

Contrast then Memorize: Semantic Neighbor Retrieval-Enhanced Inductive Multimodal Knowledge Graph Completion

Yu Zhao

zhaoyu@dbis.nankai.edu.cn
 College of Computer Science
 VCIP, TMCC, TBI Center
 Nankai University
 Tianjin, China

Xinying Qian

qianxinying@dbis.nankai.edu.cn
 College of Computer Science
 VCIP, TMCC, TBI Center
 Nankai University
 Tianjin, China

Ying Zhang*

yingzhang@nankai.edu.cn
 College of Computer Science
 VCIP, TMCC, TBI Center
 Nankai University
 Tianjin, China

Baohang Zhou

zhoubaohang@dbis.nankai.edu.cn
 College of Computer Science
 VCIP, TMCC, TBI Center
 Nankai University
 Tianjin, China

Xiangrui Cai

caixr@nankai.edu.cn
 College of Computer Science
 VCIP, TMCC, TBI Center
 Nankai University
 Tianjin, China

ABSTRACT

A large number of studies have emerged for Multimodal Knowledge Graph Completion (MKGC) to predict the missing links in MKGs. However, fewer studies have been proposed to study the inductive MKGC (IMKGC) involving emerging entities *unseen during training*. Existing inductive approaches focus on learning textual entity representations, which neglect rich semantic information in visual modality. Moreover, they focus on aggregating structural neighbors from existing KGs, which of emerging entities are usually limited. However, the semantic neighbors are decoupled from the topology linkage and usually imply the true target entity. In this paper, we propose the IMKGC task and a semantic neighbor retrieval-enhanced IMKGC framework CMR, where the *contrast* brings the helpful semantic neighbors close, and then the *memorize* supports semantic neighbor retrieval to enhance inference. Specifically, we first propose a unified cross-modal contrastive learning to simultaneously capture the textual-visual and textual-textual correlations of query-entity pairs in a unified representation space. The contrastive learning increases the similarity of positive query-entity pairs, therefore making the representations of helpful semantic neighbors close. Then, we explicitly memorize the knowledge representations to support the semantic neighbor retrieval. At test time, we retrieve the nearest semantic neighbors and interpolate them to the query-entity similarity distribution to augment the final prediction. Extensive experiments validate the effectiveness of CMR on three inductive MKGC datasets. Codes are available at <https://github.com/OreOZhao/CMR>.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '24, July 14–18, 2024, Washington, DC, USA.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
 ACM ISBN 979-8-4007-0431-4/24/07
<https://doi.org/10.1145/3626772.3657838>

CCS CONCEPTS

- Computing methodologies → Knowledge representation and reasoning;
- Information systems → Multimedia information systems; Multimedia and multimodal retrieval.

KEYWORDS

Knowledge Graph; Multimodal Representation; Knowledge Graph Completion; Information Retrieval

ACM Reference Format:

Yu Zhao, Ying Zhang, Baohang Zhou, Xinying Qian, Kehui Song, and Xiangrui Cai. 2024. Contrast then Memorize: Semantic Neighbor Retrieval-Enhanced Inductive Multimodal Knowledge Graph Completion. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3626772.3657838>

1 INTRODUCTION

Multimodal Knowledge Graphs (MKGs) [18, 44] organize multimodal factual knowledge composed of entities and relations and have been widely applied to many knowledge-driven tasks [22, 28]. Due to the incompleteness of KGs, the Multimodal Knowledge Graph Completion (MKGC) task [38, 39] has been proposed to predict the missing entities in KGs. For a query $q = (h, r, ?)$ involving the head entity and relation, MKGC aims to predict the tail entity t with multimodal information, i.e. structure, text, and visual modalities. However, in the real world, new entities emerge through time and propose challenges for the MKGC model to be inductive and generalize to unseen entities without re-training. Therefore, we propose the Inductive Multimodal Knowledge Graph Completion (IMKGC) task, aiming to complete missing links involving unseen entities with multimodal information. As shown in Figure 1(a), *WALL-E* is unseen during training and IMKGC aims to predict its *production companies*.

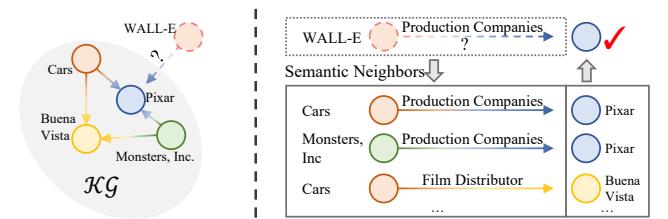
To incorporate multimodal information, many MKGC methods have been proposed based on knowledge graph embedding (KGE) methods [2, 15, 43], or pre-trained vision-language models (VLMs)

[4, 19]. However, they are not compatible with the inductive setting: the entity representation table of KGE-based methods and entity tokens of VLM-based methods all require re-training to generalize to emerging entities. The way to generalize the MKGC model to unseen entities without re-training is less explored.

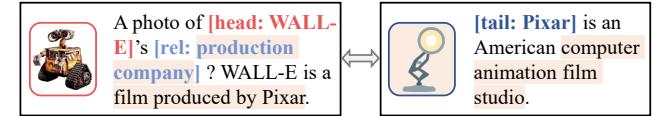
Although there have been inductive KGC studies [6, 23, 32] to predict the missing links involving unseen entities, they still have two main issues: (1) **They neglected that visual modality could help with inductive inference.** Existing inductive approaches learn the textual entity embedding based on pre-trained language models (PLMs) [8]. However, with the help of pre-trained visual models [9, 27], the visual representations could also transfer to emerging entities that are not seen during training. Moreover, the semantic correlation between entity images and queries could indicate the true entity. As shown in Figure 1(b), the texts “*production company*”, “*produced by Pixar*” from the query are not only related to the descriptions of *Pixar* but also corresponded to the image of *Pixar* logo. (2) **The helpful semantic neighbors are not explicitly aggregated.** Existing inductive approaches only aggregate structural information, which of the emerging entity is usually limited. The semantic neighbors, which *share similar semantics with the query and are not necessarily structurally linked*, usually imply the true target entity. As shown in Figure 1(a), the semantic neighbors of the query (*WALL-E, production companies*), such as (*Cars, production companies*) with similar semantics, directly indicate the target entity *Pixar*. The aggregation from such helpful semantic neighbors could help with inductive inference.

Though incorporating visual modality and semantic neighbors is intuitive for IMKGC, there still are several challenges: (1) **The multimodal contradiction.** As shown in Figure 1(b), the images of the head entity and tail entity in a triple are usually dissimilar, which contradicts the correlation in text modality. (2) **Semantic neighbor representation learning.** Unlike the structural neighbor sampling by topology linkage, the semantic neighbors are sampled by representation similarity. The helpful semantic neighbors with the same target could only be sampled to help the prediction if their representations are similar to the query. However, it is computationally expensive to train the representations of helpful semantic neighbors closer to the query than all other queries.

