Towards Multimodal Continual Knowledge Embedding with Modality Forgetting Modulation

Xiaowen Jiang[1], Jing Yang[1*], ShunDong Yang[1], Yuan Gao[1], Xinfa Jiang[2], Laurence Tianruo Yang[1, 3], Jieming Yang[1]

[1]Zhengzhou University, China
[2]Hainan University, China
[3]St. Francis Xavier University, Canada
{jxw, xinfa.jiang}@hainanu.edu.cn, yangjinghust123@gmail.com, shundongyang146@163.com,
{yuangao, jmyang}@zzu.edu.cn, ltyang@ieee.org

Abstract

The continuous emergence of new entities, relations, triples, and multimodal information drives the dynamic evolution of multimodal knowledge graph (MMKG). However, existing MMKG embedding models follow a static setting, where training from scratch for growing MMKG wastes learned knowledge, while fine-tuning on new knowledge easily leads to catastrophic forgetting, severely limiting their applicability in real-world scenarios. To address this, we propose a multimodal continual representation learning framework (MoFot) for growing MMKG. Unlike existing static multimodal embedding methods, MoFot focuses on alleviating catastrophic forgetting rather than retraining to adapt to new knowledge. Specifically, MoFot effectively mitigates catastrophic forgetting caused by parameter updates and differing forgetting rates across modalities through a multimodal collaborative modulation mechanism. The mechanism ensures consistent retention of previously learned multimodal knowledge across snapshots through multimodal weight modulation and multimodal feature modulation. MoFot outperforms existing MMKG embedding, KG continual learning, and MMKG inductive models. Experimental results demonstrate that MoFot not only avoids forgetting but also enhances old knowledge by learning new knowledge, achieving adaptation to new knowledge while mitigating forgetting of old knowledge.

Code — https://github.com/hncps6/MoFot

Introduction

Multimodal Knowledge Graph (MMKG) integrates multimodal data such as text, images, and structural information, providing a unified knowledge representation framework (Wang et al. 2019b), (Mousselly-Sergieh et al. 2018). MMKG supports knowledge reasoning capabilities in various knowledge-driven applications, including recommendation systems (Wei et al. 2024), intelligent QA (Zhang et al. 2024a), semantic search(Anderson et al. 2016), and cross-modal reasoning (Yang, Li, and Yu 2019). Throughout the lifecycle of MMKG construction and maintenance, new entities and relations with multimodal information continuously emerge

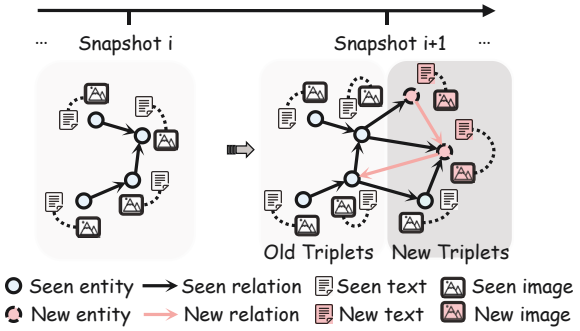

Figure 1: An example of growing MMKG. In each snapshot, new triples with new entities/relations/multimodal contexts are gradually added to the MMKG.

(Chen et al. 2024). Consequently, MMKG are rarely static; instead, they evolve and grow dynamically over time, as shown in Figure 1. Managing continuously expanding MMKG requires efficient integration and organization of newly added knowledge. However, existing multimodal knowledge embedding methods for link prediction are designed under static assumptions (Yang et al. 2024), (Li et al. 2023), (Chen et al. 2022), making them ineffective in handling continuously emerging knowledge. Models such as MKBE(Pezeshkpour, Chen, and Singh 2018), MKGC(Mousselly-Sergieh et al. 2018), VBKGC(Zhang and Zhang 2022), and NativE(Zhang et al. 2024a) aim to project modality-specific information into unified embedding spaces, thereby enriching entity representations. Similarly, IMF(Li et al. 2023) prioritizes independent learning of modality-specific features, whereas MoCi(Yang et al. 2024) captures inter-entity semantics across modalities and incorporates them. AdaMF(Zhang et al. 2024b) uses adaptive fusion to combine information from different modalities. Although the models can adapt to the expanding MMKG by retraining, they overlook the previously learned knowledge, resulting in low learning efficiency.

Researchers have proposed unimodal continual knowledge graph embedding(CKGE) models for link prediction, such as LKGE(Cui et al. 2023) and IncDE(Liu et al. 2024a). These methods are capable of acquiring new knowledge while retaining learned knowledge during continual learning. Specif-

*Corresponding author

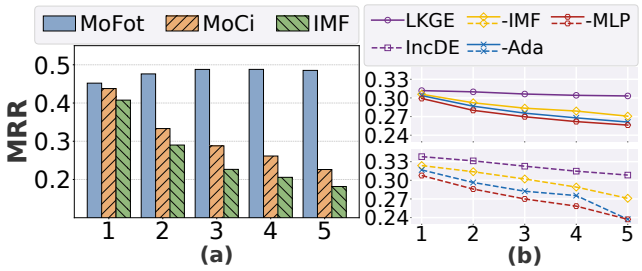

Figure 2: (a) shows the MRR forgetting trends of our model compared to the static MMKG model; (b) shows the MRR forgetting trends of CKGE model and its multimodal variant. All models are sequentially trained or fine-tuned on the new data, and then evaluated on the test set of the first snapshot.

ically, LKGE employs a regularization module to mitigate the forgetting of previously learned knowledge during parameter updates, while IncDE introduces a structure-aware incremental distillation mechanism to effectively preserve old knowledge. However, these models were originally designed for unimodal tasks and did not account for the integration of multimodal features, whereas in real-world scenarios, newly emerging data often includes multimodal content.

