Breaking the Noise Barrier: LLM-Guided Semantic Filtering and Enhancement for Multi-Modal Entity Alignment

Chenglong Lu†, Chenxiao Li†, Jingwei Cheng*, Yongquan Ji, Guoqing Chen, Fu Zhang
[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
{chenglonglu233,chenguoqing247,jyg0609}@gmail.com
{chengjingwei,zhangfu}@neu.edu.cn
chenxiaoli_joe@163.com

Abstract

Multi-modal entity alignment (MMEA) aims to identify equivalent entities between two multi-modal knowledge graphs (MMKGs). Existing methods have made substantial advancements in enhancing multi-modal fusion. However, the intrinsic noise within modalities, such as the inconsistency in visual modality and redundant attributes, has not been thoroughly investigated. Excessive noise not only weakens semantic representation but also increases the risk of overfitting in attention-based fusion methods. To address this, we propose LGEA (**LLM-G**uided **E**ntity **A**lignment), a novel LLM-guided MMEA framework that prioritizes noise reduction before fusion. Specifically, LGEA introduces two key strategies: (1) fine-grained visual filtering to remove irrelevant images at the semantic level, and (2) contextual summarization of attribute information to enhance entity semantics. To our knowledge, we are the first work to apply LLMs for both visual filtering and attribute-level semantic enhancement in MMEA. Experiments on multiple benchmarks, including the noisy FBYG dataset, show that LGEA sets a new state-of-the-art (SOTA) in robust multi-modal alignment, highlighting the potential of noise-aware strategies as a promising direction for future MMEA research [1].

1 Introduction

Multi-Modal Knowledge Graphs (MMKGs) have emerged as a key approach to modeling real-world information. Compared to typical KGs that focus on structured data, MMKGs incorporate images, texts, and triples to support multi-perspective understanding. This fusion boosts their utility in tasks like recommender systems and visual question answering (Wu et al., 2024). However, constructing MMKGs is challenging due to the need to integrate heterogeneous data from different modalities, making Multi-Modal Entity Alignment (MMEA) a crucial task.

Early MMEA methods significantly improved the accuracy of alignment results by optimizing multi-modal fusion and textual modality alignment. MEAformer (Chen et al., 2023a) dynamically determines modal coefficients by leveraging a transformer to extract relevant information from multiple modalities. Recently, AMF2SEA (Li et al., 2025) adapts to variations in image style by examining the influence of multi-modal feature fusion strategies at the entity level. TMEA (Chen et al., 2024) addresses the issue of insufficient cross-modal information by incorporating large language models (LLMs) for attribute preprocessing, leveraging variational autoencoders to reconstruct incomplete features, and applying cross-attention with orthogonal constraints for refined alignment.

Most existing MMEA methods have primarily focused on designing fusion mechanisms across modalities, while the issue of noise filtering has received only limited attention. Noise widely exists in multimodal data and can directly reduce the accuracy of entity alignment if not properly addressed. In this work, noise is defined as irrelevant, ambiguous, or inaccurate information originating from any modality, which may affect alignment performance. In the visual modality, noise often appears as entity-irrelevant or ambiguous image content, such as background clutter or unrelated objects. In the attribute modality, noise may take the form of redundant, irrelevant, or inaccurate attribute information, such as repeated values or non-discriminative descriptions.

Although previous MMEA methods have verified the importance of image noise filtering, ontology-based methods are coarse-grained and require high human annotation costs. For example, Masked-MMEA (Shi et al., 2022) and AMF2SEA (Li et al., 2025) independently employed pre-trained vision-language models for im-

*Corresponding author. † Equal contribution.

[1] Our code: https://github.com/alusang/LGEA-framework

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33142–33156
November 4-9, 2025 ©2025 Association for Computational Linguistics

age classification and compared the class conflict dictionary (CCD) to mask images identified as noise. Therefore, these methods have two primary limitations: (i) Limited granularity - even when images are classified as the same type, they may still exhibit semantic differences from their corresponding entities; (ii) Over-reliance on pre-defined ontologies and manually constructed CCDs - hindering their generalization to open-domain scenarios and unseen environments. We also observe that noise filtering, even in a simple form, consistently improves alignment performance.

TMEA feeds raw attribute–value pairs directly into LLMs for alignment, but this approach faces two key challenges. First, in datasets like YAGO15K, certain attributes (e.g., wasBornOnDate, diedOnDate) often appear multiple times for the same entity, making pairwise comparison prone to redundant matches based on trivial or duplicated values. Second, the attribute space is dominated by sparse numeric or temporal data, such as coordinates or timestamps, which carry limited semantic depth, making it difficult for LLMs to reliably distinguish entities through direct value comparison.

To tackle these challenges, we propose LGEA (**L**LM-**G**uided **E**ntity **A**lignment), a novel framework for MMEA that addresses modality noise through LLM-guided semantic filtering and representation enhancement, rather than relying solely on fusion strategies. LGEA is composed of two key components: First, we design a fine-grained visual filtering strategy at the semantic level guided by LLMs. Instead of using fixed rules, we generate image captions via BLIP (Li et al., 2022), a vision-language model capable of producing natural descriptions, and let LLMs assess their relevance to entity context, enabling more accurate filtering through language-based reasoning. Second, we leverage LLMs to summarize entity attributes into coherent textual descriptions, capturing richer contextual semantics and improving the generalization of entity representations across different MMKGs. Our main contributions are summarized as follows:

- **LLM-Guided Visual Semantic Filtering**

  We first propose a fine-grained image filtering strategy that uses BLIP-generated captions and LLM-based semantic reasoning to detect and eliminate irrelevant or noisy images to improve alignment performance.

- **LLM-Based Attribute Summarization for Semantic Enhancement**

  We convert raw attribute triples into coherent textual summaries and employ LLMs to generate enriched semantic embeddings, thereby enhancing entity representations and enabling better generalization in MMEA tasks.

- **State-of-the-art (SOTA) Performance on Challenging Benchmarks**

  Our method achieves SOTA performance across multiple benchmarks while utilizing only approximately 10% of the images required by other approaches, including three cross-lingual and two monolingual datasets. Additionally, it significantly reduces computational resource consumption and offers advantages in lightweight computing and scalability.

2 Related Work

2.1 Typical Multi-modal Entity Alignment

EVA (Liu et al., 2021) exploits visual pivots for unsupervised alignment. MSNEA (Chen et al., 2022) uses siamese networks or modality-specific encoders to integrate visual, relational, and attribute features, and MCLEA (Lin et al., 2022) employs contrastive learning to enhance inter-modal and intra-modal representation learning. MEAformer (Chen et al., 2023a) uses transformer-based prediction to adjust fusion weights. LoginMEA (Su et al., 2024) combines local and global signals through hierarchical fusion. AMF2SEA (Li et al., 2025) dynamically selects fusion methods per entity.

