

HERGC: Heterogeneous Experts Representation and Generative Completion for Multimodal Knowledge Graphs

Anonymous ACL submission

Abstract

Multimodal knowledge graphs (MMKGs) enrich traditional knowledge graphs (KGs) by incorporating diverse modalities such as images and text. multimodal knowledge graph completion (MMKGC) seeks to exploit these heterogeneous signals to infer missing facts, thereby mitigating the intrinsic incompleteness of MMKGs. Existing MMKGC methods typically leverage only the information contained in the MMKGs under the closed-world assumption and adopt discriminative training objectives, which limits their reasoning capacity during completion. Recent generative completion approaches powered by advanced large language models (LLMs) have shown strong reasoning abilities in unimodal knowledge graph completion, but their potential in MMKGC remains largely unexplored. To bridge this gap, we propose **HERGC**, a **H**eterogeneous **E**xperts **R**epresentation and **G**enerative **C**ompletion framework for MMKGs. HERGC first deploys a Heterogeneous Experts Representation Retriever that enriches and fuses multimodal information and retrieves a compact candidate set for each incomplete triple. It then uses a Generative LLM Predictor fine-tuned on minimal instruction data to accurately identify the correct answer from these candidates. Extensive experiments on three standard MMKG benchmarks demonstrate HERGC’s effectiveness and robustness, achieving state-of-the-art performance.

1 Introduction

Knowledge graphs (KGs) represent real-world facts as triples of entities and their relations, offering a structured semantic representation (Nickel et al., 2015; Ji et al., 2021). Multimodal knowledge graphs (MMKGs) (Zhu et al., 2022; Chen et al., 2024b) extend traditional KGs by incorporating additional modalities such as images and text, thereby enriching the contextual information of entities and enhancing the expressiveness of the graph. both

KGs and MMKGs have been widely adopted in various AI systems, including recommender systems (Wang et al., 2019a; Sun et al., 2020) and large language models (Pan et al., 2024). Moreover, they play an increasingly important role in scientific domains, supporting downstream tasks such as biomedical interaction prediction (Lin et al., 2020; Xiao et al., 2024).

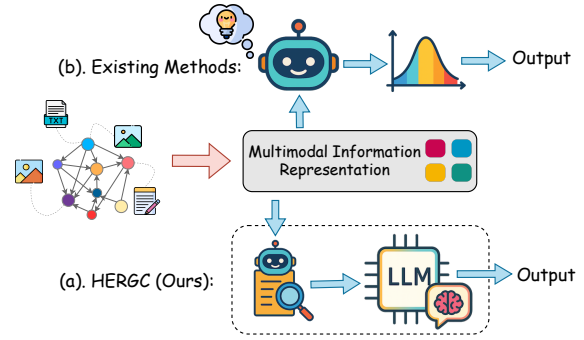


Figure 1: Comparison between (a) HERGC and (b) existing methods. Unlike prior methods, HERGC leverages the knowledge and reasoning capabilities of LLMs to generate predictions.

Like traditional KGs, MMKGs also suffer from incompleteness, often due to missing facts in the underlying data sources or facts that have yet to be discovered by humans. Unlike conventional knowledge graph completion (KGC), which primarily leverages the graph’s topological structure and local neighborhood information, multimodal knowledge graph completion (MMKGC) (Chen et al., 2024b) introduces additional complexity through the incorporation of multimodal signals. MMKGC has advanced considerably in recent years, with most work concentrating on modality fusion (Li et al., 2023; Shang et al., 2024) and modality information representation (Zhang et al., 2025a). These approaches ultimately produce joint triple embeddings or employ ensemble strategies to score candidate facts. While these MMKGC methods

offer valuable insights, their reasoning capabilities during the completion process remain limited due to their reliance solely on closed-world triples and available multimodal information.

Meanwhile, recent advances in unimodal KGC have introduced generative completion approaches powered by large language models (LLMs) (Wei et al., 2023; Zhang et al., 2024; Liu et al., 2024). By fine-tuning or leveraging in-context learning, these approaches exploit the extensive factual knowledge and reasoning capabilities that LLMs acquire during pre-training, achieving state-of-the-art results in KGC. However, this generative paradigm remains largely unexplored within the MMKGC setting. In particular, designing mechanisms to integrate multimodal information to guide LLM reasoning and constrain the search space of its responses remains a critical open challenge.

To address these challenges and fill the gap in generative MMKGC, we propose HERGC, a novel generative framework that overcomes the limitations of existing MMKGC approaches that rely on MMKG’s own contents for reasoning. Figure 1 briefly illustrates the differences between our HERGC and existing methods. Drawing on the retrieval-augmented generation (RAG) idea (Lewis et al., 2020), HERGC comprises two core components: the Heterogeneous Experts Representation Retriever (HERR) and the Generative LLM Predictor (GLP). HERR employs a Mixture of Heterogeneous Experts Network (MoHE) to enrich each modality’s embeddings from multiple and hierarchical perspectives and a Relation-aware Gated Multimodal Unit (RaGMU) to obtain high-quality fused embeddings, which are then used to score and retrieve candidate entities. GLP injects the fused multimodal embeddings into the LLM and performs LoRA fine-tuning on minimal instruction data, enabling the LLM to accurately select the correct entity from the retrieved candidates. We conduct comprehensive experiments on three public MMKG benchmarks to validate the effectiveness and robustness of HERGC. Our contributions are summarized as follows:

- We propose HERGC, the first, to the best of our knowledge, MMKGC framework based on the generative paradigm.
- We developed a novel ranking-based retrieval module, HERR, which combines the Mixture of Heterogeneous Experts Network and the

Relation-aware Gated Multimodal Unit to effectively extract cross-modal information and produce high-quality fused multimodal embeddings.

- We perform extensive experiments on three standard MMKGC benchmarks, showing that HERGC consistently outperforms strong baselines across all datasets.

