MMKGR: Multi-hop Multi-modal Knowledge Graph Reasoning

Shangfei Zheng† Weiqing Wang†† Jianfeng Qu† Hongzhi Yin‡ Wei Chen†* Lei Zhao†*
†*School of Computer Science and Technology, Soochow University*
††*Department of Data Science and AI, Monash University*
‡*School of Information Technology and Electrical Engineering, The University of Queensland*
†sfzhengsuda@stu.suda.edu.cn ††Teresa.Wang@monash.edu
‡db.hongzhi@gmail.com †{jfqu, robertchen, zhaol}@suda.edu.cn

2023 IEEE 39th International Conference on Data Engineering (ICDE) | 979-8-3503-2227-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDE55515.2023.00015

***Abstract*—Multi-modal knowledge graphs (MKGs) include not only the relation triplets, but also related multi-modal auxiliary data (i.e., texts and images), which enhance the diversity of knowledge. However, the natural incompleteness has significantly hindered the applications of MKGs. To tackle the problem, existing studies employ the embedding-based reasoning models to infer the missing knowledge after fusing the multi-modal features. However, the reasoning performance of these methods is limited due to the following problems: (1) ineffective fusion of multi-modal auxiliary features; (2) lack of complex reasoning ability as well as inability to conduct the multi-hop reasoning which is able to infer more missing knowledge. To overcome these problems, we propose a novel model entitled MMKGR (Multi-hop Multi-modal Knowledge Graph Reasoning). Specifically, the model contains the following two components: (1) a unified gate-attention network which is designed to generate effective multi-modal complementary features through sufficient attention interaction and noise reduction; (2) a complementary feature-aware reinforcement learning method which is proposed to predict missing elements by performing the multi-hop reasoning process, based on the features obtained in component (1). The experimental results demonstrate that MMKGR outperforms the state-of-the-art approaches in the MKG reasoning task.**

***Index Terms*—Multi-modal knowledge graph, Multi-hop knowledge graph reasoning, Multi-modal fusion**

I. Introduction

Knowledge Graph (KG) is essentially a kind of graph structure with entities as nodes and relations as edges, and has received extensive attention in both data mining [12] and knowledge engineering [78] areas. At present, large-scale KGs have achieved great success in assisting many applications, such as information retrieval [29], question answering [20], recommendation systems [62] [80] [72], etc. Nevertheless, most of traditional KGs only contain structural data in the form of relation triplets, i.e., (*source entity*, *relation*, *target entity*), ignoring massive amounts of multi-modal data such as text and images in reality. To integrate more diverse knowledge in KGs, multi-modal Knowledge graph (MKG) has been proposed [28] [63]. As the example presented in Fig. 1, a MKG not only contains the structural data, but also includes additional multi-modal auxiliary data (i.e., texts and images), and it is more in line with the characteristics of real-world data compared with traditional KGs [31] [49] [66]. Despite the abundant information contained by a MKG, it still suffers from the natural incompleteness of KGs, e.g., a triplet (*Titanic*, *Starred_by*, *Kate Winslet*) is missed in Fig. 1, which has significantly hindered the applications of MKGs [23].

To solve the problem of natural incompleteness of KGs, various KG reasoning approaches have been proposed [26] [79] [8] [19]. The key idea of these methods is to infer new knowledge by effectively integrating existing information in the graph [7], and they mainly focus on traditional KGs without considering multi-modal knowledge. Different from these methods, some reasoning models [64] [65] [61] [45] [53] [50] have been proposed to integrate the multi-modal knowledge on MKGs, but they are based on the model TransE [3] that focuses on completing the single-hop reasoning only. Notably, the single-hop reasoning models lack explainability and suffer from low reasoning performance since KGs have the most inferred potential knowledge within multiple hops [8]. Accordingly, there is another stream of KG reasoning methods focusing on multi-hop reasoning. A representative method in this stream is reinforcement learning (RL)-based multi-hop reasoning owing to its ability to leverage the semantic combination in KGs [56], which makes the whole reasoning process explainable [58]. For example, by connecting (*Titanic*, *Heroine*, *Rose Bukater*) and (*Rose Bukater*, *Played_by*, *Kate Winslet*), RL-based reasoning models obtain a missing triplet (*Titanic*, *Starred_by*, *Kate Winslet*). It has been proven that RL-based multi-hop KG reasoning models have not only semantic explainability, but also higher reasoning performance than single-hop reasoning models [27] [18] [81], which motivates our study to focus on multi-hop reasoning in MKGs. Note that, the existing multi-hop reasoning methods in KG area have not integrated the multi-modal information so far. An intuitive solution to conduct multi-hop reasoning in MKGs is to extend the existing multi-hop reasoning methods in traditional KGs to include the multi-modal information. However, the following two *challenges* make the direct extension ineffective.

The *challenge I* is the lack of a fine-grained multi-modal information exploiting method in KG reasoning area. Several multi-modal studies have demonstrated that fine-grained features are beneficial for obtaining accurate results in reasoning tasks [24], [73]. Typically, most of the existing MKG rea-

* These authors are corresponding authors.

2375-026X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICDE55515.2023.00015

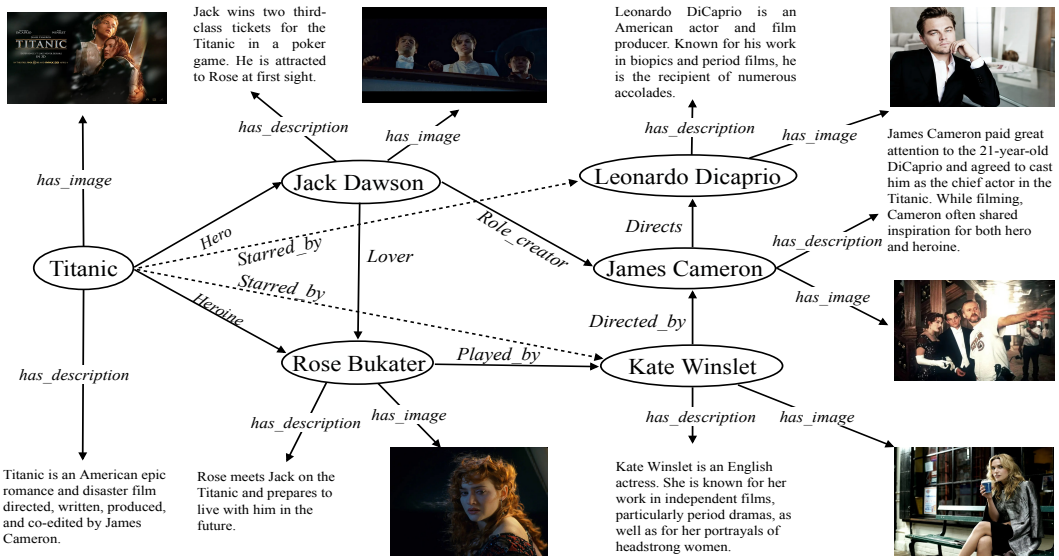


Fig. 1. A small fragment of a multi-modal knowledge graph. The ellipse represents entities, the arrow refers to the relations and dotted arrow represents missing relation. When inferring (*Titanic*, *Starred_by*, ?), the multi-modal auxiliary information such as the image or text contained in each entity can provide reasoning clues for the reasoning model. To easily follow our work, we exemplify this multi-modal knowledge graph throughout the paper.

soning methods learn separate attention distributions for only one modality information (e.g., texts or images) apart from the structure information. Despite the accomplishment that the above methods have achieved, they still remain some unsolved problems. Firstly, these methods cannot simultaneously learn the visual attention and textual attention from fine-grained representations of inter-modal interactions [50]. By way of illustration, in Fig. 1, the potential triplet (*Titanic*, *Starred_by*, *Leonardo Dicaprio*) cannot be inferred until the images and text descriptions in the 3-hop path "Titanic $\xrightarrow{Hero}$ Jack Dawson $\xrightarrow{Role\_creator}$ James Cameron $\xrightarrow{Directs}$ Leonardo Dicaprio" are simultaneously considered. Secondly, existing multi-modal reasoning methods ignore intra-modal interactions that have been shown to be the key of fine-grained learning for many unimodal learning tasks [54], [60]. For example, we can infer that "Titanic" is a love movie about two people embracing each other only from the image of the entity "Titanic". Last but not least, some irrelevant noise (e.g., black background in images) and redundant noise (compared with the image of Rose Bukater, the image of Kate Winslet is highly similar and contains less useful information) impair the robustness and generalization of the model [4], [25]. How to learn fine-grained knowledge by simultaneously solving the above problems is non-trivial in MKG reasoning area.

The *challenge II* lies in directly extending the RL-based KG reasoning method to MKG reasoning easily generates some wrong reasoning paths and degrades the reasoning performance. This is because the introduction of multi-modal auxiliary data further exacerbates the sparse reward problem, which leads to the decision bias of reinforcement learning [5], [15], [35], [76]. In fact, most RL-based KG reasoning methods have suffered from the problem of sparse reward on traditional KGs, which is manifested in the lack of feedback rewards and blind reasoning in most states [27]. Some KG reasoning methods have tried to alleviate this problem, but they still have the following technical limitations: (1) The general designed principles of density, exploration and restraint are not fully considered in the reward function, which leads to failure to converge in the late phase of training [10], [11]. (2) Lack of reward balancing mechanism to prevent reasoning models from repeatedly grabbing local rewards while ignoring the ultimate goal [43], [51]. (3) Lack of a paradigm for perceiving and exploiting multi-modal features in reinforcement learning.

