MM-FRec: Multi-Modal Enhanced Fashion Item Recommendation

Xuemeng Song, *Senior Member, IEEE*, Chun Wang, Changchang Sun, Shanshan Feng, Min Zhou, *Member, IEEE*, and Liqiang Nie, *Senior Member, IEEE*

***Abstract*—Existing studies on fashion item recommendation mainly focused on incorporating the visual signals of items to boost the user preference learning, while overlooking the semantic attributes (e.g., material and brand) of fashion items that also contain important cues about items' properties and users' preference. To bridge this gap, we aim to comprehensively explore the attribute and vision modalities of items to improve the fashion item recommendation performance. However, this is non-trivial due to the latent visual-semantic consistency, various relation types, and unique attributes with insufficient samples. To address these challenges, we propose a Multi-Modal enhanced Fashion item Recommendation scheme (MM-FRec). Specifically, to cope with the multi-modal data, we introduce a relation-oriented graph as well as a vision-oriented graph, and design MM-FRec with three key components: attribute-enhanced latent representation learning, visual representation learning, and multi-modal enhanced preference modeling. To deal with the various relation types, we present a new relation-aware propagation method for adaptively aggregating the information from neighbor nodes to promote the user and item representation learning. To cope with the unique attributes, we introduce the deep multi-task learning strategy in the relation-aware confidence assignment. Extensive experiments on a real-world dataset demonstrate the superiority of our model over state-of-the-art methods.**

***Index Terms*—Explainable recommendation, fashion recommendation, multi-modal data.**

I. Introduction

RECENT years have witnessed the flourishing of online fashion industry and the unprecedented growth of fashion data on E-commerce platforms, such as Amazon and Taobao. Surfing in the huge online fashion market, people always get overwhelmed and feel difficult to find their desired items. Therefore, personalized fashion item recommendation has become an emerging need, which can not only increase the platform's economic benefits but also improve the user's consumption experience. Different from the traditional recommendation tasks, one distinct feature of the fashion item recommendation task is that it should take into account not only the user-item historical interactions but also the visual content of fashion items. This is due to the fact that the visual property of the item is an essential factor affecting the user's preference to the item. In light of this, several visual-enhanced fashion item recommendation methods have been proposed [1], [2], [3]. Although these efforts have achieved compelling success, most of them overlook the side information of items, i.e, attributes, like "color" and "category". In fact, the attributes of an item usually convey its key features, and some attributes can be even hard to be expressed by the item's visual content, like the "material" and "brand". Therefore, incorporating the attribute information to boost the item representation learning and benefit the user preference modeling merits our special attention. Meanwhile, the intuitive semantic delivered by the attribute information can be used for improving the model interpretability and facilitating users to understand the recommendation results better. Therefore, to bridge the research gap, in this work, apart from the user-item interaction histories, we also jointly consider the attribute and vision modalities of items to boost the performance and interpretability of fashion item recommendation systems.

However, fulfilling the multi-modal enhanced fashion item recommendation is non-trivial due to the following three key challenges. 1) *C1: Latent Visual-Semantic Consistency.* Although the visual image and semantic attributes characterize the same fashion item from different perspectives, they should share certain latent consistency in terms of delivering the item's key features. Therefore, the primary challenge is how to effectively utilize the heterogeneous multi-modal data (i.e, the image and attributes) of items with the visual-semantic consistency modeling to enhance the user and item representation learning and thus promote the performance of fashion item recommendation. 2) *C2: Various Relation Types.* The relation types among our entities are various, including both the user-item relation (i.e, *interact*) and item-attribute relations (e.g., *hasColorOf*, *hasBrandOf*, and *hasCategoryOf*). Moreover, different relation types indicate different correlations among entities. Intuitively, different types of attributes contribute differently in characterizing the item and conveying the user's preference. Therefore, how to adaptively

Manuscript received 7 January 2022; revised 9 January 2023; accepted 8 April 2023. Date of publication 12 April 2023; date of current version 15 September 2023. This work was supported in part by the Shandong Provincial Natural Science Foundation under Grant ZR2022YQ59; in part by the National Natural Science Foundation of China under Grants U1936203 and 62236003; and in part by Shenzhen College Stability Support Plan under Grant GXWD20220817144428005. Recommended for acceptance by Z. Guan. *(Corresponding author: Xuemeng Song.)*

Xuemeng Song and Chun Wang are with the School of Computer Science and Technology, Shandong University, Qingdao, Shandong 266237, China (e-mail: sxmustc@gmail.com; chunwang@mail.sdu.edu.cn).

Changchang Sun is with the School of Computer Science, Illinois Institute of Technology, Chicago, IL 60616 USA (e-mail: sunchangchang123@gmail.com).

Shanshan Feng and Liqiang Nie are with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong 518055, China (e-mail: victor_fengss@foxmail.com; nieliqiang@gmail.com).

Min Zhou is with the Noah's Ark Lab, Huawei Technologies Co.,Ltd., Shenzhen, Guangdong 518129, China (e-mail: zhoum1900@163.com).

Digital Object Identifier 10.1109/TKDE.2023.3266423

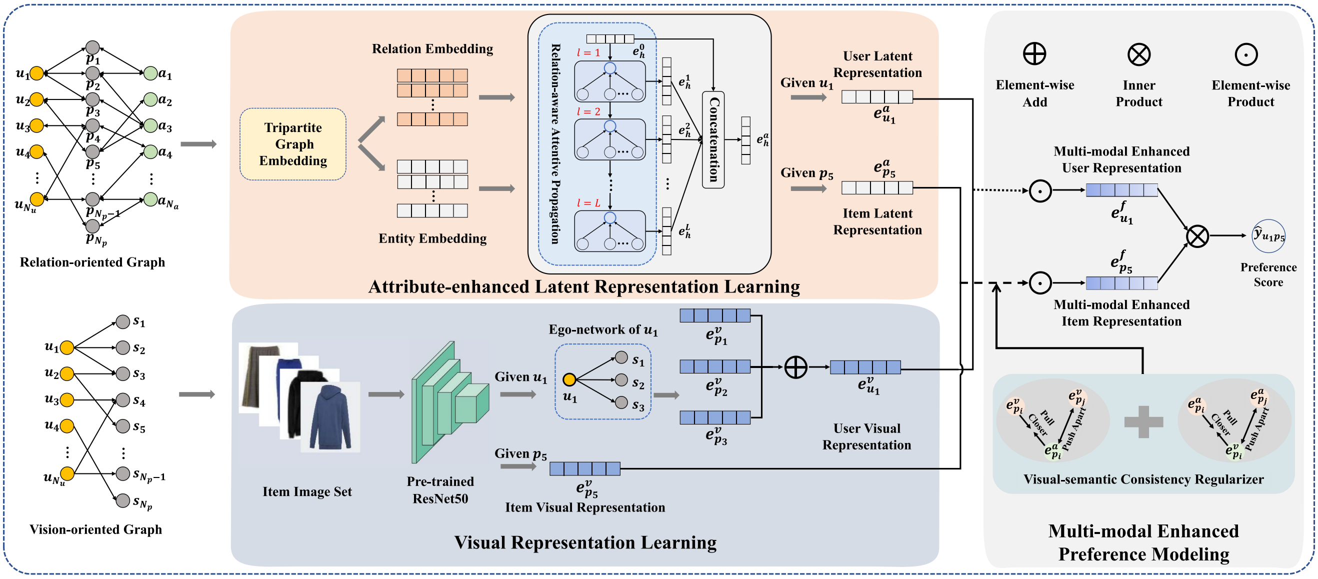


Fig. 1. Illustration of the proposed MM-FRec, which consists of three components: attribute-enhanced latent representation learning, visual representation learning, and multi-modal enhanced preference modeling.

measure the correlation among entities taking into account their relation types poses a key challenge. 3) *C3: Unique Attributes with Insufficient Samples.* In real-world applications, not all the attributes are so common that being possessed by a large number of items. For example, only certain shoes items have the "heel" attribute. Consequently, how to properly deal with these unique attributes that have insufficient samples constitutes another challenge.

To address the aforementioned challenges, as shown in Fig. 1, we propose a Multi-Modal enhanced Fashion Item Recommendation scheme, termed as MM-FRec. The heterogeneous multi-modal data of items are engaged to learn user and item embeddings, and we utilize the Bayesian Personalized Ranking (BPR) [4] loss to fulfil the user's preference learning. In particular, for the challenge **C1**, we devise a relation-oriented user-item-attribute tripartite graph and a vision-oriented user-image bipartite graph to learn the user and item embeddings from the attribute and image perspectives, respectively. We enforce the embeddings of the same item from different perspectives to be similar with a visual-semantic consistency regularizer, for the purpose of enabling the learned knowledge from the two graphs to be shared. For the challenge **C2**, we design a new relation-aware attentive propagation method to learn the user and item embeddings in the user-item-attribute tripartite graph, which employs neural networks to learn the aggregating attention of neighbors taking not only the information of head and tail nodes, but the extra information of relation types as the inputs. Towards the challenge **C3**, we treat the aggregating attention weight learning of different relation types as multiple correlated tasks and design a deep multi-task learning scheme to learn the aggregating attention weights. It consists a shared network that captures the common knowledge in different relation types, and multiple relation-specific networks that learn the specific knowledge in each relation type.