2 Related Work

2.1 Tradition Knowledge Graph Completion

Tradition Knowledge Graph Completion (KGC) primarily focuses on embedding entities and relations into continuous vector spaces to predict triples. Most of them leverage KG’s structure and design various score function to learn the embedding by maximize the positive and negative samples score difference, such as Translational-Distance approaches (TransE (Bordes et al., 2013) and RotatE (Sun et al., 2019)) and Semantic-Matching approaches (DistMult (Yang et al., 2014), ComplEX (Trouillon et al., 2016), and Tucker (Balazević et al., 2019)). To further consider neighbor aggregation and message passing to improve the representation power of embedding, graph neural network based (GNN-based) methods have been proposed, such as R-GCN (Schlichtkrull et al., 2018) and CompGCN (Vashishth et al., 2019). Besides structural information, KG, as a semantic network, naturally carries text information. Therefore, the text-based method that mainly uses text information, which encodes text information in KG through a pretrained language model (PLM), has been proposed, including KG-Bert (Yao et al., 2019) and SimKGC (Wang et al., 2022). With the recent development of LLM, the novel generative methods have come into view. They mainly use the rich external knowledge and powerful reasoning capabilities of LLMs to complete KGC in a sequence-to-sequence form, including KICGPT (Wei et al., 2023), KoPA (Zhang et al., 2024) and DIFT (Liu et al., 2019).

2.2 Multimodal Knowledge Graph Completion

While recent text-based and generative approaches have started to incorporate both structural and textual information, they often lack tight coordination between these modalities during inference (Chen et al., 2024a). Moreover, the emergence of KGs

enriched with additional modalities, such as images, audio, and video, further raises the bar for the design of dedicated MMKGC models. Initial MMKGC models, like IKRL (Xie et al., 2016), extract visual features from entities using pre-trained visual encoders and combine these with structural embeddings. Extensions such as TransAE (Wang et al., 2019b) and TBKGC (Mousselly-Sergieiev et al., 2018) incorporate both textual and visual features, enhancing entity representations. Fusion-oriented methods, including OTKGE (Cao et al., 2022) and MoSE (Zhao et al., 2022), employ sophisticated strategies like optimal transport and modality-specific representations to achieve effective multimodal integration. IMF (Li et al., 2023) utilizes an interactive fusion framework, training separate models for each modality to collaboratively infer missing links. Furthermore, MMKRL (Lu et al., 2022) employs adversarial training but focuses specifically on robustness against modality-specific perturbations. Meanwhile, approaches like MyGO (Zhang et al., 2025a) leverage fine-grained contrastive learning to enhance the granularity of multimodal embeddings. Also, there are methods that use multi-perspective ideas to enhance modal representation, such as MoMoK (Zhang et al., 2025b) that uses a mixture of expert model and information decoupling and MCKGC (Gao et al., 2025) that integrates information in a mixed curvature space.

3 Methodology

In this section, we present the HERGC framework. We begin with the preliminary, followed by a detailed description of whole workflow. An overview of HERGC is shown in Figure 2.

3.1 Preliminary

Multimodal Knowledge graph (MMKG). In this work, we focus on the most common form of MMKGs with dual visual and textual modalities. An MMKG can be represented as a directed multigraph with modal attributes, denoted as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{V}, \mathcal{D})$, where \mathcal{E} is the set of entities, \mathcal{R} is the set of relations, and $\mathcal{T} = \{(h, r, t) | h, t \in \mathcal{E}, r \in \mathcal{R}\}$ is the set of triples (i.e. (head entity, relation, tail entity)). The \mathcal{V} and \mathcal{D} are the collection of visual images and descriptive text associated with entities.

Multimodal Knowledge graph Completion (MMKGC). MMKGC aims to make full use of

the observed triples \mathcal{T} together with the visual and textual attributes of entities (\mathcal{V} and \mathcal{D}) to infer missing triples. The set of potential facts is defined as $\{(h', r', t') \mid h', t' \in \mathcal{E}, r' \in \mathcal{R}\}$, where $(h', r', t') \notin \mathcal{T}$ represents missing triples in the MMKG. In this work, we formulate MMKGC as the task of completing incomplete query triples of the form $(?, r_q, t_q)$ and $(h_q, r_q, ?)$, corresponding to head prediction and tail prediction, respectively. Here, we refer h_q or t_q the query entity and r_q the the query relation.

3.2 Multimodal Information Embedding

To enable effective fusion of multimodal information, we first encode each entity’s visual and textual modalities into embedding representations, denoted as e_V and e_D , respectively. In addition to these modalities, the graph structure in the KG provides valuable graph contextual cues. Therefore, we also embed the structural information of each entity, resulting in a representation e_S .

Image and Text Embedding. We utilize pre-trained models to encode the visual and textual information associated with each entity. Specifically, we adopt BERT (Devlin et al., 2019), an encoder-only transformer model trained on large-scale textual corpora, to obtain text embeddings, and VGG (Simonyan and Zisserman, 2014), a convolutional neural network trained on large-scale image datasets, to extract visual features. Each entity’s descriptive text and image are processed through BERT and VGG, respectively, yielding the initial modality-specific embeddings e_D and e_V .

Structure Embedding. To encode structural information from the MMKGC, we adopt Tucker (Balazević et al., 2019), a representative KGE model that learns entity and relation embeddings based on tensor factorization. We select Tucker to ensure consistency with the scoring function used in our retrieval module. The resulting embedding is taken as the structural representation e_S for each entity.

3.3 Heterogeneous Experts Representation Retriever

The retriever in KGC is typically a representation model that scores and ranks candidate triples. Under the complex multimodal setting, however, maintaining high-quality entity representations and retrieval sets becomes challenging. To address this, we introduce the Heterogeneous Experts Representation Retriever (HERR). HERR first enhances modality-specific features using a Mixture of Het-

