LBMKGC: Large Model-Driven Balanced Multimodal Knowledge Graph Completion

Yuan Guo
Dalian Maritime University
guoynow@gmail.com

Qian Ma*
Dalian Maritime University
maqian@dlmu.edu.cn

Hui Li
Dalian Maritime University
li_hui@dlmu.edu.cn

Qiao Ning
Jiangnan University
ningq669@jiangnan.edu.cn

Furui Zhan
Dalian Maritime University
izfree@dlmu.edu.cn

Yu Gu
Northeastern University, China
guyu@mail.neu.edu.cn

Ge Yu
Northeastern University, China
yuge@mail.neu.edu.cn

Shikai Guo*
Dalian Maritime University
shikai.guo@dlmu.edu.cn

Abstract

Multi-modal Knowledge Graph Completion (MMKGC) aims to predict missing entities, relations, or attributes in knowledge graphs by collaboratively modeling the triple structure and multimodal information (e.g., text, images, videos) associated with entities. This approach facilitates the automatic discovery of previously unobserved factual knowledge. However, existing MMKGC methods encounter several critical challenges: (i) the imbalance of inter-entity information across different modalities; (ii) the heterogeneity of intra-entity multimodal information; and (iii) for a given entity, the informational contributions of different modalities are inconsistent across contexts. In this paper, we propose a novel Large model-driven Balanced Multimodal Knowledge Graph Completion framework, termed LBMKGC. Initially, LBMKGC employs the Stable Diffusion XL (a Large generative vision model) to augment the imbalanced information across modalities. Subsequently, to bridge the semantic gap between heterogeneous modalities, LBMKGC aligns the multimodal embeddings of entities semantically by using the CLIP (Contrastive Language-Image Pre-Training) model. Furthermore, LBMKGC adaptively fuses multimodal embeddings with relational guidance by distinguishing between the perceptual and conceptual attributes of triples. Finally, extensive experiments conducted against 21 state-of-the-art baselines demonstrate that LBMKGC achieves superior performance across diverse datasets and scenarios while maintaining efficiency and generalizability. Our code and data are publicly available at: https://github.com/guoynow/LBMKGC.

1 Introduction

Multimodal Knowledge Graphs (MMKGs) [14] represent an advanced semantic framework that extends traditional knowledge graphs. Unlike conventional knowledge graphs, which primarily consist of structured entity-relation-attribute triples, MMKGs integrate a variety of multimodal data associated with entities, including textual descriptions, images, audio, and videos. This comprehensive approach allows MMKGs to capture and describe complex real-world phenomena across multiple

*Corresponding authors

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

dimensions. By fusing structural and multimodal information, MMKGs facilitate finer-grained reasoning in applications such as recommender systems [29], natural language processing [6], and computer vision [8]. Current mainstream MMKGC [31] methods can be categorized into two primary approaches. The first focuses on multimodal fusion mechanisms, aiming to achieve completion by systematically modeling the cross-modal correlations among structural topologies, visual representations, and textual semantics [24, 1, 35]. The second approach emphasizes negative sampling optimization algorithms, which enhance completion performance by improving the quality of negative samples through vision-text joint discrimination [2, 36, 34]. However, existing methods often overlook following critical challenges.

(1) **The imbalance of inter-entity information across different modalities.** In real-world MMKGs, modality distributions are diverse and complex. Some entities may possess rich multimodal information (e.g., images and text), while others suffer from modality scarcity (e.g., text-only entities with missing visual data). This imbalanced distribution complicates knowledge graph completion tasks, as missing modalities may lead to critical knowledge omissions, thereby compromising accuracy. Although existing MMKGC methods [38] employ frameworks like adversarial training to mitigate imbalance issues, they fail to explicitly model latent semantics in modality-missing scenarios.

(2)**The heterogeneity of intra-entity multimodal information.** Significant disparities exist in data structures and semantic comprehension across modalities. For instance, visual information is encoded as 2D pixel matrices in images, while textual modalities convey symbolic logic through word sequences. This dual gap in both data representation and semantic parsing mechanisms poses substantial challenges for cross-modal alignment. Existing MMKGC methods predominantly rely on single-modality pretrained models (e.g., VGG [21] for vision, BERT [5] for text) to independently extract visual or textual features, thereby neglecting the construction of a cohesive cross-modal representation space. Consequently, downstream tasks often struggle to fully capture the deep semantic interdependencies amongst modalities.

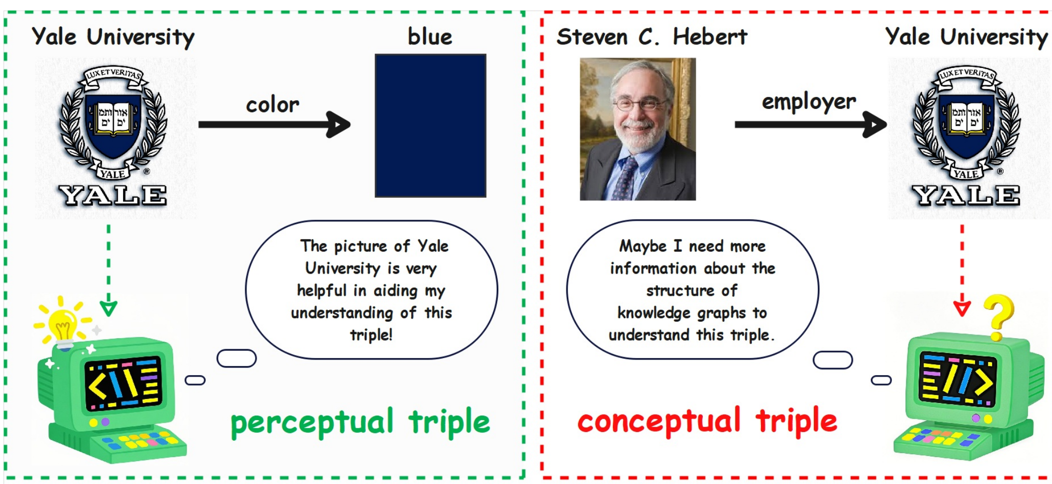


Figure 1: The salience of information carried by different modalities varies across different types of triples.

(3) **The inconsistency in the volume of information represented by various modalities for a given entity.** In knowledge graphs, triples can be categorized into perceptual (concrete) and conceptual (abstract) types. For the same entity, the salience of information carried by different modalities varies across these two types of triples. Taking the entity "Yale University" as an example, in the MKG-W dataset, the triple (Yale University, color, blue) is a perceptual triple. In this triple, the visual modality of Yale University provides discriminative evidence, and thus should be assigned higher weight during the feature fusion stage. However, the triple (Steven C. Hebert, employer, Yale University) is a conceptual triple, as it involves an employer relationship, which is a social construct and cannot be directly perceived. Therefore, Yale University in this context relies more on the structural information of the knowledge graph, and higher weight should be assigned to structural information during feature fusion. Similarly, for entities, "blue" is perceptual, where the visual modality provides discriminative evidence, while "law" is conceptual, relying more on symbolic reasoning from textual modalities. Existing MMKGC methods struggle to dynamically adjust inter-modal weights for entities based on

the perceptual or conceptual attributes of triples, leading to inefficient and imprecise utilization of multimodal information.

