GP-BPR: Personalized Compatibility Modeling for Clothing Matching

Xuemeng Song
Shandong University
sxmuste@gmail.com

Xianjing Han
Shandong University
hanxianjing2018@gmail.com

Yunkai Li
Shandong University
liyunkai@mail.sdu.edu.cn

Jingyuan Chen
Alibaba Group
jingyuanchen91@gmail.com

Xin-Shun Xu
Shandong University
xuxinshun@sdu.edu.cn

Liqiang Nie∗
Shandong University
nieliqiang@gmail.com

ABSTRACT

Owing to the recent advances in the multimedia processing domain and the publicly available large-scale real-world data provided by online fashion communities, like the IQON and Chictopia, researchers are enabled to investigate the automatic clothing matching solutions. In a sense, existing methods mainly focus on modeling the general item-item compatibility from the aesthetic perspective, but fail to incorporate the user factor. In fact, aesthetics can be highly subjective, as different people may hold different clothing preferences. In light of this, in this work, we attempt to tackle the problem of personalized compatibility modeling from not only the general aesthetics but also the personal preference perspectives. In particular, we present a personalized compatibility modeling scheme GP-BPR, comprising of two essential components: general compatibility modeling and personal preference modeling, which characterize the item-item and user-item interactions, respectively. In particular, due to the concern that both the modalities (e.g., the image and context description) of fashion items can deliver important cues regarding user personal preference, we present a comprehensive personal preference modeling method. Moreover, for evaluation, we create a large-scale dataset, IQON3000, from the online fashion community IQON. Extensive experiment results on IQON3000 verify the effectiveness of the proposed scheme. As a byproduct, we have released the dataset, codes, and involved parameters to benefit other researchers.

CCS CONCEPTS
• Information systems → Personalization; Retrieval tasks and goals;

KEYWORDS
Fashion Analysis, Personalized Compatibility Modeling, Multimodal.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:

1 INTRODUCTION

Recent years have witnessed the flourishing of the online fashion industry, whose total global value is up to 3 trillion US dollars, amounting to two percent of the world’s Gross Domestic Product†. The huge economic value reflects people’s growing demand for dressing. In fact, clothing matching, to coordinate complementary fashion items such as the tops and bottoms to make proper outfits, has become an indispensable aspect of people’s daily life. Owing to the recent proliferation of fashion-oriented online communities (e.g., IQON‡ and Chictopia§), where users can create their favorite outfits by collocating the complementary fashion items and share with the public, as shown in Figure 1, many research efforts have been dedicated to exploring the automatic clothing matching task. In a sense, most of the existing work attempts to tackle the clothing matching problem by modeling the compatibility between fashion items from the aesthetic perspective based on the visual and contextual contents of fashion items, but overlooks the role of user factor. Indeed, aesthetics can be rather subjective, as different people may have different tastes in clothing matching. For example, for the same fashion item “high-neck pullover” occurred in the first

†https://fashionunited.com/global-fashion-industry-statistics/.
‡https://www.iqon.jp/.
outfit of all three users in Figure 1, user 1 coordinates it with the “point button tweed tight skirt”, while user 3 prefers to match it with the “check flare skirt with belt”. Consequently, it is inappropriate to ignore the user context factor and access the compatibility between fashion items universally across different individuals. To bridge this gap, this work aims to tackle the personalized clothing matching problem, where without loss of generality, we focus on the compatibility modeling between the top and bottom while considering the user context.

However, the personalized compatibility modeling between fashion items is non-trivial due to the following challenges. 1) Although there are many public datasets towards the general compatibility modeling and personalized fashion item recommendation tasks, respectively, there is a lack of the large-scale benchmark dataset for personalized compatibility modeling. Accordingly, how to construct a large-scale benchmark dataset to facilitate the evaluation of the proposed method constitutes a tough challenge. 2) How to seamlessly encode the user preference on clothing matching into the personalized compatibility modeling between fashion items and thus enable the matching results not only to meet the common matching patterns but also to cater to the user personal taste poses another challenge for us. And 3) fashion items can be comprehensively characterized by multiple modalities, such as the visual images and textual descriptions, both of which may convey important cues on user preferences. For example, the visual signal can reveal the intuitive features that the user prefers, like the color and shape, while the contextual modality may deliver the user preferred item brand or fabric. Therefore, how to fully take advantage of the multi-modal data in the context of the personalized clothing matching is a crucial challenge.