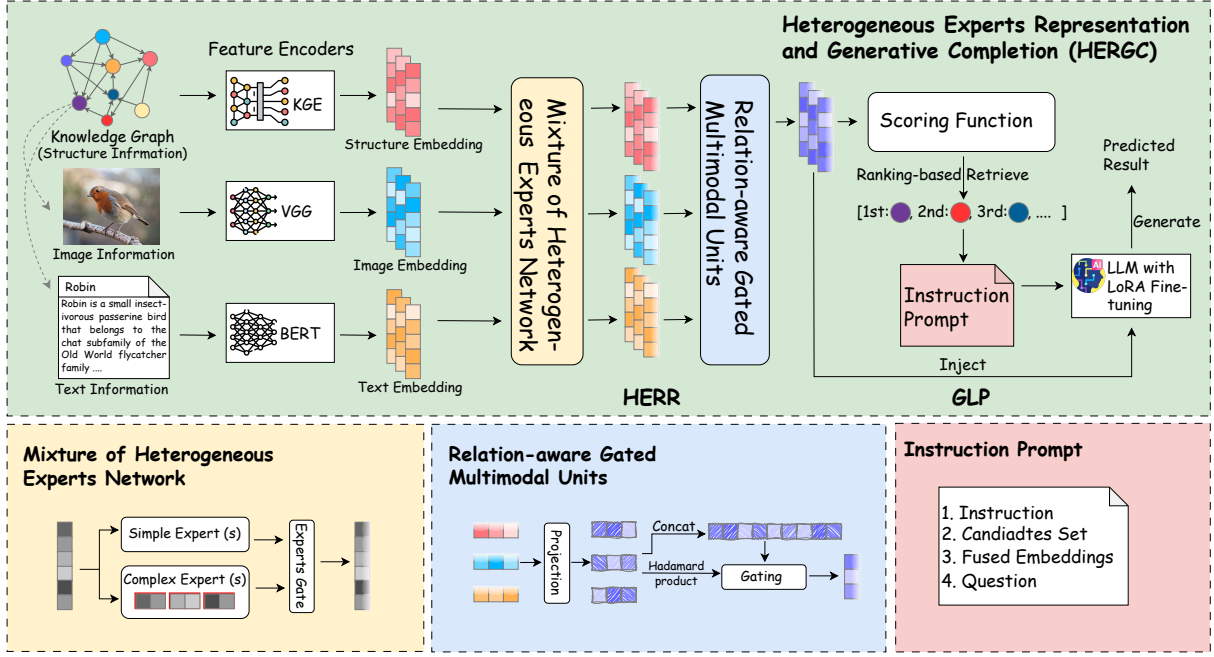


Figure 2: Overview of the HERGC framework

erogeneous Experts network. It then fuses these features via a Relation-aware Gated Multimodal Unit to produce fused embeddings. Finally, HERR employs a scoring function to compute triple scores and generate a ranked candidate list.

Mixture of Heterogeneous Experts Module. To obtain representative and multi-perspective embeddings for each modality we design a heterogeneous experts layer composed of both simple and complex experts. Given the inherent heterogeneity of modality-specific embeddings, our goal is to align them across modalities while preserving their unique characteristics during fusion. To this end, we introduce the Mixture of Heterogeneous Experts (MoHE) module.

The standard Mixture of Experts (MoE) (Shazeer et al., 2017) architecture consists of a set of independent expert networks and a gating network that determines expert assignments. Specifically, given the input feature vector of the modality, MoE outputs the weighted sum of top- κ experts outputs:

$$\mathbf{h}_{i,m} = \sum_{\kappa \in S} G_{\kappa}(\mathbf{x}_{i,m}) E_{\kappa}(\mathbf{x}_{i,m}), \quad (1)$$

where $\mathbf{x}_{i,m}$ and $\mathbf{h}_{i,m}$ denote the input and output embeddings of the i -th entity in modality m , the $\mathbf{x}_{i,m}$ comes from \mathbf{e}_m , $E(\cdot)$ is the expert network, and $S = \text{Top}\kappa(G_{\kappa}(\mathbf{x}_{i,m}))$ denotes the selected expert indices based on gating weights.

The gating weight $G_{\kappa}(\cdot)$ for the corresponding

expert E_{κ} is computed as:

$$G_{\kappa}(\mathbf{x}_{i,m}) = \text{softmax}\left(\frac{\mathbf{W}_{gate}\mathbf{x}_{i,m} + \mathbf{W}_{\epsilon}\mathbf{x}_{i,m}}{\tau}\right), \quad (2)$$

where \mathbf{W}_{gate} is the gate weight parameter matrix, \mathbf{W}_{ϵ} injects noise for exploration, and τ is the gate temperature hyperparameter.

The simplest expert in the MoE layer performs only a linear transformation and feature whitening. To better adapt heterogeneous modalities, we augment the expert set in MoHE with complex PHM experts, which are based on Block-Hypercomplex Linear Transformations (Zhang et al., 2021). Specifically, the input $\mathbf{x} \in \mathbb{R}^d$ will be partitioned into n sub-blocks of size d/n :

$$\mathbf{x} = [\mathbf{x}^{(1)}; \mathbf{x}^{(2)}; \dots; \mathbf{x}^{(n)}], \mathbf{x}^{(j)} \in \mathbb{R}^{d/n}, \quad (3)$$

then each sub-block is transformed by a shared weight matrix $\mathbf{W}_{block} \in \mathbb{R}^{\frac{d}{n} \times \frac{d}{n}}$ and a per-block weight matrix $\mathbf{H}_j \in \mathbb{R}^{\frac{d}{n} \times \frac{d}{n}}$:

$$\mathbf{h}^{(j)} = \mathbf{H}_j \mathbf{W}_{block} \mathbf{x}^{(j)}. \quad (4)$$

Finally, the PHM expert output is obtained by concatenating all transformed sub-blocks.

$$E_{PHM}(\mathbf{x}) = [\mathbf{h}^{(1)}; \mathbf{h}^{(2)}; \dots; \mathbf{h}^{(n)}] \in \mathbb{R}^d. \quad (5)$$

Relation-aware Gated Multimodal Units.

Gated Multimodal Units (Arevalo et al., 2017) provide a "multimodal projection \rightarrow gating \rightarrow

weighted fusion" idea to combine modality-specific representations. To tailor this mechanism to MMKGC, we introduced the Relation-aware Gated Multimodal Units, denoted RaGMU.

Specifically, each modality embedding \mathbf{x}_m is projected into a shared latent space:

$$\mathbf{h}_m = \tanh(\mathbf{W}_{proj,m}\mathbf{x} + \mathbf{b}_{proj,m}) \quad (6)$$

where $\mathbf{W}_{proj,m}$ and $\mathbf{b}_{proj,m}$ are the project matrix and basis of RaGMU projector for modality m .

Next, the gate vector can be calculated by:

$$\mathbf{z} = \text{softmax}(r_z(\mathbf{W}_z\mathbf{h}_{concat} + \mathbf{b}_z)) \quad (7)$$

where $\mathbf{h}_{concat} = [\mathbf{h}_1; \mathbf{h}_2; \dots; \mathbf{h}_M]$ is the concatenation of all projected modality embeddings, the \mathbf{W}_z and \mathbf{b}_z are the gating weight matrix and bias, and the r_z is the scalar that injects relation awareness.