To address the aforementioned challenges, we propose a novel Large model-driven Balanced MMKGC framework, named LBMKGC. LBMKGC comprises three key modules called **Large Generative Visual Model-based Modality Completion** (LvMC) module, **Cross-Modality Alignment** (CMoA) module, and **Context-Guided Adaptive Fusion** (CGuAF) module, respectively. To address the imbalance issue, the LvMC module utilizes the large generative visual model SDXL (Stable Diffusion XL) [16] to augment the missing modality information. To address the heterogeneity issue, the CMoA module focuses on aligning the entities between various modalities by employing the pre-trained CLIP model. To address the inconsistency issue, the CGuAF module adaptively moderates the weight of each modality with relational guidance to learn joint cross-modality representations. Our major contributions are summarized as follows.

- We propose a novel MMKGC framework, named LBMKGC, designed to address issues of modality imbalance, modality heterogeneity, and the inconsistency of information quantity for entities across different modalities.
- We innovatively tackle the modality imbalance issue in MMKGC by leveraging the capabilities of the SDXL, a state-of-the-art large generative visual model, instead of relying on conventional adversarial training methods.
- We propose a CMoA module that dynamically adjusts the alignment of multimodal embeddings, which effectively mitigates biases arising from data heterogeneity, ensuring sufficient utilization of multimodal information.
- We first consider the impact of the perceptual and conceptual attributes of triples during joint cross-modality representation learning, and accordingly propose a novel context-guided adaptive fusion module.
- To evaluate the performance of LBMKGC, we conduct comprehensive experiments and further explorations on three public benchmarks. Empirical results demonstrate that LBMKGC outperforms 21 recent baselines, achieving new state-of-the-art MMKGC results.

2 Preliminaries

2.1 Problem Statement

MMKGs [14] provide a robust foundation for knowledge representation and reasoning by integrating structured triples with rich multimodal entity attributes (e.g., images, text, audio, video). In recent years, significant progress has been made in MMKGC [31], which aims to automatically discover new knowledge from incomplete MMKGs.

We define a MMKGs as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{M})$, where $\mathcal{E}$ is the set of entities, $\mathcal{R}$ is the set of relations, $\mathcal{T} = \{(h, r, t) | h, t \in \mathcal{E}, r \in \mathcal{R}\}$ is the set of triples, and $\mathcal{M}$ is the set of modalities, encapsulating different modalities in MMKGs (such as images, text, videos, audio, etc.). For an entity $e \in \mathcal{E}$ in a modality $m \in \mathcal{M}$, its modality information is denoted as e_m, such as text modality denoted as e_t, and visual modality denoted as e_v.

The MMKGC model embeds entities and relations into a continuous vector space. For each entity $e \in \mathcal{E}$, we define its structure, text, and visual embeddings as $e_s^{\text{emb}}, e_t^{\text{emb}}, e_v^{\text{emb}}$. For each relation $r \in \mathcal{R}$, we define its structure embedding as r_s^{emb}. Additionally, the MMKGC model can use a scoring function $\mathcal{F}$ to measure the plausibility of each triple (h, r, t). $\mathcal{F}$ calculates the score through the embedding representation of triples. In the inference phase of the MMKGC task, for a given query $(h, r, ?)$ or $(?, r, t)$, the model scores each candidate entity corresponding to the triple and ranks them to make predictions.

2.2 Related Work

In recent years, numerous KGC methods have been developed. Traditional unimodal approaches, e.g., TransE [1], TransD [10], RotatE [24], PairRE [4], DistMult [35] and ComplEx [26], learn knowledge graph embeddings (KGE) by projecting entities and relations into a continuous vector space that preserves the observed triple structure. To capture heterogeneous graph topology, DGS [9] first separately embeds the cyclic instance structure and hierarchical ontology structure of two-view knowledge graphs into spherical and hyperbolic spaces, unifying them via a bridge space.

On the multi-modal front, IKRL [33] first augments TransE with visual features; TBKGC [19] further incorporates textual cues. Building on KGE, MMKGC methods such as OTKGE [3], VISTA [11], RSME [28], TransAE [32], IMF [13] and QBE [30] leverage rich multi-modal signals to refine entity representations. MMRNS [34] and MANS [36] instead focus on exploiting modalities for higher-quality negative sampling. Modality-imbalance has also been tackled by generative adversarial schemes like MMKRL [15] and AdaMF-MAT [38], which synthesize missing-modal features and use RotatE-style scores as discriminators. Yet these models (i) capture only local semantics and ignore global cross-modal correlations, and (ii) require separate adversarial networks per modality, aggravating heterogeneity. Consequently, modality imbalance, structural heterogeneity and inconsistent saliency across modalities remain largely unresolved.

In this work, we propose LBMKGC to tackle these issues and achieve state-of-the-art performance.

3 Method

In this section, we will introduce our MMKGC framework, LBMKGC. An overview of LBMKGC is shown in the Figure 2. We will introduce the main components in the following paragraphs. These components are the Large Generative Visual Model-based Modality Completion module, Cross-Modal Alignment module, Context-Guided Adaptive Fusion module, and MMKGC Module.

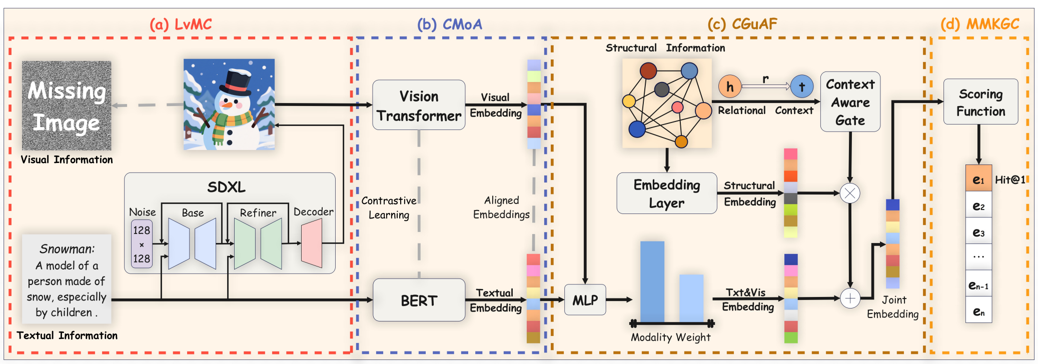


Figure 2: Architectural Overview of the LBMKGC Modal.

