Expert Systems With Applications 268 (2025) 126308

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Multimodal fusion framework based on knowledge graph for personalized recommendation

Jingjing Wang [a], Haoran Xie [b,*], Siyu Zhang [a], S. Joe Qin [b], Xiaohui Tao [c], Fu Lee Wang [d], Xiaoliang Xu [a]

[a] Hangzhou Dianzi University, 1158 2nd Ave, Qiantang district, Hangzhou, 310005, Zhejiang, China
[b] Lingnan University, 8 Castle Peak Road, Tuen Mun, New Territories, 999077, Hong Kong Special Administrative Region
[c] University of Southern Queensland, Springfield, 4300, Queensland, Australia
[d] Hong Kong Metropolitan University, 30 Good Shepherd Street, Ho Man Tin, Kowloon, 999077, Hong Kong Special Administrative Region

ARTICLE INFO

Keywords:
Knowledge graphs
Multimodal fusion framework
Recommender system

ABSTRACT

Knowledge Graphs (KGs), which contain a wealth of knowledge, have been commonly employed in recommendation systems as a valuable knowledge-driven tool for supporting high-quality representations. To further enhance the model's ability to understand the real world, Multimodal Knowledge Graphs (MKGs) are proposed to extract rich knowledge and facts among objects from text and visual content. However, existing MKG-based methods primarily focus on the reasoning relationships between entities by utilizing multimodal information as auxiliary data in the KG while overlooking the interactions between modalities. In this paper, we propose a Multimodal fusion framework based on Knowledge Graph for personalized Recommendation (Multi-KG4Rec) to address these limitations. Specifically, we systematically analyze the shortcomings of existing multimodal graph construction methods. To this end, we propose a modal fusion module to extract the user modal preference at a fine-grained level. Furthermore, we conduct extensive experiments on two real-world datasets from different domains to evaluate the performance of our model, and the results demonstrate the efficiency of the Multi-KG4Rec.

1. Introduction

Recently, recommender systems integrated with knowledge engineering have become one of the hottest research directions, primarily because Knowledge Graphs (KGs) provide comprehensive auxiliary data that supports effective recommendation results (Fan, Zhong, Zeng, & Ge, 2022; Yang, Huang, Xia, & Li, 2022). Traditional graph neural network-based methods aim to mine the collaborative filtering signals based on the bipartite graph, which is constructed from a user's historical interactions. Compared to the bipartite graph (which only includes user and item nodes), KGs are more effective in distilling the attribution-based collaborative signals, for example, users watching different movies that share the same director. However, in most existing KG-based methods, attribution information is represented as pure symbols. This limitation reduces the model's ability to understand the real-world scenarios, which are typically characterized by rich images and textual descriptions. To address this, Multimodal Knowledge Graphs (MKGs) have been proposed as a key step toward achieving human-level machine intelligence. While previous studies on MKGs (Liu, Li and Tian, 2022; Sun et al., 2020; Zhang, Yuan, Lian, Xie, & Ma, 2016) have shown promising improvements over single modal methods, these approaches still face two significant limitations.

- **Lack of a unified MKG architecture.** Existing MKG methods can be categorized into two classes, as illustrated in Fig. 1: feature-based methods and entity-based methods. Feature-based methods treat multimodal information as auxiliary data for the entity, such as entity e_2 shown in Fig. 1. While feature-based methods are efficient at enriching the entity representations, they tend to overlook interactions between different modalities. Additionally, feature-based methods impose relatively strict constraints on the collected multimodal entities, requiring the MKG to be complete. Entity-based methods consider multimodal information as newly added supplementary nodes to the original KG. Unfortunately, these entities are limited to attribute entities and do not include the item entities. This limitation arises because these newly added nodes for items are extremely sparse, as no two movies share the

* Corresponding author.
E-mail addresses: wangj3573@163.com (J. Wang), hrxie@ln.edu.hk (H. Xie), joeqin@ln.edu.hk (S.J. Qin), Xiaohui.Tao@unisq.edu.au (X. Tao), pwang@hkmu.edu.hk (F.L. Wang), xxl@hdu.edu.cn (X. Xu).

https://doi.org/10.1016/j.eswa.2024.126308
Received 3 September 2024; Received in revised form 13 December 2024; Accepted 24 December 2024
Available online 1 January 2025

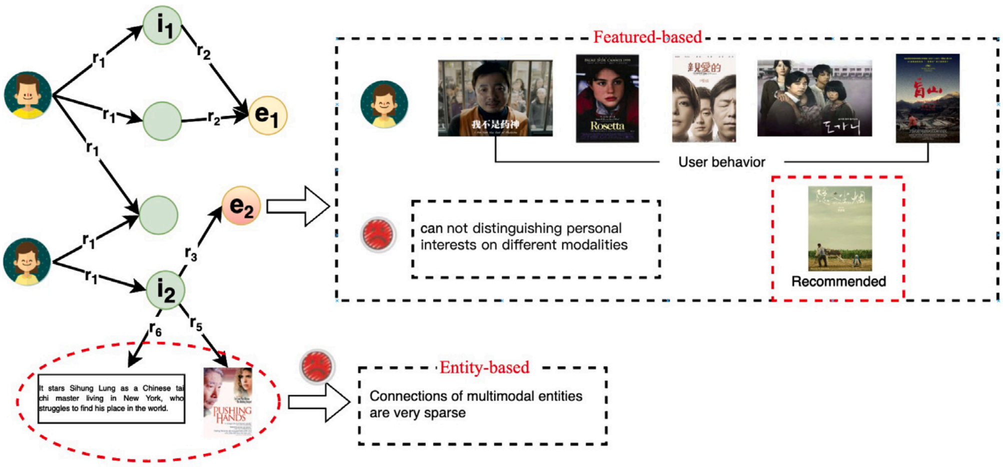

i1
r1
r2
e1
r1
r2
r1
e2
r1
r3
r1
i2
r6
r5
Featured-based
我不是药神
Rosetta
도가니
User behavior
can not distinguishing personal interests on different modalities
Recommended
It stars Sihung Lung as a Chinese tai chi master living in New York, who struggles to find his place in the world.
PUSHING HANDS
Entity-based
Connections of multimodal entities are very sparse

Fig. 1. A toy example illustrating the feature-based and entity-based methods.

same poster or text description, as illustrated by the node i_2 in Fig. 1. Consequently, existing entity-based methods struggle to extract sufficient multimodal signals from item nodes.