Finally, the fused multimodal embedding is obtained by applying the gate vector to each modality's hidden projection \mathbf{h}_m :

$$\mathbf{h}_{fuse} = \sum_m \mathbf{z}_m \odot \mathbf{h}_m, \quad (8)$$

where \odot denotes Hadamard product and \mathbf{z}_m is the m -th element gate vector corresponding to modality m .

Score Function. After getting the fused multimodal embeddings, we compute triple plausibility scores using the Tucker scoring function:

$$S(h, r, t) = \mathbf{W}_{tucker} \times_1 \mathbf{h}_h \times_2 \mathbf{r} \times_3 \mathbf{h}_t. \quad (9)$$

where \mathbf{h}_h and \mathbf{h}_t denote the fused embeddings of the head h and tail t , \mathbf{r}_r denotes the embedding of the relation r , and \times_n denotes the n -mode tensor product (Balažević et al., 2019).

To train the model, we adopt a binary classification objective that encourages higher scores for positive triples and lower scores for negative ones. The loss function is defined as:

$$\mathcal{L} = - \sum [y \log \sigma(S) + (1 - y) \log(1 - \sigma(S))], \quad (10)$$

where $y \in \{0, 1\}$ is the label indicating whether the triple is positive or negative, and $\sigma(\cdot)$ is the sigmoid function.

3.4 Generative LLM predictor

The Generative LLM Predictor (GLP) aims to predict the correct entity from a set of candidates given a query triple. The query is typically formulated as

a question derived from the query entity and relation. To simplify the design, we use an instruction-based prompt that directly asks the LLM to complete an incomplete triple by selecting an appropriate entity. Additionally, we inject the learned fused multimodal embeddings into the LLM to provide auxiliary information for prediction.

Prompt Template. Taking the tail prediction scenario as an example, we first use HERR to retrieve the ranking of all candidates based on the query $(h_q, r_q, ?)$, ensuring that the resulting triples do not already exist in the MMKG. We then select the top- k candidates, denoted as $C = [e_1, e_2, \dots, e_k]$. A natural language question Q is generated based on the query relation r_q and entity h_q . Finally, we construct a prompt P by combining the instruction I , the question Q , the candidate list C , and the fused embeddings V of h_q and each $e \in C$:

$$P = [I, Q, C, V]. \quad (11)$$

Appendix A.2 presents the detailed prompt template.

LoRA Fine-tuning. We fine-tune open-source LLMs with Low-Rank Adaptation (LoRA), using a small number of query-answer pairs. This lightweight adaptation teaches the model to follow our completion instruction while largely relying on its pretrained knowledge. The injected multimodal embeddings also help guide the LLM toward the correct choice, enabling accurate prediction.

4 Experiments

4.1 Experiment Setup

Dataset. We evaluate our proposed method on three benchmark MMKG datasets, MKG-Y (Xu et al., 2022), MKG-W (Xu et al., 2022) and DB15K (Liu et al., 2019). Dataset statistics and detailed descriptions are provided in Appendix A.1.

Baseline Methods. For MMKGC, we consider both the classic method based on unimodal design and the advanced method based on multimodal design. (1) For unimodal methods, we mainly consider several classic knowledge graph embedding methods: TransE (Bordes et al., 2013), RotatE (Yang et al., 2014), DisMult (Sun et al., 2019), GC-OTE (Tang et al., 2019) and Tucker (Balažević et al., 2019). The baseline comparisons in this paper are based on the reported performance values of these methods (2) For the multimodal methods, we selected a series of state-of-the-art multimodal KGE or KGC models: IKRL (Xie et al., 2016),

Methods	MKG-W				MKG-Y				DB15K			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
Structure-based												
TransE	29.19	21.06	33.20	44.23	30.73	23.45	35.18	43.37	24.86	12.78	31.48	47.07
RotatE	33.67	26.80	36.68	46.73	34.95	29.10	38.35	45.30	29.28	17.87	36.12	49.66
DistMult	20.99	15.93	22.28	30.86	25.04	19.33	27.80	35.95	23.03	14.78	26.28	39.59
GC-OTE	33.92	26.55	35.96	46.05	32.95	26.77	36.44	44.08	31.85	22.11	36.52	51.18
Tucker	30.39	24.44	32.91	41.25	37.05	34.59	38.43	41.45	33.86	25.33	37.91	50.38
Rule-based												
IKRL	32.36	26.11	34.75	44.07	33.22	30.37	34.28	38.26	26.82	14.09	34.93	49.09
TBKG	31.48	25.31	33.98	43.24	33.99	30.47	35.27	40.07	28.40	15.61	37.03	49.86
TransAE	30.00	21.23	34.91	44.72	28.10	25.31	29.10	33.03	28.09	21.25	31.17	41.17
MMKRL	30.10	22.16	34.09	44.69	36.81	31.66	39.79	45.31	26.81	13.85	35.07	49.39
RSME	29.23	23.36	31.97	40.43	34.44	31.78	36.07	39.09	29.76	24.15	32.12	40.29
OTKGE	34.36	28.85	36.25	44.88	35.51	31.97	37.18	41.38	23.86	18.45	25.89	34.23
IMF	34.50	28.77	36.62	45.44	35.79	32.95	37.14	40.63	32.25	24.20	36.00	48.19
QEB	33.38	25.47	35.06	45.32	34.37	29.49	37.00	42.30	28.18	14.82	36.67	51.55
VISTA	32.91	26.12	35.38	45.61	30.45	24.87	32.39	41.53	30.42	22.49	33.56	45.94
MyGO	36.10	29.78	38.54	<u>47.75</u>	38.44	35.01	39.84	44.19	37.72	30.08	41.26	52.21
MoMoK	38.89	30.38	37.54	46.31	37.91	35.09	39.20	43.20	39.54	<u>32.38</u>	43.45	54.14
MCKGC	<u>36.88</u>	<u>31.32</u>	<u>38.92</u>	47.43	<u>38.92</u>	<u>35.49</u>	<u>40.57</u>	<u>45.21</u>	<u>39.79</u>	31.92	<u>43.80</u>	<u>54.66</u>
HERGC	38.89	33.12	41.14	47.86	39.82	36.73	41.42	44.84	40.76	33.09	45.07	54.88

Table 1: Main results of the comparison between HERGC and the baselines on MKG-W, MKG-Y and DB15K. For each metric, the best performance is highlighted in **bold**, and the second-best is underlined.

