applied sciences

MDPI

*Article*

# Attention-Based Personalized Compatibility Learning for Fashion Matching

**Xiaozhe Nie [1], Zhijie Xu [1,*], Jianqin Zhang [2] and Yu Tian [1]**

[1] School of Science, Beijing University of Civil Engineering and Architecture, Beijing 102616, China; 2102520021005@stu.bucea.edu.cn (X.N.); 2107010520003@stu.bucea.edu.cn (Y.T.)
[2] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 102616, China; zhangjianqin@bucea.edu.cn
* Correspondence: xuzhijie@bucea.edu.cn

**Abstract:** The fashion industry has a critical need for fashion compatibility. Modeling compatibility is a challenging task that involves extracting (in)compatible features of pairs, obtaining compatible relationships between matching items, and applying them to personalized recommendation tasks. Measuring compatibility is a complex and subjective concept in general. The complexity is reflected in the fact that relationships between fashion items are determined by multiple matching rules, such as color, shape, and material. Each personal aesthetic style and fashion preference differs, adding subjectivity to the compatibility concept. As a result, personalized factors must be considered. Previous works mainly utilize a convolutional neural network to measure compatibility by extracting general features, but they ignore fine-grained compatibility features and only model overall compatibility. We propose a novel neural network framework called the Attention-based Personalized Compatibility Embedding Network (PCE-Net). It comprises two components: attention-based compatibility embedding modeling and attention-based personal preference modeling. In the second part, we utilize matrix factorization and content-based features to obtain user preferences. Both pieces are jointly trained using the BPR framework in an end-to-end method. Extensive experiments on the IQON3000 dataset demonstrate that PCE-Net significantly outperforms most baseline methods.

**Keywords:** fashion analysis; personalized compatibility embedding modeling; attention mechanism; multi-modal



**Citation:** Nie, X.; Xu, Z.; Zhang, J.; Tian, Y. Attention-Based Personalized Compatibility Learning for Fashion Matching. *Appl. Sci.* **2023**, *13*, 9638. https://doi.org/10.3390/app13179638

Academic Editors: Konstantinos Pliakos and Alireza Gharahighehi

Received: 27 July 2023
Revised: 17 August 2023
Accepted: 23 August 2023
Published: 25 August 2023



## 1. Introduction

The emergence of e-commerce has provided convenient shopping methods. However, the overload of data information caused by the vast amounts of products available on shopping websites has led to the need for a recommendation system to help users find items more quickly and accurately. We focus on developing an intelligent recommendation algorithm for clothing to address this demand. One of the primary challenges is generating reasonable matching suggestions for clothing styles and types. This requirement gives rise to the need for fashion compatibility modeling, which helps determine if fashion items meet specific matching criteria. As illustrated in Figure 1, compatible items satisfy particular rules, such as having matched colors and materials, while incompatible items violate those rules. Moreover, users have individual preferences, including style, texture, pattern, and more. For instance, user1 prefers pairing casual, loose tops with wide-legged pants, user 2 has a versatile fashion taste that ranges from casual sports to elegant dresses, and user 3 enjoys wearing clothes with striped patterns.

Initially, some studies only considered visual features of fashion items [1–4] when building comparison models. Subsequently, several works [5–8] modeled compatibility by fusing both visual and textual multi-modal content. Further, researchers [5,6,9,10], distinguished overall compatibility from fine-grained compatibility. Some recent studies [11–15], have considered user factors in personalized recommendation tasks. Although these works

have individual strengths, they fail to provide a comprehensive solution that addresses all the underlying problems. We aim to develop a personalized clothing recommendation system that takes visual and textual modalities as input, extracts fine-grained compatibility characteristics, and considers user preferences.


**Figure 1.** Example of fashionable outfits from IQON3000.

In fashion recommendation, the primary challenge lies in accurately predicting and providing reasonable suggestions that align with a user's preferences. Researchers must concentrate on two significant issues: firstly, how to enhance the accuracy of determining fashion item compatibility from multi-modal data. Fashion recommendation is founded on the principle of fashion compatibility, which implies that various types of fashion items can be combined to create an outfit. Developing compatible feature spaces is crucial for continually advancing fashion compatibility models. For instance, researchers often employ visual features [1,9], textual features [16], category-aware feature subspaces [5,6], and neighbor node features of graph models [13,17] as inputs. Secondly, since an outfit typically comprises multiple complementing fashion items, selecting items that satisfy the user's preferences while complementing each other is the crux of outfit construction. However, fashion is an inherently intricate and subjective concept, and defining fashion items frequently involves many complex intersubjective relationships between complementary fashion items. It is critical to note that the notion of compatibility between fashion items often spans categories and encompasses intricate interactions.

To address the abovementioned challenges, our solution is an attention-based personalized compatibility embedding network called PCE-Net for clothing matching (Figure 2). This network can evaluate the compatibility between fashion items while capturing the user's personal preference from multi-modal features and their previous preferences. The two main components of the PCE-Net include attention-based compatibility embedding modeling and attention-based personal preference modeling. To overcome the first challenge, we introduce two attention branches to model fine-grained compatibility for the multi-modal data. To address the second challenge, we are inspired by the personal preference component of GP-BPR [11] and modeled attention-based personalized preference,

which utilizes global latent preference factors and content-based preference factors. We introduced feature extractors and two attention branches to learn the compatibility embeddings of fashion items. We also learned user preference by matrix factorization and inner product. Finally, based on the Bayesian Personalized Ranking (BPR) framework [18], PCE-Net integrates attention-based compatibility embedding modeling and attention-based personal preference modeling.

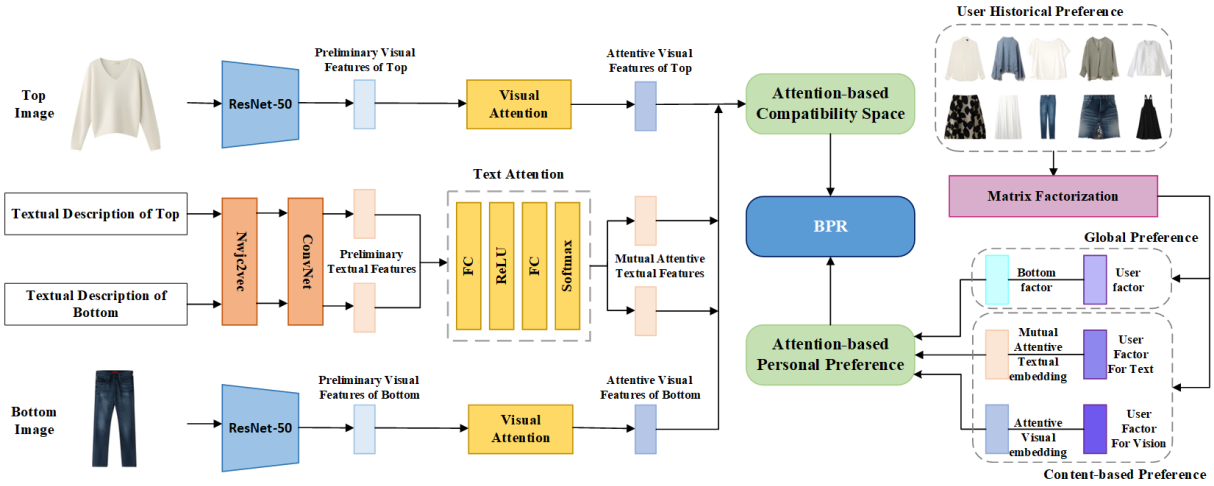


**Figure 2.** Overview of the proposed attention-based compatibility embedding network (PCE-Net) architecture. PCE-Net contains two parts: (1) attention-based compatibility embedding modeling based on feature extractor with two attention branches; (2) attention-based personal preference modeling based on matrix factorization and content-based inner product for user historical preferences.