- **Multimodal fusion.** Due to the nature of feature-based construction, most existing methods adopt connections or weighted sums to fuse multimodal information. This approach makes it challenging to effectively leverage multimodal information in the later stages of the model. For example, in Fig. 1, a user has watched various films differing in directors, actors, and genres. However, all posters share a similar style and relate to societal issues. Feature-based methods make it difficult to extract such correlations within the visual modality because these films tell different stories accompanied by varying text descriptions. The explicit relationships within the visual modality are weakened by the influence of textual information. Entity-based methods suffered from similar issues. Due to the lack of direct connections between item and attribute entities, entity-based methods cannot capture these relationships through path transmission.

To address the aforementioned issues, we propose a unified Multimodal fusion framework based on the Knowledge Graph for personalized Recommendation (named “Multi-KG4Rec”). From the perspective of KG architecture, Multi-KG4Rec divides the multimodal graph into several single modal graphs. Each entity within a modality is represented by its modality feature, enhancing the model’s ability to extract the user’s personalized modality preferences at a coarse-grained level and avoiding issues related to node sparsity. From the perspective of multimodal fusion, we first employ pre-trained models to align the corresponding content of images and texts. To leverage the advantages of large language models (LLMs) for feature extraction, we utilize a pre-trained LLM to generate the initial multimodal features. However, since such generative models (e.g., transformer decoders, in-context learning) are incompatible with graph-structured data, we design a graph neural network to align the multimodal features with the graph structure information. Following this, we adopt a cross multi-head attention module between a text transformer and a visual transformer to fuse modality information at a fine-grained level. Finally, we propagate the fused features to neighboring nodes to generate the final representation. In summary, the contributions of this paper are as follows:

- We present a unified multimodal architecture that overcomes the strict limitations of feature-based methods while addressing node sparsity issues in entity-based methods.
- We employ a pre-trained LLM to generate initial multimodal features and integrate them with graph-structured information using a graph neural network. Subsequently, we propose a multimodal fusion module to extract users’ personalized multimodal preferences at a fine-grained level.
- Extensive experiments on two real-world datasets demonstrate the effectiveness of Multi-KG4Rec.

2. Related work

In this section, we first introduce KG-based methods, followed by MKG-based methods, and finally, commonly used modal fusion methods.

2.1. KG-based methods

KG-based methods construct a heterogeneous graph with user, item, and item attributes, such as a movie’s genre and actors, and then propagate the corresponding relationships within the graph to generate the item and user representations. In text-based recommendation tasks, identifying entities is a major challenge in graph construction. For instance Li, Chiu, Feng and Wang (2020) adopted meta-learning for named entity recognition. Another challenge lies in labeling entity attributes based on text sequence inputs, requiring the detection of clear boundaries (Li et al., 2021; Li, Shang and Chen, 2020). In this work, we follow the entities identified by previous studies. Most existing KG-based methods focus on reasoning relationships between entities. For example, KGCN (Wang, Zhao, Xie, Li and Guo, 2019) incorporates GCN and KG methods to learn relationships between entities. Wang, Zhang et al. (2019) summarized existing KG-based methods as relying heavily on manual feature engineering, for example, meta-path-based methods (Hu, Shi, Zhao, & Yu, 2018; Zhao, Yao, Li, Song, & Lee, 2017), which depend on manual path definition. To address this, they constructed a user-specific weighted graph via a trainable function and incorporated a label smoothing mechanism to provide an inductive bias based on the assumption that similar users have similar or related labels. The KGAT (Wang, He, Cao, Liu and Chua, 2019) model innovatively proposed a collaborative KG method, propagating features on the collaborative KG through a GCN layer to encode high-order relationships between users and items. For document-level relation extraction, Li, Wang, Zhang and Zhang (2023) reviewed popular DocRE datasets and highlighted that entities in the text modality are vulnerable to entity mention attacks.

The above methods focus on extracting the wealth of knowledge based on KG but ignore rich knowledge and facts among objects from text and visual information, limiting the model’s representation ability.

2.2. MKG-based methods

In feature-based MKG-based methods, the model CKE (Zhang et al., 2016) divided the MKG into the bipartite graph, textual content, and visual content to extract the items' semantic features. Bipartite graphs are used to learn the structural representations based on TransR, while multimodal content is extracted via denoising autoencoders. DKN (Wang, Zhang, Xie, & Guo, 2018) proposed a convolutional neural network framework to integrate high-order relational reasoning tasks with text semantics generating. Compared to CKE, DKN explored high-order relationships under the textual modalities while not addressing other modals. CMCKG (Cao et al., 2022) is a representative method of feature-based methods. The original KG is utilized to explore structural representations, while textual descriptions are converted into newly added nodes within the KG to learn textual representations. Contrastive learning is then applied to enhance consistency between these two representations, resulting in improved model performance. MKGAT (Sun et al., 2020) is an entity-based method where only attribute entities contain multimodal information, which is transferred as newly added nodes in the KG. MMKGV (Liu, Li et al., 2022) integrated multimodality information as relationship triplets within a knowledge graph.

2.3. Multimodal fusion

Since the expressive ability of a single modality is limited, exploring multimodal fusion is essential for enhancing the model's understanding of the real world. We categorize existing multimodal fusion methods into three categories: coarse-grained attention, fine-grained, and combined attention.

Coarse-grained attention focuses on capturing modality correlation at a coarse-grained level. For example, DUALGRAPH (Li, Feng and Chiu, 2023) proposed a novel few-shot relation extraction method using a dual graph neural network to address distribution differences across domains. UVCAN (Liu, Chen, Liu and Hu, 2019) employed a co-attention mechanism to extract multimodal information from both user- and microvideo perspectives. Similar to UVCAN, MCPTR (Liu, Ma, Schubert, Ouyang and Xiong, 2022) explored cross-modal information from both user and item perspectives via the self-attention mechanism. CMBF (Chen, Lu, Wang, & Yang, 2021) adopted a cross-attention mechanism to learn the modality-level features.

Fine-grained attention methods focus on detailed correlations between modalities. POG (Chen et al., 2019) proposed an encoder–decoder model that connects user personalization with outfit recommendations, incorporating a masked item prediction task based on a self-attention mechanism. NOR (Lin et al., 2019) employed a transformer with fine-grained self-attention structures for clothing recommendations. EFRM (Hou et al., 2019) introduced an interpretable personalized fashion recommendation model based on semantic attributes, employing an attention-based fine-grained preference modeling mechanism to align user preferences with specific clothing attributes. MM-Rec (Wu, Wu, Qi, & Huang, 2021) introduced a candidate-aware attention network to evaluate cross-modal relevance between clicked and candidate items.

