Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.078547

ARTICLE

Personalized Fashion Recommendation Fusing Multi-Behavior and Multi-Modal Features

Xin Lu[1], **Jian-Hong Wang**[1,*] **and Kuo-Chun Hsu**[2,*]

[1]School of Computer Science and Technology, Shandong University of Technology, Zibo, China
[2]Department of Information Management, National Taipei University of Business, Taipei, Taiwan

*Corresponding Authors: Jian-Hong Wang. Email: jhwang@sdut.edu.cn; Kuo-Chun Hsu. Email: totoro.hsu@ntub.edu.tw

Received: 03 January 2026; Accepted: 09 March 2026; Published: 08 May 2026

ABSTRACT: Aiming at the problems of data sparsity, uneven behavior weight allocation, and insufficient timeliness modeling existing in traditional recommendation systems in the scenario of personalized fashion recommendation, this paper proposes a personalized recommendation method that integrates multi-behavior weights and multi-modal features. A dynamic weighted collaborative filtering algorithm is designed, which comprehensively considers the multi-dimensional behaviors of users, and introduces a time attenuation factor to construct a time-sensitive user-item scoring matrix, so as to more accurately depict the dynamic changes of user interests. A multi-modal deep fusion framework is built: ResNet-50 is used to extract commodity image features, and the pre-trained BERT model is combined to extract text features; meanwhile, the multi-head self-attention mechanism is adopted to realize semantic-level interaction and adaptive fusion of cross-modal features, thereby enhancing the expressive ability of commodity representation. Then, user preference score prediction is carried out based on the deep predictive network to generate a personalized recommendation list. Experimental results on real e-commerce datasets show that the method in this paper achieves 0.703 and 0.491 on HR@5 and NDCG@5, respectively, which is significantly superior to other baseline models. Ablation experiments further verify the effectiveness of each module including time attenuation, multi-behavior weights and multi-modal features. This study provides a more accurate, dynamic and transparently interpretable personalized recommendation solution for e-commerce platforms, and has certain theoretical value and practical significance.

KEYWORDS: Fashion recommendation; collaborative filtering; multi-modal fusion; personalized recommendation; deep learning

1 Introduction

In recent years, with the deep integration of Internet technologies and e-commerce, fashion e-commerce platforms have become important channels for consumers to obtain fashion products. The problem of information overload arising from the combination of massive commodities and diversified consumer demands has posed severe challenges to traditional recommendation systems [1]. As an effective means to alleviate this problem, recommendation systems analyze users' historical behaviors and preferences to provide personalized product recommendations, and have become a key technology for improving user experience and platform sales performance.

In the field of personalized fashion recommendation, early research mainly focused on two major directions: collaborative filtering and content-based recommendation. Traditional collaborative filtering methods usually rely on a single type of user behavior for modeling [2]. Although they achieve certain

effectiveness, they ignore multi-dimensional implicit behavior signals such as browsing, adding to cart, and favoriting, leading to an incomplete representation of user preferences. Moreover, these methods often fail to distinguish the importance of different behaviors and do not consider the attenuation effect of behaviors over time, which makes the recommendation results susceptible to historical noise interference and unable to reflect the current changes in user interests. On the other hand, although content-based recommendation methods can leverage product attributes such as images and texts to alleviate the cold-start problem [3], they still have limitations in feature fusion and cross-modal semantic alignment. In particular, the interaction mechanism between visual and textual features is not in-depth enough, which affects the interpretability and accuracy of recommendations.

In recent years, the integration of multi-modal fusion technology and deep learning methods has brought new possibilities to fashion recommendation. Some studies have attempted to integrate image, text, and user behavior data, and achieved certain progress by implementing cross-modal feature interaction through attention mechanisms or graph neural networks [4]. However, existing methods still have obvious deficiencies in terms of dynamic weight allocation for multi-behavior, behavioral timeliness modeling, and deep fusion of multi-modal features. Most studies have not systematically considered the weight differences among different user behaviors, nor have they conducted explicit modeling of the time attenuation effect of behaviors, resulting in considerable room for improvement in the timeliness and personalization of recommendation results. Meanwhile, the fusion of visual and textual features mostly stays at the level of shallow concatenation or simple attention weighting, failing to achieve genuine semantic-level interaction and adaptive feature selection, which limits the model's ability to understand users' complex preferences and hinders the transparency of recommendation logic.

