Multi-Modal Fact Knowledge Generation for Imbalanced Cross-Source Entity Alignment

Qian Li[1], **Cheng Ji**[2*], **Zhaoji Liang**[3,4], **Yuzheng Zhang**[1], **Zhuo Chen**[5], **Siyuan Liang**[1]
[1]School of Computer Science, Beijing University of Posts and Telecommunications
[2]Zhongguancun Laboratory
[3]Computer Network Information Center, Chinese Academy of Sciences
[4]University of Chinese Academy of Sciences
[5]Zhejiang University
{li.qian,zhangyuzheng,liang.siyuan}@bupt.edu.cn, jc@buaa.edu.cn, zjliang@cnic.cn, zhuo.chen@zju.edu.cn

Abstract

Multi-modal imbalanced cross-source entity alignment aims to identify equivalent entity pairs across multi-modal knowledge graphs (MMKGs) that encompass diverse data sources with imbalanced modality, which poses significant challenges due to the non-uniform distribution of information across different modalities. Existing methods encounter major limitations in aligning entities across MMKGs, where missing data and modality-specific inconsistencies thus create information gaps. These gaps, stemming from disparities in neighborhood structure and attribute availability, result in reduced alignment performance. To address these challenges, we propose a novel multi-modal fact knowledge generation framework to advance imbalanced cross-source entity alignment. Utilizing large language models (LLMs) for comprehensive knowledge completion, our framework enriches MMKGs by synthesizing missing neighboring entities and relational attributes, enabling precise one-to-one similarity comparisons across all relations and attributes. Specifically, neighbor entity completion generates probable neighboring entities to fill structural gaps, while attribute completion synthesizes missing relational attributes to improve alignment. The facts evaluation module assesses generated triples, ensuring that only high-quality information supports the alignment. Extensive experiments on benchmark datasets demonstrate that our framework significantly outperforms strong competitors, achieving superior entity alignment performance.

1 Introduction

Multi-modal knowledge graphs (MMKGs) are widely used for organizing and representing structured multi-modal information (Yuan et al. 2025). Due to the complexities of real-world data (Fang et al. 2023; Fang, Fang, and Wang 2025), MMKGs often suffer from missing information, making it difficult to align entities across multiple MMKGs. Multi-modal entity alignment (Liu et al. 2019; Li et al. 2023c; Liu et al. 2021a; Lin et al. 2022a; Chen et al. 2023a; Li et al. 2024; Yuan et al. 2023), which identifies equivalent entity pairs across multiple knowledge graphs that feature different modalities of attributes, such as text and images, plays a crucial role in integrating and consolidating knowledge from diverse sources. To accomplish this task, sophisticated models are required to effectively leverage information from different modalities and accurately align entities. This task is essential for various applications, such as cross-lingual information retrieval, question answering (Antol et al. 2015; Shih, Singh, and Hoiem 2016), and recommendation systems (Sun et al. 2020; Xu et al. 2021). However, existing works often suffer from information gaps in the multi-modal imbalanced cross-source entity alignment (MICEA), which are caused by the non-uniform distribution of information across different modalities, where the same entity in different KGs may be described by different relationship neighbors, leading to increased difficulty in entity alignment.

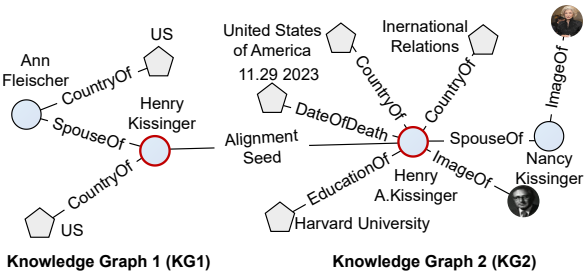


Figure 1: Example of the MICEA task, where the entity pair in KG1 and KG2 is the entity seed.

We argue that the information gap specifically refers to notable disparities between different modalities for MICEA. In other words, the challenge arises from the non-uniform distribution of information across different modalities. This issue can be elucidated through a straightforward example in Figure 1. Specifically, within KG1, the entity `Henry Kissinger` possesses only a singular textual attribute, rendering the determination of this entity challenging based solely on this attribute. In contrast, `Henry A. Kissinger` in KG2 incorporates richer attributes, such as `EducationOf` and `FieldOfWork`, facilitating a more comprehensive identification of the entity `Henry A. Kissinger`. Traditional aggregation-based methods, however, face the drawback of diluting the inherently sim-

*Corresponding Author.

ilar attribute information of two entities (between `US` and `United States of America`), thereby diminishing alignment performance. Furthermore, the presence of image information describing `Henry A. Kissinger` in KG2 but absent in KG1 complicates alignment efforts. Conventional approaches fail to leverage image information effectively to enhance the visual modality similarity between two entities, consequently impacting alignment efficacy. We categorize the challenge of inconsistent information as follows. (a) Incongruity in the quantity of attribute information. It makes alignment arduous when an entity in one KG has significantly more attributes. (b) Lack of modal attributes. It emphasizes the need for completeness in all modalities.

Previous methods (Liu et al. 2019; Chen et al. 2020; Guo et al. 2021) for entity alignment in MMKGs have not effectively addressed the challenge of information gaps between imbalanced cross-source KGs. Some approaches focus on using handcrafted rules or heuristics, which may not capture the complex relationships and attributes present in MMKGs. Other methods leverage embeddings or graph neural networks, but they struggle to capture the semantic meaning and context of entities and relationships in MMKGs. However, due to the presence of information gaps within MMKGs, this task becomes challenging. In particular, when the same entity is represented in different knowledge graphs, it may be associated with different relationship neighbors, which hampers the entity alignment process. Existing methods struggle to address this issue effectively.

To tackle these challenges, we propose a novel framework named **LLMEA**. Our framework can fill in the missing information gaps found in imbalanced cross-source KGs by completing missing neighbor entities and relationship attributes. We design prompts to generate descriptions of entities, including missing neighbors and attributes, ensuring that the structure of entity pairs in different graphs is completely aligned. We utilize LLM models to complete missing neighbor entities and relationship attributes in MMKGs. By leveraging the knowledge emergence capabilities of LLM models, we address the information gaps between knowledge graphs and enable comprehensive similarity comparisons of all relationships and attributes for each entity pair during the entity alignment process. To evaluate the generated information, we incorporate semantic consistency, structural consistency, confidence, and causal constraints to select triples that generalize well to unseen facts. To evaluate the effectiveness of our proposed approach, we design training objectives for both entity and context evaluation. Our contributions are summarized as follows:

- We propose a novel LLM-based framework for multi-modal, imbalanced cross-source entity alignment, leveraging LLMs' powerful knowledge-emergence capabilities to complete missing neighbor entities and attributes.
- We design an LLM-based knowledge completion and fact evaluation mechanism to fill in missing information gaps in MMKGs and evaluate the generated information.
- Extensive experiments on benchmark datasets demonstrate that our proposed framework outperforms strong competitors in MICEA performance.