In this paper, we propose to address the challenges with *contrast then memorize* strategy, in which the *contrast* not only avoids the modality contradiction but also brings the helpful semantic neighbors together, and the *memorize* supports the semantic neighbor sampling and aggregation at test time. We propose a novel IMKGC framework **CMR** including three steps of Contrastive learning, Memorization, and Retrieval. Specifically, (1) **unified cross-modal contrastive learning** captures the textual-textual and textual-visual correlations of the positive query-entity pairs in a unified representation space, which prevents contradiction from the visual-visual dissimilarity. We employ a query-entity bi-encoder architecture, where the entity encoder is equipped with a visual mapping network to project the images into trainable prefixes in textual space. The contrastive learning optimizes the encoders to increase the similarity between positive query-entity pairs, which thereby makes the representations of helpful semantic neighbors with the same target entity close. (2) **Knowledge representation memorization** explicitly memorizes the query embeddings



(a) IMKGC aims to predict missing entity of query (*WALL-E, production companies*) where *WALL-E* is unseen during training. The semantic neighbors of the query could help with predicting *Pixar*.



(b) Multimodal semantic correlation (shading) between query and its target entity.

Figure 1: We propose to aggregate semantic neighbors to augment inductive MKGC (Figure 1(a)). Moreover, we propose to capture both textual-textual and textual-visual correlation in the query-entity pairs (Figure 1(b)).

in a knowledge store after training. Like the KG supports structural neighbor sampling, the knowledge store supports semantic neighbor sampling with the help of search tools [12]. (3) **Retrieval-augmented inference** linearly interpolates the query-entity similarity distribution with the retrieved semantic neighbors from the knowledge store at test time, which benefits from our proposed contrastive learning and avoids expensive semantic neighbor training. Experiments on three IMKGC datasets and three transductive MKGC datasets validate the effectiveness and generalizability of CMR. The contributions of the paper can be summarized as follows:

- To the best of our knowledge, we are the first to study the IMKGC to address the emerging entity generalization problem of MKGs. We propose to enhance IMKGC with textual-visual correlations and semantic neighbors.
- We propose a framework CMR: (1) **contrastive learning** avoids modality contradiction and brings helpful semantic neighbors close, (2) **memorization** explicitly stores semantic neighbors, and (3) **retrieval** aggregates the semantic neighbors at test time to augment inference.
- Extensive experiments demonstrate the effectiveness of CMR on both inductive and transductive datasets. Experiments also show that the proposed unified cross-modal contrastive learning and semantic neighbors retrieval could help with generalization to emerging entities.

2 RELATED WORK

2.1 Multimodal Knowledge Graph Completion

MKGC [38, 39] aims to utilize auxiliary multimodal information such as texts and images to complement the structure information for missing entity prediction. In our paper, we focus on the inductive MKGC involving emerging entities unseen during training.

Some MKGC methods augment the single-modal KGE methods [1, 14] to multimodal information [2, 15, 26, 33, 36, 43], whose representation table could not generalize to unseen entities. Recently,

some methods [4, 19] based on pre-trained VLMs transform the entity as additional tokens and perform MKGC as Masked Language Modeling (MLM) problem. However, the token embeddings of unseen entities are randomly initialized without training, thus it is hard for MLM to predict unseen entities. In our paper, we focus on optimizing encoders with contrastive learning and augmenting query-entity similarity inference with semantic neighbor retrieval instead of with MLM prediction to address the unseen entity generalization problem.

Moreover, some methods point out the multimodal contradiction problem [4, 33, 43] that the images of head and tail entities show low relevance. RSME [33] proposes to filter out noisy visual information with a filter gate. MoSE [43] proposes to split the multimodal relation representations while ensemble multimodal prediction to alleviate inter-modal interference. MKGformer [4] proposes a multi-level multimodal fusion mechanism to alleviate noisy information. In our paper, we propose to capture the textual-textual and textual-visual correlation between queries and entities with unified cross-modal contrastive learning to prevent interference from visual-visual dissimilarity.

2.2 Inductive Knowledge Graph Completion

There have been developing focuses on the IKGC problem to address the unseen entity problem. Some works focus on learning unseen entity representation by aggregating the structural neighbors to existing KG [5, 34]. Recently, with the development of PLMs such as BERT [8], some works propose to utilize transferable text information. They encode the query [32] or entity [6, 23] to text embedding with BERT. Then they learn the structural information by fine-tuning BERT after structural objectives [6, 32] such as TransE [1] or aggregating subgraph neighbors [23] through GNNs [10]. However, none of these methods explore entity images in learning generalizable entity representations. Moreover, emerging entities have fewer related triples thus the structural information is limited for inference. In our paper, we propose to utilize both visual and textual information for inductive knowledge representation learning. We also propose semantic neighbor retrieval-enhanced inference, which frees the model from structural linkage and involves semantic-similar but not necessarily structural-linked entities.

2.3 Retrieval-augmented Model for KG

Retrieval-augmented language models [13] aims to improve the model performance by explicitly memorizing the patterns in a data-store rather than implicitly memorizing them in model parameters for better generalization. Inspired by them, many studies utilize retrieval-augmented methods for knowledge graph construction tasks, such as relation extraction [3], named entity recognition [35], knowledge extraction [41].

Recently, k NN-KGE [42] proposes retrieval-augmented MLM entity prediction for KGC. However, the generalization of MLM entity prediction is limited and the knowledge representations are not explicitly trained to enhance the retrieval performance. In our paper, unlike only improving generalization through memorization [13, 42], we propose contrastive learning before memorization to increase the representation similarity of semantic neighbors with the same target to improve the retrieval-augmented inference.

3 PRELIMINARIES

3.1 Multimodal Knowledge Graphs

Multimodal knowledge graphs (MKGs) are knowledge graphs with multimodal information, such as images and texts. A KG could be denoted as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{T}\}$, where \mathcal{E} is the set of entities, \mathcal{R} is the set of relations, \mathcal{T} is the set of triples $\{(h, r, t)\} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$. For MKG, each entity $e \in \mathcal{E}$ has multimodal semantics from descriptions e_d and images e_v , and each relation $r \in \mathcal{R}$ has descriptions r_d .

3.2 Multimodal Knowledge Graph Completion

For a particular query $q = (h, r, ?)$, Multimodal Knowledge Graph Completion (MKGC) aims to predict the missing entity with multimodal information, which is ranking all the candidate entities $e \in \mathcal{E}$ and making the target entity t the highest. The reversed triple (t, r^{-1}, h) are also added for unifying the head prediction and tail prediction into tail prediction [1].

3.2.1 Transductive MKGC. Under the transductive setting, the entity set in the train set and the test set are both \mathcal{E} , meaning all the entities in the test set were seen in the training phase.

3.2.2 Inductive MKGC. Under the inductive setting, emerging entities appear in the test time, which are unseen in the training phase. We define the set of seen entities as $\hat{\mathcal{E}}$, the set of unseen entities as $\tilde{\mathcal{E}}$, and $\hat{\mathcal{E}} \cup \tilde{\mathcal{E}} = \mathcal{E}$, $\hat{\mathcal{E}} \cap \tilde{\mathcal{E}} = \emptyset$. The triples in the test set all involve unseen entities as $\tilde{\mathcal{T}} = \{(\tilde{e}, r, \hat{e}) \text{ or } (\hat{e}, r, \tilde{e}) | \tilde{e} \in \tilde{\mathcal{E}}, \hat{e} \in \hat{\mathcal{E}}, r \in \mathcal{R}\}$. The IMKGC model ranks both seen and unseen entities $e \in \mathcal{E}$ for predicting the target entity. We do not assume unseen relation [6].

