Attribute-wise Explainable Fashion Compatibility Modeling

XIN YANG and XUEMENG SONG, Shandong University, China
FULI FENG, National University of Singapore, Singapore
HAOKUN WEN, Shandong University, China
LING-YU DUAN, Peking University, China
LIQIANG NIE, Shandong University, China

With the boom of the fashion market and people's daily needs for beauty, clothing matching has gained increased research attention. In a sense, tackling this problem lies in modeling the human notions of the compatibility between fashion items, i.e., Fashion Compatibility Modeling (FCM), which plays an important role in a wide bunch of commercial applications, including clothing recommendation and dressing assistant. Recent advances in multimedia processing have shown remarkable effectiveness in accurate compatibility evaluation. However, these studies work like a black box and cannot provide appropriate explanations, which are indeed of importance for gaining users' trust and improving their experience. In fact, fashion experts usually explain the compatibility evaluation through the matching patterns between fashion attributes (e.g., a *silk* tank top cannot go with a *knit* dress). Inspired by this, we devise an attribute-wise explainable FCM solution, named *ExFCM*, which can simultaneously generate the item-level compatibility evaluation for input fashion items and the attribute-level explanations for the evaluation result. In particular, ExFCM consists of two key components: attribute-wise representation learning and attribute interaction modeling. The former works on learning the region-aware attribute representation for each item with the threshold global average pooling. Besides, the latter is responsible for compiling the attribute-level matching signals into the overall compatibility evaluation adaptively with the attentive interaction mechanism. Note that ExFCM is trained without any attribute-level compatibility annotations, which facilitates its practical applications. Extensive experiments on two real-world datasets validate that ExFCM can generate more accurate compatibility evaluations than the existing methods, together with reasonable explanations.

CCS Concepts: • **Information systems** → **Retrieval tasks and goals**; *World Wide Web*;

Additional Key Words and Phrases: Fashion analysis, explainable compatibility modeling, attribute-wise learning

This work is supported by the National Key Research and Development Project of New Generation Artificial Intelligence, No.:2018AAA0102502; the National Natural Science Foundation of China, No.:61772310, No.:61702300, and No.:U1936203; the Shandong Provincial Natural Science Foundation, No.:ZR2019JQ23; the Shandong Provincial Key Research and Development Program, No.:2019JZZY010118; the Innovation Teams in Colleges and Universities in Jinan, No.:2018GXRC014.

Authors' addresses: X. Yang, X. Song (corresponding author), H. Wen, and L. Nie (corresponding author), Shandong University, 72 Binhai Road, Qingdao, China, 266237; emails: {joeyangbuer, sxmustc, whenhaokun, nieliqiang}@gmail.com; F. Feng, National University of Singapore, Singapore; email: fulifeng93@gmail.com; L.-Y. Duan, Peking University, China; email: lingyu@pku.edu.cn.

1551-6857/2021/04-ART36 $15.00
https://doi.org/10.1145/3425636

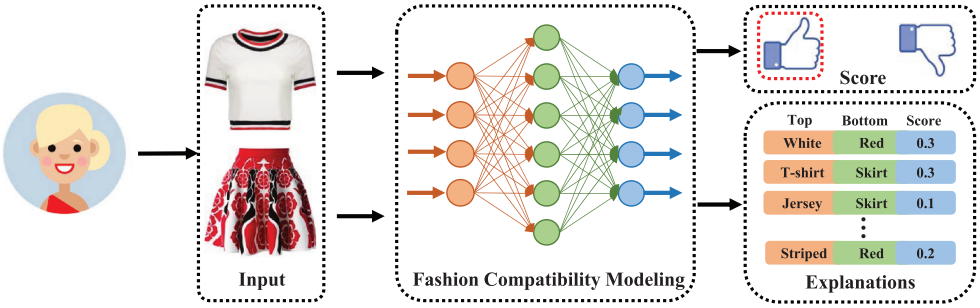


Fig. 1. Illustration of attribute-wise explainable FCM.

ACM Reference format:
Xin Yang, Xuemeng Song, Fuli Feng, Haokun Wen, Ling-Yu Duan, and Liqiang Nie. 2021. Attribute-wise Explainable Fashion Compatibility Modeling. *ACM Trans. Multimedia Comput. Commun. Appl.* 17, 1, Article 36 (April 2021), 21 pages.
https://doi.org/10.1145/3425636

1 INTRODUCTION

At present, fashion has been a flourishing industry with global production value up to three trillion dollars,[1] which demonstrates people's great demand for clothing [31]. In fact, clothing plays an essential role in people's daily lives, since a harmonious (compatible) outfit can improve the personal appearance instantly. Nevertheless, matching complementary clothes (e.g., the top, bottom, and shoes) and making proper outfits have been daily troubles for many people, especially those who have a poor sense of aesthetics. In response to this, Fashion Compatibility Modeling (FCM) that assesses the compatibility score for a given set of complementary fashion items, e.g., a blouse and a skirt, has drawn increased research attention [22, 32, 33].

Traditionally, professional fashion compatibility evaluation is manually conducted by fashion experts, magazine editors, and bloggers [12], which is infeasible for ordinary people. By contrast, owing to the extraordinary representation ability, the Deep Neural Networks (DNNs) have become the most promising automatic solutions for FCM, where the compatibility assessment can be measured by the distance between latent representations of two fashion items. Despite the powerful compatibility evaluation, the black-box DNNs cannot explain reasons of the evaluation (e.g., why two fashion items are compatible/incompatible) and hence suffer from the poor interpretability. In a sense, the explanations for FCM are crucial in practice due to the following three reasons: (1) Reliability: explanations make the evaluation results more convincing and enhance the reliability of models [37]. (2) Practicability: explanations help users learn to match fashion items and find compatible ones with fewer attempts [49]. And (3) Expandability: explanations benefit the understanding of the user's clothing matching preference and hence can be directly applied to the personalized fashion recommendation task [13, 34, 35].

To address the limitation of existing studies, we focus on devising the explainable FCM scheme, where the attribute-level interactions (e.g., a *silk* tank top cannot go with a *knit* dress) are adopted as the evaluation explanations. The underlying philosophy is that attributes (e.g., category, pattern, and color) characterize the most intuitive visual cues of fashion items in the semantic level, and their interactions are widely used to interpret the evaluation results by fashion experts [23].

[1] www.fashionunited.com/.

(a) Example 1. (b) Example 2.

Fig. 2. Illustration of the different influences of the attributes. For the same top, the pattern attribute (floral) plays a prominent role in the compatibility evaluation with the given floral bottom in the example 1. By contrast, the color attribute (white) dominates that with the black bottom in the example 2 as "white + black" is one common matching rule.

Without loss of generality, in this work, we particularly study the attribute-wise explainable FCM between items of the two most common fashion categories: the top and bottom, where both the compatibility evaluation score and explanations are generated, as shown in Figure 1.

However, exploring the explainable FCM scheme by virtue of comprehensive fashion attributes is non-trivial due to the following challenges: (1) Fashion attributes are usually unavailable in the practical usage of FCM models, as it is demanding to ask users to key in the detailed attributes of the input fashion items. Hence, how to acquire the discriminative attribute representations of fashion items for FCM constitutes a primary challenge for us. (2) Existing datasets pertaining to FCM lack of ground truth for the attribute-level compatibility. It poses another challenge on how to learn the attribute-level matching signals from the general item-level compatibility annotations and hence generate reasonable explanations. And (3) different attributes (e.g., category, color, and fabric) may contribute differently to the fashion compatibility in diverse item pairs. As can be seen from Figure 2, for the same top, the pattern attribute (floral) plays a prominent role in the compatibility evaluation with the given floral bottom in the example 1. By contrast, the color attribute (white) dominates in that with the black bottom in the example 2 as "white + black" is a common clothing matching rule. Accordingly, how to adaptively weigh the attribute influence is a crucial challenge.

To address the aforementioned challenges, we present an attribute-wise explainable FCM method, termed *ExFCM*. As shown in Figure 3, ExFCM consists of two key modules: (1) *Attribute-wise Representation Learning*. Towards the first challenge, considering fashion attributes usually associate certain regions (e.g., sleeves appear on both sides of the garment), we first resort to the *Attribute Activation Map* (AAM) [49] to align each fashion attribute with its most related region. Then, we propose a new pooling operator, named *Threshold Global Average Pooling* (TGAP), to learn the region-aware attribute representation. (2) *Attribute Interaction Modeling*. For the purpose of tackling the second challenge, we adopt the interaction mechanism [5] to capture the attribute-level matching signals and further facilitate the explainable FCM. Moreover, to cope with the third challenge, we employ the attention mechanism to explore the dynamic influence of each attribute with different input fashion items.