2 Related Work

2.1 Multi-Modal Entity Alignment

Multi-modal entity alignment has garnered considerable attention due to the inherently multi-modal nature of KGs. Numerous approaches (Zhu et al. 2022; Wang, Li, and Gu 2021; Jiang, Li, and Gu 2021; Fang et al. 2022) have been proposed to advance multi-modal entity alignment, including embedding-based methods that represent entities and their associated modalities. However, this approach may not fully capture interactions between heterogeneous modalities, thereby limiting alignment accuracy. To address this limitation, researchers have proposed MMKG embedding techniques, such as the GNN-based model by Guo et al. (Guo et al. 2021), which aggregates information across modalities for entity alignment. EVA (Liu et al. 2021a) and UMAEA (Chen et al. 2023c) incorporate uncertainty and multi-scale modality hybridization to tackle challenges like overfitting and visual noise. MEAformer (Chen et al. 2023b) uses a hierarchical self-attention block and entity-type prefix injection to preserve semantics and integrate type information. DESAlign (Wang et al. 2024) applies Dirichlet energy to ensure semantic consistency and address over-smoothing. These models, along with ACK-MMEA (Li et al. 2023a), focus on bridging information gaps and improving alignment. Nevertheless, information gaps remain a significant challenge, impeding effective alignment.

2.2 Knowledge Graph Transformer

Transformer architecture, initially designed for NLP tasks, has been successfully applied to various KG tasks (Liu et al. 2022; Wang et al. 2023; Fang et al. 2024). KGAT (Wang et al. 2019) combines graph attention mechanisms with the Transformer to capture complex relationships between entities, facilitating tasks such as link prediction and entity recommendation. K-BERT (Liu et al. 2020a) extends this approach by pre-training a Transformer on a large corpus of textual data, followed by fine-tuning on a KG, thereby enhancing entity and relation extraction. ECEformer (Fang et al. 2024) leverages the Transformer architecture to address the challenge of inferring the evolution of temporal facts in temporal KGs. Transformers excel in modeling long-range dependencies, which is particularly beneficial when entities and their associated modalities are dispersed across the KG. The attention mechanisms inherent in Transformers help prioritize relevant information, making them highly effective for aligning entities across different modalities.

3 Preliminaries

Multi-modal Imbalanced Cross-source Entity Alignment (MICEA) is to identify equivalent entity pairs across different MMKGs, which are characterized by imbalanced modality distributions. Given two distinct MMKGs, $\mathcal{G}_1 = (\mathcal{E}_1, \mathcal{R}_1, \mathcal{A}_1, \mathcal{T}_1)$ and $\mathcal{G}_2 = (\mathcal{E}_2, \mathcal{R}_2, \mathcal{A}_2, \mathcal{T}_2)$, the goal is to determine whether two entities represent the same entity. This involves assessing the similarity between entity pairs, known as alignment seeds. The entities in each graph are represented by sets $\mathcal{E}_1$ and $\mathcal{E}_2$, with sizes $N_{\mathcal{E}_1}$ and $N_{\mathcal{E}_2}$, respectively. $\mathcal{R}_1$ and $\mathcal{R}_2$ are the sets of relations, with sizes

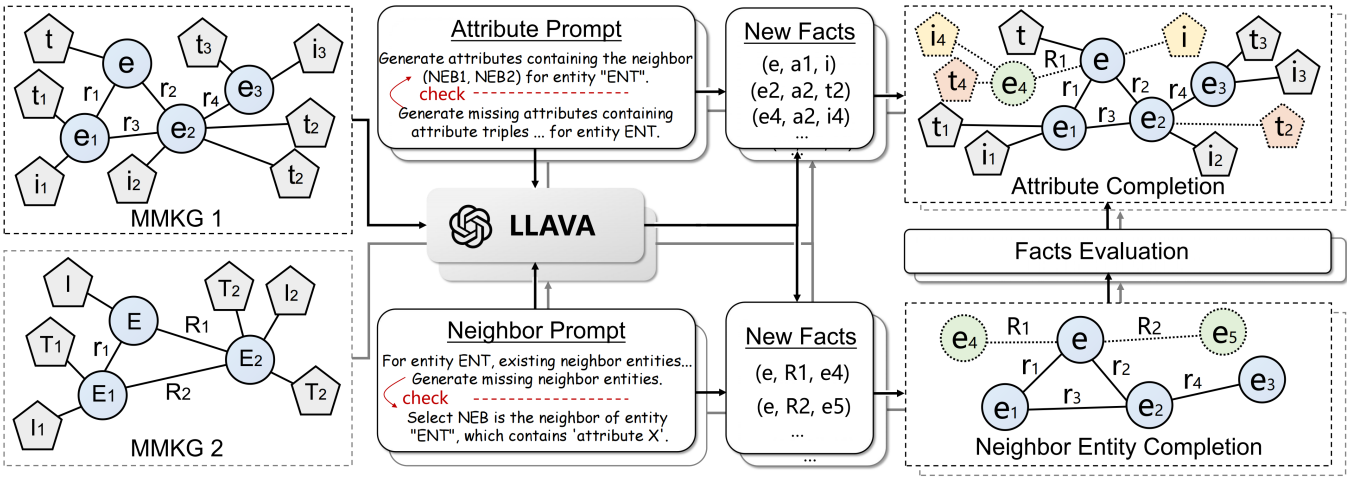


Figure 2: The framework of LLMEA. LLM-based knowledge completion addresses the problem of filling in missing information gaps in imbalanced cross-source KGs using LLM and evaluates the generation of information. Neighbor entity completion and attribute completion leverage the emergent capabilities of LLMs to fill information gaps, promoting a comprehensive comparison of relationships and attributes. A fact evaluation mechanism is incorporated to assess the generated information, ensuring the relevance, correctness, and consistency of the completed knowledge, enhancing the overall quality and reliability.

$N_{\mathcal{R}_1}$ and $N_{\mathcal{R}_2}$. These include: $\mathcal{R}_{1\mathcal{E}}$ and $\mathcal{R}_{2\mathcal{E}}$ are entity relations, $\mathcal{R}_{1T}$ and $\mathcal{R}_{2T}$ are text attribute relations, $\mathcal{R}_{1I}$ and $\mathcal{R}_{2I}$ are image attribute relations. $\mathcal{A}_1 = \mathcal{A}_{1T} \cup \mathcal{A}_{1I}$ and $\mathcal{A}_2 = \mathcal{A}_{2T} \cup \mathcal{A}_{2I}$ are the sets of multi-modal attributes: $\mathcal{A}_{1T}$ and $\mathcal{A}_{2T}$ are text attributes, $\mathcal{A}_{1I}$ and $\mathcal{A}_{2I}$ are image attributes. $\mathcal{T}_1$ and $\mathcal{T}_2$ are the sets of attribute triplets.

