HFR-MKGC: Hierarchical Fusion Reasoning with MLLMs for Multi-modal Knowledge Graph Completion

Di Wang[1,2], Junping Du[1,2*], Zhe Xue[1,2], Meiyu Liang[1,2], Guanhua Ye[1,2], Yingxia Shao[1,2], Haisheng Li[3]

[1]School of Computer Science (National Pilot School of Software Engineering), Beijing University of Posts and Telecommunications
[2]Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia
[3]School of Computer and Artificial Intelligence, Beijing Technology and Business University
{wdd123, xuezhe, meiyu1210, g.ye, shaoyx}@bupt.edu.cn,{junpingdu}@126.com, lihsh@btbu.edu.cn

Abstract

Multi-modal knowledge graph completion (MMKGC) aims to infer missing entities of triples by leveraging heterogeneous information in knowledge graph (KG). However, existing approaches often struggle with inconsistent modality alignment, limited reasoning depth, and insufficient negative sample quality. In this work, we propose HFR-MKGC, a novel framework that integrates hierarchical modal fusion and Multimodal Large Language Model (MLLM) reasoning for robust and expressive MMKGC. Specifically, we introduce a relation-guided hierarchical modal fusion module, which conducts fine-grained intra-visual fusion and relation-guided cross-modal integration to yield rich entity representations. HFR-MKGC employs a fine-tuned MLLM to perform instruction-based triple reasoning, producing candidate entities for completion. Then, it constructs hard negative samples through textual perturbation by MLLM and visual feature augmentation with rotation and noise. HFR-MKGC optimizes the model via adversarial training. Extensive experiments on three MMKGC benchmarks demonstrate that our method outperforms state-of-the-art methods, validating its effectiveness in MMKGC.

Introduction

Knowledge graphs (KGs) store real-world facts as triples (h, r, t), where h, t, and r denote the head entity, tail entity, and their relation. While widely used in question answering (Panda et al. 2024; Wang et al. 2025), recommender systems (Cui et al. 2025; Xiao et al. 2022), and semantic search (Liang et al. 2024a,b), KGs often suffer from incompleteness. Traditional knowledge graph completion (KGC) methods (Bordes et al. 2013; Sun et al. 2019; Yang et al. 2015; Trouillon et al. 2016) rely mainly on structural information, overlooking complementary semantic cues from other modalities.

To address this, MMKGC leverages additional modalities such as text and images to predict incomplete triples. Typical MMKGC methods embed entities and relations from different modalities into a shared vector space (Gao et al. 2025) and employ a scoring function to rank candidate triples. Representative approaches include early fusion strategies, which directly concatenate or average modality embeddings (Mousselly-Sergieh et al. 2018), and advanced fusion methods, which learn joint representations through attention mechanisms (Chen et al. 2022), gating modules (Wang et al. 2021), or multi-level integration (Wang et al. 2019). Furthermore, some methods construct multi-modal hard negatives for adversarial training by injecting noise or perturbations (Zhang, Chen, and Zhang 2023; Zhang et al. 2024c; Chen et al. 2025a). Recently, the rapid progress of Large Language Models (LLMs) has inspired new approaches for KGC. By converting structural and textual information from KGs into natural-language prompts (Fang et al. 2024; Barile, d'Amato, and Fanizzi 2025; Pan et al. 2025; Zhang et al. 2024b; Yao et al. 2025), LLMs can perform reasoning to infer plausible missing triples.

Despite the progress of study for MMKGC, existing approaches still face two key challenges: (1) **Lack of relation-guided multi-modal fusion.** Most fusion strategies treat all modalities equally, ignoring that the importance of each modality can vary across different relations. This relation-agnostic design makes the model vulnerable to noisy or irrelevant modalities, especially in scenarios where visual or textual information is weakly related to the relation. (2) **Limited adaptability of LLM-based reasoning.** Current LLM-assisted KGC methods primarily focus on textual and structural modalities, often relying on static fusion strategies. Such designs fail to fully exploit visual semantics and cannot dynamically adapt to the varying relevance of modalities across different triples.

To address these aforementioned challenges, we propose a method called **H**ierarchical **F**usion **R**easoning with **M**LLMs for **M**ulti-modal **K**nowledge **G**raph **C**ompletion (HFR-MKGC). To clearly illustrate the gap between conventional methods and our proposed approach, we compare three paradigms in Figure 1: conventional MMKGC framework, LLM-based KGC framework and our HFR-MKGC framework. HFR-MKGC utilizes a MLLM for auxiliary reasoning, and proposes the following two strategies: (1) **Relation-Guided Hierarchical Modal Fusion**. We employ a MLLM to generate fine-grained textual descriptions for each entity image and fuse them with visual embeddings, capturing complementary visual–semantic cues. Then, a relation-

*Corresponding author.(junpingdu@126.com)

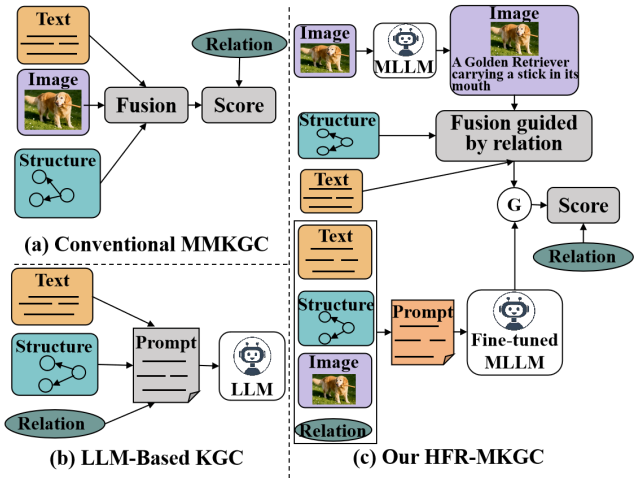

Figure 1: Comparison of three MMKGC paradigms.

guided cross-modal fusion mechanism adaptively weights textual, visual, and structural modalities based on the current relation, enabling robust fusion while suppressing irrelevant or noisy signals. (2) **MLLMs-Enhanced Knowledge Reasoning and Scoring**. We leverage a fine-tuned MLLM to perform triple-specific reasoning, generating candidate tail entities based on structural, textual, and visual cues. These reasoning outputs are integrated into the final scoring function via a gating mechanism, enabling the model to adapt its reasoning strategy to diverse inference scenarios. Finally, we construct hard negative samples via multi-modal perturbation and augmentation, and optimize the entire model under an adversarial training framework to further enhance robustness and generalization. The primary contributions of our work are summarized as follows:

- We propose HFR-MKGC, a novel framework for MMKGC which enables fine-grained intra-visual fusion and relation-guided cross-modal feature weighting, effectively suppressing irrelevant or noisy modalities. HFR-MKGC further enhances its robustness through adversarial training with multi-modal negative samples.
- We design a novel MLLMs-Enhanced Knowledge Reasoning and Scoring strategy, where a fine-tuned MLLM performs triple-specific reasoning and the reasoning outputs are integrated into the scoring process, allowing adaptive reasoning across diverse inference scenarios.
- We conduct comprehensive experiments and in-depth analysis on three public MMKGC datasets, demonstrating that HFR-MKGC achieves **SOTA** performance against 14 baselines.

