Bridging Reasoning and Retrieval: A Large-Small Model Collaboration Framework for Multimodal Entity Alignment

Anonymous ACL submission

Abstract

Multimodal Knowledge Graphs (MMKGs) provide structured multimodal grounding for Multimodal Large Language Models (MLLMs), while Multimodal Entity Alignment (MMEA) is essential for integrating heterogeneous MMKGs. However, existing MMEA methods remain vulnerable to noisy modalities and unreliable pseudo-supervision. We propose CRMEA, a large–small collaborative framework for unsupervised MMEA that unifies PTM-based retrieval and MLLM-based reasoning through a closed-loop optimization process. CRMEA introduces a text-centric adaptive cross-modal coordination mechanism to regulate noisy visual evidence, together with a retrieve-and-vote strategy for reliable pseudo-supervision. Extensive experiments show that CRMEA consistently outperforms existing unsupervised MMEA methods and achieves performance competitive with state-of-the-art supervised approaches, while maintaining strong robustness, adaptability across heterogeneous backbones, and substantially improved cost-effectiveness over exhaustive MLLM-based alignment. Code and data are available at https://anonymous.4open.science/r/CRMEA.

1 Introduction

The rapid advancement of Multimodal Large Language Models (MLLMs), such as ChatGPT Images2.0 (OpenAI, 2026) and Qwen3.6-27B (Alibaba Qwen Team, 2026), has greatly improved cross-modal understanding and reasoning (Zhao et al., 2024; E et al., 2025). However, MLLMs still suffer from hallucination due to their reliance on implicit parametric knowledge, often generating semantically plausible but factually incorrect outputs (Yang et al., 2025). To mitigate this issue, Multimodal Knowledge Graphs (MMKGs), which provide explicit multimodal representations of entities and relations (Zhang et al., 2025), have emerged as a promising external grounding source.

Effective MMKG-based grounding typically requires integrating heterogeneous MMKGs, where the same entity may be represented differently across platforms. Therefore, Multimodal Entity Alignment (MMEA) is essential for identifying equivalent entities and consolidating multimodal evidence (Chen et al., 2024b). A representative example is product-grounded shopping assistance, where Google's Shopping Graph aggregates over 50 billion products from retailers and merchants worldwide, many corresponding to the same real-world products but represented with different names, descriptions, images, and attributes (Google, 2025). MMEA-based fusion enables unified representations, providing a reliable foundation for future grounded MLLM systems.

MMEA methods have evolved from multimodal fusion to adaptive, robust alignment. Supervised approaches improve fusion through weighting, noise suppression, and auxiliary semantic signals (Li et al., 2025a; Huang et al., 2024; Feng et al., 2025), while unsupervised methods exploit cross-modal correlations via visual pivots, transformer-based fusion, and uncertainty-aware interactions (Liu et al., 2021; Chen et al., 2023a; Li et al., 2025b). More recently, LLM/MLLM-based methods have further enhanced MMEA by introducing semantic understanding and multimodal reasoning. For example, TMEA (Chen et al., 2024b) and LGEA (Lu et al., 2025a) employ LLMs for semantic abstraction and noise filtering, whereas MM-ChatAlign (Jiang et al., 2024) integrates MLLM-based multimodal reasoning with iterative refinement for alignment prediction.

However, two major challenges remain unresolved. First, how to robustly integrate noisy and incomplete multimodal evidence for MMEA. As illustrated in Figure 1 (Case A), uncontrolled multimodal fusion may overemphasize visual similarity, causing semantically different entities to be incorrectly aligned. Second, how to leverage MLLM

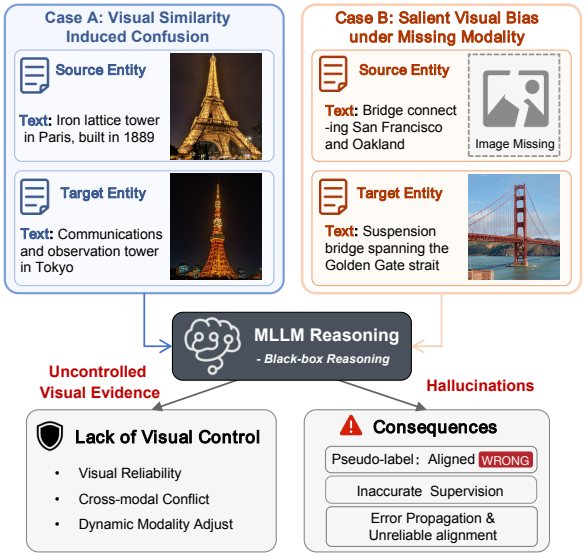


Figure 1: Motivating scenario.

reasoning to provide reliable supervision and decision in an unsupervised manner. Figure 1 (Case B) shows that missing modalities and unreliable visual evidence may mislead MLLM reasoning, resulting in hallucinated labels and error propagation.

A straightforward solution is to directly employ MLLMs for implicit multimodal fusion and pseudo-label generation in unsupervised MMEA. However, such approaches incur substantial computational overhead and remain vulnerable to hallucinated supervision, as MLLMs lack explicit mechanisms to assess visual reliability and regulate noisy or missing modalities. Consequently, unreliable multimodal evidence can be amplified during reasoning, resulting in unstable alignment decisions and cumulative error propagation.

