Tokenization, Fusion, and Augmentation: Towards Fine-grained Multi-modal Entity Representation

Yichi Zhang[1,2], Zhuo Chen[1,2], Lingbing Guo[1,2], Yajing Xu[1,2], Binbin Hu[3], Ziqi Liu[3]
Wen Zhang[4,2*], Huajun Chen[1,2,5*]
[1]College of Computer Science and Technology, Zhejiang University
[2]ZJU-Ant Group Joint Lab of Knowledge Graph
[3]Ant Group
[4]School of Software Technology, Zhejiang University
[5]Zhejiang Key Laboratory of Big Data Intelligent Computing
{zhangyichi2022, zhuo.chen, zhang.wen, huajunsir}@zju.edu.cn

Abstract

Multi-modal knowledge graph completion (MMKGC) aims to discover unobserved knowledge from given knowledge graphs, collaboratively leveraging structural information from the triples and multi-modal information of the entities to overcome the inherent incompleteness. Existing MMKGC methods usually extract multi-modal features with pre-trained models, resulting in coarse handling of multi-modal entity information, overlooking the nuanced, fine-grained semantic details and their complex interactions. To tackle this shortfall, we introduce a novel framework MyGO to **tokenize, fuse, and augment the fine-grained multi-modal representations of entities** and enhance the MMKGC performance. Motivated by the tokenization technology, MyGO tokenizes multi-modal entity information as fine-grained discrete tokens and learns entity representations with a cross-modal entity encoder. To further augment the multi-modal representations, MyGO incorporates fine-grained contrastive learning to highlight the specificity of the entity representations. Experiments on standard MMKGC benchmarks reveal that our method surpasses 19 of the latest models, underlining its superior performance.

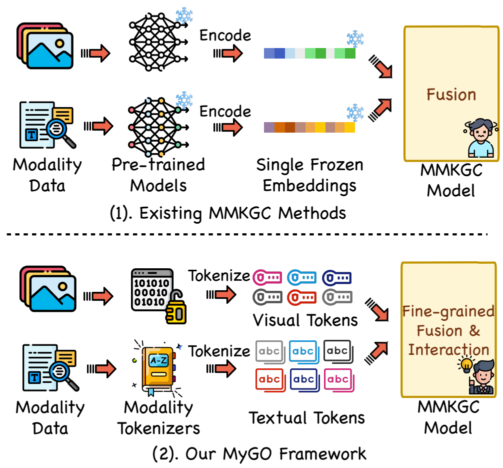

Figure 1: An intuition of existing methods and MyGO.

Introduction

Multi-modal knowledge graphs (MMKGs) (Chen et al. 2024) encapsulate diverse and complex world knowledge as structured triples (*head entity, relation, tail entity*) while incorporating multi-modal data such as images and text for additional entity context. These extensive triples, alongside their multi-modal content, form a vast multi-modal semantic network that constitutes significant infrastructures for many fields, such as recommendation (Sun et al. 2020), multi-modal understanding (Zhu et al. 2021), and large language models (Dong et al. 2024). MMKGs furnish these systems with a dependable source of factual knowledge.

MMKGs frequently grapple with the challenge of incompleteness as considerable amounts of valid knowledge remain undiscovered during their creation. This phenomenon underscores the importance of **multi-modal knowledge graph completion (MMKGC)** (Chen et al. 2024), which aims to identify new knowledge from the given MMKGs automatically. Unlike conventional knowledge graph completion (KGC) (Liang et al. 2024d) that predominantly focuses on modeling the triple structure based on the existing KGs, MMKGC needs to manage the additional multi-modal information that enriches entity description from various perspectives. Therefore, the essence of MMKGC is to harmoniously integrate structural information from triples with the rich multi-modal features associated with entities. This synergy is pivotal for informed knowledge inference within the embedding space, where the rich multi-modal information of entities serves as supplementary information and provides robust and effective multi-modal features for MMKGC.

Existing MMKGC methods (Sergieh et al. 2018) tend to represent modality information as single embedding derived from pre-trained models (Devlin et al. 2019), utilizing a fusion and prediction module to measure the triple plausibility. However, this paradigm is rather simplistic and frequently fails to capture the intricate details in the modality data. Typically, in this paradigm, the modality information extracted by pre-trained models would be frozen in later training. Moreover, when handling multiple modality instances, such as several images of an entity, these methods resort to vanilla

*Corresponding Authors.

operations like averaging, thereby stripping away potentially significant details. Considering the raw modality data houses detailed semantic units to present the crucial entity features, the common practice of generating a static embedding per modality can lead to a loss of valuable granular information, subsequently restricting MMKGC model performance. These fine-grained semantic features not only describe an entity but also embody complex cross-modal relationships. We advocate for a more fine-grained framework, allowing MMKGC models to capture the subtle, shared information embedded within the data through detailed interactions. This approach promises to significantly augment entity representations.

Aiming to solve the fine-grained information processing and leveraging problem, we propose a novel framework MYGO to achieve **fine-grained multi-modal information processing, interaction, and augmentation** in MMKGC models. Figure 1 gives a clear contrast between existing MMKGC methods and our MYGO. MYGO first employs a **modality tokenization (MT)** module to tokenize the entity modality information in MMKGs into fine-grained discrete token sequences using existing pre-trained tokenizers (Peng et al. 2022), followed by learning the MMKGC task through a **hierarchical triple modeling (HTM)** architecture. HTM consists of a cross-modal entity encoder, a contextual triple encoder, and a relational decoder to encode the fine-grained entity representation and measure the triple plausibility. To further augment and refine the entity representations, we propose a **fine-grained contrastive loss (FGCL)** to generate varied contrastive samples and boost the model performance. We conduct comprehensive experiments with public MMKG benchmarks (Liu et al. 2019a). Comparisons against 19 recent baselines demonstrate the outperforming results of MYGO. We also delve further into the nuances of MYGO's design to understand it. Our contribution is three-fold:

(1). We emphasize fine-grained multi-modal learning for MMKGC and propose a cutting-edge framework MYGO. MYGO tokenizes the modality data into fine-grained multi-modal tokens and pioneers a novel MMKGC architecture to hierarchically model the cross-modal entity representation.

