalities.

To enhance association extraction, we introduce an orthogonality loss to ensure that the residuals are uncorrelated with the original modalities. Taking $MH_{rv}(\mathbf{h}^v, \mathbf{h}^r)$ as an example, it represents the modality association between $\mathbf{h}^v$ and $\mathbf{h}^r$. The corresponding residual $\mathbf{h}^r - MH_{rv}(\mathbf{h}^v, \mathbf{h}^r)$, which reflects the part uncorrelated with $\mathbf{h}^v$, should be as orthogonal as possible to $\mathbf{h}^v$. The orthogonality loss is defined as:

$$\begin{aligned}\mathcal{L}_s = \; & S_{rth}(S_d(\mathbf{h}^r, \mathbf{h}^v), \mathbf{h}^v) + S_{rth}(S_d(\mathbf{h}^r, \mathbf{h}^a), \mathbf{h}^a) +\\ & S_{rth}(S_d(\mathbf{h}^a, \mathbf{h}^r), \mathbf{h}^r) + S_{rth}(S_d(\mathbf{h}^a, \mathbf{h}^v), \mathbf{h}^v) +\\ & S_{rth}(S_d(\mathbf{h}^v, \mathbf{h}^r), \mathbf{h}^r) + S_{rth}(S_d(\mathbf{h}^v, \mathbf{h}^a), \mathbf{h}^a),\end{aligned} \tag{23}$$

$$S_{rth}(\mathbf{x}, \mathbf{z}) = (\mathbf{x}\mathbf{z}^T)^2, \tag{24}$$

$$S_d(\mathbf{h}^*, \mathbf{h}^\#) = \mathbf{h}^* - MH_{*\#}(\mathbf{h}^\#, \mathbf{h}^*), *, \# \in \{r, a, v\}, \tag{25}$$

where $S_d(\cdot)$ is the difference computation function, $S_{\mathrm{rth}}(\cdot)$ is the orthogonality constraint function, and $\mathcal{L}_s$ serves as the overall loss function of this module, aiming to extract latent cross-modal associations and enhance modality embeddings. Detailed analysis of $\mathcal{L}_s$, please refer to the Appendix D.

4.6 Modal Fusion and Optimization

We obtain the final entity embedding by directly concatenating (Cat) the modality embeddings, and adopt a multimodal contrastive learning (Chen et al., 2022a) for optimization. The definitions are as follows:

$$\mathbf{h}^M = Concat(\hat{\mathbf{h}}^v, \hat{\mathbf{h}}^r, \hat{\mathbf{h}}^a), \tag{26}$$

$$\begin{aligned}\mathcal{L}_{mcl} = \; & \mathcal{L}_{cl}(\mathbf{h}_1^M, \mathbf{h}_2^M) + \mathcal{L}_{cl}(\mathbf{h}_1^r, \mathbf{h}_2^r)\\ & + \mathcal{L}_{cl}(\mathbf{h}_1^a, \mathbf{h}_2^a) + \mathcal{L}_{cl}(\mathbf{h}_1^v, \mathbf{h}_2^v),\end{aligned} \tag{27}$$

$$\mathcal{L}_{all} = \mathcal{L}_{TransE} + \mathcal{L}_{vae} + \alpha\mathcal{L}_m + \beta\mathcal{L}_s + \mathcal{L}_{mcl}, \tag{28}$$

where $\mathbf{h}^M$ denotes the final entity embedding, $\mathbf{h}_1^*$ and $\mathbf{h}_2^*$ represent the sets of entity embeddings for modality $*$ in KGs $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively. For a more detailed explanation of the optimization, please refer to Appendix E.

5 Experiment

In this section, we conduct extensive experiments on two real-world datasets. We first briefly introduce the experimental settings, and then present and analyze the results to verify the effectiveness of the CLAMEA model. For more detailed experimental setup, please refer to Appendix F.

Method	FB15K-DB15K				FB15K-YG15K			
	Hits@1	Hits@5	Hits@10	MRR	Hits@1	Hits@5	Hits@10	MRR
EVA (Liu et al., 2021)	0.556	0.666	0.716	0.609	0.103	0.217	0.278	0.164
MSNEA (Chen et al., 2022a)	0.653	0.768	0.812	0.708	0.443	0.626	0.698	0.529
MCLEA (Lin et al., 2022)	0.441	0.640	0.710	0.534	0.406	0.579	0.645	0.488
ACK-MMEA (Li et al., 2023)	0.304	-	0.549	0.387	0.289	-	0.496	0.360
GEEA (Guo et al., 2023)	0.343	-	0.661	0.450	0.298	-	0.585	0.393
MEAformer (Chen et al., 2023)	0.578	-	0.812	0.661	0.444	-	0.692	0.529
PMF (Li et al., 2023)	0.624	-	-	0.702	0.543	-	-	0.620
DESAlign (Wang et al., 2024b)	0.580	-	0.815	-	0.448	-	0.713	0.541
IBMEA (Su et al., 2024a)	0.631	-	0.813	0.697	0.521	-	0.708	0.584
LoginMEA (Su et al., 2024b)	0.667	-	0.854	<u>0.735</u>	<u>0.622</u>	-	<u>0.818</u>	<u>0.691</u>
SimDiff (Li et al., 2024)	0.615	-	0.820	0.678	0.530	-	0.736	0.595
PMMEA (Tang and Wang, 2024)	0.645	-	0.780	0.691	0.561	-	0.716	0.614
PCMEA (Wang et al., 2024a)	0.6763	<u>0.8214</u>	0.8872	0.7280	0.5896	<u>0.7518</u>	**0.8347**	0.6460
TMEA (Chen et al., 2024)	<u>0.786</u>	-	<u>0.903</u>	-	0.593	-	0.757	-
Ours (CLAMEA)	**0.8176**	**0.8989**	**0.9225**	**0.8551**	**0.6319**	**0.7677**	0.8122	**0.6956**

