Let Modalities Teach Each Other: Modal-Collaborative Knowledge Extraction and Fusion for Multimodal Knowledge Graph Completion

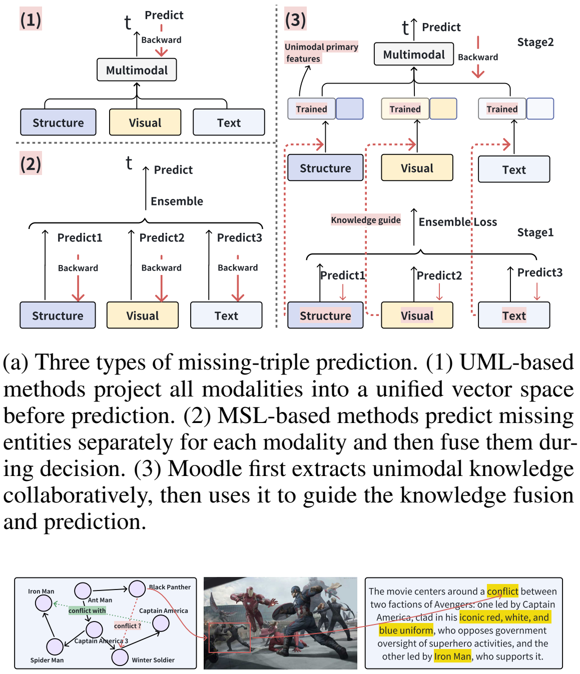
Guoliang Zhu[1,2], Tao Ren[1,2*], Dandan Wang[1,2], JUN HU[1,2*]
[1]State Key Laboratory of Intelligent Game,
Institute of Software, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{guoliang2022,rentao22,dandan,hujun}@iscas.ac.cn

Abstract

Multimodal knowledge graph completion (MKGC) aims to predict missing triples in MKGs using multimodal information. Recent research typically either extracts information from each modality separately to predict, then ensembles the predictions at the decision stage, or projects multiple modalities into a unified feature space to learn multimodal representations for prediction. However, these methods usually overlook the intrinsic correlation between modalities in MKGs which should be leveraged in both unimodal knowledge extraction and multimodal knowledge fusion. Motivated by this, we propose a noval *Mo*dal-*co*llaborative knowl*ed*ge *le*arning (Moodle) framework for MKGC, the key idea of which is to foster mutual guidance and collaboration during unimodal knowledge extraction, to let each modality acquire distinct and complementary knowledge that subsequently enhances the multimodal knowledge fusion. Specifically, Moodle preserves the representations of different modalities to learn unimodal knowledge while modeling the mutual guidance through multi-task learning. Furthermore, Moodle performs multimodal knowledge fusion and prediction guided by unimodal knowledge, capturing their synergistic relationships and acquire fine-grained semantic knowledge through contrastive learning. Extensive experiments on three real-world datasets demonstrate the advantages of Moodle over state-of-the-art methods.

1 Introduction

Recent years knowledge graphs (KGs) have experienced rapid development across various real-world applications, such as news recommendation (Liu et al., 2021; hao et al., 2023), intelligent question answering (Yasunaga et al., 2022; Wang et al., 2023; Yu et al., 2022), and social network analysis (Molokwu et al., 2020; Molokwu and Kobti, 2020), owing to their strengths in representing and organizing knowledge, typically in the form of fact triples (head entity, relation, tail entity). The growth of multimodal corpora, such as text and images, has catalyzed the emergence of multimodal knowledge graphs (MKGs) (Liu et al., 2019; Chen et al., 2024), which extend traditional KGs by incorporating diverse types of entity attributes. This enriched representation of MKGs significantly enhances their capabilities in a broader range of domains.



The movie centers around a conflict between two factions of Avengers: one led by Captain America, clad in his iconic red, white, and blue uniform, who opposes government oversight of superhero activities, and the other led by Iron Man, who supports it.

(a) Three types of missing-triple prediction. (1) UML-based methods project all modalities into a unified vector space before prediction. (2) MSL-based methods predict missing entities separately for each modality and then fuse them during decision. (3) Moodle first extracts unimodal knowledge collaboratively, then uses it to guide the knowledge fusion and prediction.

(b) An example of multimodal input consists of text description, image, and associated subgraph about the movie “Captain America 3: Civil War".

Figure 1: (a) Similarities and differences between existing methods and Moodle. (b) Correlation between modalities.

However, due to the incompleteness of facts, multimodal knowledge graphs (MKGs) often suffer from missing relations or entities, which limits

*Corresponding authors: Tao Ren and Jun Hu

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 6210–6222
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

their effectiveness in real-world applications. To address this issue, multimodal knowledge graph completion (MKGC) has been proposed to predict missing parts of MKGs by utilizing multimodal data and it has garnered significant attention from both academic and industrial communities (Xie et al., 2017; Wang et al., 2019). Traditional methods for KGC mainly adopt logical inference and statistical probability to predict missing entities and relations in KGs. However, with the emergence of MKGs, these methods face considerable challenges in effectively integrating and leveraging multiple modalities of data. This led to the development of representation learning in MKGC, where entities, relations as well as multimodal information within the MKG are projected into a low-dimensional vector space, allowing for the prediction of missing triples using scoring functions on the learned embeddings. Whereas, the introduction of representation learning also faces a critical challenge: how to effectively integrate complementary multimodal embeddings with structural embeddings.

There are primarily two paradigms for MKGC, Modality-Specific Learning (MSL-based) and Unified Multimodal Learning (UML-based). As shown in (2) of Figure 1(a), MSL-based methods has been proposed to learn modality-specific embeddings and then integrate predictions from each modality at the decision-making stage to reduce information loss resulting from modality heterogeneity. For example, MoSE (Zhao et al., 2022) proposes three different strategies to fuse the unimodal predicting results during the decision-making phase. Despite these efforts, extracting unimodal features separately could be short of effective utilization of supplementary information across modalities. In contrast, UML-based methods, as shown in (1) of Figure 1(a) , focus on projecting multimodal features into a unified vector space to capture deep semantic information. For instance, IKRL (Xie et al., 2017) pioneers the approach of fusing visual information with structural embeddings in KGC, and MKGFormer (Chen et al., 2022) modifies the internal structure of Transformers to perform cross-modal fusion, aiming to alleviate the heterogeneity issue and reduce the impact of irrelevant multimodal information on prediction performance. Although these UML-based methods can obtain a unified multimodal representation that contains richer information, they often fall short in preserving the distinct characteristics inherent to each modality and face the issue of modal conflict.

Intuitively, the multimodal data in MKGs potentially contain rich collaborated and complementary information. Take the scenario in Figure 1(b) as an example. The textual description with yellow backgrounds mentions the conflict between the Iron Man and Captain America factions, but does not explicitly refer to the conflict involving Black Panther and Winter Soldier. However, the image depicts antagonistic scenes between these characters (as seen in the red square). The visual information plays a crucial role in understanding and extracting the context of the conflict knowledge involving Black Panther and Winter Soldier, even if the text does not mention it directly. Conversely, textual information can also guide the accurate extraction of features from images. For instance, if the text describes the conflict between the Iron Man and Captain America, it will be helpful for extracting the characters and scenes involved in the conflict from the image, while disregarding irrelevant information.

