Image Attribute Completion: A Novel Task for Multi-modal Knowledge Graph

Yirui Ma, Qian Zhou, Wei Chen(✉), Xi Chen, Xiaofang Zhang and Lei Zhao

School of Computer Science and Technology, Soochow University, Suzhou, China
{20234227035,qzhou1}@stu.suda.edu.cn
{robertchen,chenxi10,xfzhang,zhaol}@suda.edu.cn

Abstract. Multi-modal knowledge graphs (MMKGs) present distinct advantages over unimodal knowledge graphs by leveraging diverse modal attributes for more comprehensive entity descriptions. Most existing studies on multi-modal knowledge graph completion focus on link prediction and numerical attribute completion, neglecting another significant issue, i.e., the incompleteness of image attribute. To fill the gap, this paper formulates a novel task **I**mage **A**ttribute **C**ompletion, and develops a dedicated model entitled **MAGIC** (Multi-modal Assisted imaGe attrIbutes Completion). Specifically, the model consists of two main components: (1) The module EEM is designed to Encode Entity with Different Modalities (i.e., structural information, discrete attributes, text, and images) by employing distinct encoders to learn modality-specific embeddings. Then, these representations are dynamically fused by calculating the entropy of each modality. (2) Based on the fused embeddings of entities, the module EIA is introduced to perform Entity-Image Alignment, which is achieved through a two-step process. First, we calculate the similarity between entities and candidate images. Second, two loss functions are combined to simultaneously enhance modality alignment and classification accuracy. The experiments conducted on three real-world datasets demonstrate the superiority of our proposed model.

Keywords: Multi-modal Knowledge Graph · Knowledge Graph Completion · Image Attribute Completion.

1 INTRODUCTION

Knowledge graphs (KGs) are represented with structural data in the form of relation triples, and have been widely used in various downstream tasks, such as recommendation systems [5, 23] and question-answering systems [9, 29]. Despite the great contributions made by traditional work, relying only on structural information misses the significant complementary information (e.g., images and text). To tackle this problem, recent studies have increasingly focused on multi-modal knowledge graphs (MMKGs) [16, 30].

Although MMKGs incorporate diverse modality information, they still suffer from the problem of incompleteness. To address the issue, the tasks of multi-modal entity alignment (MMEA) [16, 14] and multi-modal knowledge graph completion (MKGC) [12, 15, 33] have been proposed. Notably, while most current

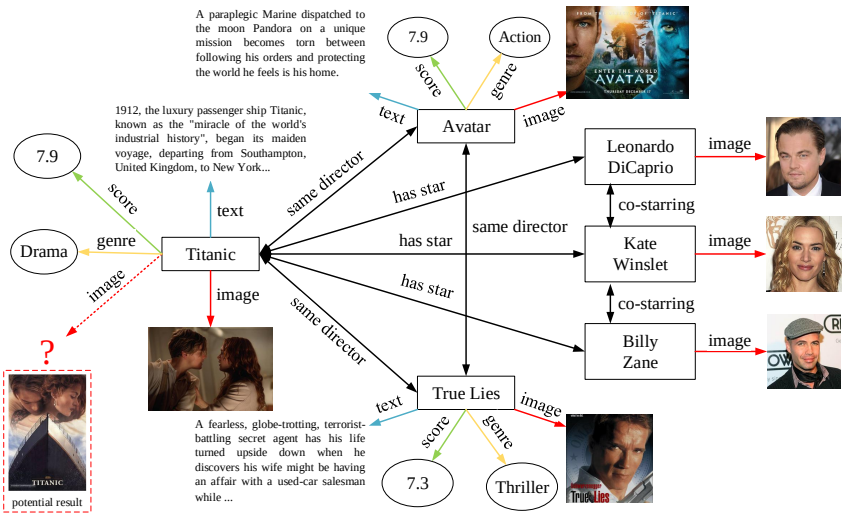


Fig. 1: An example of image attribute completion. Rectangles represent entities, black arrows denote relations, green arrows stand for numerical attributes, blue arrows depict text attributes, yellow arrows express discrete attributes, red arrows symbolize image attributes, and red dashed arrows represent the missing image attribute to be predicted.

MKGC work primarily investigates the completion of relation triples, there exist many attribute triples in knowledge graphs [27]. Intuitively, the main difference between these two types of triples is that: the former denotes different neighbors of an entity, while the latter represents different attributes of an entity in the form of text, numeric, image, etc. The existing efforts on attribute completion have focused either on completing numerical attributes of entities [21, 11, 2, 27, 26], or completing the descriptions of entities' images [30]. There has been no work on completing image attributes for entities, despite their crucial role in enhancing MMKGs. As stated in [18], high-quality MMKGs usually contain more image attributes that provide valuable visual information to enhance entity characteristics. Embedding entities with image attributes in MMKGs can significantly benefit many downstream tasks, e.g., entity alignment [6], knowledge reasoning [34], and link prediction [15].