2.2 Multi-modal Entity Alignment Based on LLMs and Data Augmentation

UMAEA (Chen et al., 2023b) employs a variational autoencoder (VAE) to reconstruct missing modalities and utilizes a parameter-freezing training strategy to enhance alignment accuracy. SimDiff (Li et al., 2024) enhances alignment indirectly via multi-modal data augmentation using latent diffusion models. TMEA (Chen et al., 2024) proposes a novel cross-modal attention mechanism with orthogonal constraints, and further incorporates LLMs for attribute abstraction, showing that language-driven preprocessing can boost alignment performance.

However, none of the above methods consider fine-grained image semantic noise. If noisy images are fed into the model, it not only increases computational cost but also raises the risk of modality overfitting, thereby compromising alignment accuracy. In contrast, our work explicitly addresses this limitation by introducing an LLM-based semantic filtering mechanism. Unlike prior studies that mainly focus on fusion or augmentation, our method performs pre-alignment data filtering on both visual and attribute modalities, ensuring that only semantically consistent information is retained for subsequent fusion. This design complements existing fusion strategies and leads to more robust and interpretable alignment performance.

3 Preliminary

Multi-modal Knowledge Graph (MMKG) A MMKG is a structured representation of real-world knowledge that integrates multiple modalities of information. Formally, a MMKG can be defined as a tuple $G = (E, R, A, \mathcal{V}, V, T)$, where E is a set of entities, R is a set of relations, A is a set of attributes, $\mathcal{V}$ is a set of attribute values (literals), V is a set of images, and T is a set of triples. The triples in T can be either relational triples (e_h, r, e_t), where $e_h, e_t \in E$ and $r \in R$, or attribute triples (e, a, v), where $e \in E$, $a \in A$, and $v \in \mathcal{V}$.

Multi-modal Entity Alignment (MMEA) MMEA is the task of identifying equivalent entities across different MMKGs. Given two MMKGs $G_1 = (E_1, R_1, A_1, \mathcal{V}_1, V_1, T_1)$ and $G_2 = (E_2, R_2, A_2, \mathcal{V}_2, V_2, T_2)$, the goal of MMEA is to find the set of aligned entity pairs $S = \{(e_1, e_2) \mid e_1 \in E_1, e_2 \in E_2, e_1 \equiv e_2\}$, where $e_1 \equiv e_2$ means that e_1 and e_2 refer to the same real-world object.

4 Method

Our approach consists of five modules, as shown in Figure 1: (1) **Semantic-Based Multi-Modal Data Filtering**, where we use LLMs to remove semantically irrelevant images and noisy attributes, providing cleaner inputs for alignment; (2) **Multi-Modal Information Encoder**, which extracts embeddings from structure, relations, attributes, and images separately; (3) **Multi-Modal Fusion**, which combines these embeddings into a unified representation; (4) **Intra-Modal Contrastive Learning**, which improves alignment by strengthening consistency within modalities; and (5) **Inter-Modal Alignment**, which aligns the fused representation with individual modalities to enhance overall performance.

4.1 Semantic-Based Multi-Modal Data Filtering

LLM-Guided Visual Semantic Filtering. Inspired by Masked-MMEA (Shi et al., 2022), to enhance the reliability of visual information, we first propose a novel fine-grained semantic filtering strategy. For each entity in the MMKG, we utilize the BLIP (Li et al., 2022) model to generate three candidate captions based on the associated image. Generating multiple captions per image helps mitigate the randomness and variability inherent in BLIP-generated descriptions.

Given an aligned entity pair, we input the corresponding captions into an LLM and prompt it to determine whether the two captions describe the same semantic topic. Based on the LLM's decision, we generate a binary mask vector $\mathbf{m}_e \in \{0, 1\}^d$ for each entity e, where d is the dimension of the visual embedding. If the captions are judged to be semantically aligned, the mask is set to all ones, preserving the visual feature; otherwise, the mask is set to all zeros, effectively filtering out the visual information.

While CLIP (Radford et al., 2021) can be used for image–image alignment based on similarity thresholds, its practical application is limited by the need to manually select these thresholds. As shown in Masked-MMEA (Shi et al., 2022), different datasets exhibit varying levels of visual noise, making it labor-intensive and dataset-specific to determine appropriate thresholds. In contrast, our proposed method achieves near-optimal filtering ratios automatically, which is further validated in subsequent experiments, Table 4. Moreover, CLIP only provides coarse similarity scores, whereas our LLM-based semantic filtering operates at a finer, more flexible granularity, allowing selective retention or removal of visual information based on semantic relevance. In summary, our approach eliminates manual threshold tuning while enabling more precise and semantically informed filtering compared with CLIP.

LLM-Based Attribute Summarization for Semantic Enhancement. We leverage LLMs to extract information from entity attributes. For each entity e in the KG, we collect its attribute triples in the form of (*entity name*, *attribute name*, *attribute value*), remove the entity name, and concatenate

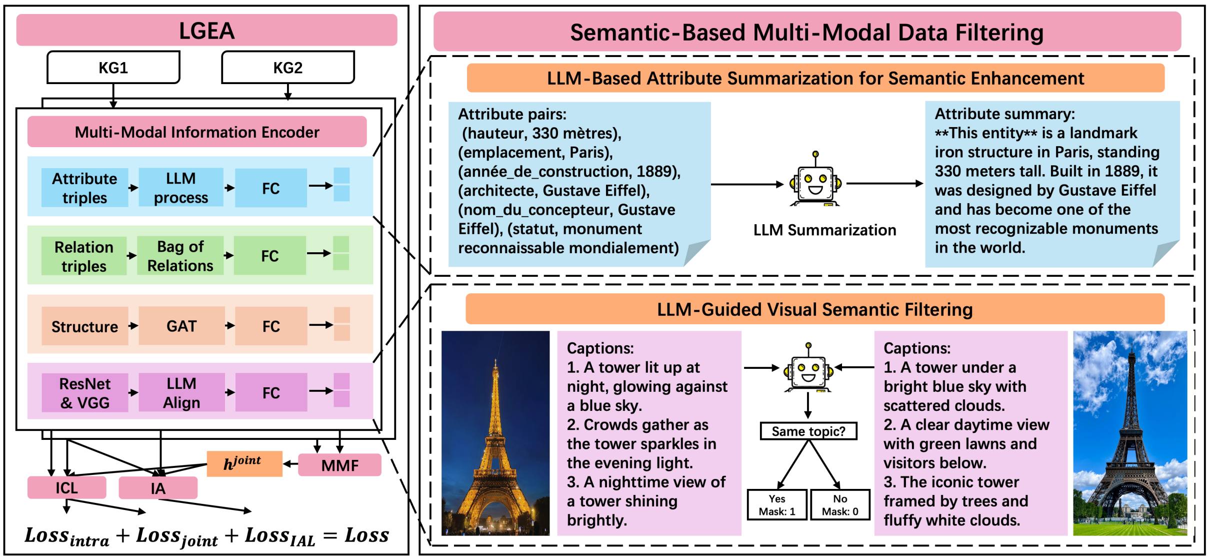
LGEA
KG1
KG2
Multi-Modal Information Encoder
Attribute triples
LLM process
FC
Relation triples
Bag of Relations
FC
Structure
GAT
FC
ResNet & VGG
LLM Align
FC
h^{joint}
MMF
ICL
IA
$Loss_{intra} + Loss_{joint} + Loss_{IAL} = Loss$
Semantic-Based Multi-Modal Data Filtering
LLM-Based Attribute Summarization for Semantic Enhancement
Attribute pairs: (hauteur, 330 mètres), (emplacement, Paris), (année_de_construction, 1889), (architecte, Gustave Eiffel), (nom_du_concepteur, Gustave Eiffel), (statut, monument reconnaissable mondialement)
LLM Summarization
Attribute summary: **This entity** is a landmark iron structure in Paris, standing 330 meters tall. Built in 1889, it was designed by Gustave Eiffel and has become one of the most recognizable monuments in the world.
LLM-Guided Visual Semantic Filtering
Captions: 1. A tower lit up at night, glowing against a blue sky. 2. Crowds gather as the tower sparkles in the evening light. 3. A nighttime view of a tower shining brightly.
Same topic?
Yes Mask: 1
No Mask: 0
Captions: 1. A tower under a bright blue sky with scattered clouds. 2. A clear daytime view with green lawns and visitors below. 3. The iconic tower framed by trees and fluffy white clouds.