(2). We propose a fine-grained contrastive learning module to augment the cross-modal entity representations. This module innovates by employing new tactics to generate high-quality comparative samples for more detailed and effective self-supervised contrastive learning.

(3). We conduct comprehensive experiments on public benchmarks and achieve new state-of-the-art performance against 19 baselines with further exploration.

Related Works

Multi-modal knowledge graphs (MMKGs) (Chen et al. 2024; Liang et al. 2024b) are knowledge graphs with rich multi-modal information like images, text descriptions, audio, and videos (Wang et al. 2023). Due to the incompleteness of the knowledge graphs, knowledge graph completion (KGC) (Liang et al. 2024a) is a popular research topic to automatically discover unobserved knowledge triples by learning from the triple structure. **Multi-modal knowledge graph completion (MMKGC)** (Liang et al. 2024c) aims to predict missing triples in the given MMKGs collaboratively leveraging the extra multi-modal information from entities. Existing MMKGC methods mainly make new improvements in three perspectives: (1) multi-modal fusion and interaction (Cao et al. 2022), (2) integrated decision (Li et al. 2023), and (3) negative sampling (Xu et al. 2022).

Task Definition

A multi-modal knowledge graph (MMKG) incorporating both visual and textual modalities can be represented as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{V}, \mathcal{D})$, where $\mathcal{E}, \mathcal{R}$ are the entity set and the relation set. $\mathcal{T} = \{(h, r, t) \mid h, r \in \mathcal{E}, r \in \mathcal{R}\}$ is the triple set, indicating that entity h is related to entity through t relation r. Besides, $\mathcal{V}, \mathcal{D}$ correspond to the collections of images and textual descriptions for each entity e.

The primary aim of **knowledge graph completion (KGC)** is to learn a score function $\mathcal{S}(h, r, t)$ which measures the plausibility of a triple (h, r, t) by **a scalar score**. In KGC models, entities and relations correspond to embeddings, and the triple score is defined on these embeddings, preferring high scores for positive triples and lower scores for negative triples. In other words, the plausibility of positive triples in the training set is maximized by a positive-negative contrast (Bordes et al. 2013) during training. Expanding to MMKGs, **multi-modal knowledge graph completion (MMKGC)** would further consider the multi-modal information $\mathcal{V}(e), \mathcal{D}(e)$ to enhance their embeddings.

Methodology

In this section, we will detailedly introduce the framework proposed by us which leverages **M**odalit**Y** information as fine-**G**rained t**O**kens (MYGO for short) to **tokenize, fuse, and augment the fine-grained multi-modal representations of entities**. We consider the mainstream MMKGC setting (Xu et al. 2022) that includes both image and text modalities (Chen et al. 2024). MYGO mainly comprises three modules: modality tokenization module, hierarchical triple modeling module, and fine-grained contrastive learning, aiming to process, fuse, and augment the fine-grained information in MMKGs respectively. Figure 2 provides an intuitive perspective of the design of MYGO.

Modality Tokenization

To capture fine-grained multi-modal information, we propose a **modality tokenization (MT)** module to process the raw multi-modal data of entities into fine-grained discrete semantic tokens, serving as semantic units to learn fine-grained entity representations. We employ the tokenizers for image and textual modality respectively, denoted as $\mathcal{Q}_{img}, \mathcal{Q}_{txt}$, to generate visual tokens $v_{e,i}$ and textual tokens $w_{e,i}$ for entity e:

$$\mathcal{U}_{img}(e) = \{v_{e,1}, v_{e,2}, \cdots, v_{e,m_e}\} = \mathcal{Q}_{img}(\mathcal{V}(e)) \quad (1)$$

$$\mathcal{U}_{txt}(e) = \{w_{e,1}, w_{e,2}, \cdots, w_{e,n_e}\} = \mathcal{Q}_{txt}(\mathcal{D}(e)) \quad (2)$$

where m_e, n_e is the number of tokens for each modality and we denote $\mathcal{U}(e)$ symbolizes the collective token set for entity e. The text tokens are from the vocabulary of a language model (Devlin et al. 2019) while the visual tokens are from

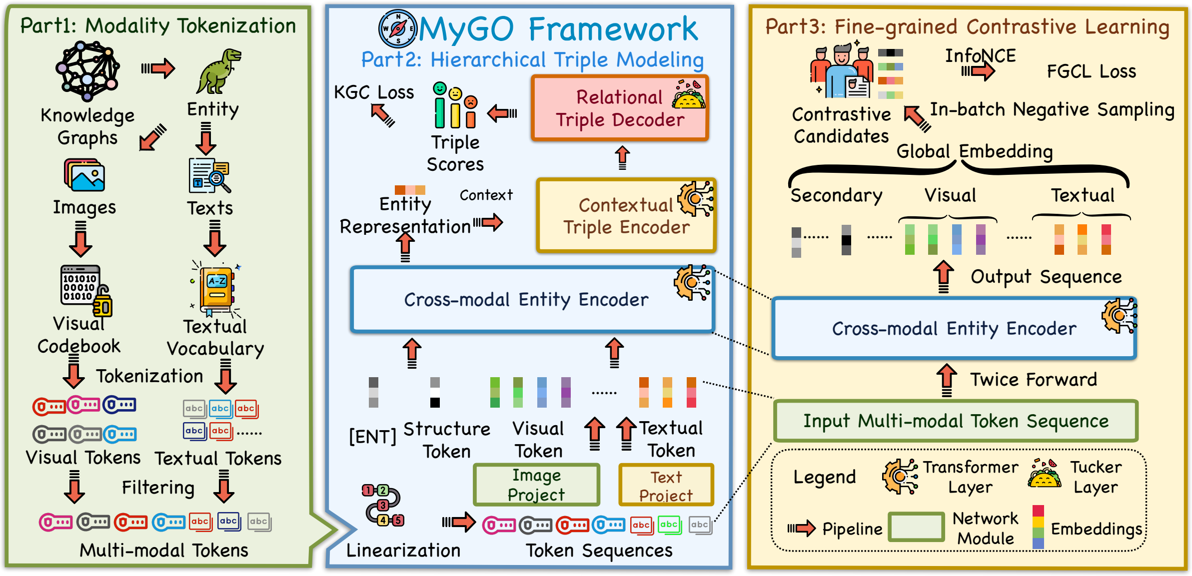

Figure 2: The overview of our MYGO framework. We mainly have three parts of new designs in MYGO to tokenize, fuse, and augment the fine-grained multi-modal semantic information in the MMKGs.

the codebook of a pre-trained visual tokenizer (Esser, Rombach, and Ommer 2021; Peng et al. 2022). Notably, $\mathcal{V}(e)$ might consist of multiple images and we process each image to accumulate the tokens in $\mathcal{U}_{img}(e)$.