3.1 Large Generative Visual Model-based Modality Completion

To address the issue of imbalanced multimodal information, we propose a modality completion module based on SDXL (Stable Diffusion XL) [16] as shown in the Figure 2 (a). In the field of knowledge graph, obtaining visual information is usually more difficult than obtaining text information. Therefore, this paper focuses on the problem of multimodal information missing in visual modality, that is, given an entity $e \in \mathcal{E}$, LvMC populates the prompt template with e_t to generate e_v.

The core component of LvMC is the generative model SDXL. SDXL employs a two-stage generative process facilitated by two specialized models: a Base model and a Refiner model. The Base model is primarily responsible for generating a coherent initial image based on the text prompt, ensuring overall semantic accuracy and composition. Subsequently, the Refiner model operates on the output (typically in the latent space) of the Base model to enhance fine-grained details and visual fidelity. For clarity, a simplified representation of the overall generation process is provided as follows:

First, LvMC generates a text embedding vector by combining two CLIP encoders:

$$\mathbf{c}_t = \mathbf{W}_1 \cdot \text{CLIP}_{\text{bigG}}(\mathbf{e}_t) + \mathbf{W}_2 \cdot \text{CLIP}_{\text{ViT-L}}(\mathbf{e}_t), \tag{1}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are projection matrices, $\text{CLIP}_{\text{bigG}}$ and $\text{CLIP}_{\text{ViT-L}}$ are respectively OpenCLIP ViT-bigG [18] and OpenAI CLIP ViT-L [17] encoders.

The image resolution (w, h) is embedded as Fourier features:

$$\mathbf{f}_{\text{size}} = [\sin(2^k \pi n_w), \cos(2^k \pi n_w), \sin(2^k \pi n_h), \cos(2^k \pi n_h)]_{k=0}^{L-1}, \tag{2}$$

where $n_w = \frac{w}{1024}$, $n_h = \frac{h}{1024}$, and L is the number of frequency bands, and the final combined embedding vector is: $\mathbf{c} = \mathbf{c}_t \oplus \mathbf{f}_{\text{size}}$.

The Base model generates the initial resolution potential through DDIM sampling [22]:

$$\mathbf{z}_{t-1}^{\text{base}} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{z}_t^{\text{base}} - \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_\theta(\mathbf{z}_t^{\text{base}}, t, \mathbf{c})}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}} \cdot \epsilon_\theta(\mathbf{z}_t^{\text{base}}, t, \mathbf{c}), \tag{3}$$

where the initial noise sampled from a standard Gaussian distribution: $\mathbf{z}_T^{\text{base}} \sim \mathcal{N}(0, \mathbf{I})$. Then, at each reverse step t, the previous state $\mathbf{z}_{t-1}^{\text{base}}$ is computed by first estimating the predicted noise $\epsilon_\theta(\mathbf{z}_t^{\text{base}}, t, \mathbf{c})$ using its UNet, and then applying the DDIM update rule based on the cumulative noise schedule $\bar{\alpha}_t$ and the conditioning embedding $\mathbf{c}$, which combines text and size information.

The Refiner model adds noise to $\mathbf{z}_0^{\text{base}}$:

$$\mathbf{z}_{t_{\text{refine}}}^{\text{refine}} = \sqrt{\bar{\alpha}_{t_{\text{refine}}}} \mathbf{z}_0^{\text{base}} + \sqrt{1-\bar{\alpha}_{t_{\text{refine}}}} \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{4}$$

where t_{refine} denotes the running timestep variable in the Refiner's reverse diffusion process.

The Refiner model refines the image step by step:

$$\begin{aligned} \mathbf{z}_{t-1}^{\text{refined}} =& \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{z}_t^{\text{refined}} - \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_\phi(\mathbf{z}_t^{\text{refined}}, t, \mathbf{c}, \mathbf{z}_0^{\text{base}})}{\sqrt{\bar{\alpha}_t}} \right) \\ &+ \sqrt{1-\bar{\alpha}_{t-1}} \cdot \epsilon_\phi(\mathbf{z}_t^{\text{refined}}, t, \mathbf{c}, \mathbf{z}_0^{\text{base}}), \end{aligned} \tag{5}$$

where ϵ_ϕ is the noise predictor of the Refiner model.

Finally, the high-resolution image in the variable $\mathbf{z}_0^{\text{refined}}$ is decoded through the VAE decoder to the real image space:

$$\mathbf{e}_v = \text{Decoder}_{\text{VAE}}(\mathbf{z}_0^{\text{refined}}). \tag{6}$$

3.2 Cross-Modal Alignment

To address the issue of heterogeneity in multimodal information within a single entity, we proposed a Cross-Modal Alignment (CMoA) module, as shown in the Figure 2 (b). This module leverages the pre-trained CLIP [17] model to achieve semantic alignment between an entity's image and text representations.

We first extract multimodal feature encodings from the original image and text description of the entity using a pre-trained model, which is a necessary step for all MMKGC methods. The text modality is encoded into text features $f_t = \{T_1, T_2, ..., T_N\}$ using BERT [5], and the visual modality is encoded into visual features $f_v = \{V_1, V_2, ..., V_N\}$ using ViT [7]. The cosine similarity matrix S between the two is calculated, where $S_{i,j}$ represents the similarity between the i-th visual feature V_i and the j-th text feature T_j:

$$S_{i,j} = \frac{V_i \cdot T_j}{\|V_i\| \cdot \|T_j\|}. \tag{7}$$

The symmetric contrastive loss function with a learnable temperature parameter τ is defined as:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2} \left(\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{S_{i,i}/\tau}}{\sum_{j=1}^{N} e^{S_{i,j}/\tau}} + \frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{S_{i,i}/\tau}}{\sum_{j=1}^{N} e^{S_{j,i}/\tau}} \right). \tag{8}$$

Based on the cross-modality feature alignment, we obtain the aligned visual features f_v and text features f_t. Subsequently, through a projection layer, each entity $e \in \mathcal{E}$ in modality $m \in \mathcal{M}$ is embedded into a representation $e_m^{\text{emb}} = \mathcal{P}_m(f_m)$. Here $\mathcal{P}_m$ is a projection layer for modality m, aimed at projecting different modality embeddings into the same vector space. Each $\mathcal{P}_m$ consists of two MLP layers with ReLU as the activation function.

3.3 Context-Guided Adaptive Fusion

To address the issue that existing MMKGC methods struggle to adaptively adjust the weights of different modalities based on the specific characteristics of the three-element entities, leading to the inability to effectively utilize different modalities, we propose a Context-Guided Adaptive Fusion (CGuAF) module, as shown in the Figure 2 (c). This mechanism is divided into two stages. The first stage performs adaptive fusion of multiple modalities, and the second stage considers the upper and lower context information to generate a joint embedding representation for each entity in the three-element set.