4 METHODOLOGY

4.1 Overview

The CMR framework is shown in Figure 2. First, we reformulate the triples (h, r, t) in the train graph as prompts from queries $q = (h, r, ?)$ and entities t . We utilize a bi-encoder architecture equipped with a visual mapping network (VMN) to encode the queries and entities. We propose a unified cross-modal contrastive knowledge representation learning to shorten the distance between textual queries and multimodal entities to capture the textual-textual and textual-visual correlations of query-entity pairs in a unified representation space. After the contrastive learning, the semantic neighbors with the same target entity are also closer.

Then we construct a knowledge store (KS) that memorizes the query embeddings and their target entity labels. We also construct an entity store (ES) with entity embeddings with all the entities. For inductive inference, with a query $q = (h, r, ?)$, we first retrieve its k -nearest semantic neighbor queries in the KS according to representation distance. The distribution based on negative distance forms the probability of KS p_{KS} . Moreover, we compute the distance between the query and all the entities in the entity store to form the probability of ES p_{ES} . We linearly interpolate the p_{KS} with p_{ES} to produce the final prediction.

4.2 Unified Cross-modal Contrastive Learning

As stated by the previous studies in MKGs [4, 33, 43], the multimodal contradictions hinder the MKG representation learning,

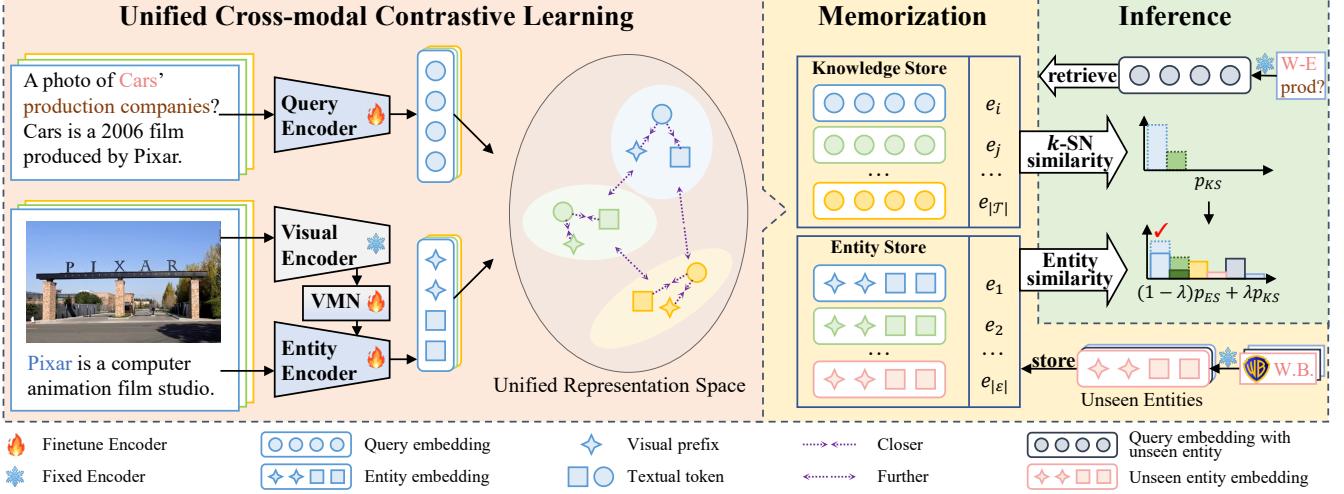


Figure 2: Contrastive learning, Memorization, and Retrieval (CMR) framework for Inductive Multimodal Knowledge Graph Completion. Contrastive learning optimizes the bi-encoders to shorten the distance between query-entity pairs and to capture multimodal semantic correlation in a unified representation space. Memorization explicitly memorizes the knowledge representations after training. The retrieval from the knowledge store aggregates the semantic neighbors and interpolates them to the query-entity similarity distribution for the final prediction at test time.

especially with the images between the head entity and tail entity not semantically similar. To effectively incorporate the visual modality, we propose to incorporate the visual modality by learning the cross-modal interactions between the textual queries and visual-textual target entities.

4.2.1 Query Encoder. The query encoder is a PLM used for encoding the texts of query $q = (h, r, ?)$. The input of the query encoder is the prompt that combines the descriptions of head entity h_d and relation r_d . The template of query is denoted as Equation (1), where h is the name of head entity.

$$T_q(h_d, r_d) = ["CLS] A photo of h's r_d? [SEP] h_d \quad (1)$$

We define the query encoder as $f_q(\cdot, \cdot)$ and the output is the L2-normalized mean aggregation over all the last-layer hidden states of the query encoder. The output of the query encoder is the query embedding, denoted as Equation (2).

$$\mathbf{q} = f_q(h, r) = f_q(T_q(h_d, r_d)) \quad (2)$$

4.2.2 Visual Prefix-based Entity Encoder. The entity encoder is a PLM used for encoding the target entity of a triple. In order to simultaneously capture the textual-textual and textual-visual correlations, inspired by the recent progress of visual prompt [17, 20] and prefix-tuning [16], we propose to convert the entity image representation from visual encoder into learnable visual prefixes in textual space. In this way, the textual tokens and their visual prefixes could interact inside the PLMs. To alleviate the modality gap from visual to text, we employ a visual mapping network (VMN) [24] between the visual encoder and the entity encoder. Then the visual prefixes and the entity description are fed into the PLM entity encoder to obtain the entity representations.

The image feature from visual encoder is $\mathbf{v}_e = f_v(e_v)$. The VMN projects the image feature into a l -length sequence of visual prefixes, denoted as Equation (3). We adopt a two-layer MLP activated with ReLU as VMN. Then, we concatenate the obtained visual prefixes embeddings to the token embeddings of entity descriptions as $[\mathbf{V}, \mathbf{D}] = (p_1, \dots, p_l, d_{l+1}, \dots, d_L)$, where L is the length of the sequence. The concatenation is fed into the PLM of the entity encoder.

$$\mathbf{V} = (p_1, \dots, p_l) = \text{VMN}(\mathbf{v}_e) = \text{VMN}(f_v(e_v)) \quad (3)$$

We also use the L2-normalized mean aggregation over all the last-layer hidden states of the entity encoder as entity embedding. We denote the multimodal fusion entity representation from the entity encoder as Equation (4). Since we utilize mean aggregation, the fusion entity representation could be denoted as the sum of the visual component and the textual component.

$$\mathbf{e}_f = f_e(e) = f_e([\mathbf{V}, \mathbf{D}]) = \mathbf{e}_v + \mathbf{e}_d \quad (4)$$

4.2.3 Training Objectives. The training objective of CMR is to optimize the encoders to increase the representation similarity between the query and its target entity in the positive triple (h, r, t) in the train graph so that the encoders could generalize to the emerging entity in the test phase. Moreover, the representations of the semantic neighbors with the same target entity would be close.