Having observed the shortcomings of existing work, we formulate a novel task in this paper, i.e., **I**mage **A**ttribute **C**ompletion (IAC), where both of the following two cases are explored: entities that have no image or contain few-shot images. Notably, entities with abundant images are not considered, since they already provide sufficient visual information for downstream tasks. To illustrate the difference between IAC and existing work (i.e., numerical attribute completion) more clearly, an example is presented in Fig. 1. In previous studies [21, 11, 2], the score of *Titanic* can be inferred using existing attribute triples in the knowledge graph, such as $(Avatar, score, 7.9)$ and $(True\ Lies, score, 7.3)$,

i.e., complete the numerical attribute for $(Titanic, score, ?)$. Technically, existing methods achieve numerical attribute completion by propagating known numerical values to missing ones with regression functions [2, 26]. Specifically, these methods first assign different weights to the scores of *True Lies* and *Avatar*, then calculate a weighted average to approximate the true score of *Titanic*. Different from the continuous numerical attributes, the non-continuous image attributes cannot be inferred by directly applying regression operations. Taking the retrieval of the potential missing image for *Titanic*, i.e., $(Titanic, image, ?)$, as an example, the reasoning approaches in numerical attribute completion may incorrectly match *Titanic* with the images of *Avatar* or other films with same stars. This misalignment underscores the limitation of existing approaches and highlights the necessity for a novel solution tailored to IAC. However, the following problems significantly complicate the development of such a solution. *(i) How to extract and fuse complementary information from diverse modalities to enhance entity representation*? Different from previous studies, which explore numerical attribute completion in unimodal knowledge graph, investigating the IAC task in MMKGs inevitably suffers from the problem of effectively utilizing multi-modal information to obtain high-quality entity embedding. *(ii) How to accurately align the representation of an entity with a candidate image*? To ensure high accuracy, the IAC task, unlike numerical attribute completion that disregards multi-modal information, naturally requires aligning entities and candidate images within a shared space, since they have different modalities.

To tackle above-mentioned problems effectively, we develop a novel model **MAGIC** (Multi-modal Assisted imaGe attrIbutes Completion), which consists of two modules, i.e., EEM (Encode Entity with Different Modalities) and EIA (Entity-Image Alignment). To address problem (i), the module EEM first processes each modality with a dedicated encoder to obtain modality-specific representations for an entity. Specifically, various pre-trained encoders and the relation-aware network are utilized to embed multi-modal information and capture the graph structural information respectively. Then, to generate a unified entity representation, a dynamic modality fusion mechanism is employed, where the entropy of each modality is computed. The modalities with higher entropy are assigned lower weights, as greater uncertainty within a modality increases the risk of inaccurate predictions [32]. To tackle problem (ii), the module EIA achieves the alignment of an entity and the accurate candidate image in two steps. At first, EIA embeds candidate images with the same method employed for embedding image attributes in EEM, ensuring consistency in data representation. After embedding candidate images and entities within a shared space, EIA utilizes binary loss to minimize the discrepancies between predictions and true labels, along with contrastive loss to bridge the gap between different modalities. In summary, the main contributions of this paper are as follows:

- To the best of our knowledge, we are the first to propose the novel task (i.e., image attribute completion) for multi-modal knowledge graphs.
- To tackle the novel task effectively, we propose an end-to-end model MAGIC that consists of the following two modules. The module EEM integrates

multi-modal and graph structural information of entities. Based on the entities' representations learned in EEM, the module EIA achieves alignment between images and entities.
- The experimental results demonstrate the superiority of our proposed model, which exhibits state-of-the-art performance across three real-world datasets.

The remainder of this paper is organized as follows. Section 2 reviews the existing work related to our study. Section 3 introduces the key terms and formulates the task IAC. In Section 4, we introduce the details of the proposed model MAGIC. The experimental results are reported in Section 5, and the paper is concluded in Section 6.

2 RELATED WORK

The two research areas: knowledge graph completion (KGC) and knowledge graph attribute completion (KGAC), are closely related to our work. Specifically, the KGC task aims at inferring the missing elements in relation triples. The KGAC task mainly focuses on predicting missing numerical attribute values for entities within KGs. The detailed efforts on them are introduced as follows.

2.1 Knowledge Graph Completion

Traditional KGC methods typically embed entities and relationships into continuous vector spaces. They can be classified into three main categories: translation-based approaches [3, 20], semantic matching methods [1, 22], and neural network-based models [8, 24]. In detail, translation-based approaches learn entity and relation embeddings by treating relations as translation operators, thereby obtaining the semantic and structural information in KGs. Semantic matching methods treat the entire KG as a three-dimensional adjacency matrix (third-order tensor), where each smaller unit tensor represents a triple of binary relational knowledge. Neural network-based solutions approach KGC as a conventional deep learning task. Despite their remarkable successes, the inherent incompleteness of traditional KGs limits the applicability of these studies. To address this, multi-modal knowledge graph completion (MKGC) is proposed, leveraging diverse data, such as images and text.

The work on MKGC can be broadly divided into two categories: fine-tuning (FT)-based and embedding-based methods. In detail, FT-based approaches utilize pre-trained Transformer models like BERT [7] and VisualBERT [13] to transform triples into token sequences for addressing the MKGC task. Embedding-based methods can be further categorized into modality fusion, modality ensemble, and negative sampling approaches. To be specific, modality fusion methods integrate multi-modal and structural embeddings of entities for triple plausibility estimation [12]. Modality ensemble approaches train separate models for different modalities, and combine them for link prediction [15]. Negative sampling methods generate false triples to improve the model's ability to distinguish

between positive and negative samples [25]. Current MKGC studies typically assume images are readily available, overlooking the scenarios where entities may suffer from insufficient image data. To address this issue, image attribute completion becomes a crucial task.