To address these challenges, we propose **C**ooperative **R**easoning for M**MEA** (CRMEA), a large–small model collaboration framework for unsupervised MMEA. CRMEA bridges efficient PTM-based retrieval and reliable MLLM-based reasoning through a closed-loop optimization process. Specifically, PTMs first retrieve compact candidate sets, after which MLLMs perform candidate-constrained reasoning and voting to generate high-confidence alignments for contrastive self-training. Through iterative collaboration, CRMEA progressively distills MLLM's reasoning capability into lightweight PTMs, enabling retrieval refinement while reducing the cost of MLLM invocations.

Furthermore, CRMEA introduces a text-centric cross-attention mechanism with adaptive gating to dynamically regulate visual evidence under noisy or missing modalities. To improve pseudo-supervision reliability, CRMEA further adopts a retrieve-and-vote strategy with multi-round majority voting to suppress hallucination-induced error propagation during iterative self-training. Together, these designs enable robust and cost-effective unsupervised MMEA.

Our contributions are summarized as follows:

- We propose CRMEA, a large–small collaborative framework for unsupervised MMEA, which enables iterative mutual refinement between PTM retrieval and MLLM reasoning in a closed-loop optimization manner.
- We introduce a text-centric adaptive cross-modal coordination that dynamically regulates visual evidence through cross-attention and adaptive gating, improving robustness against noisy and incomplete modalities.
- We develop a retrieve-and-vote pseudo-supervision strategy that stabilizes MLLM-assisted self-training through candidate-constrained reasoning and majority voting.
- Extensive experiments demonstrate that CRMEA achieves state-of-the-art unsupervised MMEA performance competitive with supervised approaches, while maintaining strong robustness and cost-effectiveness across diverse multimodal settings and heterogeneous PTM/MLLM backbones.

2 Related Work

2.1 Multimodal Fusion for Entity Alignment

Multimodal Entity Alignment (MMEA) incorporates textual and visual information to address cross-lingual textual discrepancies and structural sparsity. Early methods such as HMEA (Guo et al., 2021), EVA (Liu et al., 2021), and MSNEA (Chen et al., 2022) mainly project heterogeneous modalities into a shared embedding space and perform holistic alignment using unified representations.

Recent studies have shifted toward more interaction-aware, robustness-oriented fusion strategies. MCLEA (Lin et al., 2022) improves alignment through multimodal contrastive objectives, while MEAformer (Chen et al., 2023a) employs Transformers to capture fine-grained entity interactions and alleviate noise sensitivity. To mitigate modality inconsistency, PMF (Huang

et al., 2024) progressively suppresses alignment-irrelevant features, while UMAEA (Chen et al., 2023b) and DESAlign (Wang et al., 2024) explicitly account for noisy or missing modalities. Nevertheless, most existing approaches still rely on tightly coupled multimodal fusion, in which unreliable visual signals may be propagated into the alignment representation, leading to negative transfer. To address this limitation, our method anchors alignment in textual semantics while introducing visual evidence in a controlled manner, improving robustness against noisy and inconsistent modalities.

2.2 Seed-free Iterative Alignment

Seed-free MMEA methods learn alignments without human annotations by iteratively bootstrapping pseudo-labels from structural and multimodal similarities. Representative methods such as EVA (Liu et al., 2021) and MCLEA (Lin et al., 2022) progressively expand pseudo-alignments through self-training, while MEAformer (Chen et al., 2023a) improves candidate quality via cross-modal interaction modeling. However, similarity-based expansion strategies remain sensitive to modality noise and mismatches, causing error accumulation across iterations. RICEA (Li et al., 2025b) mitigates this issue through more robust pseudo-supervision refinement, but repeated retrieval and selection still introduce substantial overhead and residual bias under ambiguous candidates.

Building on this observation, our framework compresses the candidate space and introduces reliable pseudo-supervision to reduce training cost and stabilize iterative alignment.

2.3 (M)LLM-enhanced MMEA

LLM- or MLLM-enhanced MMEA exploits semantic reasoning to complement similarity-based retrieval. TMEA (Chen et al., 2024b) uses LLM-based in-context learning for attribute alignment and filtering under uncertain cross-modal correspondence, whereas LGEA (Lu et al., 2025a) employs LLM-guided filtering or enhancement to mitigate modality noise before fusion. MM-ChatAlign (Jiang et al., 2024) performs reasoning and reflective re-evaluation on a compressed candidate set, and converts the reasoning outcomes into alignment decisions through filtering and re-ranking.

Overall, existing (M)LLM-based approaches mainly rely on reasoning signals for filtering, re-ranking, or semantic augmentation, while their integration with representation-learning objectives remains largely heuristic. Consequently, loosely grounded reasoning outputs may introduce uncalibrated decisions and biased pseudo-labels, which can be amplified during self-training in entity alignment. Moreover, scalable deployment requires careful control of LLM invocation cost and mitigation of decision bias under uncertain reasoning. Motivated by these challenges, our framework bridges retrieval and reasoning through a cost-aware, iteration-stable pipeline, enabling reliable MLLM assistance without sacrificing scalability.

3 Problem Definition

In this section, we formally define the Multimodal Knowledge Graph (MMKG) and the Multimodal Entity Alignment (MMEA) problem.

Definition 1. (MMKG) An MMKG is defined as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{I}, \mathcal{A}, \mathcal{V}, \mathcal{T}_r, \mathcal{T}_a\}$, where $\mathcal{E}$, $\mathcal{R}$, $\mathcal{I}$, $\mathcal{A}$, $\mathcal{V}$, $\mathcal{T}_r$, $\mathcal{T}_a$ denote the set of entities, relations, images, attributes, and textual values, respectively. $\mathcal{T}_r = \{(e_i, r, e_j) | e_i, e_j \in \mathcal{E}, r \in \mathcal{R}\} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ and $\mathcal{T}_a = \{(e, a, p) | e \in \mathcal{E}, a \in \mathcal{A}, p \in (\mathcal{I} \cup \mathcal{V})\} \subseteq \mathcal{E} \times \mathcal{A} \times (\mathcal{I} \cup \mathcal{V})$ represent the set of relational and attribute triples.