The mechanism adjusts the weight ω_m of each modality based on the entity's (perception or concept) characteristics. For perceptual entities (such as color), if the visual modality can provide strong evidence, it is assigned a higher weight, and the text modality is assigned a lower weight. Conversely, for conceptual entities (such as force), it is difficult to express them with images, so a higher weight is assigned to the text modality, and a lower weight is assigned to the visual modality. Specifically, the weight of the modality m for entity e, $\omega_m(e)$, is calculated as follows:

$$\omega_m(e) = \frac{\exp\left(V \odot \tanh\left(e_m\right)\right)}{\sum_{n \in \mathcal{M}} \exp\left(V \odot \tanh\left(e_n\right)\right)}, \tag{9}$$

where V is a learnable vector and $\odot$ is the point-wise operator.

Based on these weights, we obtain the fused multimodal information h_{mm}:

$$h_{\text{mm}}^{\text{emb}} = \sum_{m \in \mathcal{M}} \omega_m e_m^{\text{emb}}. \tag{10}$$

For the same entity, the salience of information carried by different modalities varies across different types of triples (e.g., perceptual and conceptual). Accordingly, we design a relational temperature ζ_r that allows an entity to adapt its joint embedding based on the type of triple, rather than using a static one, thus providing a more effective relational context.

$$h_{\text{joint}}^{\text{emb}} = \zeta_r h_s^{emb} + (1 - \zeta_r) h_{\text{mm}}^{\text{emb}}. \tag{11}$$

Similarly, the joint embedding of the tail entity is derived using the same fusion method.

Traditional Multimodal Knowledge Graph Completion (MMKGC) approaches typically rely on static, entity-level modality fusion strategies, which overlook the specific context of the input triple. In comparison, our proposed context-guided adaptive fusion module dynamically recalibrates entity embeddings conditioned on the triple type. This fine-grained adaptive adjustment mechanism significantly enhances the model's ability to interact with multimodal features in different relationship contexts.

3.4 Multimodal Knowledge Graph Completion

As shown in the Figure 2 (d), based on the multimodal joint embedding representation of entities, we introduce the RotatE [24] score function based on complex space rotation operations to evaluate the rationality of triples (h, r, t), which has universal modeling capability for complex relationship patterns in knowledge graphs. Specifically, for a triple (h, r, t), its plausibility score is defined as the negative distance between the head entity embedding after rotation and the tail entity embedding, that is:

$$\mathcal{F}(h, r, t) = -||h_{\text{joint}}^{\text{emb}} \circ r - t_{\text{joint}}^{\text{emb}}||, \tag{12}$$

where $\circ$ is the rotation operation in complex space, and a higher score indicates greater rationality of the triple. In the training process, we use a loss function based on negative sampling to optimize the parameters, which can be expressed as:

$$\mathcal{L}_{\text{kgc}} = \sum_{(h,r,t) \in \mathcal{T}} -\log \sigma(\gamma + \mathcal{F}(h, r, t)) - \sum_{i=1}^{K} p(h_i', r_i', t_i') \log \sigma(-\mathcal{F}(h_i', r_i', t_i') - \gamma), \tag{13}$$

where σ is the sigmoid function; γ is a fixed boundary; $(h_i', r_i', t_i') \in \mathcal{T}'$, $(i = 1, 2, ..., K)$ are K negative samples of the triple (h, r, t). Additionally, $p(h_i', r_i', t_i')$ is the self-adversarial weight proposed in RotatE [24].

4 Experiments

In this section, we evaluate our method using a mainstream task in KGC - link prediction task. We first introduce the experimental setup and then present the results. We mainly explore the following four research questions (RQs) regarding LBMKGC:

- **RQ1:** Can our model LBMKGC surpass the existing baselines and make substantial progress in the MMKGC task in link prediction?
- **RQ2:** How much does the design of each module in LBMKGC contribute to performance?
- **RQ3:** When the multimodal information is imbalanced, can LBMKGC maintain stable performance in the MMKGC task?
- **RQ4:** Can LBMKGC adaptively adjust multimodal weights based on different entities?

4.1 Experimental Settings

Datesets. To better explore the MMKGC task in a diverse and complex environment, we conducted comprehensive experiments and further explorations on three public benchmarks. The DB15K [14] dataset was built by crawling search engine images and aligning them with DBpedia [12], while the MKG-W and MKG-Y [34] datasets were created by extracting subsets from Wikidata [27] and YAGO [23]. We used the pre-trained model of CLIP to extract multimodal features.

Baselines. To demonstrate the effectiveness of our method, we conducted a comprehensive comparison and analysis with 21 different state-of-the-art KGC and MMKGC models as our baselines. They can be categorized into three types:

(i) **Uni-modal KGC methods**, this paper employs 6 of the most advanced uni-modal KGC methods, including TransE [1], DistMult [35], ComplEx [26], RotatE [24], PairRE [4], GC-OTE [25], which designed elegant scoring functions and considered structural information in model design.

(ii) **Multi-modal KGC models**, this paper utilizes 12 MMKGC methods that consider multi-modal information and triple structural information, including IKRL [33], TBKGC [19], TransAE [32], MMKRL [15], RSME [28], OTKGE [3], IMF [13], AdamF [38], QEB [30], VISTA [11], AdamF-MAT [38] and MyGO [37].

(iii) **Negative sampling methods**, this paper compares our method with 3 different negative sampling methods, including KBGAN [2], MANS [36], MMRNS [34]. Among them, KBGAN is an adversarial sampling method designed for traditional KGC, applying reinforcement learning to optimize the model. MMRNS utilizes multi-modal information to enhance the negative sampling process.

Metrics. To evaluate our method, we conducted link prediction tasks on three datasets. Link prediction is an important task in knowledge graph completion, aiming to predict the missing entity for a given query $(h, r, ?)$ or $(?, r, t)$. The link prediction task consists of two parts: head prediction and tail prediction.

Following previous work, we use ranking-based metrics such as Mean Reciprocal Rank (MRR) and Hit@K (K = 1, 3, 10) to evaluate the results. MRR and Hit@K can be calculated as:

$$\text{MRR} = \frac{1}{|\mathcal{T}_{\text{test}}|} \sum_{i=1}^{|\mathcal{T}_{\text{test}}|} \left(\frac{1}{r_{h,i}} + \frac{1}{r_{t,i}} \right), \tag{14}$$

$$\text{Hit@K} = \frac{1}{|\mathcal{T}_{\text{test}}|} \sum_{i=1}^{|\mathcal{T}_{\text{test}}|} \left(1(r_{h,i} \leq K) + 1(r_{t,i} \leq K) \right), \tag{15}$$

where $r_{h,i}$ and $r_{t,i}$ are the results of head prediction and tail prediction, respectively.