Our model has extensive application prospects on e-commerce platforms and social networks. On e-commerce platforms, it can evaluate users' personalized preferences based on their purchase history and browsing history. Then, based on the clothing the user is currently browsing, it can recommend items that match the current outfit and also cater to the user's personalized preferences. On social platforms, it can provide users with compatibility assessment functionality, calculate compatibility scores based on the outfits selected by users, and provide adjustment suggestions.

Our main contributions can be summarized in threefold:

1. First itemFirstly, we present an attention-based personalized compatibility embedding scheme for personalized clothing matching, namely PCE-Net, which jointly models attention-based (item-item) compatibility and personal (user-item) preferences;
2. Secondly, we propose an innovative approach to capture the compatibility embeddings of multi-modal data using different attention branches separately, which has not been attempted before to the best of our knowledge. In addition, we demonstrate the effectiveness of different attention branches through ablation experiments;
3. Lastly, we conduct extensive experiments and use t-SNE visualization on the real-world dataset IQON3000 to validate the effectiveness of our scheme against state-of-the-art methods.

The remainder of this paper is structured as follows. We briefly review the related work in Section 2. In Section 3, we present the proposed PCE-Net in detail. The experimental results and relative analysis are provided in Section 4, followed by our concluding remarks and future work in Section 5.

## 2. Related Work

The area of fashion compatibility learning and recommendation employs two main categories of algorithms: item-item level and multi-item level approaches. The former category models compatibility interaction between two items. For example, Rendel et al. [19]

proposed the pairwise interaction tensor decomposition (PITF) factorization model, which simulates pairwise interactions between users, items, and tags. Meanwhile, refs. [20,21] utilized a single latent style space to measure compatibility solely using visual features. However, a single latent compatibility space is inadequate for the detailed modeling of complex relationships among different concepts, such as color, pattern, and category. Thus, to address this gap, Veit et al. [1] proposed the Conditional Similarity Network (CSN) model, which learns different subspaces under different similarity metrics. Vasileva et al. [5] then built on [1] by introducing a type-aware learning method to attain type-aware features in a universally shared potential space. Lastly, Katrien Laenen et al. [9] incorporated three attention mechanisms for multi-modal features in the type-aware subspace, building on the previous works. When predicting the compatibility between items, concatenation is adopted and implemented in a superior scheme. To automatically recognize the relative significance of different conditions, Tan et al. [6] leveraged the attention mechanism [22]. In a different approach, Cucurull et al. [23] suggested a graph neural network utilizing undirected graphs augmented with contextual information to predict associations between two items, thereby transforming the fashion compatibility issue into a graph edge detection problem. Singhal et al. [7] proposed a holistic approach to learning visual compatibility, encompassing TC-GAE, SAE, and search techniques modeled on a graph-based network, an autoencoder, and reinforcement learning, respectively. Finally, Song et al. [16] introduced Dual Autoencoder Network (DAE) as the first model to learn the compatibility feature space while incorporating the consistent relationship between visual-textual features and the implicit preference between items via Bayesian Personalized Ranking (BPR). Song et al. [24] presented AKD-DBPR, which employs knowledge distillation to combine fashion domain expertise with deep neural networks. Yang et al. [25] introduced TransNFCM, a fashion neural compatibility model based on translations that aim to capture complex compatibility patterns via distance functions. Lu et al. [26] developed a method for personalized outfit recommendation by training hash codes for both users and items and modeling users' preferences as the average of their preference scores for each item. Song et al. [11] proposed a joint model for general compatibility and personal preferences called GP-BPR, which combines the two characteristics. Sagar et al. [12] proposed a personalized recommendation modeling scheme named PAI-BPR, which utilizes attributes for interpretability and personalized recommendation. Finally, Taraviya et al. [27] introduced PSA-Net, which learns attribute-wise visual feature subspaces via self-attention and incorporates customer embeddings to aid in recommending item pairs in a category-based subspace.

The second category of outfit recommendation models aims to capture the interactions between multiple items in an outfit composed of three or more items. An early model developed by Han et al. [8] used a sequence approach, treating the items within an outfit as ordered, and proposed the Bi-LSTM sequence model. However, this model poses a limitation as it is order-sensitive. In contrast, Cui et al. [17] introduced the NGNN model, where a directed graph represents the complex relationships between multiple items in an outfit, providing a better model for data representation. For personalized outfit recommendation, Rendle et al. [18] proposed the widely used Matrix Factorization (MF) model, while He et al. [2] introduced the VBPR model, which is based on Bayesian Personalized Ranking [18] and incorporates user preferences for visual factors. Furthermore, He et al. [2] developed FashionNet, which recommends outfits (top, bottom, shoes) using a two-stage training strategy, where a general compatibility model with personal preferences is fine-tuned using encoding techniques. Li et al. [13] proposed the hierarchical fashion graph network (HFGN) model, which combines compatibility modeling and personalized outfit recommendation tasks. To generate personalized outfits based on users' historical click behaviors, Xu et al. [14] developed the personalized outfit generation (POG) model, which utilizes the Transformer [22] architecture to encode users' preferences for items and outfits. Dong et al. [15] proposed a personalized capsule closet creation framework (PCW-DC) based on the Bi-LSTM [8], which learns outfit compatibility, user preference, and body type information concurrently. In addition, Lin et al. [10] presented the neural outfit recommen-

dation (NOR) model, a neural network framework capable of simultaneously addressing the tasks of outfit recommendation and comment generation. The framework consists of an outfit-matching framework and a comment-generation framework. For outfit complementary item retrieval, Lin et al. [28] proposed a category-based subspace attention network and an outfit ranking loss to model the item interactions within an entire outfit. Lastly, Sarkar et al. [29] proposed OutfitTransformer, a framework based on Transformer [22], to learn an outfit-level representation.

## 3. Methodology

This section presents the problem formulation and thoroughly describes the proposed attention-based personalized compatibility embedding modeling approach.

### *3.1. Problem Formulation*

First, assume we have a set of users $U = \{u_1, u_2, \cdots, u_M\}$, a set of tops $T = \{t_1, t_2, \cdots, t_{N_t}\}$, and a set of bottoms $B = \{b_1, b_2, \cdots, b_{N_b}\}$. Each user $u_m$ is associated with a set which contains historical top-bottom pairs $O_m = \{(t_{i_1^m}, b_{j_1^m}), (t_{i_2^m}, b_{j_2^m}), \cdots, (t_{i_{N_m}^m}, b_{j_{N_m}^m})\}$, 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. Then, for each $t_i(b_j)$, we use $v_i^t(v_j^b) \in \mathbb{R}^{D_v}$ and $t_i^t(t_j^b) \in \mathbb{R}^{D_t}$ to represent its visual and textual features from different ConvNet modules. Next, we use $\widetilde{v_i^t}(\widetilde{v_j^b}) \in \mathbb{R}^{D_v}$ and $\widetilde{t_i^t}(\widetilde{t_j^b}) \in \mathbb{R}^{D_t}$ to indicate its visual and textual embeddings through different attention branches modules. $D_v$ and $D_t$ denote the dimensions of the corresponding embeddings.

In this study, we aim to develop fashion compatibility embeddings for outfit recommendations by considering user preferences and employing an attention mechanism. Consistent with previous research [11], we explore the challenge of determining "which bottom would be preferred by the user to match the given top?". Let $e_{ij}^m$ denote the preference of the user $u_m$ towards the bottom $b_j$ for the given top $t_i$, based on a generated personalized rating score list of bottoms $b_j$'s for a given top $t_i$ and hence solve the practical problem of personalized outfit matching.

To ensure accurate measurements of $e_{ij}^m$, we have designed a personalized compatibility embedding modeling network F that incorporates an attention mechanism. This network can integrate users' preferences for visual and textual aspects of items into the compatibility embedding model. The mathematical expression for this model is as follows:

$$e_{ij}^m = F(t_i, b_j, u_m | \theta_F) \tag{1}$$

where $\theta_F$ refers to the model parameters to be learned.