Fusion contrastive loss. Since the entity set \mathcal{E} is usually too large to put in a batch, we utilize the contrastive loss InfoNCE [25] to increase the similarity between positive query-entity pairs and decrease that between the negative pairs. We first propose a fusion contrastive loss \mathcal{L}_{FC} . We denote the similarity function between query embedding and entity embedding as $s = \exp(\mathbf{q} \cdot \mathbf{e}/\tau)$, where τ is a temperature hyper-parameter that adjusts the importance of negative pairs. Since the embeddings are L2-normalized, the dot product equals the cosine similarity metric. The fusion contrastive loss for the i -th triple is calculated as Equation (5), where \mathcal{N}^i is the

set of negative samples for the i -th triple.

$$\begin{aligned}\mathcal{L}_{FC}^i &= -\log \frac{s_f^i}{s_f^i + \sum_{n \in \mathcal{N}_i} s_f^n} \\ &= -\log \frac{\prod_{m \in \mathcal{M}} s_m^i}{\prod_{m \in \mathcal{M}} s_m^i + \sum_{n \in \mathcal{N}_i} s_f^n}\end{aligned}\quad (5)$$

Remark 1: Here we analyze how the \mathcal{L}_{FC} handles the hard positive samples that have contradictory similarity values between modalities. The similarity between query and fusion embedding could be seen as the product of all the similarities of multiple modalities as Equation (6), where we denote \mathcal{M} as a modality set without loss of generality for more modalities.

$$s_f = \exp(\mathbf{q} \cdot (\mathbf{e}_v + \mathbf{e}_d) / \tau) = s_v \cdot s_d = \prod_{m \in \mathcal{M}} s_m, \mathcal{M} = \{\mathcal{V}, \mathcal{D}\} \quad (6)$$

The hard positive samples are the ones that have at least one hard modality $h \in \mathcal{M}$ that s_h is small. For the modality h , the gradient from \mathcal{L}_{FC}^i to s_h^i is calculated as Equation (7). Therefore, even if only one of the modalities has a small similarity, the hard positive samples can be trained with large updates from small s_h^i and s_f^i in the denominator.

$$\begin{aligned}\frac{\partial \mathcal{L}_{FC}^i}{\partial s_h^i} &= -\frac{\sum_{n \in \mathcal{N}_i} s_f^n}{s_h^i (\prod_{m \in \mathcal{M}} s_m^i + \sum_{n \in \mathcal{N}_i} s_f^n)} \\ &= -\frac{\sum_{n \in \mathcal{N}_i} s_f^n}{s_h^i (s_f^i + \sum_{n \in \mathcal{N}_i} s_f^n)}\end{aligned}\quad (7)$$

Pre-align contrastive loss. We propose a pre-align contrastive loss \mathcal{L}_{AC} as Equation (8) for guiding the VMN module to project the image features into more accurate visual prefixes. The similarity between query embedding and the projected prefixes is denoted as $s_p = \exp(\mathbf{q} \cdot \bar{\mathbf{V}} / \tau)$ where $\bar{\mathbf{V}}$ is the mean pooling of visual prefix embeddings. The pre-align contrastive loss \mathcal{L}_{AC} captures the textual-visual correlation between queries and entities. It further enhances the fusion contrastive loss \mathcal{L}_{FC} as a middle-step constraint and improves the final prediction.

$$\mathcal{L}_{AC}^i = -\log \frac{s_p^i}{s_p^i + \sum_{n \in \mathcal{N}_i} s_p^n} \quad (8)$$

Overall objectives. In addition to the query-to-entity losses, we consider bidirectional learning objectives and add the reversed entity-to-query losses $\mathcal{L}_{FC'}$ and $\mathcal{L}_{AC'}$. For a batch \mathcal{B} , the overall training objective is as Equation (9).

$$\mathcal{L} = \sum_{\hat{\mathcal{T}} \in \mathcal{B}} (\mathcal{L}_{FC}^i + \mathcal{L}_{AC}^i + \mathcal{L}_{FC'}^i + \mathcal{L}_{AC'}^i) \quad (9)$$

4.2.4 Model Training and Adaptation. In our work, we maintain a queue of target entities encoded from preceding mini-batches [11]. For each batch, we enqueue the current batch of entities and dequeue the oldest. For a particular sample, the samples in the queue except itself are seen as negative samples, including in-batch and preceding-batch samples. In this way, the negative samples are decoupled from the mini-batch size and the encoder learns to distinguish the ground truth entity from more negative entities. Moreover, we filter out the triples with the same ground truth entity with masks to avoid contradiction. In implementation, we calculate

Algorithm 1: Contrastive learning procedure in CMR

```

Input : MKG  $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{T}\}$ , Train Triples  $\hat{\mathcal{T}}$ , Batch  $\mathcal{B}$ ,  
queue of entities in preceding batches  $\mathcal{Q}$  and  $\mathcal{Q}_v$ 
1 for  $\mathcal{B} = \{(h, r, t)\} \in loader(\hat{\mathcal{T}})$  do
2 | Calculate query embedding  $\mathbf{q}$  of  $(h, r)$  as Equation (2)
3 | Calculate visual prefixes  $\mathbf{V}$  of entity  $t$  as Equation (3)
4 | Calculate token embeddings  $\mathbf{D}$  of entity  $t$ 
5 | Calculate entity embedding  $\mathbf{e}_f$  as Equation (4)
6 | Enqueue( $\mathcal{Q}$ ,  $\mathbf{e}_f$ ), Dequeue( $\mathcal{Q}$ )
7 | Enqueue( $\mathcal{Q}_v$ ,  $\bar{\mathbf{V}}$ ), Dequeue( $\mathcal{Q}_v$ )
8 | Calculate positive logits  $s_f$  and  $s_p$ 
9 | Calculate negative logits  $s_f^n$  and  $s_p^n$ 
10 | for  $(h^i, r^i, t^i) \in \mathcal{B}$  do
11 | | Construct the negative sample mask  $M$  from  $\mathcal{N}^i$  as
12 | |  $\mathcal{N}^i = \{(h^i, r^i, n) | n \in \mathcal{Q}, (h^i, r^i, n) \notin \hat{\mathcal{T}}\}$ 
13 | | Mask the negative logits with  $M$ 
14 | | Optimize the model with overall loss as Equation (9)
15 end
```

the negative logits as $s^n = \exp(\mathbf{q} \cdot \mathbf{e}^{queue} / \tau)$ where \mathbf{e}^{queue} are the queue of preceding batch entity embeddings. Then we mask the positive and negative logits of the non-negative entities. The training procedure is illustrated in detail in Algorithm 1.

For parameter-efficient pre-trained model adaptation, we fix the parameter of the visual encoder and only fine-tune the parameters of VMN and the PLMs of the query encoder and entity encoder. The efficient adaptation of the pre-trained vision model makes the trainable parameter size of our framework similar to PLM-based methods. We explore the effect of different PLMs in Section 5.6.

4.3 Knowledge Representation Memorization

We propose to memorize the knowledge representations in a Knowledge Store (KS) as Equation (10) with query embeddings as keys and their entity labels as values.

$$(\mathcal{Q}, \mathcal{E}) = \{(f_q(h, r), e) | (h, r, e) \in \mathcal{T}\} \quad (10)$$

We also memorize entity embeddings and entity labels in the Entity Store (ES) as Equation (11).

$$(\mathcal{E}, \mathcal{E}) = \{(f_e(e), e) | e \in \mathcal{E}\} \quad (11)$$

During the store construction process, the query encoder and entity encoder are both fixed without further training.