Related Work

To provide a comprehensive context, we offer a brief introduction to MMKGC and the emerging use of LLMs, especially MLLMs, for knowledge reasoning tasks on KGs

Multi-modal Knowledge Graph Completion

MMKGC approaches integrate complementary modalities such as text (Xie et al. 2016) and images (Xie et al. 2017) to enrich entity and relation representations. Common strategies include simple vector concatenation (Mousselly-Sergieh et al. 2018), heterogeneous graph modeling (Zhao et al. 2025; Wang et al. 2025), and joint embedding learning (Li, Yu, and He 2024; Li et al. 2023). TransAE (Wang et al. 2019) jointly encodes textual and visual features via a multi-modal autoencoder. MKGformer (Chen et al. 2022) introduces hierarchical fusion to filter out irrelevant visual noise, while RSME (Wang et al. 2021) employs gating mechanisms to suppress uninformative modalities. MCKGC (Gao et al. 2025) embed modalities into diverse geometric spaces and use hierarchical gating for integration. MACO (Zhang, Chen, and Zhang 2023) tackles missing modality scenarios by adding noise to visual and structural views to simulate negative examples. AdaMF-MAT (Zhang et al. 2024c) learns adaptive fusion weights across modalities via adversarial optimization, and NativeE (Zhang et al. 2024a) expands this idea to numeric, audio, and video modalities under a unified training framework. Noise-based (Chen et al. 2025a; Jian et al. 2025) and diffusion-based models (Chen et al. 2025b) to generative negative samples further improve generalization.

Most prior works adopt relation-agnostic fusion and are susceptible to noisy. In contrast, we introduce a relation-guided hierarchical fusion mechanism that enables fine-grained alignment and adaptively emphasizes the most relevant semantic cues for relations.

Large Language Models for Knowledge Reasoning

In recent years, LLMs have shown remarkable potential in knowledge reasoning to integrate diverse information sources (Zhang et al. 2024d). GAugLLM (Fang et al. 2024) improves self-supervised learning on text-attributed graphs by prompt-based augmentation. Some studies translate KG structures into text formats understandable by LLMs for entity alignment(Jiang et al. 2024; Zhang et al. 2023) and KGC(Zhang et al. 2024b; Yao et al. 2025). LP-DIXIT (Barile, d'Amato, and Fanizzi 2025) evaluates the explainability of link prediction results using LLMs. Furthermore, LoLLM (Pan et al. 2025) integrates LLM with the structure embedding for reasoning over sparse KGs. HOLMES (Panda et al. 2024) integrates a distilled and context-aware KG into LLMs to support question answering. Recently, researchers have combined MMKGs with LLMs. UniMEL (Liu et al. 2024) proposes a unified LLM-based multi-modal framework that effectively fuses textual and visual information for entity linking. In addition, MuKDC (Li et al. 2024) enhances few-shot KGC through multi-level knowledge generation, while CATS (Li et al. 2025) improves inductive KGC via LLMs.

Existing LLM-based KG reasoning mainly focuses on textual and structural modalities with static fusion, limiting adaptability. We extend this to MLLMs for unified knowledge reasoning over structural, textual, and visual modalities, with a gating mechanism for dynamic fusion with triple-specific to boost MMKGC performance.

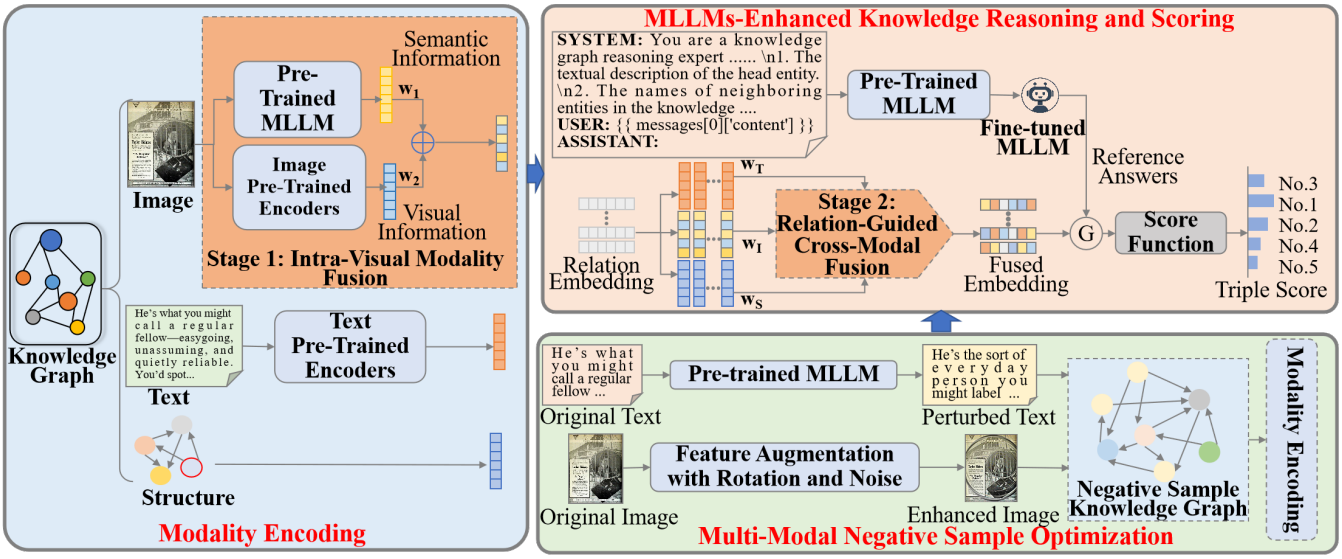

Figure 2: Overall framework of HFR-MKGC. (1) **Stage 1: Intra-Visual Modality Fusion** and **Stage 2: Relation-Guided Cross-Modal Fusion** form the Relation-Guided Hierarchical Modal Fusion module, which performs fine-grained intra-modal fusion followed by relation-guided weighting for cross-modal integration. (2) **MLLMs-Enhanced Knowledge Reasoning and Scoring**: A fine-tuned MLLM generates candidate entities for incomplete triples, whose reasoning outputs are fused with multi-modal embeddings for final scoring. (3) **Multi-modal Negative Sample Optimization**: Hard negatives are constructed via MLLM-based textual perturbation and visual feature augmentation with rotation and noise, and the model is trained using adversarial learning to enhance robustness and discriminative ability.