To address the aforementioned challenges, we present a personalized compatibility modeling scheme for clothing matching, named as GP-BPR, as shown in Figure 2, which is able to measure the compatibility between fashion items from not only the general aesthetics but also the personal preference perspectives. In particular, GP-BPR consists of two essential components: general compatibility modeling and personal preference modeling. The content-based general compatibility modeling works on learning the latent compatibility space shared by complementary items to characterize the item-item interactions towards clothing matching. Meanwhile, the personal preference modeling focuses on exploiting the latent preference factor based on the multi-modal data of fashion items and hence captures the user-item interactions comprehensively. Ultimately, based on the Bayesian Personalized Ranking (BPR) framework [32], GP-BPR jointly integrates the general compatibility and personal preference modeling. To facilitate the evaluation, we construct a large-scale dataset from the online fashion community IQON, which comprises 308,747 outfits created by 3,568 users with 672,335 fashion items.

Our main contributions can be summarized in threefold:

- We present a personalized compatibility modeling scheme for personalized clothing matching, GP-BPR, which is able to jointly model the general (item-item) compatibility and personal (user-item) preference. To the best of our knowledge, this is the first to incorporate user factor in clothing matching.

- Considering that both modalities of fashion items can deliver significant signals regarding user preferences, we introduce a comprehensive personal preference modeling scheme by integrating the multi-modal data of fashion items.

- Extensive experiments conducted on the real-world dataset demonstrate the superiority of the proposed scheme over the state-of-the-art methods. As a byproduct, we released the codes and involved parameters to benefit other researchers.

The remainder of this paper is structured as follows. Section 2 briefly reviews the related work. The proposed GP-BPR is introduced in Section 3. Section 4 details the dataset construction. Section 5 presents the experimental results and analyses, followed by our concluding remarks and future work in Section 6.

2 RELATED WORK

Owing to the recent booming of the fashion industry, increasing research attention from both the computer vision and multimedia communities has been paid to the fashion domain, especially the clothing matching problem [6–8, 22, 36], which is usually cast as the compatibility modeling task between complementary fashion items. For example, Li et al. [22] proposed an outfit quality predictor with the multi-modal multi-instance deep learning based on item appearances. In addition, Song et al. [36] introduced a content-based neural scheme towards the compatibility modeling between fashion items based on their multi-modal data. Later, Yang et al. [42] presented a translation-based neural fashion compatibility modeling framework, which jointly optimizes the fashion item embeddings and category-specific complementary relations in an end-to-end manner. Moreover, noticed that the fashion domain has accumulated various valuable knowledge that can be helpful to guide the compatibility modeling, Song et al. [35] shed light on integrating the rich fashion domain knowledge to the pure data-driven learning, where a neural compatibility modeling scheme with attentive knowledge distillation was presented. Although existing efforts have achieved compelling success, they mainly focused on modeling the compatibility between fashion items purely based on the general item-item compatibility and overlooked the user factor in the compatibility modeling, which is the major concern of our work.

In addition, personalized recommendation in fashion domain also gains great research attention [3, 12, 40]. In particular, existing personalized recommendation work in fashion domain [10, 13, 39] mainly utilized the matrix factorization (MF) framework to model user preferences based on their feedback with real-world datasets. For example, Hu et al. [13] proposed a functional tensor factorization model aiming to tackle the problem of personalized outfit recommendation based on a dataset comprising of 150 users. Although this method is effective in the whole outfit recommendation, the cold start problem constitutes a remaining issue that worths further exploring. Towards this end, He et al. [10] introduced a scalable matrix factorization model that incorporates the visual signal of items into the user preference predictors to fulfill the recommendation task. In a sense, existing efforts focus on exploring the latent user-item interactions to tackle the personalized recommendation problems. Beyond that, in this work,
we aim to fulfill the task of personalized clothing matching, where both the user-item preference and item-item compatibility need to be well explored.

3 METHODOLOGY

In this section, we first give the problem formulation and then detail the proposed personalized compatibility modeling scheme.