Definition 2. (MMEA) Given two MMKGs $\mathcal{G}_1 = \{\mathcal{E}_1, \mathcal{R}_1, \mathcal{I}_1, \mathcal{A}_1, \mathcal{V}_1, \mathcal{T}_{r_1}, \mathcal{T}_{a_1}\}$ and $\mathcal{G}_2 = \{\mathcal{E}_2, \mathcal{R}_2, \mathcal{I}_2, \mathcal{A}_2, \mathcal{V}_2, \mathcal{T}_{r_2}, \mathcal{T}_{a_2}\}$, MMEA aims to figure out each entity pair $(e_1, e_2), e_1 \in \mathcal{E}_1, e_2 \in \mathcal{E}_2$, where e_1 and e_2 are semantically equivalent and point to the same physical or virtual object.

Furthermore, MMEA can be categorized into supervised and unsupervised MMEA depending on whether the human-labeled data (i.e., seed alignments) is utilized in the training process. In this paper, we concentrate on the **unsupervised MMEA**, eliminating the need for costly human annotation.

4 Proposed Approach

In this section, we present CRMEA, an unsupervised MMEA framework based on a closed-loop optimization architecture. As shown in Figure 2, CRMEA leverages the reasoning capability of MLLMs to iteratively refine specialized PTMs without human annotation through collaborative retrieval and reasoning. The framework consists of three synergistic stages: (1) PTM-based candidate retrieval; (2) MLLM reasoning and voting for pseudo-label generation; and (3) PTM self-training for representation refinement.

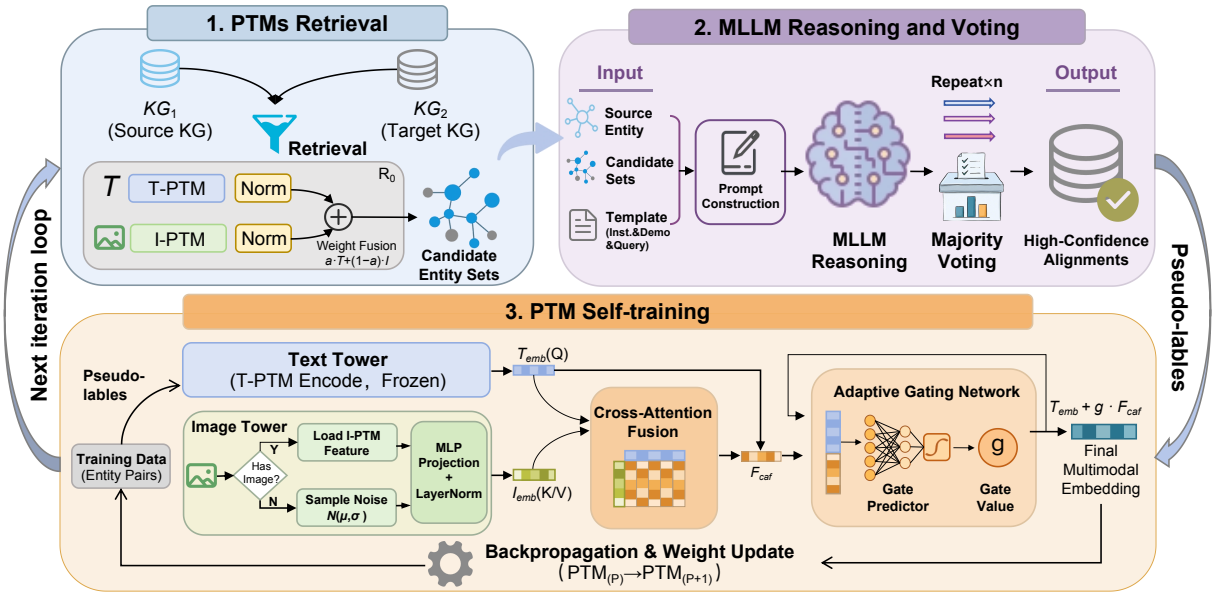


Figure 2: Overall architecture of the proposed CRMEA.

4.1 PTMs Retrieval

Given a source knowledge graph $\mathcal{G}_s$ and a target $\mathcal{G}_t$, this phase retrieves a concise candidate set from $\mathcal{G}_t$. We first flatten the structural and relational neighborhood of each entity into a token sequence, feeding it to a language PTM (e.g., LaBSE) to obtain a semantic embedding $\mathbf{h}_t = \text{Norm}(\text{L-PTM}(S_e)) \in \mathbb{R}^d$. Simultaneously, associated images are processed via an image PTM (e.g., ResNet-152) and linearly projected to derive the visual embedding $\mathbf{h}_v = \text{Norm}(\mathbf{W}_p \cdot \text{I-PTM}(I_e)) \in \mathbb{R}^d$.For self-training feedback loops ($r \geq 1$), we fuse these modalities using a text-anchored Cross-Attention module. The textual embedding serves as the Query to retrieve informative visual features:

$$\Delta\mathbf{h} = \text{Softmax}\left(\frac{(\mathbf{h}_t\mathbf{W}Q)(\mathbf{h}v\mathbf{W}K)^T}{\sqrt{d}}\right)(\mathbf{h}v\mathbf{W}_V) \tag{1}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are trainable operators, and $\Delta\mathbf{h}$ is the cross-modal refinement increment.To resolve textual ambiguities without collapsing into a text-dominant bias, we project the concatenated features through an MLP to derive an instance-specific gating intensity:

$$g = \sigma(\text{MLP}([\mathbf{h}t; \beta \cdot \Delta\mathbf{h}])) \tag{2}$$

where $g \in [0, 1]$ is the dynamic gating scalar. The integrated multi-modal representation $\mathbf{h}_{final}$ is then formulated via a scaled residual connection:

$$\mathbf{h}_{final} = \mathbf{h}_t + g \cdot (\beta \cdot \Delta\mathbf{h}) \tag{3}$$