As for the optimization of ExFCM, to enhance the feasibility and portability of ExFCM, we introduce an auxiliary dataset with rich attribute annotations to facilitate the attribute-wise representation learning. Note that the annotations are only the attribute values of each fashion item rather than the attribute-level compatibility between items. Besides, we implement the attribute interaction modeling under the *Bayesian Personalized Ranking* (BPR) framework [29], which is used to explore the relative compatibility between complementary fashion items (i.e., tops and

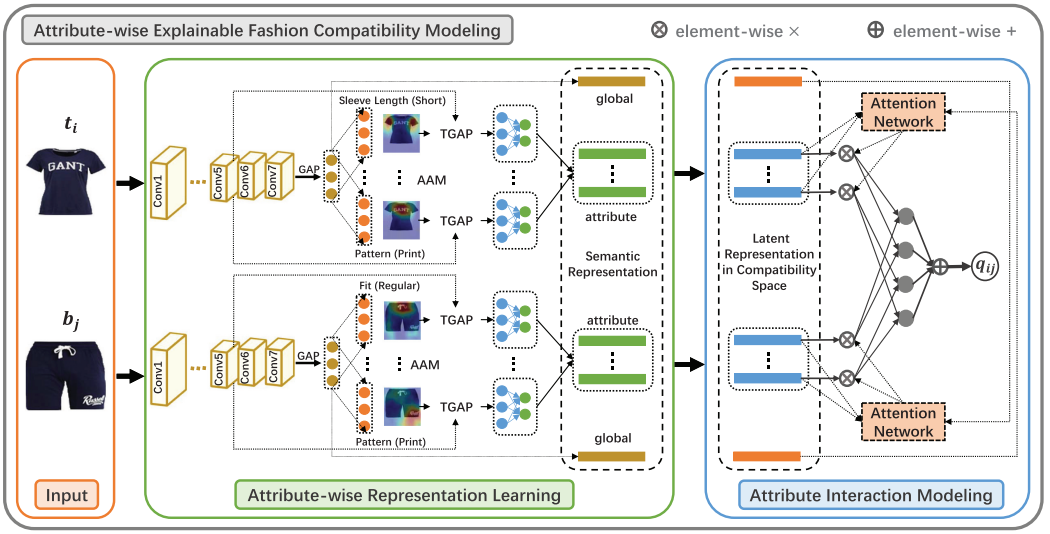


Fig. 3. Illustration of the proposed ExFCM, which consists of two key modules. The attribute-wise representation learning part aims to learn the region-aware attribute representation. The attribute interaction modeling part works on inferring the compatibility of attribute interactions between the top t_i and the bottom b_j, and then generates the overall compatibility evaluation q_{ij}.

bottoms) on our primary dataset for clothing matching. Finally, we present explanations from two aspects: (1) *Attribute Interactions*. We find the most compatible/incompatible attribute pairs (with the highest/lowest compatibility scores) for a given item pair. (2) *Individual Attributes*. We comprehensively evaluate an attribute of the fashion item based on its overall compatibility with all attributes of the complementary item.

Our main contributions can be summarized in threefold:

- We propose an interpretable scheme for FCM, which can generate comprehensive attribute-level explanations for item-level compatibility evaluations.
- The proposed model is capable of inferring attribute-level matching signals between fashion items without any attribute-level compatibility annotations.
- Extensive experiments on two real-world datasets validate that ExFCM can generate more accurate compatibility evaluations than several state-of-the-art methods, together with reasonable explanations. Codes are released.[2]

The rest of the article is organized as follows: Section 2 briefly reviews the related work. In Section 3, we expatiate the proposed ExFCM. The experimental results and detailed analyses are given in Section 4, followed by the conclusion and future work in Section 5.

2 RELATED WORK

2.1 Fashion Compatibility Modeling

Recently, there has been a growing research interest in FCM due to the huge commercial value. Existing efforts have primarily utilized the multi-modal content to learn the compatibility and performed matching in a latent space [27, 33], where compatible items are assumed to be located closer. For example, Li et al. [18] proposed an automatic composition system to score the fashion

[2]https://joeyangbuer.wixsite.com/exfcm.

outfit based on the appearances and meta-data via the multi-modal and multi-instance deep learning. In addition, Vasileva et al. [38] presented an approach to learn the type-aware embedding for fashion items, and jointly learns the item similarity and compatibility in an end-to-end manner. Moreover, Han et al. [8] presented a bidirectional LSTM [39] to sequentially model compatibility relationships among the fashion items of an outfit.

Noticing the rich knowledge referring to clothing matching in the fashion domain, Song et al. [32] compiled the domain knowledge (clothing matching rules) into a pure data-driven FCM model to boost the performance within a teacher-student network [14]. Later, Yang et al. [45] utilized the category-specific complementary relations to model the category-aware compatibility between items via a translation-based embedding space. Although these approaches have achieved compelling success, these models can only answer the question of "whether the given fashion items are compatible or not" but cannot provide explanations. Beyond that, we aim to explore an explainable FCM scheme, which can not only give accurate compatibility evaluations but also improve the interpretability of evaluation results in a comprehensive attribute-wise manner.

2.2 Explainable Fashion Analysis

In fact, plentiful studies have explored the potential of the attribute [24, 46] as the mid-level representation to bridge the long-standing semantic gap [25, 47] between the low-level visual clues and high-level intents (e.g., FCM) and enhance the explainability of fashion analysis results. In particular, Chen et al. [3] proposed a fully automated system that describes the clothing appearance with semantic attributes for fashion applications. Liao et al. [19] proposed an EI (Exclusive & Independent) tree to incorporate the structural knowledge of the fashion domain for facilitating the interpretable fashion item retrieval.

Moreover, several pioneer efforts have been made on the explainable FCM. For example, Tang et al. [36] proposed a method for quantifying how each attribute feature of each item is to the outfit compatibility score. Nevertheless, this work focused on limited attributes (only the shape, texture, and color) and overlooked the attribute interaction in the FCM, making the interpretation incomprehensive. Towards this end, Feng et al. [6] proposed a partition embedding network to learn the embedding of each attribute and then modeled the attribute-level compatibility between input fashion items. However, one key limitation of this work is that it is very dependent on the attribute-level compatibility ground truth, which is usually unavailable in practice. Besides, Wang et al. [42] aimed to learn category-specified pairwise similarities between items and diagnose the incompatible category factors with the backpropagation gradients. Different from these studies, our work is to fulfill explainable FCM task by comprehensively exploring the interactions of various different attributes without any attribute-level compatibility annotations.

3 METHODOLOGY

In this section, we first formally define the explainable FCM problem and then detail the proposed *ExFCM*.

3.1 Problem Formulation

Suppose we have a set of tops $\mathcal{T} = \{t_i\}_{i=1}^{N_t}$ and a set of bottoms $\mathcal{B} = \{b_j\}_{j=1}^{N_b}$, where N_t and N_b denote the total number of tops and bottoms, respectively. In addition, we have a set of positive top-bottom pairs $\mathcal{S} = \{(t_{i_p}, b_{j_p})\}_{p=1}^{P}$ composed by fashion experts, where P is the total number of positive pairs. In this work, we target at addressing the problem of explainable FCM by capturing the attribute-level matching signals between tops and bottoms. In particular, we predefine a set of attributes $\mathcal{U} = \{u_m\}_{m=1}^{M_t}$ for tops and $\mathcal{R} = \{r_n\}_{n=1}^{M_b}$ for bottoms, where M_t and M_b are the

Table 1. Summary of the Main Notations

Notation	Explanation
t_i	The ith top.
b_j	The jth bottom.
u_m	The mth attribute for tops.
M_t	Number of the top attributes.
r_n	The nth attribute for bottoms.
M_b	Number of the bottom attributes.
$\boldsymbol{\Theta_q}$	To-be-learned parameters.
s_{ij}^{mn}	Attribute Compatibility between u_m of top t_i and r_n of bottom b_j.
q_{ij}	Item Compatibility between the top t_i and the bottom b_j.

total number of corresponding attributes, respectively. Moreover, we denote $\mathcal{U}_m = \{u_m^c\}_{c=1}^{U_m}$ and $\mathcal{R}_n = \{r_n^c\}_{c=1}^{R_n}$ as the set of possible values for the attribute u_m and r_n (e.g., "red" and "black" for the attribute "color"), respectively. Then based on $(\mathcal{T}, \mathcal{B}, \mathcal{S})$, we focus on devising an explainable FCM network $\mathcal{Q}$ that is able to simultaneously generate both the overall compatibility q_{ij} and the attribute-wise compatibilities s_{ij}^{mn}'s for a given top-bottom pair (t_i, b_j) as follows:

$$\mathcal{Q}(t_i, b_j|\boldsymbol{\Theta_q}) \rightarrow (q_{ij}, \{s_{ij}^{mn}|\forall m, n\}), \tag{1}$$

where s_{ij}^{mn} refers to the compatibility between the attribute u_m of top t_i and attribute r_n of bottom b_j. $\boldsymbol{\Theta_q}$ denotes the to-be-learned parameters in the network $\mathcal{Q}$. Table 1 summarizes the main notations used in this article.

