

MM-ChatAlign: A Novel Multimodal Reasoning Framework based on Large Language Models for Entity Alignment

Anonymous ACL submission

Abstract

Multimodal entity alignment (MMEA) integrates multi-source and cross-modal knowledge graphs, a crucial yet challenging task for data-centric applications. Traditional MMEA methods derive the visual embeddings of entities and combine them with other modal data for alignment by embedding similarity comparison. However, these methods are hampered by the limited comprehension of visual attributes and deficiencies in realizing and bridging the semantics of multimodal data. To address these challenges, we propose MM-ChatAlign, a novel framework that utilizes the visual reasoning abilities of MLLMs for MMEA. The framework features an embedding-based candidate collection module that adapts to various knowledge representation strategies, effectively filtering out irrelevant reasoning candidates. Additionally, a reasoning and rethinking module, powered by MLLMs, enhances alignment by efficiently utilizing multimodal information. Extensive experiments on four MMEA datasets demonstrate MM-ChatAlign’s superiority and underscore the significant potential of MLLMs in MMEA tasks. The source code is available at <https://anonymous.4open.science/r/MMEA/>.

1 Introduction

Multimodal entity alignment (MMEA) aligns equivalent entities across diverse multimodal knowledge graphs (MMKGs) (Zhu et al., 2022), playing a key role in synthesizing heterogeneous data for data-centric applications. Unlike traditional entity alignment, MMEA necessitates the integration of information across various modalities and MMKGs, thereby imposing higher demands on the visual reasoning ability of MMEA methods.

Current representative MMEA methods (Liu et al., 2021; Lin et al., 2022; Zhu et al., 2023; Xu et al., 2023; Chen et al., 2023) mainly adopt knowledge representation learning (KRL) and measure the similarity of entity embeddings for MMEA.

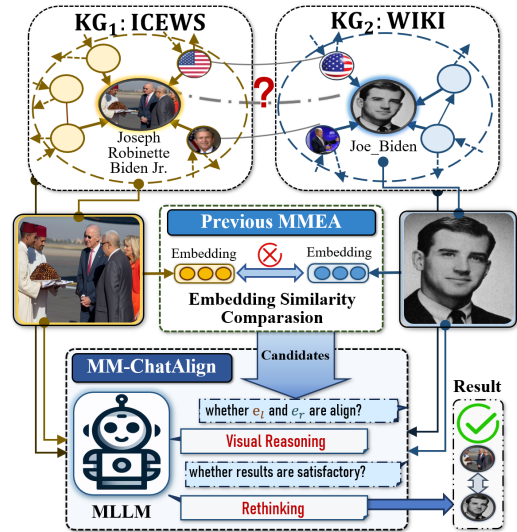


Figure 1: Comparison between the previous MMEA and MM-ChatAlign.

As shown in Figure 1, these methods face critical challenges. Firstly, their reliance on the representation learning approaches, which convert various attributes of entities into embeddings respectively, weakens the insight into underlying connections in visual attributes. As shown in Figure 1, the left image in ICEWS shows a scene of President Biden talking to leaders from other countries, and the right image in WIKI is a portrait of young Biden. Although both images are related to *Biden*, they have different embeddings for their contrasting representation. Existing embedding-based MMEA methods lack the visual reasoning capability to explicitly recognize that two images represent visual attributes of *Biden* at different times. Secondly, high-quality MMEA demands the learning of associations with visual, structural, literal, and other information in MMKGs to achieve complementary information integration. However, discrepancies in modal information within MMKGs lead to cross-modal misalignment (Zheng et al., 2023), across different modalities may limit the ef-

064 fectiveness of cross-modal information utilization
065 achieved through feature fusion (Lin et al., 2022;
066 Zhu et al., 2023; Xu et al., 2023; Chen et al., 2023).

067 Multimodal large language models (MLLMs)
068 have emerged as front-runners in comprehending
069 visual information and integrating diverse modalities,
070 especially in natural language generation and
071 visual reasoning (Huang et al., 2023). These models
072 excel in deciphering the deep semantic information
073 beyond the visual information, bringing a significant
074 advantage to multimodal knowledge reasoning tasks
075 (Huang et al., 2023). Crucially, the extensive
076 background knowledge and advanced reasoning ability
077 of MLLM open avenues for enriching the entity
078 information and bridging the semantic gap of various
079 crossmodal attributes, showcasing their potential in
080 adequately understanding and utilizing the breadth
081 of multimodal data.

082 In this paper, we propose MM-ChatAlign. Different
083 from the representation learning-based MMEA
084 paradigm of previous methods, this novel framework
085 is designed to maximize the potential of MLLMs
086 oriented to the MMEA task. MM-ChatAlign
087 utilizes the entity set derived from embedding-based
088 methods as candidates, enhancing alignment accuracy
089 through the visual reasoning capabilities of MLLMs.
090 The framework initially implements the MMKG-Code
091 translation module to effectively represent MMKG
092 in a code format (Yang et al., 2024) that is highly
093 compatible with MLLMs, thus facilitating a better
094 understanding of multi-modal information in the
095 MMKG. Furthermore, MM-ChatAlign capitalizes on
096 the background knowledge and visual reasoning
097 abilities of MLLMs by generating comprehensive
098 descriptions for entities based on their images,
099 names, and relational data, and reasons for alignment.
100 The rethinking phase evaluates the probabilities of
101 entity pair alignment, revisits results, and potentially
102 expands the search scope via an iterative candidate
103 collection process to ensure precise alignments. Extensive
104 experiment results over four representative MMEA
105 datasets demonstrate the effectiveness of MM-ChatAlign
106 and also highlight the feasibility of using MLLMs
107 for the MMEA task.

108 In general, our main contributions are as follows:

109 (1) We introduce a new paradigm in MMEA by
110 combining MLLMs with traditional embedding-based
111 methods to leverage advanced multimodal reasoning
112 and the extensive knowledge of MLLMs.

113 (2) We design MM-ChatAlign, a framework that
114 integrates MLLMs with KRL-based methods for
115

enhancing the efficiency and accuracy of MMEA. 116

(3) We conduct experiments on four representative 117
118 MMEA datasets to validate the effectiveness of MM-ChatAlign
119 and demonstrate the significant potential of MLLMs in MMEA. 120

2 Methodology 121

In this paper, we propose the MM-ChatAlign, a versatile 122
123 plug-and-play MMEA framework that capitalizes on the advanced
124 reasoning abilities and background knowledge of MLLMs, while
125 optimizing both efficiency and accuracy. The overall
126 architecture of MM-ChatAlign is shown in Figure 2. The
127 framework integrates an embedding-based candidate collection
128 module, configurable across various KRL methods, designed to
129 exclude non-relevant candidates. Moreover, it features a
130 reasoning and rethinking module, powered by MLLMs, that
131 enhances alignment by effectively leveraging multimodal
132 information. 133 134

2.1 Task Formulation 135

Formally, the MMEA task refers to the process of 136
137 identifying correspondences between entities across two
138 different MMKGs, denoted as $\mathcal{G}_1 = (\mathcal{E}_1, \mathcal{V}_1, \mathcal{R}_1, \mathcal{T}_1)$
139 and $\mathcal{G}_2 = (\mathcal{E}_2, \mathcal{V}_2, \mathcal{R}_2, \mathcal{T}_2)$. The primary challenge
140 in MMEA is to discover and establish links between pairs
141 of entities (e_1, e_2) where $e_1 \in \mathcal{E}_1$ and $e_2 \in \mathcal{E}_2$,
142 which are deemed to be equivalent in the real-world
143 context. This task is intricate due to the necessity of
144 integrating multimodal data, especially the visual data
145 contained in \mathcal{V} , to align entities between the MMKGs. 146