Our main contributions can be summarized in threefold:

- We present a novel Multi-Modal enhanced Fashion Item Recommendation scheme (MM-FRec) that jointly investigates the attribute-enhanced latent preference and visual preference of a user, where a visual-semantic consistency regularizer is introduced to mutually enhance the two preference learning.
- We design a new relation-aware propagation method for the user preference learning based on GCN, where the importance of different connections during the propagation is flexibly measured by a deep multi-task learning framework and can be used for explaining the recommendation results. As far as we know, we are the first to employ the deep multi-task learning framework to enhance the connection confidence assignment for various attributes.
- We conduct extensive experiments on a real-world dataset, and the results show that our MM-FRec achieves a prominent improvement of 28.40%, 20.22%, 24.82% in Precision@20, Precision@50, and NDCG@100, over the best baseline, respectively. As a byproduct, we have released the codes and involved parameters to benefit the community[1].

II. Related Work

In this section, we briefly present the related work, including fashion item recommendation and graph-based recommendation in Section II-A and B, respectively.

A. Fashion Item Recommendation

In recent years, fashion item recommendation has been attracting increasing attention, owing to its huge economic value. Due to the fact that the visual features of items bring much positive force to reveal user preferences, many researchers have incorporated the visual images of items to boost the fashion item

[1] https://anonymousMM-FRec.wixsite.com/MM-FRec.

recommendation performance. For example, McAuley et al. [5] proposed a linear model to learn the implicit relationships between items based on their visual images. To alleviate the cold start issues, apart from the visual feature, He et al. [2] also explored the user implicit feedback (i.e, browsing logs and the temporal information) to predict the user fashion taste. Besides, Yu et al. [3] designed a Brain-inspired Deep Network (BDN) to enhance the user preference modeling from the aesthetic perspective. Considering the semantic information of items that can intuitively express the key features of items, there are several researchers perform the multi-modal recommendation by leveraging the item images and user reviews to items [6], [7], [8]. For example, Chen et al. [6] learned an attention model over the image of fashion items and introduced user review information as a weak supervision signal to collect more comprehensive user preference. In addition, Cheng et al. [8] proposed a multi-modal aspect-aware topic model (MATM) on text reviews and item images to model users' preferences and items' features from different aspects. Although these efforts achieved significant progress, the multi-modal information lying in the review or description is usually informal, unstructured, and noisy, which largely limits the model performance. To address this issue, Hou et al. [1] proposed to extract the attribute information from the item image, and utilize the attention mechanism to find the explanations for recommendation results. Nevertheless, due to the limited expressive capability of the item images in reflecting the item attributes, this method still suffers from the sub-optimal performance. Beyond that, in this work, we exploit the explicit item attributes to boost the recommendation performance.

B. Graph-Based Recommendation

Early graph-based recommendation aims to construct a simple user-item interaction graph and exploit it to infer user preferences, such as ItemRank [9] and BiRank [10]. Specifically, ItemRank adopts a random walk-based algorithm to predict item scores for the target users and BiRank iteratively assigns scores to vertices and finally converges to a unique stationary ranking. Later, some researchers [11], [12] attempted to use the connection of entities in the knowledge graph for embedding learning to improve the recommendation performance. For example, Zhang et al. [12] adopted TransR to learn item embeddings under the guidance of structural information in the knowledge graph. One key limitation of these studies is that they overlook the higher-order connectivity [13] in the knowledge graph, which is helpful for mining the potential interests of users. Therefore, some researchers proposed to perform iterative propagation over the entire knowledge graph for user preference modeling. For instance, Wang et al. [14] and Wang et al. [15] applied graph neural networks to achieve embedding propagation on the knowledge graph. In addition, He et al. [16] developed LightGCN, which improves model efficiency without sacrificing accuracy by removing the nonlinear activation and feature transformation from NGCF [14].

Recent researches [17], [18], [13], [19], [20] propose to incorporate the semantic information into the graph learning. Most of them propose to construct a unified graph of users, items, and attributes and model their relationships through the information propagation. For example, Wang et al. [13] proposed KGAT, which jointly considers user-item and item-attribute interactions in the information propagation process. Ai et al. [20] introduced the knowledge graph to organize items' attributes, and designed a path-based soft matching algorithm to generate corresponding explanations. Similarly, Chen et al. [19] modeled the connectivity between different entities (i.e., user, item and attribute) in the knowledge graph via GCN.

Although these graph-based recommendation methods can learn the relationships between different entities, they overlook different relation types between entities. In fact, different relation types indicate different correlations among entities and contribute differently when conveying the user preference. Therefore, in this work, we propose to learn the user preference with a newly designed relation-aware attentive propagation method, which learns the different contributions of different nodes in the graph during the information propagation based on relation types.

III. Problem Formulation

Suppose we have a set of users $\mathcal{U} = \{u_i\}_{i=1}^{N_u}$, and a set of fashion items $\mathcal{P} = \{p_j\}_{j=1}^{N_p}$, where N_u and N_p are the total numbers of users and items, respectively. Meanwhile, we have a set of possible attribute values (e.g., *red color* and *wool material*) possessed by items, denoted as $\mathcal{A} = \{a_m\}_{m=1}^{N_a}$, where N_a is the total number of attribute values. Notably, each fashion item is associated with an image and a few attribute values of $\mathcal{A}$. Let $\mathcal{V} = \{v_j\}_{j=1}^{N_p}$ denote the set of item images, where v_j is the image of item p_j.

On the one hand, based on the user-item historical interactions (e.g., user u_i purchases item p_j) and the item-attribute association relations (e.g., item p_j has the attribute a_l), we can derive a user-item-attribute tripartite graph $\mathcal{G}_a = (\mathcal{E}_a, \mathcal{R}_a)$. In particular, the node set $\mathcal{E}_a = \mathcal{U} \bigcup \mathcal{P} \bigcup \mathcal{A}$ denotes the set of entities, including user entities, item entities and attribute entities, while $\mathcal{R}_a = \{r_1, \ldots, r_{2(F+1)}\}$ refers to the set of relation types held by these entities, which includes one type of user-item interaction relation as well as F types of item-attribute association relations, and their reverse ones. For instance, the relation "*InteractBy*" is the reverse form of the relation "*Interact*", while "*IsColorOf*" is the reverse one of the relation "*HasColorOf*". The motivation that we incorporate the reverse relations is to allow the information propagation in both directions. In this manner, each item entity can absorb knowledge from its attribute entities, while the attribute entity can also distill information from its related item entities, which benefits the latent representation learning of entities. For ease of the following method description, similar to previous works, we formulate the user-item-attribute graph $\mathcal{G}_a$ as a set of triplets $\mathcal{T} = \{(h, r, t) | h, t \in \mathcal{E}_a, r \in \mathcal{R}_a\}$, where each triplet (h, r, t) indicates that there is a relation of the type r from the head entity h to the tail entity t. Here we give some intuitive triplet examples, such as $(user1, bought, item1)$, $(item1, HasColorOf, black)$, and $(black, IsColorOf, item1)$.

On the other hand, based upon the user-item historical interactions, besides the relation-oriented graph $\mathcal{G}_a$, we can also derive a vision-oriented graph $\mathcal{G}_v = (\mathcal{E}_v, \mathcal{R}_v)$ to characterize the user-item interactions from the vision aspect. In particular, $\mathcal{E}_v = \mathcal{U} \bigcup \mathcal{V}$ denotes the set of entities, including the user entities and image entities, while $\mathcal{R}_v$ contains only one relation type, i.e, the interaction relation between the user and the item image. The reason why we do not incorporate the reversed relation in the vision-oriented graph is the concern that the user's visual preferences should be derived by summarizing the objective visual features of items that the user interacted before, while the bidirectional relation that allows the bidirectional information propagation may hurt such objectivity. Similar to $\mathcal{G}_a$, the vision-oriented graph $\mathcal{G}_v$ can be represented as a triplet set $\mathcal{Z} = \{(u, r', v)|u \in \mathcal{U}, v \in \mathcal{V}\}$, where each triplet (u, r', v) represents that the user u once interacted with an item that has the image of v, and $r' = interact$.

Goal. In this work, based on these two graphs, i.e, $\mathcal{G}_a$ and $\mathcal{G}_v$, we aim to train a model $\mathcal{F}$ that is able to predict the preference score of an arbitrary user u to an arbitrary item p as follows,

$$\hat{y}_{u,p} = \mathcal{F}(u, p|\mathcal{G}_a, \mathcal{G}_v), \tag{1}$$

where $\hat{y}_{u,p}$ denotes the predicted preference score of the user u towards the item p.

IV. MM-FRec

In this section, we present the proposed MM-FRec, as shown in Fig 1, which consists of three key components: 1) *Attribute-enhanced Latent Representation Learning*, 2) *Visual Representation Learning*, and 3) *Multi-modal Enhanced Preference Modeling*. The first component works on learning the attribute-enhanced latent representations for users and items over the relation-oriented tripartite graph. The second component targets on investigating the visual representations of users and items based upon the vision-oriented bipartite graph. The last component aims to learn the multi-modal enhanced user preference based on the attribute-enhanced representations and visual representations of users and items.

A. Attribute-Enhanced Latent Representation Learning

This component consists of two key modules: *Tripartite Graph Embedding* and *Relation-aware Attentive Propagation*. The former is designed to initialize the entity and relation embeddings of the user-item-attribute tripartite graph at each run. The latter is devised to adaptively perform the information propagation over the tripartite graph to derive the attribute-enhanced latent representation of users and items, where we assign the weights to neighbor nodes according to not only the head/tail nodes of a triplet but also their corresponding relation type. In particular, to deal with the unique attributes, we borrow the idea of multi-task learning and fulfil the relation-aware attention mechanism through a shared network and multiple relation-specific networks.

