MyGO: Discrete Modality Information as Fine-Grained Tokens for Multi-modal Knowledge Graph Completion

Yichi Zhang[1,2], Zhuo Chen[1,2], Lingbing Guo[1,2], Yajing Xu[1,2], Binbin Hu[3], Ziqi Liu[3]
Huajun Chen[1,2,4] and Wen Zhang[1,2*]
[1]Zhejiang University [2]Zhejiang University-Ant Group Joint Laboratory of Knowledge Graph
[3]Ant Group [4]Alibaba-Zhejiang University Joint Institute of Frontier Technology
Hang Zhou, China
{zhangyichi2022,zhang.wen}@zju.edu.cn

ABSTRACT

Multi-modal knowledge graphs (MMKG) store structured world knowledge containing rich multi-modal descriptive information. To overcome their inherent incompleteness, multi-modal knowledge graph completion (MMKGC) aims to discover unobserved knowledge from given MMKGs, leveraging both structural information from the triples and multi-modal information of the entities. Existing MMKGC methods usually extract multi-modal features with pre-trained models and employ a fusion module to integrate multi-modal features with triple prediction. However, this often results in a coarse handling of multi-modal data, overlooking the nuanced, fine-grained semantic details and their interactions. To tackle this shortfall, we introduce a novel framework MyGO to **process, fuse, and augment the fine-grained modality information from MMKGs**. MyGO tokenizes multi-modal raw data 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 20 of the latest models, underlining its superior performance. Code and data are available at https://github.com/zjukg/MyGO.

CCS CONCEPTS

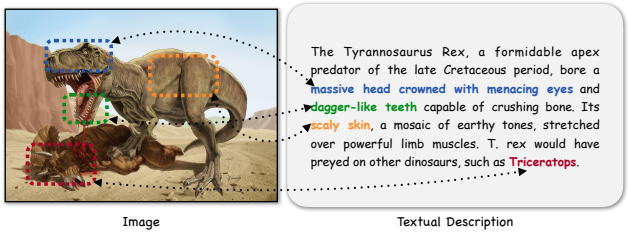
• **Information systems** → *Multimedia and multimodal retrieval*; *Data mining*; *Information integration*;

KEYWORDS

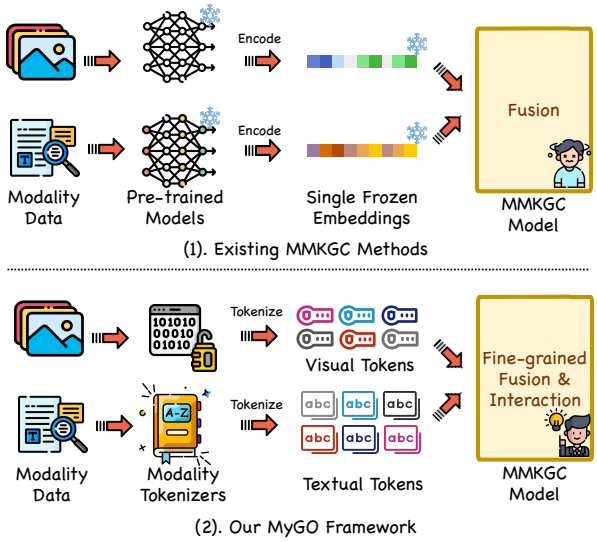
Multi-modal Knowledge Graphs, Knowledge Graph Completion, Modality Tokenization, Cross-modal Interaction

1 INTRODUCTION

Multi-modal knowledge graphs (MMKGs) [8] 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 [28], multi-modal understanding [53], and large language models [7, 10]. MMKGs furnish these systems with a dependable source of factual knowledge.



(a) Fine-grained multi-modal semantic units and their correspondences.



(b) Intuitions of existing MMKGC methods and MyGO.

Figure 1: There are fine-grained semantic interaction in the multi-modal information of MMKGs. MyGO attempts to tokenize raw modality-data into fine-grained tokens and learn the fine-grained entity representations with the multi-modal token sequences.

* Corresponding author.

ACM MM, 2024, Melbourne, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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)** [8], which aims to automatically identify new knowledge from the given MMKGs. Unlike conventional knowledge graph completion (KGC) [2, 29] 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 the triple prediction.

Existing MMKGC methods [3, 18, 20, 26] tend to represent modality information as single embedding derived from pre-trained models [9], 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 present 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 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. For instance, we present a simple case in Figure 1(a) to show these fine-grained semantic units in the image and text of an entity T-Rex, which are the image segments and the textual phrases. 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(b) 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 [9, 24], 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 [21, 43]. Comparisons against 20 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:

- 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.
- We propose a fine-grained contrastive learning module to augment the cross-modal entity representation learning process. This module innovates by employing new tactics to generate high-quality comparative samples for more detailed and effective self-supervised contrastive learning.
- We conduct comprehensive experiments on public benchmarks and achieve new state-of-the-art performance on MMKGC against 20 classic baseline methods. We perform further exploration with extensive experiments.