4 Framework

To address information gaps, we propose a novel framework utilizing LLMs to complete missing neighbor entities and relationships in imbalanced cross-source KGs, as shown in Figure 2.

4.1 Multi-Modal Fact Knowledge Generation

Multi-modal fact knowledge generation is vital to our MICEA framework, comprising neighbor entity completion and attribute completion. Neighbor entity completion fills in missing entities, while attribute completion adds missing relational attributes. Firstly, we devise two distinct positional encodings to preserve the structural integrity.

Modality Positional Encoding (MPE). To enable the model to effectively distinguish between entities, textual attributes, image attributes, and introduced entity types, we incorporate a unique position code for each modality. These position codes are then processed through the encoding layers of the model to enable it to differentiate between different modalities more accurately and learn their respective features more effectively. The multi-modal positional encoding is defined as:

$$\text{MPE} = \text{MLP}([\boldsymbol{e}; \boldsymbol{a}; \text{mod}_i]), \tag{1}$$

where $[;]$ denotes the concatenation operation, $\boldsymbol{e}$ is the entity, $\boldsymbol{a}$ is the attribute, i is the modality index (1-4 for entities, textual attributes, image attributes and entity types), and mod_i is the corresponding modality position code.

Structure Positional Encoding (SPE). To capture the positional information of neighbor nodes, we introduce a structure positional encoding that assigns a unique position code to each neighbor. For first-order neighbors, we randomly initialize a reference order and use it to assign position codes to each neighbor and its corresponding relation as $2n$ and $2n+1$, respectively. Additionally, we assign the same structure positional encoding of attributes to their corresponding entities.

$$\text{SPE} = \text{MLP}([\boldsymbol{e}; \boldsymbol{r}; \text{str}_n]), \tag{2}$$

where $\boldsymbol{r}$ is the relation, n is the neighbor index, and str_n is the structure position code. The initial entity, relation, and attribute embeddings are defined as follows:

$$\boldsymbol{e} = [\boldsymbol{e}; \text{PE}], \boldsymbol{r} = [\boldsymbol{r}; \text{SPE}], \boldsymbol{a} = [\boldsymbol{a}; \text{MPE}], \tag{3}$$

where $\text{PE} = \text{MLP}([\text{MPE}; \text{SPE}])$. To fully utilize the available information, we extract the relation triplets and the multi-modal attribute triplets for each entity. The entity representation is formed by combining the multi-modal sequences as follows:

$$\text{I}_e = [\boldsymbol{e}; (\boldsymbol{e}_1, \boldsymbol{r}_1); \ldots; (\boldsymbol{e}_n; \boldsymbol{r}_n); (\boldsymbol{a}_1; \boldsymbol{v}_1); \ldots; (\boldsymbol{a}_m; \boldsymbol{v}_m); \boldsymbol{e}T], \tag{4}$$

where $(\boldsymbol{e}_i; \boldsymbol{r}_i)$ represents the i-th neighbor of entity $\boldsymbol{e}$ and its relation. $(\boldsymbol{a}_j; \boldsymbol{v}_j)$ represents the j-th attribute of entity e and its value $\boldsymbol{v}_j$, and it contains textual and visual attributes. n and m are the numbers of neighbors and attributes. $\boldsymbol{e}_T$ is the type embedding.

Neighbor Entity Generation. To effectively complete missing neighbor entities in MMKGs, we utilize LLM and design a prompt that guides the model in generating the most probable missing entities based on contextual understanding. The neighbor entity completion can be expressed as:

$$\hat{\boldsymbol{e}}_{\text{neighbor}} = \hat{\mathcal{N}}(\boldsymbol{e}) = \text{LLM}(\boldsymbol{e}, \mathcal{N}(\boldsymbol{e}), \text{P}_{\text{entity}}), \tag{5}$$

where $\mathcal{N}(\boldsymbol{e})$ denote the set of neighbors for entity $\boldsymbol{e}$, $\hat{\mathcal{N}}(\boldsymbol{e})$ is the completed neighbor set, and $\mathrm{P}_{\text{entity}}$ is a prompt guiding the model to generate context-aware descriptions for missing neighbors. To complete the missing neighbor entities for entity `ENT` with existing neighbor entities `NEB1` and `NEB2`, the prompt $\mathrm{P}_{\text{entity}}$ concatenates the following components:

"For `ENT`, existing neighbors `NEB1` and `NEB2`. Please generate missing neighbor entities."

This prompt provides the necessary context for the LLM to understand the relationship between entity `ENT` and its existing neighbors and generate a completion that represents the most relevant missing neighbor entity. To handle missing relations explicitly and further improve the accuracy of entity alignment, we design a prompt $\mathrm{P}_{\text{check}}$ that instructs LLM to generate descriptions:

"Generate a description for "attribute X" between `ENT` and `NEB`. If it does not exist, return None."

In this prompt, we explicitly instruct the LLM to differentiate between existing and non-existing relations while generating the completion. We pass it through an LLM and generate a completion that represents the most probable missing neighbor entity for entity `ENT`.

$$\hat{\boldsymbol{e}}_{\text{check}} = \mathrm{LLM}(\boldsymbol{e}, \boldsymbol{a}, \hat{e}_{\text{neighbor}}, \mathrm{P}_{\text{check}}), \tag{6}$$

where $\hat{\boldsymbol{e}}_{\text{check}}$ represents the neighbors after validating the existence of relations, ensuring that the generated completion is accurate and contextually appropriate.

However, a challenge arises when the generation-based method produces relations/entities that do not exist in the KG. To address this, we deploy the following two steps:

Candidate Pool Mechanism. To ensure the validity of newly generated entities and relations, we predefine the entity and relation candidate pool $\mathcal{E}_{\text{pool}} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\}$ and $\mathcal{R}_{\text{pool}} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_m\}$. These pools are constructed based on the existing knowledge graphs. The generated entity $\boldsymbol{e}_{\text{gen}}$ and relation $\boldsymbol{r}_{\text{gen}}$ must match the entity candidate pool $\boldsymbol{e}_{\text{gen}} \in \mathcal{E}_{\text{pool}}$ and relation candidate pool $\boldsymbol{r}_{\text{gen}} \in \mathcal{R}_{\text{pool}}$.

Confidence-Based Filtering. To ensure the reliability of the generated entities and relations, we evaluate a confidence score for each generated entity and relation. We assign a confidence score to each by $O_C(\boldsymbol{e}_{\text{gen}}) = \mathrm{LLM}(\mathrm{P}_{\text{filter}}, \boldsymbol{e}_{\text{gen}})$, and only accept those that surpass a certain threshold. The prompt $\mathrm{P}_{\text{filter}}$ is:

"The confidence score is to give a number between [0,1], such as it is 0.9 for triple (`ENT, REL1, NEB1`) and it is 0.1 for the triple (`ENT, REL2, NEB2`). Now give the following example a confidence score."