3.2 ExFCM

As a major novelty, our proposed ExFCM is capable of modeling the attribute interactions between fashion items and hence improving the explainability of compatibility evaluation results. In particular, we first set up the *attribute-wise representation learning* network to capture the region-aware attribute representations of fashion items, and then we introduce the *attribute interaction modeling* to fulfill the explainable compatibility modeling.

Attribute-wise Representation Learning. Here, we take the attribute-wise representation learning for tops as an example, and that for bottoms can be derived in the same manner. To simplify the presentation, we temporally omit the subscript i of t_i.

Due to the concern that the fashion attributes (e.g., *neckline* and *sleeve length*) usually present the high correlations with certain regions, we particularly investigate the region-aware attribute representations of fashion items. To this end, we resort to the AAM [1], which has shown great success in locating the discriminative area of a specific attribute in an image. Specifically, we replace all the fully connected layers of AlexNet with a global average pooling (GAP) [7, 49] layer, which has proven to be effective in capturing the spatial correspondence between feature maps and the attribute [20]. To compensate the removal of fully connected layers, inspired by [1, 49], we introduce two additional convolutional layers (i.e., "conv6" and "conv7") with similar structure to "conv5", as shown in Figure 3. Similar to [1], we feed the output of the GAP layer into M_t parallel attribute classification branches, each of which corresponds to an attribute and is optimized by the cross-entropy loss. Formally, we define the above GAP-modified network as $\mathcal{P}$.

To obtain the AAM of a given top t on attribute u_m, we first derive its attribute value on u_m as follows:

$$c^* = \arg\max_c(\mathcal{P}(u_m^c|t)), \tag{2}$$

where c^* is the index of the predicted attribute value. Then define the AAM for the given top t regarding the attribute u_m as $\mathbf{M}_m^{c^*}$, whose spatial elements can be calculated by,

$$\mathbf{M}_m^{c^*}(x,y) = \sum_k w_k^{c^*} f_k(x,y), \tag{3}$$

where $f_k(x,y)$ represents the activation of the kth feature map generated by the last convolutional layer ("conv7") at the spatial point (x,y). $w_k^{c^*}$ stands for the weight of the kth feature map corresponding to the attribute value $u_m^{c^*}$, which can be derived from the attribute classification network. Intuitively, each entry $\mathbf{M}_m^{c^*}(x,y)$ indicates the contribution of the activations at the spatial point (x,y) towards the classification of the attribute u_m.

Having obtained the attribute regions, we can derive the region-aware attribute representations. Propelled by the fact that certain fashion attributes may involve several unconnected regions (e.g., sleeves appear on both sides of the garment), we devise a novel TGAP method to adaptively pool the spatial features with a threshold θ rather than the inflexible bounding box used in existing work [1]. Formally, we define TGAP with respect to the attribute u_m as follows:

$$o_m^k = \frac{\sum_{(x,y)} g_k(x,y)\,\mathbf{1}(\mathbf{M}_m^{c^*}(x,y) > \theta)}{\sum_{(x,y)} \mathbf{1}(\mathbf{M}_m^{c^*}(x,y) > \theta)}, \tag{4}$$

where $g_k(x,y)$ represents the activation of the kth feature map generated by the "conv5" layer at the spatial point (x,y), and o_m^k is the output of the kth feature map by TGAP. Note that the "conv7" layer has been used to calculate the AAM, and we thus adopt the shallower convolutional layer (i.e., "conv5") for avoiding the overfitting. $\mathbf{1}(z)$ denotes an indicator function that returns 1 when the argument z is *true* and 0 otherwise.

Ultimately, to obtain the attribute representations, similar to [1], we feed the output of TGAP $\mathbf{o}_m = [o_m^1, \ldots, o_m^K]$ into M_t multilayer perceptrons (MLPs) $\{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_{M_t}\}$, respectively. Specifically, we have,

$$\mathbf{a}_m^t = \mathcal{F}_m(\mathbf{o}_m), m \in \{1, 2, \ldots, M_t\}, \tag{5}$$

where $\mathbf{a}_m^t$ refers to the attribute representation for the top t on attribute u_m. Similarly, we can derive the attribute representation $\mathbf{a}_n^b$ for the bottom b on attribute r_n.

Optimization. In fact, most existing real-world datasets for FCM lack fine-grained attribute labels for each fashion item and hence cannot well support our attribute-wise representation learning. Towards this end, we introduce an auxiliary set of tops $\widehat{\mathcal{T}} = \{\hat{t}_i\}_{i=1}^{\hat{N}_t}$ and bottoms $\widehat{\mathcal{B}} = \{\hat{b}_j\}_{j=1}^{\hat{N}_b}$ with fine-grained attribute annotations. Similarly, here, we temporally omit the subscript i of $\hat{t}_i$ for simplicity.

Undoubtedly, it is reasonable to assume that fashion items with similar visual signals usually share similar attribute representations. To explore this underlying semantic correlation among fashion items, we adopt the triplet loss [15] as follows:

$$\mathcal{L}_T = \sum_{(\hat{t},t^+,t^-)\in\mathcal{E}} \sum_{u_m\in\mathcal{U}} max\{0, \alpha - d(\mathbf{a}_m^{\hat{t}}, \mathbf{a}_m^{t^+}) + d(\mathbf{a}_m^{\hat{t}}, \mathbf{a}_m^{t^-})\}, \tag{6}$$

where α is a margin, and $d(\cdot,\cdot)$ represents the cosine similarity between the attribute representations. $\mathcal{E}$ denotes the training triplet set, which is defined as follows:

$$\{(\hat{t},t^+,t^-)|\forall u_m \in \mathcal{U},\ \hat{u}_m = u_m^+ \wedge \hat{u}_m \neq u_m^- \wedge \{\hat{t},t^+,t^-\} \subseteq \widehat{\mathcal{T}}\}, \tag{7}$$

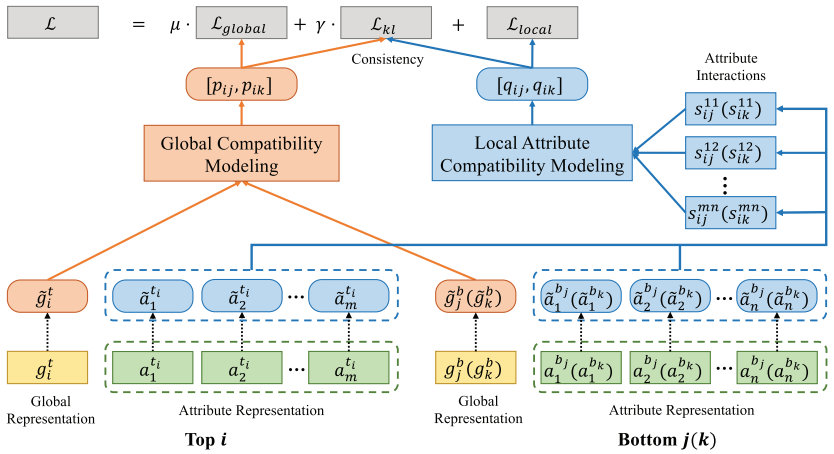


Fig. 4. Workflow of the proposed attribute interaction modeling.

where $\hat{u}_m$, u_m^+, and u_m^- denote the attribute values of tops $\hat{t}$, t^+, and t^- on attribute u_m, respectively. Intuitively, the top $\hat{t}$ shares the same attribute values with the positive sample t^+ while completely differing from those of the negative sample t^-.

Apart from preserving the semantic correlation among fashion items, we also expect that the attribution representation can well retain the discriminative cues towards the corresponding attribute classification. In light of this, we adopt the following objective function:

$$\mathcal{L}_{RC} = \sum_{\hat{t} \in \widehat{\mathcal{T}}} \sum_{u_m \in \mathcal{U}} -log(p(u_m^{\hat{c}} | \mathbf{a}_m^{\hat{t}})), \tag{8}$$

where $u_m^{\hat{c}}$ is the ground truth value of top $\hat{t}$ on attribute u_m. Ultimately, we reach the final objective function for our attribute-wise representation learning network as follows:

$$\mathcal{L}_A = \mathcal{L}_T + \mathcal{L}_{RC}. \tag{9}$$

Attribute Interaction Modeling. Having obtained the attribute representations for fashion items, we can proceed to the core of ExFCM: attribute interaction modeling, which aims to capture the attribute-level matching signals and enhance the interpretability of the compatibility evaluation. Towards this end, we lean upon the interaction mechanism [5, 44] that has been widely used in various tasks, such as the natural language inference [28, 41] and retrieve-based chatbot [43]. The key idea of interaction mechanism is to assign interaction scores between small units to infer fine-grained clues about whether two contents are matching.