Incorporating fine-grained attention often increases the model's computational complexity, which may reduce the real-time performance of the recommendation system. Therefore, it is crucial to strike a balance between computational complexity and model performance. NOVA (Liu et al., 2021) proposed an effective non-invasive self-attention mechanism under the BERT framework, leveraging side information to enhance attention distributions without directly altering item embeddings. NRPA (Liu, Wu et al., 2019) developed a personalized word-level attention mechanism to identify distinct important words and comments for different users and items, combining these attention layers through fine- and coarse-grained fusion. VLSNR (Han, Huang, & Luan, 2022) captured users' short-term and long-term interests by achieving both fine-grained and coarse-grained fusion with the multi-head attention and GRU. The multi-order attention layer in MARank (Yu, Zhang, Liang, & Zhang, 2019) was designed to capture both individual- and union-level item interactions by fusing information from multiple perspectives. This approach integrated an attention mechanism and ResNet (Deng et al., 2009) into a unified structure to achieve its objectives.

In this paper, our method falls under the category of combined-level multimodal fusion. We first employ a pretraining model, rather than an attention mechanism, to align the corresponding visual content and textual descriptions at a fine-grained level. Subsequently, we use a cross multi-head attention module between a text transformer and a visual transformer to fuse modality information at the same fine-grained level.

3. Task preliminaries

In this section, we introduce the concepts of CKG and MKG to highlight the challenges in constructing a unified MKG architecture and finally present the definition of our task.

Collaborative Knowledge Graph The collaborative KG is constructed based on a bipartite graph and the supplementary information of the corresponding items. The bipartite graph represents the user's historical behaviors, while the supplementary information typically includes real-world entities and their interconnections, creating a detailed profile for each item. The knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{M})$ consists of entities ($\mathcal{V}$ denotes the set of nodes) and relations ($\mathcal{M}$ denotes the set of edges). Specifically, there are three types of node: user $u \in \mathcal{U}$ item $i \in \mathcal{I}$ and item attribute entity $e \in \mathcal{E}$. Generally, nodes represent entities, and edges connect two entities to form a triple, expressed in the form of (head entity, relation, tail entity), typically denoted as (h, r, t), where $(h, t) \in \mathcal{V}, r \in \mathcal{R}$.

Multimodal Knowledge Graph MKGs can be viewed as a specific case of CKGs that incorporate multimodal entity nodes derived from text and visual modalities. In other words, we substitute an entity node with a multimodal entity. Due to user privacy, in this paper, only the item entity i and its corresponding attribute entity e are replaced by their multimodal entity nodes. Taking an item entity with text and visual modality information as an example, there are four types of node in the MKG: user $u \in \mathcal{U}$, text modality item entity $i_t \in \mathcal{I}$, visual modality item entity $i_v \in \mathcal{I}$, and item attribute entity $e \in \mathcal{E}$. This MKG architecture effectively extends the triple relationships in CKG to generate node features while also facilitating the extraction of node dependencies both inter- and intra-modality.

Task description Based on the above definition, our task in this paper is formulated as follows: given a CKG and its corresponding text and visual modality information, the model constructs the MKG and outputs a probability distribution predicting the item that the user is most likely to be interested in next, based on the MKG.

4. The proposed method

In this section, we illustrate the proposed Multi-KG4Rec, as shown in Fig. 2. First, we divide the multimodal KG into several single-modal graphs. Next, a text transformer and a visual transformer are employed to encode the modality information of the neighbors. To enhance the understanding of relationships across different modalities, we introduce a cross-modal multi-head attention module. Finally, GAT layers are used to propagate the multimodal information. The model predicts potential items based on the generated user and item representations.

4.1. Embedding module

The embedding module primarily consists of entity embedding and embedding optimization.

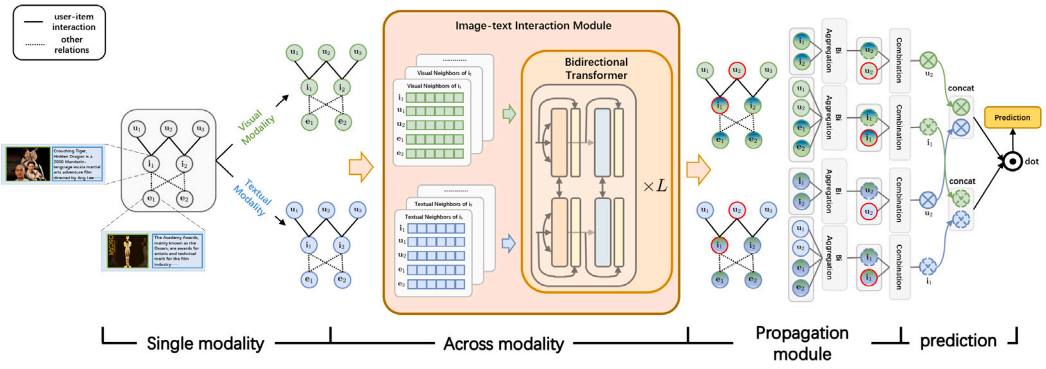


Fig. 2. The architecture of the proposed Multi-KG4Rec.

4.1.1. Entity embedding

Given a triplet (h, r, t) in KG $\mathcal{G}$, we embed the entity ID as the entity structural feature via a lookup table. For multimodal entities (items and their attribute nodes, which have the visual and text content), existing methods (Chen et al., 2021; He & McAuley, 2016; Nikzad-Khasmakhi, Balafar, Feizi-Derakhshi, & Motamed, 2021) usually adopt base encoders, such as CNN and BERT, to model the multimodality feature. However, these methods suffered from the modal misalignment. In this paper, we choose the recently commonly used pre-trained multimodal visual-text model "CLIP" (Radford et al., 2021), a contrastive learning model, to align image-text pairs and generate initial entity features. Specifically, we feed pairs of visual and textual descriptions corresponding to entities into CLIP. The outputs of these two encoders are then projected into a shared embedding space, and the output of the last layer is taken as the feature, with a dimensionality of 512.