Remark 2: Here we analyze how the contrastive learning of CMR improves the performance of semantic neighbor retrieval. In our work, the similarity between the query and its target entity embedding is enlarged after contrastive learning. Consequently, the queries with the same target entity should also be similar. That is, if a query $q = (h, r, ?)$ and $q' = (h', r', ?)$ share the same target entity t , $s(q, q')$ should be large since both $s(q, t)$ and $s(q', t)$ are large. Thus, if we retrieve the semantic neighbor q'' of q that $s(q, q'')$ is large, the target entity of q'' is highly probable to be t and helps with the inference of q . Thus our proposed contrastive learning is crucial for the retrieved semantic neighbors to help the inference.

4.4 Retrieval-augmented Inductive Inference

For inductive inference, given the query $q = (h, r, ?)$, we first get the query embedding $f_q(h, r)$. Then we get two output distributions over all the entities p_{KS} and p_{ES} .

For p_{KS} , we query the KS with $f_q(h, r)$ to retrieve its k -nearest semantic neighbors SN according to the representation distance. We implement with L2 distance retrieval with FAISS [12], which is equal to cosine similarity search. Then, we apply softmax on the negative distances and aggregate probability over all the entities in the entity set \mathcal{E} as Equation (12). The entities that are not in the SN have zero probability. In the implementation, we only contain one nearest semantic neighbor in p_{KS} for each target e_i [13, 42].

$$p_{KS}(y|(h, r)) \propto \sum_{(\mathbf{q}_i, e_i) \in SN} \mathbb{1}_{y=e_i} \exp(-d(f_q(h, r), \mathbf{q}_i)) \quad (12)$$

For p_{ES} , we retrieve all the entities in the ES and rank them according to their similarity to the query. We apply softmax on the similarity distribution to obtain the p_{ES} as Equation (13).

$$p_{ES}(y|(h, r)) \propto \exp(s(f_q(h, r), f_e(e_i))), e_i \in ES \quad (13)$$

The final output distribution is the linear interpolation of p_{ES} with p_{KS} with a hyper-parameter λ as Equation (14).

$$p(y|(h, r)) = \lambda p_{KS}(y|(h, r)) + (1 - \lambda)p_{ES}(y|(h, r)) \quad (14)$$

5 EXPERIMENTS

5.1 Experimental Setting

5.1.1 Datasets. We evaluate our proposed method on three inductive multimodal KGC datasets. For the inductive datasets, we augment the existing IKG datasets [6] **FB15K-237 ind** and **WN18RR ind** with entity images following previous studies [4, 43] and entity descriptions [6, 40]. Moreover, we follow Daza et al. to transform an existing MKGC dataset WN9 [39] to the inductive setting as **WN9 ind**. The datasets split the entity set \mathcal{E} and the triple set \mathcal{T} into disjoined train/valid/test sets. The train graph consists of triples with entities only in the \mathcal{E}_{train} . The valid or test triples consist of triples with the entities in $\mathcal{E}_{valid} \cup \mathcal{E}_{train}$ or $\mathcal{E}_{test} \cup \mathcal{E}_{valid} \cup \mathcal{E}_{train}$ respectively [6, 23]. The valid and test triples have unseen entities at the position of the head, tail, or both [6]. The dataset statistics are shown in Table 1.

We also conducted experiments on three widely used transductive MKGC datasets: **FB15K-237**, **WN18**, and **WN9**. Under the

Table 1: Dataset statistics of inductive MKGC.

\mathcal{R}	Train		Valid		Test	
	\mathcal{E}	\mathcal{T}	\mathcal{E}	\mathcal{T}	\mathcal{E}	\mathcal{T}
FB15K-237 ind	237	11,633	215,082	1,454	42,164	1,454
WN18RR ind	11	32,755	69,585	4,094	11,381	4094
WN9 ind	9	5,245	11,357	655	1,512	655
						1,528

Table 2: Dataset statistics of transductive MKGC.

Dataset	\mathcal{R}	\mathcal{E}	Train	Valid	Test
FB15K-237	237	14,541	272,115	17,535	20466
WN18	18	40,943	141,442	5,000	5,000
WN9	9	6,555	11,741	1,337	1,319

transductive setting, all the entities in the test set appear in the training set. The dataset statistics are shown in Table 2.

5.1.2 Evaluation. We evaluate each test triple with head entity prediction and tail entity prediction [1] and calculate the rank of the target entity compared to all the candidate entities from both seen and unseen entities. Four metrics are used for evaluation: Mean Reciprocal Rank (MRR), the mean of the reciprocal of the target entity, and Hits@K ($K=1,3,10$), the accuracy in the top K predicted entities. Higher MRR and higher Hits@K indicate better performance. We adopt the filtered setting [1] to exclude other target entities of the same query when calculating evaluation metrics.

5.1.3 Implementation details. We implement our framework with PyTorch and Transformers [37]. We exploit BERT [8] as query encoder and entity encoder that does not share parameters, and exploit ViT [9] as visual encoder. All entities have description information and the entities without images are padded with random values. The visual prefix length is 4 for all datasets. The total sequence length including prefixes is 50. The temperature τ is 0.05. The mini-batch size is 768 and we maintain a queue size of 3 mini-batches, consisting of 2 preceding batches and 1 current batch. We use AdamW optimizer with linear learning rate decay. The training is early stopped after 3 epochs without improvements in Hits@1 on the validation set. At test time, we fixed the encoders and memorized all the triples with both seen and unseen entities in KS. During semantic neighbor retrieval, we filter out the semantic neighbors with the same head and relation to prevent information leaks. We searched the hyperparameter λ and k in the validation set and selected the ones with the best results. We ran our experiments on 1 A6000 GPU (48GB RAM).

5.1.4 Baselines. For the inductive baselines, we employ two groups of baselines: (1) KGE-based methods: **DKRL** [38], **BLP** [6], **MoSE** [43], (2) PLM-based methods: **StAR** [32], **kNN-KGE** [42], **StATIK** [23]. In the baselines, the IKG baselines usually model two modalities: structure and text, including DKRL, BLP, StAR, kNN-KGE, and StATIK, while the MKGC baseline models three modalities: structure, text, and visual, including MoSE. Since there is no inductive MKGC methods, we extend the MKGC method MoSE to the inductive setting by utilizing text and visual embeddings and randomly initializing their structure embeddings. We did not include the VLM-based MKGC methods [4, 19] because they could not directly transfer to the inductive setting without re-training. We compare the baselines with our proposed CMR and the variant without visual modality, CMR-text, for a fair comparison with the textual IKG baselines.

For the transductive baselines, we compare CMR with two groups of representative MKGC baselines: (1) KGE-based methods: **TransAE** [36], **RSME** [33], **MoSE** [43], and (2) VLM-based methods: **MKG-former** [4], **SGMPT** [4].

5.2 Main Results

As shown in Table 3, CMR outperforms all the baselines on three inductive datasets, FB15K-237 ind, WN18RR ind, and WN9 ind.

Compared to KGE-based IKG methods [6, 38], CMR achieves 17.6%, 3.6%, and 42.4% improvements in MRR compared to BLP-TransE on three datasets respectively. Compared to MKGC methods

Table 3: Results of inductive multimodal knowledge graph completion. The best and second best results of each column are highlighted. The results with * are based on our re-implementation with their source code. Other results are from the original paper. “-text” means only text modality is used. The \mathcal{L}_{AC} denotes the pre-align contrastive loss. D denotes the description tokens in the entity encoder. The p_{KS} and p_{ES} denote the probability from knowledge store and entity store, respectively.