3.1 Problem Formulation

Formally, we first declare some notations. We use bold capital letters (e.g., \( \mathbf{X} \)) and bold lowercase letters (e.g., \( \mathbf{x} \)) to denote matrices and vectors, respectively. We employ the non-bold letters (e.g., \( x \)) to represent scalars and Greek letters (e.g., \( \beta \)) to denote the parameters. If not clarified, all vectors are in the column forms. \( \| A \|_F \) denotes the Frobenius norm of matrix \( A \).

Suppose we have a set of users \( \mathcal{U} = \{ u_1, u_2, \cdots, u_M \} \), a set of tops \( \mathcal{T} = \{ t_1, t_2, \cdots, t_{N_T} \} \) and a set of bottoms \( \mathcal{B} = \{ b_1, b_2, \cdots, b_{N_B} \} \), where \( M, N_T \) and \( N_B \) denote the total numbers of users, tops and bottoms, respectively. Each user \( u_m \) is associated with a set of historically composed top-bottom pairs \( \mathcal{O}_m = \{ (t_{i_1}^{m}, b_{j_1}^{m}), (t_{i_2}^{m}, b_{j_2}^{m}), \cdots, (t_{i_{N_T}}^{m}, b_{j_{N_B}}^{m}) \} \), where \( i_k^{m} \in [1, 2, \cdots, N_{T}] \) and \( j_k^{m} \in [1, 2, \cdots, N_{B}] \) refer to the index of the top and bottom. For each \( t_i (b_j) \), we use \( \mathbf{v}_i (\mathbf{v}_j) \in \mathbb{R}^{D_v} \) and \( \mathbf{c}_i (\mathbf{c}_j) \in \mathbb{R}^{D_c} \) to represent its visual and contextual embeddings, respectively. \( D_v \) and \( D_c \) denote the dimensions of the corresponding embeddings.

As a matter of fact, different people may have different fashion tastes and thus prefer different clothing items to make favorable outfits. Accordingly, in this work, we aim to tackle the essential compatibility modeling between fashion items for clothing matching by taking the user factor into account. Without loss of generality, we particularly investigate the problem of “which bottom would be preferred by the user to match the given top?” Let \( p_{ij}^{m} \) denote the preference of the user \( u_m \) towards the bottom \( b_j \) for top \( t_i \), based on which we can generate a personalized ranking list of bottoms \( b_j \)'s for a given top \( t_i \) and hence solve the practical problem of personalized clothing matching. In particular, to accurately measure \( p_{ij}^{m} \), we focus on devising a personalized compatibility modeling network \( \mathcal{F} \), which is capable of compiling the user preference context into the compatibility modeling between fashion items as follows,

\[
 p_{ij}^{m} = \mathcal{F}(t_i, b_j, u_m|\Theta_{\mathcal{F}}),
\]

where \( \Theta_{\mathcal{F}} \) refers to the to-be-learned model parameters.

3.2 GP-BPR

In a sense, towards personalized clothing matching (e.g., matching a bottom for a user’s top), it is natural to incorporate both the item-item compatibility and the user-item preference. In light of this, we measure the user preference towards a bottom for a given to-be-matched top based on both the general compatibility modeling and the personal preference modeling. Formally, we have,

\[
 \begin{align*}
 p_{ij}^{m} &= \mu \cdot s_{ij} + (1 - \mu) \cdot c_{mj}, \\
 s_{ij} &= \mathcal{G}(t_i, b_j|\Theta_{\mathcal{G}}), \\
 c_{mj} &= \mathcal{P}(u_m, b_j|\Theta_{\mathcal{P}}),
\end{align*}
\]

where \( \mathcal{G} \) and \( \mathcal{P} \) correspond to the general compatibility modeling and personal preference modeling networks, respectively. \( \Theta_{\mathcal{G}} \) and \( \Theta_{\mathcal{P}} \) are the corresponding model parameters. \( s_{ij} \) denotes the general compatibility between the top \( t_i \) and bottom \( b_j \), while \( c_{mj} \) represents the personal preference of user \( u_m \) towards the bottom \( b_j \). \( \mu \) is the non-negative tradeoff parameter to control the relative importance of both components.