### *3.2. PCE-Net*

To effectively address the challenge of personalized clothing matching, it is essential to account for both item-item compatibility and user-item preference. Modeling fashion item compatibility is a fundamental problem in this context. A significant issue, therefore, is how to generate compatibility embeddings that are helpful in clothing matching. To this end, we explore user preferences towards a bottom that complements a given top by modeling compatibility embeddings between fashion items and the user's personal preferences. Formally, we have:

$$e_{ij}^m = \mu \cdot c_{ij} + (1 - \mu) \cdot p_{mj} \tag{2}$$

$$c_{ij} = C(t_i, b_j | \theta_c) \tag{3}$$

$$p_{mj} = P(u_m, b_j | \theta_P) \tag{4}$$

The attention-based compatibility embeddings modeling and attention-based personal preference modeling are denoted as $C$ and $P$, respectively, with $\theta_c$ and $\theta_P$ as their corresponding model parameters. The compatibility interaction between the top $t_i$ and bottom $b_j$ is represented by $c_{ij}$, while $p_{mj}$ denotes the personal preference of user $u_m$ towards the bottom $b_j$. To balance the relative importance of both components, a non-negative trade-off parameter $\mu$ is used.

#### 3.2.1. Attention-Based Compatibility Embedding Modeling

We propose a more effective way of measuring the compatibility between the top $t_i$ and bottom $b_j$. To accomplish this, we suggest that the model learns its compatibility embeddings in latent compatibility space. In this space, complementary top-bottom pairs should be closer than incompatible pairs. Additionally, we argue that there should be a gap between the interactive features of matching top-bottom pairs and mismatched top-bottom pairs, i.e., $c_{ij}$, $c_{ik}$, thus turning the task of predicting compatibility into a classification problem.

To learn the preliminary features of items in visual and text modalities, we employ convolutional neural networks (CNN) which have demonstrated excellent performance in learning representations [30–32]. It is imperative to note that all fashion items have visual and textual modalities. For example, the information on colors, patterns, and shapes of a fashion item can be extracted from its image, while its textual description can provide information on the brand, material, and category. These two modalities provide complementary information crucial for understanding the fashion items at the feature level. Therefore, we integrate both modalities' information to learn the compatibility embeddings between fashion items.

In Section 3.1, we introduced $v_i^t(v_j^b) \in \mathbb{R}^{D_v}$ and $t_i^t(t_j^b) \in \mathbb{R}^{D_t}$ to represent the global visual and textual features of the top $t_i$ and bottom $b_j$, respectively, from various ConvNet modules. Inspired by previous works [6,9,10], we incorporated two attention branches to capture features that aid in compatibility embeddings modeling. These branches allow us to obtain attentive visual and textual features, which we denote as $\widetilde{v_i^t}(\widetilde{v_j^b}) \in \mathbb{R}^{D_v}$ and $\widetilde{t_i^t}(\widetilde{t_j^b}) \in \mathbb{R}^{D_t}$, respectively. Here, $D_t$ represents the dimensionality of the latent compatibility space.

**Visual Attention.** To enable the compatibility embeddings module to automatically capture pertinent fine-grained visual characteristics such as color, pattern, and shape, we incorporate visual dot product attention to generate attention weights based on global visual features, $v_i^t(v_j^b) \in \mathbb{R}^{D_v}$. Specifically, we employ the visual attentive representation learning of the top portions as an illustration. The visual attention weight, $\omega_v^{t_i}$, can be calculated according to Equation (5) by applying the following formula:

$$e_{t_i} = V_a^T tanh(U_a(v_i^t)^T) \tag{5}$$

where $U_a \in \mathbb{R}^{D_v \times D_v}$ and $V_a \in \mathbb{R}^{D_v}$. The visual attention weight $\omega_v^{t_i} = e_{t_i}$. Then we calculate the attentive visual features $\widetilde{v_i^t} \in \mathbb{R}^{D_v}$ of top $t_i$,

$$\widetilde{v_i^t} = \omega_v^{t_i} v_i^t \tag{6}$$

Likewise, we can calculate the attentive visual features $\widetilde{v_j^b} \in \mathbb{R}^{D_v}$ of bottom $b_j$,

$$e_{b_j} = V_a^T tanh(U_a(v_j^b)^T) \tag{7}$$

$$\omega_v^{b_j} = e_{b_j} \tag{8}$$

$$\widetilde{v_j^b} = \omega_v^{b_j} v_j^b \tag{9}$$

Next, we utilize inner products to quantify the visual compatibility interaction between the attentive visual features of the top portion $t_i$ and the bottom portion $b_j$.

$$\widetilde{v_{ij}} = \widetilde{v_i^t} \ (\widetilde{v_j^b})^T \tag{10}$$

where the inner product encodes the visual interaction scores between fashion items.

**Text Attention.** We propose the integration of a text attention branch into the compatibility embedding network. This branch aims to capture the text features of each individual top and bottom, as well as the interactive text features of top-bottom pairs. By incorporating a text attention branch, our model is able to autonomously identify the crucial text features that contribute to compatibility interaction.

For a pair of textual features $t_j^t$ and $t_j^b$, the input feature to the text attention branch is calculated as follows,

$$y_{ij} = concat\{t_i^t, t_j^b\} \tag{11}$$

where $concat\{\ldots\}$ refers to the concatenation operation.

As depicted in Figure 2, the concatenated text features are passed through a sequence of fully-connected and ReLU layers. Subsequently, a softmax function is applied to the final activation values, producing a weight vector $\omega_t^{ij}$ of dimension $D_t$. This vector is crucial in determining the significance of the textual compatibility embedding interaction. The expression for this process is as follows:

$$\omega_t^{ij} = Softmax(FC(ReLU(FC(y_{ij})))) \tag{12}$$

Then, the attentive textual features of the top $t_i$ and bottom $b_j$ are as follows,

$$\begin{cases} \widetilde{t_i^t} = \omega_t^{ij} t_i^t \\ \widetilde{t_j^b} = \omega_t^{ij} t_j^b \end{cases} \tag{13}$$

Likewise, we also use inner products to measure the textual compatibility between attentive textual features of the top $t_i$ and bottom $b_j$,

$$\widetilde{t_{ij}} = \widetilde{t_i^t} \ (\widetilde{t_j^b})^T \tag{14}$$

where the inner product encodes the textual interaction scores between fashion items.

Finally, to comprehensively measure the compatibility embeddings utilizing the aforementioned attention branches, we define the following:

$$c_{ij} = \pi \widetilde{v_{ij}} + (1 - \pi) \widetilde{t_{ij}} \tag{15}$$

where $\pi$ is a non-negative trade-off parameter that determines the relative importance of the two modalities. $c_{ij}$ denotes the interaction of compatibility embeddings of the top $t_i$ and bottom $b_j$.

#### 3.2.2. Attention-Based Personal Preference Modeling

Drawing from matrix factorization techniques, we propose a model that captures users' personalized preference for a specific type of product, a bottom, which has proven effective in personalized recommendation tasks [33–38]. The underlying principle is decomposing the user-item interaction matrix into latent factors representing users and items. Additionally, building upon the work of Song et al. [11], we expand the matrix factorization

approach to incorporate latent factors that capture users' content-based preferences. This is crucial because users' preference for fashion items may stem from visual or textual features. For instance, users may prioritize visual characteristics like color and pattern, or textual features like brand and material. To comprehensively account for users' and fashion items' latent factors, as well as their content-based factors, considering both aspects is imperative.