TBKG (Mousselly-Sergieh et al., 2018), TransAE (Wang et al., 2019b), MMKRL (Lu et al., 2022), RSME (Wang et al., 2021), OTKGE (Cao et al., 2022), IMF (Li et al., 2023), QEB (Lee et al., 2023), VISTA (Lee et al., 2023), MyGO (Zhang et al., 2025a), MoMoK (Zhang et al., 2025b), MCKGC (Gao et al., 2025). The baseline comparisons in this paper are based on the reported performance values of these methods.

Implementation Details. For modality-specific feature extraction, we use bert-base-uncased to encode text, VGG-16 to encode images, and a Tucker model trained on the training split to obtain structural embeddings, ensuring consistency with the retriever’s scoring function. For HERR training, we tune the embedding dimension from {200, 300, 400} and set the batch size to {512, 1024}. We use the Adam optimizer (Kingma and Ba, 2017), with the learning rate selected from {0.005, 0.001, 0.0005}. The MoHE module is configured with 2 simple experts and 2 complex PHM experts. The number of retrieved candidate entities is selected from {10, 20, 30, 40}. For the GLP, we employ LLaMA-3-8B and apply LoRA for parameter-efficient fine-tuning. We set the LoRA hyperparameters to $r = 64$, $\alpha = 16$, a dropout rate of 0.1, and a learning rate of 0.0002. Additional training details are provided in Appendix A.3.

Model performance is evaluated using standard ranking-based metrics: Mean Reciprocal Rank (MRR), and Hits@1, Hits@3, and Hits@10, under the “filtered” setting (Bordes et al., 2013).

All experiments were conducted on an AMD EPYC 7763 64-Core CPU, an NVIDIA A100-SXM4-40GB GPU, an and Rocky Linux 8.10.

4.2 Main Results

Table 1 presents the main results of our proposed HERGC compared with advanced unimodal KGC methods and multimodal KGC methods. HERGC achieves the best overall performance on MKG-Y, MKG-W, and DB15K across most evaluation metrics. Compared to multimodal baselines, HERGC matches MoMoK for the top MRR on MKG-W and falls marginally behind MMKRL and MCKGC in Hits@10 on MKG-Y, it consistently outperforms all other methods on the remaining metrics across the three datasets. Notably, HERGC boosts Hits@1 by an absolute 3.42%, 3.94%, and 3.67% on MKG-Y, MKG-W, and DB15K, respectively, over the strongest baseline in each case.

When compared with unimodal KGC models, HERGC demonstrates substantial gains across all metrics. Specifically, relative to Tucker, HERGC improves MRR by 27.04%, 7.48%, and 20.38% on MKG-Y, MKG-W, and DB15K, respectively,

highlighting the effectiveness of incorporating multimodal information for knowledge graph completion.

4.3 Ablation Studies

To verify the rationality of the HERGC design, we conduct an ablation study consisting of two parts: (1) ablation of modality-specific inputs to assess the contribution of each modality and the model’s ability to leverage multimodal information, and (2) ablation of key components within HERGC, including the design of each part of the retriever and the LLM predictor. The results of the ablation experiment on three datasets are shown in Table 2.

w/o	MKG-W		MKG-Y		DB15K	
	MRR	Hits@1	MRR	Hits@1	MRR	Hits@1
Modality Information						
Image Modality	36.83	31.19	38.57	34.96	39.41	30.74
Text Modality	36.17	30.59	38.42	34.72	39.59	31.18
Structure Modality	37.98	32.34	39.09	36.48	40.17	32.35
Model Components						
Complex Experts	37.95	32.26	39.04	35.51	40.26	32.30
GMU	37.02	31.41	38.96	35.19	39.97	31.28
Relation-awareness	37.56	32.02	39.21	36.14	40.34	32.41
Embedding Injection	37.94	32.26	39.00	35.84	39.05	31.16
HERGC (raw)	38.89	33.12	39.82	36.73	40.76	33.09

Table 2: Ablation study results on three datasets, with a new group of removals above the original ones.

For For modality ablation, we individually remove the textual, visual, and structural information. In all cases, performance declines, indicating that each modality contributes meaningfully to the model’s predictions and that HERGC effectively integrates multimodal information.

For component ablation, we examine the impact of removing complex PHM experts, the GMU fusion module, and relation-awareness in the retriever, as well as embedding injection in the LLM predictor. Removing any of these components results in performance degradation, highlighting their importance. Notably, omitting the embedding injection also leads to a performance drop, indicating that incorporating exogenous fused multimodal embeddings enriched with graph context indeed enhances the LLM’s reasoning capability.

4.4 Representation Visualization

We use t-SNE to visualize the entity representations learned by the HERR on DB15K, providing an intuitive view to directly assess its effectiveness. Specifically, we select entities from the following types: "Writer", "Singer", "Flim", "Com-

pany", "City", "Language" and "College". We compared the fused multimodal embeddings against individual modality embeddings, each projected from high-dimensional space to two dimensions. The visualization result is shown in Figure 3.

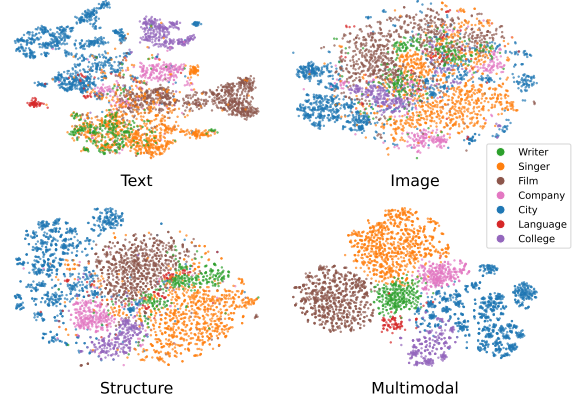


Figure 3: t-SNE data visualization of entity representations learned by the retriever on the DB15K dataset.