Intuitively, it is reasonable to argue that compatible fashion items should share certain attribute interaction patterns. For example, *tank tops* go better with *shorts* instead of the *dress*, while *red* tops better avoid the *green* bottoms. Therefore, to better characterize the attribute interaction, we first seek the latent compatibility space where compatible fashion item pairs are located closer than those are incompatible. Due to the remarkable representation ability of DNNs [4, 11, 26], we employ MLP to explore the latent compatibility space. In particular, given the attribute representation $\mathbf{a}_m^t$ of the top t, we can derive its latent attribute representation $\tilde{\mathbf{a}}_m^t \in \mathbb{R}^{D_k}$ as follows:

$$\tilde{\mathbf{a}}_m^t = \mathcal{H}_a^t\ (\mathbf{a}_m^t), \tag{10}$$

where $\mathcal{H}_a^t$ is an MLP with l hidden layers. D_k is the dimensionality of the latent attribute-wise compatibility space. In the same way, we can derive the latent attribute representation $\tilde{\mathbf{a}}_n^b$ for the bottom b with a MLP $\mathcal{H}_a^b$.

ALGORITHM 1: Attribute Interaction Modeling.

Require: $\mathcal{D}_s = (i, j, k), \mu, \gamma, l$

Ensure: Parameters in the neural networks $\mathcal{H}_a^t, \mathcal{H}_a^b, \mathcal{H}_g^t, \mathcal{H}_g^b$, and parameters $\mathbf{w}, \mathbf{U}_a, \mathbf{U}_g$ in the attention network $\mathcal{A}$.

1: Initialize parameters in the networks $\mathcal{H}_a^t, \mathcal{H}_a^b, \mathcal{H}_g^t, \mathcal{H}_g^b, \mathcal{A}$.
2: **repeat**
3: Randomly draw (i, j, k) from $\mathcal{D}_s$
4: Calculate the global and attribute representations by the trained networks $\mathcal{P}$ and $\mathcal{F}_m$.
5: Calculate the latent global and attribute representations according to Equation (10) and Equation (12).
6: Calculate the attribute influence according to Equation (13).
7: Update the parameters of neural networks $\mathcal{H}_a^t, \mathcal{H}_a^b, \mathcal{H}_g^t, \mathcal{H}_g^b, \mathcal{A}$ according to Equation (20).
8: **until** Converge

Based on the latent attribute representation, one naive approach to measure the overall compatibility q between the top t and the bottom b is averaging the pair-wise attribute interaction scores as follows:

$$q = \frac{1}{M_t M_b} \sum_{m=1}^{M_t} \sum_{n=1}^{M_b} (\tilde{\mathbf{a}}_m^t)^T \tilde{\mathbf{a}}_n^b. \tag{11}$$

Apparently, this method overlooks the fact that different attributes can flexibly contribute to the fashion compatibility in diverse matching contexts, as shown in Figure 2.

Given this, we utilize the attention mechanism [16, 17, 40] to flexibly assign the attribute influence in the overall compatibility modeling with different top-bottom contexts. Towards this end, to evaluate the attribute influence of a given top t towards different bottom contexts, we adopt the global visual representation $\mathbf{g}^b$ of the bottom b, which can be derived from the output of the GAP layer in the aforementioned GAP-modified network $\mathcal{P}$. Considering that the latent compatibility space may be highly non-linear, we further obtain the latent global visual representation of the bottom b as follows:

$$\tilde{\mathbf{g}}^b = \mathcal{H}_g^b\ (\mathbf{g}^b), \tag{12}$$

where $\mathcal{H}_g^b$ is an MLP with l hidden layers. Thereafter, given the bottom context b, we can calculate the influence of the attribute u_m of the top t as follows:

$$\begin{cases} \omega_m^t = \mathbf{w}^T \sigma(\mathbf{U}_a \tilde{\mathbf{a}}_m^t + \mathbf{U}_g \tilde{\mathbf{g}}^b + \mathbf{b}), \\ \lambda_m^t = \dfrac{exp(\omega_m^t)}{\sum_{m=1}^{M_t} exp(\omega_m^t)}, \end{cases} \tag{13}$$

where $\mathbf{U}_a \in \mathbb{R}^{D_h \times D_k}$, $\mathbf{U}_g \in \mathbb{R}^{D_h \times D_l}$, $\mathbf{b} \in \mathbb{R}^{D_h}$, $\mathbf{w} \in \mathbb{R}^{D_h}$ are parameters of the attention network, and D_h represents the hidden layer size of the attention network. λ_m^t denotes the attribute influence of u_m for top t matching with the bottom b. In the similar manner, we can derive the latent global representations of tops $\tilde{\mathbf{g}}^t$'s with an MLP $\mathcal{H}_g^t$ and the attribute influence λ_n^b of r_n for bottom b matching with the top t. $\sigma(\cdot)$ is the sigmoid function.

Based on the attribute influences, we can measure the overall compatibility q between top t and bottom b as follows:

$$\begin{cases} s^{mn} = \lambda_m^t \lambda_n^b (\tilde{\mathbf{a}}_m^t)^T \tilde{\mathbf{a}}_n^b, \\ q = \dfrac{1}{M_t M_b} \sum_{m=1}^{M_t} \sum_{n=1}^{M_b} s^{mn}, \end{cases} \tag{14}$$

Table 2. Attributes and Their Possible Values

Attributes	Attribute Values	Total
Category	Shirt, Dress, Trousers, Jacket, Shorts, ...	16
Color	Green, Navy, Blue, White, Black, ...	18
Fabric	Denim, Jersey, Canvas, Leather, Sweat, ...	14
Fit	Skinny, Straight, Regular, High waist, Oversize, ...	15
Pattern	Animal, Plain, Photo, Floral, Pinstriped, ...	16
Neckline	V-neck, Square, Round, Backless, Envelope, ...	11
Sleeve length	Long, Short, Sleeveless, Strapless, Elbow, ...	9

Attributes in bold can be only applied to tops.

where s^{mn} denotes the attribute-wise compatibility. In particular, $(\tilde{\mathbf{a}}_m^t)^T \tilde{\mathbf{a}}_n^b$ can be interpreted as the interaction score between the attribute u_m of the top t and attribute r_n of the bottom b. $\lambda_m^t \lambda_n^b$ represents the influence of the attribute pair (u_m, r_n) to the overall item compatibility.

Based on the attribute-wise compatibilities s^{mn}'s, given a top t, we can ascertain the most beneficial bottom attribute r_{n^+} or the most harmful bottom attribute r_{n^-} as follows:

$$\begin{cases} n^+ = \arg\max_n \sum_{m=1}^{M_t} s^{mn}, \\ n^- = \arg\min_n \sum_{m=1}^{M_t} s^{mn}. \end{cases} \tag{15}$$

Optimization. In a sense, we can easily derive the positive (compatible) top-bottom pairs from those that have been composed together by fashion experts. However, with respect to the non-composed fashion item pairs, we cannot safely draw the conclusion that they are incompatible, as they can also be the missing potential positive pairs (i.e., pairs can be composed in the future). Towards this end, to better model the implicit compatibility preference between the tops and bottoms, we naturally adopt the BPR framework, which has proven to be effective in the implicit preference modeling [2, 10]. In particular, we derive a positive bottom set $\mathcal{B}_i^+ = \{b_j \in \mathcal{B} | (t_i, b_j) \in \mathcal{S}\}$ for each top t_i. Then, we assume that bottoms from the positive set $\mathcal{B}_i^+$ are more compatible than those non-composed neutral bottoms for top t_i. Hence, we build the following training set:

$$\mathcal{D}_S := \{(i, j, k) | t_i \in \mathcal{T}, b_j \in \mathcal{B}_i^+ \wedge b_k \in \mathcal{B} \backslash \mathcal{B}_i^+\}, \tag{16}$$

where the triplet (i, j, k) indicates that bottom b_j is more compatible with top t_i compared to b_k.