In a similar vein, we employ the inner product to encode the latent scores for user-item interactions and the content-based scores for user-item interactions. To illustrate, let us consider the personal preference of user $u_m$ for the bottom item $b_j$. The expression is as follows:

$$p_{mj} = \alpha + \beta_m + \beta_j + \gamma_m^T \gamma_j + (\xi^v_{m\to j})^T \widetilde{v}^b_j + (\xi^t_{m\to j})^T \widetilde{t}^b_j \tag{16}$$

where $\alpha$ represents the global offset to be learned, $\beta_m$ and $\beta_j$ are the bias terms corresponding to the user $u_m$ and the bottom $b_j$, respectively. $\gamma_m$ and $\gamma_j$ are the separate latent factors of user $u_m$ and bottom $b_j$. Their inner product captures the latent preference of user $u_m$ towards bottom $b_j$. $\xi^v_{m\to j}$ and $\xi^t_{m\to j}$ denote the latent visual and textual factors of user $u_m$ for the bottom $b_j$, respectively. $\widetilde{v}^b_j$ and $\widetilde{t}^b_j$ correspond to the attentive visual and textual features of the bottom $b_j$, which were computed in Section 3.2.1. To summarize, the first four terms in Equaton (16) encode the global latent preference, while the last two terms encode the content-based latent preference of user $u_m$ towards bottom $b_j$.

#### 3.2.3. Objective Function

Based on the BPR framework [18], we utilize a model that captures the implicit interaction between users and fashion items. This model has been shown to effectively represent implicit preferences in various studies [11,12,16,17,39,40]. We construct a training set for training the BPR algorithm, ensuring its optimal performance.

$$D := \{(m, i, j, k) \mid u_m \in U \wedge (t_i, b_j) \in O_m \wedge b_k \in B \setminus b_j\} \tag{17}$$

where the quadruplet $(m, i, j, k)$ denotes that the user $u_m$ prefers the bottom $b_j$ over $b_k$ for the given top $t_i$. As for the compatibility embeddings and personal preference, the objective function is defined as follows,

$$\begin{aligned} L_{bpr} &= \sum_{(m,i,j,k)\in D} [-ln(\sigma(e^m_{ij} - e^m_{ik}))] + \frac{\lambda}{2}\|\Theta_F\|^2_F \\ &= \sum_{(m,i,j,k)\in D} [-ln(\sigma((\mu c_{ij} + (1-\mu)p_{mj}) \\ &\quad -(\mu c_{ik} + (1-\mu)p_{mk}))] + \frac{\lambda}{2}\|\Theta_F\|^2_F \end{aligned} \tag{18}$$

where $c_{ik}$ indicates the compatibility interaction between the top $t_i$ and bottom $b_k$, and $p_{mk}$ denotes the personal preference of the user $u_m$ towards the bottom $b_k$, whose specific calculation is similar to Sections 3.2.1 and 3.2.2. $\lambda$ is the non-negative hyperparameter, $\Theta_F$ refers to the set of parameters of the model, including $\omega_v^{t_i}$, $\omega_v^{b_j/b_k}$, $\omega_t^{ij/ik}$, $\alpha$, $\beta_m$, $\beta_{j/k}$, $\gamma_m$, $\gamma_{j/k}$, $\xi^v_{m\to j/k}$, $\xi^t_{m\to j/k}$. $\sigma$ is the sigmoid function $\sigma(x) = 1/(1+e^{-x})$.

## 4. Experiment

To evaluate the proposed method, we conduct comprehensive experiments on the large-scale real-world dataset IQON3000 [11] extracted from the social commerce website IQON. These experiments were conducted to showcase the effectiveness of our approach.

### *4.1. Dataset*

Our experiments were conducted on the IQON3000 real-world dataset [11], comprising 216,791 outfits created by 3568 users using 650,373 fashion items. The outfit splits

provided by the authors were utilized, including 170,601 quadruplets in the training set, 23,095 quadruplets in the validation set, and 23,095 quadruplets in the test set. Each fashion item, encompassing all tops and bottoms, is associated with a visual image and a textual description. We merged all quadruplets from the training set and incorporated signals from both modalities to train the PCE-Net model.

### 4.2. Implementation

**Visual Representation.** To understand the visual attributes of fashion items, a convolutional neural network (CNN) is employed as the feature extractor. Deep CNNs have demonstrated outstanding performance in image representation learning [41–43]. Specifically, the ResNet-50 [44] is selected as the visual representation learning module. For each fashion item image, the final global average pooling layer's output is considered the preliminary visual characteristics. These outputs are 2048-D vectors that serve as the main visual features in the vision modality. By combining these features with the visual attention branch discussed in Section 3.2.1, we obtain the ultimate visual attributes of each fashion item.

**Textual Representation.** However, we encounter a limitation that must be acknowledged here. Due to the closure of the fashion website IQON, we cannot source text descriptions directly from the provided data URLs by the authors of GP-BPR [11]. Thus, we rely on the text features previously extracted by the authors. The following section will briefly overview their approach to extracting text features. The authors utilized the category metadata and title descriptions as textual information for the fashion items. These textual inputs were tokenized using the Japanese morphological analyzer Kuromoji. Each word in the text description is then represented as a 300-D vector using Nwjc2vec [45], a Japanese Word2vec method. Subsequently, the feature matrix for the overall textual description is constructed, with each word's feature vector occupying a distinct row. This textual feature matrix is input into a single-channel CNN, comprising a convolutional layer, a max pooling layer, and an activation layer. Ultimately, the output vector of each fashion item, a 400-D representation, is obtained as the preliminary textual features. Therefore, we adopt this textual representation as the initial textual modality features of the fashion items. Subsequently, the top and bottom features are fed into the text attention branch to compute the definitive textual features for each fashion item.

**Detail Settings.** The trade-off parameters $\pi$ and $\mu$ are explored in the interval [0.0, 1.0], with $\pi = 0.5$ and $\mu = 0.1$ identified as the optimal values. During the training process, the model parameters are randomly initialized in Equation (16) using the Normal Distribution. Furthermore, the weights of the visual attention branch and textual attention branch in Equations (6), (9), and (13) are respectively initialized using the Xavier method [46] and Uniform distribution. For optimization, we utilize the Adam algorithm [47] with a learning rate set as 0.001. The learning rate is investigated in the range [0.0005, 0.001, 0.005, 0.01]. To expedite the training and promote faster convergence, a mini-batch size of 64 is employed. The proposed approach is fine-tuned for 100 epochs, and the model's performance is evaluated on the test set. Finally, the area under the ROC curve (AUC) [48] is used as a metric to assess the effectiveness of the attention-based personalized compatibility embedding network.

### 4.3. Results and Discussion

We consider the following baselines in the top-bottom pairs recommendation experiments to evaluate the proposed model.

- **POP-T:** POP is frequently used as a baseline in recommender systems [49]. POP-T simply selects the most popular bottoms for each top and vice versa. Here, "popularity" is defined as the number of tops paired with the bottom, i.e., the number of top-bottom pairs in the training set.
- **POP-U:** For this baseline [49], the number of users that used this bottom as a component of an outfit in the training set is used to determine the "popularity" of the bottom.

- **RAND:** The compatibility ratings between positive and negative pairs were randomly assigned.
- **Bi-LSTM:** The bidirectional LSTM method in [8] modeled an outfit as an ordered sequence. Its operating principle is to predict the next item conditioned on previous ones. We keep the variables constant by setting the sequence length to 2, i.e., there are only a top and a bottom.
- **BPR-DAE:** The baseline [16] uses a dual autoencoder network (DAE) to learn the potential compatibility space by jointly modeling the consistent relationship between visual and textual patterns and the implicit preference between items by Bayesian Personalized Ranking.
- **BPR-MF:** This model [18] is one of the most commonly used techniques for personalized recommendation tasks, which captures the latent user-item interaction by the Matrix Factorization (MF) method in the pairwise ranking tasks.
- **VBPR:** Unlike MF, the model [2] also considers the user's preference for visual factors. The model represents the visual characteristics of an outfit by averaging the visual characteristics of the items in the set.
- **TBPR:** The difference between TBPR [2] and VBPR is using information from different modalities.
- **VTBPR:** This model [2] uses a combination of both visual and text modalities to model user preferences.
- **GP-BPR:** This baseline [11] combines visual and textual features of clothing with personal preferences to jointly model general (item-item) compatibility and personal (user-item) preferences, where matrix factorization for the user-item interaction matrix is performed to obtain the potential user preferences.
- **PAI-BPR:** This model [12] is an attribute-based interpretable personal preference modeling scheme, where personalization is achieved by taking inspiration from GP-BPR [11] and adding attribute-wise interpretable results. Since the code is not publicly available, we directly report the experimental results of Table III in the original paper [12] for quantitative comparison.