Inspired by the above intuition, we propose a noval *Mo*dal-*co*llaborative knowle*dg*e *le*arning (*Moodle*) framework for MKGC. Moodle allows modalities to teach each other during both unimodal knowledge extracting and multimodal knowledge fusion. Specificly, we design a collaborative unimodal learner (CuLearner) that leverages multitask learning to model the mutual collaboration in unimodal knowledge extracting. The CuLearner integrates weight learning and cross-attention filter to acquire distinct knowledge across modalities. After unimodal learning, the Multimodal Semantic Learner (MsLearner) is designed to capture implicit inter-modal knowledge interactions and perform knowledge fusion across modalities guided by unimodal knowledge, enabling Moodle to acquire complementary knowledge for accurate prediction of missing triplets. We conduct extensive experiments on various benchmark datasets, showing the advantage of Moodle over the state-of-the-art methods. The contributions of this paper are summarized as follows:

- We propose to let modalities teach each other during both unimodal knowledge extraction and multimodal knowledge fusion to facilitate the prediction of missing triplets.
- We designed a collaborative unimodal learner to acquire task-relevant, distinct knowledge from each modality, enhancing the mutual guidance and collaboration of complementary information across modalities.

- We designed a multimodal semantic learner to perform knowledge-guided multimodal knowledge extraction and fusion, leading to more precise and nuanced predictions in complex multimodal scenarios.

- We conducted extensive experiments on three real-world datasets to demonstrate the superiority of Moodle over state-of-the-art methods, along with insightful analyses to validate the efficacy of Moodle.

2 Related Work

2.1 Unimodal Knowledge Graph Completion

Knowledge representation learning methods have been widely used in the KGC task, projecting entities and relations into a low-dimensional vector space and then designing appropriate scoring functions to optimize the representations. Translation-based methods, such as TransE (Bordes et al., 2013), TransH (Wang et al., 2014) and TransR (Lin et al., 2015b), model relations as distance transformation of source and target entities. DistMult (Yang et al., 2015) models relations as diagonal matrices and scores relational triples by computing a bilinear product between the entity and relation embeddings. ComplEx (Trouillon et al., 2017) extends DistMult by using complex-valued embeddings to better capture asymmetric relations in KGs. ConvE (Dettmers et al., 2018) and ConvKB (Nguyen et al., 2018) utilize a convolutional neural network to obtain the joint representation of entities and relations. RotatE (Sun et al., 2019) represents relations as rotations in a complex vector space, scoring relational triples based on the distance between the rotated head entity and the tail entity. TuckER (Lin et al., 2015a) uses tucker decomposition to model the interaction of entities and relations. However, all of the aforementioned methods rely solely on structural information, which is inadequate for addressing more complex real-world scenarios. By integrating multimodal information into the training process, MKGC enhances the representations with external knowledge, resulting in more comprehensive and robust embeddings.

2.2 Multimodal Knowledge Graph Completion

MKGC methods map multimodal information into a unified vector space for subsequent prediction, or optimize and predict each modality separately, then fuse the results at the decision level. IKRL (Xie et al., 2017) integrates visual information by combining multiple score functions between structure and visual embeddings. TransAE (Wang et al., 2019) utilizes the reconstruction loss of autoencoder to facilitate information fusion across multimodals. RSME (Wang et al., 2021) selectively filters visual information during the learning of KG embeddings. MKGFormer (Chen et al., 2022) integrates multimodal features into a unified space using the attention mechanism within the M-encoder. Additionally, it modifies the transformer's internal attention mechanism to mitigate the modality heterogeneity and align entities and relations. AdaMF-MAT (Zhang et al., 2024) proposes a modality adversarial training strategy to utilize imbalanced modality information. OTKGE (Cao et al., 2022) employs optimal transport techniques to achieve multimodal fusion, enabling efficient and effective integration of information across different modalities. MoSE (Zhao et al., 2022) uses modality-specific embeddings and dynamic ensemble methods for inference by difference strategies. IMF (Li et al., 2023) captures bilinear interactions for jointly modeling the commonality and complementarity. VBKGC (Zhang and Zhang, 2022) utilizes Visual-BERT to effectively extract multi-modal information and optimize the fusion of modalities. MMRNS (Xu et al., 2022) leverages multimodal information to improve negative sampling, leading to more effective training of KGE models. Compared with these methods, Moodle can capture both the distinct and complementary information across modalities.

3 Method

3.1 Preliminary

In this section, we provide the symbols and notations used in MKGC. A MKG is represented as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ is the set of entities and $\mathcal{R}$ is the set of relations. $\mathcal{T}$ is the set of triples, (e_i, r, e_j) denoting relations between entities where $e_i, e_j \in \mathcal{E}$ and $r \in \mathcal{R}$. MKGC aims at predicting the missing triples, which can be further divided into two types of sub-tasks: predicting the tail entity for triple $(e_i, r, ?)$ and predicting the head entity for $(?, r, e_j)$ where each entity $e \in \mathcal{E}$ has structure information e^{s} and associated multimodal data (descriptions e^{d} and images e^{v}).

3.2 Overall Architecture

To effectively utilize the multimodal information and prevent modal confliction in MKGC, we propose a Modal-collaborative knowledge learning framework, Moodle, that let modalities to teach each other during both unimodal knowledge extracting and multimodal knowledge fusion. The extraction of unimodal knowledge necessitates addressing three key issues: preserving modality-specific features while enabling mutual guidance, mitigating the impact of irrelevant information in unimodal data, and modeling the varying amounts of information across different modalities and scenarios. Additionally, multimodal knowledge fusion poses a significant challenge due to the presence of modality heterogeneity. Based on the aforementioned issues, Moodle proposes a noval Collaborative Unimodal Learner to model the mutual guidance across modalities and a Multimodal Semantic Learner to perform modal fusion in a knowledge-guided manner as show in Figure 2:

- The Collaborative Unimodal Learner employs a multi-task learning framework with an ensemble weighted learner to model the guidance and collaboration among modalities. Additionally, it utilizes a cross-attention filter to explicitly leverage textual and structural information to guide the image modality, helping to mitigate the issue of irrelevant information in images.

- The Multimodal Semantic Learner achieves multimodal fusion, after which the fused representation, along with the unimodal features, is optimized through contrastive learning to enhance the capture of fine-grained semantic information. In this process, guided by unimodal knowledge, a sufficiently simple multimodal feature extractor is employed to ensure both consistency and an adequate number of negative samples.