To address the aforementioned research gaps, this paper proposes a personalized fashion recommendation method that integrates multi-behavior weights and multi-modal features. The main contributions of this study are summarized as follows:

- A dynamic weighted collaborative filtering algorithm is designed, which comprehensively incorporates users' multi-dimensional behavior data, assigns differentiated weights according to the importance of each behavior, and introduces a time attenuation factor to construct a time-sensitive user-item scoring matrix, thereby alleviating the problems of data sparsity and insufficient behavioral timeliness.
- A multi-modal deep fusion recommendation framework is constructed: ResNet-50 and BERT are respectively adopted to extract visual and textual features, and a multi-head self-attention mechanism is designed to realize fine-grained interaction and adaptive fusion of cross-modal features. This enhances the semantic consistency and expressive power of product representations, providing a richer feature foundation for personalized recommendation.
- An end-to-end deep prediction network is proposed to achieve accurate score prediction of user preferences. Experiments demonstrate that the proposed method significantly outperforms current mainstream recommendation models on key metrics such as Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG), which verifies the effectiveness and superiority of the method.

The rest of this paper is structured as follows. Section 2 systematically reviews the key research achievements in the field of personalized fashion recommendation, clarifying the research background and academic positioning of the work presented herein. Section 3 elaborates on the proposed personalized recommendation method that integrates multi-behavior and multi-modal features in detail. Section 4 introduces the experimental design, dataset construction, and evaluation metric setting, and systematically verifies and deeply analyzes the effectiveness of the proposed method through comparative experiments and ablation experiments. Section 5 summarizes the main contributions and experimental findings of this study, and prospects future research directions.

2 Related Work

To situate the contributions of this study within the broader context of personalized fashion recommendation research, we review two key aspects of existing studies: multi-modal fusion methods and personalized recommendation technologies.

2.1 Multi-Modal Fusion Methods

In terms of multi-modal fusion methods, the importance of visual and textual information in fashion recommendation has attracted extensive attention. Early studies, such as the VBPR model proposed by He and McAuley [5], first integrated visual features extracted by convolutional neural networks into the Bayesian personalized ranking framework, which significantly enhanced the visual perception capability of recommendation and verified the effectiveness of visual features in alleviating data sparsity and the cold-start problem. Subsequent studies have further advanced on this basis: Kang et al. [6] strengthened the feature representation capability by jointly training the visual encoder and recommendation model in an end-to-end manner; Hou et al. [7] introduced a semantic attribute space to achieve fine-grained alignment between user preferences and product attributes, thereby improving the interpretability of the model. Meanwhile, cross-modal fusion has gradually become a research hotspot—Chen et al. [8] correlated user reviews with image regions through the attention mechanism, providing visual evidence for recommendation results. In recent years, multi-modal fusion technology has been further developed: Laenen and Moens [9] compared various feature fusion strategies and pointed out that the attention mechanism has advantages in fusing visual and textual features; Wu et al. [10] combined visual and textual features to enhance the interpretability of recommendation. These studies have promoted the in-depth application of multi-modal information in fashion recommendation systems and provided technical support for more accurate modeling of user preferences.

2.2 Personalized Recommendation Technologies

In terms of personalized recommendation technologies, research has mainly focused on three directions: collaborative filtering, sequential modeling, and deep learning. Traditional collaborative filtering methods, represented by matrix factorization, can effectively capture the implicit relationships between users and items. To incorporate more auxiliary information, He et al. proposed the VBPR model [5], which integrated visual features into the collaborative filtering framework to improve the accuracy of fashion recommendation. With the diversification of behavior data, multi-behavior modeling has gradually attracted attention. In the field of sequential recommendation, FPMC [11] and translation-based models [12] have been applied to capture the temporal dependencies of user behaviors. Ding et al. [13,14] further introduced immediate intention modeling to distinguish different behavior patterns such as substitution and matching. In recent years, deep learning techniques such as neural collaborative filtering have modeled nonlinear interactions through multi-layer perceptrons. Driven by graph neural networks and Transformers, the complex relationships between users and items have been depicted in a more refined manner. Li et al. [15] proposed a personalized matching recommendation method based on hierarchical graph neural networks, which effectively integrates user behaviors and product attributes; Chen et al. [16] implemented a Transformer-based personalized outfit generation system on Alibaba's iFashion platform. These studies have provided diversified methodological support for personalized fashion recommendation and achieved remarkable results in practical e-commerce scenarios.