1) Tripartite Graph Embedding: Knowledge graph embedding has been proven to be an effective way to parameterize entities and relations as vector representations by preserving the graph structure, which has been used in various knowledge enhanced recommendation methods [13], [19], [20]. Therefore, in our context, to learn the entity and relation embeddings of the user-item-attribute tripartite graph, we resort to the widely-used knowledge graph embedding method TransR [21].

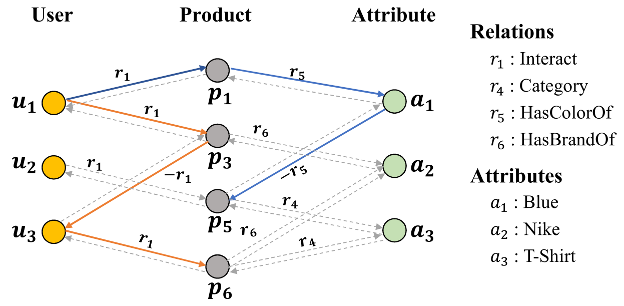


Fig. 2. Illustration of high-order connectivities.

Specifically, for each triplet (h, r, t) in $\mathcal{T}$, i.e, the relation-oriented graph, we first project the head entity h and tail entity t into the corresponding relation space as follows,

$$\mathbf{e}_{h_r} = \mathbf{M}_r \mathbf{e}_h, \ \mathbf{e}_{t_r} = \mathbf{M}_r \mathbf{e}_t, \tag{2}$$

where $\mathbf{e}_h \in \mathbb{R}^d$ and $\mathbf{e}_t \in \mathbb{R}^d$ represent the to-be-learned original embeddings of entities h and t, respectively. d is the embedding dimension. $\mathbf{M}_r \in \mathbb{R}^{d \times d}$ denotes the corresponding projection matrix of the relation type $r \in \mathcal{R}_a$. $\mathbf{e}_{h_r} \in \mathbb{R}^d$ and $\mathbf{e}_{t_r} \in \mathbb{R}^d$ are the projected representations of the head and tail entities in the latent space of the relation type r, respectively.

We then build a training set $\mathcal{Q} = \{(h, r, t, t')|(h, r, t) \in \mathcal{T}, (h, r, t') \notin \mathcal{T}\}$, where $t' \neq t$ is a negative tail entity regarding the head entity h and relation r, randomly sampled from the whole entity set $\mathcal{E}_a$. Based on the training set, we use the Bayesian ranking loss function for entity embedding of the relation-oriented graph as follows,

$$\begin{cases} \mathcal{L}_{GE} = \sum_{(h,r,t,t') \in \mathcal{Q}} \ln \sigma \left(g(h, r, t) - g\left(h, r, t'\right)\right) + \lambda \|\mathbf{\Theta}_1\|_F^2, \\ g(h, r, t) = \|\mathbf{e}_{h_r} + \mathbf{e}_r - \mathbf{e}_{t_r}\|_2^2, \end{cases} \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function, and $\mathbf{e}_r \in \mathbb{R}^d$ represents the to-be-learned embedding of relation type r. λ is a non-negative hyperparameter and $\mathbf{\Theta}_1$ refers to the parameters $\{\mathbf{M}_{r_1}, \mathbf{M}_{r_2}, \ldots, \mathbf{M}_{r_{2(F+1)}}\}$ of the graph embedding. The L_2 norm regularization on $\mathbf{\Theta}_1$ is introduced to prevent overfitting. Essentially, the lower the value of $g(h, r, t)$, the more likely the triplet exists in the graph. Then the BPR loss function encourages the positive triplet (h, r, t) to get a lower value than the negative ones. Ultimately, by optimizing the above objective function, we can obtain the embeddings for all entities (i.e, user entities, item entities, and attribute entities) and relations.

2) Relation-Aware Attentive Propagation: After obtaining the entity embedding and relation embedding of the relation-oriented graph $\mathcal{G}_a$, we proceed to the information propagation over it to uncover the user's and item's attribute-enhanced latent representation. Intuitively, one user tends to prefer items that share similar properties or have been bought by users with similar tastes, which can be reflected by the high-order connectivities in the relation-oriented graph. As illustrated by Fig. 2, based on the high-order connectivities: $u_1 \xrightarrow{r_1} p_1 \xrightarrow{r_5} a_1(black) \xrightarrow{-r_5} p_5$

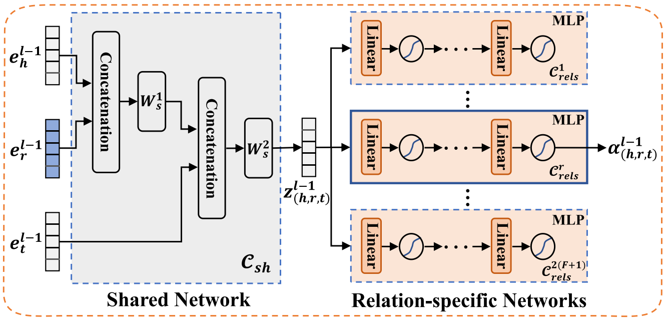


Fig. 3. Deep multi-task learning based confidence learning for each triplet.

and $u_1 \xrightarrow{r_1} p_3 \xrightarrow{-r_1} u_3 \xrightarrow{r_1} p_6$, where $r_1, r_5, -r_1$ and $-r_5$ refer to the relations "*Interact*", "*HasColorOf*", "*InteractBy*" and "*IsColorOf*", respectively, we can infer that the user u_1 may be interested in items p_5 and p_6. In light of this, we argue that it is promising to explore the high-order connectivity among entities in the relation-oriented graph to uncover the user's attribute-enhanced latent preference. Toward this end, similar to existing works, we explore the high-order connectivities among entities by conducting multiple iterations of the 1-order (i.e, one-hop) information propagation. Therefore, we describe the 1-order information propagation as an example, while the high-order information propagation can be derived iteratively.

During the 1-order information propagation, each node refines its embedding by comprehensively summarizing information from its ego-network [22], i.e, its one-hop neighbors. Formally, we regard each node as the head entity h, and define its ego-network as $\mathcal{T}_h = \{(h,r,t)|(h,r,t) \in \mathcal{T}\}$. Moreover, as a matter of fact, different attributes contribute differently in expressing the item's features, and different items also act differently in characterizing the user's latent preference. Consequently, we adopt the attention mechanism to dynamically assign the confidence of entity nodes in one's ego-network during the information propagation. Notably, even for the same head-tail entity pair, e.g., ($item1$, $Middle\_length$), there can be various relations, like "sleeve_length" and "dress_length", and the specific relation should also affect the confidence of the neighbor node (i.e, tail entity) influencing the central head entity. Accordingly, we propose to utilize all the three elements of each triplet to measure the confidence of each tail entity in the ego-network towards the central head entity representation learning.

Different from existing studies that measure the triplet confidence with the simple linear transformation, we employ the neural networks to capture the latent complex interactions among the three elements of a triplet. Moreover, considering the unique attributes that being possessed by limited samples, we adopt the multi-task learning framework to learn the triplet confidence. Specifically, we regard the confidence assignment of triplets with different relation types as a set of correlated tasks. The underlying philosophy is that there should be some common knowledge on the triplet confidence assignment regardless of relation types. In particular, as illustrated in Fig. 3, we employ a shared network and multiple relation-specific networks to learn the confidence for each triplet in the node's ego-network. The shared network works on learning the latent representation of the triplet that encodes the common knowledge on the triplet confidence assignment, while the multiple relation-specific networks target at learning the relation-specific triplet confidence. Both shared network and relation-specific networks are constructed via the multi-layer perceptron (MLP).

Specifically, we devise the shared network with two fully-connected layers, where the first layer aims to project the head and the relation embeddings into the common tail entity space, while the second one serves to learn the latent interaction among the three elements in the common space. Formally, we have,

$$\begin{aligned} \mathbf{z}^{l-1}_{(h,r,t)} &= \mathcal{C}_{sh}(\mathbf{e}_h^{l-1}, \mathbf{e}_r^{l-1}, \mathbf{e}_t^{l-1}) \\ &= \tau\left(\mathbf{W}_s^2\left(\tau\left(\mathbf{W}_s^1\left(\mathbf{e}_h^{l-1}\|\mathbf{e}_r^{l-1}\right) + \mathbf{b}_s^1\right)\|\mathbf{e}_t^{l-1}\right) + \mathbf{b}_s^2\right), \end{aligned} \tag{4}$$

where $||$ is the concatenation operation. $\mathbf{e}_h^{l-1} \in \mathbb{R}^d$, $\mathbf{e}_r^{l-1} \in \mathbb{R}^d$, and $\mathbf{e}_t^{l-1} \in \mathbb{R}^d$ denote the embedding of the head entity, relation, tail entity of the triplet (h,r,t) during the l-th propagation, $l = \{1, 2, \ldots, L\}$, respectively. L is the total number of the propagation iterations. In particular, $\mathbf{e}_h^0$, $\mathbf{e}_r^0$, and $\mathbf{e}_t^0$ are obtained by the aforementioned relation-oriented graph embedding. $\mathbf{W}_s^1 \in \mathbb{R}^{2\,d\times d}$, and $\mathbf{W}_s^2 \in \mathbb{R}^{2\,d\times d}$ are the weight matrices of the shared neural network $\mathcal{C}_{sh}$, while $\mathbf{b}_s^1 \in \mathbb{R}^d$, and $\mathbf{b}_s^2 \in \mathbb{R}^d$ are the bias vectors of $\mathcal{C}_{sh}$. $\tau(\cdot)$ refers to the tanh active function [23]. $\mathbf{z}^{l-1}_{(h,r,t)}$ denotes the learned latent representation of the triplet by the shared network in the l-th propagation.