3.3 Collaborative Unimodal Learner

We employ pre-trained unimodal encoders as frozen feature extractors to obtain structural, visual, and textual embeddings. For textual data, we utilize BERT (Devlin et al., 2019) to extract embeddings for each entity along with its description. Instead of using the [CLS] token to represent the final embedding, we use pooling, since it is hard for the [CLS] token to capture subtle differences in sentences that are sometimes crucial for MKGC. For structural data, we adopt TuckER (Lin et al., 2015a), which aggregates neighboring nodes and relationships to generate the structural embeddings. For visual data, we use VGG16 (Simonyan and Zisserman, 2015), pre-trained on ImageNet, to extract visual features and generate the corresponding visual embeddings. The formal expression for the encoding process described above is as follows:

$$h^{\mathrm{s}}, \hat{h}^{\mathrm{v}}, h^{\mathrm{t}} = \psi_{\{\mathrm{tuck,vgg,bert}\}}(e^{\mathrm{s}}, e^{\mathrm{v}}, e^{\mathrm{t}}), \quad (1)$$

where $e^{\mathrm{s}}, e^{\mathrm{v}}, e^{\mathrm{t}}$ represent the structural, visual, and textual data of entities, $h^{\mathrm{s}}, \hat{h}^{\mathrm{v}}, h^{\mathrm{t}}$ represent their corresponding embeddings, and $\psi_{\{\mathrm{tuck,vgg,bert}\}}$ denotes the the structural, visual, and textual feature extractor, respectively.

3.3.1 Irrelevant Information Filtering

In existing MKG, since the image data usually comes from web crawlers or ImageNet, each entity could be associated with multiple images that often contain redundant irrelevant information. Hence, we design a cross-attention filter to utilize textual and structural information to guide the extraction of valuable information from multiple images as follows:

$$h^{\mathrm{v}}_{\mathrm{t}} = \mathrm{softmax}\left(\frac{(h^{\mathrm{t}} w_{\mathrm{q}})(\hat{h}^{\mathrm{v}} w_{\mathrm{k}})}{\sqrt{d_{\mathrm{k}}}}\right)(\hat{h}^{\mathrm{v}} w_{\mathrm{v}}), \quad (2)$$

$$h^{\mathrm{v}} = [h^{\mathrm{v}}_{\mathrm{t}} | h^{\mathrm{v}}_{\mathrm{s}}], \quad (3)$$

where $h^{\mathrm{v}}_{\mathrm{t}}$ and $h^{\mathrm{v}}_{\mathrm{s}}$ denote text-attentioned and structure-attentioned visual embeddings respectively, and h^{v} is the filtered visual information. $h^{\mathrm{v}}_{\mathrm{s}}$ is obtained similarly with $h^{\mathrm{v}}_{\mathrm{t}}$ shown in above formulation.

3.3.2 Entity-Relation Interaction

As the same entity may exhibit different semantic characteristics across various relations, CuLearner integrate entities and relations by Tucker decomposition. Tucker decomposition is a method used for higher-order tensor decomposition. It breaks down a tensor into a core tensor and factor matrices along each dimension, enabling efficient representation and interaction of complex data. The joint representation of source entity and relation is as follows:

$$h^{m}_{ir} = \mathcal{W} \times_1 w_r \times_2 h^{m}_{i}, \quad (4)$$

where $m \in \{\mathrm{s}, \mathrm{v}, \mathrm{t}\}$ could be different modals, $\mathcal{W}$ is the core tensor, h^{m}_{i} is the entity embedding and w_r represents the relation embedding.

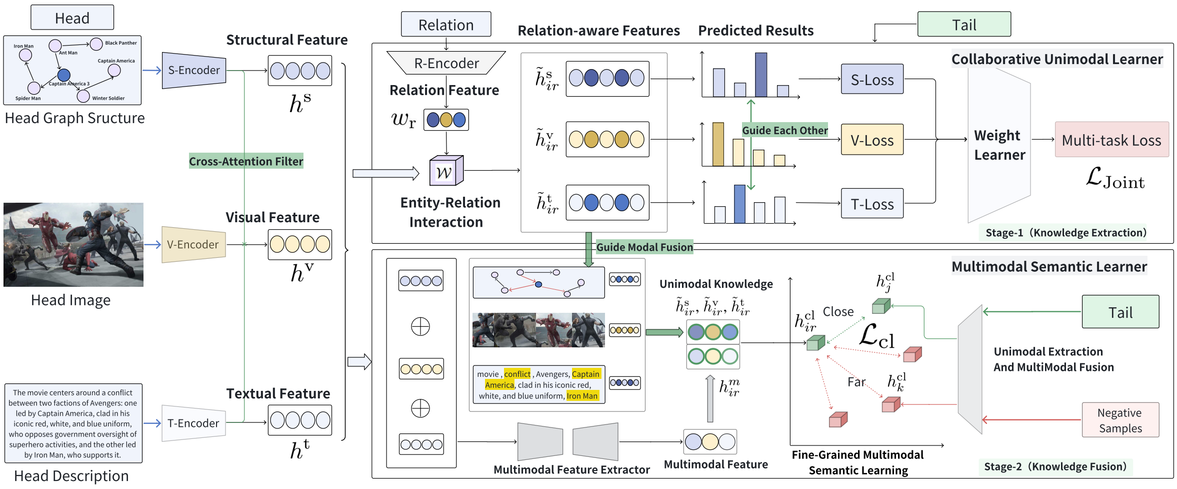

Head
Iron Man
Black Panther
Ant Man
Captain America
Captain America 3
Spider Man
Winter Soldier
Head Graph Sructure
S-Encoder
Structural Feature
h^{s}
Cross-Attention Filter
Head Image
V-Encoder
Visual Feature
h^{v}
The movie centers around a conflict between two factions of Avengers: one led by Captain America, clad in his iconic red, white, and blue uniform, who opposes government oversight of superhero activities, and the other led by Iron Man, who supports it.
Head Description
T-Encoder
Textual Feature
h^{t}
Relation
R-Encoder
Relation Feature
w_{r}
$\mathcal{W}$
Entity-Relation Interaction
Relation-aware Features
$\tilde{h}^{\mathrm{s}}_{ir}$
$\tilde{h}^{\mathrm{v}}_{ir}$
$\tilde{h}^{\mathrm{t}}_{ir}$
Predicted Results
Guide Each Other
Tail
S-Loss
V-Loss
T-Loss
Collaborative Unimodal Learner
Weight Learner
Multi-task Loss
$\mathcal{L}_{\mathrm{Joint}}$
Stage-1（Knowledge Extraction）
Guide Modal Fusion
movie , conflict , Avengers, Captain America, clad in his iconic red, white, and blue uniform, Iron Man
Unimodal Knowledge
$\tilde{h}^{\mathrm{s}}_{ir}, \tilde{h}^{\mathrm{v}}_{ir}, \tilde{h}^{\mathrm{t}}_{ir}$
h^{m}_{ir}
Multimodal Feature Extractor
Multimodal Feature
h^{cl}_{ir}
Close
h^{cl}_{j}
$\mathcal{L}_{\mathrm{cl}}$
Far
h^{cl}_{k}
Fine-Grained Multimodal Semantic Learning
Multimodal Semantic Learner
Tail
Unimodal Extraction And MultiModal Fusion
Negative Samples
Stage-2（Knowledge Fusion）

Figure 2: Overview of Moodle. The left part utilizes pre-trained models to separately extract structural, visual, and textual embeddings from the KG. The upper-right part depicts the Collaborative Unimodal Learner, which learns task-specific unimodal representations of entities and relations under the guidance of each other. The lower-right part illustrates the Multimodal Semantic Learner, which integrates unimodal knowledge with multimodal features to learn discriminative semantic information.