Task Definition

A MMKG is defined as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ is the set of entities, $\mathcal{R}$ is the set of relations, and $\mathcal{T} = \{(h, r, t) \mid h, t \in \mathcal{E}, r \in \mathcal{R}\}$ denotes the set of triples. Each entity $e \in \mathcal{E}$ can be associated with multi-modal information $\mathcal{M}_e = \{T, I, S\}$, representing textual descriptions, images, and structural context. An incomplete triple can also be referred to as a query triple. The task of MMKGC aims to predict the missing entity in a query triple (e.g., $(h, r, ?)$ or $(?, r, t)$) by leveraging these multi-modal sources of information. This is achieved by learning a scoring function $f(h, r, t)$ to rank candidate entities and select the most plausible one for completion.

Methodology

In this section, we present the details of HFR-MKGC. As shown in Figure 2, HFR-MKGC consists of three key modules: (1) Relation-Guided Hierarchical Modal Fusion. (2) MLLM-Enhanced Reasoning and Scoring. (3) Multi-Modal Negative Sample Optimization.

Relation-Guided Hierarchical Modal Fusion

Relation-Guided Hierarchical Modal Fusion aims to achieve deep integration of multi-modal features through a two-stage fusion strategy. The first stage is **intra-modal fusion level**, which employs MLLM to guide fine-grained fusion between visual and semantic representation of images. The second stage is **inter-modal fusion level**, which dynamically allocates cross-modal fusion weights based on relational information and generates a robust joint feature representation.

Modality Encoding Given an entity e with different modalities, we encode the visual features of images through CLIP (Radford et al. 2021) and encode text features through BERT (Devlin et al. 2019). Moreover, for each entity $e \in \mathcal{E}$ and modality $m \in \mathcal{M}_e$, we denote its modality-specific embedding as $\mathbf{e}^m = P_m(f_m) \in \mathbb{R}^{d_e}$, where f_m represents the initial feature extracted by the pre-trained encoder, and P_m denotes a learnable projection layer for modality m. A noteworthy point is that the generated image caption is also regarded as a type of textual modality.

Stage 1: Intra-Visual Modality Fusion To mitigate intra-modal inconsistency, we utilize a MLLM, LLaVA (Liu et al. 2023), to generate a descriptive captions as sematic information for images. We encode these captions using BERT and project them into the unified feature space by P_T, denoted as $\mathbf{e}^{I_c} \in \mathbb{R}^d$. Similarly, the visual representation of the raw image, denoted as $\mathbf{e}^{I_v} \in \mathbb{R}^d$. Then, we employ a similarity-based gating mechanism to perform the fine-grained fusion between $\mathbf{e}^{I_c}$ and $\mathbf{e}^{I_v}$, which is computed as:

$$\mathbf{e}^I = \sigma\left(\langle \mathbf{e}^{I_v}, \mathbf{e}^{I_c} \rangle\right) \cdot \mathbf{e}^{I_c} + \left(1 - \sigma\left(\langle \mathbf{e}^{I_v}, \mathbf{e}^{I_c} \rangle\right)\right) \cdot \mathbf{e}^{I_v} \quad (1)$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function, and $\langle \cdot, \cdot \rangle$ represents the dot-product similarity between $\mathbf{e}^{I_v}$ and $\mathbf{e}^{I_c}$. The resulting fused vector $\mathbf{e}^I \in \mathbb{R}^d$ integrating both low-level visual cues and high-level semantic context.

Stage 2: Relation-Guided Cross-Modal Fusion In this stage, we integrate the three projected vectors of each entity: $\mathbf{e}^S$(structural), $\mathbf{e}^I$(visual), and $\mathbf{e}^T$ (textual) based on relation-guided interactions. Let $\mathbf{r} \in \mathbb{R}^d$ be the relation

embedding corresponding to the current query triple. The head entity h of the triple (h, r, t) can be first denote as $\mathcal{H}_e = \{\mathbf{e}^S, \mathbf{e}^I, \mathbf{e}^T\} \in \mathbb{R}^{3\times d}$. To capture intrinsic correlations between different modalities, we compute a modality similarity matrix $\mathbf{A}_{i,j} = \langle \mathbf{e}^i, \mathbf{e}^j \rangle$ based on inner product, where $i, j \in \mathcal{M}_e = \{S, I, T\}$ denote different modalities. To estimate the importance of each modality, we compute a relation-guided interaction score α^m for each modality $m \in \mathcal{M}_e$ by:

$$\alpha^m = \langle \mathbf{e}^m, \mathbf{r} \rangle + \frac{1}{|\mathcal{M}_e|} \sum_{j \in \mathcal{M}_e} \mathbf{A}_{m,j} \tag{2}$$

Furthermore, the weight assigned to modality $m \in \mathcal{M}$ is computed as:

$$w^m = \frac{\exp\left(\alpha^m - \frac{1}{|\mathcal{M}_e|}\sum_{k\in\mathcal{M}_e}\alpha^k\right)}{\sum_{j\in\mathcal{M}_e}\exp\left(\alpha^j - \frac{1}{|\mathcal{M}_e|}\sum_{k\in\mathcal{M}_e}\alpha^k\right)} \tag{3}$$

where the fusion weight w^m reflects the relative importance of modality m in the context of the current relation. The final joint embedding of the head entity h is computed as a weighted aggregation of all modality-specific embeddings:

$$\mathcal{H}_{\text{joint}} = \sum_{m\in\mathcal{M}_e} w^m \cdot \mathbf{e}^m \tag{4}$$

This fusion adaptively adjusts the contribution of each modality based on its semantic consistency and relational relevance, resulting in a context-aware entity representation.

MLLMs-Enhanced Knowledge Reasoning and Scoring

MMKGC relies solely on structural or multi-modal embeddings, which is often insufficient for modeling complex semantic relationships. In this subsection, we introduce a fine-tuned MLLM to generate candidate reasoning results, which are then fused with multi-modal entity representations via a gating mechanism before scoring the triples.