Afterward, we feed the learned latent representation of each triplet to a relation-specific network according to its relation type. Ultimately, we obtain the triplet confidence as follows,

$$\alpha^{l-1}_{(h,r,t)} = \mathcal{C}^r_{rels}\left(\mathbf{z}^{l-1}_{(h,r,t)}|\mathbf{\Theta}_r^{l-1}\right), \tag{5}$$

where $\mathcal{C}^r_{rels}$ is the relation-specific network corresponding to the relation type r, composed by a MLP with Q layers. $\mathbf{\Theta}_r^{l-1}$ is the to-be-learned parameters. $\alpha^{l-1}_{(h,r,t)}$ is the confidence for the triplet (h,r,t) in the ego-network $\mathcal{T}_h$ during the l-th propagation, indicating how much information would be propagated from the tail entity t to the head entity h conditioned on the relation type r during the l-th propagation.

Thereafter, we can fulfil the information propagation and obtain each node's ego-network representation as follows,

$$\mathbf{e}^{l-1}_{\mathcal{T}_h} = \sum_{(h,r,t)\in\mathcal{T}_h} \alpha^{l-1}_{(h,r,t)} \cdot \mathbf{e}_t^{l-1}. \tag{6}$$

Ultimately, due to the remarkable capability of the Bi-Interaction aggregator function in aggregating information [13], we adopt it to aggregate the representations of the node self and its ego-network as the refined node representation after the l-th propagation as follows,

$$\mathbf{e}_h^l = \omega(\mathbf{W}_1^l\left(\mathbf{e}_h^{l-1} + \mathbf{e}^{l-1}_{\mathcal{T}_h}\right)) + \omega\left(\mathbf{W}_2^l(\mathbf{e}_h^{l-1} \odot \mathbf{e}^{l-1}_{\mathcal{T}_h})\right), \tag{7}$$

where $\omega(\cdot)$ is the LeakyReLU active function, and $\odot$ denotes the element-wise multiplication operation. $\mathbf{W}_1^l$ and $\mathbf{W}_2^l \in \mathbb{R}^{d_l \times d_{l-1}}$ are the weight matrices, where $d_l = d_{l-1}/2$, and $d_0 = d$.

After L propagation iterations, we can derive L embeddings for each entity h correspondingly, i.e, $\{\mathbf{e}_h^1, \mathbf{e}_h^2, \ldots, \mathbf{e}_h^L\}$. Ultimately, we concatenate all these representations with the initial embedding into a single vector $\mathbf{e}_h^a \in \mathbb{R}^{(d_0+\cdots+d_L)}$ to get the final representations of the entity h as follows,

$$\mathbf{e}_h^a = \mathbf{e}_h^0 \left\|\mathbf{e}_h^1\right\| \cdots \|\mathbf{e}_h^L. \tag{8}$$

As the head entity h can be a user entity, an item entity, or an attribute entity, for clarity, we further use $\mathbf{e}_u^a$, $\mathbf{e}_p^a$, and $\mathbf{e}_a^a$ to denote the final latent representation of the user entity u, the item entity p, the attribute entity a, derived according to (8), respectively.

B. Visual Representation Learning

Apart from the attribute-enhanced latent representation learning, we also introduce the visual representation learning to model the user's visual preference. In particular, we first employ the pre-trained visual representation learning model ResNet50 [24], which has been widely used for visual feature extraction [25], [26], [27], [28], to generate the initial visual representation of each item image. Let $\tilde{\mathbf{e}}_p \in \mathbb{R}^{d_i}$ be the initial visual representation of the item p, where $d_i = 2048$. We then introduce a MLP with two fully-connected layers to reduce the dimension of the representation of each item image and realize the fine-tuning as follows,

$$\begin{aligned}\mathbf{e}_p^v &= \mathcal{H}^v\left(\tilde{\mathbf{e}}_p \mid \boldsymbol{\Theta^v}\right)\\ &= \tau\left(\mathbf{W}_v^2\left(\tau\left(\mathbf{W}_v^1 \tilde{\mathbf{e}}_p + \mathbf{b}_v^1\right)\right) + \mathbf{b}_v^2\right)\end{aligned} \tag{9}$$

where $\mathbf{e}_p^v \in \mathbb{R}^{d_v}$ is the fine-tuned visual representation of the item p, and $\boldsymbol{\Theta^v}$ refers to the parameters of the MLP $\mathcal{H}^v$. Specifically, $\mathbf{W}_v^1 \in \mathbb{R}^{\frac{d_i}{4} \times d_i}$ and $\mathbf{W}_v^2 \in \mathbb{R}^{d_v \times \frac{d_i}{4}}$ are the weight matrices of the MLP $\mathcal{H}^v$, while $\mathbf{b}_v^1 \in \mathbb{R}^{\frac{d_i}{4}}$ and $\mathbf{b}_v^2 \in \mathbb{R}^{d_v}$ are the bias vectors of $\mathcal{H}^v$. $\tau(\cdot)$ refers to the tanh active function. To facilitate the following multi-modal enhanced preference learning, we make the dimension of visual representation equal to that of the attribute-enhanced latent representation $\mathbf{e}_p^a$, i.e, $d_v = (L+1)d$.

Due to the lack of direct visual cues of the user, we resort to the visual representations of images in the user entity's ego-network in graph $\mathcal{G}_v$ to obtain the user's visual embedding. In particular, we first define the user's visual ego-network as the set of triplets whose head entities are the user entity u in $\mathcal{Z}$, denoted as $\mathcal{Z}_u = \{(u, r', v) | (u, r', v) \in \mathcal{Z}\}$. Thereafter, we obtain the user's visual representation as follows,

$$\mathbf{e}_u^v = \frac{1}{|\mathcal{Z}_u|} \sum_{(u,r,v) \in \mathcal{Z}_u} \mathbf{e}_p^v, \tag{10}$$

where v is the image of the item p, and $\mathbf{e}_u^v \in \mathbb{R}^{d_v}$ is the visual embedding for user u.

C. Multi-Modal Enhanced Preference Learning

To comprehensively capture the user's preference, we fuse the user's and item's visual representation with their corresponding attribute-enhanced latent representations. Specifically, we have,

$$\mathbf{e}_u^f = \mathbf{e}_u^a \odot \mathbf{e}_u^v,\ \mathbf{e}_p^f = \mathbf{e}_p^a \odot \mathbf{e}_p^v, \tag{11}$$

Algorithm 1: The Training Procedure of MM-FRec.

```
Input: relation-oriented graph $\mathcal{G}_a$, vision-oriented graph
  $\mathcal{G}_v$, training set $\mathcal{Q}$ for the tripartite graph embedding and
  training set $\mathcal{D}$ for the user preference learning, image set
  $\mathcal{V}$, and mini-batch size $m$. Hyperparameters: $\{\lambda, \beta, \mu, \eta\}$
Output: Parameters $\{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2\}$.
Initialization
Initialize all parameters: $\{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2\}$.
repeat
  while an epoch is not end do
    Randomly sample a mini-batch of $(h, r, t, t')$'s from
    the training set $\mathcal{Q}$.
    Update $\boldsymbol{\Theta}_1$ by optimizing the loss function in (3).
  end while
  while an epoch is not end do
    Randomly sample a mini-batch of $(u, p, p')$'s from
    the training set $\mathcal{D}$.
    Update $\boldsymbol{\Theta}_2$ by optimizing the loss function in (16).
  end while
until MM-FRec converges
```

where $\odot$ denotes the element-wise multiplication operation. $\mathbf{e}_u^f$ and $\mathbf{e}_p^f$ denote the multi-modal enhanced representation of the user u and item p, respectively.

Ultimately, we employ the inner product between the multi-modal enhanced representations of a user u and an item p to measure the user u's preference to item p as follows,

$$\hat{y}_{u,p} = (\mathbf{e}_u^f)^T \mathbf{e}_p^f, \tag{12}$$

where $\hat{y}_{u,p}$ represents the predicted preference of the user u to item p.

Intuitively, the visual image and semantic attributes characterize the same fashion item, and should share certain latent consistency. In other words, the visual representation and the attribute-enhanced latent representation of the same item should be more similar than that of different items. Therefore, we utilize the contrastive loss as a visual-semantic consistency regularizer, which has shown superior performance in characterizing the inter-modal consistency and bridges the semantic gap between the low-level visual clues and high-level attribute semantics [29]. Formally, we have,

$$\begin{aligned}\mathcal{L}_{VSE} = \sum_{i \neq j} \Big(& \max\left\{0, \beta - \cos\left(\mathrm{e}_{p_i}^a, \mathrm{e}_{p_i}^v\right) + \cos\left(\mathrm{e}_{p_i}^a, \mathrm{e}_{p_j}^v\right)\right\}\\ & + \max\left\{0, \beta - \cos\left(\mathrm{e}_{p_i}^v, \mathrm{e}_{p_i}^a\right) + \cos\left(\mathrm{e}_{p_i}^v, \mathrm{e}_{p_j}^a\right)\right\}\Big),\end{aligned} \tag{13}$$

where β is a margin hyperparameter and $\cos(\cdot, \cdot)$ refers to the cosine similarity operator. $\mathrm{e}_{p_i}^a$ and $\mathrm{e}_{p_i}^v$ are the attribute-enhanced latent representation and the visual embedding of the item entity p_i, respectively.