Table 1: Performance comparison of multimodal entity alignment methods on two datasets using 20% of the aligned entity pairs. The best results are highlighted in bold, and the second-best are underlined.

5.1 Experiment Setup

5.1.1 Datasets and Evaluation Metrics

We evaluate the proposed model on two widely used monolingual datasets: FB15K-DB15K and FB15K-YAGO15K (Liu et al., 2019). We evaluate the model performance using Hits@N and Mean Reciprocal Rank (MRR) metrics. For detailed information, please refer to Appendix G.

5.1.2 Baseline Methods

Given the potential risk of data leakage (Song et al., 2024) during the training process of MMEA methods rely on large language models (LLMs), and for the sake of fairness, we compared CLAMEA with multimodal entity alignment methods that do not involve LLMs. These methods include EVA (Liu et al., 2021), MSNEA (Chen et al., 2022a), MCLEA (Lin et al., 2022), ACK-MMEA (Li et al., 2023), GEEA (Guo et al., 2023), MEAFormer (Chen et al., 2023), PMF (Li et al., 2023), DESAlign (Wang et al., 2024b), IBMEA (Su et al., 2024a), LoginMEA (Su et al., 2024b), SimDiff (Li et al., 2024), PMMEA (Tang and Wang, 2024), PCMEA (Wang et al., 2024a), TMEA (no LLMs involved version) (Chen et al., 2024). For all baselines, the experimental results are from their original papers.

5.2 Performance Comparison

We evaluate CLAMEA on FB15K-DB15K and FB15K-YG15K datasets and compare it with several representative and sota MMEA approaches. When using only 20% of the aligned entity pairs for training, the performance of each method is shown in Table 1. Our proposed CLAMEA demonstrates superior performance on both datasets. CLAMEA has made significant progress. In particular, CLAMEA achieves an absolute improvement of 3.1% in Hits@1 on FB15K-DB15K compared to the sota method. With only 20% of aligned entity pairs used for training, CLAMEA achieved impressive Hits@1 scores of 81.76% and 63.19% on the two datasets, respectively.

We observe that CLAMEA does not achieve sota performance on the Hits@10 metric for the FB15K-YG15K dataset. We speculate that this may be due to the small number of relations in the FB15K-YG15K dataset and the large number of candidate entities corresponding to the same relation, which makes many candidate entities present very close results in the ranking evaluation.

5.3 Ablation Analysis

To validate the effectiveness of modalities, modules, and components in the CLAMEA model, we conduct ablation experiments. We first remove specific modalities: relations (w/o R), visual (w/o V), and attributes (w/o A). Since there are interactions between modalities, we only remove the relevant modalities from the final modal fusion and loss function. We remove the following modules: MFE, DMAFF, DMMG, and CMAE. Additionally, we further remove key strategies and components within these modules: IT (without iterative strat-

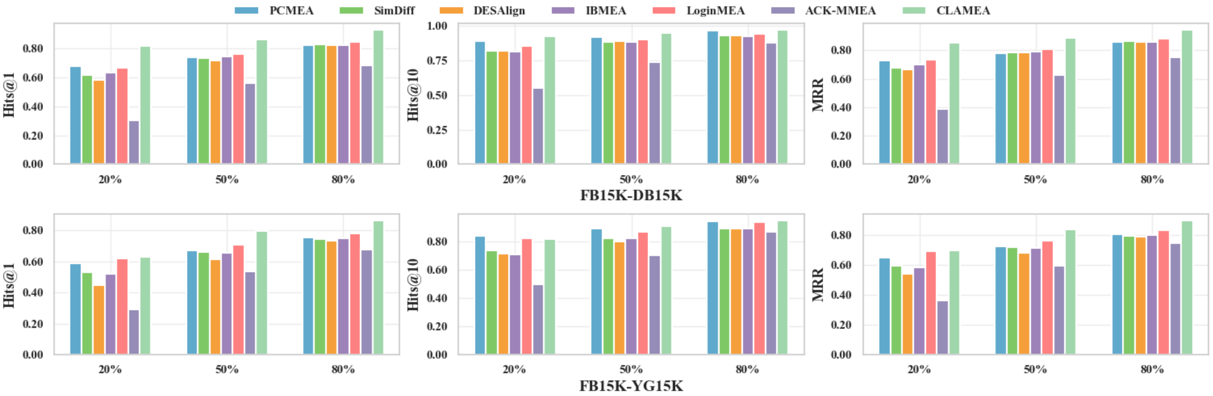

Figure 3: Comparison of different proportions of aligned entity pairs on FB15K-DB15K and FB15K-YG15K.