The fused embeddings form almost perfectly separated clusters for each entity type, with clear inter-type boundaries and uniform intra-type distributions. By contrast, image-only embeddings exhibit highly entangled regions; structure-only embeddings fail to distinguish the “Language” cluster and yield a diffuse “Writer” grouping; and text-only embeddings conflate “Writer” and “Singer” entities—likely due to their lexical similarity (e.g., names). These observations confirm that HERR effectively integrates multimodal signals to learn high-quality entity representations.

4.5 LLM Predictor Exploration

We further investigate the LLM predictor by examining two factors: (1) the effect of varying the candidate set size k on model performance and inference time, and (2) the impact of using LLMs with different parameter sizes.

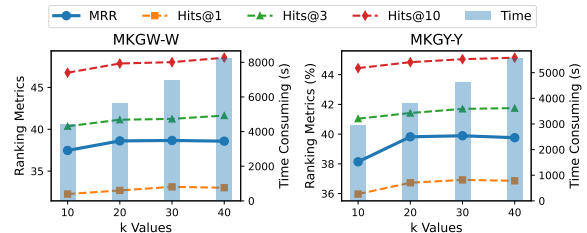


Figure 4: The performance and time consumption of the HERGC on MKG-W and MKG-Y when k takes different values.

Figure 4 shows the trends in time consumption and ranking-based metrics as k varies. As expected, inference time increases approximately linearly with larger k values due to longer prompts constructed from larger candidate sets, which has more tokens in the prompt. However, the performance gains are marginal beyond $k = 20$; only the increase from $k = 10$ to $k = 20$ yields a noticeable improvement in MRR. Considering the trade-off between effectiveness and efficiency, we set $k = 20$ in all experiments.

We also assess the effect of LLM size on the predictor. Table 3 reports the performance and time cost of HERGC using LLMs of different scales. From the results. Although the 3 B model reduces inference time by roughly 30% compared to the 8 B variant, it suffers a modest decline in accuracy, indicating that the more knowledge and better reasoning ability of the larger LLM is indeed helpful for MMKGC prediction. Additional experiment results with large visual models (LVM) and large multi-model models (LMM) are show in Appendix A.5.

Dataset	MRR (Δ)	Hits@1 (Δ)	Hits@3 (Δ)	Hits@10 (Δ)	Time (Δ %)
MKG-W	37.02 (-1.87)	31.41 (-1.71)	39.58 (-1.56)	46.71 (-1.15)	3796 (-32.8)
MKG-Y	38.72 (-1.10)	35.43 (-1.30)	40.44 (-0.98)	44.16 (-0.68)	2448 (-32.5)
DB15K	39.61 (-1.15)	31.18 (-1.91)	44.47 (-0.60)	54.34 (-0.54)	5509 (-29.6)

Table 3: HERGC Performance using Llama-3.2-3B as the LLM predictor (Δ values indicate differences from using Llama-3-8B).

4.6 Complex Environment Simulation

To evaluate HERGC’s robustness under realistic perturbations, we conduct complex environment simulations by: (i) injecting Gaussian noise into a fraction of the modality inputs, (ii) masking portions of the multimodal embeddings, and (iii) randomly removing a subset of training triples from the KG to emulate noisy modalities, missing multimodal information, and sparse graph connectivity, respectively.

Figure 5 reports how MRR degrades as we increase the proportion of corrupted modalities or removed triples. We observe that HERGC is relatively resilient to both noisy and missing multimodal inputs—its performance declines only marginally even when a substantial fraction of embeddings are perturbed or masked. In contrast, removing triples from the KG results in a visible decline in MRR, particularly on MKG-Y. When 30% of the training triples are randomly removed,

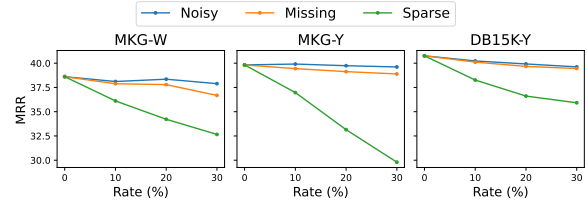


Figure 5: Changes in MRR metrics of HERGC on three datasets under different proportions of simulated interference.

HERGC experiences drops of 15.4%, 25.1%, and 11.8% on MKG-W, MKG-Y, and DB15K, respectively. Nevertheless, the performance degradation remains within a tolerable range, considering the inherent sensitivity of non-inductive KGC tasks to graph sparsity (Pujara et al., 2017). These results highlight HERGC’s robustness and practical applicability in noisy, incomplete, and sparse multimodal scenarios.

5 Conclusion

In this paper, we introduce HERGC, a novel generative framework for multimodal knowledge graph completion that adopts a generative paradigm to addresses the limitations of prior MMKGC approaches. HERGC employs a Heterogeneous Experts Representation Retriever to capture modality-specific signals, fuse them into rich entity embeddings, and retrieve a compact candidate set, thereby integrating multimodal context and narrowing the LLM’s search space. It then uses a flexible Generative LLM Predictor to select the correct entity from these candidates, leveraging the LLM’s inherent knowledge and reasoning capabilities to compensate for previous methods’ shortcomings. Extensive experiments on three standard MMKGC benchmarks demonstrate that HERGC achieves state-of-the-art performance while remaining effective and robust. By fully exploiting each modality’s information, the retriever learns high-quality fused multimodal embeddings, and the LLM predictor readily benefits from advances in LLM technology, making HERGC both powerful and easy to implement. Our HERGC bridges the gap between generative paradigm and MMKGC, offering a generalizable framework for future MMKGC research.

6 Limitations

Although HERGC demonstrates strong performance, its effectiveness is inherently bounded by

the capabilities of the underlying LLMs. In our additional experiments, the LMM and LVM underperformed, likely due to their potential not being fully explored or insufficient adaptation to the MMKG task. Moreover, our approach utilizes commonly used encoders, VGG and BERT, to extract coarse-grained multimodal embeddings. While effective, these general-purpose encoders may overlook modality-specific fine-grained features critical for accurate reasoning. Future work could investigate advanced feature extraction techniques tailored to each modality, enabling richer and more discriminative representations. Additionally, scaling HERGC to other MMKG tasks, such as paths reasoning and question answering, and exploring more efficient training paradigms, such as instruction tuning or parameter-efficient adaptation strategies, could further enhance its generalizability and practicality.