Then according to BPR, we have the objective function for local attribute-wise compatibility modeling,

$$\mathcal{L}_{local} = \sum_{(i,j,k) \in \mathcal{D}_S} -ln(\sigma(q_{ij} - q_{ik})), \tag{17}$$

where q_{ij} denotes the overall compatibility between t_i and b_j. In addition, considering that the attribute representation may fail to capture certain global characteristics (e.g., style) of fashion items, we also incorporate the global visual compatibility modeling by,

$$\mathcal{L}_{global} = \sum_{(i,j,k) \in \mathcal{D}_S} -ln(\sigma(p_{ij} - p_{ik})), \tag{18}$$

where $p_{ij} = (\tilde{\mathbf{g}}_i^t)^T \tilde{\mathbf{g}}_j^b$ stands for the global compatibility between top t_i and bottom b_j. In fact, the results derived from both perspectives should be consistent in a sense. Accordingly, we employ

Table 3. The Discriminative Ability of the Learned Attribute Representation on ACC (%) and AUC (%)

Attributes	Top		Bottom	
	ACC	AUC	ACC	AUC
Category	95.93	95.10	**97.78**	98.23
Color	89.27	84.64	88.00	97.47
Fabric	92.53	87.95	86.42	96.01
Fit	91.31	89.96	97.05	**99.59**
Pattern	93.70	91.15	85.91	98.20
Neckline	94.57	92.13	-	-
Sleeve Length	**96.96**	**96.91**	-	-
Average	**93.47**	**91.12**	**91.03**	**97.90**

the Kullback-Leibler (KL) Divergence to regularize the model results as follows:

$$\mathcal{L}_{kl} = \sum_{(i,j,k)\in\mathcal{D}_S} \left(\tilde{p}_{ij} \, log\left(\frac{\tilde{p}_{ij}}{\tilde{q}_{ij}}\right) + \tilde{p}_{ik} \, log\left(\frac{\tilde{p}_{ik}}{\tilde{q}_{ik}}\right)\right), \tag{19}$$

where $\tilde{q}_{ij}$ and $\tilde{q}_{ik}$ are the softmax output of q_{ij} and q_{ik}, respectively, and mean the sum-normalized distribution over the compatibility scores predicted from the attribute perspective. $\tilde{p}_{ij}$ and $\tilde{p}_{ik}$ can be calculated in the same way. Ultimately, we have the following objective function:

$$\mathcal{L} = \mathcal{L}_{local} + \mu\mathcal{L}_{global} + \gamma\mathcal{L}_{kl}, \tag{20}$$

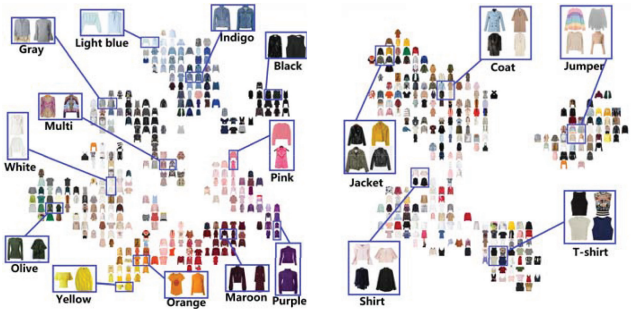
where μ and γ are the non-negative coefficients. Figure 4 illustrates the workflow of the attribute interaction modeling, while the optimization procedure is summarized in Algorithm 1.

4 EXPERIMENT

In this section, we systematically evaluated the effectiveness of the proposed ExFCM on two real-world datasets FashionVC and ExpFashion. We first introduce the experimental setting in Section 4.1 and then present the result of each experiment in the following subsections. Specifically, we evaluate the discriminative ability of the attribute-wise representation learning in Section 4.2. We next compare the proposed ExFCM with several classic FCM methods in Section 4.3. To assess the practicability, we evaluate the proposed ExFCM in the context of the complementary fashion item retrieval in Section 4.4. Besides, in Section 4.5, we conduct the explainability analysis with certain intuitive examples. Finally, to gain more deep insights, we illustrate the latent attribute matching patterns in Section 4.6.

4.1 Experimental Settings

Dataset. Existing datasets for FCM are mainly collected from either the e-commerce websites, like Amazon [27], or the fashion-oriented communities like Ployvore [8, 21, 33], where the co-purchased items of users and collocated items by fashion lovers are treated as the compatible samples, respectively. As the co-purchase relation can be very noisy and less convincing, we adopted **FashionVC** [33] and **ExpFashion** [21] as our primary datasets, both of which are crawled from Polyvore and created by fashion experts. FashionVC consists of 20,726 outfits with 14,871 tops and 13,663 bottoms, and ExpFashion comprises 200,745 outfits with 29,113 tops and 20,902 bottoms. Due to the absence of the attribute-level annotations in above two primary datasets, we introduced an auxiliary dataset **Fashion100K** [1] to pretrain the attribute-wise representation



(a). Color (b). Category

Fig. 5. The visualization of part attribute representation in terms of color and category.

learning part, based on which we can obtain the attribute representations for each item in our primary datasets. Fashion100K consists of 101,021 images and adopts a taxonomy of 12 fashion item attributes. On account of some uncommon attributes that have limited samples, we particularly chose 7 attributes, as shown in Table 2. Notably, all these 7 attributes can be used to describe the tops, while only 5 of them can be applied to the bottoms, i.e., $M_t = 7$ and $M_b = 5$.

Implementation Details. Pertaining to the attribute localization, the GAP-modified network $\mathcal{P}$ consists of seven convolutional layers, a GAP layer, and multiple parallel attribute classification branches, each of which specializes in an associated attribute classification. In addition, each attribute representation learning network (i.e., $\mathcal{F}_m$) is composed of a two-layer MLP. We divided the auxiliary dataset into three chunks: training set (80%), validation set (10%), and testing set (10%). The training of the attribute-wise representation learning part consists of two stages. We first pretrained $\mathcal{P}$ with only the attribute classification loss and then jointly trained $\mathcal{F}_m, m \in \{1, 2, \ldots, M_t\}$ and $\mathcal{P}$ according to Equation (9). We adopted the grid search strategy to determine the optimal values for the hyper-parameters (i.e., α, θ) among the values $[0.1, 0.2, 0.3, 0.4]$ and $[0.7, 0.8, 0.9]$, respectively. Besides, both learning rates of two training stages are searched in $[0.00001, 0.00005, 0.0001, 0.0005]$. We empirically found that this part achieves the optimal performance when $\alpha = 0.3, \theta = 0.8$, and the learning rates of above two training stages with 0.0001 and 0.00005, respectively.

Regarding the attribute interaction modeling, we divided the positive pair set $\mathcal{S}$ into three chunks: training set (80%), validation set (10%), and testing set (10%), and then generated triplets for above three chunks according to Equation (16). To balance the amount of the two primary datasets, three negative bottoms are sampled for each positive top-bottom pair in FashionVC while

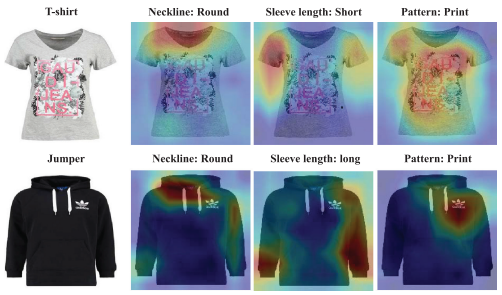


Fig. 6. The related region of attributes located by AAM.

Table 4. The Performance Comparison of FCM among Different Methods with Respect to AUC (%)

Approaches	FashionVC	ExpFashion
IBR	62.11	73.33
IBR-Attr	62.66	73.42
Bi-LSTM-V	66.29	73.09
Bi-LSTM-V-Attr	67.34	73.44
BPR-DAE-V	67.02	83.09
BPR-DAE-V-Attr	67.31	83.44
ExFCM-NoAttn	65.49	75.76
ExFCM-NoKL	67.46	82.61
ExFCM	**68.71**	**84.24**

only one negative bottom for each pair from ExpFashion. As for the optimization, we adopted the grid search strategy to determine the optimal values for the hyper-parameters (i.e., μ, γ) among the values $[0.3, 0.4, 0.5]$ and $[0.3, 0.4, 0.5]$, respectively. In addition, the learning rate and dropout rate are searched in $[0.0001, 0.0005, 0.001]$ and $[0.3, 0.4, 0.5]$, respectively. We empirically found that the proposed model achieves the optimal performance at $l = 1, \mu = 0.5, \gamma = 0.5$ with dropout rate as 0.5 and learning rate as 0.0005. Ultimately, both the above two key components of ExFCM are optimized by the Adam algorithm and the batch size is set as 64.

4.2 On Representation Learning

To comprehensively demonstrate the effectiveness of the attribute-wise representation learning, we adopted the accuracy (ACC) and the area under the ROC curve (AUC) [30] to evaluate its performance in both the attribute classification and triplet-wise attribute similarity evaluation tasks, respectively.