2.2 Knowledge Graph Attribute Completion

The KGAC task mainly focuses on numerical attribute completion, first introduced in [21], where an end-to-end multi-task neural network is proposed to perform regression-typed predictions. Subsequently, the work [11] formally defines the task of predicting entities' numerical attributes in KGs, while utilizing a K-Nearest Neighbors (KNN) graph to propagate known attribute values to the unknown ones. Following this, a multi-relational propagation (MRAP) method [2] is introduced to enable the transmission of information across various types of attributes and relationships by constructing regression functions. To capture the influence of higher-order neighbors, HOAP [26] is developed to integrate information not only from first-order neighbors but also second-order neighbors. Additionally, the study [27] incorporates language models into numerical attribute prediction for KGs, thereby enhancing the prediction accuracy with full utilization of both structural and semantic information. Due to the continuous characteristics of numerical values, numerical attribute completion is often treated as a standard regression task. Different from previous studies, the recent advance [30] proposes a new perspective for completing aspect descriptions of images. It constructs an image contrast learning model that can accurately select aspect-relevant images. Despite this, it only provides descriptions for images without enhancing the visual information of the entities themselves. Therefore, developing new approaches to enrich entity information through image attribute completion is crucial to overcoming this limitation.

3 TASK DEFINITION

In this section, we provide formal definitions for the key terms and the proposed task. The details of them are illustrated as follows.

Definition 1: **Multi-modal Knowledge Graph (MMKG)**. An MMKG is defined as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{M}\}$, where $\mathcal{E} = \{e_1, \ldots, e_{|\mathcal{E}|}\}$ indicates the entity set, $\mathcal{R} = \{r_1, \ldots, r_{|\mathcal{R}|}\}$ denotes the relation set, and $\mathcal{T} = \{(h, r, t)|h \in \mathcal{E}, r \in \mathcal{R}\}$ is the relation triple set. $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_{|\mathcal{E}|}\}$ is a superset, where $\mathcal{M}_i$ represents the multi-modal attribute set of entity e_i.

Definition 2: **Multi-modal Attribute Set** $\mathcal{M}$. For each $\mathcal{M}_i = \{\mathcal{P}, \mathcal{C}, \mathcal{D}\}$ corresponding to e_i in $\mathcal{M}$, it consists of an image set $\mathcal{P} = \{p_1, \ldots, p_{n_\mathcal{P}}\}$, a text set $\mathcal{C} = \{c_1, \ldots, c_{n_\mathcal{C}}\}$, and a discrete attribute set $\mathcal{D} = \{d_1, \ldots, d_{n_\mathcal{D}}\}$, where $n_\mathcal{P}$, $n_\mathcal{C}$, and $n_\mathcal{D}$ indicate the number of attributes in each set. By way of illustration, in Fig. 1, *Drama* in (*Titanic, genre, Drama*) and *Action* in (*Avatar, genre, Action*) are typical discrete attributes. For simplicity, we use $\mathcal{M}_i^A$ to represent a specific modal attribute set of e_i, where $A \in \{\mathcal{P}, \mathcal{C}, \mathcal{D}\}$.

Definition 3: **Image Attribute Completion (IAC)**. Given an entity e_i, a multi-modal knowledge graph $\mathcal{G}$, and a candidate image set $\tilde{\mathcal{P}}$, the IAC task aims to learn a function Φ that identifies a potential missing image $p_j \in \tilde{\mathcal{P}}$, based on the relation triples and multi-modal information in $\mathcal{G}$. The formulation of the problem is defined as:

$$\Phi(e_i, \mathcal{T}, \mathcal{M}_i, \tilde{\mathcal{P}}) \rightarrow (e_i, p_j) \tag{1}$$

where (e_i, p_j) denotes e_i possesses an image p_j.

4 METHODOLOGY

In this section, we propose an image attribute completion model MAGIC to predict an entity's potential missing image based on the existing information in MMKG. The overall architecture of the model is depicted in Fig. 2. To be specific, EEM employs a dynamic fusion strategy to aggregate an entity's modality-specific representations obtained from different encoders. EIA aligns entities and candidate images in a shared space by fully extracting the features of candidate images, along with enhancing modality alignment and classification accuracy through two loss functions. Based on these two modules, MAGIC effectively integrates multi-modal and structural information within MMKG, enabling it to achieve high performance on the IAC task.

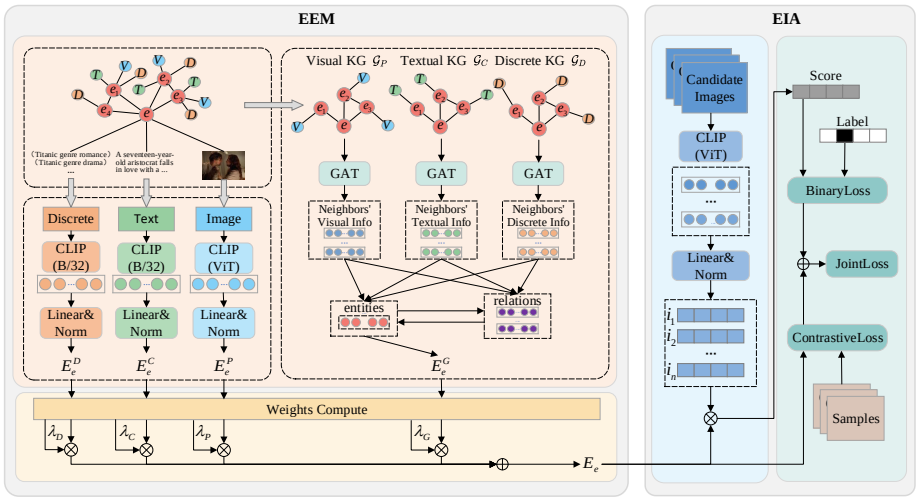


Fig. 2: Overall architecture of the proposed model MAGIC, which consists of two main modules, i.e., EEM (Encode Entity with Different Modalities) and EIA (Entity-Image Alignment).

4.1 Encode Entity with Different Modalities

The purpose of EEM is to obtain a unified representation of an entity. By simultaneously learning both the auxiliary multi-modal and graph structural information, we enhance the entity representation compared to existing models [3, 28] that only rely on structural information. To be specific, we present various dedicated encoders for different modalities, including images, text, discrete attributes, and graph structure. These encoders generate corresponding representations for an entity on each modality.