4.1.2. Embedding optimization

Here, we adopted TransR (Lin, Liu, Sun, Liu, & Zhu, 2015) to optimize the entity features training. Specifically, nodes and edges in $\mathcal{G}$ will be converted as triplets (h, r, t) and optimized as $\mathbf{e}_h^r + \mathbf{e}_r \approx \mathbf{e}_t^r$, here $\mathbf{e}_h, \mathbf{e}_t \in \mathbb{R}^d$ and $\mathbf{e}_r \in \mathbb{R}^k$ denote the embeddings of h, t and r, respectively. $\mathbf{e}_h^r, \mathbf{e}_t^r$ denote the projected representations of $\mathbf{e}_h$ and $\mathbf{e}_t$ in the space of relation r. Therefore, for a given triplet (h, r, t), the formula for the objective score is as follows:

$$g(h,r,t) = \left\|\mathbf{W}_r\mathbf{e}_h + \mathbf{e}_r - \mathbf{W}_r\mathbf{e}_t\right\|_2^2 \tag{1}$$

here $\mathbf{W}_r \in \mathbb{R}^{k\times d}$ is the transformation matrix to project entities from entity space to relation r space. The score of $g(h, r, t)$ indicates the likelihood that the triplet is true. The training of TransR distinguishes the positive $(h, r, t) \in \mathcal{G}$ and negative $(h, r, t') \notin \mathcal{G}$ triplets with the pairwise ranking loss as follows:

$$\mathcal{L}_{\mathrm{KG}} = \sum_{(h,r,t,t')\in\mathcal{T}} -\ln\sigma\left(g\left(h,r,t'\right) - g(h,r,t)\right) \tag{2}$$

$\sigma(\cdot)$ denotes the sigmoid function.

4.2. Multimodal fusion module

The above entity embedding methods are efficient in capturing the user's multimodal preferences at a coarse-grained level. In this section, we introduce the multimodal fusion module to fuse modality information at a fine-grained level. Specifically, the multimodal fusion module comprises a text transformer, a visual transformer, and a multi-head attention layer. The transformer aims to extract dependencies within a single modality, and ultimately, multimodal fusion is achieved through the multi-head attention layer.

Considering that the text transformer and the visual transformer have similar architectures, here we take the text transformer as an example. Given an item v_i, the first task is to determine the range of the sequence to be fed into the text transformer. In this paper, we treat the item's high-order neighbors as the input sequence $\mathcal{S}_i = \{v_i, u_m^1, e_p^1, e_q^2, u_n^2, e_k^2\}$, v_i is the item itself (we added a self-loop), and u_m^1 denotes that user u_m is the first-order neighbor of item v_i, ^{1} denotes the 1-order neighbors. Since the number of neighbors is different, we adopt the breadth-first method to search for neighbors until the number of $\mathcal{S}_i$ reaches n and sort them by distance. For cold-start nodes, more high-level information will be introduced to enhance their representations.

To fuse modality information at a fine-grained level, we apply a multi-head attention layer (MHA) as Fig. 3(a) shows. Given the neighbor set $\mathcal{S}_i$, the visual feature and text feature corresponding to the neighbor set $\mathcal{S}_i$, denoted as $x_v \in \mathbb{R}^{n\times d}$ and $x_t \in \mathbb{R}^{n\times d}$, respectively. Then transforms x_v into queries $Q_v \in \mathbb{R}^{n\times d}$ and key–value pairs $K_v \in \mathbb{R}^{n\times d}, V_v \in \mathbb{R}^{n\times d}$:

$$Q_v^{(i)}, K_v^{(i)}, V_v^{(i)} = x_v\mathbf{W}_Q^{v(i)}, x_v\mathbf{W}_K^{v(i)}, x_v\mathbf{W}_V^{v(i)} \tag{3}$$

Each head is parameterized by $\mathbf{W}_Q^{v(i)}, \mathbf{W}_K^{v(i)}, \mathbf{W}_V^{v(i)} \in \mathbb{R}^{d\times d_h}$ to transform the inputs into queries, keys, and values. $d_h = d/H$, where H represents the number of heads. Once we acquire the visual head and text head, we combine the K_v and V_v of the visual head with the K_t and V_t of the text head individually. The formula for calculating the visual $\mathrm{head}_i^{\mathrm{M}_v}$ is as follows:

$$\mathrm{head}_i^{\mathrm{M}_v} = \mathrm{Attn}\left(Q_v^{(i)}, \mathrm{concat}\left(K_v^{(i)}, K_t^{(i)}\right), \mathrm{concat}\left(V_v^{(i)}, V_t^{(i)}\right)\right) \tag{4}$$

$$\mathrm{head}^v = \mathrm{concat}\left(\mathrm{head}_1^{\mathrm{M}_v}, \ldots, \mathrm{head}_H^{\mathrm{M}_v}\right) W_o^v \tag{5}$$

The feedforward neural network is another key component of the transformer. It typically consists of two non-linear layers with ReLU activation. The FFN takes the result computed through layer normalization and residual connections as input and performs the following calculations:

$$\mathrm{FFN(x)}^v = \mathrm{ReLU}\left(x_v \boldsymbol{W}_1^v + \mathbf{b}_1^v\right)\boldsymbol{W}_2^v + \mathbf{b}_2^v \tag{6}$$

where $\boldsymbol{W}_1^v \in \mathbb{R}^{d\times d_m}, \boldsymbol{W}_2^v \in \mathbb{R}^{d_m\times d}$.

4.3. Information propagation module

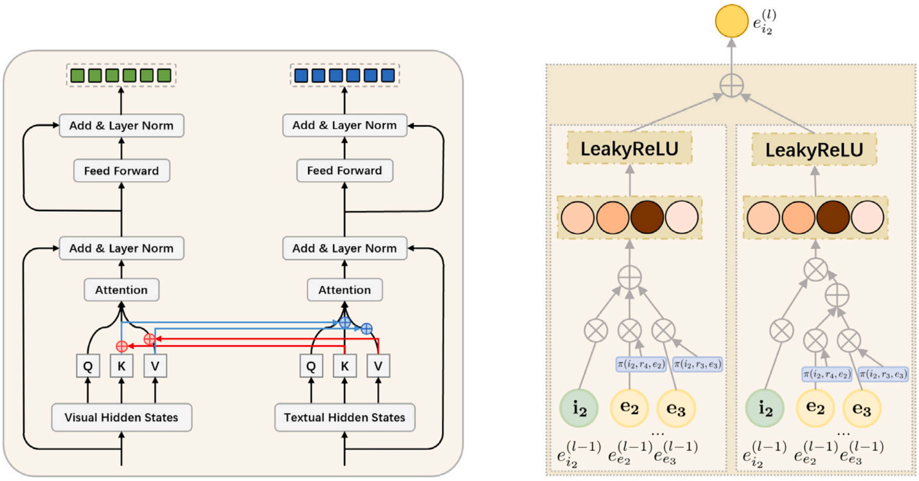
After obtaining the multimodal information, we apply a knowledge-aware graph attention layer to propagate multimodal information to higher-order neighbors.