Figure 2(a) shows the performance of MMKG models (MoCi(Yang et al. 2024) and IMF(Li et al. 2023)) on the *Snapshot 1* after being fine-tuned on each new snapshot. Although fine-tuning enables these models to adapt to new snapshots, it comes at the cost of forgetting previously learned knowledge. Specifically, as the model is continuously fine-tuned on each new snapshot, its prediction accuracy on *Snapshot 1* significantly declines, indicating that while the model absorbs new knowledge, it gradually forgets the learned knowledge. This is mainly because each new snapshot brings new data, forcing the model to constantly adjust its parameters to adapt to new knowledge, which in turn overwrites learned knowledge. Figure 2(b) shows the performance of unimodal continual learning models (LKGE(Cui et al. 2023) and IncDE(Liu et al. 2024a)) after incorporating multimodal features. We adopt three fusion strategies: bilinear pooling fusion used in IMF(Li et al. 2023), adaptive fusion from AdaMF(Zhang et al. 2024b), and MLP. It can be observed that incorporating multimodal features exacerbates the forgetting of learned knowledge, with a forgetting rate even higher than that of unimodal models, consistent with the observations in (Chen et al. 2024). The reason is that different modalities face varying forgetting rates during the continual learning, compared to unimodal models. This can lead to the different modality features shifting to varying degrees, and some modalities being marginalized, which exacerbating performance degradation on previous tasks(Chen et al. 2024). Therefore, generic multimodal fusion methods struggle to perform effectively during continual knowledge learning.

To address the above issues, we propose MoFot, a multimodal continual learning model that leverages a multimodal collaborative modulation mechanism. MoFot aims to adapt to new knowledge while retaining old knowledge in growing MMKG. The multimodal collaborative modulation mechanism consists of dual branches: multimodal weight modulation and multimodal feature modulation. Firstly, we perform multimodal weight modulation by introducing a weight interpolation update module to uniformly adjust the new and old weights of each modality in direction and magnitude. Secondly, we employ multimodal feature modulation, which leverages path semantics that preserve old structural knowledge to guide multimodal features toward consistent representations across snapshots. The collaborative modulation mechanism promotes consistent retention of previously learned multimodal knowledge across snapshots, alleviating forgetting caused by differences in forgetting rates and parameter updates. Our contributions are as follows:

- We propose the first multimodal continual representation learning framework that mitigating forgetting caused by forgetting rate differences and parameter updates through a multimodal collaborative modulation mechanism. It enables learning on growing MMKG without retraining and prevents catastrophic forgetting of historical knowledge.
- We design a weight interpolation update module that balances new and old knowledge across modalities by uniformly adjusting weights, promoting consistent retention of multimodal knowledge in the space of weights
- We propose a historical path preservation method that identifies path semantics and applies regularization to retention. Meanwhile, we leverage path semantics to guide multimodal features, ensuring consistent retention of multimodal knowledge across snapshots.

Related Works

Multimodal Knowledge Graph Embedding Models

MMKG models integrate diverse modalities of data to enhance knowledge representation and link prediction performance. For instance, MKGformer(Chen et al. 2022) utilizes a transformer-based architecture to facilitate efficient multimodal fusion, while TransAE(Wang et al. 2019b) leverages an autoencoder structure to integrate multimodal embeddings and extract complementary features for entity representation. Similarly, IMF(Li et al. 2023) incorporates a bilinear pooling module to improve multimodal fusion and MoCi(Yang et al. 2024) propose a framework designed to capture inter-entity modality semantics. Models like MKGC(Mousselly-Sergieh et al. 2018), VBKGC(Zhang and Zhang 2022), IKRL(Xie et al. 2017), QEB(Wang et al. 2023), and AdaMF(Zhang et al. 2024b) also focus on multimodal integration. However, these models are limited in their ability to handle the expansion of knowledge graphs and fail to efficiently integrate knowledge from newly added snapshots.

Continual Knowledge Graph Embedding Models

CKGE models aim to learn new knowledge while preserving old knowledge. Existing CKGE models are generally classified into three categories: dynamic architecture models such as PNN(Rusu et al. 2016) and CWR(Lomonaco and Maltoni 2017); regularization-based models like SI(Zenke, Poole, and Ganguli 2017), EWC(Kirkpatrick et al. 2017), and LKGE(Cui et al. 2023); and rehearsal-based models such as

GEM(Lopez-Paz and Ranzato 2017) and EMR(Wang et al. 2019a). Dynamic architecture-based models adapt to new tasks by dynamically expanding network structures while retaining existing knowledge. Regularization-based models identify critical parameters in old knowledge and constrain their updates to retain old knowledge. Rehearsal-based models enhance the retention of old knowledge by replaying a portion of previously seen data during a new task learning. However, these models fail to effectively integrate multimodal information in continual learning.

Task Formulation

Growing Multimodal Knowledge Graph

A multimodal knowledge graph(MMKG) is defined as $\mathcal{G} = (\mathcal{F}, \mathcal{E}, \mathcal{R})$, where $\mathcal{F}$, $\mathcal{E}$, and $\mathcal{R}$ denote the set of triples, the set of entities, and the set of relations, respectively. Each triple (s, r, o) represents a connection between a head entity s and a tail entity o via relation r. Each entity in $\mathcal{E}$ is associated with multimodal contextual information, including textual(T), visual(V), and structural(S) modalities. When the MMKG grows to snapshot i, it is defined as $\mathcal{G}^i = (\mathcal{F}^i, \mathcal{E}^i, \mathcal{R}^i)$, where $\mathcal{F}^i$, $\mathcal{E}^i$, and $\mathcal{R}^i$ represent the set of triples, entities, and relations in $\mathcal{G}^i$, respectively. Each snapshot adds new triples, entities, and relations, so $\mathcal{F}^i \subseteq \mathcal{F}^{i+1}$, $\mathcal{E}^i \subseteq \mathcal{E}^{i+1}$, and $\mathcal{R}^i \subseteq \mathcal{R}^{i+1}$. The new additions introduced in snapshot i compared to snapshot $i-1$ are denoted as $\Delta\mathcal{E}^i = \mathcal{E}^i - \mathcal{E}^{i-1}$, $\Delta\mathcal{R}^i = \mathcal{R}^i - \mathcal{R}^{i-1}$, and $\Delta\mathcal{F}^i = \mathcal{F}^i - \mathcal{F}^{i-1}$, as well as the corresponding new multimodal contexts.