During the tokenization process, it's common to encounter duplicate tokens. Therefore, we count the occurrence frequency of each token, retaining a predetermined quantity of the most common tokens for each modality. Additionally, we remove the stop words (Wilbur and Sirotkin 1992) in the textual descriptions as their contribution to the entity semantics is minimal. After the MT and refinement process, we can obtain processed token set $\mathcal{U}'_{img}$ and $\mathcal{U}'_{txt}$ with m visual tokens and n textual tokens for each entity e, featuring a collection of fine-grained tokens that embody the vital features derived from the raw multi-modal data. Subsequently, we assign a separate embedding for each token in $\mathcal{U}'_{img}$ and $\mathcal{U}'_{txt}$. This approach is tailored to the fact that different entities may share tokens, and individualized embeddings of tokens allow for a more fine-grained representation of similar features across various entities, enriching the entities' profiles with detailed multi-modal semantic units.

Hierarchical Triple Modeling

After the MT process, we further design a **hierarchical triple modeling (HTM)** module in this section. HTM leverages a hierarchical transformer architecture to capture the multi-modal entity representation and model the triple plausibility in a hierarchical manner, which consists of three components: cross-modal entity encoder, contextual triple encoder, and relational decoder.

Cross-modal Entity Encoder The cross-modal entity encoder (CMEE) aims to capture the multi-modal representation of entities by leveraging their fine-grained multi-modal tokens. Unlike existing methods (Li et al. 2023), MYGO performs fine-grained tokenization and obtains a sequence of discrete tokens. Therefore, we design a more **fine-grained feature interaction method** that allows full interaction between all the different modal messages. In MYGO, we apply a transformer (Vaswani et al. 2017) layer as the CMEE. We first linearize the multi-modal tokens as a sequence:

$$\mathcal{X}(e) = ([\mathtt{ENT}], s_e, v_{e,1}, \cdots, v_{e,m}, w_{e,1}, \cdots, w_{e,n}) \quad (3)$$

where $[\mathtt{ENT}]$ is a special token and s_e is a learnable embedding representing the structural information of the entity. $[\mathtt{ENT}]$ is analogous the $[\mathtt{CLS}]$ token in BERT (Devlin et al. 2019) to capture the sequence feature for downstream prediction. s_e is a learnable embedding to represent the structural information learned from the existing triple structures, which will be optimized during training. Besides, for the multi-modal tokens from $\mathcal{U}'_{img}$ and $\mathcal{U}'_{txt}$, we freeze their initial representations derived from the tokenizers and define linear projection layers $\mathcal{P}_{img}, \mathcal{P}_{txt}$ to project them into the same representation space as $\widehat{v}_{e,i} = \mathcal{P}_{img}(v_{e,i}) + b_{img}$ $\widehat{w}_{e,j} = \mathcal{P}_{txt}(v_{e,j}) + b_{txt}$ where b_{img}, b_{txt} are defined modality biases to enhance the labeling of information from distinct modalities. In this way, the final sequence entered into CMEE becomes $\mathcal{X}_{input}(e) = ([\mathtt{ENT}], s_e, \widehat{v}_{e,1}, \cdots, \widehat{v}_{e,m}, \widehat{w}_{e,1}, \cdots, \widehat{w}_{e,n})$. The cross-modal entity representation is obtained by $\mathbf{e} = \mathbf{Pooling}(\mathbf{Transformer}(\mathcal{X}_{input}(e)))$, where $\mathbf{Transformer}()$ represents a transformer encoder layer (Vaswani et al. 2017), $\mathbf{Pooling}$ is the pooling operation with obtains the final hidden representation of the special token $[\mathtt{ENT}]$. It allows each token in the input sequence can be dynamically highlighted by CMEE to interact and eventually learn expressive entity representations.

Contextual Triple Encoder To achieve adequate modality interaction in the relational context, we apply another transformer layer as the **contextual triple encoder (CTE)** to encode the contextual embeddings for the given query. Taking head query $(h, r, ?)$ (tail prediction) as an example, we can obtain the contextual embeddings $\tilde{\mathbf{h}}$ as: $\tilde{\mathbf{h}} = \mathbf{Transformer}([\texttt{CXT}], \mathbf{h}, \mathbf{r})$, where $[\texttt{CXT}]$ is a special token in the input sequence to capture the contextual embedding of entity, $\mathbf{h}$ is the output representation of h from CMEE, and $\mathbf{r}$ is the relation embedding for each $r \in \mathcal{R}$. The contextual embeddings of the query $(h, r, ?)$ are then processed by a relational decoder for entity prediction.