2.2 Embedding-based Candidate Collecting 147

To harness the strengths of embedding-based methods 148
149 while incorporating the advanced capabilities of MLLMs,
150 MM-ChatAlign leverages its plug-and-play capability to
151 integrate seamlessly with existing embedding-based
152 MMEA methods, such as Simple-HHEA (Jiang et al., 2023)
153 and XGEA (Xu et al., 2023). The framework is further
154 enhanced by cross-modal matching techniques (Radford
155 et al., 2021). Subsequently, it either directly generates
156 results or efficiently accumulates candidate entities.
157 This optimization is achieved through an iterative
158 process of candidate collection. 159

2.2.1 KRL-based Entity Embedding 160

This stage initializes entity embeddings as a combination 161
162 of the name, image, temporal, and structural features of
163 the entity. Specifically, it utilizes

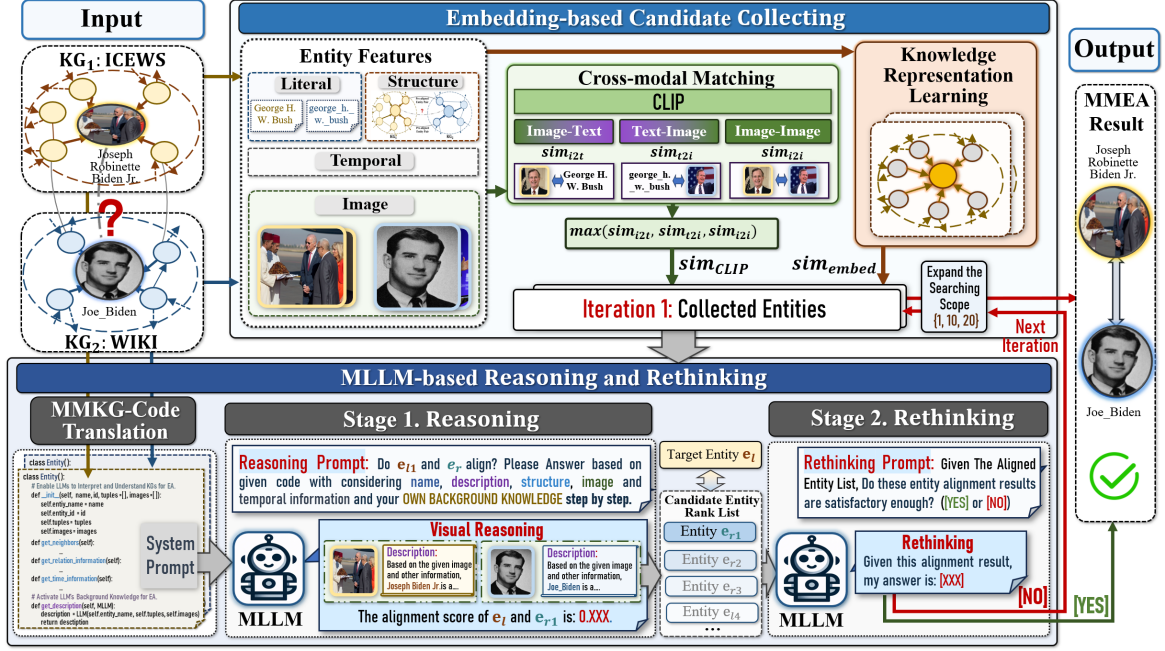


Figure 2: The comprehensive framework of MM-ChatAlign, which is designed to efficiently harness the advanced visual reasoning capabilities and intrinsic background knowledge of MLLMs.

BERT (Devlin et al., 2018) with a feature whitening transformation (Su et al., 2021) to obtain the entity name embedding $\{h_n^{name}\}_{n=1}^N$. The image features, denoted as $\{h_n^{img}\}_{n=1}^N$, of entities are derived from the CLIP model (Radford et al., 2021). The framework encapsulates temporal characteristics with Time2Vec (Goel et al., 2020), which converts time into a learnable embedding $\{h_n^{time}\}_{n=1}^N$.

The structural feature is integrated through a biased random walk (Wang et al., 2023) for precise one-hop and multi-hop relational modeling.

Furthermore, considering the plug-and-play feature of MM-ChatAlign, we have developed a variant integrated with XGEA (Xu et al., 2023), which adopts the cross-modal graph attention mechanism with graph neural network to get the structural embedding h^{struc} of the entity. The culmination of these processes results in final embeddings that merge name, temporal, and structural features into a unified representation for entities, expressed as:

$$\{h_n^{mul}\}_{n=1}^N = \{[h_n^{name} \otimes h_n^{time} \otimes h_n^{struc}]\}_{n=1}^N, \quad (1)$$

where \otimes denoted the concatenation operation. A detailed description of the KRL-based entity embedding can be found in Appendix A.3.

The entity embedding is trained by employing margin ranking loss and cross-domain similarity local scaling (CSLS) (Conneau et al., 2017) for similarity measurement.

2.2.2 Cross-modal Matching

During the candidate entity collection phase, we employed the cross-modal retrieval model CLIP (Radford et al., 2021) to expedite the comparison of cross-modal attributes between entities, taking into account efficiency considerations. Given two entities, the cross-modal similarity sim_{CLIP} is calculated by the maximum of image-to-image, image-to-text, and text-to-image similarities sim_{i2i} , sim_{i2t} , sim_{t2i} :

$$sim_{CLIP} = \max(sim_{i2i}, sim_{i2t}, sim_{t2i}) \quad (2)$$

In the context of MMKG, the max aggregation mechanism facilitates effective cross-modal information comparison even in instances where images are compromised by noise or absent entirely.

Then, the entity similarity of the given entity pairs is computed as follows:

$$sim = (1 - \alpha) \cdot sim_{embed} + \alpha \cdot sim_{CLIP}, \quad (3)$$

where sim_{embed} represents the similarity based on the entity embeddings, α is the hyper-parameter to balance the importance between the sim_{embed} and sim_{CLIP} , and these combined similarity measures are used for ranking the candidate entities.

2.3 MLLM-based Reasoning and Rethinking

To efficiently utilize the vast background knowledge and visual reasoning abilities of MLLMs, we

have integrated a multimodal reasoning module based on MLLM within the MM-ChatAlign framework, as depicted in the lower section of Figure 2.

Given the target entity e_l , the framework first gathers the potential entities as candidates by leveraging the similarity metric $sim(e_l, \{h^{r_n}\}_{n=1}^N)$. the MLLM is utilized for subsequent inference if the discrepancy in normalized similarity scores between the top two ranked candidates denoted as $sim_{embed}(e_l, e_{r_1}) - sim_{embed}(e_l, e_{r_2})$, falls below a predetermined threshold β . This approach ensures that MLLM is used only when necessary, enhancing both efficiency and accuracy to produce entity candidate list $cand = \{e_{r_1}, e_{r_2}, \dots\}$.

After selecting candidates, MM-ChatAlign conducts *MMKG-Code translation* and two-stage *Reasoning & Rethinking*. Based on prompt engineering, the MLLM estimates the alignment probability of entity pairs and decides whether to continue searching for additional candidates. The detailed pseudo-code is illustrated in Algorithm 1.

Algorithm 1 MLLM-based Reasoning and Rethinking

Input: The KG pair to be aligned $\{\mathcal{KG}_1, \mathcal{KG}_2\}$
Output: Aligned entity pairs C

- 1: //Embedding-based Candidate Collecting
- 2: $sim_{embed} \leftarrow$ KRL-BASED ENTITY EMBEDDING($\mathcal{KG}_1, \mathcal{KG}_2$)
- 3: $sim_{CLIP} \leftarrow$ CROSS MODAL MATCHING($\mathcal{KG}_1, \mathcal{KG}_2, CLIP$)
- 4: $sim \leftarrow (1 - \alpha) \cdot sim_{embed} + \alpha \cdot sim_{CLIP}$
- 5: **if** $sim(e_l, e_{r_1}) > \beta$ **then** Aligned entity pairs $C \leftarrow (e_l, e_{r_1})$
- 6: **else** //MLLM-based Reasoning and Rethinking
- 7: **for** scope $\leftarrow \{1, 10, 20\}$ **do**
- 8: cand \leftarrow COLLECT CANDIDATES($sim, scope$)
- 9: align pair \leftarrow REASONING(cand, $\mathcal{KG}_1, \mathcal{KG}_2$)
- 10: isSatisfied \leftarrow RETHINKING(align pair)
- 11: **if** isSatisfied **then**
- 12: Aligned entity pairs $C \leftarrow$ align pair
- 13: **break**
- 14: **end if**
- 15: **end for**
- 16: **end if**
- 17: **return** Aligned entity pairs C

2.3.1 MMKG-Code translation

To represent MMKGs in a format that MLLMs can comprehend the visual information and other modalities. The MMKG-Code translation module of MM-ChatAlign plays a pivotal role, which has validated the effectiveness of MLLM for understanding the MMKG (Yang et al., 2024) and improving the compatibility of MMKG with MLLMs. This module operates by encoding various modalities of the MMKG, such as entities, relations, and

visual attributes into a structured code representation through the system prompt.