2 RELATED WORKS

2.1 Multi-modal Knowledge Graph Completion

Multi-modal knowledge graphs (MMKGs) [5, 8] are knowledge graphs with rich multi-modal information like images, text descriptions, audio, and videos [38]. Due to the incompleteness of the knowledge graphs, knowledge graph completion (KGC) [2, 15, 29, 31, 45] is a popular research topic to automatically discover unobserved knowledge triples by learning from the triple structure. Multi-modal knowledge graph completion (MMKGC) 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, (2) integrated decision, and (3) negative sampling. Methods of the first category [3, 6, 26, 37, 42, 44, 48] design complex mechanisms to achieve multi-modal fusion and interaction in the representation space. For example, OTKGE proposes an optimal transport-based multi-modal fusion strategy to find the optimal weights for multi-modal fusion. The second category methods [20, 52] usually learn a discriminate model for each modality and ensemble them to make joint decisions. IMF [20] proposes an interactive multi-modal fusion method to achieve multi-modal fusion and learns four different MMKGC models with different modality information to achieve joint decisions. The third category methods [43, 47, 49, 50] aim to enhance the negative sampling process [2] with the multi-modal information of entities to generate high-quality negative samples. Overall, these MMKGC methods usually leverage the multi-modal information by extracting feature representations from pre-trained models [9, 27]. However, the feature processing them neglects the fine-grained semantic information in each modality. We will solve this problem by tokenizing the modality information into fine-grained tokens.

2.2 Multi-modal Information Tokenization

Tokenization is a widely used technology in the NLP field to process the input text into a token sequence and learn fine-grained textual representations of strings and subwords. Due to the characteristics of textual modality itself, tokenization is very effective and has

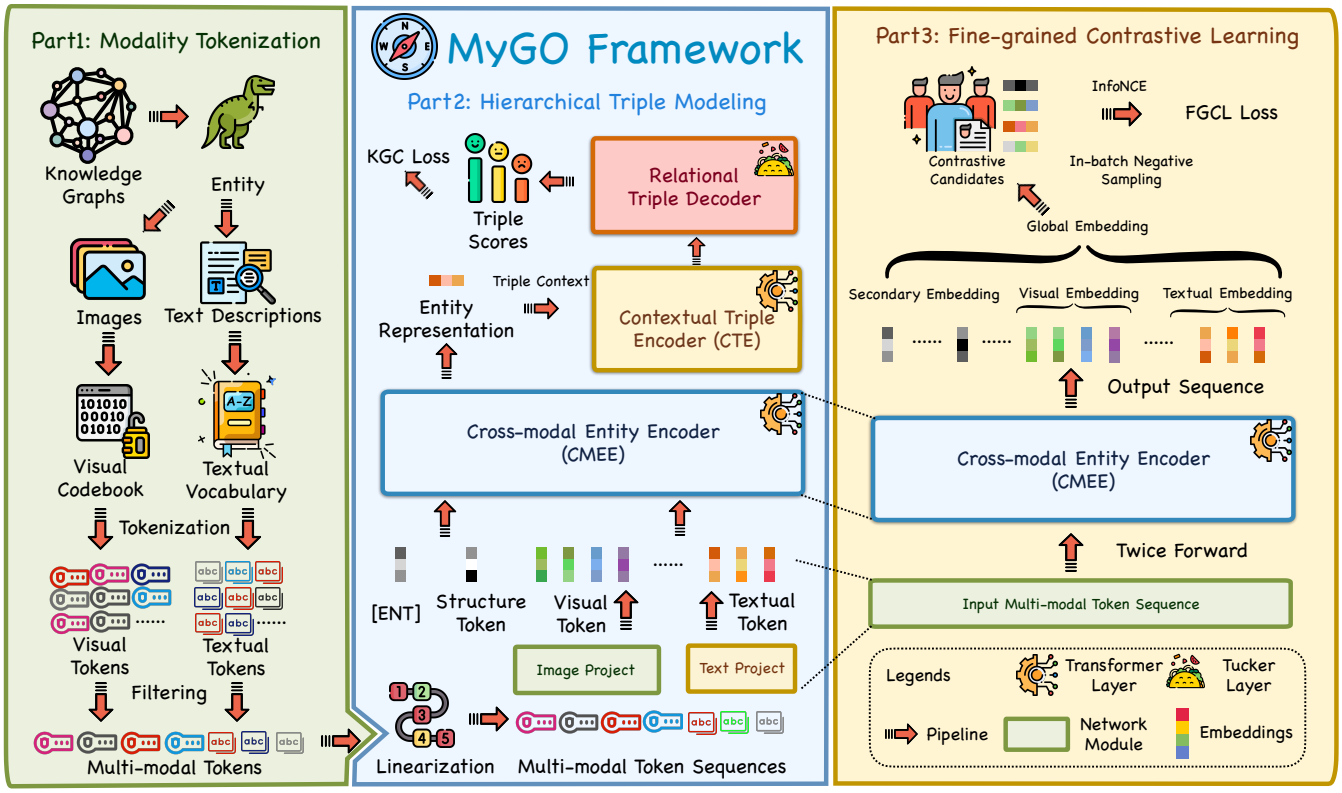


Figure 2: The overview of our MyGO framework. We mainly have three parts of new designs in MyGO to process, interact, and augment the fine-grained multi-modal semantic information in the MMKGs. They are named modality tokenization, hierarchical triple modeling, and fine-grained contrastive learning.