Instruction Fine-Tuning For each incomplete triple (taking the missing tail entity as an example), we decompose the instruction into two components: fixed instruction and variable instruction. The fixed instruction $Prom_{fix}$ remains consistent across all training samples and specifies the task type for the MLLM as well as the required output format. The variable instruction encodes triple-specific information derived from each triple (h, r, t), including textual description, a set of neighboring entities, the image of h, and the relation r. In fine-tuning process, the target entity t was employed as the supervisory information. Specifically, we adopt Low-Rank Adaptation (LoRA) to efficiently fine-tune the LLaVa while preserving its general capabilities.

Fine-tuned MLLM Reasoning Given an incomplete triple $(h, r, ?)$, our goal is to predict the missing entity by leveraging the reasoning capabilities of a fine-tuned MLLM. We construct a structured prompt $Prom_{var}(h, r)$ based on the available head entity information. The fixed prompt and $Prom_{var}$ is then fed into the fine-tuned MLLM to perform reasoning and generate the missing entity. Formally, the predicted tail entity is denoted as:

$$\hat{t}^{\text{MLLM}} = \text{MLLM}(Prom_{fix} \oplus Prom_{var}(h, r)) \tag{5}$$

where $\oplus$ is fusion operation. If the missing entity is head entity instead, the prompt $Prom_{var}(t, r)$ can be constructed analogously by using the multi-modal information of t.

Gated Fusion To effectively integrate the reasoning results generated by the fine-tuned MLLM into the triple decoding process, we design a gated fusion mechanism in the complex embedding space. The joint embeddings of the known head entity and a candidate tail entity are denoted as $\mathcal{H}_{\text{joint}}$ and $\mathcal{T}_{\text{joint}}$ respectively. The MLLM-predicted tail entity representation is $\mathcal{T}_{\text{MLLM}} = \text{BERT}(\hat{t}^{\text{MLLM}})$. We first decompose $\mathcal{H}_{\text{joint}}$ into real and imaginary parts: $\text{Re}(\mathcal{H}_{\text{joint}})$ and $\text{Im}(\mathcal{H}_{\text{joint}})$. $\mathcal{T}_{\text{joint}}$ and $\mathcal{T}_{\text{MLLM}}$ undergo a similar operation. The relation embedding $\mathbf{r} \in \mathbb{R}^d$ is scaled into a phase angle and construct the complex-valued relation as:

$$\mathbf{r}_{\mathbb{C}} = \cos\left(\frac{\mathbf{r}}{\gamma/\pi}\right) + i \cdot \sin\left(\frac{\mathbf{r}}{\gamma/\pi}\right) \tag{6}$$

where $\mathbf{r}_{\mathbb{C}} \in \mathbb{C}^d$ denotes the complex representation of r. The scalar γ controls the embedding range. To learn a soft weighting vector $\mathbf{g} \in [0, 1]^{2d}$ for fusing the joint representation and the MLLM-inferred embedding, we design a gating module parameterized as a two-layer neural network:

$$\mathbf{g} = \sigma\left(\mathbf{W}_2 \cdot \tanh\left(\mathbf{W}_1 \cdot \mathcal{T}_{\text{MLLM}} + \mathbf{b}_1\right) + \mathbf{b}_2\right) \tag{7}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d\times 2d}$ and $\mathbf{W}_2 \in \mathbb{R}^{2d\times d}$ are weight matrices, $\mathbf{b}_1 \in \mathbb{R}^d$ and $\mathbf{b}_2 \in \mathbb{R}^{2d}$ are bias terms. The final gate vector $\mathbf{g} = [\mathbf{g}_{\text{re}}; \mathbf{g}_{\text{im}}]$ is split to control real and imaginary parts independently. The fused tail entity embedding $\tilde{\mathcal{T}}_{\text{joint}} = [\text{Re}(\tilde{\mathcal{T}}_{\text{joint}}); \text{Im}(\tilde{\mathcal{T}}_{\text{joint}})]$ is computed by interpolating between the joint representation $\mathcal{T}_{\text{joint}}$ and $\mathcal{T}_{\text{MLLM}}$:

$$\text{Re}(\tilde{\mathcal{T}}_{\text{joint}}) = (1 - \mathbf{g}_{\text{re}}) \circ \text{Re}(\mathcal{T}_{\text{joint}}) + \mathbf{g}_{\text{re}} \circ \text{Re}(\mathcal{T}_{\text{MLLM}}) \tag{8}$$

$$\text{Im}(\tilde{\mathcal{T}}_{\text{joint}}) = (1 - \mathbf{g}_{\text{im}}) \circ \text{Im}(\mathcal{T}_{\text{joint}}) + \mathbf{g}_{\text{im}} \circ \text{Im}(\mathcal{T}_{\text{MLLM}}) \tag{9}$$

where $\circ$ denotes element-wise multiplication.

RotatE-Based Scoring We employ the RotatE scoring function to in the complex vector space:

$$\text{score}(h, r, t) = -\left\|\tilde{\mathcal{H}}_{\text{joint}} \circ \mathbf{r}_{\mathbb{C}} - \tilde{\mathcal{T}}_{\text{joint}}\right\| \tag{10}$$

where $\circ$ denotes element-wise complex multiplication.

Multi-Modal Negative Sample Optimization

We design a negative sample optimization strategy that generates hard negatives across modalities. Specifically, MLLM-based textual perturbation generates new entity descriptions at the semantic level, visual feature augmentation with rotation and noise introduces controlled perturbations in the visual space, and adversarial training leverages hard negatives to improve model discrimination.

MLLM-Based Textual Perturbation Given the original textual description $\mathbf{x}^t_{\text{orig}}$ of an entity, we input it into the MLLM to generate a semantically equivalent but syntactically varied version: $\tilde{\mathbf{e}}^t_{\text{pert}} = P_t\left(E_{\text{text}}\left(\text{MLLM}(\mathbf{x}^t_{\text{orig}})\right)\right)$, where $\mathbf{x}^t_{\text{orig}}$ is the original entity description, $E_{\text{text}}(\cdot)$ is textual encoder, and P_t is a projection layer for textual features.

Visual Feature Augmentation with Rotation and Noise We design a lightweight augmentation module that performs dynamic rotation and noise injection on raw visual features. Given the visual embedding $\mathbf{x}^v_{\text{orig}}$, its perturbed feature is:

$$\tilde{\mathbf{e}}^v_{\text{pert}} = ((1-\lambda)\mathbf{I} + \lambda\mathbf{R})\,\mathbf{x}^v_{\text{orig}} + \boldsymbol{\epsilon}, \tag{11}$$

where $\mathbf{I}$ is the d-dimensional identity matrix, $\mathbf{R}$ is a rotation matrix obtained via QR decomposition from randomly sampled low-rank matrices, $\lambda \in [0,1]$ is a rotation strength factor linearly increased with training epoch, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is Gaussian noise.