Continual MMKG Link Prediction

Multimodal knowledge graph(MMKG) link prediction task predicts missing head or tail entities in triples, i.e., $(s, r, ?)$ or $(?, r, o)$. We propose the continual MMKG link prediction task, in which the model is trained exclusively on new data to continually learn new knowledge, including triples, entities, relations, and multimodal contexts. For each snapshot i, the set of new triples $\Delta\mathcal{F}^i$ in each snapshot $\mathcal{G}^i$ is divided into a training set $\mathcal{T}^i$, a validation set $\mathcal{V}^i$, and a test set $\mathcal{Q}^i$ in a 3:1:1 ratio, i.e., $\Delta\mathcal{F}^i = \{\mathcal{T}^i \cup \mathcal{V}^i \cup \mathcal{Q}^i\}$. The model needs to continuously train on the training sets $\{\mathcal{T}^1, \mathcal{T}^2, ..., \mathcal{T}^I\}$ in sequence according to the snapshot order. After training, the overall evaluation is performed based on the test results accumulated from $\{\mathcal{Q}^1, \mathcal{Q}^2, ..., \mathcal{Q}^I\}$.

Methodology

Model Overview

Existing MMKG models tend to forget old knowledge when adapting to new knowledge. To address this issue, we propose MoFot. The overall framework of MoFot is shown in Figure 3. Specifically, we first utilize NBFNet(Zhu et al. 2021) to model path semantics between entity pairs via relation aggregation, enabling the extraction of structural commonalities and improving adaptability to new structural knowledge. Moreover, in continual learning, it is essential to adapting to new knowledge while preserve existing knowledge. However, NBFNet cannot preserve old structural knowledge. We thus propose a historical path preservation method that identifies relation-guided paths and applies regularization to retain old path semantics. Secondly, we use path semantics that retain old structural knowledge to guide the multimodal features, maintaining consistency of each modality's representation across snapshots and thereby ensuring consistent knowledge retention. Finally, we propose a weight interpolation update module that decomposes weights change into direction and magnitude, and fuses the old and new weights through interpolation along both components. This promotes consistent knowledge retention in the space of weights. By jointly modulating features and weights, MoFot effectively mitigates forgetting caused by parameter updates and differences in forgetting rates across modalities.

Relation-Driven Path Passing

NBFNet models path semantics for each query triple $(s, q, ?)$ in the i-th snapshot. Firstly, the embedding of the query relation q is assigned to the corresponding head entity s, while all other entities are initialized with zero vectors as their initial embeddings at the 0-th layer. Then, entity representations are updated by aggregating the embeddings of their neighboring relations and entities. Finally, the final entity representations encode the complete path semantics from the head entity s to all other entities. The entire process is defined as follows:

$$\boldsymbol{z}_{o|s,q}^{l+1} = \mathrm{Agg}\left\{\mathrm{Msg}\left(\boldsymbol{z}_{w|s,q}^{l}, \Psi_q(\boldsymbol{x}_r)\right) \mid w \in \mathcal{N}_r^i(o), r \in \mathcal{R}^i\right\} \tag{1}$$

where $\mathcal{R}^i$ is a set of existed relations in the i-th snapshot, $\boldsymbol{x}_r \in \mathbb{R}^d$ is learnable relation embedding parameters for relation r. $\mathcal{N}_r^i(o)$ denotes the neighborhoods of entity $o \in \mathcal{E}^i$ in snapshot i that are connected by relation r. $\boldsymbol{z}_{w|s,q}^{l} \in \mathbb{R}^d$ denotes embedding of entity w in the l-th layer. $\mathrm{Agg}(\cdot)$ represents sum aggregation, and $\mathrm{Msg}(\cdot)$ is message propagation function with DistMult(Yang et al. 2015). $\Psi_q(\boldsymbol{x}_r)$ denotes updating the embedding of relation r.

Historical Path Preservation

To preserve old path semantics, we design a historical path semantics preservation module in which Ψ_q, defined as a relation update function in Equation (1), enables the model to effectively distinguish path semantics guided by different query relations. This reduces interference among paths from different queries while preserving prior path semantics through regularization. We first concatenate the query relation $\boldsymbol{x}_q$ with each relation $\boldsymbol{x}_r$ observed in the i-th snapshot to construct $\Psi_q(\boldsymbol{x}_r)$. Then, the query relation is fused with all relations using a query-specific weight matrix $W_q^i \in \mathbb{R}^{2d \times d}$ and bias $b_q^i \in \mathbb{R}^d$. The specific process is as follows:

$$\Psi_q(\boldsymbol{x}_r) = \left\{\left(W_q^i[\boldsymbol{x}_r \parallel \boldsymbol{x}_q]\right) + b_q^i \mid r \in \mathcal{R}^i\right\} \tag{2}$$

where W_q^i and b_q^i are uniquely defined for each query relation. When continuing to train on the new snapshot $i (i > 1)$, relation embeddings $\boldsymbol{x} \in \mathbb{R}^{|\mathcal{R}^i| \times d}$ will be expanded:

$$\boldsymbol{x}^i = \left\{con\left(\boldsymbol{x}^{i-1}, init\left(\boldsymbol{x}_r\right)\right) \mid r \in \mathcal{R}_\Delta^i\right\} \tag{3}$$

where $\mathcal{R}_\Delta^i$ is defined as the set of newly added relations in the i-th snapshot. $init$ represents initialization and con denotes concatenating new relation embeddings in new snapshot.