To ensure fair comparisons, the filter setting [1] is applied to all results to remove candidate triples that already exist in the training set.

Implementation details. We implemented our LBMKGC model based on the famous open-source KGC library OpenKE. We conducted experiments on a Linux server with Ubuntu 24.04.01 operating system and a single NVIDIA GeForce 4090 GPU. We reproduced some advanced models, and some baseline results refer to MyGO [37].

In the LBMKGC, we fix the batch size to 512 and set the training epoch from {1000, 1250, 1500}. The embedding dimensions are tuned from {300, 400, 500} and the negative sampling number K is tuned from {32, 64, 128}. The margin γ is tuned from {8, 12, 16, 20, 24} and the temperature β is set to 4. We optimize the model with Adam and the learning rate is tuned from {1e-5, 2e-5, 5e-5}.

For baselines, we reproduce the results following the methodology and parameter setting described in the original papers and their open-source official code.

4.2 Main Results (RQ1)

We conducted link prediction experiments and presented the experimental results in Table 1. Our method, LBMKGC, achieved superior performance on all evaluation metrics across three public benchmarks. The experimental results of this method show that our approach enables the model to capture a richer set of candidate relation patterns, enhancing semantic coverage, and more stably retaining correct answers in the top region of the candidate set (Hit@3 and Hit@10), while still maintaining advanced performance in the exact first match metric (Hit@1). It has stronger practicality in real-world scenarios that require multi-candidate selection (such as recommendation systems, open-domain question answering).

4.3 Ablation Study (RQ2)

To provide a more detailed demonstration of the effectiveness of each module design and to answer RQ3, we conducted ablation studies, as shown in Table 2. We replaced the adaptive fusion guided by the context with mean fusion to verify the effectiveness of CGuAF. Furthermore, to verify the contribution of different information in MMKGC, we covered visual information (textual information/structural information) for link prediction tasks. Additionally, we used uni-modal pre-training models (VGG16 [21], BERT [5]) to independently extract visual and textual features to verify the effectiveness of cross-modal feature alignment. Finally, we randomly initialized the missing modal information to verify the importance of the real missing modal information in explicit modeling.

The results of the ablation study show that when each module or each modal information of our method is removed, the link prediction results will decrease, which demonstrates their effectiveness.

Table 1: Knowledge Graph Completion Models Comparison

Model	**DB15K**				**MKG-W**				**MKG-Y**			
	MRR	**Hit@1**	**Hit@3**	**Hit@10**	**MRR**	**Hit@1**	**Hit@3**	**Hit@10**	**MRR**	**Hit@1**	**Hit@3**	**Hit@10**
TransE	24.86	12.78	31.48	47.07	29.19	21.06	33.20	44.23	30.73	23.45	35.18	43.37
DistMult	23.03	14.78	26.28	39.59	20.99	15.93	22.28	30.86	25.04	19.33	27.80	35.95
ComplEx	27.48	18.37	31.57	45.37	24.93	19.06	26.69	36.73	28.71	22.26	32.12	40.93
RotatE	29.28	17.87	36.12	49.66	33.67	26.80	36.68	46.73	34.95	29.10	38.35	45.30
PairRE	31.13	21.62	35.91	49.30	34.40	28.24	36.71	46.04	32.01	25.53	35.84	43.89
GC-OTE	31.85	22.11	36.52	51.18	33.92	26.55	39.65	46.90	32.95	26.77	38.41	44.08
IKRL	26.82	14.09	34.93	49.09	32.36	26.11	34.75	44.07	33.22	30.37	34.28	38.26
TBKGC	28.40	15.61	37.03	49.86	31.48	25.31	33.98	43.24	33.99	30.47	35.27	40.07
TransAE	28.09	21.25	31.17	41.17	30.00	21.23	34.91	44.72	28.10	25.31	29.10	33.03
MMKRL	26.81	13.85	35.07	49.39	30.10	22.16	34.09	44.69	36.81	31.66	39.79	45.31
RSME	29.76	24.15	32.12	40.29	29.23	23.36	31.97	40.43	34.44	31.78	36.07	39.09
OTKGE	23.86	18.45	25.89	34.23	34.36	28.85	36.25	44.88	35.51	31.97	37.18	41.38
IMF	32.25	24.20	36.00	48.19	34.50	28.77	36.62	45.44	35.79	32.95	37.14	40.63
QEB	28.18	14.82	36.67	51.55	32.38	25.47	35.06	45.32	34.37	29.49	36.95	42.32
VISTA	30.42	22.49	33.56	45.94	32.91	26.12	35.38	45.61	30.45	24.87	32.39	41.53
AdaMF	32.51	21.31	39.67	51.68	34.27	27.21	37.86	47.12	38.06	33.49	40.44	45.48
KBGAN	25.73	9.91	36.95	51.93	29.47	22.21	34.87	40.64	29.71	22.81	34.88	40.21
MANS	28.82	16.87	36.58	49.26	30.88	24.89	33.63	41.78	29.03	25.25	31.35	34.49
MMRNS	32.68	23.01	37.86	51.01	35.03	28.59	37.49	47.47	35.93	30.53	39.07	45.47
AdaMF-MAT	35.14	25.30	41.11	52.92	35.77	28.87	38.85	48.71	38.57	34.34	40.59	45.76
MyGO	37.17	**29.62**	41.66	51.87	36.09	30.02	38.26	46.89	38.66	**34.85**	39.96	44.37
Ours	**37.23**	27.78	**42.75**	**54.71**	**38.46**	**31.20**	**41.78**	**51.46**	**40.03**	33.89	**43.11**	**50.81**
	+0.16%	-	+2.6%	+3.38%	+6.6%	+3.9%	+7.5%	+5.6%	+3.5%	-	+6.2%	+11.0%

Table 2: Ablation Study

Settings	**DB15K**				**MKG-W**				**MKG-Y**			
	MRR	**Hit@1**	**Hit@3**	**Hit@10**	**MRR**	**Hit@1**	**Hit@3**	**Hit@10**	**MRR**	**Hit@1**	**Hit@3**	**Hit@10**
LBMKGC	37.23	27.78	42.75	54.71	38.46	31.20	41.78	51.46	40.03	33.89	43.11	50.81
w/o LvMC	36.73	27.19	42.01	53.76	37.12	29.43	40.28	49.16	39.40	33.42	42.47	49.59
w/o CMoA	36.81	27.69	41.56	53.37	38.03	29.88	40.86	49.34	39.34	33.58	42.28	49.37
w/o CGuAF	36.36	26.55	42.39	54.48	37.53	30.51	41.76	50.77	39.29	33.01	42.92	50.86
w/o e_s^{emb}	31.21	16.96	41.12	54.60	34.76	23.84	41.86	51.97	37.79	29.33	42.43	51.02
w/o e_t^{emb}	36.51	27.09	41.45	53.19	36.72	30.01	40.02	48.79	38.33	32.72	40.95	47.67
w/o e_v^{emb}	37.08	27.34	42.01	53.92	37.89	30.63	41.05	50.89	39.24	33.13	41.76	49.73