This allows the visual modality to adaptively play a decisive role when textual attributes are uninformative. Finally, candidate pairs are prioritized using inner product affinity $\text{sim}(e_s, e_t) = \mathbf{h}^{(s)}_{final} \cdot \mathbf{h}^{(t)}_{final}$ to form the top-N candidate pool $\mathcal{C}_{cand}(e_s)$.

4.2 MLLM Reasoning and Voting

We leverage an MLLM (e.g., Qwen3-32B VL) to identify the semantically equivalent counterpart from $\mathcal{C}_{cand}(e_s)$ via in-context learning. To handle the “missing modality" issue robustly, we insert a pure white dummy image as a neutral placeholder. Prompt instructions explicitly direct the MLLM to disregard these placeholders, safeguarding the attention mechanism against visual voids.To mitigate stochastic hallucinations, we employ a multi-round majority voting mechanism. Over R independent reasoning rounds, the final predicted entity is determined by $e_{pred} = \text{argmax}_{c \in \mathcal{C}_{cand}} \sum_{r=1}^{R} \mathbb{I}(v_r = c)$. If no clear majority emerges, the sample is discarded to prevent noise propagation.

4.3 PTM Self-Training

The PTMs are optimized via contrastive learning using triplets (e_s, e_{pos}, e_{neg}), where the positive sample e_{pos} is the MLLM-validated e_{pred}, and the hard negative e_{neg} is the highest-scoring incorrect candidate in $\mathcal{C}_{cand}$ based on current PTM similarities.A pairwise margin loss enforces separation between positive and negative pairs:

$$\mathcal{L}_{align} = [\gamma - \text{sim}(\mathbf{h}_s, \mathbf{h}_{pos}) + \text{sim}(\mathbf{h}_s, \mathbf{h}_{neg})]_+ \tag{4}$$

where γ is a hyper-parameter controlling the margin, and $[\cdot]_+ = \max(0, \cdot)$ denotes the standard hinge function. This objective enforces the similarity of the validated positive pair to exceed that of the hard negative pair by at least the margin γ.

To learn a globally discriminative space, an InfoNCE loss is applied over the batch $\mathcal{B}$:

$$\mathcal{L}_{nce} = -\log \frac{\exp(\text{sim}(\mathbf{h}_s, \mathbf{h}_{pos})/\tau_{nce})}{\sum_{e_k \in \mathcal{B}} \exp(\text{sim}(\mathbf{h}_s, \mathbf{h}_k)/\tau_{nce})} \tag{5}$$

Crucially, to prevent the adaptive gating scalars g_i from collapsing into stochastic fluctuations, a gate-distribution regularization loss anchors the empirical expectation to a target prior:

$$\mathcal{L}_{reg} = \left(\frac{1}{|\mathcal{B}|} \sum e_i \in \mathcal{B} g_i - \tau_{reg} \right)^2 \tag{6}$$

The final joint objective function minimizes these components within the Hard-EM framework:

$$\mathcal{L}_{total} = \mathcal{L}_{align} + \lambda_1 \mathcal{L}_{nce} + \lambda_2 \mathcal{L}_{reg} \tag{7}$$

This establishes a principled positive feedback loop, ensuring higher-quality retrievals in subsequent iterations.

5 Experiments

Extensive experiments on multiple benchmark datasets were conducted to evaluate CRMEA for unsupervised MMEA. Specifically, we investigate the following questions: (Q1) Does CRMEA outperform state-of-the-art unsupervised MMEA methods? (Q2) Can CRMEA provide reliable MLLM-based supervision in an unsupervised setting? (Q3) Can CRMEA robustly handle noisy and incomplete multimodal evidence? (Q4) How do different components and modalities contribute to alignment performance? and (Q5) How cost-effective is CRMEA compared with exhaustive MLLM-only alignment?

5.1 Experimental Settings

Datasets. We evaluated CRMEA on two widely used MMEA benchmarks, DBP15K (Sun et al., 2017) and OpenEA (Sun et al., 2020). DBP15K contains three cross-lingual subsets (ZH–EN, JA–EN, and FR–EN), while OpenEA includes EN–FR and EN–DE. Following prior work, we adopted the image sets from EVA (Liu et al., 2021) and UMAEA (Chen et al., 2023b), respectively. Entity names were excluded, and only attribute-relation triples were used.

Experimental Environment. We implemented CRMEA using Python (version 3.9.18) with a PyTorch (version 2.1.2) backend. Experiments were conducted on a server running Ubuntu 22.04 with an AMD Ryzen 9 7950X Processor, 128GB of memory, and an NVIDIA RTX A6000 GPU.

Parameter Settings. We used LaBSE (d=768) and Qwen3-VL-32B as the main backbones with a maximum sequence length of 8,192. Visual features were projected to 768 dimensions, with missing images padded using $N(\mu, \sigma)$. The fusion module employed 8-head cross-attention with a dropout of 0.3, and the adaptive gating was initialized with a bias of 2.0 and a regularization target of 0.7. Training used AdamW with tiered learning rates ($1e^{-6}$ for backbones), allocating 20% of the data for self-training and 80% for test, with five MLLM reasoning repetitions and top-k candidates in $[1, 10]$.