Method	FB15K-DB15K			FB15K-YG15K		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
CLAMEA	**0.8176**	**0.9225**	**0.8551**	**0.6319**	**0.8122**	**0.6956**
w/o R	0.5402	0.6491	0.5783	0.3352	0.4851	0.3873
w/o V	0.3623	0.6218	0.4487	0.4248	0.6602	0.5035
w/o A	0.7867	0.8984	0.8261	0.5189	0.7031	0.5831
w/o MFE	0.8089	0.9202	0.8475	0.6081	0.7913	0.6719
w/o DMAFF	0.7924	0.9089	0.8338	0.5871	0.7739	0.6524
w/o DMMG	0.8051	0.9154	0.8443	0.6138	0.7999	0.6782
w/o CMAE	0.6998	0.8384	0.7466	0.4639	0.6485	0.5272
w/o IT	0.6926	0.8436	0.7465	0.5236	0.7239	0.5923
w/o MAF	0.8013	0.9158	0.8395	0.5913	0.7807	0.6578
w/o $\mathcal{L}_m$	0.8115	0.9170	0.8492	0.5965	0.789	0.6643
w/o $\mathcal{L}_s$	0.8105	0.9187	0.8492	0.6128	0.7944	0.6782
w/o G	0.8105	0.9143	0.8402	0.5845	0.7766	0.6499

Table 2: Ablation studies to evaluate the impact of module components across two datasets.

egy), MAF (remove Modal Attention Flow), $\mathcal{L}_m$ and $\mathcal{L}_s$ (remove loss functions), G (removing the gating mechanism in DMAFF). As shown in Table 2, we find that removing the MFE module leads to a performance drop on both datasets, which indicates that the module effectively enhances modal embeddings. After removing the DMAFF module, the performance dropped more significantly, further validating that the DMAFF effectively captures and balances modal associations, thereby improving the quality of modal embeddings. The removal of the DMMG module also results in performance degradation, suggesting that missing modality information has a significant impact on alignment tasks. Similarly, removing the CMAE module causes a notable performance decrease, further confirming the importance of modal association. It also verifies that associations exist between the fused modalities and individual modalities, and extracting these associations can effectively improve the quality of modality embeddings. We provide a detailed experimental analysis of CMAE compared to TMEA to demonstrate the efficiency of CMAE. Please refer to Appendix H.

When iterative strategy (w/o IT) is not used, the model's robustness decreases significantly, further validating the importance of this strategy. Removing the modal attention flow component (w/o MAF) also led to performance degradation, indicating the effectiveness of this component. It also shows that this component can effectively capture the latent modal associations. Detailed analysis of the MAF component, please refer to Appendix I. Finally, removing the latent modality representation optimization loss function in the DMMG module (w/o $\mathcal{L}_m$), the orthogonality loss function in the CMAE module (w/o $\mathcal{L}_s$), and the gating mechanism in DMAFF (w/o G) all lead to varying degrees of performance degradation, which emphasizes the critical role of these components in the model.

At the same time, we conduct a detailed experimental analysis of the fusion strategies involved in the model to validate the correctness of our selection. Please refer to Appendix J. For analysis of model consumption and overfitting, please refer to the Appendix K.

5.4 Performance under Varying Ratios

To further evaluate the effectiveness of the proposed CLAMEA model, we conducted experiments on two datasets using 20%, 50%, and 80% of aligned entity pairs for training. To ensure experimental rigor, we only compared against models for which performance metrics under varying alignment ratios were explicitly reported in the papers.

As shown in the Figure 3, CLAMEA consistently outperforms across all training ratios. On the FB15K-DB15K dataset, CLAMEA achieves an MRR and Hits@1 score exceeding 80% under all training settings. Moreover, on the FB15K-YG15K dataset, CLAMEA significantly surpasses other methods in terms of Hits@1 and MRR under both 50% and 80% alignment settings. The results show that, except for the Hits10 metric on the 20% aligned entities of the FB15K-YG15K dataset, where CLAMEA does not achieve optimal performance, it achieved the best performance across all other alignment ratios.

6 Conclusion

This paper proposes a novel MMEA method called CLAMEA, which aims to address the issue of modality embedding quality in input modality interactions by capturing latent modal associations. Specifically, we design a dynamic modal attention flow fusion module to capture and balance the associations between modalities, a cross-modal association extraction module to further extract latent cross-modal associations, and a modal feature enhancement module and a dynamic missing modality generation module to enhance and generate modality embeddings, respectively. Extensive experiments on two datasets demonstrate the effectiveness of CLAMEA.

Limitations

CLAMEA has contributed to the development of MMEA, but there are still shortcomings, especially when extracting latent modal associations, which are easily influenced by the structure. For example, when using same training data proportions, there is a significant difference in the final results between FB15K-DB15K and FB15K-YAGO15K. Future research should focus on addressing structural issues like those in the FB15K-YAGO15K data and enrich the structure. For example, enriching KGs information through knowledge graph completion and replacing ineffective entity modal information can help improve alignment accuracy.

Acknowledgments

The authors sincerely thank the reviewers for their valuable comments, which improved the paper. The work is supported by the National Natural Science Foundation of China (62276057).