The performance comparison of various techniques is shown in Table 1. These quantitative data allow us to draw the following conclusions:

- BPR-DAE outperforms Bi-LSTM, demonstrating that the content-based model, which captures the compatibility relationship between items by directly extracting features from multimodal data, is superior to the sequential model (predicting the following item from the previous one).
- VTBPR performs better than VBPR, TBPR, and BPR-MF, indicating the value of multimodal data in enhancing model performance.
- To solve the problem of personalized clothing matching, GP-BPR and PAI-BPR combine generalized item-item compatibility and user-item preferences using multi-modal characteristics. Since PAI-BPR uses an attribute classification network to address the interpretability of the model, performance has been slightly improved.
- PCE-NET obtains the best performance compared to the above baseline, but there is no modeling attribute classification module because PCE-NET does not focus on interpretability problems. Our model can automatically capture the compatibility features of multi-modal data using two attention branches separately, which indicates that further development and exploitation of multi-modal data is necessary for embedding learning tasks.

**Table 1.** Performance comparison among different approaches in terms of AUC.

| | Approach | AUC |
|---|---|---|
| Baselines | POP-T | 0.6042 |
| | POP-U | 0.5951 |
| | RAND | 0.5014 |
| | Bi-LSTM | 0.6675 |
| | BPR-DAE | 0.7004 |
| | BPR-MF | 0.7913 |
| | VBPR | 0.8129 |
| | TBPR | 0.8146 |
| | VTBPR | 0.8213 |
| | GP-BPR | 0.8314 |
| | PAI-BPR | 0.8502 |
| Proposed | PCE-Net | 0.8534 |

**Qualitative Results.** To visually demonstrate the superiority of our model compared with GP-BPR [11] in personalized clothing matching and to analyze the impact of each component in our model, we present multiple sets of test samples along with the model's evaluation of positives and negatives in Figure 3. As previously mentioned, the testing quadruplet $(m, i, j, k)$ indicates the user $u_m$ prefers the bottom $b_j$ than $b_k$ for the given top $t_i$. In the first example involving user1, our model determines that the bottom $b_k$ is more compatible with $t_i$ compared to $b_j$. Further analysis reveals that our model can capture the historical preference of users for brown, skirts, which stems from the contribution of the visual attention branch. In contrast, GP-BPR [11] captures users' preference for short skirts. Even though our model predicts incorrectly, we can see that there is a visually compatible relationship between $t_i$ and $b_k$, which indicates that our model is capable of generating some convincing matching suggestions. In the second example, our model learns better than GP-BPR for visual features of fashion items.

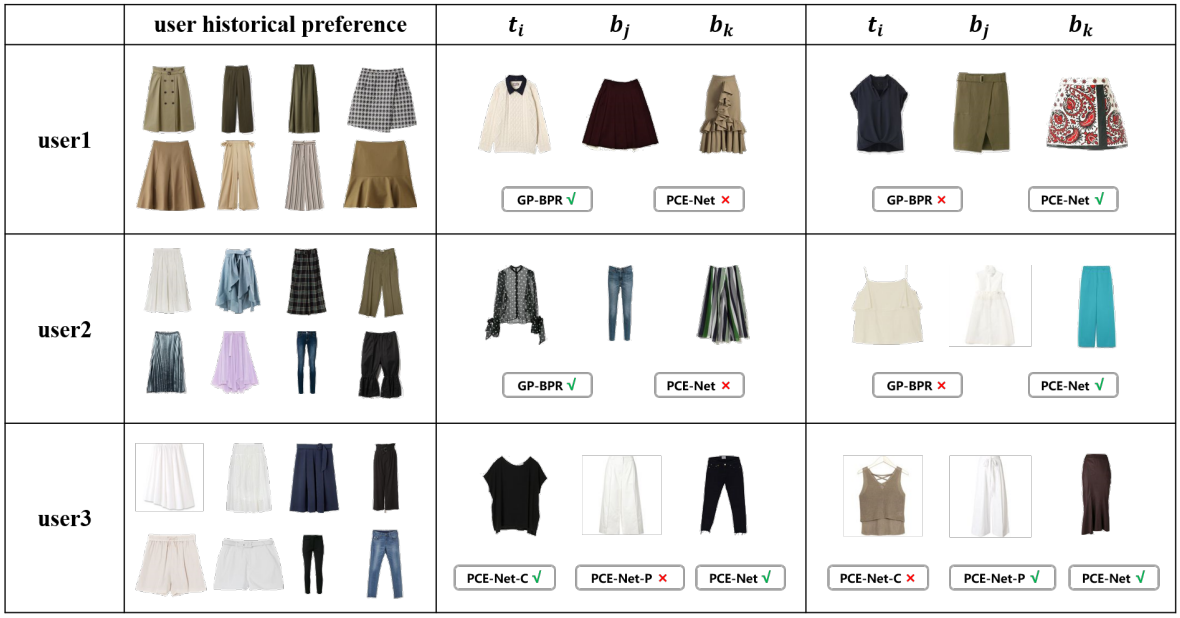


**Figure 3.** Illustration of the method comparison and the influence of the attention-based compatibility embedding and personal preference model-ing. All the quadruplets satisfy the ground truth that $\{u_m, t_i\} : b_i > b_k$. PCENET-C and PCENET-P denotes the above two components, respectively. "✓" and "×" separately indicates the correct and wrong judgements of the model.

In the first example involving user 3, the bottom $b_j$ and $b_k$ both have similar styles of clothing in the user's historical preferences, which may lead PCE-Net-P to predict

incorrectly. However, by considering the fine-grained compatibility relationships between fashion items from PCE-Net-C, black and white pairing is more common in the matching rules, and white clothing is more versatile in the matching results, i.e., the matching degree is well, so PC-E-Net finally obtains the correct evaluation result. In the second example, the top $t_i$ and the bottom $b_k$ are indistinguishable in terms of visual features such as color, which leads to an incorrect prediction of PCE-Net-C. At this point, component P (i.e., Personal Preference Modeling) enables PCE-Net to obtain the correct prediction result by capturing the user's historical preferences.

Above, we have demonstrated that our model outperforms other baseline models to a certain extent. Additionally, both components of our model are indispensable and serve as complementary sources of information to enhance its overall performance.

To assess the learning ability of the proposed PCE-Net, we visually analyze the compatibility relation for positive and negative pairs, as well as the resulting compatibility embedding space.

**Compatible Relations Visualization.** Figure 4 illustrates the application of t-SNE [50] to represent the learned compatible relation space. Each dot in the plot represents the multi-modal fusion feature of a top-bottom pair. The red dots represent compatible top-bottom pairs, while the blue dots represent incompatible pairs. The separation of the two relations learned by PCE-Net indicates that our model effectively distinguishes whether a compatible relationship exists between the top-bottom pairs and produces convincing matching results. We observe an intriguing occurrence of crossover points between the red and blue regions. This can be attributed to two factors: (1) The attention mechanism renders compatibility interaction modeling more intricate and implicit than general compatibility modeling. (2) As mentioned earlier, textual attention relies on a top-bottom feature connection, which means that the final features of an item may contain some characteristics of the other item, resulting in erroneous relationship predictions.