To deal with the above challenges, we propose a novel model entitled MMKGR (**M**ulti-hop **M**ulti-modal **K**nowledge **G**raph **R**easoning). The main difference between our model and existing ones is that MMKGR not only effectively extracts and utilizes multi-modal auxiliary features, but also completes multi-hop reasoning in MKGs. Specifically, the model contains the following two components. (1) To solve the *challenge I*, a unified gate-attention network is designed to generate multi-modal complementary features with sufficient interactions and less noise. Its attention-fusion module extracts fine-grained multi-modal features to complete inter-modal attention interactions (i.e., co-attention across different modalities) and intra-modal attention interactions (i.e., self-attention within each modality) simultaneously. An irrelevance-filtration module of this network further filters out irrelevant features and outputs more reliable multi-modal complementary features. Note that, the unified gate-attention network simultaneously aggregates low-noise information to obtain the triple query-related fine-grained representation from intra-modality and inter-modality. (2) To solve the *challenge II*, a complementary feature-aware RL method is proposed to predict the missing elements by performing the multi-hop reasoning process. More precisely, a carefully designed 3D reward mechanism, which includes

Destination reward, *Distance reward* and *Diverse reward*, is proposed in MMKGR. This study is summarized as the following key contributions.

- We are the first to point out *how to effectively leverage multi-modal auxiliary features to conduct multi-hop reasoning in KG area*, and this study provides a new perspective on KG reasoning.
- We propose a novel model MMKGR, which contains a unified gate-attention network that builds sufficient attention interactions with less noise, and a complementary feature-aware RL method that is designed to alleviate the problem of sparse rewards and conduct multi-hop reasoning in MKGs.
- Abundant experiments on two public MKG datasets have been conducted, and the experimental results show that better performance of MMKGR against the existing SOTA reasoning baselines.

Our paper is organized as follows. Firstly, we present the existing studies in Section II. Then, we add the KGR concept and formulate the problem in next section. Section IV shows MMKGR to conduct the multi-hop MKG reasoning. Experiments are conducted in Section V, which is followed by the conclusion and future work in Section VI.

II. Related Work

A. Multi-modal Knowledge Graph

A KG is essentially a structured semantic network composed of entities and relations [46] [37]. At present, the actual internet data show multi-modal characteristics [18]. MKGs are proposed to integrate multi-modal data in KGs [41] [45]. A MKG is composed of structural data (i.e., relation triplets), and multi-modal auxiliary data (i.e., texts and images) [28].

The multi-modal auxiliary data associated with early MKGs present the singularity. For example, as a precedent for MKGs, the entities of IMGpedia [13] come from a specific KG (i.e., DBpedia), and the multi-modal auxiliary data only contain images. Similar studies [28] [41] expand the existing KGs WN-9 and FB15K respectively, only adding images for each entity to further explain them. Although the above MKGs connect structural entities with images, they do not consider the diversity of images. To solve this problem, Richpedia [57] filters out similar images to ensure diversity. There are also MKGs that only contain textual descriptions. One representation is FB20K [64] that only textual descriptions are added to each entity. Although FB-Des [50] adds textual descriptions and hierarchical types for each entity on the basis of FB15k-237, the auxiliary data of this MKG is still singular. To expand the auxiliary data with one modality, two datasets WN9-IMG-TXT and FB-IMG-TXT simultaneously add a number of textual descriptions and images to each entity, aiming to further enhance the data diversity of the MKGs [45]. While these MKGs add a large amount of multi-modal auxiliary data, they also generate redundant and irrelevant data, which limits the performance of multi-modal fusion.

TABLE I
Summarization of Existing KG Reasoning Models.

Models \ KGs / Type	On traditional KGs	On multi-modal KGs
Single-hop reasoning	TransE, TransD ComplEx, HolE DistMult, RESCAL	IKRL, DKRL TransAE, MTRL KR-AMD, MKRL
Multi-hop reasoning	MINERVA, DeepPath GaussianPath, RLH GAATs, NeuralLP	**MMKGR**

B. Fusion Strategies for Multi-modal Learning

Early multi-modal studies only fuse the global features of all modalities through vector concatenation. The limitation of this method is that the multi-modal noise affects the extraction of key features [14]. Thus, some multi-modal studies [71] [30] [65] adopt conventional attention model to extract important features of auxiliary modalities. Compared with earlier fusion methods, the conventional attention mechanism aggregates essential information to obtain the key local representation [68]. Furthermore, considering that conventional attention mechanisms cannot perform feature interactions in all modalities at the same time, some studies propose co-attention models to simultaneously assign and aggregate essential information for all modalities [74] [77]. Although co-attention is extended to learn all modalities at the same time, these models, like conventional attention mechanism, still learn coarse interactions among all modalities. To address the problem of insufficient multi-modal interactions, MCAN [73] and PSAC [24] apply self-attention mechanism [54] and co-attention to complete the intra-modal and inter-modal attention interactions. The above methods have sufficient interaction, but ignore the following details: (1) redundant and irrelevant features can impair the generalization and robustness of the model [25]; (2) only self-attention or co-attention is considered in the same training stage, which limits the utilization of samples and the effective extraction of fine-grained features [54]. The inadequacy of existing methods motivates our multi-modal fusion goal to simultaneously complete the intra-modal and inter-modal interactions in a unified and low-noise way.

C. Knowledge Graph Reasoning

Since KGs are inherently incomplete, KG reasoning technology that can synthesize the original knowledge and infer the missing knowledge is particularly important [18] [17]. We summarize the existing knowledge graph reasoning models in Table I. A group of KG reasoning studies on traditional KGs aim at inferring missing elements by embedding-based methods [48] [3] and deep learning [9] [44]. For example, the embedding-based TransE [3] learns vector representations of entities and relations by minimizing the heuristic self-supervised loss functions and then the learned vectors are used to predict the probability of correct triplets. However, all the above methods are not suitable for multi-hop KG reasoning by modeling multi-step relations containing more information [8]. Moreover, these methods cannot leverage the semantic combination of relational path in KGs [56].

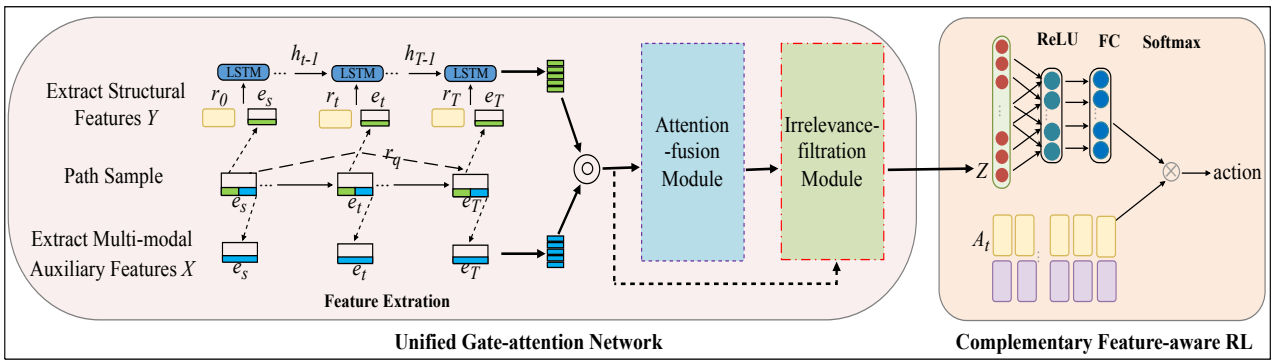


Fig. 2. Overview of MMKGR. The structural features Y in green and the multi-modal auxiliary features X in blue are obtained by feature extraction. These features at the reasoning step t are fed to the unified gate-attention network to generate multi-modal auxiliary features Z which are fed into the complementary feature-aware RL to predict action for the next reasoning step $t+1$, until reaching the target entity e_T.

Existing developments of the knowledge reasoning on traditional KGs focus on the field of RL-based KG reasoning methods. The reasoning procedure of these reasoning methods based on RL is intuitive by exploiting the symbolic compositionality of multi-step relations containing more information [8]. Typical RL-based methods include MINERVA [8], DeepPath [55], RLH [56], GaussianPath [55], etc. Also, there are other multi-hop methods, e.g., rules-based NeuralLP [70] and graph attention networks-based GAATs [59]. In spite of the superior performance of these existing multi-hop KG reasoning methods, they ignore the multi-modal data types, and cannot use multi-modal auxiliary data to complete reasoning. As presented in Table I, our MMKGR fills the gap for multi-hop reasoning on MKGs.

Focusing on single-hop reasoning on MKGs, some studies employ the conventional attention model or concatenation to fuse multi-modal features and then adopt TransE to infer missing elements, such as IKRL [65] [63] and TransAE [61]. Note that, Wang et al. have proved that the performance of TransAE on MKGs is better than that of the most traditional KG reasoning methods, such as TransE, RESCAL [39], ComplEx [52], HolE [38], and DistMult [69]. KR-AMD [53] and MKRL [50] leverage textual data as part of auxiliary data to improve reasoning performance. In addition, MTRL [45] with the state-of-the-art performance performs single-hop reasoning by concatenating the features of relation triplets and multi-modal auxiliary features that comprehensively contain textual and visual features. However, the above studies cannot simultaneously learn visual and textual attention to fully understand the semantics of the two modalities. This coarse interaction causes ineffective fusion of multi-modal features.

III. Preliminary and Definition

A KG $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{U}\}$ is a multi-relation heterogeneous graph, where $\mathcal{E}$ is a range of entities and $\mathcal{R}$ denotes the set of semantic relations. $\mathcal{U} = \{(e_s, r, e_d) \mid e_s, e_d \in \mathcal{E}, r \in \mathcal{R}\}$ presents a set of triplets in the KG $\mathcal{G}$, where e_s, e_d, and r denote a source entity, a target entity, and the semantic relation between these entities, respectively. Relation triplets in knowledge graphs are structured [3], and the corresponding structural features are different from the features learned from auxiliary data (i.e., texts and images). To better understand our methodology in this study, our definitions are introduced as follows.

Definition 1: Multi-modal knowledge graph. A multi-modal knowledge graph $\tilde{\mathcal{G}} = \{\tilde{\mathcal{E}}, \mathcal{R}, \tilde{\mathcal{U}}\}$ is an extension of the knowledge graph $\mathcal{G}$ by adding multi-modal auxiliary data, where each entity in $\tilde{\mathcal{E}}$ is attached to both structural (i.e., the relation triplets in knowledge graphs) and multi-modal data.