4.4 Modality-missing Results (RQ3)

We conducted link prediction experiments with modality dropout on the DB15K dataset. The visual modality was discarded at rates of $\eta = \{0.25, 0.5, 0.75, 1.0\}$. Specifically, for LBMKGC, we used the LvMC module to fully compensate for the missing visual modality information. For other models, we used the commonly used random initialization method [20] to compensate for the missing information. The experimental results are shown in the Figure 3, as the missing modality rate increases, the baseline models (TBKGC [19] and AdaMF-MAT [38]) exhibit significant performance degradation, whereas our proposed LBMKGC demonstrates relatively stable performance. LBMKGC's explicit modeling of missing modalities based on large-scale pre-trained models (i.e., generating real modality information) boosts MMKGC's performance and robustness more effectively than strategies using adversarial training or random initialization.

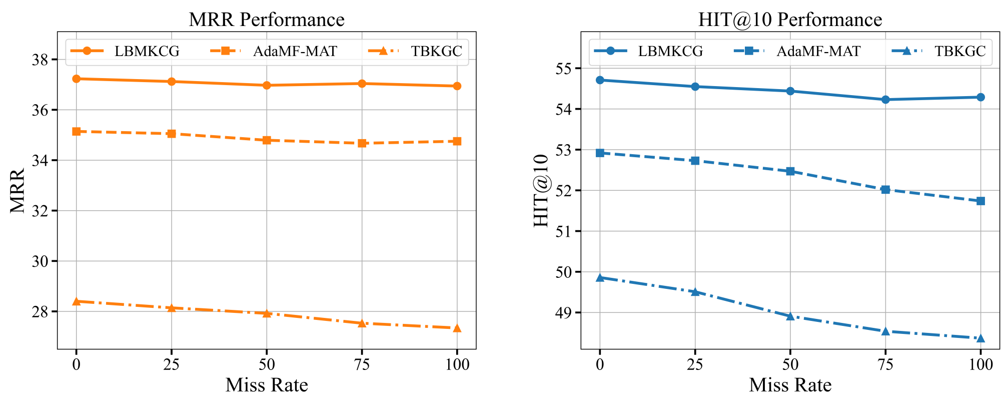
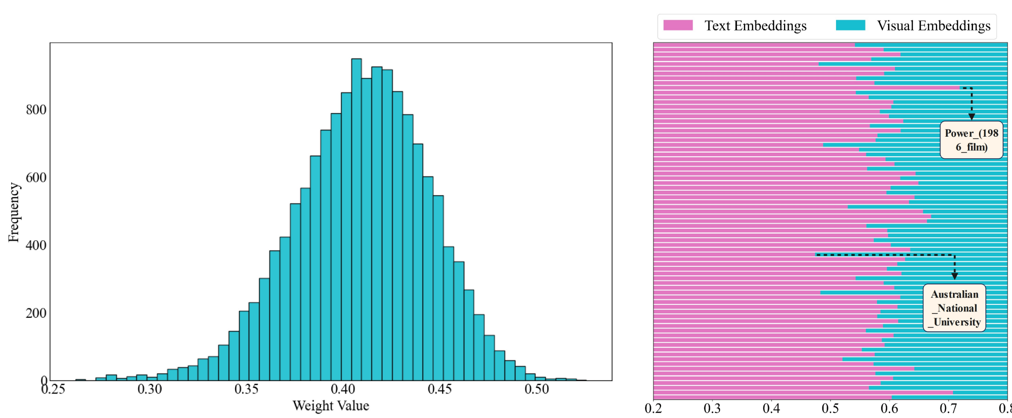


Figure 3: Link prediction results on modality-missing DB15K dataset.



(a) Histogram of Visual Modality Weights Distribution (b) Comparison Chart of Text and Visual Modality Weights

Figure 4: Fusion weights of different modalities.

4.5 Case Study (RQ4)

To intuitively demonstrate the effectiveness of CGuAF, we present the histogram of visual modality weight distribution for all entities in the MKG-W dataset in Figure 4 (a). Based on this, we randomly selected some entities' textual and visual modality weights for comparison in Figure 4 (b), for the conceptual entity "Power_(1986_film)", the weight of textual information is 71%, while that of visual modality is 29%, whereas for the perceptual entity "Australian_National_University", the weight of textual information is 47%, and that of visual modality is 53%. This reflects that CGuAF can dynamically adjust the weight of each modality according to the specific (perceptual) or abstract (conceptual) characteristics of different triad groups of entities, enhancing and retaining the truly important parts of multi-modal information, thereby significantly enhancing the model's ability to interact with multi-modal features under scenarios of complex relationships.

5 Conclusion and Future Work

In this paper, we mainly discussed the imbalance of multi-modal information among entities, the heterogeneity of multi-modal information within entities, and the issue of inconsistent information quantities across different modalities. We proposed a novel MMKGC framework, LBMKGC, to address the limitations of existing methods. LBMKGC consists of three core modules: LvMC, CMoA, and CGuAF, which sequentially solve the issues of imbalance, heterogeneity, and inconsistent information quantities. We conducted in-depth theoretical analysis to demonstrate the rationality of our design and comprehensive experiments on 21 benchmarks to show the effectiveness and robustness of our framework.

However, our study currently focuses on text and image modalities, without integrating semantically rich information such as audio and video into a unified framework. Additionally, our experiments have been conducted exclusively on general domain datasets, leaving the adaptability and robustness in specialized fields like biomedical data and social networks unverified. In the future, MMKGC tasks should more comprehensively consider these modalities in the real world and design generalizable and scalable model architectures to adapt to various tasks and modality combinations, thereby enhancing the ability to model complex relationships.

6 Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.62002039, 62472062, U23B2019).

References

[1] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795, 2013.

[2] Liwei Cai and William Yang Wang. KBGAN: adversarial learning for knowledge graph embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1470–1480. Association for Computational Linguistics, 2018.

[3] Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. OTKGE: multi-modal knowledge graph embeddings via optimal transport. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[4] Linlin Chao, Jianshan He, Taifeng Wang, and Wei Chu. Pairre: Knowledge graph embeddings via paired relation vectors. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4360–4369. Association for Computational Linguistics, 2021.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[6] Junnan Dong, Qinggang Zhang, Huachi Zhou, Daochen Zha, Pai Zheng, and Xiao Huang. Modality-aware integration with large language models for knowledge-based visual question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 2417–2429. Association for Computational Linguistics, 2024.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[8] Xin Gu, Guang Chen, Yufei Wang, Libo Zhang, Tiejian Luo, and Longyin Wen. Text with knowledge graph augmented transformer for video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18941–18951. IEEE, 2023.