The generated neighbor entities and relationships must have a sufficiently high confidence level. A language model is used to calculate the confidence scores of generated entities and relationships, and only entities and relationships whose confidence scores exceed a threshold are accepted.

$$C\left(\boldsymbol{e}_{\text{gen}}, \boldsymbol{r}_{\text{gen}}\right) \geq \tau_{\text{confidence}}, \tag{7}$$

where C represents the confidence scoring function, $\tau_{\text{confidence}}$ is the confidence threshold.

Attribute Generation. Similar to neighbor entity completion, we utilize the contextual understanding capabilities of LLMs to complete missing relationship attributes in knowledge graphs. Given an entity and its existing relationship attributes, we design a prompt $\mathrm{P}_{\text{attribute}}$ that guides the LLM model to generate the most probable missing relationship attributes. Consider an entity (`ENT`) with an existing relationship (`ENT, ATT1, VAL1`) in a knowledge graph, where (`ATT1`) is the attribute and (`VAL1`) its value. To complete a missing attribute (`ATT2`), we formulate the prompt as:

"For `ENT`, existing triples (`ENT, ATT1, VAL1`), (`ENT, ATT2, VAL2`), complete missing triple."

The relationship attribute completion is:

$$\hat{\mathcal{A}}(\boldsymbol{e}_1, \boldsymbol{e}_2) = \mathrm{LLM}(\boldsymbol{e}_1, \boldsymbol{e}_2, \mathcal{A}(\boldsymbol{e}_1, \boldsymbol{e}_2), \mathrm{P}_{\text{attribute}}), \tag{8}$$

where $\hat{\mathcal{A}}(\boldsymbol{e}_1, \boldsymbol{e}_2)$ is the completed relationship attribute set, $\mathcal{A}(\boldsymbol{e}_1, \boldsymbol{e}_2)$ is the existing relationship attribute set of $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ Furthermore, to evaluate the generated attributes, we input their neighbor information and utilize the LLM to check the generated attributes $\hat{\mathcal{A}}_{\text{check}_a} = \mathrm{LLM}(\boldsymbol{e}_1, \hat{\mathcal{A}}(\boldsymbol{e}_1, \boldsymbol{e}_2), \mathcal{N}(\boldsymbol{e}_1), \mathrm{P}_{\text{check}_a})$, where $\mathrm{P}_{\text{check}_a}$ guides the model to differentiate between existing and non-existent relations. The prompt $\mathrm{P}_{\text{check}_a}$ as follows:

"`ENT` contains (`NEB1, NEB2`) and select attributes for `ENT`. If it does not exist, return None."

Leveraging the completed knowledge graphs, we compare the relationships and attributes of entities and use these similarity metrics to measure the relatedness between entity pairs. To align entities across diverse knowledge graphs, we compute the Jaccard similarity coefficient (Ji et al. 2013) between entities based on their attributes and relationships:

$$\mathrm{Jac}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \frac{|\mathrm{At}(\boldsymbol{e}_i) \cap \mathrm{At}(\boldsymbol{e}_j)| + |\mathrm{Re}(\boldsymbol{e}_i) \cap \mathrm{Re}(\boldsymbol{e}_j)|}{|\mathrm{At}(\boldsymbol{e}_i) \cup \mathrm{At}(\boldsymbol{e}_j)| + |\mathrm{Re}(\boldsymbol{e}_i) \cup \mathrm{Re}(\boldsymbol{e}_j)|}, \tag{9}$$

where $\mathrm{At}(\boldsymbol{e})$ and $\mathrm{Re}(\boldsymbol{e})$ are the sets of attributes and relationships for entities. Specifically, $|\mathrm{At}(\boldsymbol{e}_i) \cap \mathrm{At}(\boldsymbol{e}_j)|$ and $|\mathrm{Re}(\boldsymbol{e}_i) \cap \mathrm{Rel}(\boldsymbol{e}_j)|$ denote the number of common attributes and relationships. While $|\mathrm{At}(\boldsymbol{e}_i) \cup \mathrm{At}(\boldsymbol{e}_j)|$ and $|\mathrm{Re}(\boldsymbol{e}_i) \cup \mathrm{Re}(\boldsymbol{e}_j)|$ denote the unique attributes/relationships. The denominator represents the union of the attributes and relationships, capturing the total distinct elements involved. By calculating the Jaccard coefficient, we can quantify the similarity between entities from different knowledge graphs.

4.2 Facts Evaluation

The generated neighbor entities and relationships should supplement the existing KG to ensure its integrity:

$$\begin{aligned} \mathcal{G}_{\text{new}} = \mathcal{G} \cup \{(\boldsymbol{e}_{\text{gen}}, \boldsymbol{r}_{\text{gen}}, \boldsymbol{e}_{\text{existing}}) \mid \boldsymbol{e}_{\text{gen}} \in \mathcal{E}_{\text{pool}}, \\ \boldsymbol{r}_{\text{gen}} \in \mathcal{R}_{\text{pool}}, \mathcal{C}(\boldsymbol{e}_{\text{gen}}, \boldsymbol{r}_{\text{gen}}) \geq \tau_{\text{confidence}}\}, \end{aligned} \tag{10}$$

where $\mathcal{G}_{\text{new}}$ represents the updated knowledge graph, $\mathcal{G}$ represents the original knowledge graph, and $\mathcal{C}$ denotes the confidence function with a threshold $\tau_{\text{confidence}}$. Furthermore, we proposed fact evaluation to assess the generated information's relevance, correctness, and consistency through TransE evaluation, causal evaluation, and model editing evaluation, enhancing the quality and precision of the knowledge graph.

TransE Evaluation. To incorporate relevant and trustworthy information while discarding less reliable triples, we use the TransE (Bordes et al. 2013) to assign scores and only consider triples with scores higher than a threshold a for inclusion in $\mathcal{T}$, By employing the TransE scoring function to filter out non-factual triples, we can ensure that only reliable and plausible triples are added to the MMKG.

Causal Evaluation. To select the most informative and relevant predicted triples to enhance the performance of our framework, we consider the effect of incorporating different predicted triples into the large language model. We then measure the performance of the large language model in terms of entity alignment using these predicted triples. To measure the impact of each predicted triple, we calculate the Jaccard score between the entities involved in the triple, which is the ratio of the intersection of the entities in the predicted triple and in existing triples to their union.

$$S_c(\mathcal{T}_{gen}, \mathcal{T}_{exi}) = \frac{|\mathcal{T}_{gen} \cap \mathcal{T}_{exi}|}{|\mathcal{T}_{gen} \cup \mathcal{T}_{exi}|} \geq \tau_c, \tag{11}$$

where $S_c(\mathcal{T}_{gen}, \mathcal{T}_{exi})$ is the Jaccard score between the predicted triples and existing triples, and τ_c is the threshold. A higher similarity score indicates a stronger alignment between the predicted triples and the existing triples. By analyzing how well the model performs when incorporating various predicted triples, researchers can determine which predicted triples contribute most effectively to improving the alignment of entities within the KG.