Model	FB15K-237 ind				WN18RR ind				WN9 ind*			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
GloVe-BOW	0.172	0.099	0.188	0.316	0.170	0.055	0.215	0.405	0.301	0.168	0.366	0.558
BE-BOW	0.173	0.103	0.184	0.316	0.180	0.045	0.244	0.450	0.263	0.122	0.336	0.526
GloVe-DKRL	0.112	0.062	0.111	0.211	0.115	0.031	0.141	0.282	0.191	0.085	0.228	0.402
BE-DKRL	0.144	0.084	0.151	0.263	0.139	0.048	0.169	0.320	0.241	0.153	0.260	0.425
BLP-TransE	0.195	0.113	0.213	0.363	0.285	0.135	0.361	0.580	0.434	0.314	0.496	0.672
BLP-DistMult	0.146	0.076	0.156	0.286	0.248	0.135	0.288	0.481	0.208	0.079	0.240	0.498
BLP-ComplEx	0.148	0.081	0.154	0.283	0.261	0.156	0.297	0.472	0.314	0.215	0.343	0.515
BLP-SimplE	0.144	0.077	0.152	0.274	0.239	0.144	0.265	0.435	0.328	0.218	0.377	0.541
MoSE-AI*	0.105	0.077	0.114	0.156	0.056	0.033	0.058	0.098	0.143	0.073	0.160	0.271
MoSE-BI*	0.206	0.137	0.221	0.348	0.318	0.213	0.362	0.527	0.323	0.204	0.370	0.568
MoSE-MI*	0.228	0.137	0.254	0.410	0.335	0.228	0.379	0.554	0.364	0.245	0.408	0.615
StAR	0.163	0.092	0.176	0.309	0.321	0.192	0.381	0.576	0.209	0.094	0.239	0.462
kNN-KGE	0.198	0.146	0.214	0.293	0.294	0.223	0.320	0.431	-	-	-	-
StATIK	0.224	0.143	0.248	0.381	0.516	0.425	0.558	0.690	-	-	-	-
CMR-text	0.349	0.259	0.396	0.523	0.608	0.534	0.632	0.779	0.858	0.811	0.897	0.939
CMR	0.371	0.269	0.418	0.567	0.631	0.577	0.644	0.788	0.879	0.848	0.898	0.932
w/o \mathcal{L}_{AC}	0.356	0.263	0.395	0.541	0.601	0.542	0.626	0.779	0.858	0.813	0.886	0.932
w/o D	0.298	0.215	0.338	0.466	0.531	0.503	0.547	0.577	0.842	0.835	0.843	0.853
w/o p_{ES}	0.341	0.235	0.390	0.553	0.582	0.549	0.596	0.642	0.818	0.818	0.818	0.818
w/o p_{KS}	0.258	0.177	0.281	0.422	0.424	0.300	0.489	0.664	0.509	0.386	0.579	0.740

[43], CMR achieves 14.3%, 29.6%, and 49.4% improvements in MRR compared with MoSE-MI on three datasets respectively. The reason could be that the structural information from KGE scores without re-training is limited for generalizing to unseen entities.

Compared to PLM-based methods [23, 32, 42], CMR achieves 14.7% and 11.5% improvements in MRR on FB15K-237 ind and on WN18RR ind compared with state-of-the-art baseline StATIK. Moreover, the textual version CMR-text outperforms StATIK by 12.5% and 9.2% on FB15K-237 ind and WN18RR ind. Moreover, CMR and CMR-text both outperform kNN-KGE, which utilizes retrieval-augmented MLM inference. The reason could be that the CMR has better generalization ability since our proposed contrastive learning directly optimizes the query-entity similarity and leads the semantic neighbors with the same ground truth entity to be closer. Thus the performance of retrieved semantic neighbors from CMR is boosted.

Compared with CMR-text, CMR achieves 2.2%, 2.3%, and 18.3% improvements in MRR on three datasets, demonstrating that visual modality is helpful for IKGC. Moreover, the addition of visual modality does not interfere with the performance of CMR-text, which shows that CMR could alleviate the modality contradiction. The improvements on WN9 are relatively large because (1) the image are high-qualified [39], and (2) CMR took advantage of the unremoved reversed triples like other two datasets [7, 30].

5.3 Ablation Study

The ablation study is shown in Table 3.

To evaluate the effectiveness of our proposed contrastive learning, we ablate the \mathcal{L}_{AC} and D in the entity embeddings. The performance of CMR decreases without \mathcal{L}_{AC} on three datasets. The reason could be that the pre-alignment of queries and VMN-mapped images increases the generalization of VMN. Moreover, the textual-visual alignment of \mathcal{L}_{AC} further improves the textual-multimodal alignment of \mathcal{L}_{FC} . CMR w/o D shows the results from only textual-visual contrastive learning and outperforms most IMKGC baselines. It indicates that the textual-visual correlation could help the IMKGC. However, CMR-text outperforms CMR w/o D, where the former captures textual-textual correlation. It demonstrates that the entity descriptions provide more helpful information than entity images. Moreover, CMR outperforms CMR-text and CMR w/o D, showing that CMR could effectively capture multimodal correlations.

To evaluate the effectiveness of the semantic neighbor retrieval-enhanced inference, we separately report the results of p_{KS} and p_{ES} . We report the best performance of CMR w/o p_{ES} , i.e. p_{KS} , throughout the number of semantic neighbors k . The CMR w/o p_{KS} and CMR w/o p_{ES} both underperform CMR, demonstrating the necessity of both semantic neighbors and direct query-entity similarity measures. The reason for best p_{KS} outperforms p_{ES} could be that semantic neighbors aggregate more helpful information than the triple itself. Moreover, CMR w/p p_{KS} outperforms the baselines on FB15K-237 ind and WN9 ind, demonstrating that the unified cross-modal contrastive learning effectively utilizes multimodal information for inductive MKGC.

Table 4: Results of transductive multimodal knowledge graph completion. The best and second best results of each column are highlighted. The results of the baselines are from their original paper.

Model	FB15K-237				WN18				WN9			
	MR	Hits@1	Hits@3	Hits@10	MR	Hits@1	Hits@3	Hits@10	MR	Hits@1	Hits@3	Hits@10
TransAE	431	0.199	0.317	0.463	352	0.323	0.835	0.934	17	-	-	0.942
RSME	417	0.242	0.344	0.467	223	0.943	0.951	0.957	55	0.878	0.912	0.923
MoSE-MI	<u>127</u>	0.268	<u>0.394</u>	<u>0.540</u>	<u>7</u>	<u>0.948</u>	0.962	0.974	<u>4</u>	<u>0.909</u>	<u>0.937</u>	<u>0.967</u>
MKGformer	221	0.256	0.367	0.504	28	0.944	0.961	0.972	-	-	-	-
SGMPT	238	0.252	0.370	0.510	29	0.943	<u>0.966</u>	<u>0.978</u>	-	-	-	-
CMR	123	<u>0.263</u>	0.395	0.543	<u>17</u>	0.993	0.995	0.997	1	0.916	0.983	0.997

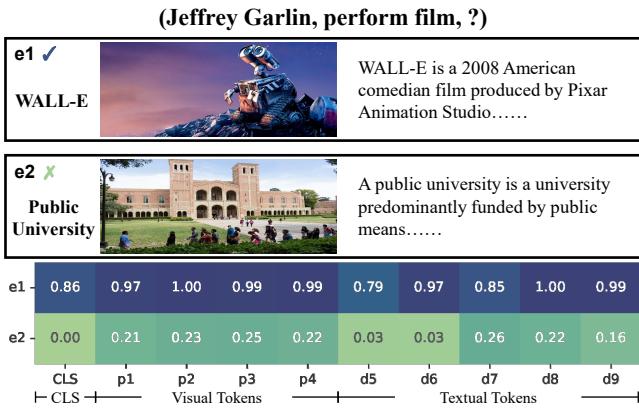


Figure 3: Token-wise similarity of query and ground-truth entity e_1 versus query and irrelevant entity e_2 .