Figure 2: Illustration of the proposed scheme. The general compatibility modeling aims to learn the visual and contextual latent embedding of the items. The personal compatibility modeling focuses on exploiting the latent user-item interaction factors to capture the user preference. These two components are integrated by the BPR framework.
where \( \pi \) is the non-negative trade-off parameter, calibrating the relative importance of the modalities. \( s_{ij} \) denotes the general compatibility between the top \( t_i \) and bottom \( b_j \).

### 3.2.2 Personal Preference Modeling

As for the personal preference modeling towards a bottom, we resort to the matrix factorization framework, which has shown great success in personalized recommendation tasks [1, 18, 21, 31]. The underlying philosophy is to decompose the user-item interaction matrix into the latent user factors and item factors, whose inner product encodes the user-item interaction scores. In our context, we model the user preference towards a bottom as follows,

\[
 c_{mj} = \alpha + \beta_m + \beta_j + \gamma_m^T \gamma_j. 
\]  

where \( c_{mj} \) represents the preference of user \( u_m \) for bottom \( b_j \). \( \alpha \) is the to-be-learned global offset, \( \beta_m \) and \( \beta_j \) are the user \( u_m \) and bottom \( b_j \) bias terms. \( \gamma_m \) and \( \gamma_j \) are the latent factors of user \( u_m \) and bottom \( b_j \), respectively, whose inner product captures the latent preference of user \( u_m \) for the bottom \( b_j \).

Apart from the latent overall preference factors, inspired by [10], we also incorporate the latent content-based preference factors. The philosophy behind lies in that the user preference for a fashion item may come from the visual characteristics, like the color and shape, or the contextual features, like the brand and material. Different from [10], we take into account of not only visual modality but also contextual modality of fashion items to comprehensively measure the user-item interactions. Accordingly, incorporating the latent visual and contextual preference factors to the matrix factorization framework, we have

\[
 c_{mj} = \alpha + \beta_m + \beta_j + \gamma_m^T \gamma_j + \eta \xi_m^T \xi_j + (1 - \eta) \xi_m^T \xi_j, 
\]  

where \( \xi_m \) and \( \xi_j \) are the latent visual factors of user \( u_m \) and bottom \( b_j \). The inner product between them conveys the visual preference interaction between the user \( u_m \) and bottom \( b_j \). Similarly, \( \xi_m^c \) and \( \xi_j^c \) stand for the latent contextual factors of user \( u_m \) and bottom \( b_j \), respectively, which compile the contextual preference interaction. In this work, we make \( \xi_j^v = \hat{v}_j^b \) and \( \xi_j^c = \hat{c}_j^b \), where \( \hat{v}_j^b \) and \( \hat{c}_j^b \) are the latent embeddings for the visual and contextual representations of bottom \( b_j \). \( \eta \) is the non-negative tradeoff parameter.

### 3.2.3 Optimization

To accurately model the implicit interaction among users and fashion items (i.e., tops and bottoms), we adopt the BPR framework, which has proven to be powerful in the pairwise implicit preference modeling [2, 27, 29]. In particular, we first construct the following training set \( \mathcal{D} := \{(m, i, j, k) | u_m \in \mathcal{U} \land \{t_i, b_j\} \in \mathcal{O}_m \land b_k \in \mathcal{B} \setminus \{b_j\} \} \), where the quadruplet \( (m, i, j, k) \) indicates that to match the given top \( t_i \) and make a proper outfit, the user \( u_m \) prefers the bottom

---

**Algorithm 1 Personalized Compatibility Modeling.**

**Input:** Training set \( \mathcal{D} = \{(m, i, j, k)\} \), learning rate \( \rho \), regularization parameter \( \lambda \), trade-off parameters \( \pi, \eta \) and \( \mu \).

**Output:** Parameters \( \Theta_F \).

1. Initialize parameters \( \Theta_F \).
2. repeat
   3. Draw \((m, i, j, k)\) from \( \mathcal{D} \).
   4. Compute \( p_{ij}^m \) according to Eqn. (2).
   5. for each parameter \( \theta \) in \( \Theta_F \) do
      6. Update \( \theta \leftarrow \theta + \rho (\mathcal{L}(\theta, \mathcal{D}) - \lambda \theta) \).
   7. end for
3. until Converge
Table 1: The number of items of each category.