The entity acts as a bridge to propagate relationships between different triplets. For a given entity h, $\mathcal{N}_h$ denotes the set of triples with h as the head entity $\mathcal{N}_h = (h, r, t \mid (h, r, t) \in \mathcal{G})$. We aggregate the neighbor information as the following formula:

$$\mathbf{e}_{\mathcal{N}_h} = \sum_{(h,r,t)\in\mathcal{N}_h} \pi(h,r,t)\mathbf{e}_t \tag{7}$$

where $\pi(h, r, t)$ controls the flow of information from entity t to entity t based on their relationship r. The $\pi(h, r, t)$ is defined as:

$$\pi(h,r,t) = \left(\mathbf{W}_r\mathbf{e}_t\right)^{\top}\tanh\left(\left(\mathbf{W}_r\mathbf{e}_h + \mathbf{e}_r\right)\right) \tag{8}$$



(a) Across-modal Attention Module

(b) Bi-Interaction Aggregator

Fig. 3. Illustration about modal interaction and high-order information propagation.

$\mathbf{W}_r$ is the trainable weight matrix. The coefficients of all triplets connected to h are then normalized using the softmax function. As shown in Fig. 3(b), we apply the Bi-Interaction to aggregate e_h and $e_{\mathcal{N}_h}$ as follows; $\odot$ denotes the element-wise product.

$$f_{\text{Bi}-Interaction} = \text{LeakyReLU}\left(\mathbf{W}_1\left(\mathbf{e}_h + \mathbf{e}_{\mathcal{N}_h}\right)\right) + \text{LeakyReLU}\left(\mathbf{W}_2\left(\mathbf{e}_h \odot \mathbf{e}_{\mathcal{N}_h}\right)\right) \tag{9}$$

4.4. Prediction

Finally, we obtain $e_u^{(1)}, \ldots, e_u^{(l)}$ and $e_i^{(1)}, \ldots, e_i^{(l)}$. We adopt the layer-aggregation mechanism (Xu et al., 2018) to concatenate the representations at each iteration into a unified vector, as follows:

$$\mathbf{e}_u^* = \mathbf{e}_u^{(0)} \| \cdots \| \mathbf{e}_u^{(l)}, \quad \mathbf{e}_i^* = \mathbf{e}_i^{(0)} \| \cdots \| \mathbf{e}_i^{(l)} \tag{10}$$

Next, the user and item representations from the visual modality are concatenated with those from the text modality to obtain the final user and item representations:

$$\mathbf{e}_u = \mathbf{e}_u^{v(*)} \| \mathbf{e}_u^{t(*)}, \quad \mathbf{e}_i = \mathbf{e}_i^{v(*)} \| \mathbf{e}_i^{t(*)} \tag{11}$$

4.5. Optimizer

To train the recommendation model, we utilize the Bayesian Personalized Ranking (BPR) loss to optimize the parameters in our recommendation based on the prediction loss.

$$\mathcal{L}_{\text{CF}} = \sum_{(u,i)\in\mathcal{R}^+,(u,j)\in\mathcal{R}^-} - \ln \sigma(\hat{y}(u,i) - \hat{y}(u,j)) + \lambda \|\Theta\|_2^2 \tag{12}$$

where $\mathcal{R}^+$ denotes observed (positive) pairs and $\mathcal{R}^-$ is the sampled unobserved (negative) pairs, $\sigma(\cdot)$ is the sigmoid function. The final loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{KG}} + \mathcal{L}_{\text{CF}} \tag{13}$$

Θ denotes all trainable model parameters, and $\|\|$ is the L2 regularization to prevent overfitting.

5. Experiments and analysis

We utilize two commonly used real-world datasets from different domains, MovieLens[1], and Amazon-Books[2] to evaluate the performance of Multi-KG4Rec. First, we introduce the datasets and experimental settings. Next, we discuss the experimental results, and finally, we provide a specific example to illustrate the rationale behind our model.

5.1. Dataset description

MovieLens: This dataset has been extensively used for evaluating recommender models. The raw data includes user IDs, item IDs, and corresponding ratings on a scale from 1 to 5. For this research, we utilized the MovieLens-1M dataset. All ratings were converted to binary, with a rating of 1 marked as 1 and all other ratings marked as 0. To enrich the MovieLens dataset with multimodal information, we followed the methodology in Zhao et al. (2019). We constructed a knowledge graph by establishing connections between items in the dataset and entities in Freebase[3]. Additionally, we retrieved corresponding movie posters and text descriptions from IMDB[4] to serve as the visual and textual multimodal information for the entities.

Amazon-Books: This dataset collects user reviews from one of the world's largest e-commerce websites. For this research, we selected the Amazon-Books subset from the entire collection. We filtered out users with fewer than 10 interactions, following the method in Wang, He et al. (2019). Multimodal information was collected using the same method as for MovieLens dataset. The detailed statistics of these two datasets are shown in Table 1.

5.2. Experimental settings

5.2.1. Evaluation metrics

Items that the user has interacted with will be treated as positive, while others will be considered as candidates. We then select the top-k ranked items as the recommendations. To measure the quality of the recommended sequences, we use three commonly used metrics: Recall@k, MRR@k, and NDCG@k. Among these, MRR@k and NDCG@k are sensitive to ranking positions, while Recall@k is not. The default value for k is 20.

[1] https://grouplens.org/datasets/movielens/.
[2] https://jmcauley.ucsd.edu/data/amazon/.
[3] https://developers.google.com/freebase.
[4] https://www.imdb.com/.

Table 1
Statistics of datasets.

Dataset	#Interactions	#Items	#Users	#Sparsity	#Entities	#Relations	#Triplets
MovieLens	834,268	3589	6040	96.15%	60,406	51	273,547
Amazon-Books	332,834	18,932	24,047	99.92%	44,935	23	192,388

5.2.2. Baselines

We conducted a comparative analysis of our proposed Multi-KG4Rec model against several baselines. These baselines include collaborative filtering methods (i.e., SpectralCF, ConvNCF), knowledge graph-based approaches (i.e., KGAT, KGCN, CKE), and multimodal methods that incorporate knowledge graphs (i.e., MKGAT).

- SpectralCF (Zheng, Lu, Jiang, Zhang, & Yu, 2018): Spectral collaborative filtering innovatively proposes a convolutional model in the spectral domain space based on the bipartite graph. By leveraging the rich connectivity information from the spectral domain, SpectralCF reveals deep connections between user-item interactions, effectively alleviating the cold-start problem.
- ConvNCF (He et al., 2018): Neural Collaborative Filtering utilizes element-wise products to explicitly capture pairwise correlations among dimensions within the embedding space.
- KGAT (Wang, He et al., 2019): This method integrates TransR (Lin et al., 2015) and GNN to generate the entity representation.
- KGCN (Wang, Zhao et al., 2019): This method utilizes a fixed number of neighbors as the receptive field and obtains the final vector representation by GNN.
- CKE (Zhang et al., 2016): CKE integrates structural information, textual data, and image data to enhance the quality of the recommender model. Structural information was obtained via TransR, while textual data and image data were extracted by stacked denoising and stacked convolutional auto-encoders, respectively.
- MKGAT (Sun et al., 2020): MKGAT is one of the representative methods in the field of multimodal recommendation models, which presents a multimodal graph attention mechanism to solve entity information aggregation and entity relationship reasoning.