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Liyi Chen, Zhi Li, Tong Xu, Han Wu, Zhefeng Wang, Nicholas Jing Yuan, and Enhong Chen. 2022a. Multi-modal siamese network for entity alignment. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 118–126.

Liyi Chen, Ying Sun, Shengzhe Zhang, Yuyang Ye, Wei Wu, and Hui Xiong. 2024. Tackling uncertain correspondences for multi-modal entity alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2016. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. *arXiv preprint arXiv:1611.03954*.

Xu Chen, Siheng Chen, Jiangchao Yao, Huangjie Zheng, Ya Zhang, and Ivor W Tsang. 2022b. Learning on attribute-missing graphs. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(02):740–757.

Zhuo Chen, Jiaoyan Chen, Wen Zhang, Lingbing Guo, Yin Fang, Yufeng Huang, Yichi Zhang, Yuxia Geng, Jeff Z Pan, Wenting Song, and 1 others. 2023. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3317–3327.

Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Yin Fang, Jeff Z Pan, Ningyu Zhang, and Wen Zhang. 2022c. Lako: Knowledge-driven visual question answering via late knowledge-to-text injection. In *Proceedings of the 11th International Joint Conference on Knowledge Graphs*, pages 20–29.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6639–6648.

Lingbing Guo, Zhuo Chen, Jiaoyan Chen, Yin Fang, Wen Zhang, and Huajun Chen. 2023. Revisit and outstrip entity alignment: A perspective of generative models. *arXiv preprint arXiv:2305.14651*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Yves R Jean-Mary, E Patrick Shironoshita, and Mansur R Kabuka. 2009. Ontology matching with semantic verification. *Journal of Web Semantics*, 7(3):235–251.

Diederik P Kingma, Max Welling, and 1 others. 2013. Auto-encoding variational bayes.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Qian Li, Shu Guo, Yangyifei Luo, Cheng Ji, Lihong Wang, Jiawei Sheng, and Jianxin Li. 2023. Attribute-consistent knowledge graph representation learning for multi-modal entity alignment. In *Proceedings of the ACM Web Conference 2023*, pages 2499–2508.

Ran Li, Shimin Di, Lei Chen, and Xiaofang Zhou. 2024. Simdiff: Simple denoising probabilistic latent diffusion model for data augmentation on multi-modal knowledge graph. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1631–1642.

Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multi-modal contrastive representation learning for entity alignment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2572–2584.

Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual pivoting for (unsupervised) entity alignment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4257–4266.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019. Mmkg: multi-modal knowledge graphs. In *The semantic web: 16th international conference, ESWC 2019, portorož, Slovenia, June 2–6, 2019, proceedings 16*, pages 459–474. Springer.

Xin Mao, Wenting Wang, Huimin Xu, Man Lan, and Yuanbin Wu. 2020. Mraea: an efficient and robust entity alignment approach for cross-lingual knowledge graph. In *Proceedings of the 13th international conference on web search and data mining*, pages 420–428.

Jinghui Peng, Xinyu Hu, Wenbo Huang, and Jian Yang. 2023. What is a multi-modal knowledge graph: A survey. *Big Data Research*, 32:100380.

Yanqi Song, Ruiheng Liu, Shu Chen, Qianhao Ren, Yu Zhang, and Yongqi Yu. 2024. Securesql: Evaluating data leakage of large language models as natural language interfaces to databases. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5975–5990.

Taoyu Su, Jiawei Sheng, Shicheng Wang, Xinghua Zhang, Hongbo Xu, and Tingwen Liu. 2024a. Ibmea: Exploring variational information bottleneck for multi-modal entity alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4436–4445.

Taoyu Su, Xinghua Zhang, Jiawei Sheng, Zhenyu Zhang, and Tingwen Liu. 2024b. Loginmea: Local-to-global interaction network for multi-modal entity alignment. In *ECAI 2024*, pages 1173–1180. IOS Press.

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

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

Wei Tang and Yuanyi Wang. 2024. Multi-modal entity alignment via position-enhanced multi-label propagation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 366–375.

Xiaobin Tang, Jing Zhang, Bo Chen, Yang Yang, Hong Chen, and Cuiping Li. 2020. Bert-int: A bert-based interaction model for knowledge graph alignment. *interactions*, 100:e1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, and 1 others. 2017. Graph attention networks. *stat*, 1050(20):10–48550.

Luyao Wang, Pengnian Qi, Xigang Bao, Chunlai Zhou, and Biao Qin. 2024a. Pseudo-label calibration semi-supervised multi-modal entity alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9116–9124.

Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 950–958.

Yuanyi Wang, Haifeng Sun, Jiabo Wang, Jingyu Wang, Wei Tang, Qi Qi, Shaoling Sun, and Jianxin Liao. 2024b. Towards semantic consistency: Dirichlet energy driven robust multi-modal entity alignment. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 3559–3572. IEEE.

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

Yifei Zhang, Olivier Morel, Marc Blanchon, Ralph Seulin, Mojdeh Rastgoo, and Désiré Sidibé. 2019. Exploration of deep learning-based multimodal fusion for semantic road scene segmentation. In *VISAPP 2019 14Th international conference on computer vision theory and applications*.