As shown in the MMKG-Code Translation part in Figure 2, The `__init__()` function enables MLLMs to process entity name, id, visual, and tuple information as input. Given an entity, the `get_neighbors()`, `get_relations()`, and `get_temporal()` functions enable MLLMs to understand neighborhoods, relations, and temporal information about entities in MMKGs.

2.3.2 Stage 1: Reasoning

The reasoning phase is designed to harness the comprehensive background knowledge and visual reasoning capabilities of MLLMs. As shown in Figure 2, different from the cross-modal matching, we first use MLLM to generate entity descriptions by meticulously extracting the pivotal semantic features in the image. Then we use the carefully designed prompt template to generate textual descriptions of entities based on the given images, entity names and tuples with the help of the rich knowledge from MLLM.

Subsequently, the MLLM conducts an in-context learning procedure to compute alignment probabilities between the target entity and its candidates. It comprehensively considers a diverse set of features for each entity pair at each step, including names, images, temporal and structural information, and generated descriptions by the MLLM. During this reasoning phase, the MLLM assesses the alignment scores for each candidate entity and re-ranks them according to their probability of correct alignment, thereby optimizing the candidate order to achieve more accurate alignments.

2.3.3 Stage 2: Rethinking

The MLLM scores the current MMEA results from the similarities of entity pairs within 4 dimensions: name, description, structure, and image, and critically evaluates all dimensions together. If the results are not unsatisfactory, it will restart the candidate collection stage to re-evaluate the alignments with more candidate entities (e.g., 1, 10, 20). This iterative refinement of the candidate list can ensure that all possible alignments are evaluated and improve the efficiency of reasoning.

3 Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of MM-ChatAlign in

MMEA tasks. Our investigation focuses on two key research questions:

- **RQ1: How does MM-ChatAlign perform in MMEA, and what is the impact of each component?** We aim to assess the overall performance of the framework and the contributions of each module to understand their utility.
- **RQ2: Does MM-ChatAlign balance accuracy with efficiency in MMEA?** This question explores the capability of MM-ChatAlign to deliver high accuracy while maintaining efficiency.

Considering the above research questions, we aim to provide comprehensive evaluations of MM-ChatAlign, highlighting its abilities for MMEA.

3.1 Datasets

We first conduct experiments on two benchmark MMEA datasets, DBP15K(EN-FR) and FB-YAGO (Sun et al., 2020; Liu et al., 2021). These datasets have been widely used in previous MMEA works, Furthermore, we extend the ICEWS-WIKI and ICEWS-YAGO datasets (Jiang et al., 2023) to multi-modal versions. These versions are specifically crafted to address the more demanding challenges of practical MMEA. They showcase significant heterogeneity between MMKG pairs, which is evident in the variance in their structural and other modality features. For the dataset construction, we use entity images from Google Image Search for ICEWS. For WIKI and YAGO, we retrieve top-3 relevant images from their Wikipedia pages. All images are manually verified to ensure quality and relevance. The detailed statistics of datasets are summarized in Appendix A.2.

3.2 Baselines

For a fair and comprehensive evaluation, we select 12 state-of-the-art methods and categorize these into three groups: *Single*, *Visual*, and *Literal*.

Single methods only utilize the structural information within MMKGs, including MTransE (Chen et al., 2017), BootEA (Sun et al., 2018), GCN-Align (Wang et al., 2018), and Dual-AMN (Mao et al., 2021). *Visual* methods enhance entity representations with images of entities, including EVA (Liu et al., 2021), MCLEA (Lin et al., 2022), MEAformer (Chen et al., 2023), XGEA (Xu et al., 2023), and MMIEA (Zhu et al., 2023). *Literal* methods introduce entity names as supplementary features, including RDGCN (Chen et al.,

2022), Dual-AMN (Mao et al., 2021), TEA (Zhao et al., 2023), BERT-INT (Devlin et al., 2018), MEAformer (Chen et al., 2023), XGEA (Xu et al., 2023), and MMIEA (Zhu et al., 2023).

To compare the impacts of different embedding-based methods, we have established two versions of MM-ChatAlign: MM-ChatAlign* refers to the version that incorporates XGEA as its base, while MM-ChatAlign uses Simple-HHEA as its base. To ensure fair comparisons and to accommodate the diverse modalities utilized by various methods, we make specific adaptations in our approach. In the visual track, MM-ChatAlign leverages structural and visual information. Meanwhile, in the Literal track, MM-ChatAlign additionally incorporates entity name information.

3.3 Experiment Settings

In our experiment setup, we utilized GPT-4V (Yang et al., 2023) for visual reasoning and LLAMA2-70b-Chat (Touvron et al., 2023) for entity alignment during the MLLM selection stage. Ablation studies were conducted using various LLMs as shown in Table 3. Data was split in a 3:7 ratio for training and testing. For image and name embeddings, we employed CLIP (Radford et al., 2021) and BERT (Su et al., 2021), respectively. The evaluation metrics used were Hits@k (for k = 1, 10) and Mean Reciprocal Rank (MRR). Detailed configurations are available in Appendix A.4.

3.4 Main Experiment Results

In response to **RQ1**, in our main experiments, as detailed in Table 1, we evaluate the performance of MM-ChatAlign from two versions: visual and literal, across the four datasets.

In the visual category, which only allows methods to leverage the structural and visual features, MM-ChatAlign demonstrates superior performance compared to other leading methods. For instance, on the DBP15K(EN-FR) and FB-YAGO datasets, MM-ChatAlign achieves a remarkable Hits@1 score of 0.940 and 0.680, which is a notable improvement over XGEA, the runner-up method with a Hits@1 score of 0.889 and 0.616. This represents 5.7% and 6.4% increases in performance. Similarly, in two challenging datasets(ICEWS-WIKI and ICEWS-YAGO), MM-ChatAlign scored a Hits@1 of 0.430 and 0.415, significantly surpassing the next-highest score of 0.263 and 0.302 on Hits@1, marking a substantial 16.7% and 11.3% improvement. This trend of out-