Then, we can obtain **multi-modal negative samples** (h^*, r, t), (h, r, t^*), and (h^*, r, t^*), and compute their plausibility scores by score$(\cdot)$.

Adversarial Training To encourage the model to discriminate between plausible and implausible triples, we adopt an adversarially weighted sigmoid loss that emphasizes harder negatives. For the i-th positive triple (h_i, r_i, t_i), a set of **structural negative triples** is generated by replacing h_i or t_i with a random entity $e' \in \mathcal{E}$. Let s_i^+ denote the model-assigned plausibility score for the i-th positive triple, and $s_{i,j}^-$ the score for its j-th negative triple. Here, each negative triple is sampled from either multi-modal negative samples or structural negative triples. Each $s_{i,j}^-$ is assigned a softmax-based weight $w_i^j = \frac{\exp(\tau \cdot s_{i,j}^-)}{\sum_{k=1}^K \exp(\tau \cdot s_{i,k}^-)}$, with τ controlling the sharpness of the distribution. During training, the negative sampling-based loss function is defined as:

$$\mathcal{L}_{kgc} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\sigma(s_i^+) + \sum_{j=1}^{K} w_i^j \log\sigma(-s_{i,j}^-)\right) \tag{12}$$

where N is the number of positive samples, K is the number of negative samples for per positive, $s_i^+ = \text{score}(h_i, r_i, t_i)$ and $s_{i,j}^-$ is the score of the j-th negative triple.

To train the generator to produce more challenging multi-modal negative samples during adversarial learning, we minimize the following loss:

$$\mathcal{L}_g = \frac{1}{M}\sum_{k=1}^{M}\mathbb{E}_{(h,r,t)\sim\mathcal{T}}\left[\max(0, c - s_k)\right] \tag{13}$$

where $\mathcal{L}_g$ is the generator loss, $M = 3$ denotes the number of generated multi-modal negative sample types, s_k is the plausibility score assigned by the scoring model for the k-th type of generated negative triples, c is a fixed margin that encourages the generator to produce harder-to-distinguish samples. The overall training objective combines the KGC loss $\mathcal{L}_{kgc}$ with adversarial learning to enhance multi-modal negative sampling. The generator is optimized by minimizing $\mathcal{L}_g$ to produce challenging multi-modal negative triples that confuse discriminator. The gradient penalty is incorporated to stabilize adversarial training and the training objectives are defined as:

$$\begin{cases} \mathcal{L}_{\mathcal{D}} = \mathcal{L}_{\text{kgc}} + \mu_1 \mathbb{E}_{\hat{f}\sim\mathbb{P}_{\hat{f}}}\left[\left(\left\|\nabla_{\hat{f}}\mathcal{D}(\hat{f})\right\|_2 - 1\right)^2\right] \\ \mathcal{L}_g = \dfrac{1}{M}\displaystyle\sum_{k=1}^{M}\mathbb{E}_{(h,r,t)\sim\mathcal{T}}\left[\max(0, c - s_k)\right] \end{cases} \tag{14}$$

where $\mathcal{D}$ is the discriminator function score$(\cdot)$, μ_1 is the weighting coefficient for the gradient penalty term, $\hat{f}$ denotes the interpolated embedding between real and fake representations, $\nabla_{\hat{f}}\mathcal{D}(\hat{f})$ denotes gradient of the interpolation samples, and $\mathbb{P}_{\hat{f}}$ denotes the sampling distribution of $\hat{f}$.

Experiments

Experiment Settings

Datasets We conduct experiments on three widely-used MMKGC benchmarks: DB15K(Liu et al. 2019), MKG-W(Xu et al. 2022), and MKG-Y(Xu et al. 2022). These datasets provide textual descriptions and entity images of entities. Detailed information about the datasets is presented in Table 2. We follow the standard train,validation and test splits as defined in previous work.

Evaluation Metrics We evaluate our model on the standard link prediction task, and report rank-based evaluation metrics including mean reciprocal rank (MRR) and Hits@K (K=1, 3, 10).

Baselines To conduct a comprehensive comparison, we adopt 14 state-of-the-art models as baselines. These baselines fall into three major categories: (1) Uni-modal KGC models that rely solely on structural information,including TransE (Bordes et al. 2013), RotatE (Sun et al. 2019), DistMult (Yang et al. 2015) and ComplEx (Trouillon et al. 2016); (2) MMKGC models that incorporate additional textual or visual modalities,including IKRL (Xie et al. 2017), RSME (Wang et al. 2021), MYGO (Zhang et al. 2025), NativeE (Zhang et al. 2024a), AdaMF-MAT (Zhang et al. 2024c), SNAG (Chen et al. 2025a), DHNS (Chen et al. 2025b) and APKGC(Jian et al. 2025); (3) MLLM-based Inference Models, including LLaMA2 (Touvron et al. 2023) and LLaVA (Liu et al. 2023). Specifically, we evaluate LLaMA-2-7B, which is prompted using textual description and structural neighbors, and LLaVA-1.5-7B, which further incorporates visual inputs alongside text and KG structure. For each incomplete triple, MLLMs are used to generate a ranked list of 10 candidate entities based on their predicted plausibility.

Implementation Details We implement our model using the PyTorch framework and conduct all experiments on a server with Ubuntu operating system equipped with 2 NVIDIA RTX A6000 GPUs (48GB). We train the model for 250 epochs. The embedding dimension d is selected from $\{128, 256, 512\}$. The number of negative samples per triple is chosen from $\{32, 64, 128\}$. The learning rate is selected from $\{1e^{-5}, 1e^{-4}, 1e^{-3}\}$. The batch size is selected