been widely used in language models (LM). For example, BPE [12], WordPiece [41], and ULM [17] are the most famous tokenization methods. For information in other modalities, tokenization becomes relatively difficult as there is no clear separation point for these modalities, which is different from texts. vector quantization (VQ) [11, 33] is an important technology proposed to map large-scale data into a fixed-length discrete codebook, where each code in the codebook is a vector representing certain specific features. Therefore, the non-textual modality information can be firstly processed into patch sequences and then each patch is mapped into a discrete code, which can be regarded as multi-modal tokens and further leveraged in many tasks [24, 25]. VQ has the advantage of compressing multimodal data while preserving a wide variety of fine-grained modal features with discrete code. For example, BEIT-v2 [24] processes each image into 196 patches in the size of 16x16 and maps each patch into a discrete code. In our work, we will also employ VQ and tokenization to process the multi-modal information in MMKGs and obtain fine-grained multi-modal representation for entities.

3 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 $\mathcal{S}(h, r, t) : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \rightarrow \mathbb{R}$ 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 [2] during training. Expanding to MMKGs, **multi-modal knowledge graph completion (MMKGC)** would further consider the multi-modal information $\mathcal{V}(e), \mathcal{D}(e)$ of each entity e to enhance their embeddings. Contemporary methods typically create embeddings for each modality and integrate these to calculate triple scores, with various multi-modal fusion techniques [3, 20, 37] employed to enhance performance.

In the inference stage, MMKGC models with predict the missing entities for a given query $(h, r, ?)$ or $(?, r, t)$. Taking tail prediction $(h, r, ?)$ for instance, MMKGC models will treat each entity $e \in \mathcal{E}$ as a candidate entity and calculate its corresponding score (h, r, e). Further, the models are evaluated by the rank of the golden answer (h, r, t) against all candidates, which means rank-based metrics [2, 29] will be employed for performance evaluation. It is mirrored for the head predictions and the overall performance usually considers both head and tail predictions on the test triples.

4 METHODLOGY

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). We consider the mainstream MMKGC setting [43] that includes both image and text modalities [8]. MyGO mainly comprises of 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.

4.1 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:

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

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

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 [9] while the visual tokens are from the codebook of a pre-trained visual tokenizer [11, 24]. 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 since certain subwords might appear multiple times within a sentence, and analogous semantic elements could recur across entity images. 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 [40] in the textual descriptions as their contribution to the entity semantics is minimal. After such refinement, we reserve m visual tokens and n textual tokens for each entity in the MMKG. For those entities with insufficient tokens to scant or absent modality data, we add a special padding token to fill the gaps. After the MT and refinement process, we can obtain processed token set $\mathcal{U}'_{img}$ and $\mathcal{U}'_{txt}$ 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.

Unlike existing MMKGC methods, the MT technique converts the modality information into more fine-grained discrete tokens. When facing multiple information in one modality (e.g. multiple images for one entity), traditional MMKGCs would make an aggregation (e.g. mean averaging [49]) on them before. However, MT preserves a sequence of tokens that represent the most prevalent features from various raw data sources, which can be more stable and scalable for increasing modality information. We will demonstrate this point in the experiments.

4.2 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 stepwise manner, which consists of three components: cross-modal entity encoder, contextual triple encoder, and relational decoder.

4.2.1 **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 [3, 20, 49] that represent each modality with a single embedding and then devise a fusion strategy to merge them, MyGO performs fine-grained tokenization of the different modalities and obtains a sequence of tokens for each modality. Therefore, we design a more **fine-grained feature interaction method** that allows for full interaction between all the different modal messages. In MyGO, we apply a transformer [35] layer as the CMEE. We first linearize the multi-modal tokens as a sequence $\mathcal{X}$ as:

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

where [ENT] is a special token and s_e is a learnable embedding representing the structural information of the entity. [ENT] is analogous the [CLS] token in BERT [9] 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:

$$\widehat{v}_{e,i} = \mathcal{P}_{img}(v_{e,i}) + b_{img} \quad \widehat{w}_{e,j} = \mathcal{P}_{txt}(v_{e,j}) + b_{txt} \tag{4}$$

where b_{img}, b_{txt} are defined modality biases to enhance the labeling of information from distinct modalities. Rather than adjusting the base token features, we aim to improve their integration by training projection layers for superior generalization. In this way, the final sequence entered into CMEE becomes:

$$\mathcal{X}_{input}(e) = ([\texttt{ENT}], s_e, \widehat{v}_{e,1}, \cdots, \widehat{v}_{e,m}, \widehat{w}_{e,1}, \cdots, \widehat{w}_{e,n}) \tag{5}$$

The cross-modal entity representation can be captured by:

$$\mathbf{e} = \textbf{Pooling}(\textbf{Transformer}(\mathcal{X}_{input}(e))) \tag{6}$$

where **Transformer**() represents a conventional transformer encoder layer with self-attention and feed-forward layers [35], **Pooling** is the pooling operation with obtains the final hidden representation of the special token [ENT]. It allows each token in the input sequence can be dynamically highlighted by CMEE to interact and eventually learn expressive entity representations.