Definition 2: Multi-hop reasoning. Given a query among three cases $(e_s, r, ?)$, $(e_s, ?, e_d)$, $(?, r, e_d)$, where "?" represents the missing element, the goal of reasoning is to infer the missing element by the relational path shorter or equal k hops, where k is an integer not less than 1.

Example: Given a triple query (*Titanic*, *Starred_by*, ?), a 3-hop reasoning path is "Titanic $\xrightarrow{Hero}$ Jack Dawson $\xrightarrow{Role\_creator}$ James Cameron $\xrightarrow{Directs}$ Leonardo Dicaprio".

Definition 3: Multi-modal auxiliary feature. Multi-modal auxiliary feature $\boldsymbol{x}$ of each entity e in entity set $\tilde{\mathcal{E}}$ is expressed as a vector, learned from multi-modal auxiliary data (text or image) and $\boldsymbol{x} \in \boldsymbol{f}_t \circ \boldsymbol{f}_i$, where "$\circ$" represents a multi-modal fusion method. The textual feature vector $\boldsymbol{f}_t$, image feature vector $\boldsymbol{f}_i$, and $\boldsymbol{x}$ are learned with some representation methods.

Problem 1: Our paper aims to infer the missing entity or relation for a given query on multi-modal knowledge graphs. The problem formulation is given as:

- Our input is a query $(?, r_q, e_d)$, $(e_s, r_q, ?)$ or $(e_s, ?, e_d)$, where r_q is a query relation, e_s and e_d are entity containing structural feature and multi-modal auxiliary feature, and the symbol "?" is the missing element in the triple query.
- Our output is a correctly predicted entity or relation obtained via RL-based multi-hop reasoning networks.

IV. Methodology

A. Overview of MMKGR

Our proposed model MMKGR, the overview of which is shown in Fig. 2, contains two components: (1) a unified gate-attention network, which is designed to conduct sufficient attention interactions and filter noises to generate more effective and reliable multi-modal complementary features encoding

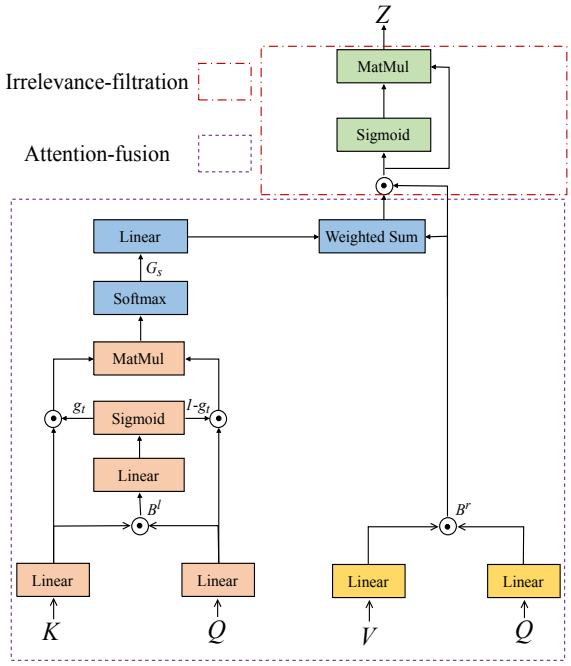


Fig. 3. The detailed schematic diagram of the unified gate-attention network to generate multi-modal complementary features.

relevant knowledge of all modalities; (2) a complementary feature-aware reinforcement learning framework, which is proposed to predict the missing elements in the multi-hop reasoning process with a carefully designed reward mechanism and the useful multi-modal complementary features.

B. Unified Gate-attention Network

The insufficient multi-modal interactions and noise interference in existing MKG reasoning methods severely limit the utilization of multi-modal data [73] [36]. To solve the multi-modal representation problem, we introduce a novel and unified gate-attention network to learn from multi-modal data in this subsection. Notably, this network is mainly inspired by (1) the significant impact of intra-modal attention interaction on fine-grained features [54], and (2) gate network which effectively filters noise [77]. Based on this, the unified gate-attention network selects features of different modalities online, and simultaneously completes intra-modal and inter-modal attention interactions with noise robustness, which is technically different from both existing KG multi-modal data modelling methods [64] [65] [61] [45] [50] and general multi-modal data modelling approaches [24] [54] that are not limited to KG area.

Specifically, the unified gate-attention network includes feature extraction, an attention-fusion module, and an irrelevance-filtration module. The extracted features of all modalities are fed into the attention-fusion module that fuses structural features and multi-modal auxiliary features together, by attending them with a carefully designed fine-grained attention scheme. Then, the irrelevance-filtration module discards irrelevant or even misleading information and generates noise robust multi-modal complementary features. Fig. 3 presents the schematic diagram of the unified gate-attention network, the details of which are illustrated as follows.

1) Feature Extraction: (1) Structural features with d_s dimensions are initialized from all entities and relations by using the TransE algorithm [3]. The source entity e_s and query relation r_q are represented as the dense vector embedding $\mathbf{e}_s$ and $\mathbf{r}_q$ respectively. In addition, the history of reasoning path that consists of the visited entities and relations is defined as $h_t = (e_s, r_0, e_1, r_1, ..., e_t)$. Next, we leverage LSTM to integrate the vector of history information $\mathbf{h}_t$ with d_s dimensions into structural features. Given the query in our multi-hop reasoning process, the structural features $\mathbf{y}$ are defined as,

$$\mathbf{y} = [\mathbf{e}_s; \mathbf{h}_t; \mathbf{r}_q] \tag{1}$$

$$Y = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m] \tag{2}$$

where $Y \in \mathcal{R}^{m \times d_y}$ represents a group of structural features, m and d_y are the number of entities and the dimension of the features in this triple query, respectively. (2) To initialize image features $\mathbf{f}_i$, we extract a d_i-dimensional vector of the last fully-connected layer before the softmax in VGG model [6]. (3) Textual features $\mathbf{f}_t$ are initialized by the word2vec framework [33] and expressed as a d_t-dimensional vector. To flexibly add multi-modal auxiliary features, we concatenate the above two groups of features on rows to form the multi-modal auxiliary features $\mathbf{x}$,

$$\mathbf{x} = [\mathbf{f}_t W_t; \mathbf{f}_i W_i] \tag{3}$$

$$X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m] \tag{4}$$

where $W_t \in \mathcal{R}^{d_t \times d_x/2}$, $W_i \in \mathcal{R}^{d_i \times d_x/2}$, and $X \in \mathcal{R}^{m \times d_x}$ represents a group of multi-modal auxiliary features, d_x is the dimension of the feature.

2) Attention-fusion Module: To obtain the complementary features for reinforcement learning, we need to fuse the structural features Y and multi-modal auxiliary features X generated in feature extraction. However, redundant features tend to have a negative impact on the prediction during the multi-modal fusion [77]. Specifically, redundant features are either shifted versions of the features related to the triple query or very similar with little or no variations [1], which can amplify the negative effects of noise [22]. For example, suppose we have "Titanic $\xrightarrow{Heroine}$ Rose Bukater $\xrightarrow{Played\_by}$ Kate Winslet" and the queries are about the movie Titanic, the features of an image containing Rose Bukater provide related information for the queries while the features coming from another very similar image or the images containing Kate Winslet (Rose's player) in other movies are regarded as redundant features. These redundant features add computational complexity and cause collinearity problems [61] [50]. Consequently, we propose the attention-fusion module that is located in the lower area of Fig. 3, with the goal of fusing the structural features and multi-modal auxiliary features effectively.

Specifically, we first utilize linear functions to generate the queries Q, keys K, and values V of the attention mechanism,

$$Q = XW_q, K = YW_k, V = YW_v \tag{5}$$

where $W_q \in \mathcal{R}^{d_x \times d}, W_k, W_v \in \mathcal{R}^{d_y \times d}$, and $Q, K, V \in \mathcal{R}^{m \times d}$ have the same shape. Then, the joint representation B^l of *Q* and *K* is learned based on MLB pooling method [21], inspired by the recent successes of it in fine-grained multi-modal fusion,

$$B^l = KW_k^l \odot QW_q^l \tag{6}$$

Similarly, we can generate the joint representation B^r of *V* and *Q* with the following equation,

$$B^r = VW_v^r \odot QW_q^r \tag{7}$$

where $W_k^l, W_q^l, W_v^r, W_q^r \in \mathcal{R}^{d \times j}$ are embedding matrices, and $\odot$ is Hadamard product.

Next, the filtration gate g_t applied to different feature vectors is defined as,

$$g_t = \sigma(B^l W_m) \tag{8}$$

where $W_m \in \mathcal{R}^{j \times d}$ is an embedding matrix and σ denotes the sigmoid activation. Based on the filtration gate g_t, we can filter out the redundant features generated during fusion and obtain a new representation with following probability distributions,

$$G_s = softmax((g_t \odot K)((1 - g_t) \odot Q)) \tag{9}$$

where g_t and $1 - g_t$ are used to trade off how many structural features and multi-modal auxiliary features are fused.

Finally, our attention-fusion module generates the attended features $\hat{V}$=$\{\mathbf{v}_i\}_{i=1}^m$ by accumulating the enhanced bilinear values of structural features and multi-modal auxiliary features,

$$\hat{V} = \sum\nolimits_{i=1}^{m} (G_s W_g^l) B_i^r \tag{10}$$

where $W_g^l \in \mathcal{R}^{d \times 1}$, and $\mathbf{v}_i \in \mathcal{R}^{1 \times j}$ denotes a row of the attended features $\hat{V} \in \mathcal{R}^{m \times j}$, feature vector $B_i^r \in \mathcal{R}^{1 \times j}$ is a row of the embedding matrix B^r.

By designing the attention-fusion module, we can complete the intra-modal and inter-modal feature interactions in a unified manner at the same time. This is because the inputs of this module are pairs from structural features and multi-modal auxiliary features, where each vector of a pair may be learned from the same modality or different ones.