Model	MKG-W				MKG-Y				DB15K			
	MRR	Hit@10	Hit@3	Hit@1	MRR	Hit@10	Hit@3	Hit@1	MRR	Hit@10	Hit@3	Hit@1
RotatE	32.92	44.23	35.91	26.68	36.67	41.11	35.91	33.92	32.07	51.40	38.57	21.23
TransE	17.01	37.27	27.92	03.35	26.27	32.98	27.92	21.32	21.95	48.09	33.81	05.56
DistMult	18.32	29.50	20.85	12.51	32.37	37.27	20.85	29.60	23.38	41.16	27.97	14.16
ComplEx	16.45	22.48	18.37	13.03	27.57	29.14	18.37	26.53	26.53	36.39	28.41	17.70
IKRL	23.53	33.99	25.35	17.75	26.35	38.11	25.35	20.18	16.22	33.44	19.66	07.21
RSME	33.41	44.31	35.80	27.59	38.03	44.02	35.80	34.28	33.72	50.39	39.23	**24.45**
MYGO	33.40	44.36	35.38	27.52	30.29	42.17	35.38	23.75	32.25	48.23	35.49	24.19
AdaMF-MAT	32.09	45.78	35.54	24.48	37.78	46.11	35.54	32.65	32.08	52.42	40.41	19.81
SNAG	28.08	41.58	31.59	20.37	17.08	28.14	31.59	11.35	26.59	39.81	29.57	19.52
DHNS	23.67	33.51	25.91	18.33	36.96	42.95	25.91	33.64	26.41	41.78	29.91	18.42
APKGC	32.67	46.20	35.83	25.24	31.48	40.68	35.83	26.17	32.91	48.91	37.15	24.39
NativeE	36.79	50.28	40.47	29.36	38.34	46.24	40.47	33.49	33.59	53.77	40.67	22.11
LLama	—	10.20	08.73	05.62	—	02.29	02.03	01.35	—	14.14	11.87	07.15
LLaVa	—	26.42	22.29	15.02	—	03.58	02.80	01.64	—	15.89	13.15	08.32
HFR-MKGC	**38.62**	**51.66**	**42.12**	**31.37**	**39.00**	**47.31**	**42.12**	**34.41**	**35.20**	**56.41**	**43.25**	23.04

Table 1: Comparison with baseline models on three MMKGC datasets (MRR and Hits@K, in %). Best results are highlighted in bold, while the second-best results are underlined."—" means the metric is not computable.

Dataset	#Entity	#Rel	#Train	#Valid	#Test	#Image	#Text
MKG-W	15000	169	34196	4276	4274	14463	14123
MKG-Y	15000	28	21310	2665	2663	14244	12305
DB15K	12842	279	79222	9902	9904	12818	9078

Table 2: Statistics of MMKG datasets.

from $\{128, 256, 1024\}$. The model is optimized using Adam (Kingma and Ba 2015). Since some existing works do not release complete data or code, we reproduce representative baseline models under our unified multi-modal setting to ensure a fair comparison. All models are evaluated using the same visual and textual encodings described in paragraph Modality Encoding, allowing us to isolate the reasoning performance of each method.

Main Results

We present the overall KGC performance of our proposed HFR-MKGC framework in Table 1 compared with 14 strong baselines across three MMKGC datasets. Our proposed method HFR-MKGC achieves the state-of-the-art performance across all three datasets and evaluation metrics (MRR, Hits@1/3/10). Specifically, on dataset MKG-W, HFR-MKGC achieves a MRR of 38.62%, outperforming the best baseline NativeE (36.79%) by +1.83%. It also achieves +2.01%, +1.65%, and +1.38% improvements in Hits@1/3/10, respectively. These consistent improvements highlight the effectiveness of our hierarchical fusion and MLLM-enhanced reasoning strategy. Our fine-tuned MLLMs not only enriches the semantic space beyond conventional embedding-based models, but also guides the model towards more accurate predictions. In contrast, previous methods often suffer from shallow fusion mechanisms or limited inference capacity due to reliance on simple scoring functions and insufficient semantic supervision. By introducing relation-guided hierarchical modal fusion and MLLMs-enhanced knowledge reasoning, HFR-MKGC can better capture multi-modal semantics and discriminate plausible but incorrect entity candidates.

Notably, we also include inference-only results from LLMs (LLaMA and LLaVA), which perform poorly without structured scoring and fusion mechanisms, confirming the necessity of tightly integrating multi-modal structure in MMKGC. In particular, LLaMA does not utilize any visual information, whereas LLaVA incorporates the image modality, leading to consistently better performance for LLaVA compared to LLaMA.

Ablation Study

To verify the contribution of each component, we conduct ablation experiments on MKG-W by removing key modules:

- **w/o RHF:** Replace the relation-guided hierarchical modal fusion with average fusion.
- **w/o MER:** Remove MLLMs-enhanced knowledge reasoning.
- **w/o PTI:** Replace the perturbation for text and image in multi-modal negative sampling with random noise.
- **w/o MNO:** Disable the adversarial training in multi-modal negative sample optimization.

The results are shown in Table 3. We observe that removing any single module leads to noticeable performance degradation. Specifically, the RHF module enables more precise multi-modal representation through relation-guided attention, while MER introduces strong semantic priors via MLLM-based inference. Additionally, PTI and MNO improves robustness by generating harder negative samples, enhancing the model's discrimination capacity. We also evaluate modality importance and remove structural **(w/o S)**, textual **(w/o T)**, visual **(w/o V)** and MLLM-generated visual semantics **(w/o VT)** respectively. Removing the structural modality causes the largest performance drop, highlighting its central role in MMKGC. Removing textual or visual features yields degradation, and eliminating image-derived semantic descriptions further confirms the complementary value of MLLM-generated visual semantics.

Setting	MRR	Hit@10	Hit@3	Hit@1
w/o S	25.67	39.42	28.81	18.02
w/o T	29.67	45.61	34.92	21.25
w/o V	31.16	47.39	36.10	22.09
w/o VT	31.78	47.33	36.41	23.22
w/o RHF	31.98	50.25	39.14	20.80
w/o MER	33.52	50.61	39.70	23.42
w/o PTI	34.59	50.95	40.03	25.12
w/o MNO	34.01	50.98	40.26	24.13
HFR-MKGC	**38.62**	**51.66**	**42.12**	**31.37**

Table 3: Ablation results on dataset MKG-W. Each component in HFR-MKGC contributes significantly.