Relational Decoder Moreover, we employ a score function $\mathcal{S}(h, r, t)$ to measure the triple plausibility by producing a scalar score, which functions as the relational decoder for query prediction. In MyGO, we employ Tucker (Ivana, Carl, and Timothy 2019) as our score function denoted as $\mathcal{S}(h, r, t) = \mathcal{W} \times_1 \tilde{\mathbf{h}} \times_2 \tilde{\mathbf{r}} \times_3 \mathbf{t}$, where $\times_i$ represents the tensor product along the i-th mode, $\mathcal{W}$ is the core tensor learned during training. We train our model with cross-entropy loss for each triple. We treat t as the golden label against the whole entity set $\mathcal{E}$, which is the same for head prediction. Therefore, the objective is a cross-entropy loss:

$$\mathcal{L}_{head} = -\sum_{(h,r,t)\in\mathcal{T}} \log \frac{\exp(\mathcal{S}(h, r, t))}{\sum_{t'\in\mathcal{E}} \exp(\mathcal{S}(h, r, t'))} \quad (4)$$

Note that we use the contextual embedding $\tilde{\mathbf{e}}_h$ of h and the multi-modal embedding $\mathbf{e}_t$ of t to calculate the score, which can expedite computation. Otherwise, we would need to extract the contextual embedding of all the candidate entities under different relations, which needs $O(|\mathcal{E}| \times |\mathcal{R}|)$-level forward passes in contextual transformer and would greatly increase the computation of the model. Besides, both head prediction and tail prediction are considered in MyGO, and the objective $\mathcal{L}_{tail}$ is similar when giving a tail query $(?, r, t)$. The overall MMKGC task objective can be denoted as $\mathcal{L}_{kgc} = \mathcal{L}_{head} + \mathcal{L}_{tail}$.

Fine-grained Contrastive Learning

Based on the above design, we have been able to train and test the MMKGC model. To further augment fine-grained and robust multi-modal entity representations, we introduce a **fine-grained contrastive learning (FGCL)** module in MyGO to achieve this goal by multi-scale contrastive learning on the entity representations.

As mentioned before, CMEE aims to capture the entity representation based on a multi-modal token sequence. Inspired by the idea of SimCSE (Gao, Yao, and Chen 2021), we augment these entity representations through contrastive learning. Specifically, given an entity e, we can get two representations $\mathbf{e}, \mathbf{e}_{sec}$ from CMEE by two forward passes. The variations between these two embeddings, induced by the dropout layer in the transformer encoder, allow for slight deactivation of multi-modal token features, effectively acting as a form of simple data augmentation. By in-batch contrastive learning across a collection of entities, MyGO is trained to extract truly significant information from token sequences, thereby enhancing the distinctiveness of each entity's representation. To deepen the granularity of this process, we further extract three additional representations from the transformer output, which can represent entity features from their perspectives. We can define the output representations of the multi-modal tokens in an input sequence $\mathcal{X}_{input}(e)$ as: $\mathcal{X}_{output}(e) = ([\texttt{ENT}]', s'_e, \widehat{v}'_{e,1}, \cdots, \widehat{v}'_{e,m}, \widehat{w}'_{e,1}, \cdots, \widehat{w}'_{e,n})$ Then we introduce three embeddings $\mathbf{s}(e), \mathbf{v}(e), \mathbf{w}(e)$ to represent the global, visual, and textual information of an entity e. $\mathbf{s}(e)$ is derived from the average of all the output representations in $\mathcal{X}_{output}(e)$. Similarly, $\mathbf{v}(e)$ and $\mathbf{w}(e)$ are the averages of the corresponding visual and textual tokens. They can be denoted as:

$$\begin{gathered}\mathbf{s}(e) = \mathbf{Mean}(\mathcal{X}_{output}(e)) \\ \mathbf{v}(e) = \frac{1}{m}\sum_{i=1}^{m} \widehat{v}'_{e,i} \quad \mathbf{w}(e) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}'_{e,i}\end{gathered} \quad (5)$$

Among these embeddings, $\mathbf{e}_{sec}, \mathbf{s}(e)$ encapsulate the global information of e and $\mathbf{v}(e), \mathbf{w}(e)$ consist of the local modality information of the entity e. For each entity e, we can collect its candidates for contrastive learning as $\mathcal{C}(e) = \{e_{sec}, \mathbf{s}(e), \mathbf{v}(e), \mathbf{w}(e)\}$, which consists of its global and local features. $(\mathbf{e}, \mathbf{e}')$ where $\mathbf{e}' \in \mathcal{C}(e)$ is regarded as a positve sample. Then we employ in-batch negative sampling to construct negative pairs and the final objective can be denoted as:

$$\mathcal{L}_{con} = -\sum_{i=1}^{\mathcal{B}} \sum_{e'_i\in\mathcal{C}(e_i)} \log \frac{\exp(\mathbf{cos}(\mathbf{e}_i, \mathbf{e}'_i)/\tau)}{\sum_{j=1}^{\mathcal{B}} \exp(\mathbf{cos}(\mathbf{e}_i, \mathbf{e}'_j)/\tau)} \quad (6)$$

where $\mathcal{B}$ is the batch size, $\mathbf{cos}(\cdot, \cdot)$ is the cosine similarity of two embeddings and τ is the temperature hyper-parameter. Through such an FGCL process, MyGO notably improves its ability to discern detailed multi-modal attributes across various entities, boosting the model performance in the MMKGC task. Finally, the overall training objective of our framework can be denoted as $\mathcal{L} = \mathcal{L}_{kgc} + \lambda\mathcal{L}_{con}$ where λ is a hyper-parameter to control the weight of $\mathcal{L}_{con}$.

Experiments

In this section, we will conduct comprehensive experiments to evaluate the performance of MyGO by presenting its **effectiveness**, **reasonablity**, **efficiency** and **explainability**.

Experiment Settings

Datasets. In this paper, we employ three public MMKGC benchmarks DB15K (Liu et al. 2019a), MKG-W and MKG-Y (Xu et al. 2022) to evaluate the model performance. The raw data for each modality are obtained from their official release sources.

Evaluation Protocol. We conduct link prediction (Bordes et al. 2013) task on the datasets, which is the mainstream MMKGC task. Following existing works, we use rank-based metrics (Sun et al. 2019)like mean reciprocal rank (MRR) and Hit@K (K=1, 3, 10) (H@K for short) to evaluate the results. Besides, we employ the filter setting (Bordes et al.