	Models	DBP15K(EN-FR)			FB-YAGO15K			ICEWS-WIKI			ICEWS-YAGO		
		Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
Single	MTransE	0.247	0.577	0.360	0.003	0.017	0.011	0.021	0.158	0.068	0.012	0.084	0.040
	BootEA	0.653	0.874	0.731	0.323	0.579	0.410	0.072	0.275	0.139	0.020	0.120	0.056
	GCN-Align	0.411	0.772	0.530	0.023	0.107	0.053	0.046	0.184	0.093	0.017	0.085	0.038
	Dual-AMN	0.840	0.965	0.888	0.403	0.662	0.499	0.077	0.285	0.143	0.032	0.147	0.069
Visual	EVA	0.793	0.942	0.847	0.171	0.417	0.260	0.081	0.203	0.119	0.019	0.075	0.038
	MCLEA	0.834	0.975	0.885	0.388	0.641	0.474	0.253	0.494	0.336	0.231	0.412	0.295
	XGEA	<u>0.889</u>	<u>0.981</u>	<u>0.924</u>	<u>0.616</u>	<u>0.794</u>	<u>0.679</u>	0.170	0.277	0.207	0.142	0.250	0.180
	MMIEA	0.830	0.962	0.870	0.536	0.712	0.599	<u>0.263</u>	<u>0.523</u>	<u>0.350</u>	<u>0.302</u>	<u>0.573</u>	<u>0.396</u>
	MEAformer	0.845	0.976	0.894	0.444	0.692	0.529	0.246	0.470	0.321	0.192	0.352	0.247
	MM-ChatAlign*	0.940	0.999	0.952	0.680	0.915	0.910	0.430	0.930	0.548	0.415	0.630	0.479
Literal	RDGCN	0.873	0.950	0.901	0.466	0.708	0.549	0.064	0.202	0.096	0.029	0.097	0.042
	Dual-AMN	0.954	0.994	0.970	0.540	0.711	0.607	0.083	0.281	0.145	0.031	0.144	0.068
	TEA	0.987	0.996	0.990	0.612	0.770	0.730	0.610	0.894	0.718	0.657	0.891	0.740
	BERT-INT	0.990	0.997	0.993	0.678	0.797	0.780	0.561	0.700	0.607	0.756	0.859	0.793
	XGEA	0.991	1.000	0.996	<u>0.835</u>	<u>0.915</u>	<u>0.869</u>	0.549	0.628	0.575	0.314	0.421	0.351
	MMIEA	0.992	0.997	0.994	0.793	0.830	0.809	0.562	0.716	0.616	0.745	0.857	0.787
	MEAformer	0.996	1.000	0.998	0.748	0.887	0.798	0.644	0.842	0.713	0.698	0.878	0.762
	Simple-HHEA	0.959	0.995	0.972	0.735	0.835	0.776	<u>0.720</u>	<u>0.872</u>	<u>0.754</u>	<u>0.847</u>	<u>0.915</u>	<u>0.870</u>
	MM-ChatAlign*	<u>0.995</u>	1.000	<u>0.996</u>	0.880	0.915	0.896	0.650	0.700	0.669	0.535	0.570	0.554
	MM-ChatAlign	0.965	1.000	0.977	0.795	0.845	0.819	0.945	0.966	0.948	0.930	0.965	0.943

Table 1: Main experiment results on the four datasets. *Bold*: the best result; *Underline*: the runner-up result.

performance by MM-ChatAlign demonstrates its robust capability in integrating visual reasoning.

In the literal category, which extra allows methods to leverage the entity name feature, MM-ChatAlign also excels other methods. On DBP15K(EN-FR) and FB-YAGO15K, MM-ChatAlign achieves a remarkable Hits@1 score of 0.990 and 0.880, which is competitive with the best baseline method. In the ICEWS-WIKI and ICEWS-YAGO datasets, MM-ChatAlign achieves a remarkable Hits@1 score of 0.945 and 0.920, significantly outperforming the score of the best baseline method (0.720 and 0.847) with 22.5% and 8.6%. This superior performance indicates proficiency of MM-ChatAlign in leveraging both visual and name information.

Notably, both MM-ChatAlign and MM-ChatAlign* have demonstrated enhancements over their base models. These improvements are particularly pronounced in both visual and literal tracks, emphasizing the effectiveness of MM-ChatAlign. The enhancements observed in the ICEWS-WIKI and ICEWS-YAGO datasets underscore MM-ChatAlign’s versatility in handling complex scenarios across diverse settings. The notable performance gains in these modalities affirm the successful integration of MLLMs, effectively bridging the gap between different types of modality representations.

Settings	ICEWS-WIKI		ICEWS-YAGO	
	Hits@1	MRR	Hits@1	MRR
MM-ChatAlign	0.945	0.948	0.930	0.943
- w/o mllm reasoning	0.735	0.789	0.840	0.872
- w/o name	0.430	0.548	0.415	0.479
- w/o image	0.915	0.924	0.905	0.938
- w/o structure	0.925	0.942	0.885	0.902
- w/o temporal	0.875	0.896	0.895	0.911
- w/o code	0.885	0.897	0.845	0.890
- w/o description	0.870	0.881	0.810	0.881
- w/o clip	0.920	0.929	0.910	0.930

Table 2: Ablation study of MM-ChatAlign.

3.5 Ablation Study

To assess the contribution of each component in MM-ChatAlign, we conduct ablation studies on the ICEWS-WIKI and ICEWS-YAGO datasets. These studies aim to determine the individual benefits of components in MM-ChatAlign and investigate their influence on the base MLLM’s performance. The results are presented in Table 2.

3.5.1 Effectiveness of Each Component

To evaluate the impact of MLLMs, MM-ChatAlign (w/o mllm reasoning) erases the MLLM component, depending exclusively on entity embeddings and cross-modal matching for MMEA. Compared to this, MM-ChatAlign demonstrates substantial performance improvements (with increases of 19% and 9% in Hits@1), which underscores the crucial contribution of MLLMs in the MMEA task.

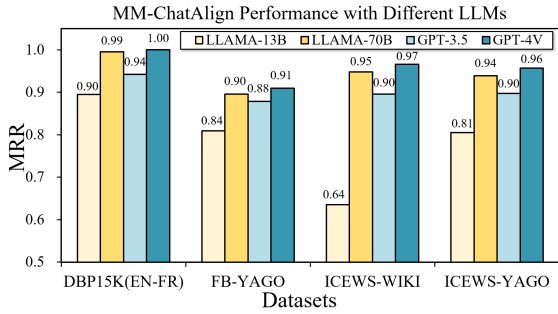


Figure 3: Performance comparison over different LLMs.

To determine the efficacy of MM-ChatAlign in utilizing name, image, and structure information, four variations are tested: MM-ChatAlign (*w/o* name, *w/o* image, *w/o* structure, and *w/o* temporal), excluding the corresponding features, respectively.

In MM-ChatAlign (*w/o* code), the MMKG-Code translation module is substituted with the direct input of entity names and tuples to the LLM. This change leads to a notable reduction in performance, thereby affirming the effective role of the MMKG-Code translation module in aiding the LLM to comprehend MMKGs effectively.

MM-ChatAlign (*w/o* description) excludes entity descriptions and also shows a performance decline. This result indicates that generating entity descriptions using the MLLM’s visual reasoning ability and background knowledge effectively harnesses visual and contextual information about entities.

In MM-ChatAlign (*w/o* clip), the CLIP is substituted with the direct entity embedding for candidate selection, the drop in the performance compared with the original version demonstrated the contribution of cross-modal matching.

In summary, this ablation study demonstrates how MM-ChatAlign capitalizes on MLLMs for MMEA. Additionally, it underscores the importance of leveraging the MLLM’s extensive background knowledge for effective MMEA.

3.5.2 Influence Over Different LLMs

Considering the versatile compatibility of MM-ChatAlign with diverse LLMs during the reasoning and rethinking phase, this study focuses on evaluating the effect of various LLMs on their performance as depicted in Figure 3. The results show that MM-ChatAlign integrated with GPT-4V achieves the best performance, which can be attributed to the advanced capability of GPT-4V in boosting the framework’s effectiveness. Additionally, MM-ChatAligns with LLAMA2 at varying scales (13b, 70b) disclose a direct relationship be-

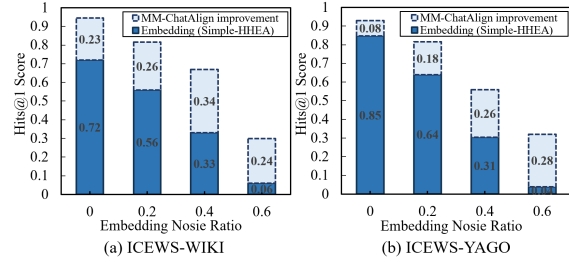


Figure 4: Performance improvement of MM-ChatAlign compared with the embedding-based method.

tween the model size and effectiveness of MM-ChatAlign. The performance of LLAMA2-13b exhibits a considerable reduction, suggesting a reevaluation of the constraints posed by smaller-scale models in MMEA. This decline is evident in shortcomings related to reasoning and output formatting.