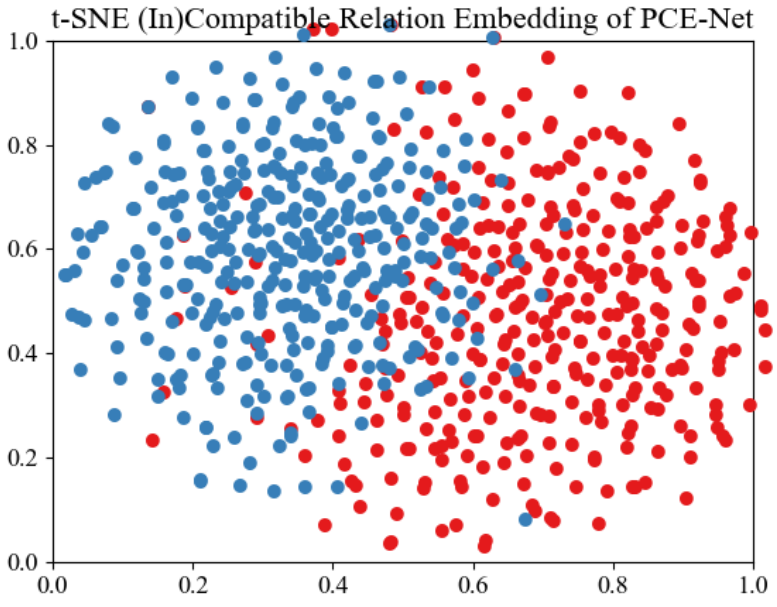


**Figure 4.** Visualization for the compatible relations between the top-positive bottom pairs (red), and the incompatible relations between the top-negative bottom pairs (blue) by using t-SNE. Best viewed in color.

**Compatibility Embeddings Visualization.** In this part, we utilize t-SNE [4] to visualize the distribution of some and all test triplets $(i, j, k)$ from IQON3000 [11] in 2-dimensional spaces. The results are shown in Figures 5 and 6. Figure 5 represents the compatibility embedding space with ten triplets. A dotted line indicates the distance between two triplets on the left. In the right part, the item enclosed in an orange box represents the given tops,

while the green and red represent compatible bottoms and incompatible bottoms, respectively. The length of the dotted line corresponds to the distance between the top and bottom. In the latent compatibility space, compatible items should be closer. Both example triplets satisfy this criterion, illustrating that our model effectively learns compatible embeddings.

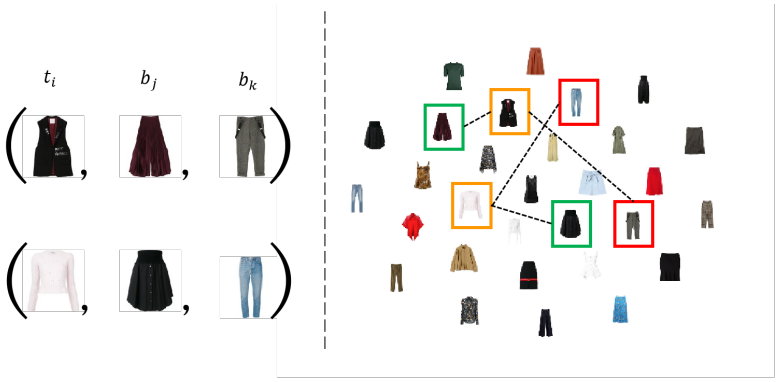


**Figure 5.** Visualization for Compatibility Embedding of some samples from PCE-Net by using t-SNE. we give two test triplet examples in the left part, and give their embedding distribution in the right part. In the compatible space, the top $t_i$ and bottom $b_i$ are compatible, and hence their distance is closer than the incompatible items $b_k$.

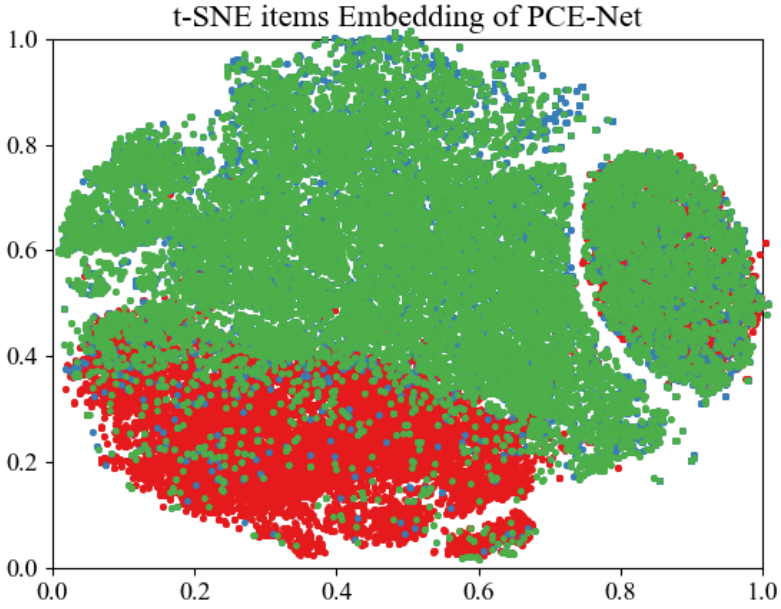


**Figure 6.** Visualization for Compatibility Embedding of all test samples from PCE-Net by using t-SNE. The blue part represents all the tops' embeddings, and the red and green part denote the posi-tive bottoms' and negative bottoms' embeddings, respectively. Best viewed in color.

Figure 6 illustrates the visualization of all feature distributions obtained by combining the multi-modal features of each test triplet $(i, j, k)$. The blue dots represent the tops, while the red and green dots denote the positive bottoms and negative bottoms, respectively. Notably, the blue region is distinct from the areas occupied by the red and green dots. Consequently, the overlapping feature distribution between the red and green dots appears reasonable since they both pertain to the same category. This similarity results from our model's ability to employ two feature extractors and two attention branches to learn the embedding of fashion items.

### *4.4. Ablation Study*

**Different Modalities.** We further assess the contribution of different input modalities in our model, specifically, the two variants of PCE-Net: PCE-Net-V and PCE-Net-T. PCE-Net-V utilizes only the visual modality to extract compatibility features, while PCE-Net-T focuses on the textual modality. Table 2 presents the performance of these modalities when used as inputs for PCE-Net. To provide a more precise comparison, we also offer the experimental results of the optimal baselines (GP-BPR [11] and PAI-BPR [12]) using only a single modality. Based on the findings in Table 2, we make the following observations: (1) PCE-Net-V and PCE-Net-T outperform GP-BPR-V and PAI-BPR-V, and GP-BPR-T and PAI-BPR-T, respectively. This validates the effectiveness of the attention branches we introduced to enhance model performance for different modalities. (2) PCE-Net performs better than both PCE-Net-V and PCE-Net-T, indicating that utilizing both modalities as complementary information improves the learning of compatibility embedding and enhances personalized preference modeling. (3) Interestingly, we note that model-T outperforms model-V in the GP-BPR [11] and PAI-BPR [12] baselines. The authors argue that critical features like pattern, style, and brand can be better summarized in the textual information. For instance, fashion items are more likely to be compatible if they share the same brand. However, our model PCE-Net-V attains equivalent performance to PCE-Net-T and even slightly outperforms the latter, suggesting that the visual attention branch effectively captures compatibility features automatically and enhances personalized preference modeling.

**Table 2.** Performance comparison among different modalities in terms of AUC.

| Modality | Approach | AUC |
|---|---|---|
| Visual modality | GP-BPR-V | 0.8239 |
| | PAI-BPR-V | 0.8413 |
| | PCE-Net-V | 0.8485 |
| Textual modality | GP-BPR-T | 0.8313 |
| | PAI-BPR-T | 0.8432 |
| | PCE-Net-T | 0.8475 |
| Muti-modal | GP-BPR | 0.8314 |
| | PAI-BPR | 0.8502 |
| | PCE-Net | 0.8534 |

In Equation (13), the non-negative parameter $\pi$ denotes the weight assigned to the visual modality. Based on the aforementioned conclusions, it is imperative to incorporate multiple modalities concurrently into the model. Thus, Figure 7 presents a line graph depicting the model's performance at various values of $\pi$. As the figure reveals, our model achieves optimal performance when $\pi = 0.5$, indicating that both modalities hold equal importance.