Baselines. We compared CRMEA with seven representative MMEA baselines. These include three recent supervised models (VRCMC (Shi et al., 2026), SGMEA (Cheng et al., 2025), and LGEA (Lu et al., 2025b)) and four unsupervised models (EVA (Liu et al., 2021), MSNEA (Chen et al., 2022), MCLEA (Lin et al., 2022), and MEAformer (Chen et al., 2023a)), ensuring a comprehensive evaluation and a fair comparison under the unsupervised multimodal setting.

Evaluation Metrics. Following standard MMEA evaluation protocols (Jin et al., 2025), we adopt Hit@k (k=1,10), Mean Reciprocal Rank (MRR), and LLM Accuracy (LA) as evaluation metrics. Hit@k measures top-k retrieval accuracy, MRR evaluates overall ranking quality, and LA (Hit@1/Hit@k) reflects the MLLM's precision within filtered candidates.

5.2 Experimental Results

(Q1) Main Results on Benchmark Datasets

We first evaluate CRMEA against representative unsupervised and supervised MMEA baselines on the DBP15K and OpenEA datasets to assess its overall alignment effectiveness. As shown in Table 1, CRMEA consistently outperforms all existing unsupervised MMEA methods on the DBP15K benchmark. On the challenging ZH–EN setting, CRMEA improves Hit@1 from 0.909 to 0.961 (+5.7%), demonstrating strong robustness under large linguistic and semantic discrepancies. On

Table 1: Performance comparison with unsupervised and supervised MMEA methods on the **DBP15K** dataset. Bold and underlined numbers indicate the best and runner-up scores among the unsupervised setting.

Setting	Method	DBP15K$_{ZH-EN}$ Hit@1	DBP15K$_{ZH-EN}$ MRR	DBP15K$_{JA-EN}$ Hit@1	DBP15K$_{JA-EN}$ MRR	DBP15K$_{FR-EN}$ Hit@1	DBP15K$_{FR-EN}$ MRR	Average Hit@1
Unsupervised	EVA (Liu et al., 2021)	.891	.917	.941	.958	.970	.982	.934
	MSNEA (Chen et al., 2022)	.858	.887	.921	.939	.954	.968	.911
	MCLEA (Lin et al., 2022)	.860	.893	.914	.938	.953	.967	.909
	MEAformer (Chen et al., 2023a)	.909	.933	.950	**.965**	.972	.983	.944
	CRMEA (Ours)	**.961**	**.962**	**.957**	.958	**.995**	**.996**	**.971**
	- improvement	+5.7%	+3.1%	+0.7%	-0.7%	+2.4%	+1.3%	+2.9%
Supervised	VRCMC (Shi et al., 2026)	.901	.931	.892	.929	.906	.936	.900
	SGMEA (Cheng et al., 2025)	.899	.931	.901	.933	.917	.945	.906
	LGEA (Lu et al., 2025b)	.970	.982	.962	.977	.975	.985	.969

Table 2: Performance comparison with unsupervised MMEA methods on the **OpenEA** dataset. Bold and underlined numbers indicate the best and runner-up scores.

Method	OpenEA$_{EN-FR}$ Hit@1	OpenEA$_{EN-FR}$ MRR	OpenEA$_{EN-DE}$ Hit@1	OpenEA$_{EN-DE}$ MRR
MSNEA	.699	.742	.788	.835
EVA	.849	.896	.956	.968
MCLEA	.888	.924	.969	.979
CRMEA (Ours)	**.982**	**.985**	**.990**	**.991**
- improvement	+10.6%	+6.6%	+2.2%	+1.2%

FR–EN, CRMEA achieves the best overall performance, reaching a near-perfect Hit@1 of 0.995 and the highest MRR of 0.996. On JA–EN, where baseline methods already approach saturation, CRMEA remains highly competitive and achieves comparable results. Overall, CRMEA improves the average Hit@1 from 0.944 to 0.971 (+2.9%), validating the effectiveness and robustness of the proposed unsupervised MMEA approach.

Notably, although CRMEA operates in a fully unsupervised setting without human intervention, its performance is highly competitive with state-of-the-art supervised MMEA methods. As shown in Table 1, CRMEA surpasses supervised methods such as VRCMC and SGMEA across most settings, and even approaches the strongest supervised baseline LGEA, with only marginal performance gaps on average Hit@1 (0.971 vs. 0.969). These results demonstrate that the proposed large–small collaborative framework can effectively compensate for the absence of supervised annotations through reliable MLLM-assisted pseudo-supervision and iterative retrieval–reasoning refinement.

As shown in Table 2, CRMEA consistently outperforms existing unsupervised MMEA methods on the OpenEA benchmark. On EN–FR, CRMEA improves Hit@1 from 0.888 to 0.982, achieving a substantial relative gain of 10.6%, while MRR increases by 6.6%. Similar improvements are observed on EN–DE, where CRMEA reaches a near-saturated Hit@1 of 0.990 and surpasses the strongest baseline by 2.2%. These results demonstrate the effectiveness of the proposed large–small model collaboration framework in achieving accurate and robust multimodal entity alignment.