3.5.3 Influence Over Embedding Methods

Initially, through the analysis presented earlier, the main experiments examining the integration of MM-ChatAlign with XGEA and Simple-HHEA demonstrated significant performance improvements. Additionally, the ablation studies, which involved removing the MLLM reasoning component, have broadly confirmed that incorporating MLLM significantly enhances performance across different embedding-based MMEA methods.

To investigate whether introducing MLLMs into MM-ChatAlign can enhance the performance over traditional MMEA methods with varying qualities of entity embeddings, we designed experiments involving embedding noise. In these experiments, random noise is injected into the dimensions of entity embeddings learned by MMEA methods (i.e., Simple-HHEA), at ratios varying from 0% to 60%. We chose Simple-HHEA as the base to observe the performance improvement brought by integrating MM-ChatAlign under various embedding conditions. As shown in Figure 4, as the noise ratio increases, the performance of Simple-HHEA declines sharply, but the proportion of performance improvement brought by introducing MM-ChatAlign expands, validating the effectiveness and adaptability of MM-ChatAlign.

3.6 Case Study

In assessing the capabilities of our MM-ChatAlign, we explore a case study from our experimental evaluations. As illustrated in Figure 5, traditional MMEA methods, which primarily depend on entity embeddings, often lead to inaccuracies due to

the misalignment of similar images, structures, and other modal information of entities. This is evident in their erroneous alignment of *Joseph Robinette Biden Jr.* with wrong case *Hunter Biden*. In contrast, MM-ChatAlign, as depicted in the reasoning process in Figure 5, initially utilizes MLLM to generate integrated entity descriptions, incorporating images and other multimodal data. This approach effectively addresses the information loss typically associated with compressing images into embeddings and capitalizes on the contextual knowledge in MLLMs. Subsequently, MM-ChatAlign executes a step-by-step reasoning process, synthesizing information across various dimensions related to the entity. It not only results in the correct alignment of *Joseph Robinette Biden Jr.* with *Joe Biden*, but also enhances explainability. The case exemplifies how MM-ChatAlign effectively leverages the MLLMs to achieve accurate and reliable MMEA.








MMEA Result	Image	Multimodal Reasoning of MM-ChatAlign
[Main Entity]: Joseph Robinette Biden Jr.	[Main Entity]: 	Step 1: [NAME SIMILARITY]
[Ground Truth]: Joe Biden	[Candidates]:  	Step 2: [PROBABILITY OF DESCRIPTION POINTING SAME ENTITY]
[Embedding based Result]: Hunter Biden	 	Descriptions: Based on the given image and other information, Entity 2 is... Description: Based on the given image and other information, Entity 1 is...
[MM-ChatAlign Result]: Joe Biden	 	Considering the roles and public duties indicated by images and their contexts - Entity 1 and Entity 2 may refer to the same person at different points in time , aligns with known facts about Joe Biden, ... [PROBABILITY OF DESCRIPTION POINTING SAME ENTITY] = 4 out of 5. Step 3: [STRUCTURE SIMILARITY] FINAL SCORE : 0.772

Figure 5: The case study of MM-ChatAlign, which overcomes the inaccuracies of traditional MMEA methods by integrating multimodal data with MLLMs.

3.7 Efficiency Analysis

In response to **RQ2**, we discuss how MM-ChatAlign optimizes efficiency while maintaining MMEA accuracy. To optimize efficiency while maintaining accuracy, MM-ChatAlign implements a three-round iterative candidate collecting, intricately tailored to adapt to the complexities of different datasets. As illustrated in Figure 6, with simpler datasets where neural methods perform better (i.e., DBP15K(EN-FR) and FB-YAGO), MM-ChatAlign tends to converge faster, leading to better utilization of resources and higher efficiency. Conversely, for more challenging datasets like ICEWS-WIKI/YAGO, the framework inclines towards collecting more candidates and conducting thorough analyses across additional iterations. This adaptive methodology guarantees the maintenance of accuracy while optimizing resource utilization. Additionally, the comparison between the original

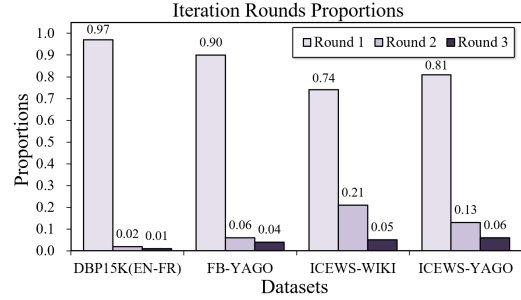


Figure 6: Proportions of iteration rounds of MM-ChatAlign’s two-stage reasoning on different datasets.

Settings	ICEWS-WIKI		ICEWS-YAGO	
	avg.tokens	avg.time	avg.tokens	avg.time
- w/ llama2-70b	13,162	84.5	7,276	50.6
- w/ llama2-13b	41,178	112.3	23,118	77.7
- w/ gpt-3.5	15,101	11.9	15,124	13.0
- w/ gpt-4	9,275	69.8	8,644	65.3
- w/o two-stage	62,825	403.3	54,403	378.1

Table 3: Efficiency analysis of MM-ChatAlign. *avg.tokens* and *avg.time* respectively denote the average tokens and time (seconds) cost per sample.

MM-ChatAlign and the *w/o* two-stage variant in Table 3 also demonstrates the superiority of the two-stage strategy in conserving over 80% computing resources and time consumption. Furthermore, MM-ChatAlign is adaptive to different MLLMs, which positions it to benefit from the ongoing evolution of MLLMs. Given the efficiency challenges associated with current MLLMs, from the perspective of application scenarios, MM-ChatAlign is now suitable for settings where accuracy in MMEA is crucial, often prioritizing result reliability over speed. However, as the efficiency of LLM improves, MM-ChatAlign’s efficiency and accuracy are expected to enhance correspondingly, as evidenced in Figure 3 and 6.

4 Conclusion

This study introduces MM-ChatAlign, an innovative framework for MMEA that leverages the advanced capabilities of MLLMs. By incorporating a code-type transformation module for MMKGs and a two-stage multimodal reasoning process, the method realizes the efficient and effective MMEA. Our experimental results not only validate MM-ChatAlign’s superior performance on newly developed MMEA datasets and classical datasets but also highlight the tremendous potential of MLLMs in challenging MMEA tasks. Future work will continue to focus on optimizing efficiency, further unleashing the potential of MLLM in MMEA.

5 Limitations

Despite MM-ChatAlign’s high accuracy in MMEA through innovative architecture and LLM integration, its application may be limited by substantial resource consumption and slow LLM inference speeds, posing challenges in time-sensitive or resource-limited environments. While enhancements in the methodology have improved the balance between precision and efficiency, further advancements such as model distillation are necessary. Additionally, the system’s reduced effectiveness with smaller-scale models highlights the need for future iterations to explore techniques like sparse fine-tuning (SFT), enabling efficient performance without reliance on large model sizes.

6 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, YAGO, ICEWS, and DBpedia. Therefore, we believe that our study complies with the ACL Ethics Policy.