References

- John Arevalo, Tamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.
- Ivana Balažević, Carl Allen, and Timothy M Hospedales. 2019. Tucker: Tensor factorization for knowledge graph completion. *arXiv preprint arXiv:1901.09590*.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.
- Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in neural information processing systems*, 35:39090–39102.
- Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, and 1 others. 2024a. Exploring the potential of large language models (llms) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 25(2):42–61.
- Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, and 1 others. 2024b. Knowledge graphs meet multi-modal learning: A comprehensive survey. *arXiv preprint arXiv:2402.05391*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Yuxiao Gao, Fuwei Zhang, Zhao Zhang, Xiaoshuang Min, and Fuzhen Zhuang. 2025. Mixed-curvature multi-modal knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11699–11707.
- Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514.
- Diederik P. Kingma and Jimmy Ba. 2017. *Adam: A method for stochastic optimization*. Preprint, arXiv:1412.6980.
- Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Whang. 2023. Vista: Visual-textual knowledge graph representation learning. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 7314–7328.
- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, and 1 others. 2015. Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023. Imf: Interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023*, pages 2572–2580.
- Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. 2020. Kgnn: Knowledge graph neural network for drug-drug interaction prediction. In *IJCAI*, volume 380, pages 2739–2745.
- Yang Liu, Xiaobin Tian, Zequn Sun, and Wei Hu. 2024. Finetuning generative large language models with discrimination instructions for knowledge graph completion. In *International Semantic Web Conference*, pages 199–217. Springer.
- Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019. Mmkg: multi-modal knowledge graphs. In *The semantic web: 16th international conference, ESWC 2019, portorož, Slovenia, June 2–6, 2019, proceedings 16*, pages 459–474. Springer.

- Xinyu Lu, Lifang Wang, Zejun Jiang, Shichang He, and Shizhong Liu. 2022. Mmkrl: A robust embedding approach for multi-modal knowledge graph representation learning. *Applied Intelligence*, pages 1–18. 765
- Hatem Mousselly-Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 225–234. 766
- Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2015. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33. 767
- Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiaapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599. 768
- Jay Pujara, Eriq Augustine, and Lise Getoor. 2017. Sparsity and noise: Where knowledge graph embeddings fall short. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1751–1756. 769
- Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, pages 593–607. Springer. 770
- Bin Shang, Yinliang Zhao, Jun Liu, and Di Wang. 2024. Lafa: Multimodal knowledge graph completion with link aware fusion and aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8957–8965. 771
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*. 772
- Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 773
- Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. 774
- Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1405–1414. 775
- Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*. 776
- Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou. 2019. Orthogonal relation transforms with graph context modeling for knowledge graph embedding. *arXiv preprint arXiv:1911.04910*. 777
- Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR. 778
- Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2019. Composition-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1911.03082*. 779
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85. 780
- Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022. Simkgc: Simple contrastive knowledge graph completion with pre-trained language models. *arXiv preprint arXiv:2203.02167*. 781
- Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. 2021. Is visual context really helpful for knowledge graph? a representation learning perspective. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2735–2743. 782
- Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019a. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 950–958. 783
- Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. 2019b. Multimodal data enhanced representation learning for knowledge graphs. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE. 784
- Yanbin Wei, Qiushi Huang, Yu Zhang, and James Kwok. 2023. KICGPT: Large language model with knowledge in context for knowledge graph completion. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8667–8683, Singapore. Association for Computational Linguistics. 785
- Yongkang Xiao, Sinian Zhang, Huixue Zhou, Mingchen Li, Han Yang, and Rui Zhang. 2024. Fuselinker: Leveraging llm’s pre-trained text embeddings and domain knowledge to enhance gnn-based link prediction on biomedical knowledge graphs. *Journal of Biomedical Informatics*, 158:104730. 786
- Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Image-embodied knowledge representation learning. *arXiv preprint arXiv:1609.07028*. 787

Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. 2022. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3857–3866.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kgbert: Bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Aston Zhang, Yi Tay, Shuai Zhang, Alvin Chan, Anh Tuan Luu, Siu Cheung Hui, and Jie Fu. 2021. Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with $1/n$ parameters. *arXiv preprint arXiv:2102.08597*.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. 2025a. Tokenization, fusion, and augmentation: Towards fine-grained multi-modal entity representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 13322–13330.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Wen Zhang, and Huajun Chen. 2024. Making large language models perform better in knowledge graph completion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 233–242.

Yichi Zhang, Zhuo Chen, Lingbing Guo, yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. 2025b. [Multiple heads are better than one: Mixture of modality knowledge experts for entity representation learning](#). In *The Thirteenth International Conference on Learning Representations*.

Yu Zhao, Xiangrui Cai, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao, and Ning Jiang. 2022. Mose: Modality split and ensemble for multi-modal knowledge graph completion. *arXiv preprint arXiv:2210.08821*.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):715–735.

A Appendix

A.1 Details of the Dataset

We evaluate our proposed method on three publicly available multimodal knowledge graph completion (MMKGC) datasets: MKG-Y (Xu et al., 2022), MKG-W (Xu et al., 2022), and DB15K (Liu et al., 2019). MKG-W (CC0 1.0 Public-Domain Dedication) and MKG-Y (CC BY 4.0) are curated subsets extracted from Wikidata (Vrandečić and Krötzsch, 2014) and YAGO (Suchanek et al., 2007) and enriched with comprehensive multimodal information including textual descriptions and associated images. DB15K (CC BY-SA 3.0) originates from DBpedia (Lehmann et al., 2015) and similarly integrates textual and visual modalities to enhance entity representations. All three datasets provide realistic and rich multimodal scenarios, suitable for rigorous benchmarking of knowledge graph completion models. Table 4 presents the statistical details of these three datasets.

Datasets	Entities	Relations	Training	Validation	Testing
MKG-W	15,000	169	34,196	4,276	4,274
MKG-Y	15,000	28	21,310	2,665	2,663
DB15K	12,842	279	79,222	9,902	9,904

Table 4: Statistics of the three datasets.

A.2 Prompt Template

Table 5 is a template with tail prediction as an example. For all three datasets, the prompt template remains consistent generally, comprising a simple instruction, a candidate set, corresponding multimodal fusion embeddings (initially represented by [Placeholder]) for reference. The only difference between the prompts for head prediction and tail prediction is that the question part is a question asking what is the head of an incomplete triple with a missing head.