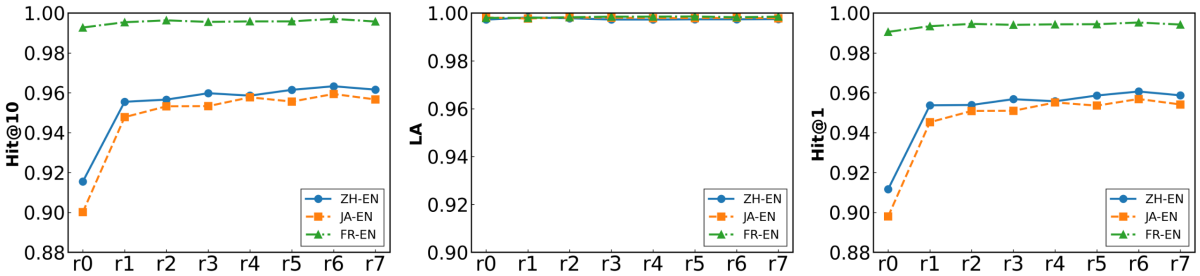
(Q2) Effectiveness of MLLM-PTM Collaboration

To further examine whether MLLM reasoning and PTM retrieval can mutually reinforce each other, we analyzed their iterative collaboration in Figure 3. As shown in Figure 3(a), PTM-based retrieval improved substantially after the first round, indicating that MLLM-generated pseudo-supervision effectively reshaped the PTM embedding space and yielded higher-quality candidate sets. Meanwhile, Figure 3(b) shows that the MLLM maintained highly stable LA across iterations, indicating that improved PTM retrieval effectively constrains the candidate space and stabilizes MLLM reasoning.

This mutual enhancement further led to steady gains in overall MMEA performance, as shown in Figure 3(c). The alignment curved closely follows the retrieval improvements and converged smoothly without degradation, demonstrating that CRMEA forms a stable positive feedback loop. Specifically, PTMs provide increasingly accurate candidates for MLLM reasoning, while MLLM supervision continuously improves PTM representations.

(Q3) Effectiveness of Adaptive Cross-Modal Coordination

We further evaluate whether CRMEA can robustly integrate multimodal evidence under noisy, incom-



(a) PTM Retrieval (b) MLLM Reasoning (c) Overall MMEA

Figure 3: Effectiveness of the iterative MLLM-PTM collaboration. The curves demonstrate the progressive improvement in retrieval and alignment accuracy (a, c) alongside the high reliability and stability of the MLLM supervision signal (b).

Table 3: Performance of CRMEA on DBP15K$_{ZH-EN}$ under different visual modality availability conditions.

Category	Quantity	Hit@1	MRR	LA
Dual-image	8089	.968	.968	.999
Single-image	2367	.955	.956	.997
No-image	1544	.941	.944	.993

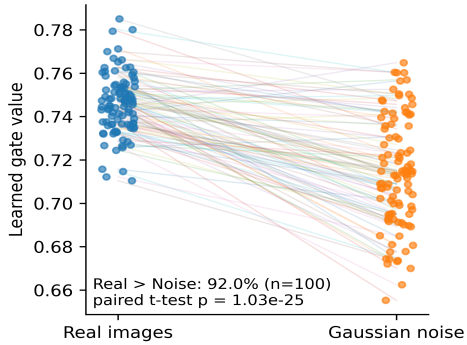


Figure 4: Comparison of learned gate value distributions for real images (0.746) and Gaussian Noise (0.714).

Table 4: Ablation study results. "w/o Att", "w/o Relation", "w/o Image" indicate removing attribute, relation, and visual features. "w/o MLLM" removes the iterative MLLM reasoning and voting phase, and "w/o Gate" disables the adaptive gating mechanism.

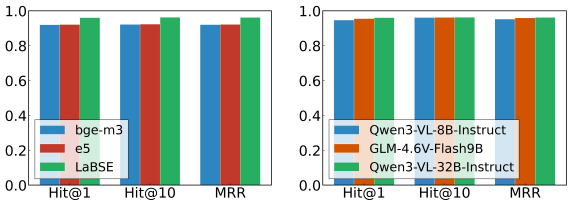
Model	Hit@1	MRR
- w/o Attribute	.879 (-8.5%)	.881 (-8.4%)
- w/o Relation	.909 (-5.4%)	.910 (-5.4%)
- w/o Image	.946 (-1.6%)	.947 (-1.6%)
- w/o MLLM	.813 (-15.4%)	.851 (-11.5%)
- w/o Gate	.951 (-1.0%)	.952 (-1.0%)
CRMEA (Full)	.961	.962

plete visual conditions. Table 3 shows the robustness of CRMEA under different visual modality availability conditions. Under the Dual-image setting, where both entities contain images, CRMEA achieves the best performance, indicating that visual information provides complementary alignment evidence beyond textual semantics alone. More importantly, under the Single-image setting, where only one entity contains an image, CRMEA still maintains strong alignment accuracy with only marginal degradation. This demonstrates that the proposed text-centric fusion and adaptive modality coordination mechanisms can robustly handle incomplete multimodal evidence without over-relying on visual inputs. In contrast, performance decreases more noticeably under the No-image setting, further confirming the positive contribution of visual modalities to multimodal entity alignment.

Figure 4 shows that our adaptive gating network adjusted cross-modal fusion in a content-aware manner. When entities are associated with real images, the learned gate values were consistently higher than those obtained under Gaussian-noise images, with a significant decrease in the mean from 0.746 to 0.714 (paired test p = $1.03e^{-25}$). This shift indicates that the gating network dynamically modulated the contribution of visual features based on semantic reliability, boosting informative visual cues while suppressing noisy ones, thereby preserving the text-anchored representation.