5.4 Transductive Results

We conducted experiments under the transductive setting to evaluate the effectiveness of CMR in Table 4. The transductive MKGC could also benefit from the generalization ability of CMR since there are always unseen or less-seen entities in KGs [29].

CMR reaches the best on most metrics compared with baselines on three transductive datasets. Compared with the KGE-based methods, CMR outperforms the SOTA method MoSE-MI in most metrics. Compared with VLM-based methods, CMR outperforms all the baselines. The reason could be that our contrastive learning could avoid modality contradiction and the semantic neighbors could help the entities with limited structural information. The transductive MKGC could also benefit from the generalization ability of CMR since there are always unseen or less-seen entities in KGs [29]. CMR significantly outperforms the baselines on WN18 and WN9 datasets. The reason could be that the WN18 and WN9 also have unremoved triples problems, thus the semantic neighbor retrieval could help predict this kind of triples.

5.5 Case Study

5.5.1 Inductive cross-modal correlations. We visualize the token-wise similarity between query embedding and the token embeddings of the ground-truth entity and irrelevant entity in Figure 3, where both entities are unseen during training. The similarity of visual tokens $p_1 - p_4$ and textual tokens $d_5 - d_9$ of ground-truth entity

Table 5: The top-3 prediction of p_{ES} , \mathcal{SN} of the query, p_{KS} and CMR with unseen entity $WALL-E$ as query and target respectively. True target entities are boldfaced.

	As query	As target
Query	(WALL-E , genre)	(Pixar , nominated)
p_{ES}	1. Humour 2. Screwball comedy 3. Comedy	1. A Bug's Life 2. Toy Story 2 3. Ratatouille
\mathcal{SN}	1. (Lilo & Stitch, genre) 2. (Shrek 2, genre) 3. (Shrek, genre)	1. (Pixar, product $^{-1}$) 2. (Pixar, product $^{-1}$) 3. (Stanton, nominated)
p_{KS}	1. Science Fiction 2. Fairy tale 3. Costume drama	1. WALL-E 2. Toy Story 3 3. Ratatouille
CMR	1. Science Fiction 2. Fairy tale 3. Costume drama	1. WALL-E 2. A Bug's Life 3. Toy Story 2

e_1 are all high while those of the irrelevant entity e_2 are all low. It shows that CMR could dynamically capture the textual-textual and textual-visual correlation between the query and multimodal tokens of entities. It also indicates that CMR could distinguish the relevant entity from the irrelevant ones by effective unified cross-modal contrastive learning.

5.5.2 Semantic neighbors Retrieval. As shown in Table 5, we conducted a case study to explore how explicit memorization and semantic neighbor retrieval helps inductive inference with unseen entities as query and as target, respectively.

With an unseen entity in the query, the retrieval of semantic neighbors finds the nearest similar queries in KS that could imply the true target entity label. For example, like *WALL-E*, *Lilo & Stitch* is also a film about space and they both belong to *Science Fiction* genre. Moreover, the semantic neighbors are not limited by the topology linkage, since *Lilo & Stitch* is not directly linked to *WALL-E*.

With an unseen entity as the target, the semantic neighbors could also imply the true target label. For example, the films *Pixar* nominated for are usually the films produced by *Pixar*. Thus the (*Pixar*, $product^{-1}$, *WALL-E*) in the test graph provide the true target entity of (*Pixar*, *nominated*). It also shows that the structural neighbors are implicitly modeled by CMR when they are also semantic neighbors of the query.

Table 6: Results of CMR on WN9 ind with different PLMs.

Model	Paras	WN9 ind			
		MRR	Hits@1	Hits@3	Hits@10
CMR-BERT _{base}	220.9M	0.858	0.813	0.886	0.932
CMR-CLIP _{base}	151.3M	0.862	0.825	0.890	0.926
CMR-RoBERTa _{base}	251.3M	0.869	0.827	0.899	0.934

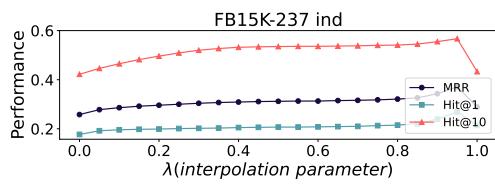
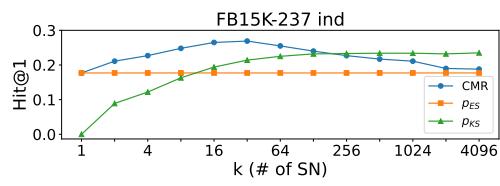


Figure 4: Performance with different semantic neighbor count k and interpolation parameter λ on FB15K-237 ind.

5.6 Effect of Pretrained Encoders

We select BERT-base as our query encoder and entity encoder for a fair comparison with the BERT-based baselines. To explore the ability of CMR with different pretrained encoders, we conducted experiments by replacing the BERT with widely-used PLMs, CLIP-base [27] and RoBERTa-base [21] in Table 6. With BERT and RoBERTa, the ViT remains fixed and only the BERT/ROBERTa and the VMN are trained, while with CLIP we finetuned both the visual and textual encoder of CLIP considering its low parameter. As shown in Table 6, the performance of different PLMs varies in a certain range on WN9 ind. The improvements from CLIP_{base} could be attributed to the visual-textual contrastive pretraining, and that from RoBERTa_{base} could be attributed to robust pretraining.

5.7 Effect of Retrieval Parameters

We conduct semantic neighbor retrieval parameter analysis of k and λ on FB15K-237 ind as shown in Figure 4.

5.7.1 Parameter k . The parameter k controls the number of semantic neighbors retrieved and aggregated to the final prediction. As shown in Figure 4(a), the performance of p_{KS} tends to increase and converge to the best performance with more semantic neighbors. The reason could be that with more semantic neighbors, the included informative triples tend to converge to the best performance. The performance of CMR first increases then decreases, and finally converges to the performance of p_{ES} . The reason could be that with more semantic neighbors, the number of noisy triples also increases, thus the combination of p_{ES} and p_{KS} tends to decrease. However, the CMR consistently outperforms the p_{ES} , indicating the effective improvements brought by the aggregation of semantic neighbors. Moreover, by adjusting the number of retrieved semantic neighbors, CMR could get the best performance and control the negative impact of the noisy semantic neighbors.