Model Editing Evaluation. The generated neighbor entities and relationships must be structurally consistent with the existing knowledge graph, ensuring they conform to existing relationship patterns. All generated relationships must be in the predefined relationship candidate pool:

$$\forall \left(\boldsymbol{e}_{\text{gen}}, \boldsymbol{r}_{\text{gen}}, \boldsymbol{e}_{\text{existing}}\right), \boldsymbol{r}_{\text{gen}} \in \mathcal{R}_{\text{pool}}. \tag{12}$$

Model editing evaluation helps filter out irrelevant or incorrect triples and retain those most likely to generalize well to unseen facts. To evaluate the generated neighbor triples, we propose a scoring mechanism. Specifically, we calculate the similarity score between each triple in $\mathcal{G}$ and the existing triples in $\mathcal{T}$ using a similarity function.

$$S_m(t, \mathcal{T}) = \sum_{t' \in \mathcal{T}} \text{sim}(t, t') \geq \tau_m, \tag{13}$$

where $\text{sim}(t, t')$ measures the similarity between triples t and t'. After calculating the scores, we select the top-k triples with the highest scores.

4.3 Training Objective

We propose a training objective function that incorporates both aligned entity and context evaluation.

Aligned Entity Evaluation. In order to align entities from two knowledge graphs, an entity similarity constraint is commonly employed:

$$\mathcal{L}_i^{EA} = \text{sim}(\mathbf{e}_i, \mathbf{e}'_i) - \text{sim}(\mathbf{e}_i, \overline{\mathbf{e}}'_i) - \text{sim}(\overline{\mathbf{e}}_i, \mathbf{e}'_i). \tag{14}$$

The given objective describes the final representations of aligned seeds $(\mathbf{e}_i, \mathbf{e}'_i)$ from two knowledge graphs. These scores are obtained by negative samples, denoted by $\overline{\mathbf{e}}_i$, $\overline{\mathbf{e}}'_i$. The entity alignment loss $\mathcal{L}^{EA}$ measures the dissimilarity between the embeddings of the aligned entities in different knowledge graphs.

Aligned Attribute Evaluation. In order to maintain consistency in the similarity of attributes within adjacent entities of the same type, we have introduced an attribute similarity constraint loss, which is a crucial component in our proposed approach. Specifically, for each entity $\mathbf{e}_i$ in the input graph, the attribute similarity constraint loss (AL) is defined as follows:

$$\mathcal{L}_i^{attr} = \sum_{A \in \{T, I\}} \text{sim}(\mathbf{e}_{i,A}, \mathbf{e}'_{i,A}), \tag{15}$$

where A represents the textual T or visual I attributes, $\mathbf{e}_{i,T}$, $\mathbf{e}_{i,I}$ are the textual/visual attribute embeddings, and $\mathbf{e}'_{i,T}$, $\mathbf{e}'_{i,I}$ are their corresponding adjacent entities' attribute embeddings. The final joint objective is:

$$\mathcal{L} = \lambda_1 \mathcal{L}^{EA} + \lambda_2 \mathcal{L}^{attr}. \tag{16}$$

5 Experiment

Dataset. (1) Monolingual MMEA: FB15K-DB15K and FB15K-YAGO15K (Liu et al. 2019). FB15K-DB15K dataset comprises a total of 12,846 alignment seeds between the FB15K and DB15K MMKGs. FB15K-YAGO15K dataset includes 11,199 alignment seeds between the FB15K and YAGO15K MMKGs. (2) Bilingual MMEA: $\text{DBP15K}_{\text{ZH-EN}}$, $\text{DBP15K}_{\text{JA-EN}}$, and $\text{DBP15K}_{\text{FR-EN}}$ from DBP15K dataset (Liu et al. 2021b), which consists of three datasets built from the multilingual versions of DBpedia.

Comparison Methods. (1) Monolingual MMEA: EA (TransE (Bordes et al. 2013), GCN-align (Wang et al. 2018), and AttrGNN (Liu et al. 2020b)), Transformer entity alignment (TransEA) (BERT (Devlin et al. 2019), ViT (Dosovitskiy et al. 2021), and CLIP (Radford et al. 2021)), and MMEA (PoE (Liu et al. 2019), Chen et al. (Chen et al. 2020), HEA (Guo et al. 2021), EVA (Liu et al. 2021a), MSNEA (Chen et al. 2022), ACK-MMEA (Li et al. 2023a), MoAlign (Li et al. 2023b), MEAformer (Chen et al. 2023b), and DESAlign (Wang et al. 2024)). (2) Bilingual MMEA: SBootEA, NAEA (Zhu et al. 2019), MCLEA (Lin et al. 2022b), MMEA-cat (Lin et al. 2022b), UMAEA (Lin et al. 2022b), PMF (Huang et al. 2024).

Implementation Details. We conducted all the experiments on a server equipped with one Tesla V100 GPU to ensure a fair and consistent comparison, and our proposed model was implemented using PyTorch. We employed the bert-large-uncased encoder from the HuggingFace library, comprising 24 layers with an embedding size of 1024, while the image attributes were initialized using the VGG16. We performed a thorough hyper-parameter tuning process via grid search, with five trials on the validation set. The training epoch was set to 200, with an L2 regularization of 0.0001 and a margin gamma of 1.0.

Methods	FB-DB (20%)			FB-DB (50%)			FB-DB (80%)			FB-YAGO (20%)			FB-YAGO (50%)			FB-YAGO (80%)		
	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10
TransE	13.4	7.8	24.0	30.6	23.0	44.6	50.7	42.6	65.9	11.2	6.4	20.3	26.2	19.7	38.2	46.3	39.2	59.5
GCN-align	8.7	5.3	17.4	29.3	22.6	43.5	47.2	41.4	63.5	15.3	8.1	23.5	29.4	23.5	42.4	47.7	40.6	64.3
AttrGNN	34.3	25.2	53.5	54.7	47.3	72.1	70.3	67.1	83.9	31.8	22.4	39.5	46.2	38.0	63.9	67.1	59.9	78.7
BERT	32.6	24.3	48.0	49.6	45.2	67.9	65.3	64.5	80.1	30.5	23.6	39.0	48.7	43.1	62.4	67.3	60.8	81.2
ViT	33.5	25.1	53.9	50.5	45.5	69.0	71.5	66.8	85.7	32.4	26.8	44.9	52.0	45.7	67.5	71.3	63.1	82.0
CLIP	35.4	27.0	55.3	54.1	48.7	71.4	73.9	68.3	86.0	34.8	29.3	47.1	56.8	49.0	70.2	72.1	65.2	85.2
PoE	17.0	12.6	25.1	53.3	46.4	65.8	72.1	66.6	82.0	15.4	11.3	22.9	41.4	34.7	53.6	63.5	57.3	74.6
Chen et al.	35.7	26.5	54.1	51.2	41.7	70.3	68.5	59.0	86.9	31.7	23.4	48.0	48.6	40.3	64.5	68.2	59.8	83.9
HEA	-	12.7	36.9	-	26.2	58.1	-	41.7	78.6	-	10.5	31.3	-	26.5	58.1	-	43.3	80.1
EVA	35.2	28.9	54.5	53.8	45.3	72.9	71.6	63.5	85.1	33.5	25.0	46.2	56.1	47.8	68.3	72.5	64.0	84.5
MSNEA	23.2	14.9	39.2	45.9	35.8	65.6	65.1	56.5	81.0	21.0	13.8	34.6	47.2	37.6	64.6	66.8	59.3	80.6
ACK-MMEA	38.7	30.4	54.9	62.4	56.0	73.6	75.2	68.2	87.4	36.0	28.9	49.6	59.3	53.5	69.9	74.4	67.6	86.4
MoAlign	40.9	31.8	56.4	63.4	57.6	74.9	77.3	69.9	88.2	37.8	29.6	52.5	61.7	55.0	71.3	76.9	68.9	88.4
MEAformer	53.4	43.4	72.8	70.4	62.5	84.7	82.5	77.3	91.8	41.6	32.5	59.8	64.0	56.0	78.0	76.8	70.5	87.4
DESAlign	66.5	58.0	81.5	62.4	56.0	73.6	75.2	68.2	87.4	36.0	28.9	49.6	59.3	53.5	69.9	74.4	67.6	86.4
LLMEA	**69.2**	**61.3**	**82.7**	**72.5**	**64.1**	**87.3**	**85.1**	**79.2**	**92.5**	**42.6**	**34.7**	**60.4**	**66.0**	**58.1**	**78.7**	**77.5**	**71.2**	**89.6**