3) Irrelevance-filtration Module: To further improve the robustness of the model, we design an irrelevance-filtration module, which is located in the upper area of Fig. 3. The attended features $\hat{V}$ obtained by attention-fusion module may contain irrelevant features [16]. Specifically, irrelevant features are irrelevant to the triple query in the reasoning process. Since the attention mechanism assigns weights to all features, these features tend to participate in model computation and mislead the reasoning policy [42]. For example, the features from the black background of images in Fig. 1 are regarded as irrelevant features. This motivates our model to weight more on the most related complementary features and dynamically filter irrelevant ones. This is achieved by a well designed irrelevance-filtration gate function. The output of this gate is a scalar, the value range of which is [0,1]. The multi-modal complementary features *Z* are obtained as follows,

$$G_f = \sigma(B^r \odot \hat{V}) \tag{11}$$

$$Z = G_f(B^r \odot \hat{V}) \tag{12}$$

where σ and G_f denote the sigmoid activation function and irrelevance-filtration gate, respectively.

C. Complementary Feature-aware Reinforcement Learning

The existing KG reasoning methods based on RL are not suitable to be directly applied to reasoning in MKGs due to the dilemma of sparse rewards [27]. Sparse rewards (i.e., the agent cannot get enough rewards in a short period of time) are more likely to generate wrong reasoning paths, and the multi-modal auxiliary features of entities on these paths aggravate the introduction of noise, thereby further affecting the performance of reasoning. To solve this problem, we propose a novel reward mechanism in this subsection. Compared with existing RL frameworks, the main technical difference of MMKGR lies in the below points. (1) A carefully-designed 3D reward mechanism that combines reward design principles with domain-knowledge of KGR is proposed to eliminate reward sparsity. (2) A novel method, where the policy function is used as a multi-modal perception interface, is first introduced in RL to fully utilize the multi-modal features.

Our proposed model MMKGR transforms the process of reasoning about missing elements into the Markov Decision Process (MDP) where the reasoning target is to select some optimal reasoning decisions (i.e., optimal choices of reasoning paths) to earn the accumulated rewards (e.g., finding the target elements in MKGs). Thus, MMKGR trains an agent to interact with the sample environment of MKGs in the form of the four-tuple of MDP (State, Action, Transition, Reward).

States: State of the RL-based agent corresponds to a set that includes some elements in MKGs. Formally, each state at reasoning step *t* is denoted as s_t= (e_t, (e_s, r_q), $\mathcal{N}_t$, $\mathcal{E}_t$) $\in \mathcal{S}$, where $\mathcal{S}$ is a state space and e_t is the entity that is accessed at step *t*. e_s and r_q are the source entity and query relation respectively. In addition, to ensure the design of a complete state, we also consider the set of neighboring entities $\mathcal{N}_t$ and all edges $\mathcal{E}_t$ connected to e_t.

Actions: For the given state s_t, its action space $\mathcal{A}_t$ is the set of usable actions A_t at reasoning step *t* and a abort action *STOP*. A_t denotes the set of leaving edges connected to e_t, and extends the reasoning path until getting to next node in MKGs. Formally, A_t is expressed as A_t = $\{(r_{t+1}, e_{t+1}) | (e_t, r_{t+1}, e_{t+1}) \in \mathcal{G}\}$. To avoid infinite unroll in the reasoning process, the *STOP* action is executed when the reasoning step *t* increases to the maximum step *T*.

Transition: $\mathcal{P}_r$ is set to transform current state s_t to the next state s_{t+1}. $\mathcal{P}_r$: $\mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ is defined as $\mathcal{P}_r$ (s_t, A_t) = $\mathcal{P}_r$ (e_t, (e_s, r_q), $\mathcal{N}_t$, $\mathcal{E}_t$, A_t). When the reasoning process is unfolded to a constant pace *T*, s_t is denoted as s_T.

Rewards: Reward is a key factor affecting the performance of all RL-based models [47]. Since directly extending the RL-based KG reasoning method to MKG reasoning will exacerbate sparse rewards and degrades reasoning performance, we propose a **3D** reward mechanism (**Destination reward**, **Distance reward**, and **Diverse reward**) inspired by [67].

According to the general principles of reward design, we employ reward shaping to maintain density of reward and introduce domain-knowledge (i.e. some inter-path and intra-path information in KGR) to present constraints and explorations of reward. Note that, the 3D reward mechanism not only preserves the three general principles of reward design, but also extracts intra-path and inter-path influences from the KGR domain, which ensures the completeness of reward design. (1) **Destination reward**. When the agent unrolls to a maximum reasoning pace *T* or reaches e_T, this agent receives a returned destination reward. For existing RL-based methods, the destination reward is 1 if e_T is the ground truth target entity e_d. Otherwise, the destination reward is 0. We argue that this setting makes the rewards more sparse with the increase of the length of the reasoning path. For example, in Fig. 1, if (*Titanic*, *Starred_by*, *Kate Winslet*) or (*Titanic*, *Starred_by*, *Leonardo Dicaprio*) still are not inferred at the maximum step *T*, the above methods fail to obtain an effective reward (i.e., the value of destination reward is 0) and learn slowly or even ineffectively from multi-modal data. To alleviate the problem, we employ a shaping method [27] to design our destination reward when e_d is not reached,

$$R_{destination} = \begin{cases} 1 & e_T = e_d \\ l(e_s, r_q, e_T) & e_T \neq e_d \end{cases} \tag{13}$$

where *l* denote the score mechanism using ConvE [9] and can estimate the likelihood over (e_s, r_q, e_T). (2) **Distance reward**. The degree of sparse rewards tends to be positively correlated with the length of the reasoning path. As shown in Fig. 1, the triplet (*Titanic*, *Starred_by*, *Kate Winslet*) obtained through 2-hop reasoning path gets the terminal reward faster than (*Titanic*, *Starred_by*, *Leonardo Dicaprio*) obtained through 3-hop reasoning path. In fact, when the number of hops exceeds 3, the reasoning performance may be at risk of degradation [8]. Therefore, we use the $R_{distance}$ as part of the reward function to alleviate sparse rewards within a shorter path,

$$R_{distance} = \begin{cases} \frac{1}{k} & k \leqslant 3 \\ -\frac{1}{k^2} & k > 3 \end{cases} \tag{14}$$

where the symbol *k* denotes the amount of hops for the agent at step t, 3 is a threshold of *k*. (3) **Diverse reward**. The lack of exploration further exacerbates sparse rewards. For example, in Fig. 1, given a reasoning task (*Titanic*, *Starred_by*, ?), the agent can successfully complete reasoning via "Titanic $\xrightarrow{Heroine}$ Rose Bukater $\xrightarrow{Played\_by}$ Kate Winslet". The path found early will be biased, which limits exploration of novel paths. Thus, $R_{diversity}$ based on the Gaussian kernel is encouraged to explore a diverse set of paths and prevents the agent from falling into the locally optimal path that negatively impacts on rewards,

$$R_{diversity} = -\left|\frac{1}{V}\right| exp(-\frac{\|\mathbf{p} - \mathbf{p}_i\|}{2u^2}) \tag{15}$$

where $\mathbf{p}$ denotes the representation of the reasoning path, *V* presents the known amount of reasoning paths, and *u* is a hyper parameter. In all, to balance the impact of each component of the 3D reward on reasoning performance, the reward *R* presents the explicit union of the below rewards, and it is denoted as,

$$R = \lambda_1 R_{destination} + \lambda_2 R_{distance} + \lambda_3 R_{diversity} \tag{16}$$

where λ_i is a discount factor and $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

Policy Network: We design a policy network to drive the interaction between the explicitly defined MDP and MKGs. This policy network inputs the multi-modal complementary features and outputs the next action with the highest executable probability. In fact, the policy network is a feed-forward neural network. Specifically, π selects the promising relation in $\mathcal{A}_t$ with the maximum likelihood, and π is denoted as follow,

$$\pi_\theta(a_t|s_t) = softmax(\mathbf{A}_t(\mathbf{W}_2 \mathrm{ReLu}(Z))) \tag{17}$$

where $\mathcal{A}_t$ can be encoded to $\mathbf{A}_t$ by stacking the representations of existing available actions. Multi-modal complementary features Z are obtained from the unified gate-attention network.

To maximize the accumulated rewards of our model and obtain the optimal policy, the objective function is as follow,

$$J(\theta) = E_{(e_s,r,e_d)\sim\mathcal{G}_f} E_{a_1,\ldots,a_T\sim\pi_\theta}[R(S_T \mid e_s, r)] \tag{18}$$

Finally, the stochastic gradient is used to conduct optimization.

$$\nabla_\theta J(\theta) = \nabla_\theta \sum_{t=1}^{T} R(S_T|e_s, r) log \pi_\theta(a_t|s_t) \tag{19}$$

V. Experiments

A. Experimental Setup

1) Datasets: We report the reasoning results of MMKGR on existing public MKG datasets[1], WN9-IMG-TXT and FB-IMG-TXT. Both of them are MKGs widely adopted by existing reasoning studies [63] [61] [45]. Notably, we completed data cleaning before training. Each entity in these MKGs contains three modal information: structure, image, and text. Specifically, the relation triplets and textual descriptions of two datasets are extracted from WordNet [34] and Freebase [2]. To extract the image features of the entities, 10 images and 100 images are crawled for each entity in WN9-IMG-TXT and FB-IMG-TXT, respectively [45]. Detailed statistics are shown in Table II.

TABLE II
Statistics of the experimental datasets.

Dataset	#Ent	#Rel	#Train	#Valid	#Test
WN9-IMG-TXT	6,555	9	11,747	1,337	1,319
FB-IMG-TXT	11,757	1,231	285,850	29,580	34,863

2) Evaluation Protocol: Entity link prediction [3] and relation link prediction [40] are used to evaluate the performance of MMKGR. For **entity link prediction**, we use the following two metrics. (1) Mean reciprocal rank (MRR) of all correct entities and (2) the proportion of correct entities that rank no larger than N (Hits@N) [27], [32]. For **relation link prediction**, mean average precision (MAP) score is adopted as the metric [56].