3 Methods

3.1 Core Idea

This paper proposes a recommendation method integrating multi-behavior weights and multi-modal features, aiming to construct a more accurate and dynamic personalized fashion recommendation system. First, this method comprehensively analyzes users' multi-dimensional behavioral data, including browsing, adding to cart, favoriting, and purchasing. Differentiated weights are assigned according to the user decision-making costs and interest intensity reflected by different behaviors; meanwhile, a time attenuation factor is introduced to dynamically adjust the behavior weights, so that recent behaviors exert a greater impact on user preference modeling, and behavior frequency is taken into account to strengthen the signals of sustained interests. On this basis, a time-sensitive user-item scoring matrix is constructed to fully capture the dynamic changes of user interests. Furthermore, to make full use of product information, a deep multi-modal fusion model is designed: ResNet-50 is employed to extract visual features, the BERT model is used to capture textual semantic features, and the multi-head self-attention mechanism is adopted to achieve deep fusion of cross-modal features, thereby obtaining more expressive product representations. Finally, an end-to-end deep prediction network combines user preferences with product multi-modal features to perform rating prediction and generate personalized recommendation lists, satisfying the personalized demands of different users.

The overall architecture of the system is illustrated in Fig. 1.

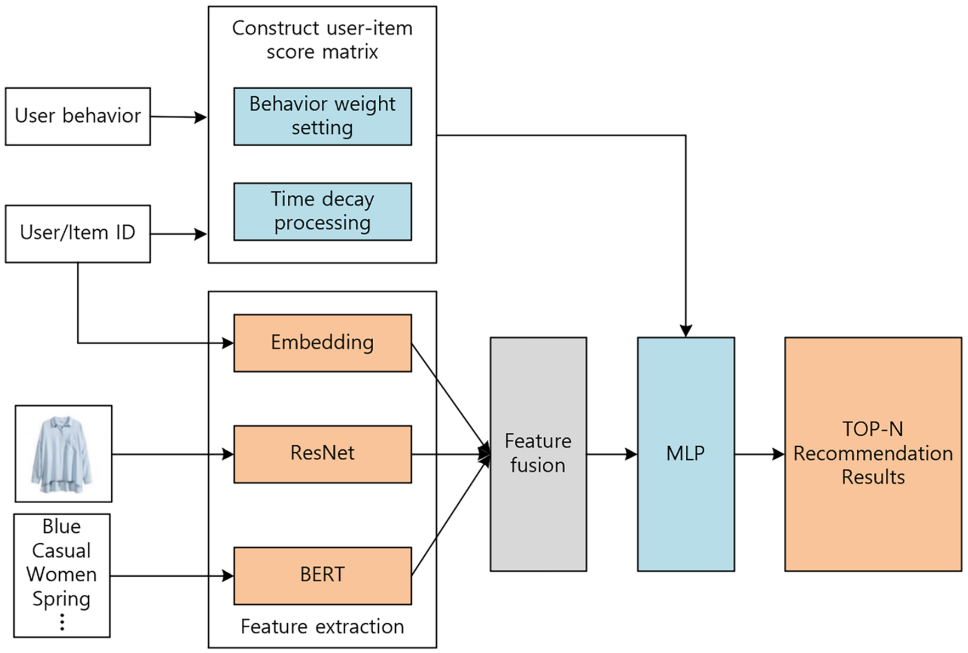

Figure 1: System overall architecture diagram.