Image Encoder and Text Encoder. As the pre-trained model CLIP [19] can fully understand the relationship between images and text by mapping them into a common feature space, it is chosen as encoder here. Detailedly, the CLIP text and CLIP image encoders are first utilized to encode the text and image attributes of an entity e respectively. Then, these encoded representations are fed into a linear layer, which maps the input vectors from their original feature space to a new one:

$$E_e^P = Norm(W_P \cdot CLIPImageEncoder(\mathcal{M}_e^P)), \tag{2}$$

$$E_e^C = Norm(W_C \cdot CLIPTextEncoder(\mathcal{M}_e^C)), \tag{3}$$

where $W_P \in \mathbb{R}^{d \times d_p}$ and $W_C \in \mathbb{R}^{d \times d_c}$ are learnable transformation matrices, and $Norm$ denotes the normalization function. E_e^P and E_e^C denote the representations of entity e on image and text modalities respectively.

Discrete Attributes Encoder. The discrete attributes in MMKG are represented in the form of triples, such as "(*Titanic, genre, Drama*)" and "(*Avatar, genre, Action*)" in Fig. 1. Prior to encoding these attributes, it is essential to interpret them semantically. This involves generating sentences that encapsulate the attribute information, such as "*The genre of Titanic is Drama*" and "*The genre of Avatar is Action*". Then, the sentences are embedded to obtain the representation of entity e on discrete attributes, which is denoted as E_e^D, using the CLIP text encoder:

$$E_e^D = Norm(W_D \cdot CLIPTextEncoder(Semantic(\mathcal{M}_e^D))). \tag{4}$$

Notably, the original multi-modal information of different entities exhibits inconsistency, since not all entities contain full modalities. To minimize the impact of missing modalities on the experimental results, the missing modal attributes of entities are set to zero vectors.

Graph Structure Encoder. After encoding entities with their own attributes based on the above encoders, we attempt to learn an additional representation E_e^G for each entity e from a graph structural perspective, by leveraging the multi-modal information from the neighbors of e. To achieve this, we suffer

from the following two main challenges: (1) the varying contributions of neighboring entities' attributes for learning E_e^G, depending on their modalities, and (2) the differing significance of information provided by the neighboring entities themselves. For instance, while learning the representation of $Titanic$ to infer the missing image for $(Titanic, image, ?)$ in Fig. 1, (1) the image attributes of $Titanic$'s neighboring entities, which are more likely to contain highly similar information with the potential missing image, play a more crucial role than other attributes like text and genre. (2) The neighboring entities of *Titanic* connected by relation *has star* are more relevant than those linked by other relations, since they share some actors (i.e., have similar visual information) with *Titanic*.

During the learning of the structural representation E_e^G, we process attributes from different modalities independently. Specifically, an MMKG is first decomposed into three sub-graphs $\mathcal{G}_P$, $\mathcal{G}_C$, and $\mathcal{G}_D$, where each sub-graph consists of nodes (i.e., entities) that exclusively contain attributes from a specific modality. For instance, the entities in $\mathcal{G}_C$ only include text attributes. To tackle challenge (1), we first learn the structural representation of e within each sub-graph by aggregating its neighboring nodes' embeddings. Next, the representations of e from different sub-graphs (i.e., modalities) are aggregated with dynamic weights, and the output is the unified initial embedding of e. To tackle challenge (2), the initial representation of e is enhanced through a relation-aware network, which integrates relationship information into the iterative process.

Formally, the representation of a neighboring entity on different attributes is defined as E_v^A, where v denotes a neighbor of the current entity e and $A \in \{P, C, D\}$ indicates the modality. In particular, when the attribute of v from a specific modality is missing, the representation of v on this attribute is generated based on the attributes of $v's$ neighbors denoted as $\mathcal{N}(v)$, using a GAT network:

$$E_v^A = \sigma(W_G \cdot GAT(\{E_u^A | u \in \mathcal{N}(v)\}), \tag{5}$$

where $W_G \in R^{d \times d}$ is a learnable parameter, d is the dimension of E_u^A, and σ is $ReLU(\cdot)$ function. The initial representation entity e (i.e., $E_e^{G,0}$) is obtained by aggregating the representations of neighboring nodes across different sub-graphs:

$$E_e^{G,0} = \sigma(\sum_A \sum_{v \in \mathcal{N}} w_A a_v E_v^A W_{e,A}), \tag{6}$$

where $W_{e,A} \in \mathbb{R}^{d \times d}$ is learnable transformation matrices, w_A represents the weight of attribute A, and a_v denotes the weight neighbor v. Following existing work [14], we utilize the attribute-consistent relation representation encoder and relation-aware entity representation encoder to enhance the representation of e. First, the initial representation of the relation between e and v is generated:

$$R^{(0)}[e, v] = \sum_{A \in \{P,C,D\}} (R_A[e, v] W_{0,A}), \tag{7}$$

where $W_{0,A} \in \mathbb{R}^{d \times d}$. Then, the representation of relation between e and v on attribute A, denoted as $R_A[e, v]$, is obtained based on TransE [4]:

$$R_A[e, v] = |E_v^A - E_e^{G,0}|, \tag{8}$$

Next, we update the representations of relations and entity e:

$$E_e^{G,l} = W_{h,E}[E_e^{G,(l-1)}||\frac{1}{D_e}\sum_{A}\sum_{v\in\mathcal{N}(e)} w_A[E_v^A||R^{(l-1)}[e,v]]], \tag{9}$$

$$R^l[e,v] = ReLU(W_{R,R}^{(l)}R^{(l-1)}[u,v] + \sum_{A\in(P,C,D)} W_{R,A}^{(l)}([E_e^{G,(l-1)}||E_v^A]), \tag{10}$$

where $W_{R,R}, W_{R,A}, W_{h,E} \in \mathbb{R}^{d\times d}$ are learnable parameter, D_e denotes the degree of e, and the max iterations number is defined as L. $E_e^{G,L} = E_e^G$ is the final represent of entity e on structural information.