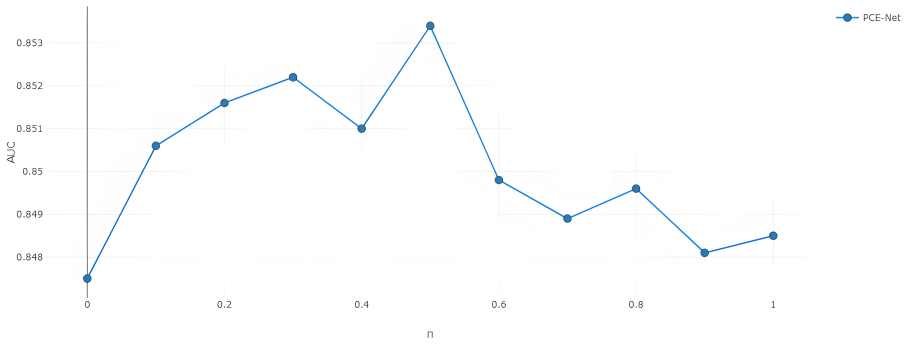


**Figure 7.** Performance of PCE-Net with respect to the trade-off parameter $\pi$, $\pi = 0.5$ is the best.

**Attention Branch.** The Attention Branch plays a crucial role in our study. To evaluate its impact on the model across different modalities, we present quantitative data in Table 3. The findings highlight that employing two separate attention branches to encode the preprocessing features of visual and textual modalities effectively captures compatible interaction features. Furthermore, these features serve as complementary data, ultimately enhancing the model's overall performance.

**Table 3.** Performance comparison of the two attention branches in terms of AUC.

| | Approach | AUC |
|---|---|---|
| Attention branch | Visual Attention | 0.8432 |
| | Text Attention | 0.8508 |
| | (V + T) Attention | 0.8534 |

**Different Component.** To evaluate the individual contributions of each component in our model, we present the results in Table 4, specifically focusing on two components: compatibility embedding modeling and personal preference modeling within PCE-Net. Our observations are as follows: (1) Our comprehensive model surpasses the performance of the two derived models containing only one component. This substantiates the vital role each component plays in our model. (2) PCE-Net-P outperforms PCE-Net-C, signifying that users' historical preferences effectively capture their personalized preferences, thereby influencing the outcome of the personalized clothing matching task.

**Table 4.** Performance comparison among each component in terms of AUC.

| | Approach | AUC |
|---|---|---|
| Different component | PCE-Net-P | 0.8376 |
| | PCE-Net-C | 0.6982 |
| ours | PCE-Net | 0.8534 |

Additionally, our model's performance in Equation (2) is evaluated by showcasing its performance as a line graph in Figure 8, within the range [0.0, 1.0]. For the sake of clarity in comparing with the baseline GP-BPR [11], we also include the experimental results of both models in the same line graph, highlighting the varying parameter $\mu$. However, we cannot compare our model with PAI-BPR [12] due to the original paper's absence of publicly available code and relevant experimental results. The figure demonstrates that, for most parameter values, our model outperforms GP-BPR, thereby affirming the validity of our approach. It is worth noting that PCE-Net exhibits lower performance than GP-BPR when $\mu = 1.0$, likely because our text attention interaction branch constructs interactions between the top-bottom pairs, resulting in the fusion of their respective features. Consequently, without the personalized preference component's guidance, our model's performance in item recommendation is compromised.

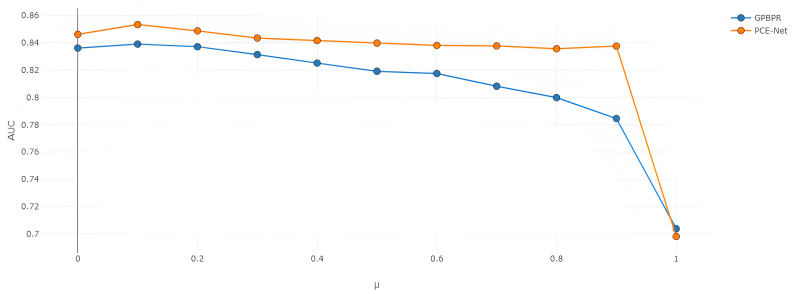


**Figure 8.** Performance of PCE-Net with respect to the trade-off parameter $\mu$. Best viewed in color.

### *4.5. User Study*

An additional user study is conducted to further validate our model's effectiveness. A total of 100 participants are selected for this study. They are then presented with seven questions, depicted in Figure 9 (only six are displayed). As an illustration, for the first question, participants are provided with five pairs representing their historical style preferences from top to bottom. Following that, they are shown a triplet in the test set, consisting of a top, a positive bottom, and a negative bottom. The top serve as the condition, while the positive and negative bottoms are given as two options. It should be noted that the order in which the options are presented is unrelated to the positive or negative bottom. Subsequently, all participants are asked to choose a compatible bottom aligned with their historical preference for the given top. More than half of the participants choose the "positive bottom" option for each question, except the last one.


**Figure 9.** Some examples of the questions presented to the participants for the user study.

In Q1, the user exhibits a preference for black bottoms and primarily favors black-black matching rules. Notably, 86% of the participants chose the positive bottom A, indicating that our model successfully simulates user preferences. It is worth mentioning that bottom A and bottom B in Q3 possess similar visual styles and align with the user's historical preferences. This fact influenced 43% of the participants to choose the negative bottom B. Upon consultation, it became apparent that they largely overlooked the compatibility features of bottom A and the top (please refer to the enlarged pink patterns for color reference). In contrast, our model adequately considered these compatibility features, resulting in accurate prediction outcomes. This further illustrates the capability of our model to learn fine-grained compatibility embeddings and enhance the performance of downstream tasks. As for Q6, due to an incorrect prediction by our model, this question was excluded from the user study to determine whether the model could generate convincing recommendation suggestions. Surprisingly, 70% of the participants also selected option B, the negative bottom. This implies that our model's misguided choice could be attributed to its assertion that the user prefers black bottoms to complement the given top. Despite the incorrect prediction in this particular case, the fact that 70% (>50%) of the participants selected the same option as suggested by our model reinforces the notion that our model can provide persuasive bottom recommendations that exhibit compatibility with the given top.

## 5. Conclusions and Future Work

Our research addresses the task of modeling compatibility embeddings in the context of fashion. To achieve this, we propose an attention-based personalized compatibility embedding network called PCE-Net, which consists of two components: attention-based compatibility embedding modeling and attention-based personalized preference modeling. By incorporating multiple attention branches for visual and textual modalities of fashion

items, our model automatically captures features relevant to compatibility embedding, thereby benefiting downstream tasks such as top-bottom matching. To evaluate the effectiveness of our model, we conducted various experiments on the IQON3000 dataset. These experiments encompassed quantitative and qualitative comparisons, ablation studies for each modality/component/attention branch, t-SNE visualization, and user studies. The results of these experiments corroborated the model's ability to learn compatibility embeddings and generate convincing matching results.

However, it is essential to acknowledge a limitation of our work: we only model users' personalized preferences using potential preference factors for bottoms and content-based preference factors. In the future, we intend to address this limitation by incorporating a component that can search for visually or textually similar tops in the candidate pool, enabling the identification of compatible bottoms for personalized recommendations and ultimately enhancing the performance of the recommendation task for fashion collections.

**Author Contributions:** Conceptualization, X.N., Z.X. and Y.T.; methodology, X.N.; software, X.N. and Y.T.; validation, X.N. and J.Z.; formal analysis, X.N. and J.Z.; investigation, Y.T.; resources, Z.X.; data curation, J.Z.; writing—original draft preparation, X.N.; writing—review and editing, Y.T. and Z.X.; visualization, X.N. and Y.T.; supervision, Z.X. and J.Z.; project administration, Z.X.; funding acquisition, Z.X. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Beijing Natural Science Foundation [No. 8202013], and the Beijing University of Civil Engineering and Architecture, Graduate Student Innovation Projects [No. PG2023147].