Model	Fusion Strategy	DB15K				MKG-W				MKG-Y			
		MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
TransE	None	24.86	12.78	31.48	47.07	29.19	21.06	33.20	44.23	30.73	23.45	35.18	43.37
DistMult	None	23.03	14.78	26.28	39.59	20.99	15.93	22.28	30.86	25.04	19.33	27.80	35.95
ComplEx	None	27.48	18.37	31.57	45.37	24.93	19.09	26.69	36.73	28.71	22.26	32.12	40.93
RotatE	None	29.28	17.87	36.12	49.66	33.67	26.80	36.68	46.73	34.95	29.10	38.35	45.30
TuckER	None	<u>33.86</u>	<u>25.33</u>	37.91	50.38	30.39	24.44	32.91	41.25	37.05	<u>34.59</u>	38.43	41.45
IKRL	Static	26.82	14.09	34.93	49.09	32.36	26.11	34.75	44.07	33.22	30.37	34.28	38.26
TBKGC	Static	28.40	15.61	37.03	49.86	31.48	25.31	33.98	43.24	33.99	30.47	35.27	40.07
TransAE	Static	28.09	21.25	31.17	41.17	30.00	21.23	34.91	44.72	28.10	25.31	29.10	33.03
MMKRL	Static	26.81	13.85	35.07	49.39	30.10	22.16	34.09	44.69	36.81	31.66	39.79	45.31
RSME	Adaptive	29.76	24.15	32.12	40.29	29.23	23.36	31.97	40.43	34.44	31.78	36.07	39.09
VBKGC	Static	30.61	19.75	37.18	49.44	30.61	24.91	33.01	40.88	37.04	33.76	38.75	42.30
OTKGE	Adaptive	23.86	18.45	25.89	34.23	34.36	<u>28.85</u>	36.25	44.88	35.51	31.97	37.18	41.38
MACO	Static	27.41	14.61	35.59	50.00	31.74	25.23	34.23	44.37	34.98	31.59	36.68	40.51
IMF	Adaptive	32.25	24.20	36.00	48.19	34.50	28.77	36.62	45.44	35.79	32.95	37.14	40.63
QEB	Static	28.18	14.82	36.67	51.55	32.38	25.47	35.06	45.32	34.37	29.49	36.95	42.32
VISTA	Adaptive	30.42	22.49	33.56	45.94	32.91	26.12	35.38	45.61	30.45	24.87	32.39	41.53
AdaMF	Adaptive	32.51	21.31	<u>39.67</u>	<u>51.68</u>	34.27	27.21	<u>37.86</u>	47.21	<u>38.06</u>	33.49	**40.44**	**45.48**
MANS	Static	28.82	16.87	36.58	49.26	30.88	24.89	33.63	41.78	29.03	25.25	31.35	34.49
MMRNS	Adaptive	32.68	23.01	37.86	51.01	<u>35.03</u>	28.59	37.49	<u>47.47</u>	35.93	30.53	39.07	<u>45.47</u>
MyGO	Adaptive	**37.72**	**30.08**	**41.26**	**52.21**	**36.10**	**29.78**	**38.54**	**47.75**	**38.44**	**35.01**	<u>39.84</u>	44.19
		+11.4%	+18.4%	+4.0%	+1.0%	+3.1%	+3.2%	+1.8%	+0.6%	+0.9%	+1.2%	-	-

Table 1: The main MMKGC results. We list the type of fusion strategy (none / static / adaptive) considered by each method in the table. The best results are marked as **bold** and the second best results are <u>underlined</u>.

2013) in the prediction results to remove the candidate triples existing in the training data for fair comparisons.

Baselines. To make a comprehensive performance evaluation, we employ 19 different state-of-the-art MMKGC baselines (Bordes et al. 2013; Yang et al. 2015; Sun et al. 2019; Trouillon et al. 2016; Xie et al. 2017; Sergieh et al. 2018; Cao et al. 2022; Wang et al. 2019; Lu et al. 2022; Wang et al. 2021; Zhang and Zhang 2022; Zhang, Chen, and Zhang 2023b; Li et al. 2023; Wang et al. 2023; Lee et al. 2023; Zhang et al. 2024; Xu et al. 2022; Zhang, Chen, and Zhang 2023a) in our experiments.

Implementation Details. In our experiments, we implement MyGO with PyTorch (Paszke et al. 2019). For modality tokenization, we employ the tokenizer of BEIT (Peng et al. 2022) and BERT (Devlin et al. 2019) as our visual/textual tokenizers. The codebook size of BEIT (Peng et al. 2022) is 8192 and the vocabulary size of BERT tokenizer is 32000. During training, we set the training epoch to 2000, the batch size to 1024, and the embedding dimension to 256. The max token number m and n are tuned in $\{4, 8, 12\}$ and the weight λ is tuned in $\{1, 0.1, 0.01, 0.001\}$. We optimize the model with Adam (Kingma and Ba 2015) optimizer.

Main Results

The primary experiment results of MMKGC are depicted in Table 1. We list the strategies in which all methods utilize the modal information in addition to the statistical performance metrics. Unimodal approaches do not incorporate multi-modal information of entities, whereas the multi-modal and negative sampling-based KGC methods are categorized as either static or adaptive based on their multi-modal fusion methodology. MyGO employs self-attention in the encoders so that it is an adaptive method as well. **Outperforming results.** Firstly, we can observe that MyGO outperforms all baseline methods on all evaluation metrics, achieving new state-of-the-art performance on two datasets. The adaptive MMKGC methods, as indicated by the results, generally outperform static ones. Meanwhile, different from other adaptive approaches, MyGO employs more fine-grained feature processing and fusion with modality tokenization and hierarchical triple modeling. As existing methods tend to set only one feature (embedding) for each modality, MyGO obtains more fine-grained features by tokenization of existing raw data and improves the model

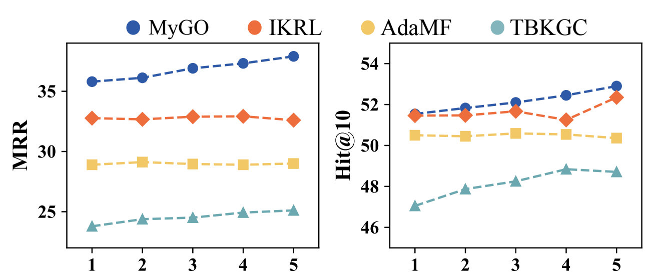

Figure 3: The MMKGC results of MyGO and several baselines with different numbers of entity images.