3.3.3 Multi-task Learning

After obtaining relation-aware entity representations, CuLearner employs a multitask learning approach to model the collaboration across different modalities. Given the presence of one-to-many relationships in KGs, we frame MKGC as a multi-label classification problem for each modality. Each modality corresponds to a subtask within the multitask learning framework, and the model is optimized using cross-entropy loss.

$$y_i^m = h_{ir}^m \cdot [h_1^m, h_2^m, \dots, h_k^m]^{\mathrm{T}}, e_k \in \mathcal{E}, \tag{5}$$

$$\mathcal{L}_m = -\frac{1}{N}\sum_{i=1}^{N}(t_i \cdot \log(y_i^m) + (1-t_i)\cdot\log(1-y_i^m)), \tag{6}$$

where t_i is the label of the current triple, and y_i^m is the predicted result of each modality which can be obtained by performing dot product between the relation-aware entity embedding and the embeddings of all entities. On account of the varying impact of different modalities in different contexts, we design a self-learning weight learner to integrate the losses of each modality. The total loss for this part can be expressed as:

$$\mathcal{L}_{\mathrm{Joint}} = \gamma_{\mathrm{s}}\mathcal{L}_{\mathrm{s}} + \gamma_{\mathrm{v}}\mathcal{L}_{\mathrm{v}} + \gamma_{\mathrm{t}}\mathcal{L}_{\mathrm{t}}, \tag{7}$$

where $\mathcal{L}_{\mathrm{s}}, \mathcal{L}_{\mathrm{v}}, \mathcal{L}_{\mathrm{t}}$ denote the losses of different modalities and $\gamma_{\mathrm{s}}, \gamma_{\mathrm{v}}, \gamma_{\mathrm{t}}$ are the self-learned weights.

The CuLearner enables the modalities to guide and complement each other, effectively preserving their respective features and filtering the irrelevant information. After training, unimodal knowledge can be obtained. The trained features of the source entity and relation for three modalities are denoted as $\tilde{h}^{\mathrm{s}}_{ir}$, $\tilde{h}^{\mathrm{v}}_{ir}$, $\tilde{h}^{\mathrm{t}}_{ir}$, respectively. For target and other entities, these features are denoted as $\tilde{h}^{\mathrm{s}}_{j}$, $\tilde{h}^{\mathrm{v}}_{j}$, $\tilde{h}^{\mathrm{t}}_{j}$ and $\tilde{h}^{\mathrm{s}}_{k}$, $\tilde{h}^{\mathrm{v}}_{k}$, $\tilde{h}^{\mathrm{t}}_{k}$.

3.4 Multimodal Semantic Learner

In this section, we aim to obtain multimodal representations that capture the complex interactions between different modalities. Many existing multimodal fusion methods have achieved promising results. However, most of them overlook the use of task-relevant unimodal knowledge to guide the fusion process. As a result, they often face issues of modal conflict and information loss. To alleviate the above issues, we propose a noval Multimodal Semantic Learner to realize this guidance and learn fine-grained semantic knowledge. Specificly, we utilize contrastive learning (CL) to optimize our task. The objective is to maximize the similarity between the source relation-aware embeddings and the target entity embeddings, while minimizing the similarity with other negative samples.

3.4.1 Fine-Grained Semantic Learning

In MKGC task, a challenging issue arises when dealing with samples that are semantically similar, making them difficult to distinguish. Therefore, it is necessary to learn more fine-grained features for prediction. This involves ensuring that semantically similar entities have sufficient distance in the vector space. To address this, we employ contrastive learning to extract more granular semantic information. Specifically, we use InfoNCE loss to optimize the representations as show in following formulation:

$$\mathcal{L}_{\text{cl}} = -\log \frac{e^{(\mathcal{F}(e_i,r,e_j)-\gamma)/\tau}}{e^{(\mathcal{F}(e_i,r,e_j)-\gamma)/\tau} + \sum_{k=1}^{|\mathcal{N}|} e^{\mathcal{F}(e_i,r,e_k)/\tau}}, \tag{8}$$

where $\mathcal{N}$ denotes the set of negative samples, and τ is the temperature parameter that adjusts the relative importance of negative samples. A smaller τ makes the loss function focus more on hard negatives but also increases the risk of overfitting to label noise. γ is the margin value which aims to increase the score distance between the positive and negative samples. We adopt in-batch negative sampling, where for triples within the same batch, the joint embeddings of head and relation are used as queries, and all tail embeddings are used as keys. If the triple is not in the knowledge graph, it is considered a negative example and needs to be distinguished from positive samples. Cosine similarity is used for the calculation of scoring function $\mathcal{F}$ as follows:

$$\mathcal{F}(e_i, r, e_j) = \cos(h_{ir}^{\text{cl}}, h_j^{\text{cl}}) = \frac{h_{ir}^{\text{cl}} \cdot h_j^{\text{cl}}}{\|h_{ir}^{\text{cl}}\| \|h_j^{\text{cl}}\|}, \tag{9}$$

where h_{ir}^{cl} and h_j^{cl} denoted the joint embedding of entity-relation and embedding of entities respectively, which will be discussed in detail in the following sections.

3.4.2 Multimodal Feature Extractor

To effectively learn discriminative semantic knowledge, a consistent feature extractor and sufficient negative samples during contrastive learning (CL) are essential. SimKGC (Wang et al., 2022) utilizes BERT (Devlin et al., 2019) to extract features and optimizes all parameters through CL. Although this method has a consistent feature extractor, it requires significant memory resources, especially with large batch sizes. This issue is exacerbated in multimodal scenarios, as each modality necessitates a separate feature extractor to be finetuned. To address this, we design a lightweight feature extractor akin to a stacked autoencoder, consisting of an encoder and a decoder. The encoder maps the input features x into a lower-dimensional representation z as follows:

$$z = \sigma(W_2\sigma(W_1x + b_1) + b_2). \tag{10}$$

The decoder reconstructs the input from the lower-dimensional representation z:

$$h^m = \sigma(W_4\sigma(W_3z + b_3) + b_4), \tag{11}$$

where x is the input represents the concatenation of features from each modality denoted as $[h^{\text{s}}|h^{\text{v}}|h^{\text{t}}]$, z is encoded representation and h^m is the reconstructed multimodal features. W, b are the trainable parameters and σ is the sigmoid activation function.