[1]https://public.ukp.informatik.tu-darmstadt.de/starsem18-multimodalKB

TABLE III
REASONING PERFORMANCES ON TWO MKGS.

Model	WN9-IMG-TXT				FB-IMG-TXT			
	MRR	Hits@1	Hits@5	Hits@10	MRR	Hits@1	Hits@5	Hits@10
MTRL	48.3	45.6	69.8	83.8	25.2	21.3	32.4	47.2
NeuralLP	41.3	36.5	60.4	80.7	22.1	18.0	25.7	34.8
MINERVA	47.2	43.1	65.6	83.2	23.4	19.2	30.6	43.9
FIRE	56.4	52.8	77.6	86.8	42.8	37.9	49.5	57.1
GAATs	58.2	54.6	79.4	87.7	45.4	41.2	54.3	61.8
RLH	62.4	58.3	81.3	89.4	50.6	44.5	60.2	68.4
MMKGR	**80.2**	**73.6**	**87.8**	**92.8**	**71.3**	**65.8**	**77.5**	**82.6**
Improv.	17.8%	15.3%	6.5%	3.4%	20.7%	21.3%	17.3%	14.2%

TABLE IV
MAP OF RELATION LINK PREDICTION ON WN9-IMG-TXT AND FB-IMG-TXT.

Tasks	MTRL	NeuralLP	MINERVA	FIRE	GAATs	RLH	MMKGR
has_part	65.6	55.2	64.7	75.9	77.8	84.6	**98.2**
derivationally_related	62.7	53.3	60.5	73.6	74.4	80.7	**95.4**
domain_topic	60.2	48.7	58.3	70.1	70.5	78.6	**92.5**
....							
Overall	63.8	54.3	61.6	74.0	75.2	83.4	**97.1**
place_founded	41.5	36.7	40.2	63.5	67.3	70.4	**88.3**
producer_type	50.3	43.6	46.0	65.9	73.1	79.4	**93.2**
registering_agency	45.7	42.4	42.9	64.7	69.8	72.7	**91.6**
....							
Overall	48.7	43.1	45.4	67.8	70.4	74.6	**93.6**

3) Hyper-parameters: In our training stage, some core settings are as follows. The embedding dimension d_s of entity, relation and history is set to 200, the embedding dimension d_i of image feature is set to 128 and 4096 on FB-IMG-TXT and WN9-IMG-TXT respectively, and the embedding dimension d_t of textual feature is 1000 [45]. The maximum reasoning step *T* is 4. The size of batches *N* is 128. The bandwidth u in Eq. (15) is set to 3. λ_1, λ_2, and λ_3 are fixed as 0.1, 0.8, and 0.1, respectively.

B. Baselines

To investigate the performance of MMKGR, two categories of methods are compared: 1) single-hop MKG reasoning methods MTRL [45]; 2) multi-hop reasoning methods MINERVA [8], FIRE [75], GAATs [59], NeuraILP [70] and RLH [56] in traditional KGs.

C. Performance Comparisons

Link prediction results are represented as the above Table III (score is shown in percentage) and IV. We have the following analysis of experiment.

1) Entity Link Prediction: We study entity link prediction from Table III. Firstly, the scores on FB-IMG-TXT are generally lower than those on WN9-IMG-TXT as shown in Table III, which is consistent with previous work. The essential reason is that the dataset FB-IMG-TXT is more sparse and complex than the dataset WN9-IMG-TXT [45] as shown in Table II. In addition, Hits@1 has more improvement than Hits@10 or Hits@5, which indicates that MMKGR tends to rank the ground-truth entity higher and has superior reasoning ability than other models.

Secondly, although multi-modal features are not used, some RL-based methods (e.g., RLH) and the graph neural network-based reasoning method (i.e., GAATs) exceed MTRL that uses multi-modal features in overall performance. There are two potential reasons. On the one hand, the performance of MTRL is limited by TransE-based single-hop model. On the other hand, these novel models make better use of structural data (e.g., neighbor entities) in the multi-hop reasoning process.

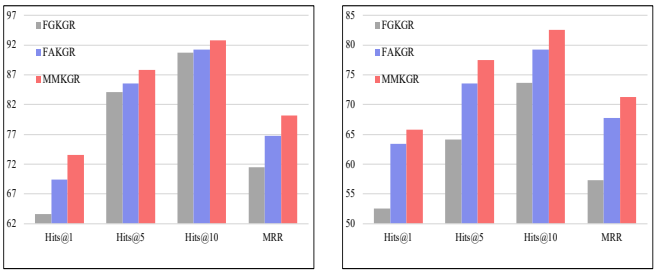
Finally, MMKGR achieves the best reasoning results in all reasoning baselines. Observed from the last row of Table III, the most significant improvements on the two datasets are 17.8% and 21.3%, respectively. In general, the above results prove that our model significantly outperforms existing methods in most metrics.

2) Relation Link Prediction: Relation prediction is shown in Table IV. The upper and lower parts of the table are the MAP scores on WN9-IMG-TXT and FB-IMG-TXT, respectively. We compare the MAP score of each relation (e.g., has_part and place_founded), and then count the overall MAP score (i.e., "Overall") of all relations in the test set. Observed from Table IV, the MAP scores of different models on WN9-IMG-TXT are higher than those on the FB-IMG-TXT, which is similar to the distribution of the experimental results of entity prediction. In additon, the overall improvements of MMKGR compared with RLH on the two datasets are 13.7% and 19.0%, respectively. Regardless of the comparison over each relation or overall relations, the performance of MMKGR surpasses other models.

Based on the experimental results of entity and relation link prediction, we can summarize that MMKGR has made great progress in the multi-modal knowledge graph reasoning. This is because MMKGR makes full use of all modal data

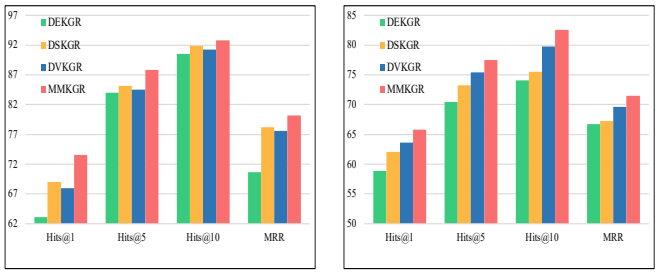
TABLE V
EFFECTS OF DIFFERENT MULTI-MODAL AUXILIARY FEATURES ON ENTITY LINK PREDICTION.

Model	WN9-IMG-TXT				FB-IMG-TXT			
	MRR	Hits@1	Hits@5	Hits@10	MRR	Hits@1	Hits@5	Hits@10
OSKGR	66.0	61.5	82.5	90.5	55.1	47.8	63.1	73.2
STKGR	71.2	65.1	84.6	91.3	60.1	52.3	64.9	75.3
SIKGR	74.7	68.8	85.8	91.9	66.8	59.7	69.4	78.6
MMKGR	**80.2**	**73.6**	**87.8**	**92.8**	**71.3**	**65.8**	**77.5**	**82.6**



(a) WN9-IMG-TXT (b) FB-IMG-TXT

Fig. 4. Ablation on different components of the unified gate-attention network.



(a) WN9-IMG-TXT (b) FB-IMG-TXT

Fig. 5. Ablation on different components of 3D reward mechanism.

through the unified gate-attention network and effectively leverages these data to help multi-hop reasoning without being negatively affected by the sparse reward dilemma.

D. Ablation Studies

1) Impact of Different Components in the Unified Gate-attention Network: We conduct the ablation experiment via separately deleting different MMKGR' elements. (1) **FAKGR**: the irrelevance-filtration module is removed in this variant version. The attended features $\hat{V}$ generated by the attention-fusion module are directly fed into the feature-aware RL framework. (2) **FGKGR**: after multi-modal fusions are completed by Eq. (6), only the irrelevance-filtration module is used to generate features that are fed into the complementary feature-aware RL. This variant version is to evaluate the effectiveness of the well designed attention-fusion module. The results are shown in Fig. 4. Several observations are made from the results. First, MMKGR has the best performance compared with both variant versions. This validates the effectiveness of both the attention-fusion module and the irrelevance-filtration module. Second, the results of FAKGR are closer to those of MMKGR and consistently better than those of FGKGR. The above results indicate that the main improvement of MMKGR comes from the attention-fusion module, and the improvement of the irrelevance-filtration module is relatively small. This is because WN9-IMG-TXT and FB-IMG-TXT have eliminated part of the influence of noise when the multi-modal data are crawled from Web [45]. Third, even though FGKGR performs worse compared with the variant versions of our own models FAKGR and MMKGR, combined with the scores in Table III, we also observe that the scores of FGKGR are better than those of MTRL and RLH (i.e., the state-of-the-art baseline among MKG reasoning and KG reasoning methods). For example, the Hits@1 of FGKGR is about 27.4% and 8.1% higher than that of MTRL and HRL, respectively. This demonstrates that the irrelevance-filtration module makes full use of multi-modal features and positively affects reasoning performance in MKGs.

2) Impact of Different Components in the Reward Function: The 3D reward mechanism can solve the sparse reward problem. It includes three components: the destination reward, the distance reward, and the diverse reward. Note that. the destination reward is indispensable reward setting used to drive the reasoning agent to the target entity. Based on this, we perform the ablation study for the reward function by separately removing different components of the 3D reward mechanism: (1) **DEKGR**, a variant version only leveraging the destination reward as the reward function; (2) **DSKGR**, a variant version where the distance reward is added on the basis of the destination reward; (3) **DVKGR**, in which diverse reward is added on the basis of the destination reward.