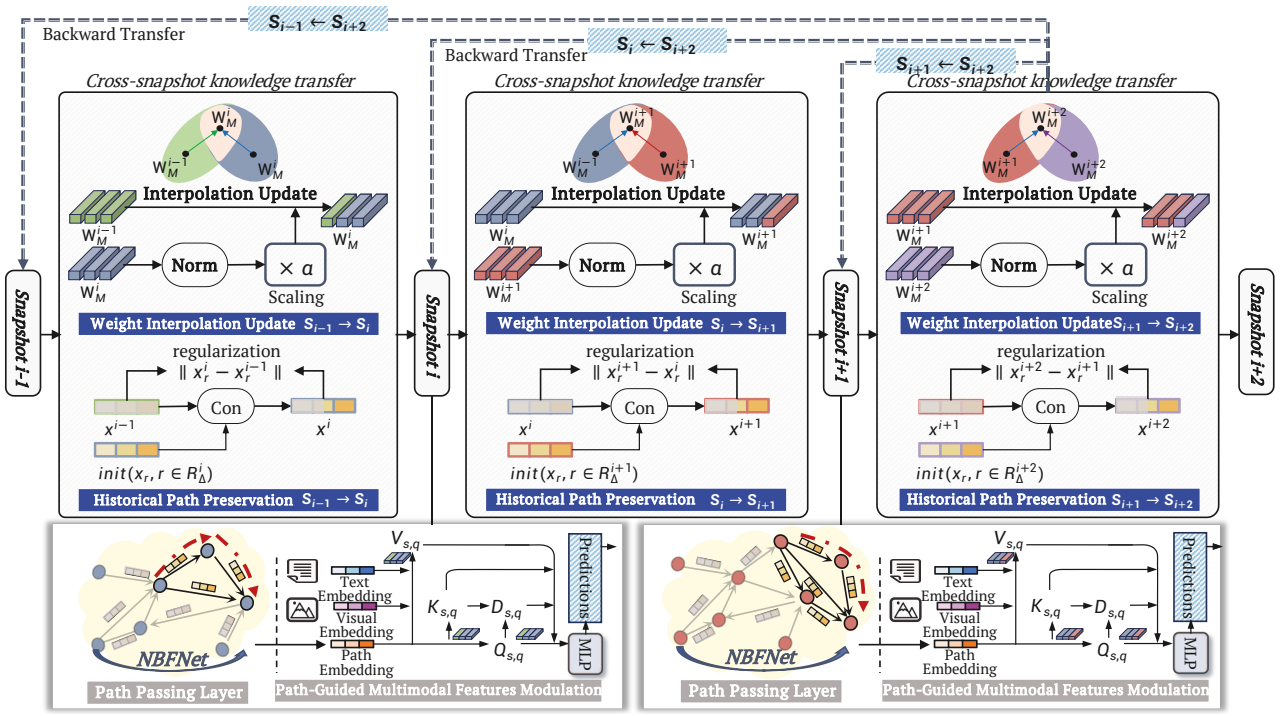

Figure 3: The overall framework of multimodal continual representation learning framework (MoFot).

Finally, we use regularization constraints on the relation embeddings to align them with previous optimization directions, thereby preserving old path semantics.

$$\Delta P_1 = \sum_{r \in \mathcal{R}^{i-1}} \alpha \left\| \boldsymbol{x}_r^i - \boldsymbol{x}_r^{i-1} \right\|_2^2 \tag{4}$$

where α denotes the hyperparameter.

Path-Guided Multimodal Features Modulation

Due to differences in data sources and feature extraction, features from different modalities converge at varying rates, leading to inconsistent forgetting rates across modalities(Chen et al. 2024). Inspired by the fact that textual and visual modalities rely on specific structural semantics to effectively align with the semantic space of the target KG (Liu et al. 2024c)(Liang et al. 2023), we propose a novel multimodal feature modulation method that uses the path semantics capable of preserving old structural knowledge as a central guide for multimodal fusion through an attention mechanism. This encourages other modality features to align with path semantics across snapshots, promoting consistent retention of multimodal knowledge and deeper interactions between modalities. Specifically, a learnable weight $W_f^i \in \mathbb{R}^{3d \times d}$ is used to fuse the multimodal features with the path semantics:

$$\boldsymbol{Z}_F^i = \left\{ W_f \left([\boldsymbol{Z}_V^i \,||\, \boldsymbol{Z}_T^i \,||\, \boldsymbol{Z}_{s,q}^i] \right) \right\} \tag{5}$$

where $\boldsymbol{Z}_V \in \mathbb{R}^{|\mathcal{E}^i| \times d}, \boldsymbol{Z}_T \in \mathbb{R}^{|\mathcal{E}^i| \times d}$ represent the visual and textual embeddings obtained by transforming the raw visual and textual features using $W_V^i \in \mathbb{R}^{d \times d}$ and $W_T^i \in \mathbb{R}^{d \times d}$, respectively. Then, we project the path representations $\boldsymbol{Z}_{s,q}^i = (\boldsymbol{z}_{o|s,q}^L)_{o \in \mathcal{E}^i}$ into shared query and key matrices using learnable weights: $\boldsymbol{Q}_{s,q} = W_1^i \boldsymbol{Z}_{s,q}^i$ and $\boldsymbol{K}_{s,q} = W_2^i \boldsymbol{Z}_{sq}^i$. Subsequently, we normalize the query and key matrices and define the value matrix as $\boldsymbol{V}_{s,q} = \boldsymbol{Z}_F^i$. Finally, we adopt the attention mechanism from (Liu et al. 2024b) to encourage fused embeddings to align with path semantics, while ensuring linear time complexity with related to the number of entities and triples, enabling scalability to large-scale knowledge graphs. The detailed process is as follows:

$$\boldsymbol{D}_{s,q} = \operatorname{diag}\left(\mathbf{1} + \frac{\boldsymbol{Q}_{s,q} \left(\boldsymbol{K}_{s,q}^T \mathbf{1} \right) + |\mathcal{E}^i|}{|\mathcal{E}^i|} \right) \tag{6}$$