4.2.2 **Contextual Triple Encoder**. To achieve adequate modality interaction in the relational context, we apply another transformer layer sd **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:

$$\tilde{\mathbf{h}} = \textbf{Transformer}([\texttt{CXT}], \mathbf{h}, \mathbf{r}) \tag{7}$$

where [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.

4.2.3 **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 [1] as our score function which is denoted as:

$$\mathcal{S}(h, r, t) = \mathcal{W} \times_1 \tilde{\mathbf{h}} \times_2 \tilde{\mathbf{r}} \times_3 \mathbf{t} \tag{8}$$

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 training objective is a cross-entropy loss:

$$\mathcal{L}_{head} = -\sum_{(h,r,t)\in\mathcal{T}} \log \frac{\exp(\mathcal{S}(h, r, t))}{\sum_{t'\in\mathcal{E}} \exp(\mathcal{S}(h, r, t'))} \tag{9}$$

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 and tail prediction are considered in MyGO, and the objective $\mathcal{L}_{tail}$ is similar when giving a tail query $(?, r, t)$:

$$\mathcal{L}_{tail} = -\sum_{(h,r,t)\in\mathcal{T}} \log \frac{\exp(\mathcal{S}(h, r, t))}{\sum_{h'\in\mathcal{E}} \exp(\mathcal{S}(h', r, t))} \tag{10}$$

The overall MMKGC task objective can be denoted as:

$$\mathcal{L}_{kgc} = \mathcal{L}_{head} + \mathcal{L}_{tail} \tag{11}$$

4.3 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 [13], 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:

$$\mathcal{X}_{output}(e) = ([\texttt{ENT}]', s'_e, \widehat{v}'_{e,1}, \cdots, \widehat{v}'_{e,m}, \widehat{w}'_{e,1}, \cdots, \widehat{w}'_{e,n}) \tag{12}$$

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:

$$\mathbf{s}(e) = \textbf{Mean}(\mathcal{X}_{output}(e)) \tag{13}$$

$$\mathbf{v}(e) = \frac{1}{m}\sum_{i=1}^{m} \widehat{v}'_{e,i} \quad \mathbf{w}(e) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}'_{e,i} \tag{14}$$

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.

For each entity e, we can collect its candidates for contrastive learning as $C(e) = \{e_{sec}, \mathbf{s}(e), \mathbf{v}(e), \mathbf{w}(e)\}$, which consists of its global and local features. $(\mathbf{e}, \mathbf{e}')$ where $\mathbf{e}' \in C(e)$ is regarded as a positve sample. Then we employ in-batch negative sampling to construct negative pairs and InfoNCE [32] as the contrastive backbone. The final FBCL objective can be denoted as:

$$\mathcal{L}_{con} = -\sum_{i=1}^{\mathcal{B}} \sum_{e'_i \in C(e_i)} \log \frac{\exp(\textbf{cos}(\mathbf{e}_i, \mathbf{e}'_i)/\tau)}{\sum_{j=1}^{\mathcal{B}} \exp(\textbf{cos}(\mathbf{e}_i, \mathbf{e}'_j)/\tau)} \tag{15}$$

where $\mathcal{B}$ is the batch size, $\textbf{cos}(\cdot,\cdot)$ 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:

$$\mathcal{L} = \mathcal{L}_{kgc} + \lambda \mathcal{L}_{con} \tag{16}$$

where λ is a hyper-parameter to control the weight of the contrastive loss $\mathcal{L}_{con}$.

Table 1: Statistical information of the datasets.

Dataset	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{R}\rvert$	#Train	#Valid	#Test	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{D}\rvert$
DB15K	12842	279	79222	9902	9904	12818	12842
MKG-W	15000	169	34196	4276	4274	14463	14123

5 EXPERIMENTS

In this section, we will conduct comprehensive experiments to evaluate the performance of MyGO. We begin by detailing our experimental setup and subsequently present an analytical assessment of the results. We aim to address the following research questions (RQs) in our study:

RQ1. Can MyGO outperform the existing baseline methods and make meaningful progress in MMKGC?

Table 2: The main MMKGC results on DB15K [21] and MKG-W [43]. 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 underlined.