5.2.3. Parameter settings

We randomly split the whole interactions with $8:1:1$ to train, validate, and test our model. We apply the Xavier initializer to initialize model parameters and use the Adam optimizer for model optimization. Mini-batch sizes and learning rates were searched within the sets $\{1,024,\ 5,120,\ 10,240\}$ and $\{0.0001,\ 0.0005,\ 0.001,\ 0.005,\ 0.01\}$, respectively. The coefficient of λ is set in $\{10^{-5}, 10^{-4}, \dots, 10^{-1}\}$. For visual and text entities, we utilized the 512-dimensional features from the last layer of the CLIP model and then reduced the dimension to 64 via a non-linear transformation with the $LeakReLU$ activation function. We stacked 3 blocks in the multimodal fusion module, each with 8 attention heads. For the information propagation, we set 3 layers of the knowledge-aware graph neural network to encode high-order connectivity, with the output dimension of each layer being $\{64,\ 32,\ 16\}$. We implement our Multi-KG4Rec model in PyTorch. All experiments were conducted on a Windows PC equipped with an RTX 3090 GPU.

5.3. Experimental results

5.3.1. Comprehensive comparison

The results for all models are shown in Table 2, and it can be observed that our Multi-KG4Rec model outperforms all baseline models on both the MovieLens and Amazon-Books datasets. We have the following findings:

Table 2
Overall performance comparison.

Models	MovieLens			Amazon-Books		
	Recall	MRR	NDCG	Recall	MRR	NDCG
SpectralCF	0.2199	0.3714	0.2082	0.1327	0.0541	0.0602
ConvNCF	0.1815	0.3405	0.1794	0.0404	0.0148	0.0175
KGAT	0.2489	0.3941	0.2303	0.1431	0.0553	0.0702
KGCN	0.2268	0.3783	0.2165	0.1418	0.0528	0.0677
CKE	0.2217	0.3754	0.2128	0.1324	0.0491	0.0612
MKGAT	0.2513	0.3963	0.2311	0.1477	0.0560	0.0707
Multi-KG4Rec	**0.2552**	**0.4077**	**0.2383**	**0.1498**	**0.0572**	**0.0727**
Improv.	1.55%	2.88%	3.12%	1.42%	2.83%	2.14%

- Multi-KG4Rec consistently performs the best across the two datasets. Specifically, compared to MKGAT, Multi-KG4Rec improved recall@20 by up to 3.12% and 2.14% for MovieLens and Amazon-Books, respectively. MKGAT is a representative feature-based method, and this result verifies the importance of modality fusion. The multimodal fusion in MKGAT enables the model to integrate information from various sources, such as text, images, and interactions, providing a richer view of user-item interactions. This comprehensive perspective is crucial for making accurate and personalized recommendations, especially when user preferences are influenced by multiple factors.
- KGAT and KGCN can be viewed as the entity-based method, with KGAT outperforming KGCN. KGCN applied a more extensive entity receptive field to aggregate the heterogeneous and high-order neighborhood information. However, compared to KGAT, the method of extending the receptive field introduced noise. In terms of model architecture, CKE is similar to our model, as both divide text and images into separate modes. However, CKE performed the worst among the KG-based methods. This is because CKE does not use the GNN to aggregate high-order neighbor information, thus lacking sufficient attention on the interaction between different modes.
- Among all comparative methods, KG-based methods (i.e., CKE, KGAT, KGCN, MKGAT) outperform collaborative filtering-based methods (i.e., SpectralCF, ConvNCF) on both datasets. This demonstrates that KGs are useful in helping GNNs encode relational reasoning between attribute entities. It is undeniable that the multimedia information serving as supplementary features or nodes not only alleviates the issues of data sparsity and cold start but also enhances the model's ability to understand the relationships between users and items, thereby improving the quality of recommended results.

5.3.2. Modality effectiveness analyses

To evaluate the effectiveness of different modalities, we compare the results of MKGAT and our Multi-KG4Rec on the MovieLens dataset. Table 3 presents the results of the model performance; here, "w/o t&i" denotes the multimodal fusion module is disabled, "w/o v" represents the case where only the visual information is activated, and "w/o t" is similar to the "w/o v." From Table 3, we have the following observations:

It is obvious that models incorporating multimodal features achieve superior performance compared to those relying on a single modality. This is because multimodal information offers rich and diverse item characteristics from different perspectives, thus improving the model's