$$\tilde{\boldsymbol{Z}}_{s,q}^i = \boldsymbol{D}_{s,q}^{-1} \left[\boldsymbol{V}_{s,q} + \frac{\mathbf{1}^T \boldsymbol{V}_{s,q} + \boldsymbol{Q}_{s,q} \left(\boldsymbol{K}_{s,q}^T \boldsymbol{V}_{s,q} \right)}{|\mathcal{E}^i|} \right] \tag{7}$$

where $\mathcal{E}^i$ represents the set of entities in snapshot i. $|\mathcal{E}^i|$ is the number of entities. $\operatorname{diag}(\cdot)$ denotes the diagonal operator. Then, $\tilde{\boldsymbol{Z}}_{s,q}^i \in \mathbb{R}^{|\mathcal{E}^i| \times d}$ is the fused entity embeddings obtained by corresponding query $(s, q, ?)$.

Weight Interpolation Update

In multimodal continual learning, the model needs to continuously update the weights of each modality to adapt to new knowledge. However, such weight updates may disrupt

previously learned representations and increase the risk of catastrophic forgetting(Cui et al. 2023). Moreover, different modalities have independent weights with varying update rates, leading to inconsistent forgetting across modalities and further exacerbating the forgetting problem. To address this, we propose a weight interpolation update method, which aims to mitigate the interference of weight updates on learned knowledge and achieve consistent retention of multimodal knowledge through unified and fine-grained modulation of the direction and magnitude of weights across modalities.

First, the learnable weights of each modality are used to update their corresponding knowledge representations during multimodal continual learning. Specifically, as follows:

$$\boldsymbol{y}^i_{out|M} = W^i_M \boldsymbol{y}^i_{in|M} \tag{8}$$

where W^i_M, $M \in \{V, T, S\}$ denotes the weights used for different modalities, including the weight W^i_q, as well as the modality-specific weights W^i_V and W^i_T for visual and textual features, respectively. The input features $y^i_{in|M}$ consist of the relation embeddings $\boldsymbol{x}_r$, $r \in \mathcal{R}^i$, and the entity embeddings of the visual and textual modalities $\boldsymbol{z}_{o|V}$ and $\boldsymbol{z}_{o|T}$, $o \in \mathcal{E}^i$.

Then, our model learns new weights for the new snapshot while freezing the original weights to preserve old knowledge. We apply a weighted interpolation method to combine the new and old weights, producing the final weights.

$$W^i_M = \left(1 - \frac{1}{i}\right) W^{i-1}_M + \frac{1}{i}\mathcal{H}(\Delta W^i_M) \tag{9}$$

where $i (i > 1)$ denotes the i-th snapshot, and ΔW^i_M represents a set of re-initialized weights in the new snapshot. We perform weighted interpolation between old and new weights to guide their optimization toward a direction that balances both new and old knowledge(Zhou et al. 2025). $\mathcal{H}(\cdot)$ adjusts the magnitude of the weights.

After adjusting the directions, we apply a uniform scaling to all new weights of all modalities. Specifically, we first normalize the weights of each modality, then apply a unified scaling factor σ to scale all the normalized weights equally:

$$\mathcal{H}(\Delta W^i_M) = \sigma(\overline{\Delta W^i_M}) \tag{10}$$

where σ is a learnable parameter and $\overline{\Delta W^i_M}$ represents the normalized weights newly initialized in new snapshot. Since the weights need to be reinitialized and fused in each new snapshot, we apply the weight interpolation update mechanism only to the independent weights corresponding to each modality to reduce training overhead. For the multimodal shared W^i_f, W^i_1 and W^i_2 components, we adopt a simplified regularization method to achieve efficient balancing i.e., $\Delta P_2 = \alpha \left\| W^i_j - W^{i-1}_j \right\|^2_2$, where $j \in \{f, 1, 2\}$. By jointly modulating features and weights, we address catastrophic forgetting caused by parameter updates and inconsistent forgetting rates across modalities.

Model Training

During continual learning, we train training set $\mathcal{T}^i$ from the newly added triples $\Delta\mathcal{F}^i$ in snapshot i. After obtaining the $\tilde{\boldsymbol{Z}}^i_{s,q}$, the scoring function f (i.e., multi-layer MLP) maps the d-dimensional embeddings to one-dimensional logits $f(s, q, o)$. The loss is defined as follows:

$$\mathcal{L} = -\log f(s, q, o) - \frac{1}{n}\sum_{j=1}^{n} \log\left(1 - f\left(s'_j, q, o'_j\right)\right) + \Delta P_1 + \Delta P_2 \tag{11}$$

where n is a hyperparameter denoting the number of negative samples for each positive sample. (s, q, o) and (s'_j, q, o'_j) are the positive sample and the j-th negative sample, respectively. The negative samples are generated by corrupting either the head entity s or the tail entity o of the positive sample.

Experiments

Experiment Settings

Datasets: To fully evaluate the performance of the related models, we created five MMKG datasets by integrating textual and visual modality information into five benchmark datasets: ENTITY (Cui et al. 2023), RELATION(Cui et al. 2023), FACT(Cui et al. 2023), HYBRID(Cui et al. 2023), and GraphEqual(Liu et al. 2024a), where growth is driven by entities, relations, triples, or mixed strategies. Each dataset consists of five time steps. Following IMF(Li et al. 2023), we use encoders BERT(Devlin et al. 2019) and VGG16(Simonyan and Zisserman 2014) to extract textual and visual features.

Metrics for Link Prediction: We used four metrics to evaluate our model's performance on the link prediction task: MRR and Hits@1, Hits@3, Hits@10. MRR is the mean reciprocal rank, while Hits@K measures the proportion of correct answers ranked in the top K (K = 1, 3, 10).