Although the above extraction process has strong feature extraction capabilities, it does not account for the specific knowledge required in the MKGC task and the model training can easily get stuck in a local optimum. To overcome this, we combine the trained modal-specific features with the output of the multimodal feature extractor as the initial features of CL to realize the knowledge guidance. The formulation is as follows:

$$h_{ir}^{\text{cl}} = h_{ir}^{m}|(\tilde{h}_{ir}^{\text{s}}|\tilde{h}_{ir}^{\text{v}}|\tilde{h}_{ir}^{\text{t}}), \tag{12}$$

where h_{ir}^{cl} represents the joint embedding of entities and relations in CL and h_j^{cl}, h_k^{cl} are obtained similarly. After obtaining the aforementioned hybrid representation, we optimize it with a large number of negative samples during contrastive learning. This process helps the model learn more fine-grained semantic knowledge and enhances its capacity to discriminate between similar samples.

4 Experiment Setup

4.1 Datasets

In this study, we assess the performance of Moodle on three widely recognized benchmarks: DB15K, FB15K-237 and YAGO15K. They are derived from MMKG (Liu et al., 2019), which is a collection of multimodal knowledge graphs. And we adopt two types of evaluation metrics: Mean Reciprocal Rank (MRR) and Hits@K (K = 1, 3, 10). Key statistics about these datasets are detailed in Table 1.

Datasets	#Ent.	#Rel.	#Train	#Valid	#Test
DB15K	14,777	279	69,319	9,903	19,806
FB15K-237	14,541	237	272,115	17,535	20,466
YAGO15K	15,283	32	86,020	12,289	24,577

Table 1: Dataset Statistics

Model	DB15K				FB15K-237				YAGO15K			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
TransE	.256	.137	.329	.469	.279	.198	.376	.441	.161	.051	-	.384
DistMult	.351	.271	.393	.502	.301	.214	.370	.476	.324	.246	.372	.469
ConvE	.312	.219	-	.507	.312	.225	.341	.497	.267	.168	-	.426
ConvKB	.342	.276	.379	.460	.343	.259	.372	.514	.307	.241	.366	.454
RotatE	.382	.317	.414	.508	.304	.213	.335	.491	.342	.258	367	.473
ComplEx	.374	.305	.408	.505	.322	.229	.353	.511	.326	.250	.361	.471
IKRL	.268	.141	.349	.491	.309	.232	-	.493	.139	.048	-	.317
TransAE	.281	.213	.312	.412	-	.199	.317	.463	.253	.184	.289	.445
RSME	.297	.242	.321	.403	-	.242	.344	.467	.277	.226	.309	.412
MoSE-MI	.318	.252	.367	.486	.353	**.268**	.394	.540	.299	.223	.326	.449
MKGFormer	.346	.222	.398	.506	-	.256	.367	.504	301	.245	.338	.464
AdaMF-MAT	.351	.253	.411	.529	.343	.241	.360	.478	.344	.262	.369	.486
Moodle	**.434**	**.374**	**.462**	**.548**	**.360**	.266	**.396**	**.548**	**.377**	**.303**	**.411**	**.518**

Table 2: Results of various models on DB15K, FB15K-237, and YAGO15K datasets. The best results are marked bold and the second-best results are underlined in each column.

5 Experiments Results

5.1 Overall Performance

As shown in Table 2, we can observe that:

- Moodle achieves state-of-the-art or competitive performance across three widely-recognized benchmarks for MKGC. Specifically, Moodle consistently surpasses all baseline models on three datasets, achieving improvements on MRR of 5.2%, 0.7% and 3.3%, respectively. These findings substantiate that the incorporation of the CuLearner and MsLearner enables Moodle to attain exceptional performance on various MKGs.

- *In comparison with UML-based methods*, which may lose unimodal information, Moodle exhibits superior robustness. For instance, while AdaMF-MAT (Zhang et al., 2024) achieves the second-best performance on Hits@10, it underperforms several unimodal methods on Hits@1 and MRR metrics. In contrast, Moodle maintains consistent performance across all metrics.

- *Compared to MSL-based methods*, such as MoSE (Zhao et al., 2022) that learns knowledge from multiple modalities separately, Moodle employs a two-stage learning framework, achieving noticeable improvements, enabling the detection of subtle semantic differences between similar entities.

5.2 Ablation Study

To validate the effectiveness of the experimental design, we conducted ablation experiments as shown in Table 3.

Model	MRR	H@1	H@3	H@10
Moodle	**.377**	**.303**	**.411**	**.518**
Moodle w/o v	.351	.276	.384	.482
Moodle w/o t	.329	.255	.361	.473
Moodle w/o vt	.320	.245	.353	.464
Moodle w/o WL	.344	.275	.372	.487
Moodle w/o CAF	.352	.279	.384	.494
Moodle w/o MSL	.316	.241	.347	.465
Moodle w/ MSL with MS	.209	.136	.242	.350
Moodle w/o MFE	.197	.054	.031	.073
Moodle w/ MFE with ATT	.218	.144	.242	.371

Table 3: Ablation study on the benchmark of YAGO15K

5.2.1 Role of Multimodal Information

In order to validate the supplementary enhancement provided by the interaction of multimodal information for MKGC, ablation experiments were conducted on the YAGO15K dataset. These experiments involved three settings: using only structural information denoted as ‘w/o vt’, excluding textual information denoted as ‘w/o t’, and excluding visual information denoted as ‘w/o v’. The results are shown in the Table 3. We find that the absence of textual information led to a noticeable decline, and using only structural information showed similar results. These findings highlight the critical role of multimodal information in improving MKGC accuracy.

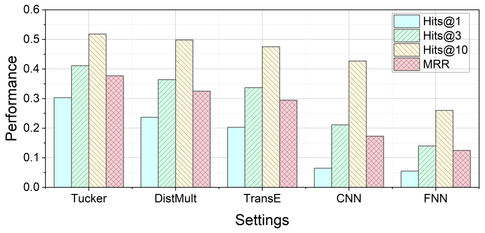


Figure 3: Various settings for the interaction methods between entities and relations.

5.2.2 Effect of Collaborative Unimodal Learner

To validate the role of the CuLearner, we conducted ablation experiments shown in Table 3. Firstly, we replaced the weight learner with direct addition, denoted as 'w/o WL'. The results indicated a decrease in all metrics, with Hits@10 dropping by 3.1%. Then we removed the cross-attention filter, denoted as 'w/o CAF,' and the results showed a decrease of 2.4% in Hits@10. Additionally, we explored various configurations for the interaction methods between entities and relations, presented on the Figure 3. Results indicated that Tucker decomposition outperformed all other methods, surpassing the second-highest by 2.0% in Hits@10.

5.2.3 Role of Multimodal Semantic Learner

To demonstrate the impact of the MsLearner on results, we conducted a series of ablation experiments. 'w/o MSL' refers to predictions using only unimodal knowledge. 'w/ MSL with MS' represents conducting MSL with multimodal features only, excluding distinct unimodal features. 'w/o MFE' denotes directly concatenating unimodal features as multimodal feature without feature extraction, while 'w/ MFE with ATT' uses the self-attention mechanism of Transformer for multimodal feature extraction.

As shown in Table 3, excluding or relying solely on unimodal knowledge both reduces performance due to inadequate interaction between modalities, lacking well-initialized features and task-relevant knowledge guidance during fusion. The results also show a noticeable performance drop when MFE is removed, with even Transformer-based extraction underperforming predictions using only unimodal knowledge. This highlights the crucial role of MFE in Moodle.