Multi-modal Fusion. To better leverage the multi-modal and structural information of entities, we propose a dynamic modality fusion strategy. Unlike traditional hand-crafted fusion methods, which are constrained to specific domains [10, 17], dynamic modality fusion techniques are more adaptable to complex scenarios. This adaptability is crucial for our model as the importance of different modalities fluctuates dynamically based on entity type and data availability. By dynamically computing modality-specific weights, we generate a unified entity representation that effectively reflects the varying reliability of each modality.

Based on the assumption of [32], i.e., the higher uncertainty of a modality increases the likelihood of incorrect predictions, we first calculate the entropy of each modality to assess its uncertainty. Then, distinct weights are assigned to different modalities. Formally, the final output E_e, which denotes the embedding of entity e and has been presented in Fig. 2, of the module EEM is obtained with following equation:

$$E_e = \sigma(\sum_{m\in M\cup\{G\}} \lambda_{m,e}E_e^m), \tag{11}$$

where $M = \{P, C, D\}$ denotes the multi-modal information set, G represents the graph structural information, $\lambda_{m,e}$ is the weight of the entity e on modality m, and σ denotes the $ReLU(\cdot)$ function. To compute $\lambda_{m,e}$, we need to obtain the entropy $EP_{m,e}$ of each modality:

$$EP_{m,e} = -p_{m,e}^T \log p_{m,e}, \quad p_{m,e} = Softmax(E_e^m), \tag{12}$$

where the purpose of the $Softmax$ function is to normalize the target vector into a probability distribution p. Higher entropy indicates greater uncertainty of a modality, suggesting a higher risk of incorrect predictions. Therefore, a lower weight should be assigned to it to reduce its unreliability. After obtaining entropy $EP_{m,e}$, we calculate $\lambda_{m,e}$ with following equation:

$$\lambda_{m,e} = \frac{exp(\max_{m\in M\cup\{G\}} EP_{m,e} - EP_{m,e})}{\sum_{n\in M\cup\{G\}} exp(\max_{m\in M\cup\{G\}} EP_{m,e} - EP_{n,e})}. \tag{13}$$

4.2 Entity-Imgae Alignment

The purpose of EIA is to establish alignment between the entity and its potentially missing image. To ensure data consistency, we also apply the CLIP image encoder, which has been utilized to embed image attributes in EEM, to learn representations for candidate images. After calculating the similarity, we propose a weighted joint loss to enhance the model's learning capability, promoting better alignment between modalities and improving classification accuracy.

Formally, the query entity and candidate image are defined as e_q and i_c respectively, the similarity score $f(e_q, i_c)$ between them is calculated after embedding the candidate images into a shared space with entities:

$$f(e_q, i_c) = E_{e_q} \cdot \phi(CLIPImageEncoder(i_c)), \tag{14}$$

where E_{e_q} denotes the representation of e_q, and ϕ is a linear function designed to transform the raw embeddings of candidate images into a shared space with entities. We use the dot product method to calculate similarity scores.

Notably, the module EIA has two key objectives during optimization: the modality alignment between entities and candidate images, as well as the accuracy in matching candidate images to entities. Therefore, we design two loss functions, which are combined through weighted addition to form the final loss of the model. In detail, to enable the model to learn the former, we introduce a contrastive loss to reduce the modality differences, defined as:

$$\mathcal{L}_{ContrastiveLoss} = -\log \frac{E_{e_q} \cdot k^+/\tau}{E_{e_q} \cdot k^+/\tau + \sum_{i=0}^{K} E_{e_q} \cdot k_i^-/\tau}, \tag{15}$$

where k^+ denotes the positive sample, k_i^- indicates the i-th negative sample, and τ is a temperature hyperparameter. The module learns entity representations by maximizing the similarity of positive sample pairs and minimizing the similarity of negative sample pairs.

On the other hand, to improve the accuracy of matching, we train the model using a binary loss function, which can quantify the difference between the predicted probability and the actual binary label, effectively measuring how well the model's predictions align with the true outcomes:

$$\mathcal{L}_{BinaryLoss} = -\frac{1}{N}\sum_{i=1}^{N}(y_{e,i} \log f(e,i) + (1 - y_{e,i})(1 - \log f(e,i))), \tag{16}$$

where N is the number of samples, $y_{e,i}$ is the label of the pair (e, i) of entity e and image i (1 indicates a match, while 0 indicates a non-match), and $f(e, i)$ is defined in Eq. (14). Finally, the overall training objective is to minimize the following joint loss:

$$\mathcal{L}_{Joint} = \alpha \mathcal{L}_{ContrastiveLoss} + (1 - \alpha)\mathcal{L}_{BinaryLoss}, \tag{17}$$

where α is a hyperparameter.