Optimization. For optimization, we adopt the widely-used BPR loss, which has been proven to be powerful in the pairwise implicit preferences modeling [30], [31], [32]. Specifically, we first construct the following training set $\mathcal{D}$=

$$\left\{(u, p, p') \mid u \in \mathcal{U},\ p, p' \in \mathcal{P}, (u, r, p) \in \mathcal{T}, (u, r, p') \notin \mathcal{T}\right\}, \tag{14}$$

where the negative item p' was randomly sampled from $\mathcal{T}$, and r refers to the "*interact*" relation. We then have the BPR loss function as follows,

$$\mathcal{L}_{BPR} = -\sum_{(u,p,p')\in\mathcal{D}} \ln \sigma\left(\widehat{y}_{(u,p)} - \widehat{y}_{(u,p')}\right) + \mu \|\mathbf{\Theta}_2\|_F^2, \tag{15}$$

where μ is the non-negative hyperparameter. The last term is introduced to avoid overfitting and $\mathbf{\Theta}_2$ refers to the set of model parameters (i.e, $\{\mathbf{W}_s^1, \mathbf{b}_s^1, \mathbf{W}_s^2, \mathbf{b}_s^2, \mathbf{W}_1^l, \mathbf{W}_2^l, \mathbf{\Theta}_r^l, \mathbf{\Theta}^v\}$). Intuitively, we expect that a user's preference to the positive item should be larger than that to the negative item.

Finally, we have the following objective function,

$$\mathcal{L} = \mathcal{L}_{BPR} + \eta \mathcal{L}_{VSE}, \tag{16}$$

where η is the non-negative hyperparameter.

During the training phase, similar to KGAT [13], we optimize $\mathcal{L}$ and $\mathcal{L}_{GE}$ alternatively. The overall procedure of the optimization is briefly summarized in Algorithm 1. As can be seen, in each run, we first optimize the entity embeddings of the user-item-attribute graph and the projection matrices $\mathbf{M}_r$'s of TransR according to (3), and then optimize the other parameters of our model, including those in the relation-aware attentive propagation, visual representation learning, and multi-modal enhanced preference learning, according to (16).

V. Experiments

In this section, we first present the dataset and experimental setting, and then introduce the extensive experiments that we conducted on a real-world dataset by answering the following research questions:

RQ1. Does the proposed MM-FRec outperform state-of-the-art methods?

RQ2. What is the effect of each component on the performance of our MM-FRec?

RQ3. How is the sensitivity of our model?

RQ4. What is the intuitive performance of our model?

A. Dataset

Towards the fashion item recommendation, there have been several public benchmark datasets, such as POG [33], Amazon [1], and Tradesy [2]. Nevertheless, most of them (e.g., Tradesy) lack the attribute information of items. Although POG provides a set of attributes for each item, it lacks the corresponding attribute types. In addition, the Amazon dataset only contains the item's size attribute and color attribute. These deficiencies lead to the limitations of their applications in our context. Therefore, we resorted to the dataset IQON3000 [34], which was originally constructed for the personalized outfit recommendation. To be specific, IQON3000 comprises 308,747 outfits of 3,568 users, involving 672,335 items. Each item is associated with a visual image, and several fine-grained attributes. In particular, in total, there are 11 attribute types, including *color*, *price*, *brand*, *category*, *variety*, *material*, *pattern*, *design*, *heel*, *dress_length*, and *sleeve_length*.

TABLE I
Statistics of IQON10

#Users	#Items	#User-item
3,568	23,363	459,146
#Attribute Types	**#Attribute Values**	**#Item-attribute**
11	3,019	153,020

TABLE II
Attribute Distribution of Different Types in IQON10

Attribute Type	#Triplets	Attribute Type	#Triplets
Color	46,292	Pattern	2,371
Price	23,363	Design	1,624
Brand	23,360	Heel	929
Category	23,334	Dress_length	550
Variety	22,929	Sleeve_length	518
Material	7,750		

To adapt IQON3000 to our task of fashion item recommendation, we treated items in an outfit composed by a user as the positive items for this user. We argue that if a user publicly shares an outfit, then he/she should prefer the composing items of the outfit. To ensure the quality of the dataset, we filtered out the items that are preferred by less than 10 users. Ultimately, the final derived dataset, named as IQON10, comprises 3,568 users, 23,363 fashion items and 459,146 user-item interactions. In addition, there are 3,019 attribute values, and 153,020 item-attribute association relations. Table I summarizes the statistics of our dataset. In addition, Table II shows the detailed statistics for different attribute types. As can be seen, the number of the triplets with different attribute types varies in a large range. Concretely, a large number of items have the "color" attribute, while limited items have the attributes of "Dress_length" and "Sleeve_length", which confirms the existence of the unique attributes in practice.

B. Experimental Settings

In this part, we present the evaluation metrics, baselines, and implementation details.

Evaluation Metrics. We evaluated our method with the task of Top-K recommendation [35], [36], and adopted three widely-used evaluation metrics [37], [38]: *Recall*@K, $NDCG@K$ and $Precision@K$. In this work, we set $K = 20, 50, 100$, respectively.

Baselines. To prove the effectiveness of MM-FRec, we compared the proposed MM-FRec with five state-of-the-art baselines.

- *CFKG* [20]: This model only relies on the user-item-attribute graph to model the user's preference, which adopts TransE [39] to embed the heterogeneous entities in the user-item-attribute graph.
- *KGAT* [13]: This model explores the high-order connectivity in the collaborative knowledge graph, i.e, the user-item-attribute graph in our context, with graph neural network and the attention mechanism. Notably, this method overlooks the user's visual preference modeling.

TABLE III
PERFORMANCE COMPARISON AMONG DIFFERENT APPROACHES. THE NUMBERS IN BOLD INDICATE THE BEST RESULTS, WHILE THOSE UNDERLINED REFER TO THE SECOND BEST RESULTS. * DENOTES THAT THE p-VALUE OF THE SIGNIFICANT TEST BETWEEN OUR RESULT AND THE BEST BASELINE RESULT IS LESS THAN 0.01. "R":"RECALL"; "P": "PRECISION"; "IMPROV.": "PERFORMANCE IMPROVEMENT"

Model	R@20	NDCG@20	P@20	R@50	NDCG@50	P@50	R@100	NDCG@100	P@100
CFKG	0.1860	0.0940	0.0172	0.2837	0.1154	0.0122	0.3589	0.1383	0.0097
KGAT	0.1643	0.0900	0.0199	0.2287	0.1011	0.0134	0.3081	0.1204	0.0109
JNSKR	0.2073	0.1005	0.0116	0.2850	0.1462	0.0083	0.3126	0.1543	0.0073
VBPR	0.0932	0.0505	0.0092	0.1243	0.0769	0.0062	0.1506	0.0821	0.0045
SAERS	0.1863	0.1252	0.0243	0.2202	0.1373	0.0178	0.2479	0.1460	**0.0167**
A-FRec	0.2071	0.1201	0.0274	0.2941	0.1441	0.0192	0.3662	0.1634	0.0146
MM-FRec	**0.2317***	**0.1486***	**0.0312***	**0.3157***	**0.1732***	**0.0214***	**0.3857***	**0.1926***	0.0161*
Improv.	11.77%↑	18.69%↑	28.40%↑	10.77%↑	18.47%↑	20.22%↑	7.47%↑	24.82%↑	-

- *JNSKR* [19]: This method uses an efficient Non-Sampling (NS) optimization algorithm for the user-item-attribute graph embedding, which is able to train the model by the whole training data with a low time complexity. This method also ignores the visual modality of items.
- *VBPR* [2]: This is a multi-modal based model, which incorporates the visual modality of fashion items to enhance the factorization-based user preference modeling.
- *SAERS* [1]: This is also a multi-modal based approach that characterizes the preference of users from both the visual and semantic attribute perspectives. Notably, since our IQON10 contains specific attributes of fashion items, we directly adopted these ground-truth attribute labels without pre-training an additional attribute extraction network in the reproduction of this model.

Implementation Details. We randomly sampled 80%, 10%, and 10% of the positive user-item interactions in our dataset to form the training set, validation set and testing set, respectively. Regarding the user-item-attribute graph, there are 24 relation types, corresponding to the user-item interaction relation type and 11 attribute types, as well as their reverse ones. Accordingly, for the relation-aware attentive propagation, apart from the shared network, we deployed 24 independent relation-specific networks for the confidence assignment toward nodes in the ego-network. The size of item images is unified to 150 × 150. Pertaining to the optimization, we utilized the adaptive moment estimation method (Adam) [40], and all the parameters in the neural networks are initialized by Xavier initializer [41]. The grid search strategy is adopted to determine the optimal values for the regularization hyperparameters λ and μ among values $\{10^r|r \in \{-5, \cdots, -2\}\}$. In addition, the learning rate lr, the embedding dimension d, the number of layers in the relation-specific network Q, the number of iterations of information propagation L, the margin of contrastive loss β, and the hyperparameter for contrastive loss η are searched in $[0.001, 0.0005, 0.0001]$, $[16, 32, 64]$, [1,2,3,4], [1,2,3,4], $\{10^r|r \in \{-5, \cdots, 1\}\}$, and $[0.1, 0.5, 1, 5, 10]$, respectively.

Finally, we empirically found that the proposed model achieves the optimal performance with regularization hyperparameters $\lambda = \mu = 10^{-5}$, the learning rate $lr = 0.0005$, the embedding dimension $d = 64$, the number of layers in the relation-specific network $Q = 3$, the number of iterations of information propagation $L = 3$, the margin of contrastive loss $\beta = 10^{-3}$, and the hyperparameter for contrastive loss $\eta = 5$.