Model		Fusion Strategy	DB15K				MKG-W			
			MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
Uni-modal	**TransE** [2]	None	24.86	12.78	31.48	47.07	29.19	21.06	33.20	44.23
	DistMult [45]	None	23.03	14.78	26.28	39.59	20.99	15.93	22.28	30.86
	ComplEx [31]	None	27.48	18.37	31.57	45.37	24.93	19.09	26.69	36.73
	RotatE [29]	None	29.28	17.87	36.12	49.66	33.67	26.80	36.68	46.73
	PairRE [4]	None	31.13	21.62	35.91	49.30	34.40	28.24	36.71	46.04
	GC-OTE [30]	None	31.85	22.11	36.52	51.18	33.92	26.55	35.96	46.05
	Tucker [1]	None	33.86	25.33	37.91	50.38	30.39	24.44	32.91	41.25
Multi-modal	**IKRL** [42]	Static	26.82	14.09	34.93	49.09	32.36	26.11	34.75	44.07
	TBKGC [26]	Static	28.40	15.61	37.03	49.86	31.48	25.31	33.98	43.24
	TransAE [39]	Static	28.09	21.25	31.17	41.17	30.00	21.23	34.91	44.72
	MMKRL [22]	Static	26.81	13.85	35.07	49.39	30.10	22.16	34.09	44.69
	RSME [37]	Adaptive	29.76	24.15	32.12	40.29	29.23	23.36	31.97	40.43
	VBKGC [51]	Static	30.61	19.75	37.18	49.44	30.61	24.91	33.01	40.88
	OTKGE [3]	Adaptive	23.86	18.45	25.89	34.23	34.36	28.85	36.25	44.88
	IMF [20]	Adaptive	32.25	24.20	36.00	48.19	34.50	28.77	36.62	45.44
	QEB [38]	Static	28.18	14.82	36.67	51.55	32.38	25.47	35.06	45.32
	VISTA [18]	Adaptive	30.42	22.49	33.56	45.94	32.91	26.12	35.38	45.61
	AdaMF [49]	Adaptive	32.51	21.31	39.67	51.68	34.27	27.21	37.86	47.21
Negative Sampling	**MANS** [47]	Static	28.82	16.87	36.58	49.26	30.88	24.89	33.63	41.78
	MMRNS [43]	Adaptive	32.68	23.01	37.86	51.01	35.03	28.59	37.49	47.47
MyGO		Adaptive	**37.72**	**30.08**	**41.26**	**52.21**	**36.10**	**29.78**	**38.54**	**47.75**
			+11.4%	+18.4%	+4.0%	+1.0%	+3.1%	+3.2%	+1.8%	+0.6%

RQ2. Can MyGO leverage more modal information compared to baseline methods? How does the number of tokens retained during Modality Tokenization affect the results?

RQ3. Is the design of each module in MyGO necessary and reasonable to unlock the model potential?

RQ4. Can we visually demonstrate the quality of the multi-modal token representations learned by the model?

5.1 Experiment Settings

5.1.1 **Datasets**. In this paper, we employ two public MMKGC benchmarks DB15K [21] and MKG-W [43] to evaluate the model performance. DB15K derives from DBpedia [19] and MKG-W is a subset of WikiData [36]. Both of them consist of image and textual descriptions, providing a rich multi-modal context. Detailed information about the datasets is presented in Table 1. The raw data for each modality are obtained from their official release sources.

5.1.2 **Task and Evaluation Protocols**. We conduct link prediction [2] task on the datasets, which is the mainstream MMKGC task. Following existing works, we use rank-based metrics [29] like mean reciprocal rank (MRR) and Hit@K (K=1, 3, 10) to evaluate the results. Besides, we employ the filter setting [2] in the prediction results to remove the candidate triples existing in the training data for fair comparisons. As mentioned in Section 3, the final results are the average of both head prediction and tail prediction. MRR and Hit@K can be denoted as:

$$\text{MRR} = \frac{1}{|\mathcal{T}_{test}|} \sum_{i=1}^{|\mathcal{T}_{test}|} \left(\frac{1}{r_{h,i}} + \frac{1}{r_{t,i}}\right) \tag{17}$$

$$\text{Hit@K} = \frac{1}{|\mathcal{T}_{test}|} \sum_{i=1}^{|\mathcal{T}_{test}|} \left(\mathbf{1}(r_{h,i} \leq K) + \mathbf{1}(r_{t,i} \leq K)\right) \tag{18}$$

where $r_{h,i}$ and $r_{t,i}$ are the results of head prediction and tail prediction, and $\mathcal{T}_{test}$ is the test triple set.

5.1.3 **Baseline Methods**. To make a comprehensive performance evaluation, we employ 20 different state-of-the-art baselines in our experiments. We focus on the ability of the methods to exploit multi-modal information in our comparison and select three types of baselines:

(1). **Uni-modal KGC methods**: TransE [2], DistMult [45], ComplEx [31], RotatE [29], PairRE [4], GC-OTE [30], and Tucker [1]. These conventional methods only consider the structure information in their model design.