5 EXPERIMENT

5.1 Datasets and Baselines

The performance of MAGIC is evaluated on following three real-world datasets: two publicly available datasets across diverse domains and one newly collected dataset in specific domain (i.e., film domain). **(1) DB15K.** This publicly available dataset is derived from the open-source project DBpedia. We supplement its multi-modal information from Wikipedia[1]. **(2) AspectMMKG.** It is provided by [30], where image attributes related to various aspects are constructed for entities. Building upon this, we also incorporate additional multi-modal information from Wikipedia. **(3) IMDB_MMKG.** This novel dataset is crawled from IMDB[2], consisting of relation triples (h, r, t) that represent films, actors, and their relationships. For each entity (film or actor), we collect its images, textual descriptions, and discrete attributes. All datasets encompass information across four modalities: relation triples, images, text, and discrete attributes. However, some entities may lack data in one or more modalities. In such cases, missing information is represented by an empty vector. Each dataset is divided into training, validation, and testing sets in an 80%-10%-10% ratio. The detailed statistics of datasets are presented in Table 1.

Table 1: Statistics of datasets

Dataset	#Ent.	#Rel.	#Train	#Valid	#Test
DB15K	12838	279	9024	1120	1120
AspectMMKG	1617	93	1280	160	160
IMDB_MMKG	91591	3	38094	4752	4752

To demonstrate the effectiveness of MAGIC, several highly representative methods on MMKG completion and image-text retrieval are utilized as baselines.

- **ControlNet** [31] is a neural network architecture that adds control to pretrained diffusion models, allowing users to guide the image generation with additional inputs. This improves the model's accuracy and precision.
- **MRAP** [2] is a numerical attribute prediction method that builds regression functions to facilitate information propagation across various attribute types and relationships, and it can be viewed as a message-passing framework.
- **IMF** [15] is an MMKG link prediction method, which learns knowledge separately in each modality and jointly models the complicated interactions between different modalities with a two-stage fusion.
- **AIR** [30] is developed to complete the knowledge graph AspectMMKG, which builds an image contrast learning model that selects aspect-relevant image from the original retrieved result.

[1] Wikipedia: https://en.wikipedia.org/wiki/
[2] IMDB: https://www.imdb.com/

Table 2: Experimental results of all methods

Methods	DB15K				AspectMMKG				IMDB_MMKG			
	MR	MRR	HITS@1	HITS@10	MR	MRR	HITS@1	HITS@10	MR	MRR	HITS@1	HITS@10
ControlNet	638	0.037	0.011	0.071	79	0.048	0.016	0.100	568	0.048	0.023	0.090
MRAP	762	0.004	0.002	0.008	91	0.010	0.004	0.024	2308	0.006	0.002	0.011
IMF	302	0.108	0.074	0.165	29	0.185	0.075	0.387	376	0.077	0.163	0.223
AIR	198	0.12	0.055	0.247	25	0.283	0.187	0.4	308	0.159	0.116	0.237
MAGIC	**152**	**0.176**	**0.106**	**0.316**	**15**	**0.371**	**0.225**	**0.637**	**139**	**0.252**	**0.168**	**0.417**
Improv.	4.1%	5.6%	5.1%	6.9%	6.2%	8.8%	3.8%	23.7%	3.5%	9.3%	5.2%	17.7%

5.2 Implementation Details and Evaluation Metrics

The model parameters are optimized by Adam optimizer with a learning rate of 0.01. For both image and text pre-trained models, the embedding dimensions are configured to 512. The hidden layer has a dimension of 300, and the size of batches N is set to 16. The hyperparameter α in Eq. (17) is set to 0.4.

Following previous studies [15, 30], three evaluation metrics are utilized to investigate the model performance, i.e., mean rank (MR), mean reciprocal rank (MRR), and HITS@K.

5.3 Overall Performance

We first calculate the similarity scores between entities and candidate images, then rank the obtained results. The model is evaluated based on the ranking of the correct image, with 'Improv.' denoting the achieved enhancement. Notably, MR improvement is measured as a relative percentage reduction from the baseline, as it can not be directly calculated. The experimental results of all methods are presented in Table 2, and we provide the following analysis.

Observed from Table 2, the proposed model MAGIC presents superior performance compared to other baselines, achieving a 23.7% improvement in the HITS@10 evaluation metric on dataset AspectMMKG. This suggests that traditional methods for text-image generation and knowledge graph completion show varying degrees of inadequacy when applied to image attribute completion tasks. The image generation models, such as ControlNet, are unable to leverage the structural information within knowledge graph. Although MRAP achieves competitive results on numerical attributes, it struggles to generalize effectively to multi-modal data. Although IMF incorporates both structural and multi-modal attributes, it is designed for completing relations [15], and performs poorly when applied to align entities and candidate images. Moreover, AIR exhibits certain limitations, as it focuses on completing the descriptions of images rather than linking them to entities [30]. In summary, our proposed model MAGIC significantly demonstrates substantial improvements over baseline methods on the task of image attribute completion, particularly in handling multi-modal data.

Table 3: Ablation on different modalities

Methods	DB15K				AspectMMKG				IMDB_MMKG			
	MR	MRR	HITS@1	HITS@10	MR	MRR	HITS@1	HITS@10	MR	MRR	HITS@1	HITS@10
P	326	0.075	0.038	0.144	35	0.291	0.206	0.431	785	0.167	0.129	0.234
P+C	169	0.156	0.088	0.283	21	0.324	0.214	0.5	337	0.191	0.132	0.308
P+C+D	158	0.165	0.089	0.313	20	0.332	0.218	0.55	225	0.212	0.142	0.356
P+C+G	153	0.165	0.091	0.315	18	0.346	0.225	0.575	174	0.235	0.156	0.392
P+D+G	169	0.169	0.105	0.293	19	0.306	0.223	0.543	184	0.211	0.141	0.346
C+D+G	183	0.121	0.061	0.249	23	0.296	0.212	0.468	183	0.151	0.079	0.291
P+C+D+G	**152**	**0.176**	**0.106**	**0.316**	**15**	**0.371**	**0.237**	**0.637**	**139**	**0.252**	**0.168**	**0.417**