A.3 Model Training

We train the retriever HERR using the training and test sets following the standard dataset splits of MKG-W, MKG-Y, and DB15K. For training the GLP, we fine-tune the LoRA module with a small number of samples. Specifically, we employ a consistent prompt template to transform the sample triples into query–candidates formats for training. Notably, since the retriever is trained on the training set, the correct entity often receives a high score and is consistently ranked first. To prevent the LLM

Prompt Template for GLP

You are an excellent linguist. The task is to predict the head or tail based on the given incomplete triple, and you only need to answer one entity. The answer must be in ('candidate1', 'candidate2', 'candidate3', 'candidate4', 'candidate5', 'candidate6', 'candidate7', 'candidate8', 'candidate9', 'candidate10', 'candidate11', 'candidate12', 'candidate13', 'candidate14', 'candidate15', 'candidate16', 'candidate17', 'candidate18', 'candidate19', 'candidate20').

You can refer to the entity embeddings: 'query entity': [Placeholder], 'candidate1': [Placeholder], 'candidate2': [Placeholder], 'candidate3': [Placeholder], 'candidate4': [Placeholder], 'candidate5': [Placeholder], 'candidate6': [Placeholder], 'candidate7': [Placeholder], 'candidate8': [Placeholder], 'candidate9': [Placeholder], 'candidate10': [Placeholder], 'candidate11': [Placeholder], 'candidate12': [Placeholder], 'candidate13': [Placeholder], 'candidate14': [Placeholder], 'candidate15': [Placeholder], 'candidate16': [Placeholder], 'candidate17': [Placeholder], 'candidate18': [Placeholder], 'candidate19': [Placeholder], 'candidate20': [Placeholder].

Question: What is the tail in ('query entity', 'query relation', tail)?

Answer:

Table 5: Prompt template for the LLM in predictor GLP (tail prediction example).

from overfitting to this shortcut—i.e., learning the retriever’s ranking pattern rather than making predictions based on textual content—we follow previous work (Wei et al., 2023; Liu et al., 2024) and use the validation set to construct the fine-tuning data for the LLM. Concretely, for MKG-W and MKG-Y, we split the original validation set into a training/validation split for LLM fine-tuning at a 9:1 ratio. For DB15K, we randomly sample 5,000 triples from its original validation set and similarly divide them into training and validation subsets using a 9:1 ratio. The test sets remain identical to the original benchmarks, and we perform both head and tail entity prediction for each test triple, in line with standard KGC evaluation protocols.

A.4 Evaluation Metrics

We employ widely-used ranking metrics in knowledge graph completion: Mean Reciprocal Rank (MRR) and Hits@k.

For each test query triple $(h, r, ?)$ or $(?, r, t)$, the model scores every candidate entity, producing a ranked list. All metrics are reported under the *filtered* setting, where corrupted triples that already exist in the KG are removed (Bordes et al., 2013).

Mean Reciprocal Rank (MRR). Let rank_i denote the position of the correct entity for the i -th query in the filtered list. The reciprocal rank is $1/\text{rank}_i$.

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^N \frac{1}{\text{rank}_i},$$

where N is the total number of test queries. MRR ranges from 0 to 1; higher values indicate better overall ranking quality.

Hits@k. Hits@k measures the proportion of queries whose correct entity appears within the top k positions:

$$\text{Hits@k} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[\text{rank}_i \leq k],$$

where $\mathbb{1}[\cdot]$ is the indicator function. Throughout the paper we report Hits@1, Hits@3, and Hits@10, providing a fine-grained view of top-rank accuracy under varying tolerance levels.

A.5 Evaluating LVM and LMM as Predictors

To further investigate the GLP component, we experimented with replacing the LLM in GLP with a large vision model (LVM) or a large multimodal model (LMM), enabling the predictor to directly incorporate the image of the query entity in addition to textual input. The results, presented in Table 6, show that this modification did not lead to the expected improvements. Specifically, substituting the LLM with an LVM resulted in a marked reduction in overall performance, whereas replacing it with an LMM offered no substantial benefit, yielding only marginal gains in MRR (+1.1%) and Hits@10 (+0.8%).

Dataset	MRR	Hits@1	Hits@3	Hits@10
Llama-3-8B	39.82	36.73	41.42	44.84
Llava-1.5-7B	27.87	15.79	38.72	42.68
Llama-3.2-11B-Vision	40.26	36.31	40.91	45.22

Table 6: HERGC Performance using Llava-1.5-7B and Llama-3.2-11B-Vision as the LLM predictor on MKG-Y.

The performance degradation observed when replacing the LLM component in GLP with LLaVA-1.5-7B may be attributed to limitations in its backbone architecture and pre-training objectives. Specifically, LLaVA-1.5-7B utilizes Llama-2-7B as its backbone, which inherently possesses weaker language modeling capabilities compared to more

advanced models such as Llama-3. Moreover, LLaVA-1.5-7B is fine-tuned primarily using CLIP-based visual features and visual-language instructions, with its pre-training tasks heavily centered on image-text alignment and visual question-answering, rather than structured relational reasoning. Consequently, even after subsequent LoRA fine-tuning, the limited number of training examples might be insufficient to effectively transition the model from merely "understanding images" toward "leveraging images for relational inference in knowledge graph completion."

Similarly, the modest performance gains achieved by replacing the LLM component with Llama-3.2-11B-Vision might be due to the already mature textual reasoning capability of its underlying model, Llama-3-8B. Given the strong inherent language modeling performance of Llama-3, the additional inclusion of visual features likely provides minimal incremental benefit for relational prediction. Although large multimodal models (LMMs) generally excel at capturing visual semantics due to extensive pre-training on image-text corpora, they are not typically fine-tuned for structured relational inference tasks such as KGC. Therefore, it remains challenging for these models to accurately extract and leverage KGC-relevant relational signals from images with only a limited number of fine-tuning samples (as imposed by the LoRA rank constraints). Another potential factor is that visual information within MMKG datasets might inherently have weak correlations with the relational semantics required by the KGC task. As a result, effectively utilizing fine-grained relational clues from entity images for MMKGC remains an open and promising research direction.