Bin Zhu, Meng Wu, Yunpeng Hong, Yi Chen, Bo Xie, Fei Liu, Chenyang Bu, and Weiping Ding. 2023. Mmiea: Multi-modal interaction entity alignment model for knowledge graphs. *Information Fusion*, 100:101935.

Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Iterative entity alignment via joint knowledge embeddings. In *IJCAI*, volume 17, pages 4258–4264.

A Entity Alignment

As knowledge graphs (KGs) expand in real-world applications, research has gradually shifted from rule-based matching methods (Jean-Mary et al., 2009) to more expressive embedding-based approaches (Chen et al., 2016; Zhu et al., 2017; Bordes et al., 2013; Sun et al., 2018). These methods map entities in KGs to vector spaces, ensuring that semantically similar entities are closer in geometric space. Among these approaches, models based on relation translation mechanisms leverage the structural path properties between entities and emphasize the geometric similarity of aligned entities during optimization (Chen et al., 2016; Bordes et al., 2013). Graph neural network (GNN)-based methods (Wang et al., 2018; Sun et al., 2020; Velickovic et al., 2017; Kipf and Welling, 2016; Chen et al., 2023) further explore the neighborhood information of entities, enhancing their contextual representations through multi-level aggregation of graph structures. However, these methods struggle to accommodate diverse modal knowledge.

B Modal Embedding

B.1 Relational Embedding

The form of the relational modality is $(h_h, r, h_t) \in T^R$. We use the TransE (Bordes et al., 2013) model to encode the relational modality. The score function $f(h_h, r, h_t)$ and the margin-based loss function $\mathcal{L}_{TransE}$ are defined as follows:

$$f(h_h, r, h_t) = \|\mathbf{h}_h^r + \mathbf{r} - \mathbf{h}_t^r\|_2^2, \tag{29}$$

$$\mathcal{L}_{TransE} = \textstyle\sum_{\tau \in T^R} \sum_{\tau^- \in T_R^-} \max\left(0, \gamma + f(\tau) - f(\tau^-)\right), \tag{30}$$

where $\mathbf{h}_h^r$ and $\mathbf{h}_t^r$ are initial relational feature vectors of the head and tail entities, respectively, and $\mathbf{r}$ is the embedding of the relation r. $\|\cdot\|_2$ is the L2 norm, γ is the margin hyperparameter, and T_R^- is the set of negative examples. In this way, we obtain the entity relational modality embedding $\mathbf{h}_o^r$.

B.2 Visual Embedding

We use the pre-trained visual model ResNet (He et al., 2016) to process the image features x^v to obtain the visual embedding. It can be formulated as:

$$\mathbf{h}_o^v = \mathbf{W}_v \cdot \mathbf{x}^v + \mathbf{b}_v, \tag{31}$$

where $\mathbf{W}^v$ is the weight matrix for the linear transformation of the features, and $\mathbf{b}^v$ is the bias term.

B.3 Attribute Embedding

We convert attribute triples into corresponding short sentences (Chen et al., 2024). For example, given two attribute triples related to “Trump” — (“Trump”, “birthday”, “1946-6-14”) and (“Trump”, “age”, “79”) — we construct the short sentences H: “birthday is 1946-6-14, age is 79”.

Then, we use a pre-trained BERT (Devlin et al., 2019; Tang et al., 2020) model to encode these short sentences, thereby forming the attribute embedding of the corresponding entity. The computation is as follows:

$$\mathbf{h}_o^a = \mathbf{W}_a \cdot \frac{1}{n} \sum_{i=1}^{n} \mathrm{BERT}(w_i) + \mathbf{b}_a, \tag{32}$$

where w_i denotes the i-th word in the short sentence H, $\mathrm{BERT}(\cdot)$ represents the hidden feature vector from the last layer of BERT, $\mathbf{W}_a$ is the weight matrix, and $\mathbf{b}_a$ is the bias vector.

C Dynamic Modal Attention Flow Fusion Weighted

To generate the fused modality representation , we assign dynamic weights w_m to the modality features obtained from the Dynamic Intra-Modality Attention Flow module. These weights are dynamically calculated based on the interaction strength between modalities, defined as follows:

$$w_m = \frac{\exp(\sum\limits_{j \in \mathcal{M}} \sum\limits_{i=0}^{N_h} \beta_{mj}^{(i)} / \sqrt{|\mathcal{M}| \times N_h})}{\sum\limits_{k \in \mathcal{M}} \exp(\sum\limits_{j \in \mathcal{M}} \sum\limits_{i=0}^{N_h} \beta_{kj}^{(i)} \sqrt{|\mathcal{M}| \times N_h}),} \tag{33}$$

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

$$\beta_{mj} = \frac{\exp\left(\mathbf{Q}_m^\top \mathbf{K}_j / \sqrt{d_h}\right)}{\sum\limits_{n \in \mathcal{M}} \exp\left(\mathbf{Q}_m^\top \mathbf{K}_n / \sqrt{d_h}\right),} \tag{34}$$

where $\mathcal{M} \in \{a, r, v\}$ denotes the set of modalities, N_h is the number of attention heads, $d_h = d/N_h$, and $\beta_{mj}^{(i)}$ represents the interaction weight between modality m and j under the i-th attention head.