(2). **MMKGC methods**: IKRL [42], TBKGC [26], TransAE [39], MMKRL [22], RSME [37], VBKGC [51], OTKGE [3], IMF [20], QEB [38], VISTA [18], and AdaMF [49]. These methods consider image and textual information in the MMKGC models and achieve multi-modal fusion with different settings.

(3). **Negative sampling methods**: MANS [47], and MMRNS [43]. These methods utilize the multi-modal information in the MMKGs to generate high-quality negative samples for training.

Some of the methods like KG-BERT [46] for fine-tuning pre-trained language models are orthogonal to our design, so we did not compare our method to them.

5.1.4 **Implementation Details**. In our experiments, we implement MyGO with PyTorch [23]. For modality tokenization, we employ the tokenizer of BEIT [9] and BERT [9] as our visual/textual tokenizers. The codebook size of BEIT [24] is 8192 and the vocabulary size of BERT tokenizer is 32000. We keep 3 images for each entity with visual information in the main experiments. The feature dimensions of the visual and textual tokens are 32 and 768 designed by the original models. During training, we set the training epoch to 2000, the batch size to 1024, and the embedding dimension to 256. For the transformer layers, we employ 1 transformer layer for both CMEE and CTE with 4 attention heads. We set the dropout layer with $p \in \{0.3, 0.4, 0, 5\}$. The max token number m and n are tuned in $\{4, 8, 12\}$ and the contrastive loss weight λ is tuned in $\{1, 0.1, 0.01, 0.001\}$. The contrastive temperature τ is tuned in $\{0.1, 0.5, 1.0\}$. We optimize the model with Adam [16] optimizer and the learning rate is searched in $\{1e^{-4}, 3e^{-4}, 5e^{-4}\}$. For the baseline models, we reuse their official code or reproduce their results based on OpenKE [14]. All the experiments are conducted on a Linux server with one NVIDIA A800 GPU, taking 1 to 5 hours to finish the training and evaluation on different datasets.

5.2 Main Results (RQ1)

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 transformer encoders so that it is an adaptive method as well.

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 performance through fine-grained interactive fusion with transformer-based encoders in HTM. We think that's what makes MyGO outperform the adaptive fusion baselines. 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 multi-modal tokenization and fusion process.

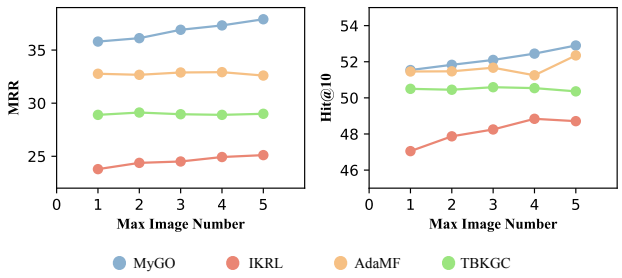


Figure 3: The MMKGC results (MRR, Hit@10) of MyGO and baselines with differen number of entity images on DB15K.

5.3 Exploration on Modality Tokenization (RQ2)

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. As mentioned in **RQ2**, 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 different token amounts.

5.3.1 **Multiple Modality Information.** As outlined in Section 4.1, 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 pre-processing 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 on the MMKGC results 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 experimental 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 performance of other methods is somewhat erratic, displaying fluctuations as the image amount increases, and their overall effectiveness does not match ours. 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 in an image 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.

5.3.2 **Impact of Token Amount.** Another intriguing aspect to investigate pertains to the two token amount hyper-parameters m and n that we set in MT. These parameters dictate the number of high-frequency multi-modal tokens retained and processed by

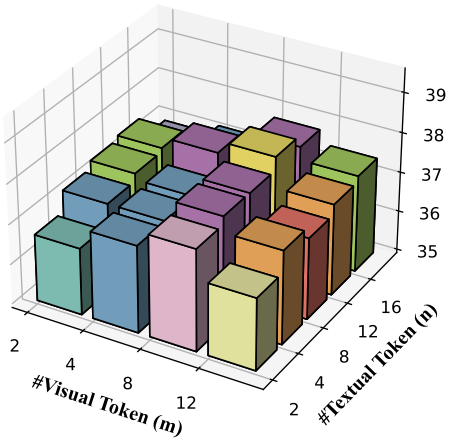


Figure 4: The MRR results with different m and n on DB15K.

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. Simultaneously, while the number of tokens increased, we measured that the model efficiency gradually increased from 4s/epoch to about 20s/epoch. In conclusion, the optimal number of tokens should be $m = 8, n = 12$, which takes about 10s for one training epoch. Relative to existing methods, this efficiency induces minimal latency, while potentially realizing state-of-the-art performance.

5.4 Ablation Study (RQ3)

To confirm the effectiveness of each module in MyGO, we further conduct an ablation study from three perspectives: the modality information, the model design, and the fine-grained contrastive loss. We removed the corresponding modules across different settings and performed MMKGC experiments. The experimental results are presented in Table 3.