Metrics for Knowledge Retention and Transfer: To evaluate the model's ability to retain and transfer knowledge, we adopted the backward transfer (BWT) and forward transfer(FWT) metrics(Lopez-Paz and Ranzato 2017). The negative BWT indicate that the model forgets prior knowledge, while the positive BWT indicates effective retention or even enhancement of prior knowledge. FWT indicates whether the model can transfer learned knowledge to new tasks.

Baselines: We compare our model with several baseline models, including: (i) unimodal continual learning models: PNN(Rusu et al. 2016), EWC(Kirkpatrick et al. 2017), SI(Zenke, Poole, and Ganguli 2017), LKGE(Cui et al. 2023) and IncDE(Liu et al. 2024a); (ii) static MMKG models: MoCi(Yang et al. 2024) and IMF(Li et al. 2023); (iii) inductive multimodal model: IndMKG(Yang et al. 2025); and (iv) continual learning model using the bilinear pooling fusion method of IMF: M-LKGE and M-IncDE. PNN, EWC, SI, LKGE, IncDE, and our model are all capable of handling continual tasks. For static multimodal models, we continuously fine-tune the parameters on the newly added triples in each snapshot. The main results are average metrics across all snapshot test sets, evaluated using the model trained or fine-tuned on the last snapshot. Inductive MMKG models are trained on the first snapshot and evaluated across all test sets, with the average result as the final result. After training the model on $\mathcal{G}^i$, we consider all entities appearing in $\mathcal{G}^1$ to $\mathcal{G}^i$ as candidates for ranking.

Model	M-ENTITY			M-GraphEqual			M-FACT			M-HYBRID			M-RELATION		
	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10
PNN	0.229	0.130	0.425	0.214	0.120	0.407	0.157	0.084	0.290	0.185	0.101	0.349	0.167	0.096	0.305
SI	0.154	0.072	0.311	0.176	0.092	0.350	0.172	0.088	0.343	0.111	0.049	0.229	0.113	0.055	0.224
EWC	0.229	0.130	0.423	0.209	0.115	0.401	0.201	0.113	0.382	0.186	0.102	0.350	0.165	0.093	0.306
LKGE	0.234	0.136	0.425	0.215	0.120	0.406	0.210	0.122	0.387	0.207	0.121	0.379	0.192	0.106	0.366
IncDE	0.253	0.151	0.448	0.234	0.134	0.432	0.216	0.126	0.394	0.223	0.131	0.399	0.199	0.110	0.368
LKGE-M	0.213	0.123	0.387	0.207	0.115	0.393	0.197	0.116	0.362	0.185	0.105	0.344	0.153	0.080	0.301
IncDE-M	0.220	0.127	0.396	0.218	0.120	0.411	0.169	0.098	0.304	0.168	0.095	0.302	0.134	0.058	0.277
IMF	0.197	0.140	0.315	0.176	0.114	0.307	0.177	0.115	0.308	0.119	0.080	0.197	0.062	0.042	0.099
MoCi	0.221	0.157	0.345	0.196	0.124	0.344	0.187	0.123	0.319	0.132	0.093	0.210	0.066	0.046	0.103
IndMKG	0.372	0.269	0.574	0.279	0.197	0.446	0.262	0.182	0.426	0.314	0.217	0.511	0.323	0.221	**0.525**
MoFot	**0.413**	**0.308**	**0.615**	**0.362**	**0.253**	**0.581**	**0.309**	**0.212**	**0.505**	**0.353**	**0.255**	**0.545**	**0.338**	**0.245**	0.520

Table 1: Performance comparison of our model and baseline models on continual MMKG link prediction tasks.

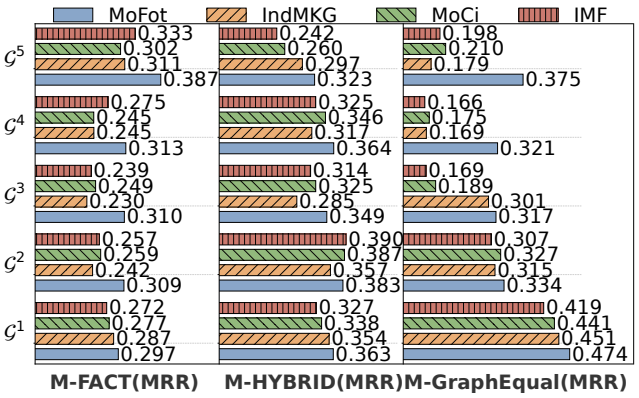

Figure 4: The performance on individual snapshots.

Main Results

Table 1 presents the performance comparison of each model. Compared to MMKG models, MoFot demonstrates significant improvements. This is primarily because existing MMKG models struggle to effectively handle the dynamically growing MMKG. Although they can adapt to new data through fine-tuning, they suffer from catastrophic forgetting. Moreover, even when comparing performance on individual snapshots (IMF and MoCi fine-tune on $\mathcal{G}^i$ and test only on snapshot i; MoFot and IndMKG use their original reported results without averaging), our model still outperforms the others, as shown in Figure 4. IndMKG generalizes well by capturing transferable relational structures, achieving strong performance on the relation dataset. However, the performance of IndMKG inevitably declines when dealing with scenarios where data continues to grow. Compared to unimodal continual learning model, MoFot achieves significant improvements. This is attributed not only to the integration of multimodal information but also to its multimodal collaborative modulation mechanism for multimodal continual learning. MoFot extracts path semantics that can preserve old structural knowledge and uses it to promote consistent retention of multimodal knowledge. Additionally, MoFot further enhances knowledge retention by adjusting the weights.

To evaluate the performance evolution of MoFot, we conducted experiments snapshot i and all previous snapshots using the model $\mathcal{M}^i$ trained on $\mathcal{G}^i$. As shown in Figure 5 in the M-ENTITY, M-GraphEqual and M-HYBRID datasets. The inference performance of MoFot showed little decline and even improved in some snapshots. Overall, MoFot not only demonstrates excellent retention of old knowledge, but also possesses the ability to update it using new knowledge.