[9] Roshni G. Iyer, Yunsheng Bai, Wei Wang, and Yizhou Sun. Dual-geometric space embedding model for two-view knowledge graphs. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 676–686. ACM, 2022.

[10] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 687–696, 2015.

[11] Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Jiyoung Whang. VISTA: visual-textual knowledge graph representation learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7314–7328. Association for Computational Linguistics, 2023.

[12] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.

[13] Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. IMF: interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 2572–2580. ACM, 2023.

[14] Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. MMKG: multi-modal knowledge graphs. In *The Semantic Web - 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2-6, 2019, Proceedings*, volume 11503, pages 459–474, 2019.

[15] Xinyu Lu, Lifang Wang, Zejun Jiang, Shichang He, and Shizhong Liu. MMKRL: A robust embedding approach for multi-modal knowledge graph representation learning. *Appl. Intell.*, 52(7):7480–7497, 2022.

[16] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[18] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[19] Hatem Mousselly Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 225–234. Association for Computational Linguistics, 2018.

[20] Hatem Mousselly Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 225–234. Association for Computational Linguistics, 2018.

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

[22] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[23] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy, editors, *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 697–706. ACM, 2007.

[24] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[25] Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou. Orthogonal relation transforms with graph context modeling for knowledge graph embedding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2713–2722. Association for Computational Linguistics, 2020.

[26] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2071–2080. JMLR.org, 2016.

[27] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.

[28] Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. Is visual context really helpful for knowledge graph? A representation learning perspective. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 2735–2743. ACM, 2021.

[29] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. KGAT: knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 950–958. ACM, 2019.

[30] Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. TIVA-KG: A multimodal knowledge graph with text, image, video and audio. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 2391–2399. ACM, 2023.

[31] Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. Multimodal data enhanced representation learning for knowledge graphs. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019.

[32] Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. Multimodal data enhanced representation learning for knowledge graphs. In *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*, pages 1–8. IEEE, 2019.

[33] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Image-embodied knowledge representation learning. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3140–3146. ijcai.org, 2017.

[34] Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 3857–3866. ACM, 2022.

[35] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[36] Yichi Zhang, Mingyang Chen, and Wen Zhang. Modality-aware negative sampling for multi-modal knowledge graph embedding. In *International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, June 18-23, 2023*, pages 1–8. IEEE, 2023.

[37] Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Tokenization, fusion, and augmentation: Towards fine-grained multi-modal entity representation. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 13322–13330. AAAI Press, 2025.

[38] Yichi Zhang, Zhuo Chen, Lei Liang, Huajun Chen, and Wen Zhang. 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, 20-25 May, 2024, Torino, Italy*, pages 17120–17130. ELRA and ICCL, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We stated our contributions and the problems we addressed in the Abstract and Introduction, and compared our results with more than twenty baselines, achieving state-of-the-art (SOTA) performance.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitations of our work in Section 5 Conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We have provided the corresponding formulas in the Section 3, as well as a complete set of assumptions and correct proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have provided the code and data, and introduced the experimental settings in the Section 4.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provided link to access the code and data in the Abstract and detailed the experimental settings.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We detailed the experimental settings in the Section 4.1 and provided link to the code and data.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We utilized three open-source benchmark datasets and conducted training/testing splits and other initialization tasks in a manner consistent with other papers (such as MyGO [37], AdaMF-MAT [38], MMRNS [34]) to ensure the validity of our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We described the server model and operating system used, along with other experimental settings, in the Section 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The data we used are all from open-source datasets, and we ensure that our research complies with the NeurIPS Code of Ethics in all respects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Our research focuses on knowledge graph completion, which can be applied to tasks such as recommendation systems, and does not involve social impacts.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: Our paper does not involve such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The resources we used are all open-source, and we have cited them in our references.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provided a link to the code in the Abstract, and all the resources we used are open-source.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We described the details and principles of the LLMs and explained how to use them in the provided link.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Technical Appendices and Supplementary Material

A.1 Dataset Analysis

In real-world multimodal knowledge graphs, the modal information is diverse and complex. Some entities may have rich multimodal information such as images and text, while others may have scarce modalities (e.g., containing only text modality with the absence of image modality). This uneven distribution makes the task of knowledge graph completion exceptionally challenging. The missing modal information can lead to the omission of key knowledge, thereby affecting the accuracy of the completion process. As shown in Figure 5, we present the phenomenon of imbalance across different datasets. Specifically, we focus on the imbalance of the image modality and calculate the proportion of entities containing image modality within the entire set of entities. Regarding the text modality, even if an entity lacks a corresponding textual description, the entity itself contains textual information (for example, the triple <snowman, state, stillness>, which consists of three words and is also considered a form of textual information), so we consider the text modality to be balanced.

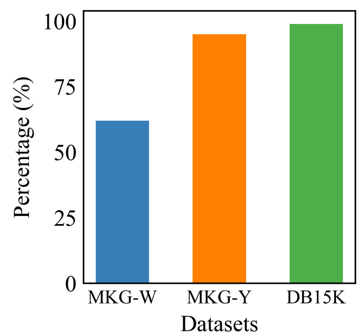


Figure 5: Imbalance Situations in Different Datasets

A.2 Generalization Experiments

We conducted generalization experiments on four different MMKGC models (namely TBKGC, IKRL, AdaMF and QEB) using the multimodal features obtained from the LvMC and CMoA modules. We employed the same parameters as LBMKGC on the DB15K dataset. In fact, with more meticulous parameter tuning tailored to different MMKGC models, these models could potentially exhibit even better performance. We report the Mean Reciprocal Rank (MRR) and Hit@10 results on the DB15K dataset, as shown in Figure 6. Based on the experimental results, we can conclude that by addressing the imbalance of multimodal information between entities and the heterogeneity of multimodal information within entities, MMKGC models can achieve significant performance gains. This indicates that the LvMC and CMoA modules can serve as universal components applicable to various models in downstream tasks.