Table 1: Main experiments on FB15K-DB15K (left) and FB15K-YAGO15K (right) with different proportions of MMEA seeds.

Methods	$\text{DBP15K}_{\text{ZH-EN}}$			$\text{DBP15K}_{\text{JA-EN}}$			$\text{DBP15K}_{\text{FR-EN}}$		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
SBootEA	62.9	84.7	70.3	62.2	85.4	70.1	65.3	87.4	73.1
NAEA	65.0	86.7	72.0	64.1	87.3	71.8	67.3	89.4	75.2
EVA*	74.6	91.0	80.7	74.1	91.8	80.5	76.7	93.9	83.1
MSNEA*	64.3	86.5	71.9	57.2	83.2	66.0	58.4	84.1	67.1
MCLEA*	81.1	95.4	86.5	80.6	95.3	86.1	81.1	95.4	86.5
MMEA-cat	62.4	84.5	70.2	64.1	86.9	72.3	72.5	91.4	79.3
UMAEA	75.8	95.1	82.9	77.5	96.3	84.5	79.2	97.0	85.9
MoAlign	84.7	97.0	89.2	84.2	97.4	89.2	84.5	97.6	89.4
PMF	83.5	-	88.4	83.5	-	88.5	85.0	-	89.8
LLMEA	**85.2**	**97.6**	**89.8**	**85.1**	**97.9**	**89.7**	**85.2**	**98.3**	**90.0**

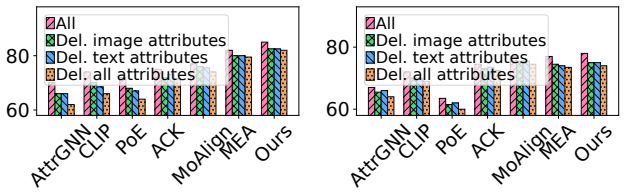
Table 2: Main experiments on the bilingual MMEA datasets.

Variants	FB15K-DB15K			FB15K-YOGA15K			$\text{DBP15K}_{\text{ZH-EN}}$		
	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10
LLMEA	**85.1**	**79.2**	**92.5**	**77.5**	**71.2**	**89.6**	**89.8**	**85.2**	**97.6**
w/o LKC	82.3	77.0	89.4	75.5	68.2	87.6	87.3	84.6	96.1
w/o NEC	83.4	78.2	90.5	75.1	69.2	88.7	86.4	84.0	96.8
w/o RAC	83.0	78.2	90.5	76.3	69.6	88.2	87.6	84.1	96.3
w/o AL	83.1	78.4	91.5	76.2	70.0	88.4	86.4	84.6	96.2
w/o TA	82.5	77.2	89.1	74.3	68.5	86.6	85.7	84.5	96.4
w/o IA	82.4	78.9	89.2	75.5	69.7	87.3	85.6	84.2	97.0

Table 3: Variant Experiments. "w/o" means removing the corresponding module from the complete model.

5.1 Main Results

To verify the effectiveness of our model, we report the overall results in Table 1 and 2. We can observe that: **1)** On FB15K-DB15K and FB15K-YAGO15K datasets, LLMEA performs best with varying seed splits. It demonstrates LLMEA's robustness and strong performance in few-shot learning scenarios. **2)** LLMEA shows a significant improvement in MRR over EA baselines on both datasets, highlighting the benefit of multi-modal context in enhancing alignment performance. **3)** Compared to multi-modal transformers, LLMEA consistently delivers superior results, suggesting that transformer architectures capture multi-modal information effectively. **4)** Against multi-modal entity alignment baselines, LLMEA achieves better performance in MRR, Hits@1, and Hits@10, respectively, owing to its LLM-based knowledge completion mechanism, which enhances performance over transformer-based entity alignment. All observations confirm the effectiveness of LLMEA.



(a) FB15K-DB15K. (b) FB15K-YOGA15K.

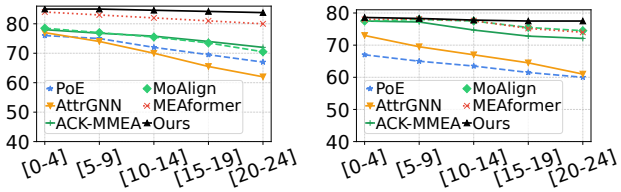
Figure 3: Results (MRR) of deleting attributes.

5.2 Ablation Study

From Table 3, we can observe that: **1)** The impact of LLM-based knowledge completion (LKC) tends to be more significant. The reason is that the consistent introduction of multi-modal attributes and neighbors captures more clues. **2)** By removing the neighbor entity or relationship attribute completion (NEC or RAC), the performance decreased significantly. It demonstrates that the completion captures more effective multi-modal information. **3)** When all image attributes (IA) are removed, LLMEA decreases 1.7% on average. This underscores that image attributes contribute to improving the performance, and LLMEA effectively leverages them to capture more alignment knowledge.

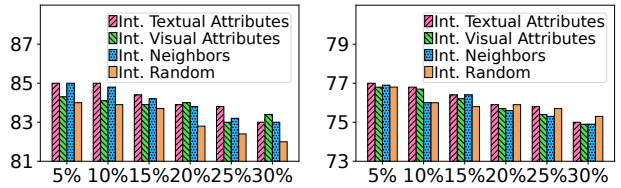
5.3 Impacts of Multi-Modal Attributes

Figure 3 provides several insightful observations: **1)** The variants that lack text or image exhibit a significant decline. The result highlights the importance and efficacy of multi-modal attributes in MMEA. **2)** Our model is less affected



(a) FB15K-DB15K. (b) FB15K-YOGA15K.