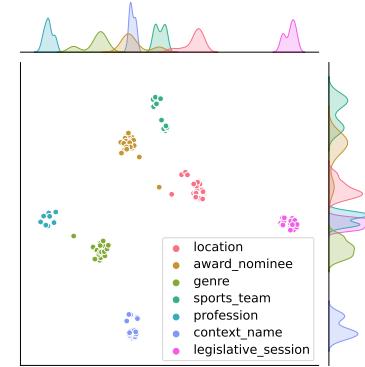


Figure 5: The t-SNE [31] visualization of unseen entity embeddings with different entity types.

5.7.2 Parameter λ . The parameter λ controls the interpolation ratio of p_{KS} to the p_{ES} according to Equation 14. We set $k = 32$ to get the best performance of CMR according to Figure 4(a) and conducted experiments with various λ . As shown in Figure 4(b), with larger λ , the performance of CMR first increases. After reaching the best performance when $\lambda = 0.95$, the performance decreases. It shows that by choosing the appropriate interpolation ratio of the semantic neighbor retrieval distribution, CMR could adjust the combination of both distributions and reach the best performance.

5.8 Visualization

We show the t-SNE visualization of unseen entities in Figure 5. We visualize the representations of unseen entities with different entity types (colors) to demonstrate different semantic information. It shows that CMR could encode the entities with similar semantics closer in representation space even without being seen during training. It indicates our generalizable representation ability.

6 CONCLUSION

In this paper, we propose to study the IMKGC task to address the emerging unseen entity problem in MKGs. We propose a semantic neighbor retrieval-enhanced IMKGC framework, CMR. Specifically, we propose unified cross-modal contrastive learning to capture multimodal semantic correlation in a unified representation space and lead the helpful semantic neighbors closer, explicit knowledge representation memorization to support semantic neighbor aggregation, and semantic neighbor retrieval-enhanced inference to augment the query-entity similarity prediction. Experiments validate the effectiveness and generalization ability of CMR on both inductive and transductive MKGC datasets. Future work directions include efficient semantic neighbor training, pretrained VLM exploration, and retrieval strategy enhancement.

ACKNOWLEDGEMENTS

This research is supported by the National Natural Science Foundation of China (No. 62272250, U22B2048, 62077031), and the Natural Science Foundation of Tianjin, China (No. 22JCJQJC00150, 22JCQNJC01580).

REFERENCES

- [1] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* 26 (2013).
- [2] Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. Ottge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in Neural Information Processing Systems* 35 (2022). 39090–39102.
- [3] Xiang Chen, Lei Li, Ningyu Zhang, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Relation Extraction as Open-book Examination: Retrieval-enhanced Prompt Tuning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2443–2448.
- [4] 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*. 904–915.
- [5] Damai Dai, Hua Zheng, Fuli Luo, Pengcheng Yang, Baobao Chang, and Zhifang Sui. 2020. Inductively representing out-of-knowledge-graph entities by optimal estimation under translational assumptions. *arXiv preprint arXiv:2009.12765* (2020).
- [6] Daniel Daza, Michael Cochez, and Paul Groth. 2021. Inductive entity representations from text via link prediction. In *Proceedings of the Web Conference 2021*. 798–808.
- [7] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *AAAI*, Vol. 32.
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.
- [10] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*. PMLR, 1263–1272.
- [11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.
- [12] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.
- [13] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations*.
- [14] Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. 2018. Canonical tensor decomposition for knowledge base completion. In *ICML*. PMLR, 2863–2872.
- [15] 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*. 2572–2580.
- [16] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 4582–4597. <https://doi.org/10.18653/v1/2021.acl-long.353>
- [17] Yunxin Li, Baotian Hu, Chen Xinyu, Yuxin Ding, Lin Ma, and Min Zhang. 2023. A Multi-Modal Context Reasoning Approach for Conditional Inference on Joint Textual and Visual Clues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 10757–10770. <https://aclanthology.org/2023.acl-long.601>
- [18] Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, and Fuchun Sun. 2022. Reasoning over different types of knowledge graphs: Static, temporal and multi-modal. *arXiv preprint arXiv:2212.05767* (2022).
- [19] Ke Liang, Sihang Zhou, Yue Liu, Lingyuan Meng, Meng Liu, and Xinwang Liu. 2023. Structure Guided Multi-modal Pre-trained Transformer for Knowledge Graph Reasoning. *arXiv preprint arXiv:2307.03591* (2023).
- [20] Sheng Liang, Mengjie Zhao, and Hinrich Schütze. 2022. Modular and Parameter-Efficient Multimodal Fusion with Prompting. In *Findings of the Association for Computational Linguistics: ACL 2022*. 2976–2985.
- [21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [22] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*. 3195–3204.
- [23] Elan Markowitz, Keshav Balasubramanian, Mehrnoosh Mirtaheri, Murali Annaram, Aram Galstyan, and Greg Ver Steeg. 2022. StATIK: Structure and text for inductive knowledge graph completion. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 604–615.
- [24] Ron Mokady, Amit Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734* (2021).
- [25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
- [26] Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. 2018. Embedding Multimodal Relational Data for Knowledge Base Completion. In *EMNLP*. 3208–3218.
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [28] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *CIKM*. 1405–1414.
- [29] Zhaoxuan Tan, Zilong Chen, Shangbin Feng, Qingyue Zhang, Qinghua Zheng, Jundong Li, and Minnan Luo. 2023. KRALC: contrastive learning with graph context modeling for sparse knowledge graph completion. In *Proceedings of the ACM Web Conference 2023*. 2548–2559.
- [30] Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *EMNLP*. 1499–1509.
- [31] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [32] Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, Ying Wang, and Yi Chang. 2021. Structure-augmented text representation learning for efficient knowledge graph completion. In *Proceedings of the Web Conference 2021*. 1737–1748.
- [33] 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*. 2735–2743.
- [34] Peifeng Wang, Jialong Han, Chenliang Li, and Rong Pan. 2019. Logic attention based neighborhood aggregation for inductive knowledge graph embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7152–7159.
- [35] Shuhu Wang, Xiaoya Li, Yuxian Meng, Tianwei Zhang, Rongbin Ouyang, Jiwei Li, and Guoyin Wang. 2022. k NN-NER: Named Entity Recognition with Nearest Neighbor Search. *arXiv preprint arXiv:2203.17103* (2022).
- [36] 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)*. IEEE, 1–8.
- [37] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *EMNLP*. 38–45.
- [38] Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. Representation learning of knowledge graphs with entity descriptions. In *AAAI*, Vol. 30.
- [39] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied knowledge representation learning. In *IJCAI*. 3140–3146.
- [40] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. *arXiv preprint arXiv:1909.03193* (2019).
- [41] Yunzhi Yao, Shengyu Mao, Ningyu Zhang, Xiang Chen, Shumin Deng, Xi Chen, and Huajun Chen. 2023. Schema-aware reference as prompt improves data-efficient knowledge graph construction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 911–921.
- [42] Ningyu Zhang, Xin Xie, Xiang Chen, Yongheng Wang, Xu Cheng, and Huajun Chen. 2022. Reasoning through memorization: Nearest neighbor knowledge graph embeddings. *arXiv preprint arXiv:2201.05575* (2022).
- [43] 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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 10527–10536.
- [44] Xiangrui Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-Modal Knowledge Graph Construction and Application: A Survey. *arXiv preprint arXiv:2202.05786* (2022).