Figure 1: The framework of LGEA consists of five key components, (1) Semantic-Based Multi-Modal Data Filtering: This component filters out noisy data by leveraging semantic cues to enhance the quality of the input data before further processing. (2) Multi-Modal Information Encoder: It encodes the filtered data from different modalities into a shared representation space, enabling a better fusion of multi-modal information. (3) Multi-Modal Fusion (MMF): This component combines the representations from different modalities to create a unified multi-modal feature space that captures the most relevant information for alignment. (4) Intra-Modal Contrastive Learning (ICL): It enforces consistency within each modality by applying contrastive learning techniques to minimize the difference between similar instances and maximize the gap between dissimilar ones. (5) Inter-Modal Alignment (IA): Utilizes Inter-Modal Alignment Loss (IAL) to ensure accurate alignment of semantically similar entities across different modalities by minimizing the distance between aligned representations.

the remaining pairs into a textual sequence $\mathbf{t}_e$, e.g., $(att_1, value_1)$, $(att_2, value_2)$, ... This forms a raw attribute description that reflects the factual content associated with the entity. Specifically, each $\mathbf{t}_e$ is summarized into a concise text (limited to 100 words) for embedding.

Constraining the summary length encourages the model to focus on the most salient and discriminative attributes while filtering out redundant or ambiguous information. Furthermore, since entity names are excluded and no external prompts are used to invoke factual knowledge, the risk of hallucination (Zhang et al., 2023; Chen et al., 2025) and information leakage is minimized. The prompt is shown in Table 11.

4.2 Multi-Modal Information Encoder

We employ four encoders to extract modality-specific embeddings: structure, attribute, relation, and vision. Each embedding is projected into a shared latent space via a fully connected layer.

Structure Encoder We use Graph Attention Networks (GATs) (Veličković et al., 2018) to capture local graph structure through attention-based aggregation. For a node i corresponding to entity e, its structure-based embedding is computed as:

$$\mathbf{h}_i^{\text{structure}} = \text{FC}\left(\sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}\mathbf{h}_j\right)\right), \quad (1)$$

where α_{ij} is the attention weight between node i and neighbor j, and σ denotes a non-linear activation function. The final output is obtained via a fully connected layer $\text{FC}(\cdot)$.

Attribute Encoder We leverage an LLM to encode the textual attributes of each entity. Given a summarized attribute description $\mathbf{t}_e$, the attribute embedding is:

$$\mathbf{h}_e^{\text{A}} = \text{FC}\left(\mathcal{F}_{\text{LLM}}(\mathbf{t}_e)\right), \quad (2)$$

where $\mathcal{F}_{\text{LLM}}(\cdot)$ denotes the LLM-based encoder.

Relation Encoder We represent the relational context of an entity e as a sparse bag-of-relations (Yang et al., 2019) vector $\mathbf{u}_e^{\text{R}}$. This feature is transformed into a dense embedding using:

$$\mathbf{h}_e^{\text{R}} = \text{FC}\left(\mathbf{u}_e^{\text{R}}\right). \quad (3)$$

Vision Encoder Let $\mathbf{v}_e$ be the image embedding from a pretrained ResNet (He et al., 2016) and VGG (Simonyan and Zisserman, 2015). We apply an entity-specific mask $\mathbf{m}_e$ to filter irrelevant visual signals:

$$\tilde{\mathbf{v}}_e = \mathbf{m}_e \odot \mathbf{v}_e, \tag{4}$$

$$\mathbf{h}_e^{\mathrm{V}} = \mathrm{FC}\left(\tilde{\mathbf{v}}_e\right). \tag{5}$$

This semantic-guided filtering helps suppress visual noise and enhances the quality of visual representations.

4.3 Multi-Modal Fusion

To integrate multiple modalities effectively, we adopt a **Multi-Head Cross-Modal Attention (MHCA)** mechanism inspired by MEAformer (Chen et al., 2023a). Given a set of available modality embeddings $\{\mathbf{h}^m\}_{m\in M}$, where $M = \{structure, R, A, V\}$, we first stack them into a matrix $\mathbf{H}^m$ and feed them into a shared MHCA module:

$$\tilde{\mathbf{H}}^m = \mathrm{MHCA}(\mathbf{H}^m), \tag{6}$$

where each $\tilde{\mathbf{H}}^m$ contains contextualized information from all other modalities.

We extract the attention weights α_m from the MHCA module by summing over heads and normalizing:

$$\alpha_m = \mathrm{softmax}\left(\sum_{i=1}^{A_h} \mathrm{Attention}^{(i)}\right), \tag{7}$$

where $\mathrm{Attention}^{(i)} \in \mathbb{R}^{B\times|M|\times|M|}$ is the attention matrix from head i, and softmax is applied across modalities.

Finally, we compute the joint embedding as a weighted concatenation of the original modality embeddings:

$$\mathbf{h}^{joint} = \bigoplus_{m\in M} \alpha_m \cdot \mathrm{Norm}(h^m), \tag{8}$$

where $\mathrm{Norm}(\cdot)$ denotes ℓ_2 normalization and $\bigoplus$ indicates concatenation across modalities.