5.3 Case Study

Here, we demonstrate how Moodle recognizes semantically similar samples, thereby improving the Hits@K metric. Figure 4 presents the different prediction results of Moodle and AdaMF-MAT (Zhang et al., 2024) on DB15K. When using MsLearner, Moodle accurately identified "Sean_Bean" as the most relevant result, indicating its understanding of the semantic relationships and experiences of "Sean_Bean" to infer his birthplace. In contrast, AdaMF-MAT replaced "Sean_Bean" with "Sheffield_Wednesday_F.C.", possibly because it focused solely on literal similarity and did not grasp semantic connections. In a similar case, Moodle correctly predicted "Art_director", while AdaMF-MAT replaced it with "Audio_engineer", possibly influenced by the fact that "Peter_Lamont" was involved in the production of several films in MKG.

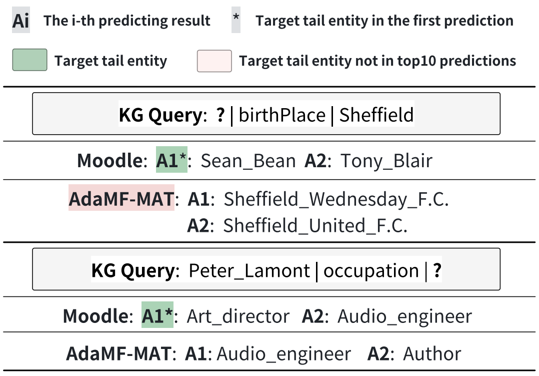

Ai The i-th predicting result
* Target tail entity in the first prediction
Target tail entity
Target tail entity not in top10 predictions
KG Query: ? | birthPlace | Sheffield
Moodle: A1*: Sean_Bean A2: Tony_Blair
AdaMF-MAT: A1: Sheffield_Wednesday_F.C.
A2: Sheffield_United_F.C.
KG Query: Peter_Lamont | occupation | ?
Moodle: A1*: Art_director A2: Audio_engineer
AdaMF-MAT: A1: Audio_engineer A2: Author

Figure 4: Comparison of Moodle and AdaMF-MAT in case study.

6 Conclusion

In this paper, we propose a Modal-collaborative knowledge learning framework for MKGC, Moodle, that enhances mutual collaboration of modalities during both unimodal knowledge extraction and multimodal knowledge fusion. Experimental results on several benchmark datasets demonstrate the effectiveness of our method. Additionally, we conducted an in-depth experimental analysis to validate the rationale behind our method design and a case study to show its potential value in applications. In the future, we plan to adopt pretrained vision-language models (VLM) to harness the idea of modal collaboration for more diverse MKG applications, such as multimodal knowledge reasoning, multimodal question answering, etc.

7 Limitations

Although Moodle shows promising results, several limitations exist. First, the initialization of features from different modalities is relatively simple and does not deeply leverage the knowledge from pre-trained models, which may limit the model's ability to capture richer representations. Second, the presence of false negative samples can hinder the model's convergence, and no specific design has been implemented to address this issue. Additionally, Moodle currently lacks thorough investigation into cross-modal interactions at deeper levels. Lastly, Moodle has not been extensively tested in diverse, real-world scenarios, which may limit its generalizability across various domains. We plan to solve these issues for Moodle as future work.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2787–2795, Red Hook, NY, USA. Curran Associates Inc.

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

Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22. ACM.

Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, Jiaqi Li, Xiaoze Liu, Jeff Z. Pan, Ningyu Zhang, and Huajun Chen. 2024. Knowledge graphs meet multi-modal learning: A comprehensive survey. *Preprint*, arXiv:2402.05391.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. *Preprint*, arXiv:1707.01476.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Chen hao, Xie Runfeng, Cui Xiangyang, Yan Zhou, Wang Xin, Xuan Zhanwei, and Zhang Kai. 2023. Lkpnr: Llm and kg for personalized news recommendation framework. *Preprint*, arXiv:2308.12028.

Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023. Imf: Interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023*, WWW '23. ACM.

Yankai Lin, Zhiyuan Liu, Huanbo Luan, Maosong Sun, Siwei Rao, and Song Liu. 2015a. Modeling relation paths for representation learning of knowledge bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 705–714, Lisbon, Portugal. Association for Computational Linguistics.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015b. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2181–2187. AAAI Press.

Danyang Liu, Jianxun Lian, Zheng Liu, Xiting Wang, Guangzhong Sun, and Xing Xie. 2021. Reinforced anchor knowledge graph generation for news recommendation reasoning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 1055–1065, New York, NY, USA. Association for Computing Machinery.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S. Rosenblum. 2019. Mmkg: Multi-modal knowledge graphs. *Preprint*, arXiv:1903.05485.

Bonaventure C. Molokwu and Ziad Kobti. 2020. Social network analysis using RLVECN: representation learning via knowledge-graph embeddings and convolutional neural-network. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 5198–5199. ijcai.org.

Bonaventure C. Molokwu, Shaon Bhatta Shuvo, Ziad Kobti, and Narayan C. Kar. 2020. Social network analysis using knowledge-graph embeddings and convolution operations. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 6351–6358. IEEE.

Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. 2018. A novel embedding model for knowledge base completion based on convolutional neural network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *Preprint*, arXiv:1409.1556.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *Preprint*, arXiv:1902.10197.

Théo Trouillon, Christopher R. Dance, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2017. Knowledge graph completion via complex tensor factorization. *Preprint*, arXiv:1702.06879.

Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022. Simkgc: Simple contrastive knowledge graph completion with pre-trained language models. *Preprint*, arXiv:2203.02167.

Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. 2021. Is visual context really helpful for knowledge graph? a representation learning perspective. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 2735–2743, New York, NY, USA. Association for Computing Machinery.

Yujie Wang, Hu Zhang, Jiye Liang, and Ru Li. 2023. Dynamic heterogeneous-graph reasoning with language models and knowledge representation learning for commonsense question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14048–14063. Association for Computational Linguistics.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, page 1112–1119. AAAI Press.

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

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied knowledge representation learning. *Preprint*, arXiv:1609.07028.

Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. 2022. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 3857–3866, New York, NY, USA. Association for Computing Machinery.

Bishan Yang, Wen tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. *Preprint*, arXiv:1412.6575.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2022. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *Preprint*, arXiv:2104.06378.

Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. 2022. Kg-fid: Infusing knowledge graph in fusion-in-decoder for open-domain question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4961–4974. Association for Computational Linguistics.

Yichi Zhang, Zhuo Chen, Lei Liang, Huajun Chen, and Wen Zhang. 2024. Unleashing the power of imbalanced modality information for multi-modal knowledge graph completion. *Preprint*, arXiv:2402.15444.

Yichi Zhang and Wen Zhang. 2022. Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling. *Preprint*, arXiv:2209.07084.