From the first set of experiments, the utilization of both modalities significantly enhances performance, validating that MyGO effectively learns from multi-modal information. According to the second set of 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 FCGL also contributes to the model performance, with a ccontrastivecandidate in $C(e)$ 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 $\lambda = 0.01$.

Table 3: The ablation study results on DB15K. We validated the design in three ways including modality information, model design, and constative loss (FGCL). The weight of FGCL λ is set to 0.01 for groups not involving this parameter.

Setting		MRR	Hit@10	Hit@3	Hit@1
Modality Information	w/o Image	36.31	51.08	39.81	28.72
	w/o Text	37.52	51.98	40.94	29.97
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
FBCL	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**

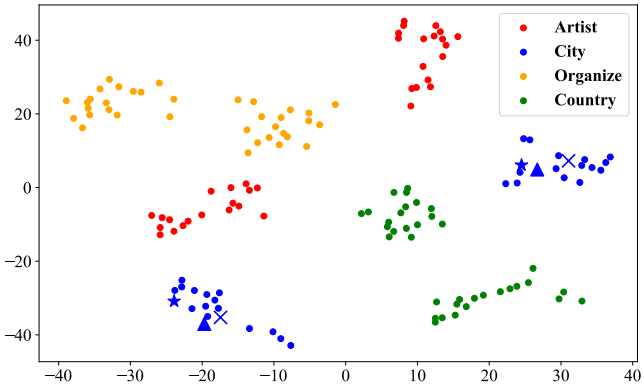


Figure 5: The embedding visualization results of the multi-modal tokens from entities with different categories (Artist, City, Organization, and Country) in DB15K.

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 modality-aware triple prediction. The FCGL module further makes more of further boosts model performance based on the backbone.

5.5 Token Embedding Visualization (RQ4)

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 [34] algorithm. For the city category, we uniquely mark the same

tokens in both city entities using special marks (△, ★, ×) to demonstrate the contextual multi-modal embeddings of the same tokens under different token sequences.

Figure 5 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.

6 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 the raw multi-modal information into multi-modal token sequences and extract fine-grained multi-modal entity representation with a hierarchical multi-modal fusion transformer architecture. We also propose a fine-grained contrastive learning module to further augment the quality of entity representations to achieve better performance. 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 and try to integrate such a tokenization technology in more complex MMKG downstream application tasks.

REFERENCES

[1] Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. 2019. TuckER: Tensor Factorization for Knowledge Graph Completion. In *EMNLP/IJCNLP (1)*. Association for Computational Linguistics, 5184–5193.

[2] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NIPS*. 2787–2795.

[3] Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. OTKGE: Multi-modal Knowledge Graph Embeddings via Optimal Transport. In *NeurIPS*.

[4] Linlin Chao, Jianshan He, Taifeng Wang, and Wei Chu. 2021. PairRE: Knowledge Graph Embeddings via Paired Relation Vectors. In *Proc. of ACL*.

[5] Zhuo Chen, Jiaoyan Chen, Wen Zhang, Lingbing Guo, Yin Fang, Yufeng Huang, Yichi Zhang, Yuxia Geng, Jeff Z. Pan, Wenting Song, and Huajun Chen. 2023. MEAformer: Multi-modal Entity Alignment Transformer for Meta Modality Hybrid. In *ACM Multimedia*. ACM, 3317–3327.

[6] Zhuo Chen, Yin Fang, Yichi Zhang, Lingbing Guo, Jiaoyan Chen, Huajun Chen, and Wen Zhang. 2024. The Power of Noise: Toward a Unified Multi-modal Knowledge Graph Representation Framework. arXiv:2403.06832 [cs.CL]

[7] Zhuo Chen, Wen Zhang, Yufeng Huang, Mingyang Chen, Yuxia Geng, Hongtao Yu, Zhen Bi, Yichi Zhang, Zhen Yao, Wenting Song, Xinliang Wu, Yi Yang, Mingyi Chen, Zhaoyang Lian, Yingying Li, Lei Cheng, and Huajun Chen. 2023. Tele-Knowledge Pre-training for Fault Analysis. In *ICDE*. IEEE, 3453–3466.

[8] Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, Jiaqi Li, Xiaoze Liu, Jeff Z. Pan, Ningyu Zhang, and Huajun Chen. 2024. Knowledge Graphs Meet Multi-Modal Learning: A Comprehensive Survey. *CoRR* abs/2402.05391 (2024).

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 4171–4186.

[10] Junnan Dong, Qinggang Zhang, Huachi Zhou, Daochen Zha, Pai Zheng, and Xiao Huang. 2024. Modality-Aware Integration with Large Language Models for Knowledge-based Visual Question Answering. *CoRR* abs/2402.12728 (2024).

[11] Patrick Esser, Robin Rombach, and Björn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*. Computer Vision Foundation / IEEE, 12873–12883.

[12] Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal* 12, 2 (1994), 23–38.

[13] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP (1)*. Association for Computational Linguistics, 6894–6910.