This allows the model to softly emphasize more informative modalities and produce a unified representation for downstream tasks.

4.4 Intra-modal Contrastive Learning (ICL)

Inspired by MEAformer (Chen et al., 2023a), to improve both modality-specific quality and cross-modal alignment, we adopt a dual-level contrastive learning framework.

Given a set of aligned entity pairs $(e_1, e_2) \in S$, we compute contrastive probabilities using in-batch negatives:

$$p(e_1, e_2) = \frac{\exp(\mathrm{sim}(f(e_1), f(e_2))/\tau)}{\sum\limits_{e'\in N(e_1,e_2)} \exp(\mathrm{sim}(f(e_1), f(e'))/\tau)}, \tag{9}$$

where $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity and τ is a temperature parameter.

We define a contrastive loss as:

$$\begin{aligned} L_m = -\mathbb{E}_{(e_1,e_2)\in S}\big[&\log p_m(e_1, e_2) \\ &+ \log p_m(e_2, e_1)\big]. \end{aligned} \tag{10}$$

Intra-modal Contrastive Loss. For each modality $m \in M$, we compute an intra-modal contrastive loss L_m to encourage alignment within the same modality. These individual losses are then combined using a learnable weighted sum to form the total intra-modal contrastive loss:

$$L_{\text{intra}} = \sum \lambda_m L_{m\in M}. \tag{11}$$

Joint Contrastive Loss. To enhance cross-modal consistency, the same objective is applied to the fused embeddings:

$$L_{\text{joint}} = L_{\sum m} \tag{12}$$

4.5 Inter-modal Alignment

To ensure consistency between each modality and the joint embedding, we introduce an Inter-modal Alignment Loss (IAL) (Lin et al., 2022) by minimizing the bidirectional KL divergence (Schulman et al., 2017) between their alignment distributions.

$$\begin{aligned} L_{\text{IAL}}^m = \mathbb{E}_{(e_1,e_2)\in S}\frac{1}{2}(&\mathrm{KL}(q_o(e_1,e_2)\|q_m(e_1,e_2)) \\ &+ \mathrm{KL}(q_o(e_2,e_1)\|q_m(e_2,e_1))), \end{aligned} \tag{13}$$

where q_o and q_m represent the alignment probability distributions from the joint embedding space and modality m, respectively. Specifically, $q_o(e_1, e_2)$ and $q_o(e_2, e_1)$ represent the alignment probabilities of entity pairs (e_1, e_2) and (e_2, e_1) obtained by the joint embedding, while $q_m(e_1, e_2)$ and $q_m(e_2, e_1)$ represent the alignment probabilities of the corresponding entity pairs in modality m.

We ensure the consistency between each modality and the joint embedding by minimizing the bidirectional KL divergence between these two probability distributions. The KL divergence measures the difference between the two probability distributions, and the bidirectional calculation ensures the symmetry and sufficiency of the alignment process.

The total IAL is computed by aggregating over all modalities:

$$L_{\text{IAL}} = \lambda \cdot \sum_{m \in M} L_{\text{IAL}}^{m}, \tag{14}$$

where λ controls the overall contribution of IAL.

The final training objective combines all aforementioned loss components to enable both intra- and inter-modal learning. Specifically, the overall loss is defined as:

$$L = L_{\text{intra}} + L_{\text{joint}} + L_{\text{IAL}}. \tag{15}$$

5 Experiments

5.1 Experimental setup

Datasets We evaluate our method on two types of datasets: (i) **Bilingual datasets:** We use DBP15K (Chen et al., 2023a), which includes three cross-lingual subsets derived from the multilingual versions of DBpedia: DBP15K$_{\text{ZH-EN}}$, DBP15K$_{\text{JA-EN}}$, and DBP15K$_{\text{FR-EN}}$. Each subset contains approximately 400K knowledge graph triples and 15K pre-aligned entity pairs, of which 30% are used as seed alignments. (ii) **Monolingual datasets:** We select FBDB15K and FBYG15K from the MMKG (Liu et al., 2019) benchmark, of which 20% are used as seed alignments. Detailed information about these datasets is shown in Table 10 (see Appendix).

Evaluation Metrics We assess the alignment probability between entities from different MMKGs based on cosine similarity. Hits@N and Mean Reciprocal Rank (MRR) are used as evaluation metrics for all models. Higher values of Hits@N and MRR indicate better performance.

Experimental Setup All experiments are conducted on entity alignment tasks with 20% of entity pairs used as supervision for the FBDB15K and FBYG15K datasets, and 30% for the DBP15K dataset. For each setting, models are trained iteratively, where alignment predictions from previous rounds are gradually incorporated to refine the embedding space and improve alignment quality. The entire experimental process does not involve any surface forms (i.e., entity names).

Implementation Details We utilize **LLaMA-3-8B-Instruct** as the core LLM for both attribute summarization and embedding. For image-text matching tasks, we employ **GPT-3.5-Turbo** via API-based inference. All experiments are implemented using the **PyTorch** framework. Inference with LLaMA-3-8B-Instruct is conducted on a computing server equipped with **NVIDIA A100 GPU (40GB)**. No fine-tuning is applied to either model; instead, we adopt task-specific prompt engineering to adapt the LLMs to various alignment subtasks.

5.2 Main Experimental Results

We evaluate LGEA on five standard multi-modal MMEA benchmarks, covering two FB15K-based datasets (FBDB15K and FBYG15K) and three multilingual DBP15K datasets (ZH-EN, JA-EN, FR-EN). The results are reported in Table 2 and Table 1, with comparisons against a comprehensive set of baselines, including early multi-modal models (MCLEA, MEAformer), and recent SOTA approaches such as SimDiff, LoginMEA, and TMEA.

On the multilingual DBP15K datasets, as shown in Table 1, LGEA achieves substantial and consistent improvements over all baselines. On DBP15K$_{\text{ZH-EN}}$, LGEA achieves 0.970 on Hits@1, 0.998 on Hits@10, and 0.982 on MRR—outperforming LoginMEA (0.873/0.978/0.913) by over 9 points on Hits@1 and nearly 7 points on MRR. Similarly, on DBP15K$_{\text{JA-EN}}$, LGEA records (0.962/0.999/0.977), and on DBP15K$_{\text{FR-EN}}$, it attains (0.975/0.999/0.985), achieving new SOTA performance in all metrics across all three language pairs.

While recent methods such as UMAEA and MEAformer show strong performance on DBP15K, they still fall short in consistently achieving top Hits@1 scores across datasets. LGEA's strong performance on Hits@1 indicates its superior precision in top-ranked predictions, which is particularly critical in resource-constrained alignment scenarios. Furthermore, LGEA exhibits high Hits@10 across all datasets, demonstrating not only accurate top-1 matching but also broader contextual ranking quality.