D CMAE Loss Function Analysis

It is worth noting that we did not include the fused modality in the loss function, as existing loss function has already effectively optimized the unimodal embeddings. We hypothesize that incorporating

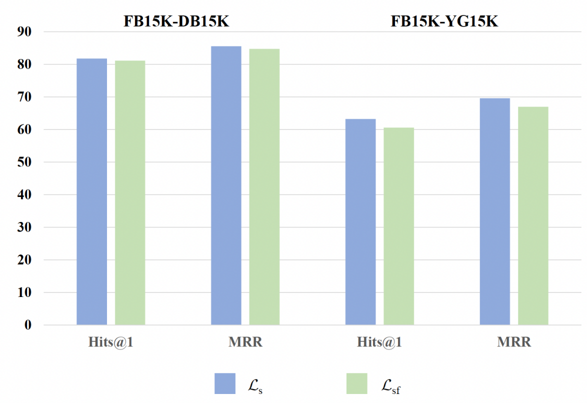

Figure 4: Comparison results of optimized the fused modality and non-optimized fused modality. $\mathcal{L}_\mathrm{s}$ denotes the loss function that does not optimize the fused modality, while $\mathcal{L}_\mathrm{sf}$ denotes the loss function that does

the fused modality into the loss function may lead to ineffective or even adverse optimization of each modality, ultimately affecting the overall performance of the model. To verify this, we conduct relevant experiments.

We conduct experiments on the FB15K-DB15K and FB15K-YG15K datasets using 20% of the aligned entity pairs. As shown in the Figure 4, $\mathcal{L}_\mathrm{s}$ consistently outperforms $\mathcal{L}_\mathrm{sf}$ on both datasets, which supports our hypothesis.

E Optimization

E.1 Multimodal Contrastive Learning Approach

The multimodal contrastive learning approach guides the model to learn by comparing the similarity between positive and negative entity pairs from multiple modality perspectives. Specifically, positive entity pairs refer to correctly aligned entity pairs, while negative entity pairs refer to misaligned entity pairs. The specific definition of contrastive representation learning is as follows:

$$\mathcal{L}_\mathrm{cl}(\mathbf{h}_1^*, \mathbf{h}_2^*) = \frac{1}{2|T|} \sum_{(\mathbf{h}^1, \mathbf{h}^2) \in T} \Big[(1-y) \cdot d^2(\mathbf{h}_1^*, \mathbf{h}_2^*) + y \cdot \max\left(\gamma_\mathrm{cl} - d(\mathbf{h}_1^*, \mathbf{h}_2^*),\, 0\right)^2 \Big], \quad (35)$$

where $* \in \{M, r, a, v\}$ represents the modality type—*M* represents the final fusion modality, *r* represents the relational modality, *a* represents the attribute modality, and *v* represents the visual modality, $\mathbf{h}_1^*$ and $\mathbf{h}_2^*$ represent the entity embeddings of modality $*$ in KGs $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively. T is the set of entity pairs including both positive and negative samples, $d(\cdot, \cdot)$ is the cosine similarity function, y indicates whether a given entity pair is aligned, and γ_{cl} is the margin hyperparameter used in contrastive learning.

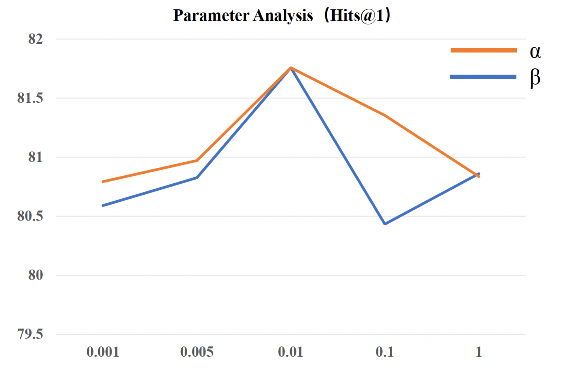

Figure 5: The results of the parameters on Hits@1 for the FB15K-DB15K

E.2 Optimization Strategy Description

To integrate all the designed loss functions, the final optimization objective $\mathcal{L}_{all}$ is defined as:

$$\mathcal{L}_{all} = \mathcal{L}_{TransE} + \mathcal{L}_{vae} + \alpha \mathcal{L}_m + \beta \mathcal{L}_s + \mathcal{L}_{mcl}, \quad (36)$$

where α and β are hyperparameters. During the training phase, we minimize $\mathcal{L}_{all}$ and update the model parameters via backpropagation.

Meanwhile, we adaopt a bidirectional iterative strategy (Mao et al., 2020) to enhance the effectiveness of model training.

E.3 Parameter Analysis

We introduced α and β in the objective function to regulate the influence of different loss terms. We experiment with five different values for each parameter and validate on 20% of the aligned entity pairs on the FB15K-DB15K datasets. As shown Figure 5, the model achieves optimal performance when both α and β are set to 0.01. This indicates that excessively large or small loss weights significantly affect the model's performance.

F Experimental Details

The dimensionality of all entity modality features is set to 100. For modality feature enhancement, variational autoencoder (VAE), and cross-modal association extraction, the number of attention heads is set to 2. The latent representation dimension of the VAE is 64. For the dynamic weight calculation in attention, the number of heads N_h is set to 5. In $\mathcal{L}_{TransE}$, we set $\gamma = 1$, and in $\mathcal{L}_\mathrm{cl}$, $\gamma_\mathrm{cl} = 2$.