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
<th>Category</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outerwear</td>
<td>35,765</td>
<td>Top</td>
<td>119,895</td>
</tr>
<tr>
<td>Bottom</td>
<td>77,813</td>
<td>Shoes</td>
<td>106,598</td>
</tr>
<tr>
<td>One Piece</td>
<td>25,816</td>
<td>Accessories</td>
<td>306,448</td>
</tr>
</tbody>
</table>

$b_j$ to $b_k$. Then according to the BPR loss [32], we thus have the following objective function,

\[
L = \sum_{(m,i,j,k) \in \mathcal{D}} l_{BP}(p_{ij}^m, p_{ik}^m) + \lambda \left\| \Theta_F \right\|^2_F,
\]

where $\lambda$ is the non-negative hyperparameter, the last term is designed to avoid overfitting and $\Theta_F$ refers to the set of parameters (i.e., $W^F_m$, $b^F_m$, $\alpha$, $\beta_j$, $Y_m$, $Y_j$, $\xi_{mk}^m$ and $\xi_{jk}^m$) of the model. Figure 3 illustrates the workflow of our model, and the optimization procedure of our framework is summarized in Algorithm 1.

4 DATASET

In fact, several fashion datasets have been collected for different research purposes, for instance, the WoW [26], Fashion-136k [15], Amazon [30], DeepFashion [28], PolyvoreDataset [8], and FashionVC [36]. However, most of the existing publicly available datasets lack the user context, which makes it intractable to tackle the personalized clothing matching problem. It is worth noting that although the dataset Amazon [30] contains the valuable user contexts but it focuses more on the item recommendation based on the user preference and hence lacks the ground truth regarding the coordination among fashion items. Moreover, the dataset used in [13] contains only 150 users, which hinders the practical evaluation. Therefore, to bridge this gap, we created a new large dataset for personalized clothing matching. In particular, we crawled our data from the popular fashion web service IQON.

In particular, we first collected a set of popular outfits on IQON as the seeds, and by tracking them, we obtained 6,191 users. Thereafter, we crawled the latest 500 historical outfits of each user due to the following twofold concerns. 1) Extremely active users have created thousands of outfits, where according to our pilot study, the most active user has 4,562 outfits, and would result in the imbalanced dataset. And 2) users’ tastes on clothing matching may shift gradually and it thus should be more reasonable to be reflected by their latest outfits. To ensure the quality of the dataset, we filtered out the users with less than 5 historical outfits and only retained the items belonging to the six common categories: Coat, Top, Bottom, One Piece\(^6\), Shoes and Accessories. Thereafter, we obtained the dataset, IQON3000, comprising 308,747 outfits created by 3,568 users with 672,335 fashion items. Table 1 lists the statistics of our dataset. For each fashion item, we particularly crawled its profile, including the visual image, categories, attributes and item description, as shown in Figure 4. In addition, each outfit is associated with its price and number of likes.

\(^6\)One piece refers to the dress and tunic.

Figure 4: Screenshot of the item profile. We particularly collected the information highlighted by the boxes. Notably, the text has been translated for illustration.

5 EXPERIMENT

To evaluate the proposed method, we conducted extensive experiments on the real-world dataset IQON3000 by answering the following research questions:

- Does the proposed GP-BPR achieve better performance than the state-of-the-art methods?
- What is the contribution of the personal preference modeling as compared to that over the general compatibility?
- How do GP-BPR perform in the application of the personalized complementary fashion item retrieval?

5.1 Implementation

Contextual Representation. As a pioneering attempt of the personalized clothing matching, here we only consider the title description and category metadata as the contextual information of the fashion item. We first tokenized the text with the help of the Japanese morphological analyzer Kuromoji\(^7\). To obtain the effective contextual representation, instead of the traditional linguistic features [37, 38], we adopted the CNN architecture [19], which has achieved compelling success in various natural language processing tasks [14, 33]. In particular, we first represented each contextual description as a concatenated word vector, where each row represents one constituent word. To represent each word, we employed the 300-D vector provided by the Japanese word2vec word2vec in the search mode, which is created from NIHJAL Web Japanese Corpus [34]. We then deployed the single channel CNN, consisting of a convolutional layer on top of the concatenated word vectors and a max pooling layer. In particular, we used four kernels with sizes of 2, 3, 4, and 5, respectively. For each kernel, we had 100 feature maps. We employed the rectified linear unit (ReLU) as the activation function. Ultimately, we obtained a 400-D contextual representation for each item.