References

- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Adv Neural Inf Process Syst*, 26.
- Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization.
- Zhibin Chen, Yuting Wu, Yansong Feng, and Dongyan Zhao. 2022. Integrating manifold knowledge for global entity linking with heterogeneous graphs. *Data Intelligence*, 4(1):20–40.
- Zhuo Chen, Jiaoyan Chen, Wen Zhang, Lingbing Guo, Yin Fang, Yufeng Huang, Yichi Zhang, Yuxia Geng, Jeff Z Pan, Wenting Song, et al. 2023. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3317–3327.
- Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Rishab Goel, Seyed Mehran Kazemi, Marcus Brubaker, and Pascal Poupart. 2020. Diachronic embedding for temporal knowledge graph completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3988–3995.
- Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, et al. 2023. Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science*, 15(1):29.
- Xuhui Jiang, Chengjin Xu, Yinghan Shen, Fenglong Su, Yuanzhuo Wang, Fei Sun, Zixuan Li, and Huawei Shen. 2023. Rethinking gnn-based entity alignment on heterogeneous knowledge graphs: New datasets and a new method. *arXiv preprint arXiv:2304.03468*.
- Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multi-modal contrastive representation learning for entity alignment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2572–2584.
- Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual pivoting for (unsupervised) entity alignment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4257–4266.
- Zhiyuan Liu, Yixin Cao, Liangming Pan, Juanzi Li, and Tat-Seng Chua. 2020. Exploring and evaluating attributes, values, and structures for entity alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6355–6364.
- Xin Mao, Wenting Wang, Yuanbin Wu, and Man Lan. 2021. Boosting the speed of entity alignment 10 ×: Dual attention matching network with normalized hard sample mining. In *Proceedings of the Web Conference 2021*, pages 821–832. ACM.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

689	Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu.	relation extraction from a translation point of view.	745
690	2018. Bootstrapping entity alignment with knowl-	In <i>Proceedings of the 61st Annual Meeting of the</i>	746
691	edge graph embedding . In <i>Proceedings of the Twenty-</i>	<i>Association for Computational Linguistics (Volume</i>	747
692	<i>Seventh International Joint Conference on Artificial</i>	<i>1: Long Papers)</i> , pages 6810–6824.	748
693	<i>Intelligence</i> , volume 18, pages 4396–4402. Interna-		
694	tional Joint Conferences on Artificial Intelligence	Ziyue Zhong, Meihui Zhang, Ju Fan, and Chenxiao	749
695	Organization.	Dou. 2022. Semantics driven embedding learning for	750
696	Zequn Sun, Qingheng Zhang, Wei Hu, Chengming	effective entity alignment. In <i>2022 IEEE 38th Inter-</i>	751
697	Wang, Muhao Chen, Farahnaz Akrami, and Chengkai	<i>national Conference on Data Engineering (ICDE)</i> ,	752
698	Li. 2020. A benchmarking study of embedding-based	pages 2127–2140. IEEE.	753
699	entity alignment for knowledge graphs . <i>Proceedings</i>		
700	<i>of the VLDB Endowment</i> , 13(12):2326–2340.	Bin Zhu, Meng Wu, Yunpeng Hong, Yi Chen, Bo Xie,	754
701	Xiaobin Tang, Jing Zhang, Bo Chen, Yang Yang,	Fei Liu, Chenyang Bu, and Weiping Ding. 2023.	755
702	Hong Chen, and Cuiping Li. 2020. BERT}-{INT:	Mmiea: Multi-modal interaction entity alignment	756
703	A {BERT}-based interaction model for knowledge	model for knowledge graphs. <i>Information Fusion</i> ,	757
704	graph alignment. <i>interactions</i> , 100:e1.	100:101935.	758
705	Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-	Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang,	759
706	bert, Amjad Almahairi, Yasmine Babaei, Nikolay	Penglei Sun, Xuwu Wang, Yanghua Xiao, and	760
707	Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti	Nicholas Jing Yuan. 2022. Multi-modal knowledge	761
708	Bhosale, et al. 2023. Llama 2: Open founda-	graph construction and application: A survey. <i>IEEE</i>	762
709	tion and fine-tuned chat models. <i>arXiv preprint</i>	<i>Transactions on Knowledge and Data Engineering</i> .	763
710	<i>arXiv:2307.09288</i> .		
711	Chenxu Wang, Zhenhao Huang, Yue Wan, Junyu Wei,		
712	Junzhou Zhao, and Pinghui Wang. 2023. FuAlign:		
713	Cross-lingual entity alignment via multi-view repre-		
714	sentation learning of fused knowledge graphs . <i>In-</i>		
715	<i>form. Fusion</i> , 89:41–52.		
716	Zhichun Wang, Qingsong Lv, Xiaohan Lan, and		
717	Yu Zhang. 2018. Cross-lingual knowledge graph		
718	alignment via graph convolutional networks . In <i>Pro-</i>		
719	<i>ceedings of the 2018 Conference on Empirical Meth-</i>		
720	<i>ods in Natural Language Processing</i> , pages 349–357.		
721	Association for Computational Linguistics.		
722	Baogui Xu, Chengjin Xu, and Bing Su. 2023. Cross-		
723	modal graph attention network for entity alignment.		
724	In <i>Proceedings of the 31st ACM International Con-</i>		
725	<i>ference on Multimedia</i> , pages 3715–3723.		
726	Ke Yang, Jiateng Liu, John Wu, Chaoqi Yang, Yi R		
727	Fung, Sha Li, Zixuan Huang, Xu Cao, Xingyao		
728	Wang, Yiquan Wang, et al. 2024. If llm is the wizard,		
729	then code is the wand: A survey on how code em-		
730	powers large language models to serve as intelligent		
731	agents. <i>arXiv preprint arXiv:2401.00812</i> .		
732	Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng		
733	Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan		
734	Wang. 2023. The dawn of llms: Preliminary		
735	explorations with gpt-4v (ision). <i>arXiv preprint</i>		
736	<i>arXiv:2309.17421</i> , 9(1).		
737	Yu Zhao, Yike Wu, Xiangrui Cai, Ying Zhang, Haiwei		
738	Zhang, and Xiaojie Yuan. 2023. From alignment to		
739	entailment: A unified textual entailment framework		
740	for entity alignment. In <i>Findings of the Association</i>		
741	<i>for Computational Linguistics: ACL 2023</i> , pages		
742	8795–8806.		
743	Changmeng Zheng, Junhao Feng, Yi Cai, Xiaoyong Wei,		
744	and Qing Li. 2023. Rethinking multimodal entity and		

A Appendix

A.1 Related Works of Entity Alignment

Entity alignment (EA) has historically seen diverse methodologies. Translation-based methods such as MTransE (Chen et al., 2017), and BootEA (Sun et al., 2018), based on the TransE framework (Bordes et al., 2013), excel in knowledge representation by embedding entities and relations. Graph Neural Network (GNN) (Kipf and Welling, 2016) based methods, such as GCN-Align (Wang et al., 2018), RDGCN (Chen et al., 2022), and Dual-AMN (Mao et al., 2021), leverage neighborhood aggregation mechanism for modeling structural information. Other methods, such as BERT-INT (Tang et al., 2020), TEA (Zhao et al., 2023), AttrGNN (Liu et al., 2020), SDEA (Zhong et al., 2022), address KG heterogeneity using multi-view information.

Multimodal entity alignment (MMEA) extends EA to multimodal domains, which is more challenging due to the request of modeling visual information of entities. Current MMEA methods primarily employ representation learning models to calculate similarities between entity embeddings based on multiple modalities, EVA (Liu et al., 2021) leverages visual similarities for pre-alignment and introduces a multimodal fusion module. MCLEA (Lin et al., 2022) and XGEA (Xu et al., 2023) capture cross-modal relationships for measuring entity similarity. MMIEA (Zhu et al., 2023) extends the BERT-INT to suit the MMEA task. MEAformer (Chen et al., 2023) proposes an attention mechanism that dynamically fosters modality preferences adaptable to entities.

Despite their widespread adoption, these methods heavily rely on the quality of input MMKG data and entity embeddings derived from knowledge representation learning (KRL). This phenomenon becomes a bottleneck, particularly evident in handling more challenging but practical EA scenarios (Jiang et al., 2023). Besides, MMKG-derived modal representations limit their ability to leverage visual comprehension and broader background knowledge, underscoring the need for more advanced MMEA methods. Consequently, there is a growing interest in exploring new paradigms for MMKG tasks, with MLLMs emerging as a promising supplement. Leveraging extensive parametric knowledge and visual reasoning abilities, MLLMs offer potential solutions to overcome the limitations of previous methods, processing MMKGs without solely relying on KRL.