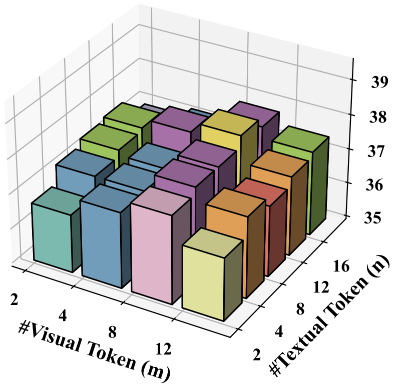

Figure 4: The MRR results with different modality token amount m and n on DB15K.

performance through fine-grained interactive fusion with transformer-based encoders in HTM.

Accurate reasoning ability. Further, comparing the improvement of each metric horizontally, we can see that MyGO's improvement for Hit@1 and MRR is significantly higher than that of Hit@10 and other metrics. For example, on DB15K, MyGO achieves an 18.4% increase on Hit@1 but a 1% increase on Hit@10. This underscores MyGO's capability to significantly improve accurate reasoning through its sophisticated design.

Exploration on Modality Tokenization

Compared to current methods, our innovative new design in MyGO introduces the Modality Tokenization (MT). Therefore, we aim to delve deeply into the principles of MT in this section. We mainly focus on two considerable problems: the model's capability to handle multiple pieces of information within a single modality, and the impact of token amounts.

Multiple Modality Information Existing methods usually obtain the multi-modal embeddings by consolidating the multiple information within a single modality like mean pooling across various. This operation results in a loss of crucial features and frequently happens during the preprocessing phase. Conversely, MT transforms multiple data particles into a token sequence, preserving common features as much as possible. To demonstrate the effectiveness of MT in this scenario, we conduct another experiment using varying amounts of entity images. As an entity's textual description usually only comprises one paragraph, dividing it is a challenge. Therefore, we evaluate the MMKGC performance of different models on different numbers of entity images, keeping $N = 1, 2, 3, 4, 5$ images for each entity as far as possible. The results depicted in Figure 3 highlight that in comparison to other baselines, MyGO can achieve consistent and impressive performance enhancements, even when faced with increasing multi-modal data, as the MMKGC performance shows a clear trend of increasing. Contrarily, the baseline performance is somewhat erratic, displaying fluctuations as the image amount increases, and the overall effectiveness does not match MyGO. We attribute this phenomenon to the fact that different methods handle modal information differently. Current models typically generate an embedding for a specific modality from an entity's multiple raw data, thus losing essential original information from the initial features. However, MyGO processes the information into fine-grained semantic units through the design of modality tokenization, retaining the most recurring components. Through this technique, **MyGO masters the ability to retain the general and uniform information** in a modality even when the modality-specific data volume expands, making our method more stable and scalable.

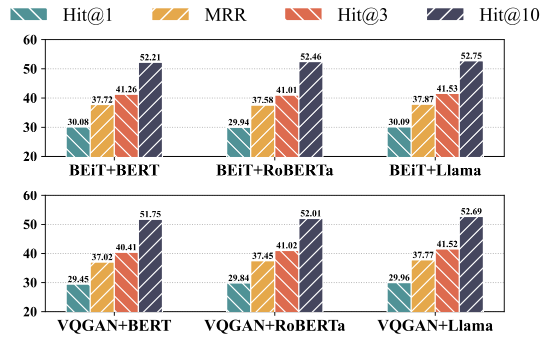

Figure 5: MMKGC results using different tokenizers.

Impact of Token Amount Another intriguing aspect to investigate pertains to the two token amount hyperparameters m and n that we set in MT. These parameters dictate the number of high-frequency multi-modal tokens retained and processed by CMEE. More tokens correlate with more intricate interactions in the model with an $O((m+n)^2)$-level increase in time efficiency. This is because the time complexity of the Transformer layer used in CMEE is positively related to the quadratic of the sequence length. Therefore, we explored the performance of the MMKGC concerning the variation of the parameters m and n. The experimental results are depicted in Figure 4 as a 3D bar chart. From the figure we can observe that the model performance shows an increasing and then decreasing trend with the increase in the number of tokens m and n. This pattern is more distinct with increasing visual tokens, while the text modality experiences minor variations.

Impact of Different Tokenizers To further explore the robustness of MyGO over different tokenizers, we conduct the experiments on DB15K with more kinds of visual tokenizers (BEiT (Peng et al. 2022) and VQGAN (Esser, Rombach, and Ommer 2021)) and textual tokenizers (BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019b), and Llama (Touvron et al. 2023)). The MMKGC results with their combinations are presented in Figure 5. We can conclude that MyGO is stable and robust across different tokenizers, This indicates that MyGO is a generalizable and universal framework, which can integrate different multi-modal backbones. Besides, we can find that larger backbones like Llama (Touvron et al. 2023) could bring better performance, demonstrating the possibilities of combining our approach with the latest LLM technology.

Setting		MRR	H@10	H@3	H@1
Model Design	w/o MT	35.48	50.89	39.09	27.48
	w/o Refine	36.61	50.97	39.89	29.13
	w/o CMEE	34.78	50.44	38.32	26.66
	w/o CTE	34.71	50.72	38.37	26.59
FGCL	w/o $\mathcal{L}_{con}$	35.99	51.31	39.68	27.98
	w/o e_{sec}	36.82	51.75	40.55	29.02
	w/o $\mathbf{s}(e)$	37.62	52.46	40.91	29.97
	w/o $\mathbf{v}(e)$	37.24	51.18	40.70	29.58
	w/o $\mathbf{w}(e)$	37.64	52.16	41.24	29.96
	$\lambda = 1$	37.43	52.03	40.75	29.83
	$\lambda = 0.1$	37.48	52.16	41.22	29.72
	$\lambda = 0.001$	36.91	52.10	39.89	27.99
Full Model	$\lambda = 0.01$	**37.72**	**52.21**	**41.26**	**30.08**

Table 2: The ablation study results on DB15K.