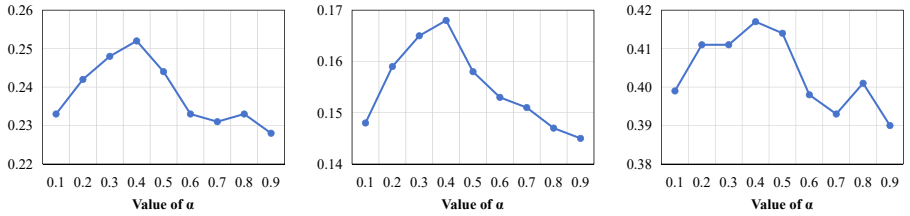
Table 4: Ablation on different fusion strategies

Methods	DB15K				AspectMMKG				IMDB_MMKG			
	MR	MRR	HITS@1	HITS@10	MR	MRR	HITS@1	HITS@10	MR	MRR	HITS@1	HITS@10
ADD	210	0.089	0.036	0.183	26	0.229	0.118	0.431	182	0.134	0.067	0.278
CONCAT	196	0.115	0.059	0.231	21	0.286	0.175	0.525	154	0.214	0.136	0.379
LSTM	188	0.089	0.039	0.185	26	0.207	0.093	0.443	174	0.235	0.156	0.392
MAGIC	**152**	**0.176**	**0.106**	**0.316**	**15**	**0.371**	**0.225**	**0.637**	**139**	**0.252**	**0.168**	**0.417**

5.4 Ablation Study and Parameter Analysis

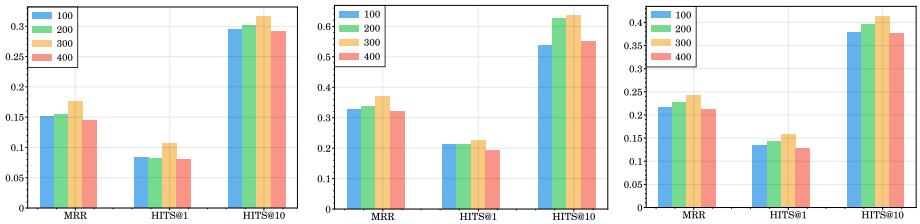
Impact of Different Modality Information. To assess the impact of different modalities on the model performance, we conduct ablation experiments by successively removing modality and retraining the model on three datasets. The results are reported in Table 3, where P denotes image attributes, C represents text attributes, D stands for discrete attributes, and G refers to the graph structure information. From this table, we have the following observations: (1) When exploiting information from all modalities for prediction, the best results are achieved. This demonstrates that the inclusion of multi-modal information significantly contributes to the improvement of model performance. (2) While different modalities are removed individually, we can observe that removing the image modality has the greatest impact on experimental results. This might be because the image modality provides the most information when predicting new attributes of the same modality. (3) However, this does not mean that the image modality alone can predict new attributes effectively. On the one hand, removing any of the other modalities also leads to a decline in model performance, indicating that information from these modalities is crucial for accurate predictions. On the other hand, when the model relies solely on the image modality for prediction, performance decreases even more drastically, further emphasizing the importance of utilizing multiple modalities for reliable prediction.

Impact of Different Fusion Strategies. To explore the impact of different multi-modal fusion strategies on model performance, we conduct comparative experiments by replacing our entropy-based fusion strategy with following three common fusion methods. (1) **ADD**: Project the information from different



(a) MRR (b) HITS@1 (c) HITS@10

Fig. 3: Performance of MAGIC w.r.t. varied α



(a) DB15K (b) AspectMMKG (c) IMDB_MMKG

Fig. 4: Performance of MAGIC w.r.t. varied dimension of hidden layer

modalities into a unified vector space, sum them, and apply a *ReLU* activation function to produce the final output. (2) **CONCAT**: Concatenate the information from different modalities and perform a linear transformation to generate a unique representation of the entity. (3) **LSTM**: Sequentially feed the discrete attributes, structural information, text, and images into an LSTM network, allowing the model to capture information from different modalities layer by layer, ultimately producing the final representation. The results on three datasets are reported in Table 4. Observed from it, our proposed model MAGIC presents a significant improvement compared to other fusion methods, which also validates the previous assumption: higher uncertainty in a modality increases the likelihood of incorrect predictions.

Impact of Hyperparameters. To investigate the impact of two key hyperparameters, i.e., α in Eq. (17) and the dimension of hidden layer, on model performance, we conduct extensive experiments and present the results in Fig. 3. (1) Seen from the figure, when α varies from 0.1 to 0.9, the proposed model MAGIC achieves the best performance at α= 0.4. Any increase or decrease of α leads to a noticeable decline in performance. This demonstrates that the weighted loss function we employed has a positive impact on model performance. It allows the model to learn distinct information from the two types of loss functions, thereby enhancing the overall effectiveness of MAGIC. (2) For the dimension of hidden layer, we conduct experiments on three datasets with dimensions of 100,

200, 300, and 400. The results are reported in Fig. 4. As can be seen from the experimental outcomes, the best performance is achieved when the dimension of the hidden layer is set to 300. When given a too-small dimension, the model is unable to effectively capture features. However, this does not imply that increasing the dimension will continuously improve performance. When the dimension becomes too large, the model performance begins to degrade, possibly due to the introduction of redundant information. Notably, the analysis of other parameters is omitted here, due to the space limitation.