(Q4) Analysis of Key Components and Modalities

To evaluate the adaptability and model-agnostic nature of CRMEA, we conduct backbone substitution experiments across both PTMs and MLLMs. As shown in Figure 5(a), CRMEA consistently maintains strong alignment performance under different text encoders, including BGE (Chen et al.,



(a) PTMs (b) MLLMs

Figure 5: Robustness analysis of CRMEA under different PTM and MLLM backbones.

Table 5: Cost comparison between CRMEA and an exhaustive MLLM-only alignment strategy on DBP15K. CRMEA reduced overall cost by approximately 685.7×. The estimation assumes 10 candidate pairs, 2 seconds, and 3K tokens per MLLM query on average.

Metric	**CRMEA (0.15%)**	MLLM-only
MLLM query	1.05M	720M
Token consumption	315M	216B
Runtime	210K seconds	144M seconds

2024a), E5 (Wang et al., 2022), and LaBSE, demonstrating that the proposed text-centric fusion and adaptive coordination mechanisms generalize well across heterogeneous embedding spaces. Although stronger PTMs yield higher absolute performance, the overall performance trend remains stable, indicating that CRMEA does not rely on a specific encoder architecture.

Figure 5(b) shows that CRMEA achieves robust alignment accuracy across different MLLM backbones, ranging from lightweight Qwen3-VL-8B and GLM-4V-9B (Glm et al., 2024) to larger Qwen3-VL-32B. Notably, the performance gaps between MLLMs remain relatively small, suggesting that CRMEA effectively stabilizes reasoning quality and mitigates stochastic hallucinations through candidate-space compression and majority voting. These results demonstrate that the proposed large–small collaborative framework possesses strong adaptability, modularity, and flexibility under diverse PTM and MLLM configurations.

To test the contribution of different design choices, we conduct ablation studies by selectively removing textual, visual, and collaborative reasoning components. As shown in Table 4, removing textual attributes (w/o Attribute) causes a substantial performance drop, reducing Hit@1 by 8.5%, which confirms that attribute-level semantics serve as the primary alignment anchor beyond structural information. Removing relation information (w/o Relation) also leads to noticeable degradation (-5.4%), highlighting the importance of structural context for entity alignment.

Beyond modality information, the collaborative reasoning mechanism also plays a critical role. Removing the MLLM collaboration (w/o MLLM) results in the greatest degradation, reducing Hit@1 from 0.961 to 0.813 (-15.4%). This demonstrates that PTM retrieval alone is insufficient for unsupervised MMEA, while MLLM reasoning and majority voting provide essential high-confidence supervision for stable representation learning. Furthermore, disabling the adaptive gating mechanism (w/o Gate) consistently reduces performance, validating its effectiveness in regulating visual information and preserving text-centric representations.

(Q5) Cost-effectiveness Analysis

To verify the practical efficiency of CRMEA, we compared its computational cost with exhaustive MLLM-only alignment on DBP15K. As shown in Table 5, CRMEA substantially reduces the cost of MLLM-based alignment by first compressing the candidate space through PTM retrieval and then selectively applying MLLM reasoning to informative candidate pairs. Specifically, CRMEA requires only 1.05M MLLM queries, incurring 315M tokens and 210K seconds, whereas exhaustive MLLM-only alignment requires 720M MLLM queries, 216B tokens, and 144M seconds. Overall, CRMEA reduces the cost by about 685.7 times, demonstrating strong cost-effectiveness and scalability while maintaining superior alignment performance.

6 Conclusion

In this paper, we propose CRMEA, a large–small collaborative framework for unsupervised multimodal entity alignment that unifies PTM retrieval and MLLM reasoning through closed-loop optimization. By combining text-centric adaptive fusion and reliable MLLM-assisted pseudo-supervision, CRMEA achieves robust, scalable, and cost-effective multimodal entity alignment under diverse multimodal conditions and heterogeneous backbone configurations. Extensive experiments demonstrate its strong effectiveness across diverse datasets and heterogeneous backbones.

Future work will explore richer multimodal settings and uncertainty-aware reasoning mechanisms to further improve robustness and scalability in real-world MMKG applications.

Limitations

Despite its effectiveness, CRMEA still has several limitations. First, CRMEA relies on PTM-based candidate retrieval before MLLM reasoning. Although candidate-space compression substantially improves efficiency, the framework remains sensitive to retrieval recall, since missing the correct counterpart during candidate generation may irreversibly affect subsequent reasoning and self-training. Second, CRMEA adopts a text-centric fusion strategy to improve robustness against noisy visual evidence. However, when textual descriptions are highly sparse or semantically ambiguous, excessive reliance on textual anchors may limit the contribution of visual modalities. Third, although CRMEA is significantly more cost-effective than exhaustive MLLM-based alignment, its practical deployment may still depend on external MLLM services, including inference latency, API pricing, and model-version stability. Finally, our experiments mainly focus on standard MMEA benchmarks. Further evaluation on larger-scale and real-world MMKGs with dynamic modality availability and evolving entity distributions is needed to fully verify the generalizability of CRMEA.

AI Assistants in Research or Writing

We used AI assistants, including ChatGPT and Gemini, to assist with language polishing, grammar correction, phrasing refinement, LaTeX formatting, and code debugging during the preparation of this manuscript. All AI-generated suggestions were manually reviewed and revised by the authors. The AI assistants were not used to generate research ideas, conduct experiments, or make final scientific claims.

References

Alibaba Qwen Team. 2026. Qwen3.6-27B: Dense 27b multimodal open-weight model. https://huggingface.co/Qwen/Qwen3.6-27B. Accessed: 2026-05-26.