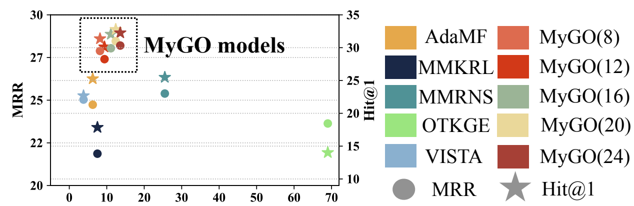

Figure 6: The efficiency-performance trade-off analysis.

Ablation Study

To confirm the effectiveness of each module in MYGO, we further conduct an ablation study from two perspectives:the model design, and the FGCL loss. We removed the corresponding modules across different settings and performed MMKGC experiments. The experimental results are presented in Table 2. According to the first part of the experimental results, all of the core modules we designed in the backbone network, the modality tokenization process, and the filtering process critically influence the final prediction. Besides, the design of FGCL also contributes to the model performance, with a contrastive candidate in $\mathcal{C}(e)$ being essential for achieving the SOTA performance. Meanwhile, we explore the influence of the loss weight λ of FGCL. As we tuned λ in $\{1, 0.1, 0.01, 0.001\}$, the MMKGC results show an increasing and then decreasing trend and reaches state-of-the-art at $\lambda = 0.01$. Overall, we can find that the most pivotal modules affecting the overall performance are CMEE and CTE, which extract fine-grained and contextual entity representations to make a modality-aware triple prediction. The FCGL module further makes more of further boosts model performance based on the backbone.

Efficiency Analysis

To validate the efficiency of MYGO, we conduct an efficiency experiment. As shown in Figure 6, we compare the training efficiency and final prediction of MYGO with different m and n against several recent baselines. We find that MYGO achieves the best performance while maintaining relatively good efficiency, realizing a trade-off between performance and efficiency. Besides, we can observe that the total amount of tokens $(m + n)$ has a small effect on efficiency. As $m+n$ increases from 8 to 24, MYGO's efficiency goes from about 8s to about 13s, which is still acceptable in practice.

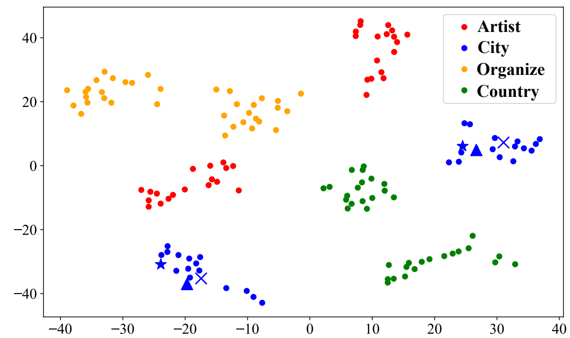

Figure 7: The embedding visualization results of the multi-modal tokens from entities with different categories (Artist, City, Organization, and Country).

Token Embedding Visualization

To give an intuitive view of the learned representations, we conduct an embedding visualization demonstration experiment in this section. We choose four diverse categories of entities (artist, city, organize, country) in DB15K. For each category, we select two entities and perform dimensionality reduction of their multi-modal tokens from the entire input sequence $\mathcal{X}(e)$ employing the t-SNE (Van der Maaten and Hinton 2008) algorithm. For the city category, we uniquely mark the same tokens in both city entities using special marks $(\triangle, \star, \times)$ to demonstrate the contextual multi-modal embeddings of the same tokens under different token sequences. Figure 7 reveals that each small cluster in the figure represents the tokens of an entity. Furthermore, tokens in entities of the same category can easily form clusters, displaying a certain degree of distinction between tokens from diverse entities. This indicates that the learned token embeddings are distinguishable. Also, identical tokens under differing contexts revealed slight variances, which underscores their potential to provide unique roles for different entities, even if originated from the same token feature. Altogether, this visualization validates the effectiveness of our approach by revealing the distribution of multi-modal tokens.

Conclusion

In this paper, we focus on the problem of capturing fine-grained semantic information in MMKGs. We propose a new framework MYGO to tokenize, fuse, and augment multi-modal entity representations. Experiments on public benchmarks demonstrate the effectiveness, reliability, reasonableness, and interpretability of our design. In the future, we will focus on processing and interpreting the fine-grained multi-modal information in MMKGs.

Acknowledgements

This work is founded by the National Natural Science Foundation of China (NSFCU23B2055 / NSFCU19B2027 / NSFC62306276), Zhejiang Provincial Natural Science Foundation of China (No. LQ23F020017), Yongjiang Talent Introduction Programme (2022A-238-G), and Fundamental Research Funds for the Central Universities (226-2023-00138). This work was supported by AntGroup.

References

Bordes, A.; Usunier, N.; García-Durán, A.; Weston, J.; and Yakhnenko, O. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NIPS*, 2787–2795.

Cao, Z.; Xu, Q.; Yang, Z.; He, Y.; Cao, X.; and Huang, Q. 2022. OTKGE: Multi-modal Knowledge Graph Embeddings via Optimal Transport. In *NeurIPS*.

Chen, Z.; Zhang, Y.; Fang, Y.; Geng, Y.; Guo, L.; Chen, X.; Li, Q.; Zhang, W.; Chen, J.; Zhu, Y.; Li, J.; Liu, X.; Pan, J. Z.; Zhang, N.; and Chen, H. 2024. Knowledge Graphs Meet Multi-Modal Learning: A Comprehensive Survey. *CoRR*, abs/2402.05391.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, 4171–4186. Association for Computational Linguistics.

Dong, J.; Zhang, Q.; Zhou, H.; Zha, D.; Zheng, P.; and Huang, X. 2024. Modality-Aware Integration with Large Language Models for Knowledge-based Visual Question Answering. *CoRR*, abs/2402.12728.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*, 12873–12883. Computer Vision Foundation / IEEE.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP (1)*, 6894–6910. Association for Computational Linguistics.