6 CONCLUSION AND FUTURE WORK

In this paper, we propose a novel task, i.e., image attribute completion, and develop a model called MAGIC with two modules to tackle it. Specifically, the module EEM is designed to encode entities, by aggregating modality-specific representations of an entity with a dynamic fusion strategy. In addition, the module EIA optimizes the alignment of entities and images by calculating similarities and combining two distinct loss functions. The extensive experiments conducted on three datasets demonstrate the effectiveness of the proposed model MAGIC. Nevertheless, there remains potential for further exploration, especially in the simultaneous completion of attributes across different modalities.

Acknowledgments. This research is supported by the National Natural Science Foundation of China No. 62272332 and the Major Program of the Natural Science Foundation of Jiangsu Higher Education Institutions of China No. 22KJA520006.

References

1. Balazevic, I., Allen, C., Hospedales, T.M.: Tucker: Tensor factorization for knowledge graph completion. In: EMNLP. pp. 5184–5193 (2019)
2. Bayram, E., García-Durán, A., West, R.: Node attribute completion in knowledge graphs with multi-relational propagation. In: ICASSP. pp. 3590–3594 (2020)
3. Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NeurIPS. p. 2787–2795 (2013)
4. Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NeurIPS. p. 2787–2795 (2013)
5. Chen, J., Fan, W., Zhu, G., Zhao, X., Yuan, C., Li, Q., Huang, Y.: Knowledge-enhanced black-box attacks for recommendations. In: SIGKDD. p. 108–117 (2022)
6. Chen, Z., et al: Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In: ACM MM. p. 3317–3327 (2023)
7. Devlin, J., et al: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL. pp. 4171–4186 (2019)
8. Gregucci, C., et al: Link prediction with attention applied on multiple knowledge graph embedding models. In: WWW. p. 2600–2610 (2023)
9. Huang, X., Zhang, J., Li, D., Li, P.: Knowledge graph embedding based question answering. In: WSDM. p. 105–113 (2019)

10. Joze, H.R.V., Shaban, A., Iuzzolino, M.L., Koishida, K.: Mmtm: Multimodal transfer module for cnn fusion. In: CVPR. pp. 13286–13296 (2019)
11. Kotnis, B., GarcÃa-DurÃ¡n, A.: Learning numerical attributes in knowledge bases. In: AKBC. pp. 1–20 (2019)
12. Lee, J., Chung, C., Lee, H., Jo, S., Whang, J.: Vista: Visual-textual knowledge graph representation learning. In: EMNLP. pp. 7314–7328 (2023)
13. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. CoRR **abs/1908.03557** (2019)
14. Li, Q., et al: Attribute-consistent knowledge graph representation learning for multi-modal entity alignment. In: WWW. p. 2499–2508 (2023)
15. Li, X., Zhao, X., Xu, J., Zhang, Y., Xing, C.: Imf: Interactive multimodal fusion model for link prediction. In: WWW. p. 2572–2580 (2023)
16. Lin, Z., Zhang, Z., Wang, M., Shi, Y., Wu, X., Zheng, Y.: Multi-modal contrastive representation learning for entity alignment. In: COLING. pp. 2572–2584 (2022)
17. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. In: NeurIPS. pp. 14200–14213 (2021)
18. Oñoro-Rubio, D., et al: Answering visual-relational queries in web-extracted knowledge graphs. In: AKBC. p. 1–20 (2017)
19. Radford, A., et al: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
20. Sun, Z., Deng, Z., Nie, J.Y., Tang, J.: Rotate: Knowledge graph embedding by relational rotation in complex space. In: ICLR. pp. 1–18 (2019)
21. Tay, Y., Luu, A.T., Phan, M.C., Hui, S.C.: Multi-task neural network for non-discrete attribute prediction in knowledge graphs. In: CIKM. p. 1029–1038 (2017)
22. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.: Complex embeddings for simple link prediction. In: ICML. p. 2071–2080 (2016)
23. Wang, H., et al: Ripplenet: Propagating user preferences on the knowledge graph for recommender systems. In: CIKM. pp. 417–426 (2018)
24. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: CVPR. pp. 12692–12702 (2019)
25. Xu, D., Xu, T., Wu, S., Zhou, J., Chen, E.: Relation-enhanced negative sampling for multimodal knowledge graph completion. In: ACM MM. p. 3857–3866 (2022)
26. Xu, J., Zhang, W., Duan, Q., Li, S.: Hoap: Node attribute completion of knowledge graph based on high-order neighbor attribute propagation. Preprint at *ArXiv* https://doi.org/10.21203/rs.3.rs-1937079/v1 (2022)
27. Xue, B., Li, Y., Zou, L.: Introducing semantic information for numerical attribute prediction over knowledge graphs. In: ISWC. pp. 3–21 (2022)
28. Yang, B., tau Yih, W., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. In: ICLR. pp. 1–12 (2014)
29. Yih, W.t., et al: Semantic parsing via staged query graph generation: Question answering with knowledge base. In: ACL. pp. 1321–1331 (2015)
30. Zhang, J., Wang, J., Wang, X., Li, Z., Xiao, Y.: Aspectmmkg: A multi-modal knowledge graph with aspect-aware entities. In: CIKM. p. 3361–3370 (2023)
31. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3813–3824 (2023)
32. Zhang, X., Yoon, J., Bansal, M., Yao, H.: Multimodal representation learning by alternating unimodal adaptation. In: CVPR. pp. 27446–27456 (2024)
33. Zhang, Y., et al: Native: Multi-modal knowledge graph completion in the wild. In: SIGIR. pp. 91–101 (2024)
34. Zheng, S., Wang, W., Qu, J., Yin, H., Chen, W., Zhao, L.: Mmkgr: Multi-hop multi-modal knowledge graph reasoning. In: ICDE. pp. 96–109 (2023)