On the monolingual FB15K-based datasets (Table 2), LGEA delivers competitive or superior performance. For FBDB15K, LGEA achieves a Hits@1 of 0.801, Hits@10 of 0.910, and MRR of 0.842. These results are behind the current top-performing model TMEA (0.867/0.944/0.895), but still represent a clear improvement over all other baselines, including LoginMEA (0.667/0.854/0.735) and SimDiff

Models	DBP15K$_{\text{ZH-EN}}$			DBP15K$_{\text{JA-EN}}$			DBP15K$_{\text{FR-EN}}$		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
Masked-MMEA	0.612	0.837	0.693	0.627	0.858	0.711	0.712	0.901	0.779
MSNEA	0.643	0.865	0.719	0.572	0.832	0.660	0.584	0.841	0.671
AMF2SEA	0.691	0.879	0.751	0.696	0.871	0.757	0.767	0.914	0.818
EVA	0.746	0.910	0.807	0.741	0.918	0.805	0.767	0.939	0.831
MCLEA	0.811	0.954	0.865	0.806	0.953	0.861	0.811	0.954	0.865
SimDiff	0.829	0.963	0.877	0.835	0.966	0.883	0.861	0.980	0.905
MEAformer	0.847	0.970	0.892	0.842	0.974	0.892	0.845	0.976	0.894
UMAEA	0.856	0.974	0.900	0.857	0.980	0.904	0.873	0.988	0.917
LoginMEA	0.873	0.978	0.913	0.866	0.981	0.911	0.881	0.988	0.924
LGEA	**0.970**	**0.998**	**0.982**	**0.962**	**0.999**	**0.977**	**0.975**	**0.999**	**0.985**

Table 1: Main experimental results of LGEA on DBP15K datasets.

Models	FBDB15K			FBYG15K		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
MMEA	0.265	0.451	0.357	0.234	0.398	0.317
MoAlign	0.318	0.564	0.296	0.378	0.000	0.000
MCLEA	0.441	0.640	0.534	0.406	0.579	0.488
EVA	0.556	0.666	0.609	0.103	0.217	0.164
SimDiff	0.615	0.820	0.678	0.530	0.736	0.595
MSNEA	0.653	0.768	0.708	0.443	0.626	0.529
LoginMEA	0.667	0.854	0.735	0.758	0.898	0.810
TMEA	**0.867**	**0.944**	**0.895**	0.818	0.916	0.853
LGEA	0.801	0.910	0.842	**0.850**	**0.942**	**0.883**

Table 2: Main experimental results of LGEA on FBDB15K and FBYG15K datasets.

(0.615/0.820/0.678), showcasing LGEA's strong modality integration capability.

On the more challenging FBYG15K dataset, LGEA attains SOTA results with a Hits@1 of 0.850, Hits@10 of 0.942, and MRR of 0.883. This surpasses all baselines, including the previously best-performing TMEA (0.818/0.916/0.853), with margins of +3.2, +2.6, and +3.0 points on Hits@1, Hits@10, and MRR respectively. It is worth noting that many models (e.g., EVA, MSNEA) suffer significant performance drops on this dataset, while LGEA remains robust and accurate, demonstrating its strong generalization across diverse graph and modality structures.

In summary, LGEA achieves the best performance on 4 out of 5 datasets and ranks second only to TMEA on the remaining one, while still outperforming TMEA on the more challenging FBYG15K benchmark, which is characterized by fewer attribute and relation types but significantly more noise. These results confirm the effectiveness and robustness of LGEA in integrating and aligning multi-modal information, and establish it as a new SOTA for MMEA tasks across both monolingual and multilingual settings.

5.3 Ablation Study

To better understand the contribution of each modality and architectural component within LGEA, we conduct a comprehensive ablation study on two benchmark datasets: FBDB15K and FBYG15K. The results are summarized in Table 3. We divide the ablation into two parts: the upper half of the table explores the impact of removing individual modalities (vision (`wo/img`), relations (`wo/rel`), structure (`wo/structure`), and attributes (`wo/att`)), while the lower half focuses on core design choices of the model, LLM integration (`wo/llm`), iterative refinement (`wo/it`), and visual masking (`wo/mask`).

Modality Ablation Each modality brings distinct and complementary signals. Removing structural information (`wo/structure`) leads to the most significant performance drop, especially

Models	FBDB15K			FBYG15K		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
LGEA	**0.801**	**0.910**	**0.842**	**0.850**	**0.942**	**0.883**
w/o img	0.740	0.890	0.791	0.767	0.896	0.811
w/o rel	0.737	0.890	0.791	0.813	0.917	0.851
w/o structure	0.440	0.615	0.502	0.678	0.844	0.737
w/o att	0.655	0.840	0.720	0.563	0.755	0.628
w/o llm	0.682	0.855	0.745	0.558	0.756	0.626
w/o mask	0.714	0.868	0.772	0.818	0.917	0.854
w/o it	0.684	0.868	0.749	0.785	0.914	0.832

Table 3: Ablation results of LGEA on FBDB15K and FBYG15K datasets, showing modal and method ablations.

Models	FBDB15K			FBYG15K		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
LGEA	**0.801**	**0.910**	**0.842**	**0.850**	**0.942**	**0.883**
0%mask	0.714	0.868	0.772	0.818	0.917	0.854
50%mask	0.790	0.908	0.834	0.842	0.934	0.878
80%mask	0.766	0.890	0.812	0.834	0.928	0.868

Table 4: Ablation results of different masking proportion on FBDB15K and FBYG15K.

on FBDB15K, where Hits@1 falls from 0.801 to 0.440. This underscores the importance of neighborhood-based reasoning for MMEA. Visual information (wo/img) and relational triples (wo/rel) also provide important support, as their removal results in 6 $\sim$ 7% drops in Hits@1 and MRR, indicating their synergy with other modalities. Notably, removing attribute information (wo/att) causes a substantial drop, especially on FBYG15K (Hits@1 drops from 0.850 to 0.563), highlighting the key role of rich semantic clues from attributes in alignment.

Architecture Ablation Beyond evaluating individual modalities, we examine the impact of core architectural components in LGEA. The LLM-based attribute summarization module (wo/llm) proves essential, particularly in settings with noisy or inconsistent attribute information. Its removal leads to a dramatic drop in performance—for example, Hits@1 falls from 0.850 to 0.558 on FBYG15K—highlighting that raw structured attributes often carry noise and redundancy, whereas LLM-generated summaries capture more coherent and informative semantics for alignment. The visual semantic filtering module (wo/mask) also contributes notably by eliminating irrelevant or misleading images. As shown in Table 3, omitting this component causes performance to decline (e.g., Hits@1 drops from 0.801 to 0.714 on FBDB15K and from 0.850 to 0.818 on FBYG15K). Finally, removing the iterative refinement mechanism (wo/it) leads to noticeable decreases across metrics (e.g., MRR drops from 0.842 to 0.749 on FBDB15K and from 0.883 to 0.832 on FBYG15K), indicating its effectiveness in progressively improving alignment quality. Together, these results underscore the importance of semantic enhancement and noise filtering in advancing MMEA performance.