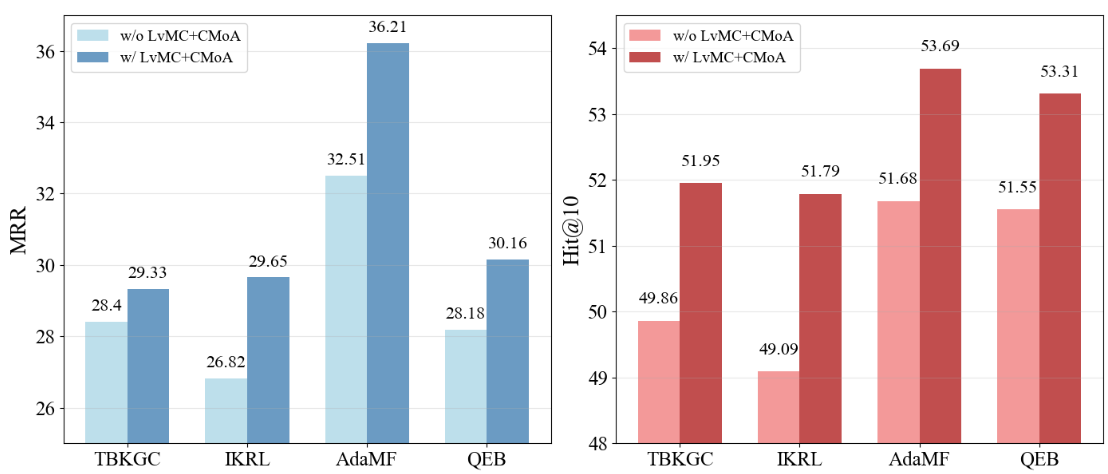


Figure 6: Imbalance Situations in Different Datasets

A.3 Statistical Testing

To rigorously evaluate the stability and reproducibility of our proposed LBMKGC method, we have conducted five independent experiments on each of the three benchmark datasets (MKG-W, MKG-Y, and DB15K) without fixing the random seed during training.

The comprehensive results across all datasets, as depicted in Figure 7, demonstrate highly consistent performance patterns throughout the evaluation metrics. On the MKG-W dataset, MRR values remain confined to a narrow interval of 0.3840–0.3866, while Hit@1 scores exhibit stability, ranging from 0.3119 to 0.3152. A similar degree of consistency is observed on the MKG-Y dataset, where MRR varies between 0.3992 and 0.4031, accompanied by Hit@1 scores spanning 0.3387–0.3449. The DB15K dataset also reflects minimal performance deviation, with MRR consistently maintained between 0.3707 and 0.3734. The standard deviations across all evaluation metrics on MKG-W, MKG-Y, and DB15K datasets remain below 0.0026, demonstrating the exceptional algorithmic stability of our proposed method.

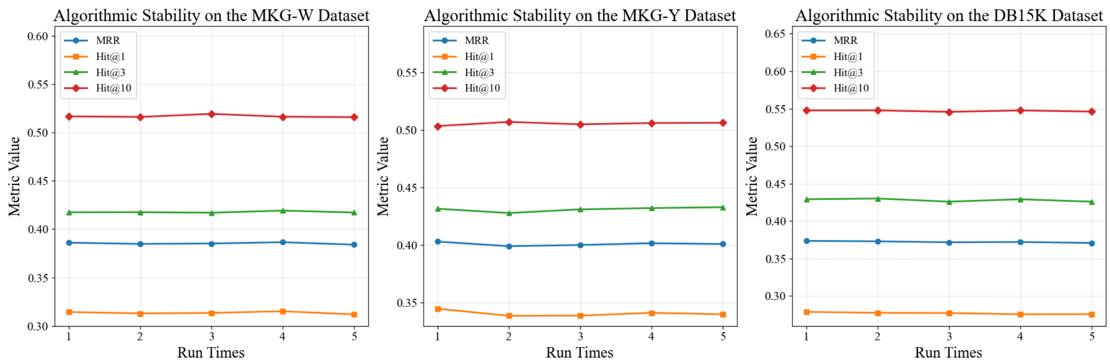


Figure 7: Algorithmic Stability on Different Datasets

A.4 Hyperparameter Sensitivity

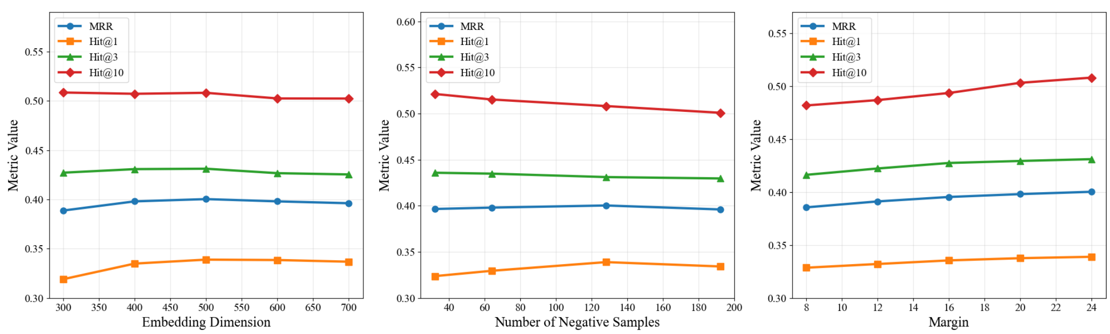


Figure 8: Hyperparameter Sensitivity on MKG-W Dataset

We have conducted experiments on the core hyperparameters (including embedding dimension, number of negative samples, and margin) using the MKG-Y dataset. To ensure the fairness of these experiments, we fixed the random seed at 42 and employed a controlled variable approach for multiple independent experiments. The experimental results are shown as Figure 8.

The experimental results indicate that while there is some variability in the outcomes under different hyperparameter settings, this variability is minor and does not significantly impact the model's performance.

A.5 Time Efficiency

We have conducted time efficiency experiments on the DB15K dataset using an NVIDIA GeForce 4090 GPU, employing the LBMKGC, TBKGC, QEB, RSME, and IKRL methods. The results are as Figure 9, this deceleration is within an acceptable range and leads to substantial performance improvements. The increased computational time is primarily due to the additional processing required for assigning weights to the text and image modalities, as well as to the graph structural information regulated by relational context.

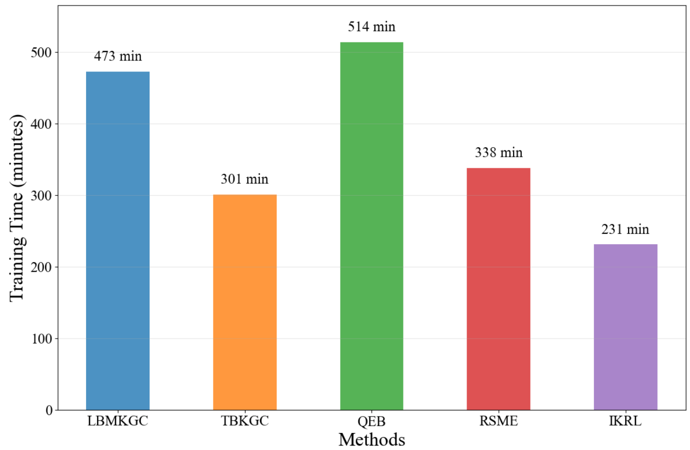


Figure 9: Training Time Efficiency on DB15K Dataset

Overall, the LBMKGC model achieves a commendable balance between efficiency, performance, and stability. The additional computational investment yields disproportionately valuable returns in terms of prediction accuracy and model robustness, making it particularly suitable for applications where performance is prioritized over minimal training time.