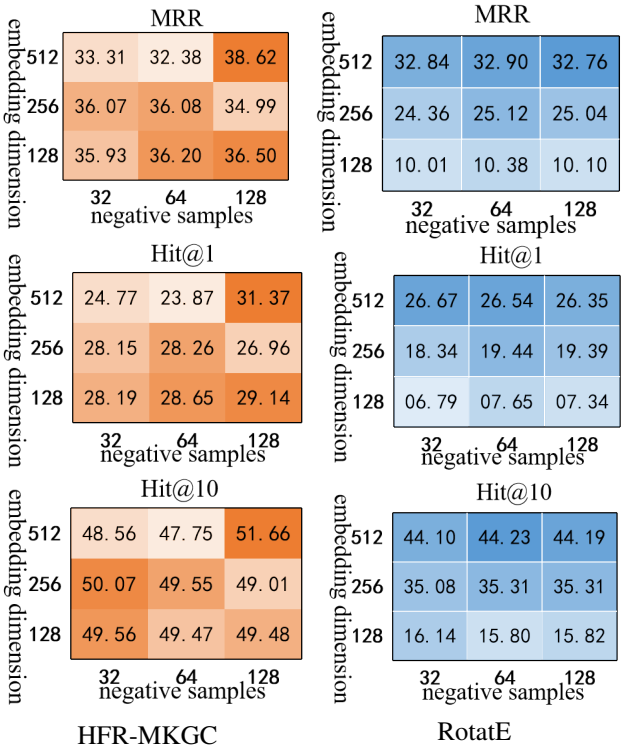

Figure 3: Comparison of HFR-MKGC and RotatE on MKG-W under different embedding dimensions and numbers of negative samples. Darker color indicates better performance.

Hyper-parameters Sensitivity Analysis

We analyze how embedding dimension and the number of negative samples affect model performance on dataset MKG-W. Figure 3 shows results under different combinations for both HFR-MKGC and RotatE. Three key observations emerge: (1) Higher dimensions (e.g., 512) benefit more from larger negative sets (e.g., 128), while medium dimensions (256) perform best with moderate negatives (64). For low dimensions (128), changes in negative samples have little effect. (2) High dimensions need enough negatives. Without enough negatives, 256D may outperform 512D due to better optimization. (3) Metric sensitivity differs. MRR and Hit@1 vary more with hyper-parameters, while Hit@10 remains relatively stable. These results suggest careful tuning of both hyper-parameters is essential in multi-modal models.

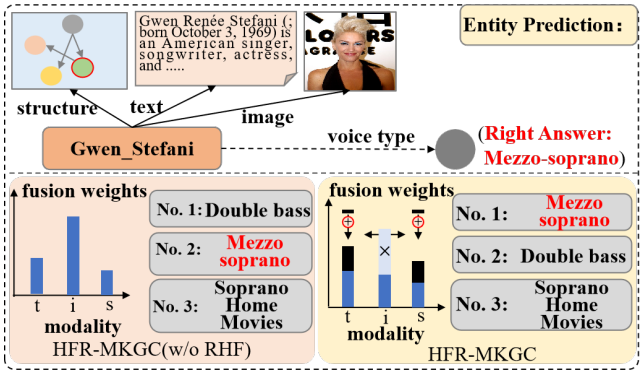

Figure 4: A case example of HFR-MKGC. The bar charts indicate modality weights, and the gray blocks show candidate rankings.

Case Study

To further investigate how our proposed Relation-Guided Hierarchical Fusion (RHF) enhances multi-modal reasoning, we conduct a qualitative case study on dataset MKG-W. As shown in Figure 4, we select a triple with a missing tail entity: *(Gwen_Stefani, voice type, ?)*. We compare the inference results under two settings: (1) HFR-MKGC w/o RHF, where modality fusion is implemented via randomly initialized weights without relation guidance; and (2) our full HFR-MKGC framework that incorporates RHF to adaptively weigh modalities based on relation semantics.

The results show that without RHF, the model fails to prioritize the correct modality cues and ranks the target entity "Mezzo-soprano" only second. In contrast, HFR-MKGC correctly places it at the top. This is because the image of Gwen Stefani contains only appearance information with no visual cues about voice type, while the accompanying text hints at her identity as a singer. RHF therefore assigns higher weight to text and lower weight to image via Eq. 2–3, enabling the accurate "Mezzo-soprano" prediction and validating the effectiveness of our design.

Conclusion

In this paper, we propose HFR-MKGC, a novel framework for MMKGC. Our method introduces a relation-guided hierarchical modal fusion module and leverages fine-tuned MLLMs for instruction-based triple inference. Furthermore, we construct challenging negative samples via textual disturbance and visual perturbation, and jointly optimize the model using adversarial training. Extensive experiments on multiple benchmarks validate the effectiveness and superiority of our approach in capturing multi-modal semantics and improving reasoning performance. In future work, we aim to explore how to more precisely utilize MLLMs for multi-modal knowledge representation and reasoning.

Acknowledgments

This work was supported by National Key Research and Development Program of China (2023YFF0725103), National Natural Science Foundation of China (62192784, U22B2038, 62422202, 62272054), Beijing Nova Program (No. 20230484319).

References

Barile, R.; d'Amato, C.; and Fanizzi, N. 2025. LP-DIXIT: Evaluating Explanations for Link Predictions on Knowledge Graphs using Large Language Models. In *Proceedings of the ACM on Web Conference 2025*, 4034–4042.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Chen, X.; Zhang, N.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; and Chen, H. 2022. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 904–915.

Chen, Z.; Fang, Y.; Zhang, Y.; Guo, L.; Chen, J.; Pan, J. Z.; Chen, H.; and Zhang, W. 2025a. Noise-powered Multi-modal Knowledge Graph Representation Framework. In *Proceedings of the 31st International Conference on Computational Linguistics*, 141–155.

Chen, Z.; Fang, Y.; Zhang, Y.; Guo, L.; Chen, J.; Pan, J. Z.; Chen, H.; and Zhang, W. 2025b. Noise-powered Multi-modal Knowledge Graph Representation Framework. In *Proceedings of the 30th International Conference on Database Systems for Advanced Applications*.

Cui, Z.; Weng, Y.; Tang, X.; Lyu, F.; Liu, D.; He, X.; and Ma, C. 2025. Comprehending knowledge graphs with large language models for recommender systems. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1229–1239.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186.

Fang, Y.; Fan, D.; Zha, D.; and Tan, Q. 2024. Gaugllm: Improving graph contrastive learning for text-attributed graphs with large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 747–758.

Gao, Y.; Zhang, F.; Zhang, Z.; Min, X.; and Zhuang, F. 2025. Mixed-Curvature Multi-Modal Knowledge Graph Completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11699–11707.

Jian, Y.; Luo, X.; Li, Z.; Zhang, M.; Zhang, Y.; Xiao, K.; and Hou, X. 2025. APKGC: Noise-enhanced Multi-Modal Knowledge Graph Completion with Attention Penalty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 15005–15013.