These results confirm that both modality-level and architecture-level components are critical to LGEA's success. Notably, the LLM-guided semantic filtering and summarization modules contribute the most, as they effectively reduce visual and attribute noise while enriching entity semantics. Their synergy, together with iterative refinement, enables LGEA to achieve strong generalization across diverse and noisy datasets. This highlights the importance of semantic-level reasoning and noise-aware design in advancing MMEA.

5.4 Mask Analysis

Effect of Visual Semantic Masking

We analyze the impact of varying the proportion of images considered valid (i.e., mask = 1) in the visual semantic filtering module. As shown in Table 4, the masking mechanism plays a critical role in balancing information quality and quantity in the visual modality.

When no masking is applied (`0%mask`), the model includes all images regardless of their relevance. This leads to a noticeable performance degradation, especially on FBDB15K, where Hits@1 drops from 0.801 (full LGEA) to 0.714, and MRR drops from 0.842 to 0.772. This suggests that noisy or irrelevant images can introduce confusion and negatively affect alignment quality.

Applying a moderate mask ratio (`50%mask`) achieves the best performance across both datasets, reaching 0.790 Hits@1 and 0.834 MRR on FBDB15K, and 0.842 Hits@1 and 0.878 MRR on FBYG15K. This demonstrates that selectively filtering less informative images while retaining valuable visual cues enhances model effectiveness.

However, an overly aggressive masking strategy (`80%mask`) slightly reduces performance compared to 50% masking, with Hits@1 falling to 0.766 on FBDB15K and 0.834 on FBYG15K. This indicates that discarding too many images may lead to the loss of useful semantic signals.

In summary, these results highlight the importance of image filtering mechanism. Proper visual masking mitigates noise and improves alignment, while excessive masking can result in information loss. Importantly, our BLIP+LLM-based semantic filtering demonstrates the effectiveness of automatically identifying noisy visual information across different datasets, achieving suitable masking ratios without the need for manually defined thresholds for each dataset. This adaptability not only validates the robustness of our approach but also reduces reliance on human intervention.

5.5 Ablation Study on Loss Functions

FBDB15K			
settings	**Hits@1**	**Hits@10**	**MRR**
LGEA	0.801	0.910	0.842
w/o L_{intra}	0.711	0.868	0.765
w/o L_{joint}	0.742	0.876	0.791
w/o L_{IAL}	0.770	0.894	0.816

Table 5: Ablation results of different loss components on the FBDB15K dataset.

From Table 5, we observe that removing any loss term leads to a performance drop compared to the complete LGEA model, demonstrating that each component plays a crucial role. Specifically, eliminating L_{intra} causes the largest decline, with Hits@1 decreasing from 0.801 to 0.711 and MRR from 0.842 to 0.765, highlighting the importance of intra-modal consistency in learning robust representations. Removing L_{joint} also results in a considerable reduction, showing that global alignment across modalities is essential for accurate entity matching. The performance decrease caused by removing L_{IAL} is relatively smaller but still noticeable, indicating that inter-modal alignment further enhances cross-modal correspondence. Overall, these results verify the complementary effect of the three loss terms and confirm that their joint optimization significantly improves model performance.

6 Conclusion

In this paper, we present LGEA, a novel method for MMEA that tackles a fundamental but underexplored challenge: the prevalence of noisy or semantically irrelevant information across modalities. Instead of focusing solely on fusion strategies, LGEA shifts the perspective toward semantic-level understanding and noise reduction, guided by LLMs. Specifically, we introduce two key techniques: (1) a fine-grained visual semantic filtering method that leverages BLIP and LLMs to discard off-topic images based on entity context; and (2) an attribute summarization strategy that transforms structured attributes into coherent textual descriptions to enhance the semantics of entities. Extensive experiments across both monolingual and multilingual MMEA benchmarks validate the effectiveness of our approach, with LGEA achieving SOTA performance and demonstrating strong robustness under noisy settings. These results highlight the value of LLM-guided reasoning in advancing MMEA.

Limitations

Although the proposed LGEA framework achieves strong results across diverse datasets, there remain certain limitations worth further investigation. For instance, on the structurally rich FBDB15K dataset, LGEA underperforms compared to TMEA. This is largely due to TMEA's powerful fusion mechanism that better leverages abundant relational information. However, in noisier and more structurally sparse scenarios, such as FBYG15K, LGEA outperforms TMEA by a notable margin, thanks to its LLM-guided semantic filtering and enhancement strategies. These results suggest that future research should advance noise reduction and multimodal fusion together to achieve more robust and accurate MMEA.

On the other hand, in the semantic filtering process of the visual modality, the original design considered using a multi-modal large model (MLLM) for more refined image semantic alignment to further improve the accuracy of visual quality control. However, in the experimental preparation stage, considering the high computational and resource costs of the current mainstream MLLM, we had to abandon this solution at this stage. Nevertheless, with the development of multi-modal model technology and the gradual reduction of inference costs, this direction still has great potential. In the future, the introduction of lightweight or task-customized MLLM modules will be an important direction worthy of further research.

Ethics Statement

To the best of our knowledge, this work does not involve any discrimination, social bias, or private data. All the datasets are constructed from open-source KGs such as Wikidata, YAGO15K, FB15K, and DBpedia. Therefore, we believe that our work complies with the EMNLP Ethics Policy.

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

Guoqing Chen, Fu Zhang, Jinghao Lin, Chenglong Lu, and Jingwei Cheng. 2025. Rrhf-v: Ranking responses to mitigate hallucinations in multimodal large language models with human feedback. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6798–6815.

Liyi Chen, Zhi Li, Tong Xu, Han Wu, Zhefeng Wang, Nicholas Jing Yuan, and Enhong Chen. 2022. Multimodal 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*.

Zhuo Chen, Jiaoyan Chen, Wen Zhang, Lingbing Guo, Yin Fang, Yufeng Huang, Yichi Zhang, Yuxia Geng, Jeff Z Pan, Wenting Song, and 1 others. 2023a. 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, Lingbing Guo, Yin Fang, Yichi Zhang, Jiaoyan Chen, Jeff Z Pan, Yangning Li, Huajun Chen, and Wen Zhang. 2023b. Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In *International Semantic Web Conference*, pages 121–139. Springer.

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.