A.2 Detailed Statistics of the MMEA datasets

The detailed statistics of the four MMEA datasets in our experiments are shown in Table 4

A.3 Detailed KRL-based Entity Embedding

This stage initializes entity embeddings as a combination of the name, image, temporal, and structural features of the entity. Specifically, it utilizes BERT (Devlin et al., 2018) with a feature whitening transformation (Su et al., 2021) to obtain the entity name embedding $\{h_n^{name}\}_{n=1}^N$. The image features, denoted as $\{h_n^{img}\}_{n=1}^N$, of entities are derived from the CLIP model (Radford et al., 2021). The framework encapsulates temporal characteristics with Time2Vec (Goel et al., 2020), which converts time into a learnable embedding $\{h_n^{time}\}_{n=1}^N$.

Furthermore, the structural feature is integrated through a biased random walk (Wang et al., 2023) for precise one-hop and multi-hop relational modeling. Let e_j represent the node selected at the j -th step of random walks, and define $(e_1, r_1, e_2, \dots, e_{j-1}, r_{j-1}, e_j)$ as the path generated during this process. The selection probability of an entity is as follows:

$$P_r(e_{j+1} | e_j) = \begin{cases} \beta, & d(e_{j-1}, e_{j+1}) = 2 \\ 1 - \beta, & d(e_{j-1}, e_{j+1}) = 1 \end{cases}, e_{j+1} \in \mathcal{N}_{e_j}^-, \quad (4)$$

where $\mathcal{N}_{e_j}^-$ denotes the set of 1-hop neighbors \mathcal{N}_{e_j} of entity e_j , excluding e_{j-1} . $d(e_{j-1}, e_{j+1})$ denotes the shortest path length between e_{j-1} and e_{j+1} . Here, $\beta \in (0, 1)$ is a hyper-parameter that balances BFS and DFS search strategies (Wang et al., 2023). Then, the Skip-gram *SkipGram*(\cdot) is adopted to learn entity embeddings $\{h_n^{struc}\}_{n=1}^N$ based on the generated random walk paths.

Furthermore, considering the plug-and-play feature of our proposed framework, we have developed a variant integrated with XGEA (Xu et al., 2023), which adopts the cross-modal graph attention mechanism with graph neural network, expressed as:

$$h_{e_t}^{l+1} \leftarrow \text{AGG}_{\forall (e_s, r, e_t) \in \mathcal{Q}} (\text{ATT}(h_{e_s}^l, h_r^l, h_{e_t}^l) \cdot \text{MSG}(h_{e_s}^l, h_{e_t}^l)), \quad (5)$$

where $h_{e_t}^{l+1}$ denotes the learned entity embeddings of e_t at layer l , MSG, ATT, and AGG denote message passing, cross-modal attention, aggregation, and self-loop mechanism of XGEA (Xu et al., 2023), respectively. Finally, the output from the fi-

Dataset		#Entities	#Relations	#Facts	Density	#Anchors	Image	Temporal
DBP15K(EN-FR)	EN	15,000	193	96,318	6.421	15,000	15,000	No
	FR	15,000	166	80,112	5.341		15,000	No
FB-YAGO15K	FB	14,951	1345	592,213	39.481	11,199	13,444	No
	YAGO	15,404	32	122,886	8.192		11,194	No
ICEWS-WIKI(V)	ICEWS	11,047	272	3,527,881	319.352	5,058	33,141	Yes
	WIKI	15,896	226	198,257	12.472		47,688	Yes
ICEWS-YAGO(V)	ICEWS	26,863	272	4,192,555	156.072	18,824	80,589	Yes
	YAGO	22,734	41	107,118	4.712		68,202	Yes

Table 4: The detailed statistics of the datasets. *Temporal* denotes whether the dataset contains temporal information.

nal layer is used as the structural embedding h^{struc} of the entity.

The culmination of these processes results in final embeddings that merge name, temporal, and structural features into a unified multi-view representation for each entity, expressed as:

$$\{h_n^{mul}\}_{n=1}^N = \{[h_n^{name} \otimes h_n^{time} \otimes h_n^{struc}]\}_{n=1}^N,$$

where \otimes denoted the concatenation operation.

A.4 Detailed Experiment Settings

A.4.1 Model Configuration

For LLM selection, during the candidate collecting stage, we adopt CLIP (Radford et al., 2021) to realize cross-modal retrieval.

For MLLM selection, during the reasoning & rethinking stage, we choose GPT-4V (Yang et al., 2023) to generate descriptions for visual reasoning of entities based on the given images and MLLM’s background knowledge. Then, we adopt the open-source LLAMA2-70b-Chat (Touvron et al., 2023) for aligning entities. We also validate other representative LLMs (i.e., directly adopt GPT-4V) for MMEA in ablation studies 3. To ensure fairness in our evaluation, baseline models are configured according to their original hyper-parameter settings, except for setting hidden dimensions of the learned entity embedding to 64. Through extensive experimentation, we respectively set the hyper-parameter α and β to 0.3 and 0.2 to achieve optimal performance. Datasets are split following a 3:7 ratio for training and testing, respectively, and identical preprocessing steps were applied to all models for initial feature. The experiments are conducted with four 40GB NVIDIA A100 GPUs.

A.4.2 Initial Feature Setup

For a fair comparison, all image embeddings are obtained by CLIP (Radford et al., 2021). All MMEA models that utilize entity names share the same

name embeddings. For DBP15K(EN-FR), we obtain entity names using machine translation. For FB-YAGO, we map the IDs of Freebase and YAGO into entity names. For ICEWS-WIKI/YAGO, we use the original entity names. After that, we employ BERT (Su et al., 2021) to obtain the name embeddings. Structure-based MMEA methods that do not utilize entity name information were initialized according to their original method-specific configurations. This process involved the random initialization of embeddings.

A.4.3 Evaluation Metrics

In line with standard practices in prior MMEA research, we use two metrics for evaluation: (1) Hits@k, measuring the percentage of correct entity alignments within the top k ($k = 1, 10$) matches. (2) Mean Reciprocal Rank (MRR), reflecting the average inverse ranking of correct results. Higher values in both Hits@k and MRR indicate better MMEA performance.

A.5 Detailed prompts of MM-ChatAlign

In this section, we illustrate the prompts of MM-ChatAlign in Table, 5, 6, 7, and 8.

A.6 Details about the Case Study of MM-ChatAlign

The details of the Case Study of MM-ChatAlign, including input prompt and model output, are illustrated in Table 9 and 10.

Table 5: Prompt for getting descriptions

Prompt for getting descriptions

Given following informations: 1.[Entity] {{ Name }}; 2.[Knowledge Tuples] = {{ Tuples }}; 3.IMAGES related to [Entity]. Please answer the question:

[Question]: What is {{ Name }}? Please give a two-sentence brief introduction. The first sentence is to simply describe what is {{ Name }}, combining the identity features in IMAGES. The second sentence is to give additional description about {{ Name }} based on IMAGES, [Knowledge Tuples] and YOUR OWN KNOWLEDGE. Give [answer] strictly in format: [Entity] is

[answer]:

Table 6: Prompt for rethinking

Prompt for rethinking

Now given the following entity alignments:
[Main Entity]: {{ Name }} -> {{ Align Pairs }}

Please answer the question: Do these entity alignments are satisfactory enough ([YES] or [NO])?

Answer [YES] if they are relatively satisfactory, which means the alignment score of the top-ranked candidate meet the threshold, and is far higher than others; otherwise, answer [NO] which means we must search other candidate entities to match with [Main Entity].

NOTICE, Just answer [YES] or [NO]. Your reasoning process should follow [EXAMPLE]s:

{{ Examples }}

Just directly answer [YES] or [NO], don't give other text.

Table 7: Prompt for reasoning

Prompt for reasoning

Now given [Main Entity] $l_e = \text{Entity}(\{\{ ID, Name \text{ and } Tuples \}\})$, and [Candidate Entity] $r_e = \text{Entity}(\{\{ ID, Name \text{ and } Tuples \}\})$,

- Do [Main Entity] and [Candidate Entity] align or match? Think of the answer STEP BY STEP with name, description, structure, time, YOUR OWN KNOWLEDGE:

Step 1, think of [NAME SIMILARITY] = A out of 5, using `self.entity_name`.