Visual Representation. Regarding the visual modality, we applied the deep CNNs, which has proven to be the state-of-the-art model for image representation learning [4, 17, 23, 24]. In particular, we chose the 50-layer residual network (ResNet50) in [9]. We fed the image of each fashion item to the network, and adopted the

\[^7\]http://www.atulika.org/
output of the last average pooling layer as the visual representation. Thereby, we represented the visual modality of each item with a 2048-D vector.

**Experiment Settings.** In our context of matching bottoms for a given top, we only considered the outfits that either contain a top and a bottom, or a coat plus a bottom/dress, where we treated the coat as the 'top' while the bottom/dress as the 'bottom'. As one user may coordinate different shoes or accessories for the same top-bottom pair to make different outfits, we removed the duplicated top-bottom pairs from the dataset, resulting in 217,806 unique top-bottom pairs.

Regarding the evaluation, we adopted the leave-one-out strategy, where we randomly sampled one top-bottom pair for each user and retained it as the testing sample. Then we generated the quadruple set \( D_{\text{train}}, D_{\text{valid}} \) and \( D_{\text{test}} \) according to Eqn.(7), where for each positive top-bottom pair \((t_i, b_j)\) of the user \( u_m \), we randomly sampled a negative bottom \( b_k \) from the whole bottom dataset (i.e., \( B \)) to comprise a quadruplet \((m, i, j, k)\). Finally, we adopted the area under the ROC curve (AUC) [45] as the evaluation metric.

For optimization, we employed the adaptive moment estimation method (Adam) [20]. We adopted the grid search strategy to determine the optimal values for the regularization parameter \( \lambda \) and trade-off parameters \((\pi, \eta, \mu)\). In addition, the mini-batch size, the number of hidden units and learning rate were searched in \([32, 64, 128], [256, 512, 1024], \) and \([0.0005, 0.001, 0.005, 0.01]\), respectively. The proposed model was fine-tuned for 40 epochs, and the performance on the testing set was reported. We empirically set the number of hidden layers in representation learning \( K = 1 \).

### 5.2 On Model Comparison (RQ1)

We chose the following state-of-the-art methods as the baselines to evaluate the proposed model.

- **POP-T**: We used the “popularity” of the bottom to measure its compatibility with top, which is defined as the number of outfits that the bottom appeared in the training set.
- **POP-U**: Similarly, in this baseline, we defined the “popularity” of the bottom as the number of users who once interacted with the bottom in the training set.
- **RAND**: We randomly assigned the compatibility scores of \( m_{ij} \) and \( m_{ik} \) between items.
- **Bi-LSTM**: We used the bidirectional LSTM method in [8] which sequentially models the outfit compatibility by predicting the next item conditioned on previous ones. Here, we adapted Bi-LSTM to deal with an outfit comprising only a top and a bottom.
- **BPR-DAE**: We selected the content-based neural scheme introduced by [36] that is capable of jointly modeling the coherent relation between different modalities of fashion items and the implicit preference among items via a dual autoencoder network. It is worth noting that BPR-DAE overlooks the user factor in the compatibility modeling.
- **BPR-MF**: We used the pairwise ranking method introduced in [32], where the latent user-item relations are captured by the MF method.
- **VBPR**: We adopted the VBPR in [10], which exploits the visual data of fashion items with the factorization method to recommend an item for the user.
- **TBPR**: We derived TBPR from VBPR by replacing the visual signals with the textural modality of fashion items.
- **VTBPR**: We extended VBPR in [10] by further introducing the context factor to comprehensively characterize the user’s preference from both the visual and contextual perspectives.

Table 2 shows the performance comparison among different approaches. From this table, we have the following observations: 1) BPR-DAE shows superiority over Bi-LSTM, which implies that the content-based scheme performs better than the sequential model in the general compatibility modeling between fashion items. 2) VTBPR outperforms VBPR, TBPR and BPR-MF, which confirms the advantage of considering both the visual and contextual modalities in the personal preference modeling. Interestingly, we found that TBPR slightly surpasses VBPR, demonstrating the great potential of contextual data in characterizing users’ personal preference of items. 3) GP-BPR achieves better performance than all the other methods that focus on either the general compatibility modeling or person preference modeling, validating the necessity of incorporating both the general item-item compatibility and user-item preference in the context of personalized clothing matching.