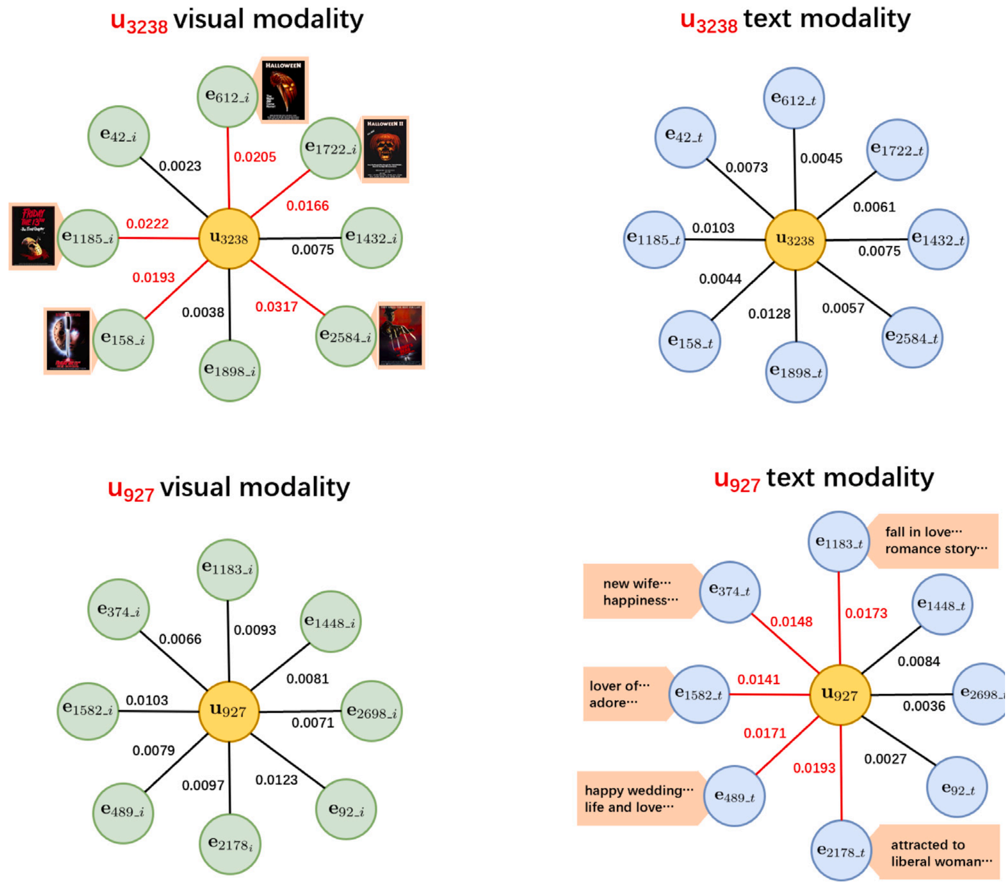


Fig. 4. Attention distribution for u_{3238} and u_{927}.

Table 3
Impact of different modalities on MovieLens.

Models	MKGAT			Multi-KG4Rec		
	Recall	MRR	NDCG	Recall	MRR	NDCG
w/o t&v	0.2453	0.3907	0.2251	0.2489	0.3941	0.2303
w/o v	0.2477	0.3949	0.2272	0.2518	0.4014	0.2327
Improv.	1.00%	1.07%	0.93%	1.16%	1.85%	1.04%
w/o t	0.2479	0.3951	0.2285	0.2531	0.4016	0.2340
Improv.	1.06%	1.13%	1.51%	1.69%	1.90%	1.61%
Multi-KG4Rec	**0.2488**	**0.3963**	**0.2311**	**0.2542**	**0.4033**	**0.2371**
Improv.	1.42%	1.43%	2.67%	2.13%	2.33%	2.95%

ability to understand user intention. Under a single modal, we find that the visual modality is more effective than the text modality. This conclusion aligns with most multimodal models (Li, Feng et al., 2023; Liu, Ma et al., 2022) where images typically contain more information than text. In other words, the visual modality has a higher weight on the users' final decisions compared to text content. This also motivates us to explore additional modalities, such as video, web pages, and tables, to provide more contextual information for user profile generation. Additionally, compared to the MKGAT, Multi-KG4Rec achieves superior performance, proving that Multi-KG4Rec has strong expressive power to perceive the implicit relationship between images and texts. This performance improvement is attributed to the multimodal fusion module, which effectively extracts cross-modal information at a fine-grained level.

5.3.3. Ablation study

Table 3 has verified the effectiveness of the multimodal fusion module. To further analyze the effectiveness of the Bi-Transformer, we modified the bi-directional attention mechanism (denoted "Bi-Trans") into two variants: only text-to-image attention activated (denoted as Bi-Trans$_{t2v}$) and only image-to-text attention activated (denoted as Bi-Trans$_{v2t}$). Here, "w/o t&v" is consistent with Table 4, which denotes that the multimodal fusion module is disabled. The results are shown in Table 4. Based on the results, we have the following findings:

Table 4
Ablation study about the multimodal fusion module.

Dataset	MovieLens			Amazon-Books		
	Recall	MRR	NDCG	Recall	MRR	NDCG
w/o t&v	0.2453	0.3907	0.2251	0.1473	0.0566	0.0716
Bi-Trans$_{t2v}$	0.2437	0.3917	0.2244	0.1428	0.0514	0.0674
Bi-Trans$_{v2t}$	0.2444	0.3944	0.2227	0.1436	0.0521	0.0662
Multi-KG4Rec	**0.2552**	**0.4077**	**0.2383**	**0.1498**	**0.0572**	**0.0727**

Multi-KG4Rec performs worse on both datasets in terms of Recall, MRR, and NDCG. This result further confirms the conclusion in the previous subsection that integrating multimodal information enhances the model's expressiveness. Whether the text-to-visual attention is disabled or the visual-to-text attention is deactivated, performance decreases. This may be because unidirectional transformers consider correlations from one side, potentially resulting in the loss of text information relevant to image features, and vice versa. In contrast, a idirectional transformer has several advantages in multimodal tasks. They first independently extract significant features from each modality. Then, the weight of these significant features is adjusted through the cross-modality attention module in the bidirectional transformers. This mechanism enhances the interaction between modalities, leading to better performance. Additionally, bidirectional transformers exhibit more robustness to noisy data or irrelevant information within a single modality, thereby improving overall performance.

5.4. Case study

To validate the significance of modalities in influencing user preferences, we selected a user from both the MovieLens dataset and the Amazon-Books dataset and gathered 10 items with which the user had interacted. Using the attention mechanism, we computed correlation scores between user-item pairs. The higher the correlation score, the greater the impact of the current item's modality on user preferences. The result is visualized in Fig. 4.

Fig. 4 shows that different users exhibit varying preferences for visual and text modalities. Specifically, user u_{3238} from the MovieLens dataset exhibits a significantly higher score for the visual modality compared to the text modality, while u_{927} shows the opposite trend. This validates the rationality and necessity of discussing modal fusion at a fine-grained level. The underlying reason could be that u_{3238} from the MovieLens dataset prioritizes movie posters over text descriptions when selecting a movie, while u_{927} collected from the book sales websites exhibits the opposite preference. Furthermore, we visualized the multimodal content with a high attention score and found that u_{3238} prefers posters with scary elements while u_{927} tends to be romantic-themed books.

6. Conclusion

In this study, we proposed Multi-KG4Rec, a personalized recommendation multimodal fusion framework based on a knowledge graph. This framework leverages the Bi-Transformer to effectively learn the potential relationships and interactions between textual and visual modalities. Subsequently, it employs a GNN layer to propagate high-order information. Extensive experiments conducted on two real-world datasets validate the effectiveness of the proposed Multi-KG4Rec model. Furthermore, we believe that web pages can serve as an additional modality to offer contextual information for items. There have been relatively few research studies in this field. In the future, we aim to collect such web page datasets and design models to validate our thoughts.

CRediT authorship contribution statement

Jingjing Wang: Conceptualization, Methodology, Visualization, Writing – original draft, Writing – review & editing, Validation. **Haoran Xie:** Writing – review & editing, Visualization, Validation, Supervision, Funding acquisition. **Siyu Zhang:** Data curation, Validation. **S. Joe Qin:** Writing – review & editing, Validation, Supervision, Funding acquisition. **Xiaohui Tao:** Validation, Supervision. **Fu Lee Wang:** Writing – review & editing, Supervision. **Xiaoliang Xu:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research has been supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (R1015-23) and the Faculty Research Grant (DB24A4) of Lingnan University, Hong Kong.