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

A Additional Resources

A.1 Pseudo-code

The training process of Moodle includes two stages, the first of which leverages multi-task learning to model mutual collaboration in unimodal knowledge extraction, while the second performs multimodal knowledge fusion guided by the unimodal knowledge obtained in the first. The pseudo-codes are presented in Algorithm 1 and 2, respectively.

In the first stage as shown in Algorithm 1, Moodle first utilizes three unimodal encoders ψ_{tuck}, ψ_{vgg} and ψ_{bert} to generate the entity embeddings $h^{\text{s}}, h^{\text{v}}, h^{\text{t}}$, during which a cross-attention filter is integrated to filter out irrelevant information from the visual embedding (*Line 1-7*). Then, Moodle employs Tucker decomposition to produce joint entity-relation embeddings $h^{\text{s}}_{ir}, h^{\text{v}}_{ir}, h^{\text{t}}_{ir}$ (*Line 8*). Next, these embeddings are optimized by $\mathcal{L}_{\text{joint}}$ using multi-task learning (*Line 9-11*).

In the second stage as shown in Algorithm 2, The three unimodal encoders ψ_{tuck}, ψ_{vgg} and ψ_{bert} are re-employed to generate entity embeddings that are concatenated and utilized to guide the knowledge fusion (*Line 1-5*). The multimodal features are then combined with the unimodel knowledge obtained in the first stage, producing the final relation-aware entity representation denoted as h^{cl}_{ir} (*Line*

6-7). The target and negative entity representations $h_j^{\mathrm{cl}}, h_k^{\mathrm{cl}}$ are obtained similarly except without the entity-relation interaction (*Line 8*). Last, these representations within a batch are used for contrastive learning, optimizing both unimodal and multimodal feature representations guided by the unimodal knowledge (*Line 9-10*).

Algorithm 1 Unimodal Knowledge Extraction

```
Input: Multimodel Knowledge Graph $\mathcal{G}$
Output: Trained Model $\mathcal{M}_1$
 1: Build structure encoder $\psi_{\mathrm{tuck}}$ by training TuckER model on $\mathcal{G}$ in Equation (1)
 2: Build visual encoder $\psi_{\mathrm{vgg}}$ and textual encoder $\psi_{\mathrm{bert}}$ using VGG16 and BERT-base
 3: Initialize all entity embeddings with the outputs of the above encoders
 4: while not converge do
 5:     Sample a batch of triples from $\mathcal{G}$
 6:     for each head entity $e_i$, relation $r$ in the batch do
 7:         Obtain the structural, visual, textual embeddings $h^{\mathrm{s}}, h^{\mathrm{v}}, h^{\mathrm{t}}$ of entity $e_i$
 8:         Compute the relation-aware entity embeddings $h_{ir}^{\mathrm{s}}, h_{ir}^{\mathrm{v}}, h_{ir}^{\mathrm{t}}$ by Equation (5)
 9:         Obtain all entity embeddings in entity set $h_k^{\mathrm{s}}, h_k^{\mathrm{v}}, h_k^{\mathrm{t}}$
10:         Compute the loss $\mathcal{L}_{\mathrm{s}}, \mathcal{L}_{\mathrm{v}}, \mathcal{L}_{\mathrm{t}}$ and $\mathcal{L}_{\mathrm{joint}}$ with unimodal scorers via Equation (6), (7) and (8)
11:         Update model parameters of $\mathcal{M}_1$
12:     end for
13: end while
14: Return $\mathcal{M}_1$
```

A.2 Baselines

To evaluate the performance of Moodle, we compare with both unimodal and multimodal baselines.

The unimodal baselines include:

- **TransE** (Bordes et al., 2013) models relations as translations between entities and uses an energy function to score relational triples.
- **DistMult** (Yang et al., 2015) models relations as diagonal matrices and scores relational triples by computing a bilinear product between entity and relation embeddings.
- **ConvE** (Dettmers et al., 2018) applies a convolutional neural network to 2D reshaped entity and relation embeddings.

Algorithm 2 Multimodal Knowledge Fusion

```
Input: MKG $\mathcal{G}$, Trained Model $\mathcal{M}_1$
Output: Trained Model $\mathcal{M}_2$
 1: while not converge do
 2:     Sample a batch of entities from $\mathcal{G}$
 3:     for each head entity $e_i$, relation $r$ in the batch do
 4:         Reacquire $h^{\mathrm{s}}, h^{\mathrm{v}}, h^{\mathrm{t}}$ of entity $e_i$ by $\psi_{\mathrm{tuck}}$, $\psi_{\mathrm{vgg}}$, $\psi_{\mathrm{bert}}$
 5:         Obtain trained relation-aware entity embeddings $\tilde{h}_{ir}^{\mathrm{s}}, \tilde{h}_{ir}^{\mathrm{v}}, \tilde{h}_{ir}^{\mathrm{t}}$ by Equation (5)
 6:         Obtain the multimodal feature $h_{ir}^{\mathrm{m}}$ by Equation (11) and (12)
 7:         Obtain $h_{ir}^{\mathrm{cl}}$ based on $h_{ir}^{\mathrm{m}}$, $\tilde{h}_{ir}^{\mathrm{s}}, \tilde{h}_{ir}^{\mathrm{v}}, \tilde{h}_{ir}^{\mathrm{t}}$ by Equation (13)
 8:         Obtain $h_j^{\mathrm{cl}}$ and $h_k^{\mathrm{cl}}$ similar to $h_{ir}^{\mathrm{cl}}$
 9:         Compute the loss $\mathcal{L}_{\mathrm{cl}}$ by Equation (9) and (10)
10:         Update model parameters of $\mathcal{M}_2$
11:     end for
12: end while
13: Return $\mathcal{M}_2$
```

- **ConvKB** (Nguyen et al., 2018) captures the complex interactions between entities and relations through the sliding window of convolutional filters to score relational triples.
- **RotatE** (Sun et al., 2019) represents relations as rotations in a complex vector space, scoring relational triples based on the distance between the rotated head entity and tail entity.
- **ComplEx** (Trouillon et al., 2017) extends DistMult to the complex vector space, allowing it to capture asymmetric relations by utilizing the Hermitian dot product for scoring relational triples.

The multimodal baselines include:

- **IKRL** (Xie et al., 2017) integrates visual information from images to enhance the representation of entities.
- **TransAE** (Wang et al., 2019) utilizes the reconstruction loss of autoencoder to facilitate information fusion across multimodals.
- **RSME** (Wang et al., 2021) automatically enhance or filter the influence of visual context during the representation learning by designing a Relation Sensitive Multi-modal Embedding model.

- **MoSE** (Zhao et al., 2022) learns modality-split relation embeddings for each modality and makes predictions then exploits various ensemble methods to combine the predictions in the inference phase.

- **AdaMF-MAT** (Zhang et al., 2024) proposes a modality-adversarial training strategy to generate synthetic multimodal embeddings and construct adversarial examples.