Figure 4: Results (MRR) of differences in attribute number.



(a) FB15K-DB15K. (b) FB15K-YOGA15K.

Figure 5: Interference with attributes or neighbors (MRR). "Int." means interference with some proportion of them.

by the removal of all multi-modal attributes, likely due to the utilization of LLM-based Knowledge Completion and the Attribute loss, which contribute to obtaining superior entity representations. **3)** LLEMA achieves superior results, regardless of whether all or some of the multi-modal attributes are abandoned when compared to others. It indicates that LLMEA is capable of fully utilizing multi-modal attributes.

5.4 Impact of the Number of Attributes

We examine the effects of varying degrees of information gap between EA seeds. We select alignment seeds with the same number of image attributes and vary the gaps of text number within the range of [0, 24]. From Figure 4, it is noticed that: **1)** The decline in performance becomes more pronounced for all methods as the gaps between alignment seeds widen. This difficulty arises from the increased challenge of matching entities when there are larger gaps between alignment seeds. **2)** Our model showcases a more gradual decline in performance, illustrating the superiority of our approach in mitigating the information gap.

5.5 Impact of Interference Data

To examine the influence of interference data, we conduct experiments where we randomly substitute part of the neighbor entities and attribute information, as illustrated in Figure 5. The experimental results highlight that our method displays superior tolerance to interference. This indicates that our approach, integrating completed information through LLM-based knowledge completion, effectively manages interference data and enhances the overall robustness of the model. Specifically, as the percentage of interference increases, our method shows a smaller degradation in performance. This robustness is attributed to the effective completion of missing entities and attributes, ensuring that

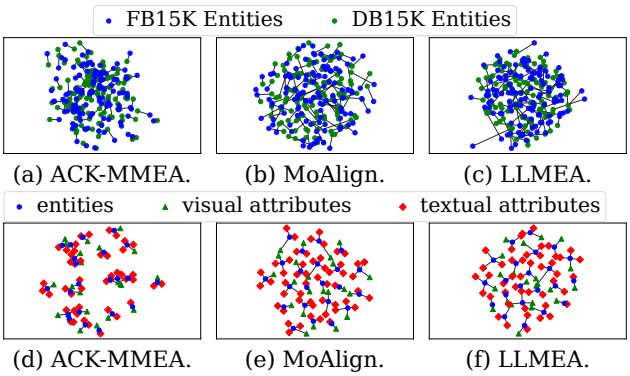


Figure 6: Visualization of embeddings and attributes.

the model maintains a comprehensive understanding of entity relationships even in the presence of noisy data.

5.6 Visualization of Embeddings

To intuitively demonstrate LLMEA's effectiveness, we selected aligned entity pairs from FB15K-DB15K and visualized their embeddings using t-SNE (Van der Maaten and Hinton 2008), as shown in Figure 6. Blue and green nodes represent the same entities on different KGs, with black lines connecting aligned pairs. Our model exhibits shorter lines than other models, indicating closer proximity of aligned pairs in the embedding space. Additionally, the increased distance between non-aligned pairs suggests enhanced differentiation of unrelated entities. The compact clustering of aligned pairs and the distinct separation of non-aligned entities further highlight the strength of our embeddings in capturing structural relationships.

5.7 Visualization of Multi-modal Attributes

To further illustrate the effectiveness of LLEMA, we selected entities on FB15K-DB15K and visualized their multi-modal attribute embeddings, as shown in Figure 6. The blue nodes represent entities, red nodes represent visual attributes, and green nodes represent textual attributes. The black lines connect the entities and their attributes. From Figure 6, it is evident that the embeddings generated by the LLMEA model are better compared to ACK-MMEA and MoAlign. In the visualization of our LLMEA embeddings, the entities, visual attributes, and textual attributes are closely clustered together, indicating a high level of integration and coherence in the multi-modal embeddings.

6 Conclusion

In this paper, we present a novel framework to address the information gap challenges in MICEA. By focusing on the completion of missing neighbor entities and attributes within MMKGs, our framework facilitates a more comprehensive and accurate comparison of similarity between entities across diverse relationships and attributes. Through the effective use of LLMs, LLMEA bridges these gaps, enhancing the quality and precision of the alignment process. Experimental results demonstrate the effectiveness of ours.

Acknowledgments

We thank the anonymous reviewers for their insightful comments and suggestions. The corresponding author is Cheng Ji. The authors of this paper were supported by the NSFC through grants No.62402054, No.62425203, and No.62032003, the China Postdoctoral Science Foundation through grant 2024M760279, and the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation under grant BX20250390.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *ICCV*, 2425–2433.

Bordes, A.; Usunier, N.; García-Durán, A.; Weston, J.; and Yakhnenko, O. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NeurIPS*, 2787–2795.

Chen, L.; Li, Z.; Wang, Y.; Xu, T.; Wang, Z.; and Chen, E. 2020. MMEA: Entity Alignment for Multi-modal Knowledge Graph. In *KSEM*, volume 12274, 134–147. Springer.

Chen, L.; Li, Z.; Xu, T.; Wu, H.; Wang, Z.; Yuan, N. J.; and Chen, E. 2022. Multi-modal Siamese Network for Entity Alignment. In *ACM SIGKDD*, 118–126.

Chen, Z.; Chen, J.; Zhang, W.; Guo, L.; Fang, Y.; Huang, Y.; Zhang, Y.; Geng, Y.; Pan, J. Z.; Song, W.; and Chen, H. 2023a. MEAformer: Multi-modal Entity Alignment Transformer for Meta Modality Hybrid. 3317–3327.

Chen, Z.; Chen, J.; Zhang, W.; Guo, L.; Fang, Y.; Huang, Y.; Zhang, Y.; Geng, Y.; Pan, J. Z.; Song, W.; and Chen, H. 2023b. MEAformer: Multi-modal Entity Alignment Transformer for Meta Modality Hybrid. In *MM*, 3317–3327. ACM.

Chen, Z.; Guo, L.; Fang, Y.; Zhang, Y.; Chen, J.; Pan, J. Z.; Li, Y.; Chen, H.; and Zhang, W. 2023c. Rethinking Uncertainly Missing and Ambiguous Visual Modality in Multi-Modal Entity Alignment. In *ISWC*, volume 14265, 121–139.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *ACL*, 4171–4186. Minneapolis, Minnesota.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Fang, X.; Fang, W.; and Wang, C. 2025. Hierarchical Semantic-Augmented Navigation: Optimal Transport and Graph-Driven Reasoning for Vision-Language Navigation. In *Advances in Neural Information Processing Systems*.

Fang, X.; Liu, D.; Zhou, P.; and Hu, Y. 2022. Multi-modal cross-domain alignment network for video moment retrieval. *IEEE Transactions on Multimedia*.

Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2448–2460.

Fang, Z.; Lei, S.; Zhu, X.; Yang, C.; Zhang, S.; Yin, X.; and Qin, J. 2024. Transformer-based Reasoning for Learning Evolutionary Chain of Events on Temporal Knowledge Graph. In *SIGIR*, 70–79. ACM.

Guo, H.; Tang, J.; Zeng, W.; Zhao, X.; and Liu, L. 2021. Multi-modal entity alignment in hyperbolic space. *Neurocomputing*, 461: 598–607.

Huang, Y.; Zhang, X.; Zhang, R.; Chen, J.; and Kim, J. 2024. Progressively Modality Freezing for Multi-Modal Entity Alignment. In *ACL*, 3477–3489.

Ji, J.; Li, J.; Yan, S.; Tian, Q.; and Zhang, B. 2013. Min-Max Hash for Jaccard Similarity. In *ICDM*, 301–309.

Jiang, J.; Li, M.; and Gu, Z. 2021. A Survey on Translating Embedding based Entity Alignment in Knowledge Graphs. In *DSC*, 187–194.

Li, Q.; Guo, S.; Luo, Y.; Ji, C.; Wang, L.; Sheng, J.; and Li, J. 2023a. Attribute-Consistent Knowledge Graph Representation Learning for Multi-Modal Entity Alignment. In *ACM Web Conference*, 2499–2508.

Li, Q.; Ji, C.; Guo, S.; Liang, Z.; Wang, L.; and Li, J. 2023b. Multi-Modal Knowledge Graph Transformer Framework for Multi-Modal Entity Alignment. In *EMNLP*, 987–999.

Li, Q.; Li, J.; Wu, J.; Peng, X.; Ji, C.; Peng, H.; Wang, L.; and Yu, P. S. 2024. Triplet-aware graph neural networks for factorized multi-modal knowledge graph entity alignment. *Neural Networks*, 179: 106479.

Li, Y.; Chen, J.; Li, Y.; Xiang, Y.; Chen, X.; and Zheng, H. 2023c. Vision, Deduction and Alignment: An Empirical Study on Multi-modal Knowledge Graph Alignment. *CoRR*, abs/2302.08774.

Lin, Z.; Zhang, Z.; Wang, M.; Shi, Y.; Wu, X.; and Zheng, Y. 2022a. Multi-modal Contrastive Representation Learning for Entity Alignment. In *COLING*, 2572–2584.

Lin, Z.; Zhang, Z.; Wang, M.; Shi, Y.; Wu, X.; and Zheng, Y. 2022b. Multi-modal Contrastive Representation Learning for Entity Alignment. In *CCL*, 2572–2584.

Liu, F.; Chen, M.; Roth, D.; and Collier, N. 2021a. Visual Pivoting for (Unsupervised) Entity Alignment. In *AAAI*, 4257–4266.

Liu, F.; Chen, M.; Roth, D.; and Collier, N. 2021b. Visual pivoting for (unsupervised) entity alignment. In *AAAI*, volume 35, 4257–4266.

Liu, W.; Zhou, P.; Zhao, Z.; Wang, Z.; Ju, Q.; Deng, H.; and Wang, P. 2020a. K-BERT: Enabling Language Representation with Knowledge Graph. In *AAAI*, 2901–2908.

Liu, X.; Zhao, S.; Su, K.; Cen, Y.; Qiu, J.; Zhang, M.; Wu, W.; Dong, Y.; and Tang, J. 2022. Mask and Reason: Pre-Training Knowledge Graph Transformers for Complex Logical Queries. In *ACM SIGKDD*, 1120–1130.

Liu, Y.; Li, H.; García-Durán, A.; Niepert, M.; Oñoro-Rubio, D.; and Rosenblum, D. S. 2019. MMKG: Multi-modal Knowledge Graphs. In *ESWC*, volume 11503, 459–474. Springer.

Liu, Z.; Cao, Y.; Pan, L.; Li, J.; and Chua, T. 2020b. Exploring and Evaluating Attributes, Values, and Structures for Entity Alignment. In *EMNLP*, 6355–6364.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763.

Shih, K. J.; Singh, S.; and Hoiem, D. 2016. Where to look: Focus regions for visual question answering. In *CVPR*, 4613–4621.

Sun, R.; Cao, X.; Zhao, Y.; Wan, J.; Zhou, K.; Zhang, F.; Wang, Z.; and Zheng, K. 2020. Multi-modal Knowledge Graphs for Recommender Systems. In *CIKM*, 1405–1414.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, J.; Meng, F.; Zheng, D.; Liang, Y.; Li, Z.; Qu, J.; and Zhou, J. 2023. Towards Unifying Multi-Lingual and Cross-Lingual Summarization. In *ACL*, 15127–15143.

Wang, X.; He, X.; Cao, Y.; Liu, M.; and Chua, T. 2019. KGAT: Knowledge Graph Attention Network for Recommendation. In *SIGKDD*, 950–958.

Wang, Y.; Sun, H.; Wang, J.; Wang, J.; Tang, W.; Qi, Q.; Sun, S.; and Liao, J. 2024. Towards Semantic Consistency: Dirichlet Energy Driven Robust Multi-Modal Entity Alignment. In *ICDE*, 3559–3572. IEEE.

Wang, Z.; Li, M.; and Gu, Z. 2021. A Review of Entity Alignment based on Graph Convolutional Neural Network. In *DSC*, 144–151.

Wang, Z.; Lv, Q.; Lan, X.; and Zhang, Y. 2018. Cross-lingual Knowledge Graph Alignment via Graph Convolutional Networks. In *EMNLP*, 349–357.

Xu, G.; Chen, H.; Li, F.; Sun, F.; Shi, Y.; Zeng, Z.; Zhou, W.; Zhao, Z.; and Zhang, J. 2021. AliMe MKG: A Multi-modal Knowledge Graph for Live-streaming E-commerce. In *CIKM*, 4808–4812.

Yuan, L.; Cai, Y.; Shen, X.; Li, Q.; Huang, Q.; Deng, Z.; and Wang, T. 2025. Collaborative Multi-LoRA Experts with Achievement-based Multi-Tasks Loss for Unified Multi-modal Information Extraction. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, 6940–6948. International Joint Conferences on Artificial Intelligence Organization.

Yuan, L.; Cai, Y.; Wang, J.; and Li, Q. 2023. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11051–11059.

Zhu, Q.; Zhou, X.; Wu, J.; Tan, J.; and Guo, L. 2019. Neighborhood-aware attentional representation for multilingual knowledge graphs. In *IJCAI*, 1943–1949.

Zhu, X.; Li, Z.; Wang, X.; Jiang, X.; Sun, P.; Wang, X.; Xiao, Y.; and Yuan, N. J. 2022. Multi-Modal Knowledge Graph Construction and Application: A Survey. *CoRR*, abs/2202.05786.