In Fig. 5, we first observe that all three variant versions perform worse than MMKGR which demonstrates that both diverse reward and distance reward can improve reasoning performance. Second, the reasoning performance fluctuates when any reward is removed. Furthermore, although DEKGR has the lowest performance among the variants, it still outperforms the state-of-the-art methods MTRL and RLH, which proves the effectiveness of destination reward in MKG reasoning. Finally, the performance of DSKGR and DVKGR is closer to that of MMKGR than the other variants on WN9-IMG-TXT and FB-IMG-TXT, respectively. This experimental result illustrates that more diverse paths need to be explored by diverse reward in larger dataset (i.e., FB-IMG-TXT), while lightweight datasets (i.e., WN9-IMG-TXT) rely more on the distance reward. A potential reason is that there are not enough diverse paths to be inferred in lightweight datasets.

E. Effects of Multi-modal Features

In this subsection, we focus on the effect of different multi-modal auxiliary features over the reasoning result. To evaluate

TABLE VI
HITS@1 OF MMKGR WITH THE CHANGING REASONING STEP T AND DIFFERENT REWARD THRESHOLD OF k.

Hits@1 / Th. \ T	WN9-IMG-TXT					FB-IMG-TXT				
	T=2	T=3	T=4	T=5	T=6	T=2	T=3	T=4	T=5	T=6
2	45.7	69.8	71.8	67.4	64.8	47.9	60.5	62.8	57.8	55.1
3	—	**73.1**	**73.6**	**73.5**	**73.3**	—	**65.3**	**65.8**	**64.9**	**64.1**
4	—	—	72.1	71.5	71.1	—	—	63.3	62.4	61.6
5	—	—	—	71.4	70.8	—	—	—	61.7	61.1
6	—	—	—	—	70.7	—	—	—	—	60.7

TABLE VII
THE PERFORMANCE CHANGE OF MODELS AFTER FUSION OF MULTI-MODAL INFORMATION ON FB-IMG-TXT.

Method	Attention		Concatenation	
	Rewards	Hits@1	Rewards	Hits@1
GAATs	—	-2.1%	—	-3.7%
NeuralLP	—	-3.3%	—	-5.4%
MINERVA	-2.2%	-6.3%	-2.7%	-7.1%
FIRE	-1.8%	-5.9%	-2.3%	-6.5%
RLH	-1.1%	-3.8%	-1.7%	-4.9%

TABLE VIII
COMPARISON OF HITS@1 ON TEST SETS WITH DIFFERENT PROPORTIONS.

Proportion	WN9-IMG-TXT		FB-IMG-TXT	
	MMKGR	OSKGR	MMKGR	OSKGR
20%	**85.6**	74.1	**60.8**	40.2
40%	**75.5**	65.0	**71.8**	59.3
60%	**72.3**	60.4	**68.7**	54.9
80%	**69.4**	60.1	**57.6**	41.1
100%	**73.6**	61.5	**65.8**	47.8

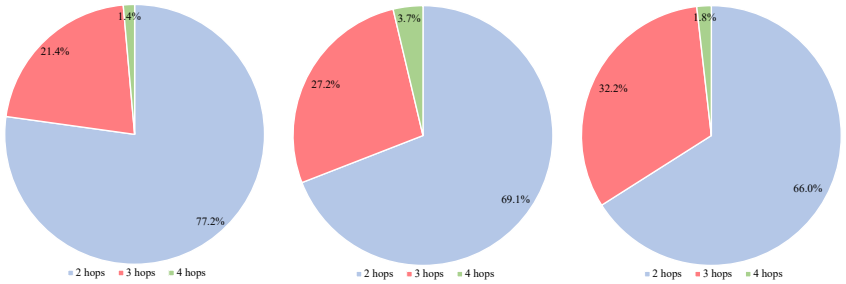
this, we compare three versions where features of a type of modality are removed: (1) **OSKGR**: a version where only structural features are considered in Eq. (17); (2) **STKGR**: a version where image features are not calculated by the unified gate-attention network. (3) **SIKGR**: a version that textual features are not input into the unified gate-attention network. Observed from Table V, OSKGR still outperforms existing multi-hop reasoning methods in Table III, even though the multi-modal features are not added. For example, the MRR of OSKGR is 3.6% and 4.5% higher than that of RLH on both datasets, respectively. This is because OSKGR retaining 3D reward mechanism can eliminate the negative impact of sparse rewards and improve reasoning performance. In addition, the performance of MMKGR is significantly better than that of the other three variant versions which validates the benefit brought by both image and text features and confirms the necessity for including multi-modal data in KG reasoning. Finally, the performance of SIKGR is better than that of STKGR. This is because each entity is connected to a larger number of pictures (i.e., 100 images for each entity on FB-IMG-TXT) containing more useful information.

In fact, performing multi-hop reasoning in MKGs is challenging. To investigate the impact of this challenge on existing multi-hop reasoning models, we conduct the following experimental setup. Specifically, we first utilize Attention and Concatenation (i.e., two multi-modal fusion method derived from existing single-hop methods on MKGs). Then, existing multi-hop reasoning methods (e.g. GAATs) are we combined to conduct multi-hop reasoning on the MKG. As shown in Table VII, when multi-modal data are added to the existing multi-hop reasoning models, all performance percentages of RL-based models decrease compared with the absence of these data. This is because sparse rewards reduce reward accumulation and limit reasoning performance on the MKG. Although non-RL methods (e.g., GAATs) are not affected by the sparse rewards, the Hits@1 still declines compared with the absence of multi-modal data. A reasonable explanation is that existing fusion methods are not suitable for multi-hop reasoning methods.

To further verify the global impact of multi-modal features on multi-hop reasoning, we compare the changes of Hits@1 before and after fusing multi-modal data (i.e., feature fusion settings like OSKGR and MMKGR) on test sets. We first randomly sample different proportions of test data (e.g., 20%), and then use MMKGR and OSKGR to perform reasoning. The experimental results are presented in Table VIII. We can observe that regardless of the proportion of test data, the Hits@1 of MMKGR is much larger than that of OSKGR. This expected result once again proves that multi-modal auxiliary features can help the model to improve the reasoning performance. Furthermore, although it is challenging to integrate multi-modal auxiliary features with multi-hop reasoning, MMKGR is able to stably and effectively utilize these features.

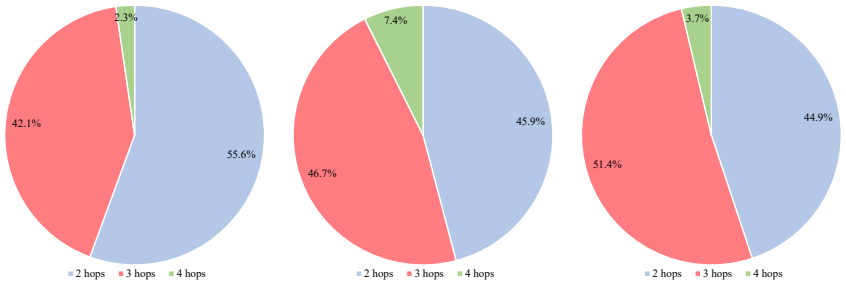
F. Studies for Multi-hop Paths

In this section, we study the factors that affect multi-hop paths. The first is the effect of distance reward in MMKGR. To investigate the plausibility that the threshold of k is set to 3 in the distance reward (i.e., Eq.(14)), we present corresponding experimental results in Table VI. The horizontal line represents the null value because the threshold cannot be greater than the maximum reasoning step T. Observed from Table VI, MMKGR achieves the best reasoning performance on both datasets when threshold of k is set to 3. Note that when k is the same as T, there is no negative penalty term in Eq. (14). In addition, Hits@1 has the fastest growth rate when the maximum reasoning step T=3 for all models as shown in Fig. 8. Thus, we believe that long reasoning paths are also not required on MKGs. After T>4, the change of Hits@1 is relatively stable on the small-scale WN9-IMG-TXT, because most of the test triplets are successfully inferred before T=4.



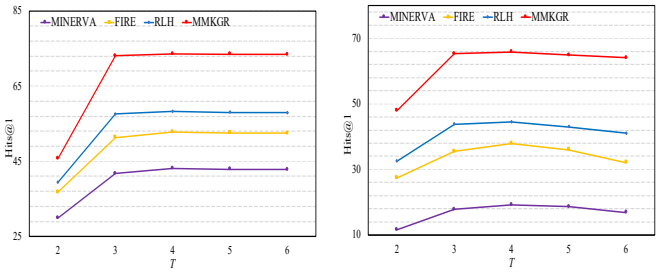
(a) MMKGR (b) DVKGR (c) OSKGR

Fig. 6. The proportions of test triplets successfully inferred by different path lengths on WN9-IMG-TXT.



(a) MMKGR (b) DVKGR (c) OSKGR

Fig. 7. The proportions of test triplets successfully inferred by different path lengths on FB-IMG-TXT.



(a) WN9-IMG-TXT (b) FB-IMG-TXT

Fig. 8. Hits@1 of the changing reasoning step T for RL-based models.

(a) WN9-IMG-TXT (b) FB-IMG-TXT

Fig. 9. The convergence rate of different methods.

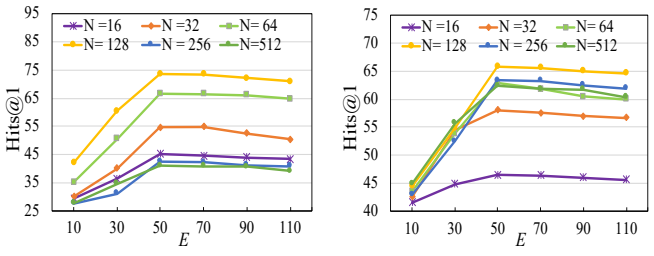
Furthermore, the Hits@1 of all models gradually decreases when T>4 on FB-IMG-TXT. One potential reason is that some test triples negatively affected by noise are inferred incorrectly after T>4. Although the number of reasoning step exceeding the threshold 3 may degrade reasoning performance as mentioned in Section IV-C, we can observe that reasoning performances of all models are still slightly improved when T=4. Therefore, T is set to 4 in this study.