Chenxiao Li, Jingwei Cheng, Qiang Tong, and Fu Zhang. 2025. Exploring the impacts of feature fusion strategy in multi-modal entity alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7809–7818.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

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*, 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 *European Semantic Web Conference*, pages 459–474. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Yinghui Shi, Meng Wang, Ziheng Zhang, Zhenxi Lin, and Yefeng Zheng. 2022. Probing the impacts of visual context in multimodal entity alignment. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 255–270. Springer.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Wei Wu, Chao Wang, Dazhong Shen, Chuan Qin, Liyi Chen, and Hui Xiong. 2024. Afdgcf: Adaptive feature de-correlation graph collaborative filtering for recommendations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1242–1252.

Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, and Xu Sun. 2019. Aligning cross-lingual entities with multi-aspect information. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4430–4440. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

A Appendix

A.1 Statistical Data of Datasets

Table 10 presents the statistics of the datasets used in our experiments, covering both multilingual and cross-KG scenarios.

Dataset: Each dataset consists of a pair of knowledge graphs (KGs), either from different languages (e.g., ZH-EN) or different sources (e.g., FB15K vs. DB15K/YAGO15K).

KG: This field specifies the identity of the individual KG within the dataset, such as Chinese, English, Freebase (FB15K), DBpedia (DB15K), or YAGO (YAGO15K).

# Ent.: The number of unique entities in each KG.

# Rel.: The number of distinct relation types used to form relation triples.

# Attr.: The number of distinct attribute types used to describe entity properties.

# Rel. Triples: The total number of relation triples, each representing a factual connection in the form of (head entity, relation, tail entity).

# Attr. Triples: The total number of attribute triples, each describing a property in the form of (entity, attribute, value).

# Image: The number of entities that are associated with image data. Not all entities are linked with an image.

# EA pairs: The number of pre-aligned entity pairs across the two KGs, used as ground truth for training or evaluation in the entity alignment task.

Note that the values may vary significantly across datasets due to differences in language, domain, and data richness. Additionally, some entities may lack image data or aligned counterparts in the other KG.

A.2 Comparison of parameters of different MMEA methods

We report the parameter sizes of different methods on the DBP15K dataset, as shown in the Figure 2. Although LGEA has slightly more parameters than the previous classical MMEA methods, it remains significantly lighter than TMEA, whose parameter size is over five times larger. The increase in LGEA's parameters primarily comes from the use of higher-dimensional attribute vectors generated by the LLM-based encoder.

A.3 Prompts

To facilitate the use of LLMs in entity representation and cross-modal alignment, we design specific prompts targeting two key tasks: textual summarization and image-based similarity judgment. As shown in Table 11, the summary prompt guides the model to generate a concise English description of an entity based on its structured attribute-value pairs. This enables abstraction and normalization of entity semantics from heterogeneous knowledge graphs. Meanwhile, Table 12 presents the image matching prompt, which is used to determine

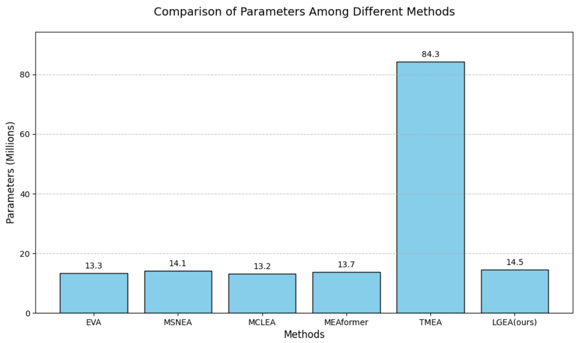

Figure 2: Comparison of parameters of different MMEA methods.

whether two image captions refer to the same entity or concept. If they match, the model is further asked to extract a shared theme, aiding in cross-modal entity alignment by highlighting common visual semantics. These carefully crafted prompts ensure task-specific responses while minimizing irrelevant or verbose output.

A.4 Comparative Study

Comparison and Analysis of Iterative Alignment Results. Table 13 reports the performance of several representative multimodal entity alignment methods under different proportions of training seeds (20%, 50%, and 80%) on two widely used cross-KG datasets: FBDB15K and FBYAGO15K. We compare the proposed method **LGEA** against four baselines: EVA, MSNEA, MCLEA, and MEAformer, using standard evaluation metrics including Hits@1, Hits@10, and MRR.

Across all settings, **LGEA consistently outperforms all baselines**, demonstrating strong generalization and alignment capability. When only 20% of the reference entities are provided, LGEA achieves 0.771 Hits@1, 0.894 Hits@10, and 0.816 MRR on FBDB15K, significantly surpassing the previously best-performing MEAformer by 0.193, 0.082, and 0.155, respectively. Similarly, on FBYAGO15K, LGEA reaches 0.850 Hits@1, 0.942 Hits@10, and 0.883 MRR, showing substantial improvements of 0.406, 0.250, and 0.354 over MEAformer.

As the proportion of training seeds increases to 50% and 80%, the performance of all models improves, but LGEA maintains a clear advantage. For instance, at the 80% level, LGEA achieves 0.866 Hits@1, 0.946 Hits@10, and 0.894 MRR on FBDB15K, as well as 0.925 Hits@1, 0.978 Hits@10, and 0.944 MRR on FBYAGO15K, consistently setting new state-of-the-art results.

These results confirm that LGEA is not only effective in low-resource scenarios but also scales well with more supervision. Moreover, its superiority is more pronounced on FBYAGO15K, a dataset characterized by greater structural and modality heterogeneity, which further validates the robustness and adaptability of our proposed approach across different types of knowledge graphs.

A.5 Impact of Different LLMs

FBDB15K			
settings	**Hits@1**	**Hits@10**	**MRR**
LGEA$_{\text{Llama3-8B}}$	0.801	0.910	0.842
LGEA$_{\text{Llama2-7B}}$	0.793	0.898	0.836
LGEA$_{\text{Mistral-7B}}$	0.791	0.893	0.833

Table 6: Performance comparison of LGEA with different LLMs on the FBDB15K dataset.

From Table 6, we observe that all three LLM variants achieve competitive results, with Hits@1 ranging from 0.791 to 0.801 and MRR ranging from 0.833 to 0.842. Among them, LGEA$_{\text{Llama3-8B}}$ achieves the best performance, reaching 0.801 in Hits@1 and 0.910 in Hits@10, slightly outperforming the other two backbones. LGEA$_{\text{Llama2-7B}}$ and LGEA$_{\text{Mistral-7B}}$ show marginally lower results but remain close, demonstrating the robustness of the framework to different LLMs.

These results suggest that while the choice of LLM can influence performance, the gap between different backbones is relatively small. This indicates that the proposed approach does not rely heavily on a specific model, but rather benefits from the general semantic reasoning capabilities of modern LLMs. Nonetheless, the advantage of Llama3-8B highlights that stronger LLMs with better alignment ability can provide slight but consistent improvements in entity alignment tasks.