3.2 User-Item Scoring Matrix with Multi-Behavior Weights and Time Attenuation

To comprehensively depict user preferences and enhance the timeliness of recommendations, this section proposes a user-item interaction modeling method that integrates multi-behavior weights and a time attenuation mechanism. This method comprehensively considers multiple user behaviors on e-commerce platforms, including browsing, favoriting, adding to cart, and purchasing. Corresponding weights are assigned according to the contribution degree of different behaviors to preference expression, and a time

attenuation function [17] is introduced to dynamically adjust the impact of historical behaviors on current interests. Thus, a user-item scoring matrix that can both reflect users' long-term interests and capture their recent preferences is constructed.

3.2.1 Definition of Multi-Behavior Weights

Different user behaviors imply differentiated preference intensities. In general, the higher the cost users pay to complete a certain behavior, the stronger the indicative effect of this behavior on their preferences. Based on this, this study assigns basic weight scores to four common e-commerce behaviors: a single browsing action is assigned a preference score of 1; favoriting a product is assigned a preference score of 3; adding a product to the shopping cart is assigned a preference score of 5; and purchasing a product is assigned a preference score of 10. Purchasing a product represents the most explicit preference signal of users, so it is assigned the highest weight; adding a product to the shopping cart indicates that users have the intention to purchase in the future, so it is assigned the second-highest weight; favoriting a product reflects users' potential interest in the product, so it is assigned a lower weight than cart-adding; browsing only represents preliminary attention, thus it is assigned the lowest weight. The weight system formed in this way is consistent with the psychological costs and behavioral intentions in the process of user decision-making, ensuring that the model can distinguish the preference signals behind different behaviors.

3.2.2 Temporal Attenuation Modeling

User interests evolve over time, and recent behaviors are more indicative of current preferences. To characterize this phenomenon, an exponential decay function is introduced to temporally adjust the behavior weights. The temporal attenuation formula adopts an exponential decay model [17], incorporating a time attenuation factor to address the timeliness of user behaviors. Temporal attenuation is expressed as Eq. (1):

$$\text{decay} = e^{-\lambda \cdot \text{daysAgo}} \tag{1}$$

where daysAgo denotes the number of days elapsed since the behavior occurred, and λ represents the attenuation coefficient, set to a value of 0.05. A larger value of λ corresponds to a faster attenuation rate. For instance, for a behavior occurring on the current day, daysAgo = 0 and decay = 1.0, meaning the weight remains unchanged without attenuation; for a behavior occurring 7 days prior, decay = $e^{-0.05\times7} \approx 0.7$, with the weight attenuated to 70% of its original value; for a behavior occurring 30 days prior, decay = $e^{-0.05\times30} \approx 0.22$, with the weight attenuated to 22% of its original value. Through the temporal attenuation mechanism, the model can adaptively emphasize the impact of recent behaviors, thereby capturing the dynamic changes in user interests more effectively.

3.2.3 Construction of the User-Item Scoring Matrix

Based on the aforementioned behavior weights and temporal attenuation mechanism, a preference score can be calculated for each user-item interaction pair. For each behavior of user u toward item i, the score calculation formula is defined as follows:

$$s_{u,i}^{(t)} = w_{\text{action}} \cdot \text{decay}(t) \tag{2}$$

where w_{action} denotes the basic weight corresponding to the specific behavior. If a user performs the same or different behaviors on the same item at multiple time points, the scores of these behaviors are accumulated. For example, if User A browsed Item P twice 3 days ago, the preference score of this behavior would be

$2 \cdot e^{-0.05*3} \approx 1.7214$, and this score is added to the interaction term between User A and Item P. The cumulative preference score of the user for the item is thus formulated as:

$$S_{u,i} = \sum_{t \in T_{u,i}} s_{u,i}^{(t)} \tag{3}$$

After iterating through all user behavior logs, a user-item scoring matrix with dimensions of $u \times i$ can be constructed, where u represents the total number of users and i represents the total number of items. Each element $S_{u,i}$ in the matrix comprehensively reflects the preference intensity of user u for item i across multiple behaviors and time points. It incorporates both the information of behavior types and time sensitivity, thereby providing high-quality initial interaction representations for subsequent recommendation tasks.