As the first work for performing multi-hop reasoning in MKGs, it is necessary to investigate the affect of multi-modal data on the path length of MMKGR. For this purpose, we compare MMKGR with the ablated model OSKGR that only utilizes structural data. As another variable affecting reasoning length in this study, the role of the distance reward that encourages the agent to infer the target within 3 hops also needs to be investigated. Thus, we also compare MMKGR with the ablated model DVKGR that removes the distance reward from 3D reward in this subsection. The different proportions of triplets in test sets are inferred with different hops in Fig. 6 and 7, based on which we have the following two observations. (1) the proportion of 4 hops in DVKGR is the largest compared with MMKGR on both datasets. This shows that distance reward can further encourage the agent to find the target entity within the 3 hops most relevant to the query. (2) MMKGR has more 2-hop ratio and less 3-hop ratio than OSKGR. This indicates that multi-modal data provide beneficial reasoning clues to improve reasoning performance.

G. Convergence Rate Analysis

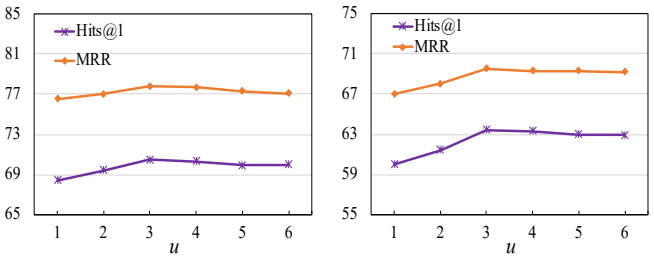
We investigate the convergence rate of our proposed model in Fig. 9, where only the effect of 3D reward mechanism on convergence rate is presented, due to the space limitation. A new variant of MMKGR named ZOKGR is added. Notably, in ZOKGR, the 3D reward is completely replaced by the "0-1 reward" that is commonly used in existing RL-based reasoning methods (e.g. MINERVA and RLH).

Observed from Fig. 9, ZOKGR fluctuates greatly and has not converged. A reliable explanation is that the "0-1 reward" is easy to fall into the dilemma of sparse reward in MKGs. It is



(a) WN9-IMG-TXT (b) FB-IMG-TXT

Fig. 10. The Hits@1 of MMKGR w.r.t. varied epoch E and batch size N.



(a) WN9-IMG-TXT (b) FB-IMG-TXT

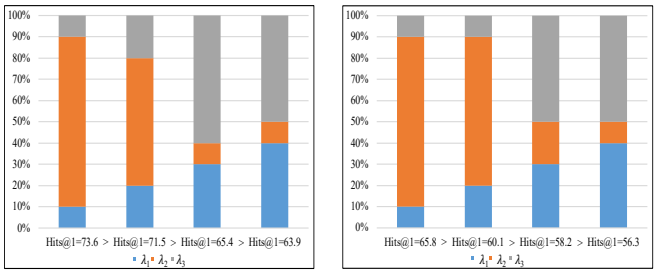
Fig. 11. Performance of MMKGR w.r.t. varied u on different datasets.

interesting that the MRR value of DEKGR on FB-IMG-TXT is closer to that of MMKGR. A potential reason is that our destination reward based on reward shaping plays a pivotal role in convergence rate on large datasets that are more likely to result in sparse rewards. In addition, a fair concern for diverse reward is that it can slow down the convergence rate, since this reward encourages the agent to explore a more diverse set of paths. The good convergence of DVKGR eliminates the above concern. The convergence rates of DSKGR and DVKGR are both greater than that of DEKGR, which shows that both the distance reward and the diverse reward can accelerate the convergence. In a word, the carefully designed 3D reward mechanism can boost reasoning performance and increase convergence rate on all datasets.

H. Parameter Interpretability

1) Impact of Different Epochs E and Batch Sizes N : We investigate the impact of different epochs E and batch sizes N in Fig. 10. Here, we set the number of epochs $E \in$ [10, 30, 50, 70, 90, 110] and the size of batches $N \in$ [16, 32, 64, 128, 256, 512]. We can observe that the performance of MMKGR will increase firstly and then decreases steadily in most cases with the increase of E and N. This is because the under-training and over-fitting have a negative impact on the model. The results suggest that a suitable scope of train parameters can improve the reasoning effectiveness of models. Consequently, the best parameters are chosen as E = 50 and N = 128.

2) Impact of Different Bandwidths u: We evaluate the influence of bandwidth u on MMKGR. From the results in Fig. 11, we can observe that 3 is the optimal value of u on both two datasets. If the value of u exceeds 3, the performance will be relatively stable. One potential reason is that the range of the local influence of the Gaussian kernel function increases



(a) WN9-IMG-TXT (b) FB-IMG-TXT

Fig. 12. Performance of MMKGR w.r.t. varied discount factor.

as the bandwidth value increases. Beyond this range, the value of the kernel function is almost unchanged, and the impact on the reward becomes stable.

3) Impact of Different Discount Factors λ: In this study, we linearly combine destination reward, distance reward, and diverse reward to form our 3D mechanism. The discount factor is equivalent to the combined weight of the reward. Since the destination reward will return a relatively large reward value (i.e., 1) when it predicts the correct entity, a smaller discount factor of destination reward is an appropriate choice. Otherwise, false positive target entities will be generated. Thus, the value of λ_1 is set to some small values, i.e., 0.1, 0.2, 0.3, 0.4. Fig. 12 shows the highest Hits@1 in different number λ combinations. We can observe that the optimal values of λ_1, λ_2, and λ_3 on different datasets are 0.1, 0.8, and 0.1 respectively. In addition, the increase of λ_1 will lead to the increase of λ_3 in a combination. This is because high reward for reaching the target entity will make it easier to fall into the local optimal path, and the diverse reward can solve the problem.

VI. Conclusion And Future Work

In this work, we study the problem *how to effectively leverage multi-modal auxiliary features to complete multi-hop KG reasoning*, which is an unexplored problem. An effective model MMKGR is proposed, its reasoning results outperform the SOTA baselines in public MKG datasets. In MMKGR, we first perform feature extraction and multi-modal fine-grained fusion for structural data and multi-modal auxiliary data. In addition, a novel attention-based representation method is used to produce multi-modal complementary representation. Next, these features are fed into a complementary feature-aware RL framework to predict the missing elements. Extensive experiments demonstrate the rationality and effectiveness of MMKGR. In future work, we would like to study few-shot reasoning and inductive reasoning to predict unseen entities in MKGs. How to infer missing triplets over neighbor structure on MKGs, still awaits further exploration.

VII. Acknowledgment

This work was supported by the National Natural Science Foundation of China No. 61902270, the Major Program of the Natural Science Foundation of Jiangsu Higher Education Institutions of China No. 19KJA610002, Australian Research Council Nos.FT210100624, DP190101985.

REFERENCES

[1] B. O. Ayinde and J. M. Zurada, "Building efficient convnets using redundant feature pruning," *CoRR*, vol. abs/1802.07653, pp. 1–9, 2018.

[2] K. D. Bollacker, C. Evans, P. K. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *SIGMOD*, 2008, pp. 1247–1250.

[3] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *NeurIPS*, 2013, pp. 2787–2795.

[4] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[5] D. S. Chaplot, L. Lee, R. Salakhutdinov, D. Parikh, and D. Batra, "Embodied multimodal multitask learning," in *IJCAI*, 2020, pp. 2442–2448.

[6] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014, pp. 1–11.

[7] X. Chen, S. Jia, and Y. Xiang, "A review: Knowledge reasoning over knowledge graph," *Expert Systems With Applications*, vol. 141, pp. 1–21, 2020.

[8] R. Das, S. Dhuliawala, M. Zaheer, L. Vilnis, I. Durugkar, A. Krishnamurthy, A. Smola, and A. McCallum, "Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning," in *ICLR*, 2018, pp. 1–16.

[9] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2d knowledge graph embeddings," in *AAAI*, 2018, pp. 1811–1818.

[10] R. Devidze, G. Radanovic, P. Kamalaruban, and A. Singla, "Explicable reward design for reinforcement learning agents," in *NeurIPS*, 2021, pp. 20 118–20 131.

[11] D. Dewey, "Reinforcement learning and the reward engineering principle," in *AAAI*, 2014, pp. 1–7.

[12] S. Di, Q. Yao, Y. Zhang, and L. Chen, "Efficient relation-aware scoring function search for knowledge graph embedding," in *ICDE*, 2021, pp. 1104–1115.

[13] S. Ferrada, B. Bustos, and A. Hogan, "Imgpedia: A linked dataset with content-based analysis of wikimedia images," in *ISWC*, vol. 10588, 2017, pp. 84–93.

[14] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *EMNLP*, 2016, pp. 457–468.

[15] H. Hu, D. Yarats, Q. Gong, Y. Tian, and M. Lewis, "Hierarchical decision making by generating and following natural language instructions," in *NeurIPS*, 2019, pp. 10 025–10 034.

[16] L. Huang, W. Wang, J. Chen, and X. Wei, "Attention on attention for image captioning," in *ICCV*, 2019, pp. 4633–4642.

[17] N. Q. V. Hung, C. T. Duong, T. T. Nguyen, M. Weidlich, K. Aberer, H. Yin, and X. Zhou, "Argument discovery via crowdsourcing," *VLDB J.*, vol. 26, no. 4, pp. 511–535, 2017.

[18] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 2, pp. 494–514, 2022.

[19] Z. Jiang, J. Gao, and X. Lv, "Metap: Meta pattern learning for one-shot knowledge graph completion," in *SIGIR*, 2021, pp. 2232–2236.

[20] M. Kaiser, R. S. Roy, and G. Weikum, "Reinforcement learning from reformulations in conversational question answering over knowledge graphs," in *SIGIR*, 2021, pp. 459–469.

[21] J. Kim, K. W. On, W. Lim, J. Kim, J. Ha, and B. Zhang, "Hadamard product for low-rank bilinear pooling," in *ICLR*, 2017, pp. 1–14.

[22] P. Li, H. Liu, Y. Si, C. Li, and F. Li, "EEG based emotion recognition by combining functional connectivity network and local activations," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 10, pp. 2869–2881, 2019.