Table 3 shows the performance of the representation learning for each attribute in both tasks. In terms of the top attributes, we observed that the attribute "sleeve length" gains the best performance regarding both the ACC and AUC. One possible reason is that the "sleeve length" intrinsically has more discriminative spatial property and hence is easier to be captured by the AAM, compared to the other attributes. As shown in Figure 6, the *short sleeves* are most likely to appear around the upper shoulder region, while the *long sleeves* often cover the lower part of the top. In addition, we noticed that the performance of the attribute "category" is also promising, which may be attributed to the fact that the "category" is often closely correlated with the "sleeve length". Meanwhile, our attribute-wise representation learning presents the worst performance for the attribute "color". The possible explanations are twofold: (1) the attribute "color" is usually widely distributed over the whole item with no distinct discriminative area; (2) the value labels of this attribute are much fine-grained, such as the *light blue* and *indigo*, sharing high visual similarity and making the representation learning more challenging. As for the bottom attributes, similar observations can be found. The attributes, such as "category" and "fit", that have clear spacial distribution features gain better performance than the other attributes, such as the "color" and "fabric". Overall, as we can see, our attribute representation learning achieves the satisfactory performance with an average accuracy of 92.45% and AUC of 93.95% in the contexts of attribute classification and triplet-wise attribute similarity evaluation, respectively.

To gain a deep understanding of the performance, we performed the dimensionality reduction to visualize the learned attribute representations. For illustration, we chose the attributes "color" and

Bottom	Category	Color	Fabric	Fit	Neckline	Pattern	Sleeve Length
b_j	0.0904	0.1977	0.0436	0.1405	0.1378	**0.2175**	0.1726
b_k	0.1315	**0.2477**	0.0259	0.1663	0.1414	0.0442	0.2430

Fig. 7. Illustration of attentive attribute influence. Each number indicates the influence of a top attribute when the top is paired with a bottom (one row for each bottom).

"category", which have 19 and 5 possible values, respectively. We randomly sampled 40 fashion items for each value of the attribute "color" and 120 for "category". As shown in Figure 5,[3] the learned representations of similar attribute values are much closer than those of dissimilar ones, which is consistent with the quantitative evaluations.

4.3 On Model Comparison

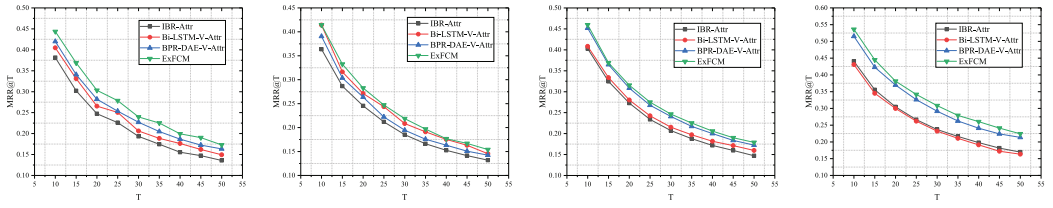
We compared our proposed ExFCM with the following baselines on FCM:

- **IBR:** IBR is an image-based recommendation method proposed by [27], which aims to model the relations between objects based on their visual appearance. In particular, a latent visual style space is learned based on which related objects can be retrieved using nearest-neighbor search.
- **IBR-Attr:** We extend IBR by incorporating the attribute information (i.e., $\tilde{\mathbf{a}}_m^s$) extracted from our pretrained attribute representation learning network to embed each fashion item as follows:

$$\mathbf{v}^s = \tilde{\mathbf{g}}^s + \frac{1}{M_s}\sum_{m=1}^{M_s}\tilde{\mathbf{a}}_m^s,\ s \in \{t, b\}, \tag{21}$$

  where $\tilde{\mathbf{g}}^s$ denotes the global visual representation of the item.
- **BPR-DAE-V:** BPR-DAE is a content-based neural scheme introduced by [34], which explores the multi-modal data fusion of fashion items towards FCM. For the sake of fairness, here, we use the variant of this model (BPR-DAE-V) with only the visual modality as ours.
- **BPR-DAE-V-Attr:** Similarly, we introduce the attribute semantic representation into BPR-DAE-V according to Equation (21).
- **Bi-LSTM-V:** Bi-LSTM in [8] models the compatibility of multiple fashion items in a sequential way. Similarly, we only feed the model with the visual modality of garments.
- **Bi-LSTM-V-Attr:** In the same way, we integrate the attribute representation into Bi-LSTM according to Equation (21).
- **ExFCM-NoAttn:** This baseline is a derivative of our ExFCM, where the attention mechanism is disabled and the influence for different attribute pairs is assigned uniformly.

[3]For better illustration, some overlapped images are randomly removed.



(a) (b) (c) (d)

Fig. 8. Performance of different models with respect to MRR at different numbers of the bottom candidates T. (a) and (c) show the results of observed top query on FashionVC and Expfashion, respectively. (b) and (d) are the results of unobserved top query on FashionVC and Expfashion, respectively.

- **ExFCM-NoKL:** This method is derived from our ExFCM by eliminating the KL loss, where the FCM only focuses on the attribute perspective without considering the global visual cues.

We adopted AUC as the evaluation metric, and the comparison results of ExFCM and baselines on both FashionVC and ExpFashion are given in Table 4. From this table, we have the following observations:

(1) IBR achieves the worst performance compared to the other methods. This may be attributed to the fact that the factors contributing to compatibility range from style and color, to material and shape, and their relations can be rather sophisticated. However, IBR exploits the compatibility relations in a plain linear style space, which can be insufficient to model the highly complicated compatibility between fashion items. On the contrary, BPR-DAE-V resorts to advanced DNNs to seek the non-linear latent compatibility space, and thus improves the performance.

(2) Incorporating the attribute representations (-Attr) into IBR, Bi-LSTM-V, and BPR-DAE-V does boost the performance, which implies that the attribute semantic cues and global visual signals complement each other in the FCM.

(3) ExFCM consistently outperforms all baselines in terms of AUC on both datasets, demonstrating the superiority of the attribute interaction modeling in evaluating the compatibility between fashion items. Notably, ExFCM surpasses all the enhanced baselines (i.e., IBR-Attr, Bi-LSTM-V-Attr, and BPR-DAE-V-Attr) that also incorporate the attribute representations learned with the help of the auxiliary dataset, reflecting the necessity of exploring comprehensive attribute interactions.

(4) ExFCM shows remarkable superiority over ExFCM-NoAttn on both datasets, which enables us to draw the conclusion that it is advisable to assign the influence of attribute pairs attentively rather than uniformly. Accordingly, this confirms the assumption that different attributes of an item can contribute differently in the compatibility with diverse matching items.

(5) The performance of ExFCM is consistently better than its derivative ExFCM-NoKL across different datasets, suggesting that it is appropriate to consider the consistency between the compatibility evaluation results from the global visual view and local attribute perspective in the FCM context.

To gain deeper insights, we checked the influence assignment results for different attributes of the same top towards different bottoms. As can be seen from Figure 7, given the bottom b_j, the attribute "pattern" of the top t_i gains the highest attention value, which is consistent to the clothing matching rule that a *striped* top can go better with a *striped* bottom. Then, given the bottom b_k, the influence assigned to the attribute "color" of the top is the largest one, which is reasonable due to the color consistency between t_i and b_k.

Fig. 9. Illustration of the ranking results of ExFCM on *Observed Tops* and *Unobserved Tops* scenarios for the given testing tops. The bottoms highlighted in the red boxes are positive.

4.4 On Complementary Fashion Item Retrieval

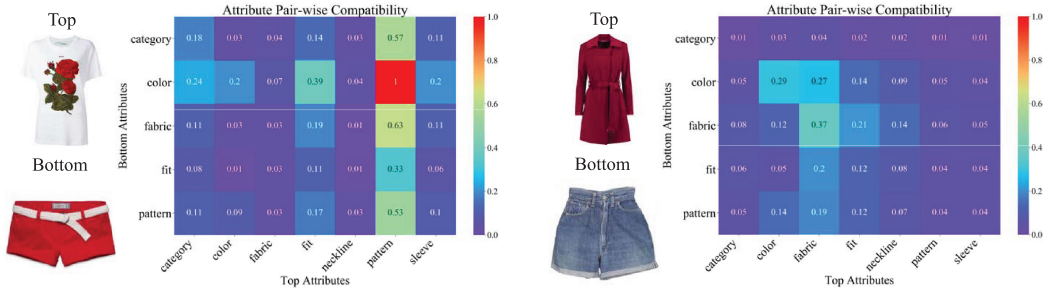
To assess the practical value of our proposed ExFCM, we evaluated the performance with respect to complementary fashion item retrieval, where we employed the Mean Reciprocal Rank (MRR) as the evaluation metric. In particular, considering the fact that it is time-consuming to rank all bottoms for each given top, we adopted the protocol in [9], i.e., we fed each top that appeared in our testing set as a query and randomly selected T bottoms as the ranking candidates with only one positive bottom. Then, we generated a ranking list of the bottoms for each given top query based on their compatibility scores. In total, there are 1,954 unique tops and 6,343 unique tops in the testing set of FashionVC and ExpFashion, respectively. In view of the sparsity of these real-world datasets, we found that there are 1,262 tops and 2,198 tops that never appeared in the training set of FashionVC and ExpFashion, respectively. To comprehensively evaluate the proposed model, we compared it with different FCM methods in two testing scenarios: *Observed Tops* and *Unobserved Tops*.