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors are grateful to other colleagues who work or practice in Ping An for their help.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Veit, A.; Belongie, S.; Karaletsos, T. Conditional similarity networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 830–838.
2. He, R.; McAuley, J. VBPR: Visual bayesian personalized ranking from implicit feedback. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.
3. He, T.; Hu, Y. FashionNet: Personalized outfit recommendation with deep neural network. *arXiv* **2018**, arXiv:1810.02443.
4. Shih, Y.S.; Chang, K.Y.; Lin, H.T.; Sun, M. Compatibility family learning for item recommendation and generation. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
5. Vasileva, M.I.; Plummer, B.A.; Dusad, K.; Rajpal, S.; Kumar, R.; Forsyth, D. Learning type-aware embeddings for fashion compatibility. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 390–405.
6. Tan, R.; Vasileva, M.I.; Saenko, K.; Plummer, B.A. Learning similarity conditions without explicit supervision. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10373–10382.
7. Singhal, A.; Chopra, A.; Ayush, K.; Govind, U.P.; Krishnamurthy, B. Towards a unified framework for visual compatibility prediction. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 3607–3616.
8. Han, X.; Wu, Z.; Jiang, Y.G.; Davis, L.S. Learning fashion compatibility with bidirectional lstms. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1078–1086.
9. Laenen, K.; Moens, M.F. Attention-based fusion for outfit recommendation. In *Fashion Recommender Systems*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 69–86.
10. Lin, Y.; Ren, P.; Chen, Z.; Ren, Z.; Ma, J.; De Rijke, M. Explainable outfit recommendation with joint outfit matching and comment generation. *IEEE Trans. Knowl. Data Eng.* **2019**, *32*, 1502–1516. [CrossRef]
11. Song, X.; Han, X.; Li, Y.; Chen, J.; Xu, X.S.; Nie, L. GP-BPR: Personalized compatibility modeling for clothing matching. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 320–328.

12. Sagar, D.; Garg, J.; Kansal, P.; Bhalla, S.; Shah, R.R.; Yu, Y. Pai-bpr: Personalized outfit recommendation scheme with attribute-wise interpretability. In Proceedings of the 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), New Delhi, India, 24–26 September 2020; pp. 221–230.
13. Li, X.; Wang, X.; He, X.; Chen, L.; Xiao, J.; Chua, T.S. Hierarchical fashion graph network for personalized outfit recommendation. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Xi'an, China, 25–30 July 2020; pp. 159–168.
14. Chen, W.; Huang, P.; Xu, J.; Guo, X.; Guo, C.; Sun, F.; Li, C.; Pfadler, A.; Zhao, H.; Zhao, B. POG: Personalized outfit generation for fashion recommendation at Alibaba iFashion. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2662–2670.
15. Dong, X.; Song, X.; Feng, F.; Jing, P.; Xu, X.S.; Nie, L. Personalized capsule wardrobe creation with garment and user modeling. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 302–310.
16. Song, X.; Feng, F.; Liu, J.; Li, Z.; Nie, L.; Ma, J. Neurostylist: Neural compatibility modeling for clothing matching. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 753–761.
17. Cui, Z.; Li, Z.; Wu, S.; Zhang, X.Y.; Wang, L. Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 307–317.
18. Rendle, S.; Freudenthaler, C.; Gantner, Z.; Schmidt-Thieme, L. BPR: Bayesian personalized ranking from implicit feedback. *arXiv* **2012**, arXiv:1205.2618.
19. Rendle, S.; Schmidt-Thieme, L. Pairwise interaction tensor factorization for personalized tag recommendation. In Proceedings of the Third ACM International Conference on Web Search and Data Mining, New York, NY, USA, 4–6 February 2010; pp. 81–90. [CrossRef]
20. McAuley, J.; Targett, C.; Shi, Q.; Van Den Hengel, A. Image-based recommendations on styles and substitutes. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 9–13 August 2015; pp. 43–52.
21. Veit, A.; Kovacs, B.; Bell, S.; McAuley, J.; Bala, K.; Belongie, S. Learning visual clothing style with heterogeneous dyadic co-occurrences. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 4642–4650.
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Springer: Long Beach, CA, USA, 2017; Volume 30.
23. Cucurull, G.; Taslakian, P.; Vazquez, D. Context-aware visual compatibility prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12617–12626.
24. Song, X.; Feng, F.; Han, X.; Yang, X.; Liu, W.; Nie, L. Neural compatibility modeling with attentive knowledge distillation. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 5–14.
25. Yang, X.; Ma, Y.; Liao, L.; Wang, M.; Chua, T.S. Transnfcm: Translation-based neural fashion compatibility modeling. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 403–410.
26. Lu, Z.; Hu, Y.; Jiang, Y.; Chen, Y.; Zeng, B. Learning binary code for personalized fashion recommendation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10562–10570.
27. Taraviya, M.; Beniwal, A.; Lin, Y.L.; Davis, L. Personalized compatibility metric learning. In Proceedings of the KDD 2021 International Workshop on Industrial Recommendation Systems, Singapore, 14–18 August 2021.
28. Lin, Y.L.; Tran, S.; Davis, L.S. Fashion outfit complementary item retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 3311–3319.
29. Sarkar, R.; Bodla, N.; Vasileva, M.I.; Lin, Y.L.; Beniwal, A.; Lu, A.; Medioni, G. Outfittransformer: Learning outfit representations for fashion recommendation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–6 January 2023; pp. 3601–3609.
30. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
31. Wang, Z.; Bai, X.; Ye, M.; Satoh, S. Incremental deep hidden attribute learning. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 72–80.
32. Liu, M.; Wang, X.; Nie, L.; Tian, Q.; Chen, B.; Chua, T.S. Cross-modal moment localization in videos. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 843–851.
33. Lee, D.; Seung, H.S. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2000; Volume 13.
34. Mnih, A.; Salakhutdinov, R.R. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2007; Volume 20.
35. Koren, Y.; Bell, R.; Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **2009**, *42*, 30–37. [CrossRef]
36. Koren, Y.; Bell, R. Advances in Collaborative Filtering. In *Recommender Systems Handbook*; Ricci, F., Rokach, L., Shapira, B., Eds.; Springer: New York, NY, USA, 2015; pp. 77–118. [CrossRef]

37. Kim, D.; Park, C.; Oh, J.; Lee, S.; Yu, H. Convolutional matrix factorization for document context-aware recommendation. In Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, 15–19 September 2016; pp. 233–240.
38. Packer, C.; McAuley, J.; Ramisa, A. Visually-aware personalized recommendation using interpretable image representations. *arXiv* **2018**, arXiv:1806.09820.
39. Loni, B.; Pagano, R.; Larson, M.; Hanjalic, A. Bayesian personalized ranking with multi-channel user feedback. In Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, 15–19 September 2016; pp. 361–364.
40. Cao, D.; Nie, L.; He, X.; Wei, X.; Zhu, S.; Chua, T.S. Embedding factorization models for jointly recommending items and user generated lists. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Japan, 7–11 August 2017; pp. 585–594.
41. Khosla, A.; Das Sarma, A.; Hamid, R. What makes an image popular? In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Republic of Korea, 7–11 April 2014; pp. 867–876.
42. Kim, Y. Convolutional neural networks for sentence classification. *arXiv* **2014**, arXiv:1408.5882.
43. Severyn, A.; Moschitti, A. Twitter sentiment analysis with deep convolutional neural networks. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 9–13 August 2015; pp. 959–962.
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
45. Asahara, M. *NWJC2Vec: Word Embedding Dataset from 'NINJAL Web Japanese Corpus'*; John Benjamins Publishing Company: Amsterdam, The Netherlands; Philadelphia, PA, USA, 2018; Volume 24, pp. 7–22.
46. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, Sardinia, Italy, 13–15 May 2010; pp. 249–256.
47. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
48. Zhang, H.; Zha, Z.J.; Yang, Y.; Yan, S.; Gao, Y.; Chua, T.S. Attribute-augmented semantic hierarchy: Towards bridging semantic gap and intention gap in image retrieval. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; pp. 33–42.
49. He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; Chua, T.S. Neural collaborative filtering. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 173–182.
50. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.