[23] R. Li and X. Cheng, "DIVINE: A generative adversarial imitation learning framework for knowledge graph reasoning," in *EMNLP*, 2019, pp. 2642–2651.

[24] X. Li, J. Song, L. Gao, X. Liu, and W. Huang, "Beyond rnns: Positional self-attention with co-attention for video question answering," in *AAAI*, 2019, pp. 8658–8665.

[25] X. Li, Y. Wang, and R. Ruiz, "A survey on sparse learning models for feature selection," *IEEE Trans. Cybern.*, vol. 52, no. 3, pp. 1642–1660, 2022.

[26] Y. Li, T. Ge, and C. X. Chen, "Online indices for predictive top-k entity and aggregate queries on knowledge graphs," in *ICDE*, 2020, pp. 1057–1068.

[27] X. V. Lin, R. Socher, and C. Xiong, "Multi-hop knowledge graph reasoning with reward shaping," in *EMNLP*, 2018, pp. 3243–3253.

[28] Y. Liu, H. Li, A. García-Durán, M. Niepert, D. Oñoro-Rubio, and D. S. Rosenblum, "MMKG: multi-modal knowledge graphs," in *ESWC*, vol. 11503, 2019, pp. 459–474.

[29] Z. Liu, C. Xiong, M. Sun, and Z. Liu, "Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval," in *ACL*, 2018, pp. 2395–2405.

[30] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *CVPR*, 2017, pp. 3242–3250.

[31] M. Luggen, J. Audiffren, D. E. Difallah, and P. Cudré-Mauroux, "Wiki2prop: A multimodal approach for predicting wikidata properties from wikipedia," in *WWW*, 2021, pp. 2357–2366.

[32] X. Lv, Y. Gu, X. Han, L. Hou, J. Li, and Z. Liu, "Adapting meta knowledge graph information for multi-hop reasoning over few-shot relations," in *EMNLP-IJCNLP*, 2019, pp. 3374–3379.

[33] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR*, 2013, pp. 1–12.

[34] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[35] D. K. Misra, J. Langford, and Y. Artzi, "Mapping instructions and visual observations to actions with reinforcement learning," in *EMNLP*, 2017, pp. 1004–1015.

[36] D. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," in *CVPR*, 2018, pp. 6087–6096.

[37] T. T. Nguyen, M. Weidlich, D. C. Thang, H. Yin, and N. Q. V. Hung, "Retaining data from streams of social platforms with minimal regret," in *IJCAI*, 2017, pp. 2850–2856.

[38] M. Nickel, L. Rosasco, and T. A. Poggio, "Holographic embeddings of knowledge graphs," in *AAAI*, 2016, pp. 1955–1961.

[39] M. Nickel, V. Tresp, and H. Kriegel, "A three-way model for collective learning on multi-relational data," in *ICML*, 2011, pp. 809–816.

[40] G. Niu, Y. Li, C. Tang, R. Geng, J. Sun, F. Huang, and L. Si, "Relational learning with gated and attentive neighbor aggregator for few-shot knowledge graph completion," in *SIGIR*, 2021, pp. 213–222.

[41] P. Pezeshkpour, L. Chen, and S. Singh, "Embedding multimodal relational data for knowledge base completion," in *EMNLP*, 2018, pp. 3208–3218.

[42] T. Rahman, S. Chou, L. Sigal, and G. Carenini, "An improved attention for visual question answering," in *CVPR*, 2021, pp. 1653–1662.

[43] J. Ren, S. Guo, and F. Chen, "Orientation-preserving rewards' balancing in reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021.

[44] M. S. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *ESWC*, vol. 10843, 2018, pp. 593–607.

[45] H. M. Sergieh, T. Botschen, I. Gurevych, and S. Roth, "A multimodal translation-based approach for knowledge graph representation learning," in *NAACL*, 2018, pp. 225–234.

[46] Y. Shen, J. Chen, P. Huang, Y. Guo, and J. Gao, "M-walk: Learning to walk over graphs using monte carlo tree search," in *NeurIPS*, 2018, pp. 6787–6798.

[47] D. Silver, S. Singh, D. Precup, and R. S. Sutton, "Reward is enough," *Artif. Intell.*, vol. 299, pp. 103 535–103 548, 2021.

[48] G. Stoica, O. Stretcu, E. A. Platanios, and T. M. Mitchell, "Contextual parameter generation for knowledge graph link prediction," in *AAAI*, 2020, pp. 3000–3008.

[49] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, and K. Zheng, "Multi-modal knowledge graphs for recommender systems," in *CIKM*, 2020, pp. 1405–1414.

[50] X. Tang, L. Chen, J. Cui, and B. Wei, "Knowledge representation learning with entity descriptions, hierarchical types, and textual relations," *Inf. Process. Manag.*, vol. 56, no. 3, pp. 809–822, 2019.

[51] A. S. Therrien, D. M. Wolpert, and A. J. Bastian, "Effective reinforcement learning following cerebellar damage requires a balance between exploration and motor noise," *Brain*, vol. 139, no. 1, pp. 101–114, 2016.

[52] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *ICML*, vol. 48, 2016, pp. 2071–2080.

[53] Y. Uo, Q. Fang, S. Qian, X. Zhang, and C. Xu, "Representation learning of knowledge graphs with entity attributes and multimedia descriptions," in *BigMM*, 2018, pp. 1–5.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[55] G. Wan and B. Du, "Gaussianpath: A bayesian multi-hop reasoning framework for knowledge graph reasoning," in *AAAI*, 2021, pp. 4393–4401.

[56] G. Wan, S. Pan, C. Gong, C. Zhou, and G. Haffari, "Reasoning like human: Hierarchical reinforcement learning for knowledge graph reasoning," in *IJCAI*, 2020, pp. 1926–1932.

[57] M. Wang, H. Wang, G. Qi, and Q. Zheng, "Richpedia: A large-scale, comprehensive multi-modal knowledge graph," *Big Data Res.*, vol. 22, pp. 159–170, 2020.

[58] Q. Wang, H. Yin, W. Wang, Z. Huang, G. Guo, and Q. V. H. Nguyen, "Multi-hop path queries over knowledge graphs with neural memory networks," in *DASFAA*, vol. 11446, 2019, pp. 777–794.

[59] R. Wang, B. Li, S. Hu, W. Du, and M. Zhang, "Knowledge graph embedding via graph attenuated attention networks," *IEEE Access*, vol. 8, pp. 5212–5224, 2019.

[60] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[61] Z. Wang, L. Li, Q. Li, and D. Zeng, "Multimodal data enhanced representation learning for knowledge graphs," in *IJCNN*, 2019, pp. 1–8.

[62] Y. Xian, Z. Fu, S. Muthukrishnan, G. de Melo, and Y. Zhang, "Reinforcement knowledge graph reasoning for explainable recommendation," in *SIGIR*, 2019, pp. 285–294.

[63] R. Xie, S. Heinrich, Z. Liu, C. Weber, Y. Yao, S. Wermter, and M. Sun, "Integrating image-based and knowledge-based representation learning," *IEEE Trans. Cogn. Dev. Syst.*, vol. 12, no. 2, pp. 169–178, 2020.

[64] R. Xie, Z. Liu, J. Jia, H. Luan, and M. Sun, "Representation learning of knowledge graphs with entity descriptions," in *AAAI*, 2016, pp. 2659–2665.

[65] R. Xie, Z. Liu, H. Luan, and M. Sun, "Image-embodied knowledge representation learning," in *IJCAI*, 2017, pp. 3140–3146.

[66] J. Xiong, G. Liu, Y. Liu, and M. Liu, "Oracle bone inscriptions information processing based on multi-modal knowledge graph," *Comput. Electr. Eng.*, vol. 92, pp. 173–186, 2021.

[67] W. Xiong, T. Hoang, and W. Y. Wang, "Deeppath: A reinforcement learning method for knowledge graph reasoning," in *EMNLP*, 2017, pp. 564–573.

[68] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, and R. S. Zemel, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, vol. 37, 2015, pp. 2048–2057.

[69] B. Yang, W. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in *ICLR*, 2015, pp. 1–12.

[70] F. Yang, Z. Yang, and W. W. Cohen, "Differentiable learning of logical rules for knowledge base reasoning," in *NeurIPS*, 2017, pp. 2319–2328.

[71] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," in *CVPR*, 2016, pp. 21–29.

[72] J. Yu, H. Yin, J. Li, M. Gao, Z. Huang, and L. Cui, "Enhance social recommendation with adversarial graph convolutional networks," *CoRR*, vol. abs/2004.02340, 2020.

[73] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *CVPR*, 2019, pp. 6281–6290.

[74] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 12, pp. 5947–5959, 2018.

[75] C. Zhang, L. Yu, M. Saebi, M. Jiang, and N. V. Chawla, "Few-shot multi-hop relation reasoning over knowledge bases," in *EMNLP*, 2020, pp. 580–585.

[76] J. Zhang, T. Zhao, and Z. Yu, "Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog," in *SIGDIAL*, 2018, pp. 140–150.

[77] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets," in *AAAI*, 2018, pp. 5674–5681.

[78] Y. Zhang, Q. Yao, W. Dai, and L. Chen, "Autosf: Searching scoring functions for knowledge graph embedding," in *ICDE*, 2020, pp. 433–444.

[79] Y. Zhang, Q. Yao, Y. Shao, and L. Chen, "Nscaching: Simple and efficient negative sampling for knowledge graph embedding," in *ICDE*, 2019, pp. 614–625.

[80] K. Zhao, Y. Zhang, H. Yin, J. Wang, K. Zheng, X. Zhou, and C. Xing, "Discovering subsequence patterns for next poi recommendation," in *IJCAI*, 2020, pp. 3216–3222.

[81] S. Zheng, W. Chen, P. Zhao, A. Liu, J. Fang, and L. Zhao, "When hardness makes a difference: Multi-hop knowledge graph reasoning over few-shot relations," in *CIKM*, 2021, pp. 2688–2697.