C. On Model Comparison (RQ1)

Table III shows the performance of different methods in the Top-K recommendation task. To make a fair comparison with baseline methods that only adopt the attribute information of each item, we also derived A-FRec from our MM-FRec by removing the visual modality. From Table III, we have the following observations. 1) Our MM-FRec outperforms all the other baselines in the two recommendation tasks on almost all evaluation metrics. In particular, compared with the best baseline, MM-FRec achieves a significant improvement of 28.40%, 20.22%, 24.82% in Precision@20, Precision@50, and NDCG@100, respectively. This demonstrates the superiority of our MM-FRec and reflects that our model tends to rank the positive items at the top places, as compared to the baseline methods. 2) Notably, we also conducted a pairwise significant test, and the results verify that all the improvements of MM-FRec are statistically significant with p-value < 0.01. 3) MM-FRec surpasses A-FRec in all cases, confirming the advantage of considering the visual content of items in the fashion item recommendation task. And 4) even removing the visual modality, our A-FRec still outperforms KGAT and JNSKR regarding almost all metrics, which indicates the effectiveness of our relation-aware attentive propagation.

D. On Ablation Study (RQ2)

In the ablation study, we investigated the three key components of our model: relation-aware attention mechanism, visual-semantic consistency regularizer, and multi-modal fusion manner.

1) Effect of Relation-Aware Attention Mechanism: As aforementioned, the proposed relation-aware attention mechanism simultaneously considers the three elements of each triple, i.e, the head entity, the relation, and the tail entity, to assign the node confidence. Meanwhile, considering the existence of unique attributes, we devised a shared network and multiple relation-specific networks for the attention calculation. Accordingly, to comprehensively evaluate the relation-aware attention mechanism in our model, we introduced the following five variant models of our MM-FRec:

TABLE IV
EFFECT OF THE RELATION-AWARE ATTENTION MECHANISM ON OUR MODEL

Model	Recall@20	NDCG@20	Precision@20
A-FRec-NoRA	0.1850	0.1068	0.0247
A-FRec-NoSH	0.1988	0.1169	0.0264
A-FRec-NoRS	0.1973	0.1148	0.0263
A-FRec-Linear	0.1754	0.0983	0.0206
A-FRec-NoR	0.1981	0.1168	0.0263
A-FRec	**0.2056**	**0.1205**	**0.0268**
MM-FRec-NoRA	0.2128	0.1327	0.0282
MM-FRec-NoSH	0.2281	0.1426	0.0302
MM-FRec-NoRS	0.2310	0.1468	0.0308
MM-FRec-Linear	0.1998	0.1210	0.0272
MM-FRec-NoR	0.2085	0.1238	0.0272
MM-FRec	**0.2317**	**0.1486**	**0.0312**

TABLE V
EFFECT OF VISUAL-SEMANTIC CONSISTENCY REGULARIZER ON OUR MODEL

Model	Recall@20	NDCG@20	Precision@20
MM-FRec-NoVSE	0.2215	0.1312	0.0299
MM-FRec	**0.2317**	**0.1486**	**0.0312**

- *MM-FRec-NoRA*: The whole relation-aware attention mechanism is disabled.
- *MM-FRec-NoSH*: The shared network of our model is removed.
- *MM-FRec-NoRS*: The relation-specific networks of our model are deleted.
- *MM-FRec-Linear*: To justify the benefit of using neural networks, we introduced this method by replacing the neural networks with the linear transformations.
- *MM-FRec-NoR*: To verify the necessity of incorporating the relation element of a triplet in the confidence assignment, we removed the $\mathbf{e}_r^{l-1}$ from (4).

Table IV shows the performance of our MM-FRec and A-FRec with different attention configurations. From this table, we have the following observations. 1) MM-FRec-NoRA and A-FRec-NoRA perform worse than our MM-FRec and A-FRec, respectively. This verifies the necessity of the dynamic triplet confidence assignment in the information propagation over the user-item-attribute tripartite graph. 2) MM-FRec and A-FRec outperform MM-FRec-NoSH (MM-FRec-NoRS) and A-FRec-NoSH (A-FRec-NoRS), respectively. This confirms the benefit of incorporating the deep multi-task learning strategy to handle the triplet confidence assignment of various attributes. 3) MM-FRec and A-FRec surpass MM-FRec-NoR and A-FRec-NoR, respectively. This verifies the necessity of introducing relation element in the confidence assignment during information propagation. 4) MM-FRec and A-FRec outperform MM-FRec-Linear and A-FRec-Linear, respectively. This proves the effectiveness of neural networks in learning the latent complex interactions among the three elements of a triplet towards the neighbor node's confidence assignment.

2) Effect of Visual-Semantic Consistency Regularizer: In this part, we aim to examine the effect of the visual-semantic consistency regularizer on our model. Therefore, we removed the contrastive loss function from (16) by setting $\eta = 0$ and obtained a variant of MM-FRec, termed as MM-FRec-NoVSE. Table V shows the performance comparison between MM-FRec-NoVSE and MM-FRec. As can be seen, MM-FRec shows superiority over MM-FRec-NoVSE across all metrics. This verifies the benefit of bridging the semantic gap between the low-level visual clues and high-level attribute semantics.

TABLE VI
PERFORMANCE OF DIFFERENT MODALITY FUSION METHODS

Model	Recall@20	NDCG@20	Precision@20
MM-FRec$_+$	0.2018	0.1190	0.0279
MM-FRec$_{\|\|}$	0.2106	0.1289	0.0273
MM-FRec	**0.2317**	**0.1486**	**0.0312**

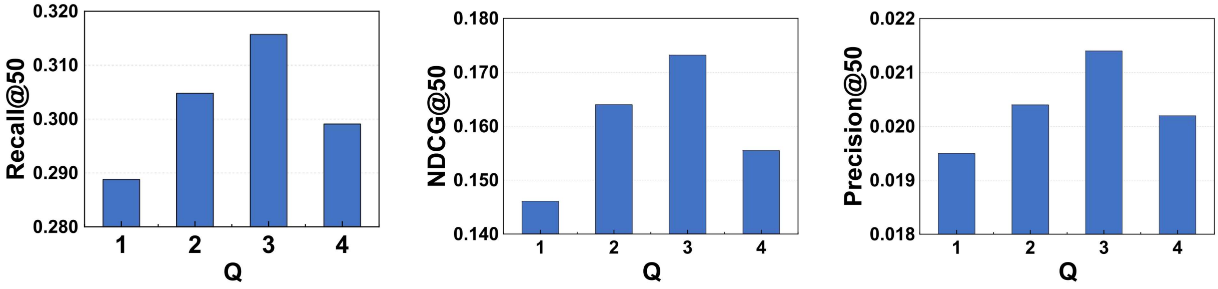
3) Effect of Modality Fusion Manner: In previous experiments on model comparison, we have demonstrated the advantage of utilizing the multiple modalities of items in the fashion item recommendation. In this part, we aim to further explore the effect of the multi-modal fusion method. In particular, we introduced two variant models of MM-FRec, termed as MM-FRec$_+$ and MM-FRec$_{||}$, which adopt the addition and concatenation fusion strategies in (11), respectively. Table VI shows the performance of our MM-FRec and its two variant models. It is obvious that our MM-FRec consistently outperforms the two variants over all evaluation metrics, implying that the element-wise product fusion strategy in MM-FRec is more powerful in our context of fashion item recommendation.

E. On Sensitivity Analysis (RQ3)

Besides, we studied the sensitivity of our model pertaining to the number of layers in the relation-specific neural networks, the number of propagation iterations, and the density of semantic attributes.

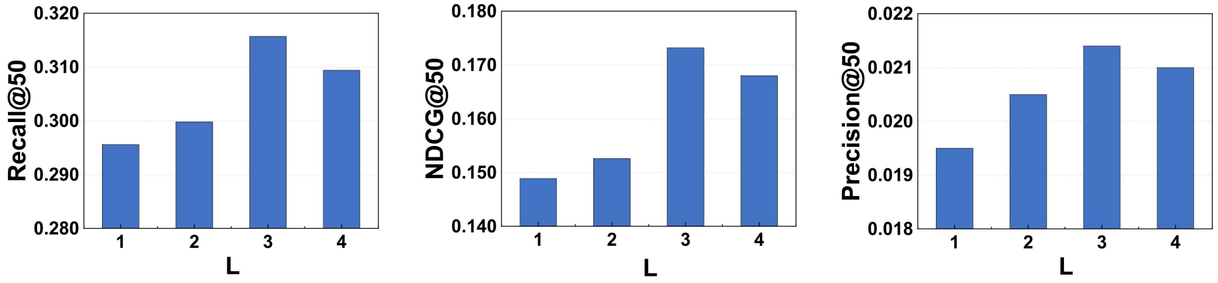
1) On the Number of Layers in the Relation-Specific Neural Networks: First, we varied the number of layers in the relation-specific neural networks, i.e, Q, in the range of $\{1, 2, 3, 4\}$. Fig. 4 shows the performance of our model with different Q in terms of Recall@50, NDCG@50, and Precision@50. As we can see, with the increase of the number of layers, the performance of our model increases first and then decreases. Overall, our model achieves the optimal performance when the number of layers in each relation-specific neural network is 3. When the number of layers is larger than 3, the model's performance begins to drop. The reason may be that too many layers may lead to the overfitting issue.

2) On the Number of Propagation Iterations: To learn the impact of the high-order connectivity, we changed the number of propagation iterations, i.e, L, from 1 to 4. Fig. 5 illustrates the performance of our model with different L in terms of Recall@50, NDCG@50, and Precision@50. As we can see, similar to the previous experiment on Q, the performance of our model increases first and then decreases, with the increase of the number of propagation iterations. Our model performs best when $L = 3$. This demonstrates that the utility of the high-order connectivity information in the user and item latent representation learning. As for the performance drop of our model when $L = 4$, one likely reason is that with the richer information propagated with the increasing of L, certain noise can be also disseminated.