For the sake of fairness, we compared our proposed ExFCM with baselines' derivative, all of which incorporate the attribute information according to Equation (21). As shown in Figure 8, we observed that ExFCM consistently achieves the best retrieval performance in terms of MRR@T at different number of bottom candidates (i.e., T), demonstrating the effectiveness of our ExFCM in the complementary item retrieval. In particular, regarding the MRR@10, ExFCM improves the performance on observed tops by 6.0%, 4.5%, and 1.6%, while that on unobserved tops by 7.3%, 5.3%, and 2.2%, as compared to IBR-Attr, Bi-LSTM-V-Attr, and BPR-DAE-V-Attr, respectively, which suggests that ExFCM can cope well with the practical cold-start problem.

Figure 9 shows several intuitive ranking results in different scenarios. For illustration, we adopt the case of $T = 10$, where for each top query, we selected 10 bottom candidates including 1 positive one and 9 negative ones. The positive bottoms have been highlighted by red boxes. As we can see, ExFCM tends to recommend the positive bottoms with the high priority. Although ExFCM fails to accurately rank the positive item at the exact first place in some cases, the recommended bottoms ranked before the positive bottom also seem to be compatible with the given top. As can be seen from the first example of unobserved tops, the black jeans at the first place are not the positive ground truth, but still goes well with the given white T-shirt due to the harmonious style and color.

4.5 On Explainability Analysis

To illustrate the explainability of our model, we show the pair-wise attribute compatibility of a compatible outfit and an incompatible one in Figure 10. As we can see, in the first compatible



(a) A compatible outfit.

(b) An incompatible outfit.

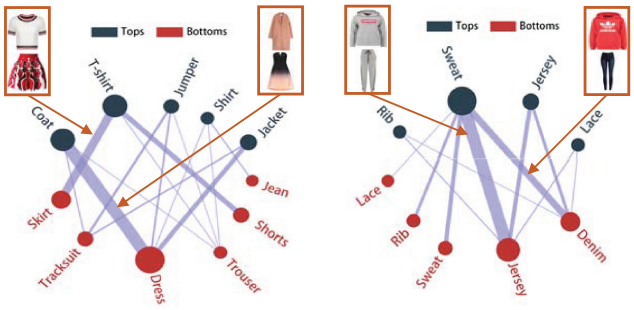
Fig. 10. The visualization of the attribute compatibility of a compatible outfit and an incompatible outfit.

example, the interaction between the top's "pattern" and the bottom's "color" obtains the highest score, which may be due to that the top's pattern (floral) shares the same red color with the bottom and can make the coordinated outfit. As for the incompatible example, most of the attribute pairs between the top and bottom suffer from the relatively low compatibility. Interestingly, we noticed that our model gives the highest compatibility to the attribute pair "fabric + fabric". Intuitively, this is reasonable, as the "fleece" top can go with a "denim" bottom, e.g., a fleece coat plus long jeans. According to Equation (15), taking the summation over all pair-wise compatibility of one attribute, we found that the most harmful attribute of the top for the bottom is the "category", while that of the bottom for the top is the "sleeve length". Both results are reasonable according to our common sense in fashion domain.

To quantitatively verify the explanations provided by ExFCM, we manually built a testing dataset comprising 100 incompatible top-bottom pairs. Specifically, we first randomly sampled some tops and then employed a fashion expert to manually choose one incompatible bottom for each top. Notably, to facilitate the evaluation, we required the fashion expert to particularly choose the incompatible bottom that has only one harmful/incompatible attribute for matching the given top. To guarantee the quality of the testing dataset, we further employed another fashion expert to check the selected incompatible bottoms and only kept the bottoms that are supported by both fashion experts. Finally, we obtained 100 incompatible top-bottom pairs. Then, we employed our ExFCM to ascertain the most harmful bottom attribute for each top-bottom pair according to Equation (15). Due to the fact that there is limited work on ascertaining the most harmful attribute for incompatible outfit, here, we introduced the most common baseline, i.e., random (Rand) strategy, which identifies the most harmful bottom attribute randomly. Table 5 shows the comparison results in terms of the accuracy (ACC). As can be seen, ExFCM significantly outperforms the random method, demonstrating the superior explainability of ExFCM.

Table 5. Performance Comparison on Locating the Most Harmful Attribute

Approaches	ACC (%)
Rand	19.00
ExFCM	44.00



(a) Category (b) Fabric

Fig. 11. Attribute matching patterns on the "category" and "fabric". Examples of the most common two matching patterns are given for each attribute.

4.6 On Knowledge Discovery

We further conducted experiments on the knowledge discovery to mine the latent attribute matching patterns in the fashion domain. In particular, we merged the positive top-bottom pairs on both primary datasets and employed our ExFCM to obtain the pair-wise attribute compatibility for each pair. To ensure the quality of the learned attribute matching patterns, for each top-bottom pair, we only considered the attribute interaction assigned with the highest compatibility score as a matching pattern candidate. In a sense, the matching attributes of a top and a bottom usually come from the same type. Accordingly, here, we only considered the attribute matching patterns that involve the same attribute of the top and the bottom. For illustration, Figure 11 visualizes the attribute matching patterns regarding the attributes "category" and "fabric". Each edge, linking a top attribute value and a bottom attribute value, corresponds to a matching pattern. The width of the edge reflects the frequency of the pattern occurred in our dataset. Intuitively, the larger the frequency, the stronger the attribute matching pattern. Notably, for clear illustration, we only keep matching patterns with the top 50% frequency. From this figure, we observed that: (1) the top category "coat" goes better with the bottom category "dress", while the "t-shirt" matches "skirt" and "shorts" well in most cases; (2) The most versatile bottom categories are "dress" and "trousers", as they match more types of tops, ranging from the "coat" to "t-shirt". (3) A top with the fabric attribute of "sweat" (e.g., a sweatshirt) is more likely to make a compatible outfit with a "denim" bottom (e.g., the jeans) or a "jersey" bottom (e.g., the sport pants). Overall, the mined attribute matching patterns are reasonable according to our common sense and can facilitate people to dress properly.

To comprehensively assess our model in knowledge discovery, apart from the above objective evaluation, we further conducted the subjective user study. In this part, we invited 40 fashion-lovers (20 males and 20 females) to participate the psycho-visual test over the top-5 "category" and "fabric" matching patterns. In particular, each fashion-lover was asked to judge whether these 10 attribute matching patterns mined by our ExFCM are reasonable. We illustrate the female, male, and average support rates of the psycho-visual test in Table 6. As we can see, overall, the fashion-lovers supported the mined attribute matching patterns, which is consistent with the above objective evaluation result. Besides, for some gender-specific matching patterns, such as "coat+dress" and "jacket+dress", we noticed that the fashion-lovers with corresponding gender will give a higher support rate. This is also reasonable due to the difference in the clothing matching styles and habits of different genders.

Table 6. Support Rate (SR) of Fashion-lovers over the Top-5 Attribute Matching Patterns on the "Category", "Fabric"

Id	Top	Bottom	Female SR (%)	Male SR (%)	Average SR (%)
P1	coat	dress	80.0	35.0	57.5
P2	t-shirt	skirt	95.0	70.0	82.5
P3	t-shirt	shorts	100.0	95.0	97.5
P4	jacket	dress	65.0	35.0	50.0
P5	jumper	tracksuit	85.0	65.0	75.0
P6	sweat	jersey	80.0	75.0	77.5
P7	sweat	denim	85.0	70.0	77.5
P8	sweat	sweat	80.0	85.0	82.5
P9	jersey	jersey	90.0	90.0	90.0
P10	jersey	denim	100.0	80.0	90.0

The first five rows are the results of "category," and the last five rows are of "fabric."