Jianly Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Liyi Chen, Zhi Li, Tong Xu, Han Wu, Zhefeng Wang, Nicholas Jing Yuan, and Enhong Chen. 2022. Multi-modal siamese network for entity alignment. In *Proc. of ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 118–126.

Liyi Chen, Ying Sun, Shengzhe Zhang, Yuyang Ye, Wei Wu, and Hui Xiong. 2024b. Tackling uncertain correspondences for multi-modal entity alignment. *Advances in Neural Information Processing Systems*, 37:119386–119410.

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 *Proc. of ACM Int'l Conf. on Multimedia (ACMMM)*, 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 *Proc. of Int'l Semantic Web Conf. (ISWC)*, pages 121–139. Springer.

Jingwei Cheng, Mingxiao Guo, and Fu Zhang. 2025. Sgmea: Structure-guided multimodal entity alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7851–7861.

Shaojun E, Yuchen Yang, Jiaheng Wu, Yan Zhang, Tiejun Zhao, and Ziyan Chen. 2025. Mage: Multimodal alignment and generation enhancement via bridging visual and semantic spaces. In *Proc. of Int'l Joint Conf. on Artificial Intelligence (IJCAI)*, pages 954–962.

Guan Dong Feng, Tao Ren, Jun Hu, and Dan dan Wang. 2025. Cateea: Enhancing entity alignment via implicit category supervision. In *Proc. of Int'l Conf. on Computational Linguistics (COLING)*, pages 5975–5986.

Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Google. 2025. AI transforms shopping in search. https://business.google.com/uk/think/search-and-video/google-shopping-ai-mode-virtual-try-on-update/. Accessed: 2026-05-26.

Hao Guo, Jiuyang Tang, Weixin Zeng, Xiang Zhao, and Li Liu. 2021. Multi-modal entity alignment in hyperbolic space. *Neurocomputing*, 461:598–607.

Yani Huang, Xuefeng Zhang, Richong Zhang, Junfan Chen, and Jaein Kim. 2024. Progressively modality freezing for multi-modal entity alignment. In *Proc. of Int'l Conf. on Computational Linguistics (ACL)*, pages 3477–3489.

Xuhui Jiang, Yinghan Shen, Zhichao Shi, Chengjin Xu, Wei Li, Huang Zihe, Jian Guo, and Yuanzhuo Wang. 2024. Mm-chatalign: A novel multimodal reasoning framework based on large language models for entity alignment. In *Proc. of Findings of the Assoc. for Computational Linguistics: EMNLP (Findings of EMNLP)*, pages 2637–2654.

Xiongnan Jin, Zhilin Wang, Jinpeng Chen, Liu Yang, Byungkook Oh, Seung-won Hwang, and Jianqiang Li. 2025. Hlmea: Unsupervised entity alignment based on hybrid language models. In *Proc. of AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 11888–11896.

Chenxiao Li, Jingwei Cheng, Qiang Tong, and Fu Zhang. 2025a. Exploring the impacts of feature fusion strategy in multi-modal entity alignment. In *Proc. of Int'l Conf. on Computational Linguistics (ACL)*, pages 7809–7818.

Chenxiao Li, Jingwei Cheng, Qiang Tong, Fu Zhang, and Cairui Wang. 2025b. Probing relative interaction and dynamic calibration in multi-modal entity alignment. In *Proc. of Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 28504–28516.

Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multi-modal contrastive representation learning for entity alignment. In *Proc. of Int'l Conf. on Computational Linguistics (COLING)*, pages 2572–2584.

Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual pivoting for (unsupervised) entity alignment. In *Proc. of AAAI Conf. on Artificial Intelligence (AAAI)*, volume 35, pages 4257–4266.

Chenglong Lu, Chenxiao Li, Jingwei Cheng, Yongquan Ji, Guoqing Chen, and Fu Zhang. 2025a. Breaking the noise barrier: Llm-guided semantic filtering and enhancement for multi-modal entity alignment. In *Proc. of Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 33141–33155.

Chenglong Lu, Chenxiao Li, Jingwei Cheng, Yongquan Ji, Guoqing Chen, and Fu Zhang. 2025b. Breaking the noise barrier: Llm-guided semantic filtering and enhancement for multi-modal entity alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33141–33155.

OpenAI. 2026. Introducing ChatGPT images 2.0. https://openai.com/index/introducing-chatgpt-images-2-0/. Accessed: 2026-05-26.

Xinchen Shi, Bin Li, and Xiaowei Zhang. 2026. Vrcmc: a semi-supervised method for multi-modal entity alignment with visual refinement and cross-modal calibration. *Journal of King Saud University Computer and Information Sciences*.

Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *Proc. of Int'l Semantic Web Conf. (ISWC)*, pages 628–644. Springer.

Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. 2020. A benchmarking study of embedding-based entity alignment for knowledge graphs. *arXiv preprint arXiv:2003.07743*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Yuanyi Wang, Haifeng Sun, Jiabo Wang, Jingyu Wang, Wei Tang, Qi Qi, Shaoling Sun, and Jianxin Liao. 2024. Towards semantic consistency: Dirichlet energy driven robust multi-modal entity alignment. In *Proc. of IEEE Int'l Conf. on Data Engineering (ICDE)*, pages 3559–3572. IEEE.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. 2025. Multiple heads are better than one: Mixture of modality knowledge experts for entity representation learning. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*.

Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, and Xiangyu Zhang. 2024. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. In *Proc. of Int'l Joint Conf. on Artificial Intelligence (IJCAI)*, pages 1743–1752.