To evaluate the contribution of each modality in our model, we further compared GP-BPR with its two derivatives: GP-BPR-V and GP-BPR-T, where only the visual and contextual modality of fashion items were explored, respectively. Table 3 shows the performance comparison of our model with different modalities. We observed that our model outperforms both GP-BPR-V and GP-BPR-T, which suggests that the visual and contextual signals do complement each other and both contribute to the personalized compatibility modeling. In addition, similar to above TBPR and VBPR, we found that GP-BPR-T achieves better performance than GP-BPR-V. This may be due to two reasons: 1) The contextual information of fashion items can summarize the key features, such as the pattern and

<table>
<thead>
<tr>
<th>Approach</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>POP-T</td>
<td>0.6042</td>
</tr>
<tr>
<td>POP-U</td>
<td>0.5951</td>
</tr>
<tr>
<td>RAND</td>
<td>0.5014</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>0.6739</td>
</tr>
<tr>
<td>BPR-DAE</td>
<td>0.7096</td>
</tr>
<tr>
<td>BPR-MF</td>
<td>0.7958</td>
</tr>
<tr>
<td>VBPR</td>
<td>0.8170</td>
</tr>
<tr>
<td>TBPR</td>
<td>0.8190</td>
</tr>
<tr>
<td>VTBPR</td>
<td>0.8232</td>
</tr>
<tr>
<td>GP-BPR</td>
<td>0.8388</td>
</tr>
</tbody>
</table>

Table 3: Performance comparison among different modalities in terms of AUC.

<table>
<thead>
<tr>
<th>Approach</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP-BPR-V</td>
<td>0.8239</td>
</tr>
<tr>
<td>GP-BPR-T</td>
<td>0.8313</td>
</tr>
<tr>
<td>GP-BPR</td>
<td>0.8388</td>
</tr>
</tbody>
</table>
material of fashion items more concisely. And 2) the contextual data usually convey some high-level semantic cues, like the item brand, which obviously can facilitate not only the personal preference modeling but also the general compatibility modeling, as items of the same brand are more likely to be compatible.

5.3 On Component Comparison (RQ2)

To gain a better understanding with respect to the contribution of each component in our model, we introduced two derivatives: G-BPR and P-BPR, where we only consider the general compatibility and personal preference modeling component of our model, respectively. Figure 5 shows the performance of our model with different component configurations. It can be seen that our model surpasses the derivative models, confirming the importance of each component in our model. In addition, interestingly, we noticed that P-BPR outperforms G-BPR, which suggests that the personal preference is the dominant factor affecting the individual's personalized clothing matching. To gain more detailed insights, we further checked the performance of our model with different components on seven popular bottom categories. As shown in Figure 5, GP-BPR outperforms the G-BPR and P-BPR consistently across different bottom categories, which reconfirms the effects of both two components. In addition, it is interesting to observe that by incorporating the general compatibility modeling, GP-BPR achieves the greatest improvement over the pure personal preference modeling component P-BPR in terms of the category "Long Skirt". One plausible explanation is that long skirts are usually critical of tops to make compatible outfits. Accordingly, taking the general compatibility modeling into account can boost the performance of P-BPR significantly. On the contrary, even with the help of the general compatibility modeling, GP-BPR shows limited superiority over P-BPR regarding the category "Denim Pants". That can be attributed to the fact that denim pants can go with various tops, ranging from coats to T-shirts, which makes incorporating the general item-item compatibility less helpful.

Moreover, we also illustrate the performance of our GP-BPR with respect to the trade-off parameter $\mu$ in Figure 6, where $\mu$ represents the weight of the general compatibility modeling component. As we can see, when $\mu = 0.3$ and P-BPR gets a higher weight than G-BPR, our GP-BPR achieves the optimal performance, indicating the dominant effect of P-BPR to GP-BPR. In addition, we noticed that when the value of $\mu$ ranges from 0.8 to 1.0 and our GP-BPR degenerates into the G-BPR, there is a sharp performance decrease on GP-BPR. In a sense, this is consistent with the above observation that the general compatibility modeling component alone suffers from the poor performance in the context of personalized compatibility modeling.