Step 2, think of [PROBABILITY OF DESCRIPTION POINTING SAME ENTITY] = B out of 5, using `get_description()` and YOUR OWN KNOWLEDGE.

Step 3, think of [STRUCTURE SIMILARITY] = C out of 5, using `self.tuples`, `get_neighbors()` and `get_relation_information()`.

Step 4, think of [IMAGE SIMILARITY] = D out of 5, using `self.images`.

Step 5, think of [TIME SIMILARITY] = E out of 5, using `get_time_information()`.

NOTICE, the information provided above is not sufficient, so use YOUR OWN KNOWLEDGE to complete them.

Output answer strictly in format: [NAME SIMILARITY] = A out of 5, [PROBABILITY OF DESCRIPTION POINTING SAME ENTITY] = B out of 5, [STRUCTURE SIMILARITY] = C out of 5, [IMAGE SIMILARITY] = D out of 5, [TIME SIMILARITY] = E out of 5.

Table 8: Prompt for MMKG-Code translation, which is also the system prompt.

Prompt for MMKG-Code translation

A Knowledge Graph Entity is defined as follows:

Class Entity:

```
def __init__(self, name, id, tuples=[], images=[]):
    self.entity_name = name
    self.entity_id = id
    self.tuples = tuples
    self.images = images
def get_neighbors(self):
    neighbors = set()
    for head_entity, _, tail_entity, _, _ in self.tuples:
        if head_entity == self.entity_name:
            neighbors.add(tail_entity)
        else:
            neighbors.add(head_entity)
    return list(neighbors)
def get_relation_information(self):
    relation_info = []
    for _, relation, _, _, _ in self.tuples:
        relation_info.append(relation)
    return relation_info
def get_time_information(self):
    time_info = []
    for _, _, _, start_time, end_time in self.tuples:
        time_info.append((start_time, end_time))
    return time_info
def get_description(self, LLM):
    description = LLM(self.entity_name, self.tuples, self.images)
    return description
```

You are a helpful assistant, helping me align or match entities of knowledge graphs according to name information (self.entity_name), description information (get_description()), structure information (self.tuples, get_neighbors(), get_relation_information()), image information (self.images), time information (get_time_information()), YOUR OWN KNOWLEDGE.

Your reasoning process for entity alignment should strictly follow this case step by step:

{{ *reasoning case* }}

[Output Format]: [NAME SIMILARITY] = A out of 5, [PROBABILITY OF DESCRIPTION POINTING SAME ENTITY] = B out of 5, [STRUCTURE SIMILARITY] = C out of 5, [IMAGE SIMILARITY] = D out of 5, [TIME SIMILARITY] = E out of 5. NOTICE, A,B,C,D,E are in range [1, 2, 3, 4, 5], which respectively means [VERY LOW], [LOW], [MEDIUM], [HIGH], [VERY HIGH]. NOTICE, you MUST strictly output like [Output Format].

Table 9: Detailed input prompt of the Case Study.

PROMPT

Now given [Main Entity] l_e = Entity('2846', 'Joséphine_de_Bade', 'Joséphine_de_Bade is Joséphine de Bade, depicted here in a portrait showcasing her as a woman of nobility with a poised and elegant demeanor. She was the consort of Charles-Antoine de Hohenzollern-Sigmaringen and the mother of Carol Ier, contributing to the lineage of the Hohenzollern family.', [(Charles-Antoine_de_Hohenzollern-Sigmaringen, conjoint, Joséphine_de_Bade), (Joséphine_de_Bade, enfants, Carol_Ier), (Joséphine_de_Bade, enfants, Stéphanie_de_Hohenzollern-Sigmaringen), (Marie_de_Hohenzollern-Sigmaringen, mère, Joséphine_de_Bade), (Joséphine_de_Bade, sépulture, Hedingen)]),

and [Candidate Entity] r_e = Entity('13346', 'Princess_Josephine_of_Baden', 'Princess_Josephine_of_Baden is a historical figure depicted in a 19th-century photograph, dressed in attire typical of European nobility of that era. She was a member of the Grand Duchy of Baden's royal family, married Karl Anton, Prince of Hohenzollern, and was the mother of Carol I of Romania and Princess Marie of Hohenzollern-Sigmaringen.', [(Princess_Josephine_of_Baden, placeOfBirth, Grand_Duchy_of_Baden), (Princess_Josephine_of_Baden, issue, Carol_I_of_Romania), (Princess_Josephine_of_Baden, mother, Stéphanie_de_Beauharnais), (Princess_Josephine_of_Baden, placeOfDeath, German_Empire), (Princess_Josephine_of_Baden, after, Infanta_Antónia_of_Portugal)]),

- Compared with other Candidate Entities, do [Main Entity] and [Candidate Entity] align or match? Think of the answer STEP BY STEP with name, description, structure, images, YOUR OWN KNOWLEDGE:
 - Step 1, think of [NAME SIMILARITY] = A out of 5, using self.entity_name.
 - Step 2, think of [PROBABILITY OF DESCRIPTION POINTING SAME ENTITY] = B out of 5, using self.entity_description.
 - Step 3, think of [STRUCTURE SIMILARITY] = C out of 5, using self.tuples, get_neighbors() and get_relation_information().
 - Step 4, think of [IMAGE SIMILARITY] = D out of 5, using self.images.
 - NOTICE, the information provided above is not sufficient, so use YOUR OWN KNOWLEDGE to complete them.
 - Output answer strictly in format: [NAME SIMILARITY] = A out of 5, [PROBABILITY OF DESCRIPTION POINTING SAME ENTITY] = B out of 5, [STRUCTURE SIMILARITY] = C out of 5, [IMAGE SIMILARITY] = D out of 5.
-

Table 10: Detailed model output of the Case Study

OUTPUT

[NAME SIMILARITY] = 4 out of 5,
[PROBABILITY OF DESCRIPTION POINTING SAME ENTITY] = 4 out of 5,
[STRUCTURE SIMILARITY] = 3 out of 5,
[IMAGE SIMILARITY] = 4 out of 5.

Explanation:

Step 1: NAME SIMILARITY

The names "Joséphine de Bade" and "Princess Josephine of Baden" are similar, with the same first name and a similar surname. The difference in spelling between "Bade" and "Baden" is minor, and it is likely that these names refer to the same person. Therefore, [NAME SIMILARITY] = 4 out of 5.

Step 2: PROBABILITY OF DESCRIPTION POINTING SAME ENTITY

The descriptions of both entities mention that they were members of the royal family and married to a prince. They also mention that they had children, including a son named Carol I of Romania. These similarities suggest that the two entities are likely to be the same person. Therefore, [PROBABILITY OF DESCRIPTION POINTING SAME ENTITY] = 4 out of 5.

Step 3: STRUCTURE SIMILARITY

The structures of the two entities are similar in that they both have a parent-child relationship with a prince, and they both have a child named Carol I of Romania. However, the difference in the spellings of the names and the fact that one entity has an additional child, Princess Marie of Hohenzollern-Sigmaringen, reduces the similarity score. Therefore, [STRUCTURE SIMILARITY] = 3 out of 5.

Step 4: IMAGE SIMILARITY

The images of the two entities are similar in that they have the same hairstyle and similar hair accessories, and their facial features are also very similar. Although their clothing are different, the images suggest they could represent the same individual, based on the high similarity in hairstyle and facial features. Therefore, [IMAGE SIMILARITY] = 4 out of 5.

Overall, the combination of the similar names, descriptions, and structures suggests that [Main Entity] Joséphine de Bade and [Candidate Entity] Princess Josephine of Baden are likely to be the same person. Therefore, the answer is [NAME SIMILARITY] = 4 out of 5, [PROBABILITY OF DESCRIPTION POINTING SAME ENTITY] = 4 out of 5, [STRUCTURE SIMILARITY] = 3 out of 5, [IMAGE SIMILARITY] = 4 out of 5.