(a) (b) (c)

Fig. 4. Performance of our model with different numbers of layers in the relation-oriented network (i.e, Q).



(a) (b) (c)

Fig. 5. Performance of our model with different numbers of propagation iterations (i.e, L).

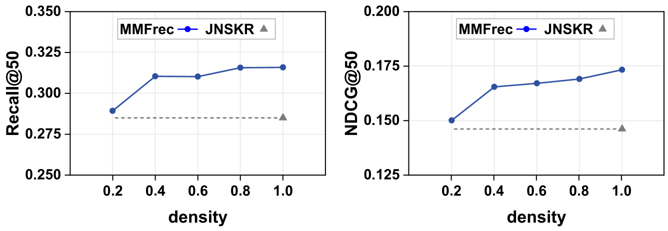


Fig. 6. Performance of our model with different dataset configurations, i.e, different densities of semantic attribute.

3) On the Density of Semantic Attributes: As aforementioned, semantic attributes of items play an important role in characterizing the item features and benefiting the user preference modeling. Nevertheless, in real-world applications, the semantic attributes of items may not be rich enough. Therefore, we evaluated our model on different dataset configurations, which have different attribute densities. In other words, different proportions of semantic attribute labels (i.e, item-attribute edges) in our original dataset were made available. In particular, we adopted five different proportions: 0.2, 0.4, 0.6, 0.8, and 1.0. Fig. 6 shows the performance of our model with different attribute density configurations in terms of Recall@50 and NDCG@50. As we can see, the more attribute tags we have, the better the performance of our model will be, which is in line with our expectation. In addition, although lacking attribute tags hurts the model performance, the degree of degradation is limited, reflecting a good tolerance of our model towards the density of semantic attributes. Notably, we also displayed the performance of the best baseline JNSKR with all attributes reserved by grey triangles in Fig. 6. As can be seen, our model with only 20% available semantic attributes can also surpass the best baseline that has 100% semantic attributes. This reconfirms the superiority of our model over existing methods.

F. On Case Study (RQ4)

To intuitively demonstrate the effectiveness of our model, apart from the quantitative evaluation, we also conducted the case study on both the item recommendation results and the recommendation interpretability.

To get the intuitive understanding of our MM-FRec in the task of item recommendation, we compared the recommendation results of MM-FRec and JNSKR for three testing users in Fig. 7. The reason why we chose JNSKR as the baseline lies in the fact that it is the best baseline for most metrics according to Table III. Due to limited space, we list the top 10 recommended items' images of MM-FRec and JNSKR, respectively. To facilitate the understanding of the results, we also randomly sampled 10 item images that are historically interacted by these users in the training set to display the users' historical habits. First, as can be seen, our model is able to return more ground-truth items that the user likes, highlighted in red boxes, than JNSKR, which reconfirms the superior recommendation performance of our model. Second, we noticed that our model is able to return items with unique attributes, highlighted in blue boxes, like the zebra pattern and round frame of glasses in the first case. One possible explanation for this observation is that MM-FRec copes well with the unique attributes, and hence promotes the recommendation performance. Similar conclusions can be found in the second and third cases.

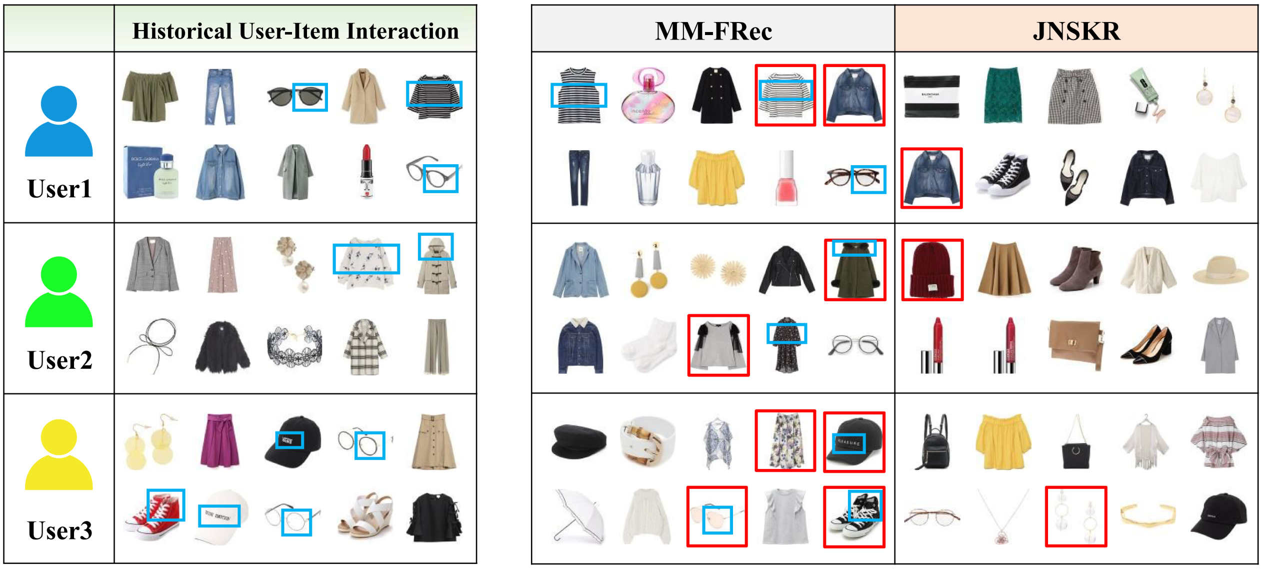


Fig. 7. Illustration of the returned Top-*10* recommendation results for three testing users of MM-FRec and JNSKR, where ground-truth images are highlighted by red boxes. Meanwhile, certain unique attributes captured by our model are also highlighted by blue boxes.

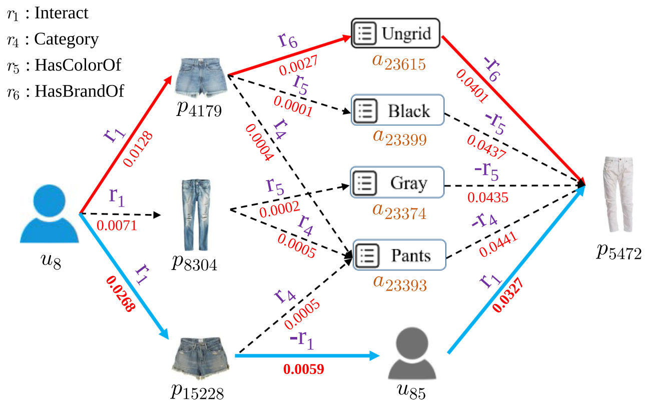


Fig. 8. Illustration of a recommendation result. All the paths are the accessible connections between the given user (i.e, u_8) and the recommended item (i.e, p_{5472}). The solid path in blue has the largest weight, while that in red has the second largest weight.

To demonstrate the interpretability of our model, we randomly sampled a user-item pair, i.e, (u_8, p_{5472}), from the testing set, which is correctly predicted by our MM-FRec. We illustrated all accessible connections from u_8 to p_{5472} in the user-item-attribute tripartite graph $\mathcal{G}_a$ in Fig. 8. Notably, we omitted some weak connections according to the attention weights for clear illustration. As can be seen from the example, there are multiple paths linking the user u_8 and item p_{5472}, all of which indicate that u_8 has a tendency to like p_{5472}. After going through all paths, we noticed that the path $u_8 \xrightarrow{r_1} p_{15228} \xrightarrow{-r_1} u_{85} \xrightarrow{r_1} p_{5472}$, labeled by the blue solid line, gets the highest attention weight. This enables us to explain the recommendation result by the fact that the item p_{5472} has been interacted by user u_{85}, who once expressed the same interest in the item p_{15228} with the user u_8. The other connection paths in Fig. 8 can provide the supplement explanation for recommending item p_{5472} to user u_8. For example, based on the path $u_8 \xrightarrow{r_1} p_{4179} \xrightarrow{r_6} a_{23615} \xrightarrow{-r_6} p_{5472}$, labeled by the red solid line, we can conclude that one reason why user u_8 would be interested in item p_{5472} is that the item p_{5472} has the same brand (i.e, Ungrid) with the item p_{4179}, which is previously liked by user u_8.

VI. Conclusion and Future Work

In this work, we present a Multi-Modal enhanced Fashion item Recommendation scheme, named MM-FRec, which jointly considers the visual images and semantic attributes of items to boost the recommendation performance. Meanwhile, we introduce a new relation-aware attention mechanism to adaptively measure how much information should be propagated from one's neighbor node according to not only the head and tail nodes but also the relation type. In particular, the multi-task learning framework is adopted to promote the triplet confidence assignment for unique attributes. Extensive experiments have been conducted on the real-world dataset and the results demonstrate the superiority of MM-FRec over existing methods and validate the benefits of jointly utilizing the attribute and vision modalities in the context of fashion item recommendation. In addition, experiments show that MM-FRec can not only provide explanations for the recommendation results according to the user-item-attribute tripartite graph, but also discover the user preference on unique attributes.

Currently, one limitation of our work is that we deem the user's preference temporally unchanged. However, in practice, the user's preference can be dynamically varied. Accordingly, in the future, we plan to incorporate the user's behavior trajectory to dynamically analyze user preferences for fashion items. Additionally, we are interested in integrating more clothing images of different views, such as front and back, and exploring more fine-grained visual preference mining methods to further improve the performance of fashion item recommendation.