Ivana, B.; Carl, A.; and Timothy, M. H. 2019. TuckER: Tensor Factorization for Knowledge Graph Completion. In *EMNLP/IJCNLP (1)*, 5184–5193. Association for Computational Linguistics.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

Lee, J.; Chung, C.; Lee, H.; Jo, S.; and Whang, J. J. 2023. VISTA: Visual-Textual Knowledge Graph Representation Learning. In *EMNLP (Findings)*, 7314–7328. Association for Computational Linguistics.

Li, X.; Zhao, X.; Xu, J.; Zhang, Y.; and Xing, C. 2023. IMF: Interactive Multimodal Fusion Model for Link Prediction. In *WWW*, 2572–2580. ACM.

Liang, K.; Liu, Y.; Zhou, S.; Tu, W.; Wen, Y.; Yang, X.; Dong, X.; and Liu, X. 2024a. Knowledge Graph Contrastive Learning Based on Relation-Symmetrical Structure. *IEEE Trans. Knowl. Data Eng.*, 36(1): 226–238.

Liang, K.; Meng, L.; Liu, M.; Liu, Y.; Tu, W.; Wang, S.; Zhou, S.; Liu, X.; Sun, F.; and He, K. 2024b. A Survey of Knowledge Graph Reasoning on Graph Types: Static, Dynamic, and Multi-Modal. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12): 9456–9478.

Liang, K.; Meng, L.; Liu, Y.; Liu, M.; Wei, W.; Liu, S.; Tu, W.; Wang, S.; Zhou, S.; and Liu, X. 2024c. Simple Yet Effective: Structure Guided Pre-trained Transformer for Multi-modal Knowledge Graph Reasoning. In *ACM Multimedia*, 1554–1563. ACM.

Liang, K.; Meng, L.; Zhou, S.; Tu, W.; Wang, S.; Liu, Y.; Liu, M.; Zhao, L.; Dong, X.; and Liu, X. 2024d. MINES: Message Intercommunication for Inductive Relation Reasoning over Neighbor-Enhanced Subgraphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10645–10653.

Liu, Y.; Li, H.; García-Durán, A.; Niepert, M.; Oñoro-Rubio, D.; and Rosenblum, D. S. 2019a. MMKG: Multi-modal Knowledge Graphs. In *ESWC*, volume 11503 of *Lecture Notes in Computer Science*, 459–474. Springer.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Lu, X.; Wang, L.; Jiang, Z.; He, S.; and Liu, S. 2022. MMKRL: A robust embedding approach for multi-modal knowledge graph representation learning. *Appl. Intell.*, 52(7): 7480–7497.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E. Z.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, 8024–8035.

Peng, Z.; Dong, L.; Bao, H.; Ye, Q.; and Wei, F. 2022. BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers. *CoRR*, abs/2208.06366.

Sergieh, H. M.; Botschen, T.; Gurevych, I.; and Roth, S. 2018. A Multimodal Translation-Based Approach for Knowledge Graph Representation Learning. In **SEM@NAACL-HLT*, 225–234. Association for Computational Linguistics.

Sun, R.; Cao, X.; Zhao, Y.; Wan, J.; Zhou, K.; Zhang, F.; Wang, Z.; and Zheng, K. 2020. Multi-modal Knowledge Graphs for Recommender Systems. In *CIKM*, 1405–1414. ACM.

Sun, Z.; Deng, Z.-H.; Nie, J.-Y.; and Tang, J. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *ICLR (Poster)*. OpenReview.net.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; and Bouchard, G. 2016. Complex Embeddings for Simple Link Prediction. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, 2071–2080. JMLR.org.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*, 5998–6008.

Wang, M.; Wang, S.; Yang, H.; Zhang, Z.; Chen, X.; and Qi, G. 2021. Is Visual Context Really Helpful for Knowledge Graph? A Representation Learning Perspective. In *ACM Multimedia*, 2735–2743. ACM.

Wang, X.; Meng, B.; Chen, H.; Meng, Y.; Lv, K.; and Zhu, W. 2023. TIVA-KG: A Multimodal Knowledge Graph with Text, Image, Video and Audio. In *ACM Multimedia*, 2391–2399. ACM.

Wang, Z.; Li, L.; Li, Q.; and Zeng, D. 2019. Multimodal Data Enhanced Representation Learning for Knowledge Graphs. In *IJCNN*, 1–8. IEEE.

Wilbur, W. J.; and Sirotkin, K. 1992. The automatic identification of stop words. *J. Inf. Sci.*, 18(1): 45–55.

Xie, R.; Liu, Z.; Luan, H.; and Sun, M. 2017. Image-embodied Knowledge Representation Learning. In *IJCAI*, 3140–3146. ijcai.org.

Xu, D.; Xu, T.; Wu, S.; Zhou, J.; and Chen, E. 2022. Relation-enhanced Negative Sampling for Multimodal Knowledge Graph Completion. In *ACM Multimedia*, 3857–3866. ACM.

Yang, B.; Yih, W.-t.; He, X.; Gao, J.; and Deng, L. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *ICLR (Poster)*.

Zhang, Y.; Chen, M.; and Zhang, W. 2023a. Modality-Aware Negative Sampling for Multi-modal Knowledge Graph Embedding. In *IJCNN*, 1–8. IEEE.

Zhang, Y.; Chen, Z.; Liang, L.; Chen, H.; and Zhang, W. 2024. Unleashing the Power of Imbalanced Modality Information for Multi-modal Knowledge Graph Completion. *CoRR*, abs/2402.15444.

Zhang, Y.; Chen, Z.; and Zhang, W. 2023b. MACO: A Modality Adversarial and Contrastive Framework for Modality-Missing Multi-modal Knowledge Graph Completion. In *NLPCC (1)*, volume 14302 of *Lecture Notes in Computer Science*, 123–134. Springer.

Zhang, Y.; and Zhang, W. 2022. Knowledge Graph Completion with Pre-trained Multimodal Transformer and Twins Negative Sampling. *CoRR*, abs/2209.07084.

Zhu, Y.; Zhao, H.; Zhang, W.; Ye, G.; Chen, H.; Zhang, N.; and Chen, H. 2021. Knowledge Perceived Multi-modal Pretraining in E-commerce. In *ACM Multimedia*, 2744–2752. ACM.