A.3 Datasets

DB15K, derived from DBPedia, comprises approximately 15,000 entities, each paired with images sourced from web searches, thereby offering valuable visual context for tasks based on multimodal knowledge graphs (MKGs). The dataset encompasses 279 distinct relation types and spans a wide array of domains, including individuals, organizations, and locations. Similarly, YAGO15K is constructed from the YAGO database and contains around 15,000 entities, each accompanied by images collected from online sources. This dataset features 32 relation types and covers various domains such as countries, films, and sports. FB15k-237 represents a curated subset of the Freebase database, containing roughly 15,000 entities and 237 relation types. This dataset addresses the issue of data leakage existing in the original FB15k, making it a more robust and reliable resource for tasks like link prediction in MKGs.

A.4 Evaluation metrics

We adopt two types of evaluation metrics: Mean Reciprocal Rank (MRR) and Hits@K (K = 1, 3, 10). MRR is calculated as:

$$\mathrm{MRR} = \frac{1}{2|\mathcal{T}_{test}|} \sum_{i=1}^{|\mathcal{T}_{test}|} \left(\frac{1}{\mathrm{r}_i^{\mathrm{h}}} + \frac{1}{\mathrm{r}_i^{\mathrm{t}}}\right), \qquad (13)$$

where $\mathcal{T}_{test}$ represents the test dataset and $|\mathcal{T}_{test}|$ denotes the number of triples in the test dataset. The $\mathrm{r}_i^{\mathrm{h}}$ and $\mathrm{r}_i^{\mathrm{t}}$ refer to the position at which the target entity (either the head or tail) is found in the ranked list generated by the KGC model. MRR ranges from 0 to 1, with higher values indicating better performance. Hits@K measures the proportion of triples for which the target entity appears in the top K positions of the ranked list. Hits@K is calculated as:

$$\mathrm{Hits@K} = \frac{1}{2|\mathcal{T}_{test}|} \sum_{i=1}^{|\mathcal{T}_{test}|} \mathrm{I}(\mathrm{r}_i^{\mathrm{h}} \leq K) + \mathrm{I}(\mathrm{r}_i^{\mathrm{t}} \leq K), \qquad (14)$$

where I denotes the indicate function. If the condition is true, the function value is 1, otherwise 0. K takes the value 1, 3 or 10. All metrics are computed by averaging over two directions: head entity prediction and tail entity prediction.

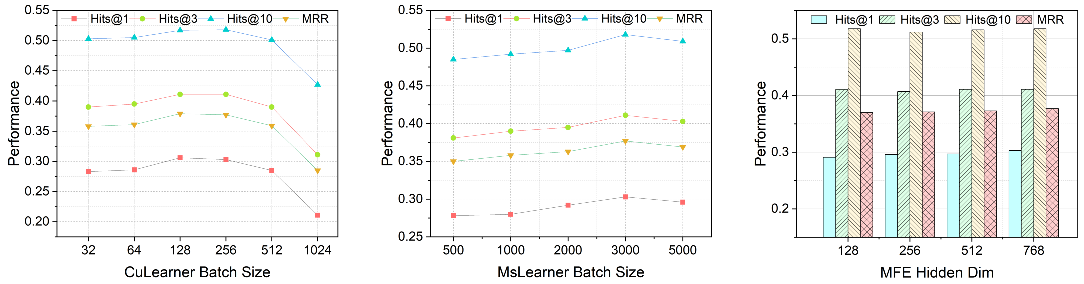
A.5 Implementation Details

We conducted all experiments on Intel(R) Xeon(R) CPU E5-2673 v4 CPU and a single NVIDIA GeForce RTX3060-12G GPU, using Ubuntu 9.4.0, Python 3.8.10 and PyTorch 1.11.0. Model parameters were initialized using Xavier initialization and optimized with the Adam optimizer. The hyperparameters were set as follows: The batch sizes are tuned in {64, 128, 256, 512} and {500, 1000, 2000, 3000, 5000} for CuLearner and MsLearner, respectively. For each modality, we set the embedding dim of entities and relations to 256, with a learning rate of 5e-3 and a dropout rate of 0.3. The learning rate of MsLearner is tuned in {1e-5, 2e-5, 2e-4, 2e-3}. The hidden dimension of multimodal feature extractor is tuned in {125, 256, 512, 768}. We follow the filtered setting in (Bordes et al., 2013) to evaluate MKGC performances.

A.6 Parameter Analysis

We empirically verify the imapct of the batch size in Collaborative Unimodal Learner (CuLearner) and Multimodal Semantic Learner (MsLearner) as well as the hidden dimension of multimodal feature extractor (MFE) on MKGC performances through a series of experiments. The results are shown in Figure 5(a), (b), (c).

Firstly, we explore the impact of the batch size of the Collaborative Unimodal Learner on model performance, setting it to 32, 64, 128, 256, 512, and 1024. As shown in Figure 5(a), with the growth of the batch size, the model performance initially improves but subsequently declines. When the batch size of the CuLearner exceeds 256, the model performance drops significantly. This may be due to limited training data, where an excessively large batch size can lead to local optimum, preventing the model from learning generalized semantic knowledge.



(a) Batch size of CuLearner (b) Batch size of MsLearner (c) Hidden dim of MFE

Figure 5: Performance analysis of Moodle w.r.t. different hyper-parameters.

Then, we investigate the impact of the batch size of Multimodal Semantic Learner (also the number of negative samples) on model performance, setting it to 500, 1000, 2000, 3000 and 5000. The experimental results are shown in Figure 5(b), where it is observed that the performance of Moodle improves with the increase of the number of negative samples, and stabilizes after reaching 3000.

Finally, we examine the impact of the hidden dimension in the multimodal feature extractor (MFE). The hidden dimension was varied across 128, 256, 512, 768, and 1024. The results, illustrated in Figure 5(c), reveal that the hidden layer dimensionality in the MFE model has a relatively minor impact on overall performance. While increasing the dimensionality leads to slight improvements in the Hits@1 and MRR metrics, other performance indicators remain largely unaffected. This suggests that beyond a certain point, further increasing the dimensionality offers limited benefits to the model's effectiveness.

A.7 Visualization Study

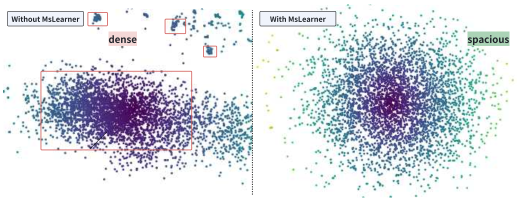


Figure 6: Comparison of Moodle with and without Multimodal Semantic Learner in visualization study.

To visually demonstrate the enhanced discriminative power of Moodle with MsLearner, we projected the entity vectors from test datasets of FB15K-237 into a two-dimensional space using PCA, as illustrated in Figure 6. The left part of Figure 6 shows embeddings obtained without MsLearner, while the right illustrates embeddings obtained with MsLearner. With the same scale applied, it is clear that embeddings without MsLearner tend to cluster closely together, making differentiation challenging. In contrast, embeddings trained with MsLearner display a more dispersed pattern, spreading from the center to the periphery, which indicates improved discriminative ability.