References

[1] M. Hou, L. Wu, E. Chen, Z. Li, V. W. Zheng, and Q. Liu, "Explainable fashion recommendation: A semantic attribute region guided approach," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 4681–4688.

[2] R. He and J. J. McAuley, "VBPR: Visual Bayesian personalized ranking from implicit feedback," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 144–150.

[3] W. Yu, H. Zhang, X. He, X. Chen, L. Xiong, and Z. Qin, "Aesthetic-based clothing recommendation," in *Proc. ACM Int. Conf. World Wide Web Conf.*, 2018, pp. 649–658.

[4] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. Conf. Uncertainty Artif. Intell.*, 2009, pp. 452–461.

[5] J. J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in *Proc. Int. ACM SIGIR Conf. Res. Develop.Inf. Retrieval*, 2015, pp. 43–52.

[6] X. Chen et al., "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2019, pp. 765–774.

[7] C. Xu et al., "Recommendation by users' multimodal preferences for smart city applications," *IEEE Trans. Ind. Inform.*, vol. 17, no. 6, pp. 4197–4205, 2021.

[8] Z. Cheng, X. Chang, L. Zhu, R. C. Kanjirathinkal, and M. S. Kankanhalli, "MMALFM: Explainable recommendation by leveraging reviews and images," *ACM Trans. Inf. Syst.*, vol. 37, no. 2, pp. 16:1–16:28, 2019.

[9] M. Gori and A. Pucci, "ItemRank: A random-walk based scoring algorithm for recommender engines," in *Proc. Int. Joint Conf. Artif. Intell.*, 2007, pp. 2766–2771.

[10] X. He, M. Gao, M. Kan, and D. Wang, "BiRank: Towards ranking on bipartite graphs," *IEEE Trans.*, vol. 29, no. 1, pp. 57–71, Jan. 2017.

[11] Y. Cao, X. Wang, X. He, Z. Hu, and T. Chua, "Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences," in *Proc. ACM Int. Conf. World Wide Web Conf.*, 2019, pp. 151–161.

[12] F. Zhang, N. J. Yuan, D. Lian, X. Xie, and W. Ma, "Collaborative knowledge base embedding for recommender systems," in *Proc. Int. ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2016, pp. 353–362.

[13] X. Wang, X. He, Y. Cao, M. Liu, and T. Chua, "KGAT: Knowledge graph attention network for recommendation," in *Proc. Int. ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2019, pp. 950–958.

[14] X. Wang, X. He, M. Wang, F. Feng, and T. Chua, "Neural graph collaborative filtering," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2019, pp. 165–174.

[15] H. Wang, M. Zhao, X. Xie, W. Li, and M. Guo, "Knowledge graph convolutional networks for recommender systems," in *Proc. ACM Int. Conf. World Wide Web Conf.*, 2019, pp. 3307–3313.

[16] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "LightGCN: Simplifying and powering graph convolution network for recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 639–648.

[17] X. Huang, Q. Fang, S. Qian, J. Sang, Y. Li, and C. Xu, "Explainable interaction-driven user modeling over knowledge graph for sequential recommendation," in *Proc. Int. Conf. Multimedia*, 2019, pp. 548–556.

[18] Y. Liu et al., "Pre-training graph transformer with multimodal side information for recommendation," in *Proc. Int. Conf. Multimedia*, 2021, pp. 2853–2861.

[19] C. Chen, M. Zhang, W. Ma, Y. Liu, and S. Ma, "Jointly non-sampling learning for knowledge graph enhanced recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 189–198.

[20] Q. Ai, V. Azizi, X. Chen, and Y. Zhang, "Learning heterogeneous knowledge base embeddings for explainable recommendation," *Algorithms*, vol. 11, no. 9, 2018, Art. no. 137.

[21] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 2181–2187.

[22] J. Qiu, J. Tang, H. Ma, Y. Dong, K. Wang, and J. Tang, "DeepInf: Social influence prediction with deep learning," in *Proc. Int. ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2018, pp. 2110–2119.

[23] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–12.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[25] L. Sun, Z. Lian, J. Tao, B. Liu, and M. Niu, "Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism," in *Proc. Int. ACM Conf. Multimedia*, 2020, pp. 27–34.

[26] W. Liu, C. Zhang, G. Lin, T. Hung, and C. Miao, "Weakly supervised segmentation with maximum bipartite graph matching," in *Proc. Int. ACM Conf. Multimedia*, 2020, pp. 2085–2094.

[27] O. Gune, B. Banerjee, S. Chaudhuri, and F. Cuzzolin, "Generalized zero-shot learning using generated proxy unseen samples and entropy separation," in *Proc. Int. ACM Conf. Multimedia*, 2020, pp. 4262–4270.

[28] N. Pu, W. Chen, Y. Liu, E. M. Bakker, and M. S. Lew, "Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification," in *Proc. Int. ACM Conf. Multimedia*, 2020, pp. 2149–2158.

[29] X. Yang, X. Song, X. Han, H. Wen, J. Nie, and L. Nie, "Generative attribute manipulation scheme for flexible fashion search," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 941–950.

[30] X. Han, X. Song, J. Yin, Y. Wang, and L. Nie, "Prototype-guided attribute-wise interpretable scheme for clothing matching," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2019, pp. 785–794.

[31] D. Cao, L. Nie, X. He, X. Wei, S. Zhu, and T. Chua, "Embedding factorization models for jointly recommending items and user generated lists," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2017, pp. 585–594.

[32] J. Chen, C. Wang, and J. Wang, "Modeling the intransitive pairwise image preference from multiple angles," in *Proc. Int. ACM Conf. Multimedia*, 2017, pp. 351–359.

[33] W. Chen et al., "POG: Personalized outfit generation for fashion recommendation at alibaba iFashion," in *Proc. Int. ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2019, pp. 2662–2670.

[34] X. Song, X. Han, Y. Li, J. Chen, X. Xu, and L. Nie, "GP-BPR: Personalized compatibility modeling for clothing matching," in *Proc. Int. ACM Conf. Multimedia*, 2019, pp. 320–328.

[35] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T. Chua, "Neural collaborative filtering," in *Proc. ACM Int. Conf. World Wide Web*, 2017, pp. 173–182.

[36] Z. Zhou, X. Di, W. Zhou, and L. Zhang, "Fashion sensitive clothing recommendation using hierarchical collocation model," in *Proc. Int. ACM Conf. Multimedia*, 2018, pp. 1119–1127.

[37] C. Chen, M. Zhang, Y. Liu, and S. Ma, "Social attentional memory network: Modeling aspect- and friend-level differences in recommendation," in *Proc. Int. Conf. Web Search Data Mining*, 2019, pp. 177–185.

[38] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, E. J. Yannakoudakis, N. J. Belkin, P. Ingwersen, and M. Leong, Eds., 2000, pp. 41–48.

[39] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Conf. Workshop Neural Inf. Process. Syst.*, 2013, pp. 2787–2795.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[41] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

Xuemeng Song (Senior Member, IEEE) received the BE degree from the University of Science and Technology of China, in 2012, and the PhD degree from the School of Computing, National University of Singapore, in 2016. She is currently an associate professor with Shandong University, China. She has published several papers in the top venues, such as ACM SIGIR, MM, and TOIS. Her research interests include information retrieval and social network analysis. She has served as a reviewer for many top conferences and journals. She is also an AE of IEEE TCSVT and IET Image Processing.

Chun Wang received the BE degree from the School of Computer Science and Technology, Shandong University, Shandong, in 2019. He is currently working toward the graduation degree with the Department of Computer Science and Technology, Shandong University. His research interests include natural language generation, information retrieval and data mining.

Changchang Sun received the BE and MS degrees from Shandong University, China, in 2018 and 2021, respectively. She is currently working toward the PhD degree with the School of Computer Science, Illinois Institute of Technology, IL, USA. Her research interests include information retrieval and computer vision.

Shanshan Feng received the BS degree form the University of Science and Technology of China in 2011 and the PhD degree from Nanyang Technological University in 2017. He is currently an associate professor with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen). Prior to that, he was a research scientist with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. Before that, he was a senior researcher with Tencent company, and a research scientist with the Institute of High-performance Computing, Singapore. His current research interests include Big Data analytics, spatio-temporal data mining, artificial intelligence, and financial technologies.

Min Zhou (Member, IEEE) received the BS degree in automation from the University of Science and Technology of China, and the PhD degree from Industrial Systems Engineering and Management Department, National University of Singapore, respectively. She is currently a principal research engineer of Huawei Noah's Ark Lab, Shenzhen, China. Her interests include pattern mining and machine learning, and their applications in sequence and graph data. Her several works related to graph learning and mining were published at top conferences, including KDD, WWW, ICDE, and SIGIR.

Liqiang Nie (Senior Member, IEEE) received the BEng degree from Xi'an Jiaotong University and the PhD degree from the National University of Singapore (NUS), respectively. He is currently the dean with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen). His research interests lie primarily in multimedia computing and information retrieval. He has co-/authored more than 100 papers and 4 books, received more than 15,000 Google Scholar citations. He is an AE of *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Circuits and Systems for Video Technology*, *ACM Transactions on Multimedia Computing, Communications, and Applications*, and *Information Science*. Meanwhile, he is the regular area chair of ACM MM, NeurIPS, IJCAI and AAAI. He is a member of ICME steering committee. He has received many awards, like ACM MM and SIGIR best paper honorable mention in 2019, SIGMM rising star in 2020, TR35 China 2020, DAMO Academy Young Fellow in 2020, and SIGIR best student paper in 2021.