Knowledge Retention and Transfer Capability

To evaluate MoFot's ability for knowledge transfer and retention, we present the FWT and BWT scores, as shown in Table 2. Since PNN and indMKG do not update their old parameters during continual learning, we do not report their BWT scores. MoFot achieves higher FWT scores compared to other models. Its BWT score also outperforms all the baseline models, which demonstrates that it effectively retains the knowledge it has learned. Notably, MoFot can achieve positive BWT scores, which indicates that it enhances old knowledge through continual learning of new knowledge.

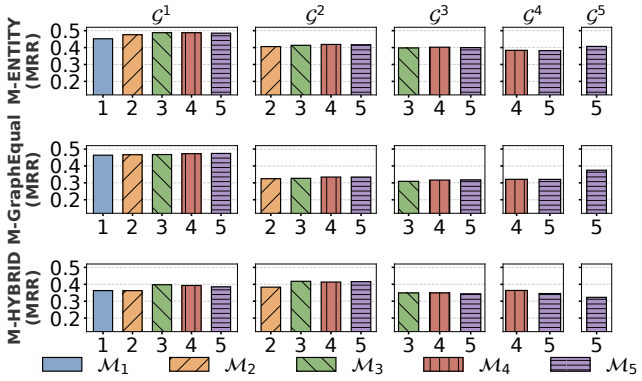

Figure 5: The MRR performance trend of MoFot, $\mathcal{M}^i$ denotes the model obtained through continual training up to snapshot i, evaluated on the test sets from snapshots 1 to i

Model	M-ENTITY		M-GraphEqual		M-HYBRID	
	BWT	FWT	BWT	FWT	BWT	FWT
SI	-0.114	0.044	-0.041	0.104	-0.112	0.037
EWC	-0.001	0.046	-0.005	0.108	-0.008	0.038
LKGE	-0.006	0.097	-0.004	0.134	-0.006	0.109
IncDE	-0.017	0.043	-0.003	0.107	-0.015	0.043
LKGE-M	-0.028	0.083	-0.007	0.128	-0.041	0.101
IncDE-M	-0.036	0.079	-0.004	0.130	-0.068	0.044
IMF	-0.247	NA	-0.095	NA	-0.232	NA
MoCi	-0.239	NA	-0.091	NA	-0.229	NA
MoFot	**0.011**	**0.387**	**0.007**	**0.330**	**0.008**	**0.164**

Table 2: The BWT and FWT performance.

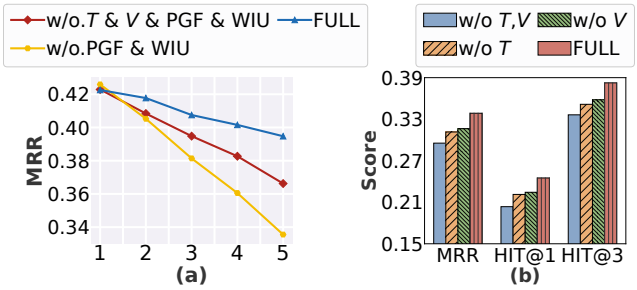

Figure 6: (a) compares the MRR of MoFot in unimodal and multimodal settings, with and without the PGF and WIU modules. (b) shows the impact of each modality on MoFot.

Ablation Study

Modality: We conduct modality ablation on the M-RELATION dataset to analyze the contribution of multimodal information and validate the effectiveness of Path-Guided Multimodal Feature Modulation(PGF) and Weight Interpolation Update (WIU) module. Figure 6(a) shows the MRR changes of different MoFot variants on *Snapshot 1* after training on each subsequent snapshot. After removing these modules, the unimodal MoFot exhibits significantly less forgetting than multimodal MoFot, highlighting that inconsistent forgetting rates across modalities exacerbate catastrophic forgetting. With the PGF and WIU modules, MoFot achieves significantly better performance and slower forgetting. Figure 6(b) illustrates the impact of sequentially removing the textual modality (T) and visual modality (V) on model performance, confirming the contribution of each modality.

Module: We conducted ablation studies to validate the effectiveness of each module. Table 3 shows the link prediction performance of each variant, while Table 4 presents their corresponding BWT scores. We constructed three variants of MoFot by removing the Historical Path Preservation(HPP), PGF, and WIU. (i) w/o HPP leads to a significant drop in the model's overall performance and BWT score. Without the ability to distinguish path information guided by different query relations and retain prior structural knowledge, maked the model unable to balance new and old structural knowledge. (ii) Both w/o PGF and w/o WIU weaken the model's ability to mitigate catastrophic forgetting caused by inconsistent forgetting rate of modalities and parameter updates.

Model	M-FACT		M-HYBRID		M-RELATION	
	MRR	Hits@1	MRR	Hits@1	MRR	Hits@1
w/o HPP	0.239	0.149	0.254	0.159	0.148	0.080
w/o PGF	0.299	0.204	0.312	0.220	0.302	0.214
w/o WIU	0.302	0.207	0.333	0.237	0.315	0.221
MoFot	**0.309**	**0.212**	**0.353**	**0.255**	**0.338**	**0.245**

Table 3: Ablation study of each module.

Model	M-FACT BWT	M-HYBRID BWT	M-RELATION BWT
w/o HPP	-0.0826	-0.0806	-0.1292
w/o PGF	-0.0223	-0.0315	-0.0337
w/o WIU	-0.0256	-0.0151	-0.0386
MoFot	**-0.0180**	**0.0076**	**-0.0121**

Table 4: Ablation study of module impact on BWT.

Learning Efficiency

Figure 7 shows the training time on the M-ENTITY dataset, with all models trained under the same conditions. The retraining of IMF and MoCi is the most time-consuming, as they discard previously learned knowledge, leading to low efficiency. Even compared to the fine-tuning of IMF and MoCi, our model remains the most efficient while also achieving better long-term prediction performance.