5 CONCLUSION AND FUTURE WORK

In this work, we presented an attribute-wise explainable FCM scheme, dubbed ExFCM, which is able to simultaneously generate the compatibility evaluation for input fashion items and explanations for the evaluation result. In particular, we utilized the interaction mechanism to infer attribute-level matching signals between fashion items without any attribute-level compatibility annotations. Considering that the same attribute can have different influence levels in different item contexts. The matching signals are dynamically aggregated into the overall evaluation by the attention mechanism. Extensive experiments conducted on two real-world datasets demonstrate that ExFCM can generate evaluations more accurately than several state-of-art methods, together with reasonable explanations.

One limitation of our work is that we ignore the matching preference of different users (e.g., some users are fond of matching "coat" with "jeans", while others prefer to match with "dress"). In view of the matching habits of different users are not exactly the same, we plan to incorporate the user preference into FCM in the future. In addition, we would like to encode more users' side information to conduct more well-rounded FCM, such as gender, occupation, and body shapes.

REFERENCES

[1] Kenan E. Ak, Ashraf A. Kassim, Joo-Hwee Lim, and Jo Yew Tham. 2018. Learning attribute representations with localization for flexible fashion search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7708–7717.

[2] Da Cao, Liqiang Nie, Xiangnan He, Xiaochi Wei, Shunzhi Zhu, and Tat-Seng Chua. 2017. Embedding factorization models for jointly recommending items and user generated lists. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 585–594.

[3] Huizhong Chen, Andrew C. Gallagher, and Bernd Girod. 2012. Describing clothing by semantic attributes. In *Proceedings of the European Conference on Computer Vision*. Springer, 609–623.

[4] Peng Cui, Shaowei Liu, and Wenwu Zhu. 2018. General knowledge embedded image representation learning. *IEEE Trans. Multimedia* 20, 1 (2018), 198–207.

[5] Cunxiao Du, Zhaozheng Chin, Fuli Feng, Lei Zhu, Tian Gan, and Liqiang Nie. 2019. Explicit interaction model towards text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 6359–6366.

[6] Zunlei Feng, Zhenyun Yu, Yezhou Yang, Yongcheng Jing, Junxiao Jiang, and Mingli Song. 2018. Interpretable partitioned embedding for customized fashion outfit composition. In *Proceedings of the ACM International Conference on Multimedia Retrieval*. ACM, 143–151.

[7] Xintong Han, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S. Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1472–1480.

[8] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S. Davis. 2017. Learning fashion compatibility with bidirectional LSTMs. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1078–1086.
[9] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the International Conference on World Wide Web*. ACM, 173–182.
[10] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 549–558.
[11] Yonghao He, Shiming Xiang, Cuicui Kang, Jian Wang, and Chunhong Pan. 2016. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Trans. Multimedia* 18, 7 (2016), 1363–1377.
[12] Wei-Lin Hsiao and Kristen Grauman. 2018. Creating capsule wardrobes from fashion images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7161–7170.
[13] Yang Hu, Xi Yi, and Larry S. Davis. 2015. Collaborative fashion recommendation: A functional tensor factorization approach. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 129–138.
[14] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard H. Hovy, and Eric P. Xing. 2016. Harnessing deep neural networks with logic rules. In *Proceedings of the Meeting of the Association for Computational Linguistics*. The Association for Computer Linguistics, 2410–2420.
[15] Junshi Huang, Rogério Schmidt Feris, Qiang Chen, and Shuicheng Yan. 2015. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1062–1070.
[16] Dong Li, Ting Yao, Ling-Yu Duan, Tao Mei, and Yong Rui. 2019. Unified spatio-temporal attention networks for action recognition in videos. *IEEE Trans. Multimedia* 21, 2 (2019), 416–428.
[17] Linghui Li, Sheng Tang, Yongdong Zhang, Lixi Deng, and Qi Tian. 2018. GLA: Global-local attention for image description. *IEEE Trans. Multimedia* 20, 3 (2018), 726–737.
[18] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. 2017. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Trans. Multimedia* 19, 8 (2017), 1946–1955.
[19] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. 2018. Interpretable multimodal retrieval for fashion products. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1571–1579.
[20] Min Lin, Qiang Chen, and Shuicheng Yan. 2014. Network in network. In *Proceedings of the International Conference on Learning Representations*.
[21] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten de Rijke. 2019. Explainable fashion recommendation with joint outfit matching and comment generation. *IEEE Trans. Knowl. Data Eng.* 32, 8 (2019), 1502–1516.
[22] Jinhuan Liu, Xuemeng Song, Zhumin Chen, and Jun Ma. 2019. Neural fashion experts: I know how to make the complementary clothing matching. *Neurocomputing* 359 (2019), 249–263.
[23] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan. 2012. Hi, magic closet, tell me what to wear! In *Proceedings of the ACM International Conference on Multimedia*. ACM, 619–628.
[24] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1096–1104.
[25] Yi-Jie Lu, Linjun Yang, Kuiyuan Yang, and Yong Rui. 2015. Mining latent attributes from click-through logs for image recognition. *IEEE Trans. Multimedia* 17, 8 (2015), 1213–1224.
[26] Lei Ma, Hongliang Li, Fanman Meng, Qingbo Wu, and King Ngi Ngan. 2017. Learning efficient binary codes from high-level feature representations for multilabel image retrieval. *IEEE Trans. Multimedia* 19, 11 (2017), 2545–2560.
[27] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 43–52.
[28] Martin Mirakyan, Karen Hambardzumyan, and Hrant Khachatrian. 2018. Natural language inference over interaction space: ICLR 2018 reproducibility report. In *Proceedings of the International Conference on Learning Representations*.
[29] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 452–461.
[30] Steffen Rendle and Lars Schmidt-Thieme. 2010. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the Conference on Web Search and Web Data Mining*, Brian D. Davison, Torsten Suel, Nick Craswell, and Bing Liu (Eds.). ACM, 81–90.
[31] Sijie Song and Tao Mei. 2018. When multimedia meets fashion. *IEEE Trans. Multimedia* 25, 3 (2018), 102–108.
[32] Xuemeng Song, Fuli Feng, Xianjing Han, Xin Yang, Wei Liu, and Liqiang Nie. 2018. Neural compatibility modeling with attentive knowledge distillation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 5–14.

[33] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. 2017. NeuroStylist: Neural compatibility modeling for clothing matching. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 753–761.

[34] Xuemeng Song, Xianjing Han, Yunkai Li, Jingyuan Chen, Xin-Shun Xu, and Liqiang Nie. 2019. GP-BPR: Personalized compatibility modeling for clothing matching. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 320–328.

[35] Guang-Lu Sun, Zhi-Qi Cheng, Xiao Wu, and Qiang Peng. 2018. Personalized clothing recommendation combining user social circle and fashion style consistency. *Multimedia Tools Applic.* 77, 14 (2018), 17731–17754.

[36] Pongsate Tangseng and Takayuki Okatani. 2020. Toward explainable fashion recommendation. In *Proceedings of the Winter Conference on Applications of Computer Vision*. IEEE, 2153–2162.

[37] Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. In *Proceedings of the International Conference on Data Engineering Workshops*. IEEE, 801–810.

[38] Mariya I. Vasileva, Bryan A. Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David A. Forsyth. 2018. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European Conference on Computer Vision*. Springer, 405–421.

[39] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image captioning with deep bidirectional LSTMs. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 988–997.

[40] Qiurui Wang, Chun Yuan, Jingdong Wang, and Wenjun Zeng. 2019. Learning attentional recurrent neural network for visual tracking. *IEEE Trans. Multimedia* 21, 4 (2019), 930–942.

[41] Shuohang Wang and Jing Jiang. 2016. Learning natural language inference with LSTM. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. The Association for Computational Linguistics, 1442–1451.

[42] Xin Wang, Bo Wu, and Yueqi Zhong. 2019. Outfit compatibility prediction and diagnosis with multi-layered comparison network. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 329–337.

[43] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 496–505.

[44] Jun Xiao, Hao Ye, Xiangnan He, Hanwang Zhang, Fei Wu, and Tat-Seng Chua. 2017. Attentional factorization machines: Learning the weight of feature interactions via attention networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*. ijcai.org, 3119–3125.

[45] Xun Yang, Yunshan Ma, Lizi Liao, Meng Wang, and Tat-Seng Chua. 2019. TransNFCM: Translation-based neural fashion compatibility modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 403–410.

[46] Xin Yang, Xuemeng Song, Xianjing Han, Haokun Wen, Jie Nie, and Liqiang Nie. 2020. Generative attribute manipulation scheme for flexible fashion search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 941–950.

[47] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. 2013. Attribute-augmented semantic hierarchy: Towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 33–42.

[48] Yongfeng Zhang and Xu Chen. 2018. Explainable recommendation: A survey and new perspectives. arxiv:cs.IR/1804.11192.

[49] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2921–2929.

Received March 2020; revised August 2020; accepted September 2020