To intuitively show the impact of both components, we further illustrate the comparison among G-BPR, P-BPR and GP-BPR with several testing quadruplets in Figure 7. Notably, as aforementioned, each testing quadruplet $(m, i, j, k)$ indicates that the user $u_m$ prefers the bottom $b_j$ than $b_k$ to match the given top $t_i$. As we can see, bottoms $b_j$ and $b_k$ in the first example of user2 share the similar style with the items in the user’s historical preference, making the user preference to these two bottoms hard to tell and resulting the failure of P-BPR. However, taking the general item-item compatibility into account, where the "Ocean logo T-shirt" seems to go better with the shorts rather than the long jeans, GP-BPR can get the correct evaluation result. In addition, we also found that the personal preference can boost the performance especially when the general compatibility is hard to model. As can be seen, in the second example of user1, the general compatibility between the top and bottom candidates should be difficult to distinguish. Fortunately, resorting to the historical preference of user1, our GP-BPR can also reach the right result. Overall, both the general compatibility modeling and personal preference modeling are pivotal in our model and the cooperation of these two components can boost the performance of each component.

5.4 On Fashion Item Retrieval (RQ3)

To assess the practical value of our work, we evaluate our model towards the personalized complementary fashion item retrieval. Similar to [11], we fed each user-top pair $(u_m, t_i)$ in $D_{test}$ as the query and randomly selected $T$ bottoms as the ranking candidates with only one positive (ground truth) bottom. Thereafter, by passing them to the trained models and calculating the compatibility score, we generated a ranking list of these bottoms for each query. In our setting, we focused on the average position of the positive bottom in the ranking list and thus adopted the mean reciprocal rank (MRR) metric [16, 43, 44].

Figure 8 shows the performance of different models in terms of MRR at different numbers of the bottom candidates $T$. As
Figure 7: Illustration of the effect of the general compatibility and personal preference modeling. All the quadruplets satisfy the ground truth that \( u_{mt}, t_j \triangleright b_k \). “G”, “P” and “GP” are the abbreviations for G-BPR, P-BPR and GP-BPR, respectively. We represent the correct judgments of the model with the green circle and that of the wrong with the red cross.

Figure 8: Performance of different approaches with respect to MRR at different numbers of the bottom candidates \( T \).

can be seen, our GP-BPR shows superiority over all the other baselines consistently at different numbers of bottom candidates, demonstrating the effectiveness of our model in the personalized complementary fashion item retrieval. Moreover, to get a better understanding of our GP-BPR in this context, in Figure 9, we listed the ranking results of GP-BPR and its derivatives G-BPR and P-BPR for a given query. For the query "red knit pullover", G-BPR that simply relies on the general compatibility modeling, does rank the compatible bottoms at first places, including the positive one. Then further taking the user (historical) preference factor into account, we found that GP-BPR can boost the rank of the positive bottom from the fourth place to the first one, which verifies the importance of the user factor.

Figure 9: Illustration of the ranking results. The bottoms highlighted in the red boxes are the positive ones.

the visual and contextual modalities of fashion items can deliver valuable information regarding personal preference, we integrate the visual and contextual data of fashion items into the personal preference modeling. Moreover, we create a large-scale real-world dataset, IQON3000, which has been released to benefit the research community. Extensive experiments have been conducted on the created dataset IQON3000. The encouraging experiment results verify the effectiveness of the proposed scheme and indicate the necessity of integrating both the general item-item compatibility and personal user-item preference in the context of personalized clothing matching. One limitation of our work is that currently we fuse the two components of general compatibility modeling and personalized preference modeling linearly. In the future, we plan to devise a more advanced fusion strategy, such as the attentive fusion, to boost the performance.

6 CONCLUSION AND FUTURE WORK

In this work, we present a personalized compatibility modeling scheme towards personalized clothing matching, termed GP-BPR, which measures the compatibility between fashion items from not only the general aesthetics but also the personal preference perspectives. In particular, motivated by the fact that both modalities (i.e.,