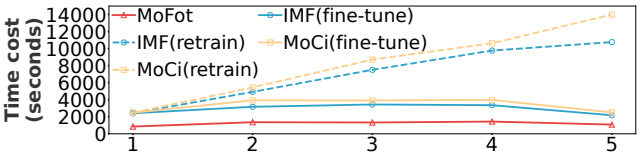

Figure 7: Time efficiency comparison.

Conclusion

Existing MMKG models often struggle to maintain long-term predictive accuracy when adapting to new knowledge, primarily due to catastrophic forgetting. To address this issue, we propose MoFot, the first multimodal continual representation learning framework. MoFot tackles the catastrophic forgetting caused by inconsistent forgetting rates across modalities and parameter updates, through two key mechanisms: multimodal weight modulation and multimodal feature modulation, thereby effectively mitigating performance degradation resulting from catastrophic forgetting. MoFot provides an effective solution for the continual expansion of MMKG in real-world scenarios.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62302131, the Postdoctoral Fellowship Program (Grade B) of the China Postdoctoral Science Foundation under Grant GZB20250405, and the Fellowship Program (Grade C) of the China Postdoctoral Science Foundation under Grant GZC20251050.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 382–398. Springer.

Chen, X.; Zhang, J.; Wang, X.; Zhang, N.; Wu, T.; Wang, Y.; Wang, Y.; and Chen, H. 2024. Continual Multimodal Knowledge Graph Construction. In *IJCAI*.

Chen, X.; Zhang, N.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; and Chen, H. 2022. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *ACM SIGIR*, 904–915.

Cui, Y.; Wang, Y.; Sun, Z.; Liu, W.; Jiang, Y.; Han, K.; and Hu, W. 2023. Lifelong embedding learning and transfer for growing knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4217–4224.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, 4171–4186.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Li, X.; Zhao, X.; Xu, J.; Zhang, Y.; and Xing, C. 2023. IMF: interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023*.

Liang, K.; Zhou, S.; Liu, Y.; and Meng, L. 2023. Structure guided multi-modal pre-trained transformer for knowledge graph reasoning. *arXiv preprint arXiv:2307.03591*.

Liu, J.; Ke, W.; Wang, P.; Shang, Z.; Gao, J.; Li, G.; Ji, K.; and Liu, Y. 2024a. Towards continual knowledge graph embedding via incremental distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8759–8768.

Liu, J.; Mao, Q.; Jiang, W.; and Li, J. 2024b. KnowFormer: Revisiting Transformers for Knowledge Graph Reasoning. 31669–31690.

Liu, K.; Zhao, F.; Yang, Y.; and Xu, G. 2024c. Dysarl: Dynamic structure-aware representation learning for multimodal knowledge graph reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8247–8256.

Lomonaco, V.; and Maltoni, D. 2017. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on robot learning*, 17–26. PMLR.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

Mousselly-Sergieh, H.; Botschen, T.; Gurevych, I.; and Roth, S. 2018. A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the seventh joint conference on lexical and computational semantics*, 225–234.

Pezeshkpour, P.; Chen, L.; and Singh, S. 2018. Embedding Multimodal Relational Data for Knowledge Base Completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3208–3218.

Rusu, A. A.; Rabinowitz, N. C.; Desjardins, G.; Soyer, H.; Kirkpatrick, J.; Kavukcuoglu, K.; Pascanu, R.; and Hadsell, R. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wang, H.; Xiong, W.; Yu, M.; Guo, X.; Chang, S.; and Wang, W. Y. 2019a. Sentence embedding alignment for lifelong relation extraction.

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 *Proceedings of the 31st ACM international conference on multimedia*, 2391–2399.

Wang, Z.; Li, L.; Li, Q.; and Zeng, D. 2019b. Multimodal data enhanced representation learning for knowledge graphs. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Wei, Z.; Wang, K.; Li, F.; and Ma, Y. 2024. M3KGR: A Momentum Contrastive Multi-Modal Knowledge Graph Learning Framework for Recommendation. *Information Sciences*, 120812.

Xie, R.; Liu, Z.; Luan, H.; and Sun, M. 2017. Image-embodied Knowledge Representation Learning. In *The 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*.

Yang, B.; Yih, W.-t.; He, X.; Gao, J.; and Deng, L. 2015. Embedding entities and relations for learning and inference in knowledge bases. *International Conference on Learning Representations*.

Yang, J.; Yang, S.; Gao, Y.; Yang, J.; and Yang, L. T. 2024. Multimodal contextual interactions of entities: A modality circular fusion approach for link prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8374–8382.

Yang, S.; Li, G.; and Yu, Y. 2019. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4145–4154.

Yang, S.; Yang, J.; Jiang, X.; Gao, Y.; Yang, L. T.; Luo, R.; and Yang, J. 2025. Towards Multimodal Inductive Learning: Adaptively Embedding MMKG via Prototypes. In *Proceedings of the ACM on Web Conference 2025*, 109–118.

Zenke, F.; Poole, B.; and Ganguli, S. 2017. Continual learning through synaptic intelligence. In *International conference on machine learning*, 3987–3995. PMLR.

Zhang, Y.; Chen, Z.; Guo, L.; Xu, Y.; Hu, B.; Liu, Z.; Zhang, W.; and Chen, H. 2024a. NativE: Multi-modal Knowledge Graph Completion in the Wild. In *SIGIR*, 91–101. ACM.

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

Zhang, Y.; and Zhang, W. 2022. Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling. *KDD 2022 Undergraduate Consortium*.

Zhou, D.-W.; Zhang, Y.; Wang, Y.; Ning, J.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2025. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhu, Z.; Zhang, Z.; Xhonneux, L.-P.; and Tang, J. 2021. Neural bellman-ford networks: A general graph neural network framework for link prediction. *Advances in Neural Information Processing Systems*, 34: 29476–29490.