A.6 Impact of Learning Rate

As shown in Table 7, the performance of LGEA remains stable across different learning rates. The optimal result is obtained when the learning rate is set to 5×10^{-4}, but the differences compared with 1×10^{-4} and 1×10^{-3} are relatively minor. This indicates that our approach is not overly sensitive to the choice of learning rate, and the reported improvements are not dependent on careful hyperparameter tuning. These results further confirm the robustness and reliability of our method.

FBDB15K			
settings	**Hits@1**	**Hits@10**	**MRR**
LGEA$_{\text{5e-4}}$	0.801	0.910	0.842
LGEA$_{\text{1e-4}}$	0.794	0.901	0.829
LGEA$_{\text{1e-3}}$	0.772	0.896	0.814

Table 7: Parameter sensitivity analysis of LGEA on FBDB15K under different learning rates.

A.7 Efficiency and Scalability Analysis

To evaluate the efficiency of our method, we measured the processing time of LLM-based components. As shown in Table 8, the most time-consuming stage is the Visual Semantic Filtering, which relies on GPT API calls (about 1–2 seconds per call, 15k calls in total), resulting in 5–8 hours of processing. This stage, however, can be significantly accelerated by parallel requests, thus scaling well across multiple datasets. The Attribute Summarization and entity information embedding, both completed locally by the LLaMA3-8B model, require only about 10 minutes even for the full set of 30k entities. Overall, the total LLM processing time for a complete dataset is about 5–8 hours, and processing all five datasets requires only 6–9 hours, demonstrating the good efficiency and scalability of our approach.

A.8 Cost Analysis

As shown in Table 9, the cost analysis of LLM usage in our LGEA framework includes two main parts. Attribute Summarization and embedding are processed by Llama3-8B on an A100-40G GPU, costing about $0.3. Visual Semantic Filtering is performed via the GPT-3.5-turbo API, costing about $0.2. Therefore, the maximum total cost for the entire DBP15K$_{\text{ZH-EN}}$ dataset is approximately $0.5.

LLM Component	**Setting**	**Time Cost**
Visual Semantic Filtering	GPT API (15k calls, 1–2s per call)	5–8 h (parallelizable)
Attribute Summarization	LLaMA3-8B (local)	$\approx$ 10 min (total)
Embedding of entity information	LLaMA3-8B (local)	
Total (per dataset)		5–8 h
Total (all 5 datasets)		6–9 h

Table 8: Processing time of LLM components in our method (measured on DBP15K$_{\text{ZH-EN}}$ with 15,000 entity pairs).

Module	**Setup**	**Computation**	**Cost**
Visual Semantic Filtering	GPT-3.5-turbo API	15,000 calls, $\sim$600k tokens	\$0.2
Attribute Summarization & embedding	Llama3-8B (A100-40G)	15,000 pairs, 10–20 min	\$0.3
Total	–	–	**\$0.5**

Table 9: Cost analysis of LLM processing in the LGEA framework on DBP15K$_{\text{ZH-EN}}$.

Dataset	**KG**	**# Ent.**	**# Rel.**	**# Attr.**	**# Rel. Triples**	**# Attr. Triples**	**# Image**	**# EA pairs**
DBP15K$_{\text{ZH-EN}}$	ZH (Chinese)	19,388	1,701	8,111	70,414	248,035	15,912	15,000
	EN (English)	19,572	1,323	7,173	95,142	343,218	14,125	
DBP15K$_{\text{JA-EN}}$	JA (Japanese)	19,814	1,299	5,882	77,214	248,991	12,739	15,000
	EN (English)	19,780	1,153	6,066	93,484	320,616	13,741	
DBP15K$_{\text{FR-EN}}$	FR (French)	19,661	903	4,547	105,998	273,825	14,174	15,000
	EN (English)	19,993	1,208	6,422	115,722	351,094	13,858	
FBDB15K	FB15K	14,951	1,345	116	592,213	29,395	13,444	12,846
	DB15K	12,842	279	225	89,197	48,080	12,837	
FBYG15K	FB15K	14,951	1,345	116	592,213	29,395	13,444	11,199
	YAGO15K	15,404	32	7	122,886	23,532	11,194	

Table 10: Statistics for datasets

Summary Prompt

```
prompt = """
You are an expert who can provide concise explanations based on entity information.
I will give you the properties of attributes and values of an entity in the form of (predicate object).
Using these information, please provide a short description of the entity.

- The explanation should be no longer than 100 words.
- Focus on summarizing the entity based on the given information.
- Do not include unnecessary details or explanations beyond the entity description.
- Do not include entity name.

Example:

Entity Information: (occupation Mathematician), (notable work Principia Mathematica)
Explanation: The entity was a prominent mathematician recognized for authoring Principia
Mathematica, a foundational work in the history of science and mathematics.

Now, please summarize the following entity information and return an desctription in English:
"""
```

Table 11: Prompt for summary

Image Matching Prompt

```
prompt = """
Here are two captions of different images:

Caption 1:
"{a group of photos shows different buildings}"

Caption 2:
"{many different images of the various building styles}"

Do they describe the same thing or topic? Answer only "Yes" or "No".
If Yes, summarize the shared theme in one sentence.
If No, leave the theme blank.
"""
```

Table 12: Prompt for image similarity judgment and shared theme extraction

Seeds	**Models**	**FBDB15K**			**FBYAGO15K**		
		Hits@1	**Hits@10**	**MRR**	**Hits@1**	**Hits@10**	**MRR**
20%	EVA	0.231	0.488	0.318	0.188	0.403	0.260
	MSNEA	0.149	0.392	0.232	0.138	0.346	0.210
	MCLEA	0.395	0.656	0.487	0.322	0.546	0.400
	MEAformer	0.578	0.812	0.661	0.444	0.692	0.529
	LGEA	**0.801**	**0.910**	**0.842**	**0.850**	**0.942**	**0.883**
50%	EVA	0.364	0.606	0.449	0.325	0.560	0.404
	MSNEA	0.358	0.656	0.459	0.376	0.646	0.472
	MCLEA	0.620	0.832	0.696	0.563	0.751	0.631
	MEAformer	0.690	0.871	0.755	0.612	0.808	0.682
	LGEA	**0.814**	**0.922**	**0.853**	**0.883**	**0.947**	**0.906**
80%	EVA	0.491	0.711	0.573	0.493	0.695	0.572
	MSNEA	0.565	0.810	0.651	0.593	0.806	0.668
	MCLEA	0.741	0.900	0.802	0.681	0.837	0.737
	MEAformer	0.784	0.921	0.834	0.724	0.880	0.783
	LGEA	**0.866**	**0.946**	**0.894**	**0.925**	**0.978**	**0.944**

Table 13: Iterative results on two cross-KG datasets are presented.