Data availability

Data will be made available on request.

References

Cao, X., Shi, Y., Wang, J., Yu, H., Wang, X., & Yan, Z. (2022). Cross-modal knowledge graph contrastive learning for machine learning method recommendation. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 3694–3702).

Chen, W., Huang, P., Xu, J., Guo, X., Guo, C., Sun, F., et al. (2019). POG: personalized outfit generation for fashion recommendation at Alibaba iFashion. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2662–2670).

Chen, X., Lu, Y., Wang, Y., & Yang, J. (2021). CMBF: Cross-modal-based fusion recommendation algorithm. *Sensors*, *21*(16), 5275.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Fan, H., Zhong, Y., Zeng, G., & Ge, C. (2022). Improving recommender system via knowledge graph based exploring user preference. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, 1–13.

Han, S., Huang, W., & Luan, X. (2022). VLSNR: Vision-linguistics coordination time sequence-aware news recommendation. arXiv preprint arXiv:2210.02946.

He, X., Du, X., Wang, X., Tian, F., Tang, J., & Chua, T.-S. (2018). Outer product-based neural collaborative filtering. arXiv preprint arXiv:1808.03912.

He, R., & McAuley, J. (2016). VBPR: Visual bayesian personalized ranking from implicit feedback. *vol. 30*, In *Proceedings of the AAAI conference on artificial intelligence*.

Hou, M., Wu, L., Chen, E., Li, Z., Zheng, V. W., & Liu, Q. (2019). Explainable fashion recommendation: A semantic attribute region guided approach. arXiv preprint arXiv:1905.12862.

Hu, B., Shi, C., Zhao, W. X., & Yu, P. S. (2018). Leveraging meta-path based context for top-n recommendation with a neural co-attention model. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1531–1540).

Li, J., Chiu, B., Feng, S., & Wang, H. (2020). Few-shot named entity recognition via meta-learning. *IEEE Transactions on Knowledge and Data Engineering*, *34*(9), 4245–4256.

Li, J., Feng, S., & Chiu, B. (2023). Few-shot relation extraction with dual graph neural network interaction. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, J., Han, P., Ren, X., Hu, J., Chen, L., & Shang, S. (2021). Sequence labeling with meta-learning. *IEEE Transactions on Knowledge and Data Engineering*, *35*(3), 3072–3086.

Li, J., Shang, S., & Chen, L. (2020). Domain generalization for named entity boundary detection via metalearning. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(9), 3819–3830.

Li, J., Wang, Y., Zhang, S., & Zhang, M. (2023). Rethinking document-level relation extraction: A reality check. arXiv preprint arXiv:2306.08953.

Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. *vol. 29*, In *Proceedings of the AAAI conference on artificial intelligence*.

Lin, Y., Ren, P., Chen, Z., Ren, Z., Ma, J., & De Rijke, M. (2019). Explainable outfit recommendation with joint outfit matching and comment generation. *IEEE Transactions on Knowledge and Data Engineering*, *32*(8), 1502–1516.

Liu, S., Chen, Z., Liu, H., & Hu, X. (2019). User-video co-attention network for personalized micro-video recommendation. In *The world wide web conference* (pp. 3020–3026).

Liu, C., Li, X., Cai, G., Dong, Z., Zhu, H., & Shang, L. (2021). Noninvasive self-attention for side information fusion in sequential recommendation. *vol. 35*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 4249–4256).

Liu, H., Li, C., & Tian, L. (2022). Multi-modal graph attention network for video recommendation. In *2022 IEEE 5th international conference on computer and communication engineering technology* (pp. 94–99). IEEE.

Liu, Z., Ma, Y., Schubert, M., Ouyang, Y., & Xiong, Z. (2022). Multi-modal contrastive pre-training for recommendation. In *Proceedings of the 2022 international conference on multimedia retrieval* (pp. 99–108).

Liu, H., Wu, F., Wang, W., Wang, X., Jiao, P., Wu, C., et al. (2019). NRPA: neural recommendation with personalized attention. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval* (pp. 1233–1236).

Nikzad-Khasmakhi, N., Balafar, M. A., Feizi-Derakhshi, M. R., & Motamed, C. (2021). BERTERS: Multimodal representation learning for expert recommendation system with transformers and graph embeddings. *Chaos, Solitons & Fractals*, *151*, Article 111260.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.

Sun, R., Cao, X., Zhao, Y., Wan, J., Zhou, K., Zhang, F., et al. (2020). Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 1405–1414).

Wang, X., He, X., Cao, Y., Liu, M., & Chua, T.-S. (2019). Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 950–958).

Wang, H., Zhang, F., Xie, X., & Guo, M. (2018). DKN: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 world wide web conference* (pp. 1835–1844).

Wang, H., Zhang, F., Zhang, M., Leskovec, J., Zhao, M., Li, W., et al. (2019). Knowledge-aware graph neural networks with label smoothness regularization for recommender systems. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 968–977).

Wang, H., Zhao, M., Xie, X., Li, W., & Guo, M. (2019). Knowledge graph convolutional networks for recommender systems. In *The world wide web conference* (pp. 3307–3313).

Wu, C., Wu, F., Qi, T., & Huang, Y. (2021). Mm-rec: multimodal news recommendation. arXiv preprint arXiv:2104.07407.

Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.-i., & Jegelka, S. (2018). Representation learning on graphs with jumping knowledge networks. In *International conference on machine learning* (pp. 5453–5462). PMLR.

Yang, Y., Huang, C., Xia, L., & Li, C. (2022). Knowledge graph contrastive learning for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval* (pp. 1434–1443).

Yu, L., Zhang, C., Liang, S., & Zhang, X. (2019). Multi-order attentive ranking model for sequential recommendation. *vol. 33*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 5709–5716).

Zhang, F., Yuan, N. J., Lian, D., Xie, X., & Ma, W.-Y. (2016). Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 353–362).

Zhao, W. X., He, G., Yang, K., Dou, H., Huang, J., Ouyang, S., et al. (2019). Kb4rec: A data set for linking knowledge bases with recommender systems. *Data Intelligence*, *1*(2), 121–136.

Zhao, H., Yao, Q., Li, J., Song, Y., & Lee, D. L. (2017). Meta-graph based recommendation fusion over heterogeneous information networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 635–644).

Zheng, L., Lu, C.-T., Jiang, F., Zhang, J., & Yu, P. S. (2018). Spectral collaborative filtering. In *Proceedings of the 12th ACM conference on recommender systems* (pp. 311–319).