Dataset	KGs	Ent	Rel	Attr	Rel-Triples	Attr-Triples	Ent-Image	Ent-pairs
FB15K-DB15K	DB15K	12842	279	225	89197	48080	12837	12846
	FB15K	14951	1345	116	592213	29395	13444	
FB15K-YG15K	YG15K	15404	32	7	122886	23532	11194	11199

Table 3: The statistics of datasets FB15K-DB15K and FB15K-YG15K

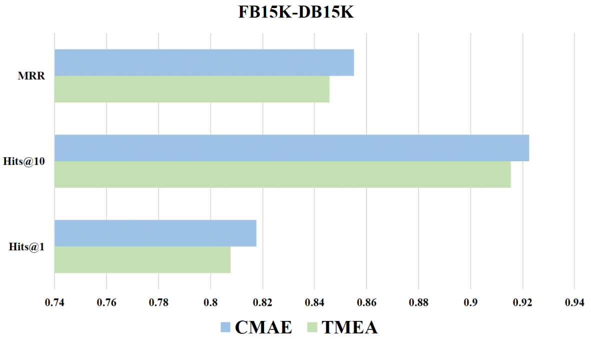

Figure 6: Performance comparison between improved CMAE and TMEA

We use the Adam optimizer with a learning rate of 0.001 and a batch size of 5000. All experiments are performed on an NVIDIA GeForce RTX 3090 GPU.

G Datasets and Evaluation Metrics

G.1 Datasets

Table 3 presents the detailed data statistics of the FB15K-DB15K and FB15K-YG15K datasets. These datasets cover different entity alignment ratios (20%, 50%, and 80%) to comprehensively assess the robustness of models under varying levels of supervision.

G.2 Evaluation Metrics

Hits@N measures the proportion of correctly aligned entities ranked in the top-N candidates, while MRR (Mean Reciprocal Rank) reflects the average ranking quality of the correctly aligned entities. Higher values of both metrics indicate better entity alignment performance.

H Analysis of Cross-modal Association Extraction

We further validate the effectiveness of the CMAE module while also verifying a conclusion we proposed. As mentioned earlier, CMAE is inspired by the Multi-Modal Commonality Enhancement module in TMEA (Chen et al., 2024), and improvements were made based on this. We conduct experiments on the FB15K-DB15K dataset using 20% of the aligned entity pairs, comparing the pre- and post-improvement modules. As shown in the Figure 6, the results show clear advantages across all metrics, validating the effectiveness of CMAE and confirming our conclusion that associations exist not only between individual modalities but also between fused and individual modalities.

I Further Analysis of Modal Attention Flow

To fully demonstrate the effectiveness of the Modal Attention Flow (MAF), We conducted experiments on both datasets, using 20%, 50%, and 80% of the aligned entity pairs respectively.The results, as shown in Figure 7, indicate that removing the attention flow mechanism leads to a significant performance drop. A clear decline in performance was observed across all datasets and evaluation metrics. This strongly suggests that the proposed dynamic modal attention flow mechanism can effectively capture and balance the complex associations of the modalities. Different modalities are capable of guiding information updates within each other. These results provide strong evidence for the effectiveness of MAF.

J Analysis Modal Fusion Strategies

We experimented with five specific fusion strategies using 20% of the aligned entity pairs on the FB15K-DB15K dataset: attention-based dynamic weighted concatenation (Dcat), static weighted concatenation (Wcat), direct concatenation (Cat), attention-based dynamic weighted summation (Dsum), and static weighted summation (Wsum), to evaluate the effectiveness of the data fusion strategies used in our model. The evaluation metrics used are Hits@1 and MRR. There are three modality fusions involved in the CLAMEA model: Attention Flow Fusion (AFF), Modality Generation Fusion (MGF), and Final Modal Fusion (FMF). In both AFF and MGF, we used the Dsum strat-