Jiang, X.; Shen, Y.; Shi, Z.; Xu, C.; Li, W.; Li, Z.; Guo, J.; Shen, H.; and Wang, Y. 2024. Unlocking the Power of Large Language Models for Entity Alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7566–7583.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Li, A.; Li, Y.; Shao, Y.; and Liu, B. 2023. Multi-view scholar clustering with dynamic interest tracking. *IEEE Transactions on Knowledge and Data Engineering*, 35(9): 9671–9684.

Li, M.; Yang, C.; Xu, C.; Song, Z.; Jiang, X.; Guo, J.; Leung, H.-f.; and King, I. 2025. Context-aware inductive knowledge graph completion with latent type constraints and subgraph reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 12102–12111.

Li, Q.; Chen, Z.; Ji, C.; Jiang, S.; and Li, J. 2024. LLM-based multi-level knowledge generation for few-shot knowledge graph completion. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, volume 3, 2135–2143.

Li, Y.; Yu, Z.; and He, D. 2024. Text-Rich Graph Neural Networks With Subjective-Objective Semantic Modeling. *IEEE Transactions on Knowledge and Data Engineering*, 36(9): 4956–4967.

Liang, M.; Du, J.; Liang, Z.; Xing, Y.; Huang, W.; and Xue, Z. 2024a. Self-supervised multi-modal knowledge graph contrastive hashing for cross-modal search. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 13744–13753.

Liang, M.; Li, Y.; Yu, Y.; Cao, X.; Xue, Z.; Li, A.; and Lu, K. 2024b. Structures aware fine-grained contrastive adversarial hashing for cross-media retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 36(7): 3514–3528.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Q.; He, Y.; Xu, T.; Lian, D.; Liu, C.; Zheng, Z.; and Chen, E. 2024. Unimel: A unified framework for multimodal entity linking with large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 1909–1919.

Liu, Y.; Li, H.; Garcia-Duran, A.; Niepert, M.; Onoro-Rubio, D.; and Rosenblum, D. S. 2019. MMKG: multi-modal knowledge graphs. In *European Semantic Web Conference*, 459–474. Springer.

Mousselly-Sergieh, H.; Botschen, T.; Gurevych, I.; and Roth, S. 2018. A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the seventh joint conference on lexical and computational semantics*, 225–234.

Pan, Y.; Hong, J.; Zhao, T.; Song, L.; Liu, J.; and Shang, X. 2025. Logic-Aware Knowledge Graph Reasoning for Structural Sparsity under Large Language Model Supervision. In

Proceedings of the ACM on Web Conference 2025, 4531–4542.

Panda, P.; Agarwal, A.; Devaguptapu, C.; Kaul, M.; and Ap, P. 2024. HOLMES: Hyper-Relational Knowledge Graphs for Multi-hop Question Answering using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13263–13282.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Sun, Z.; Deng, Z.-H.; Nie, J.-Y.; and Tang, J. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *International Conference on Learning Representations*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; and Bouchard, G. 2016. Complex embeddings for simple link prediction. In *International conference on machine learning*, 2071–2080. PMLR.

Wang, J.; Li, Y.; Shao, Y.; Xue, Z.; Guan, Z.; Li, A.; and Ye, G. 2025. Reinforcement Active Client Selection for Federated Heterogeneous Graph Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21117–21125.

Wang, M.; Wang, S.; Yang, H.; Zhang, Z.; Chen, X.; and Qi, G. 2021. Is visual context really helpful for knowledge graph? A representation learning perspective. In *Proceedings of the 29th ACM international conference on multimedia*, 2735–2743.

Wang, Z.; Li, L.; Li, Q.; and Zeng, D. 2019. Multimodal data enhanced representation learning for knowledge graphs. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Xiao, S.; Shao, Y.; Li, Y.; Yin, H.; Shen, Y.; and Cui, B. 2022. LECF: recommendation via learnable edge collaborative filtering. *Science China Information Sciences*, 65(1): 112101.

Xie, R.; Liu, Z.; Jia, J.; Luan, H.; and Sun, M. 2016. Representation learning of knowledge graphs with entity descriptions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Xie, R.; Liu, Z.; Luan, H.; and Sun, M. 2017. Image-embodied Knowledge Representation Learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 3140–3146. International Joint Conferences on Artificial Intelligence Organization.

Xu, D.; Xu, T.; Wu, S.; Zhou, J.; and Chen, E. 2022. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *Proceedings of the 30th ACM international conference on multimedia*, 3857–3866.

Yang, B.; Yih, S. W.-t.; He, X.; Gao, J.; and Deng, L. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*.

Yao, L.; Peng, J.; Mao, C.; and Luo, Y. 2025. Exploring large language models for knowledge graph completion. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zhang, R.; Su, Y.; Trisedya, B. D.; Zhao, X.; Yang, M.; Cheng, H.; and Qi, J. 2023. Autoalign: Fully automatic and effective knowledge graph alignment enabled by large language models. *IEEE Transactions on Knowledge and Data Engineering*, 36(6): 2357–2371.

Zhang, Y.; Chen, Z.; Guo, L.; Xu, Y.; Hu, B.; Liu, Z.; Zhang, W.; and Chen, H. 2024a. Native: Multi-modal knowledge graph completion in the wild. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, 91–101.

Zhang, Y.; Chen, Z.; Guo, L.; Xu, Y.; Hu, B.; Liu, Z.; Zhang, W.; and Chen, H. 2025. Tokenization, fusion, and augmentation: Towards fine-grained multi-modal entity representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 13322–13330.

Zhang, Y.; Chen, Z.; Guo, L.; Xu, Y.; Zhang, W.; and Chen, H. 2024b. Making large language models perform better in knowledge graph completion. In *Proceedings of the 32nd ACM international conference on multimedia*, 233–242.

Zhang, Y.; Chen, Z.; Liang, L.; Chen, H.; and Zhang, W. 2024c. Unleashing the Power of Imbalanced Modality Information for Multi-modal Knowledge Graph Completion. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 17120–17130.

Zhang, Y.; Chen, Z.; and Zhang, W. 2023. MACO: A modality adversarial and contrastive framework for modality-missing multi-modal knowledge graph completion. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 123–134. Springer.

Zhang, Z.; Wang, X.; Zhang, Z.; Li, H.; Qin, Y.; and Zhu, W. 2024d. LLM4DyG: can large language models solve spatial-temporal problems on dynamic graphs? In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4350–4361.

Zhao, Y.; Yu, J.; Cheng, Y.; Yu, C.; Liu, Y.; Li, X.; and Wang, S. 2025. Variational Graph Autoencoder for Heterogeneous Information Networks with Missing and Inaccurate Attributes. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 2067–2078.