[14] Xu Han, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. OpenKE: An Open Toolkit for Knowledge Embedding. In *Proc. of EMNLP*.

[15] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge Graph Embedding via Dynamic Mapping Matrix. In *ACL (1)*. The Association for Computer Linguistics, 687–696.

[16] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

[17] Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *ACL (1)*. Association for Computational Linguistics, 66–75.

[18] Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Jiyoung Whang. 2023. VISTA: Visual-Textual Knowledge Graph Representation Learning. In *EMNLP (Findings)*. Association for Computational Linguistics, 7314–7328.

[19] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* (2015).

[20] Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023. IMF: Interactive Multimodal Fusion Model for Link Prediction. In *WWW*. ACM, 2572–2580.

[21] Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. 2019. MMKG: Multi-modal Knowledge Graphs. In *ESWC (Lecture Notes in Computer Science, Vol. 11503)*. Springer, 459–474.

[22] Xinyu Lu, Lifang Wang, Zejun Jiang, Shichang He, and Shizhong Liu. 2022. MMKRL: A robust embedding approach for multi-modal knowledge graph representation learning. *Appl. Intell.* 52, 7 (2022), 7480–7497.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*. 8024–8035.

[24] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers. *CoRR* abs/2208.06366 (2022).

[25] Michael S. Ryoo, A. J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. 2021. TokenLearner: Adaptive Space-Time Tokenization for Videos. In *NeurIPS*. 12786–12797.

[26] Hatem Mousselly Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A Multimodal Translation-Based Approach for Knowledge Graph Representation Learning. In **SEM@NAACL-HLT*. Association for Computational Linguistics, 225–234.

[27] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

[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*. ACM, 1405–1414.

[29] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *ICLR (Poster)*. OpenReview.net.

[30] Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou. 2020. Orthogonal Relation Transforms with Graph Context Modeling for Knowledge Graph Embedding. In *Proc. of ACL*.

[31] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *ICML (JMLR Workshop and Conference Proceedings, Vol. 48)*. JMLR.org, 2071–2080.

[32] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR* abs/1807.03748 (2018).

[33] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. In *NIPS*. 6306–6315.

[34] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*. 5998–6008.

[36] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[37] 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 *ACM Multimedia*. ACM, 2735–2743.

[38] Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. 2023. TIVA-KG: A Multimodal Knowledge Graph with Text, Image, Video and

Audio. In *ACM Multimedia*. ACM, 2391–2399.
[39] Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. 2019. Multimodal Data Enhanced Representation Learning for Knowledge Graphs. In *IJCNN*. IEEE, 1–8.
[40] W. John Wilbur and Karl Sirotkin. 1992. The automatic identification of stop words. *J. Inf. Sci.* 18, 1 (1992), 45–55.
[41] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR* abs/1609.08144 (2016).
[42] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied Knowledge Representation Learning. In *IJCAI*. ijcai.org, 3140–3146.
[43] Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. 2022. Relation-enhanced Negative Sampling for Multimodal Knowledge Graph Completion. In *ACM Multimedia*. ACM, 3857–3866.
[44] Derong Xu, Jingbo Zhou, Tong Xu, Yuan Xia, Ji Liu, Enhong Chen, and Dejing Dou. 2023. Multimodal Biological Knowledge Graph Completion via Triple Co-Attention Mechanism. In *ICDE*. IEEE, 3928–3941.
[45] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *ICLR (Poster)*.
[46] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for Knowledge Graph Completion. *CoRR* abs/1909.03193 (2019).
[47] Yichi Zhang, Mingyang Chen, and Wen Zhang. 2023. Modality-Aware Negative Sampling for Multi-modal Knowledge Graph Embedding. In *IJCNN*. IEEE, 1–8.
[48] Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. 2024. NativE: Multi-modal Knowledge Graph Completion in the Wild. *Authorea Preprints* (2024).
[49] Yichi Zhang, Zhuo Chen, Lei Liang, Huajun Chen, and Wen Zhang. 2024. Unleashing the Power of Imbalanced Modality Information for Multi-modal Knowledge Graph Completion. *CoRR* abs/2402.15444 (2024).
[50] Yichi Zhang, Zhuo Chen, and Wen Zhang. 2023. MACO: A Modality Adversarial and Contrastive Framework for Modality-Missing Multi-modal Knowledge Graph Completion. In *NLPCC (1) (Lecture Notes in Computer Science, Vol. 14302)*. Springer, 123–134.
[51] Yichi Zhang and Wen Zhang. 2022. Knowledge Graph Completion with Pre-trained Multimodal Transformer and Twins Negative Sampling. *CoRR* abs/2209.07084 (2022).
[52] 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 *EMNLP*. Association for Computational Linguistics, 10527–10536.
[53] Yushan Zhu, Huaixiao Zhao, Wen Zhang, Ganqiang Ye, Hui Chen, Ningyu Zhang, and Huajun Chen. 2021. Knowledge Perceived Multi-modal Pretraining in E-commerce. In *ACM Multimedia*. ACM, 2744–2752.