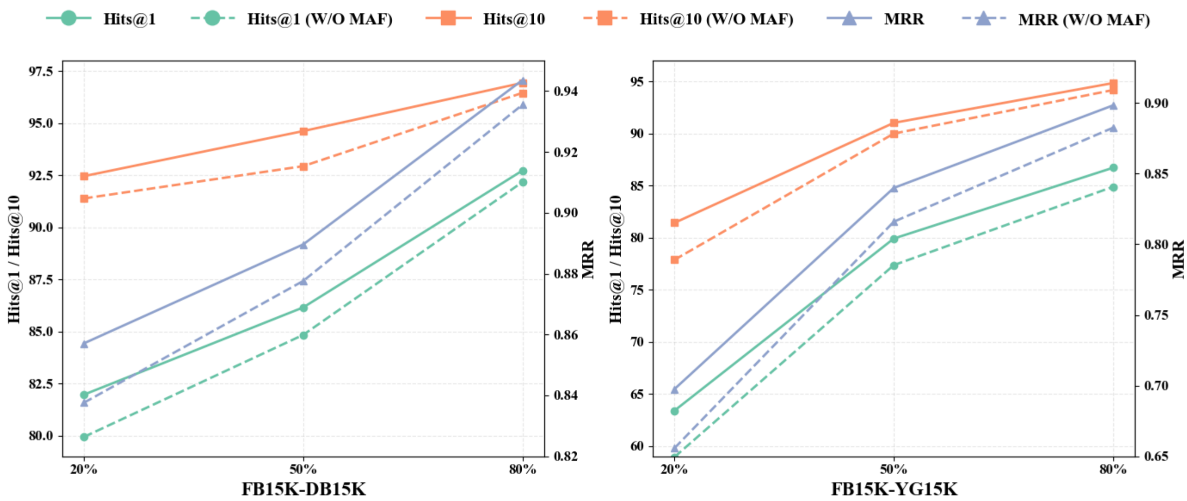

Figure 7: Demonstrating the impact of modal information flow on model performance across different ratios of training data

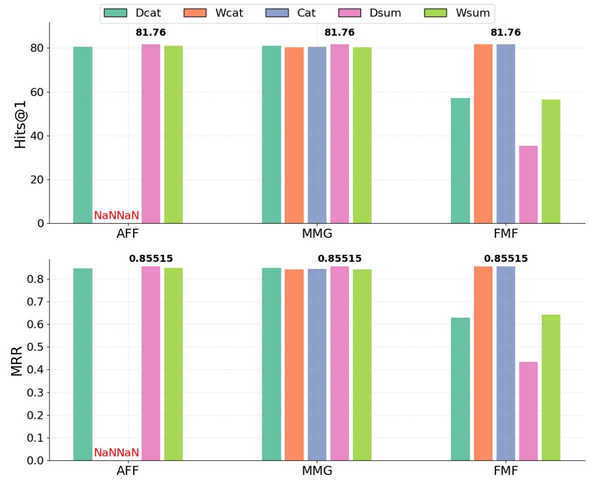

Figure 8: Results of Different Fusion Strategies

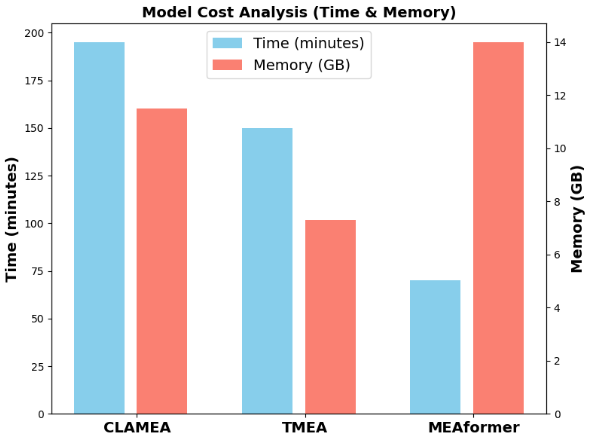

Figure 9: Time and memory comparison of different models

egy. In FMF, we adopted the Cat strategy. As shown in Figure 8, the Dsum strategy achieved the best performance in both the AFF and MGF. In the FMF, the Cat strategy exhibited the highest performance. Comparative analysis shows that concatenation is generally more effective in generating the final fused entity representation than summation. We speculate that the summation strategy may mask the unique features of each modality, leading to information redundancy or interference, which negatively affects the quality of the final entity embedding and thus the alignment performance.

From the Figure 8, we observe that when the fusion strategy in the AFF module is replaced with Wcat or Cat, the code encounters NaN issues and fails to run. We conduct an analysis and found that this problem does not occur when the learning rate is sufficiently low or when gradient clipping is applied. Based on this observation, we judge that the NaN issue may be caused by gradient explosion under high learning rates.

We do not use too low learning rate and gradient clipping because too low learning rate or using gradient clipping will significantly limit the model optimization speed, increase training time, and may cause the model to fall into local optimum, reducing model performance.

K Model Cost and Overfitting Analysis

We compare the time and memory consumption of the model with TMEA (Chen et al., 2024) and MEAformer (Chen et al., 2023). As shown in the

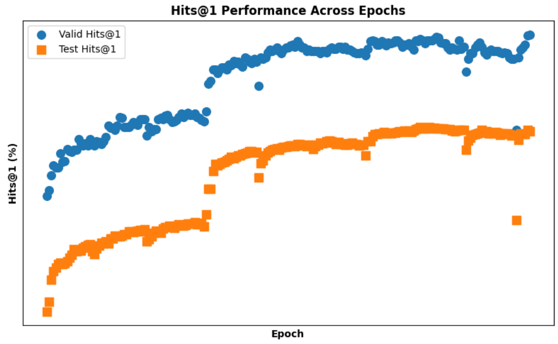

Figure 10: Hits@1 Performance on Validation and Test Sets

Figure 9, although CLAMEA consumes more time and memory than TMEA, it outperforms TMEA comprehensively in terms of model performance. Meanwhile, compared to MEAformer, CLAMEA requires less memory and achieves absolute improvements of 23.96% and 18.79% in Hits@1 on the two datasets, fully demonstrating the effectiveness of our model.

During the model training process, we use the MRR metric on the validation set to select and save the model with the best performance. As shown in the Figure 10. We observe that the variation trends of Hits@1 scores on the validation and test sets are generally consistent. Overfitting has never occurred during our experiments.