3.3 Multi-Modal Fusion Recommendation

To construct a recommendation system that comprehensively leverages user behaviors, product attributes, and visual information, we design a multi-modal fusion framework whose core workflow includes embedding representation learning, image and text feature extraction, cross-modal feature interaction, and final user preference prediction. The specific steps are described as follows.

3.3.1 User and Item Embedding Representation

Discrete identifiers (IDs) of users and items are first mapped to low-dimensional continuous vectors through an embedding layer [18], so as to capture their latent semantic correlations. Let the user set be U, the item set be I, and the embedding dimension be d. For a user $u \in U$ and an item $i \in I$, their embedding vectors are defined, respectively as:

$$\mathrm{e}_u = \mathrm{Emb}_{\mathrm{user}}(u) \in \mathbb{R}^d, \mathrm{e}_i = \mathrm{Emb}_{\mathrm{item}}(i) \in \mathbb{R}^d \tag{4}$$

After the above processing, high-dimensional sparse symbolic features are converted into dense vectors, which facilitates high-order interaction modeling by subsequent neural networks, mitigates the curse of dimensionality problem, and lays a foundation for the fusion of multi-modal features.

3.3.2 Product Visual Feature Extraction

Visual information plays a crucial role in fashion recommendation, as it can reflect the intuitive attributes of products such as style, color, and texture. To fully leverage image information, we adopt ResNet-50 [19], a deep convolutional neural network based on residual structures, for feature extraction. ResNet-50 alleviates the gradient vanishing problem in deep networks through residual connections and exhibits strong feature representation capability. Prior to feature extraction, the original images need to be preprocessed to meet the input requirements of the network.

Given an original product image $I_{\mathrm{raw}} \in \mathbb{R}^{H \times W \times C}$, where H, W and C denote height, width, and the number of channels, respectively. First, the image is resized to a fixed size of 224×224 pixels, yielding $I_{\mathrm{resized}} \in \mathbb{R}^{224 \times 224 \times 3}$. Subsequently, normalization is performed [20], which linearly transforms the pixel values from the range $[0, 255]$ to $[0, 1]$:

$$\hat{I}(i, j, k) = \frac{I_{\mathrm{resized}}(i, j, k)}{255} \tag{5}$$

To further stabilize the training process, standardization is implemented using the statistics of the ImageNet dataset [21]:

$$\tilde{I}(i,j,k) = \frac{\hat{I}(i,j,k) - \mu_k}{\sigma_k} \tag{6}$$

where μ_k and σ_k represent the mean and standard deviation of the k-th channel in the ImageNet dataset, respectively.

The preprocessed image is fed into the ResNet-50 network, and high-level visual features are extracted through multiple layers of convolution and pooling operations. Let $\mathrm{W}_{\mathrm{conv}}$ and $\mathrm{b}_{\mathrm{conv}}$ denote the weights and biases of the convolutional layers, respectively. The feature extraction process can be expressed as:

$$\mathrm{F}_{\mathrm{conv}} = \mathrm{ConvNet}(\tilde{I}; \mathrm{W}_{\mathrm{conv}}, \mathrm{b}_{\mathrm{conv}}) \in \mathbb{R}^{H' \times W' \times D} \tag{7}$$

where H' and W' are the spatial dimensions of the feature map, and D is the number of channels. To convert the feature map into a fixed-length vector, we first employ Global Average Pooling [22] to compress it into a D-dimensional vector, followed by nonlinear transformation and dimensionality reduction through two fully connected layers [23]:

$$\mathrm{h}_v^{(1)} = \mathrm{GELU}(\mathrm{W}_1 \cdot \mathrm{GAP}(\mathrm{F}_{\mathrm{conv}}) + \mathrm{b}_1) \tag{8}$$

$$\mathrm{v}_i = \mathrm{W}_2 \cdot \mathrm{h}_v^{(1)} + \mathrm{b}_2 \in \mathbb{R}^{300} \tag{9}$$

where $\mathrm{W}_1 \in \mathbb{R}^{d_m \times D}$ is the weight matrix of the first fully connected layer, $\mathrm{b}_1 \in \mathbb{R}^{d_m}$ is the corresponding bias term, $\mathrm{W}_2 \in \mathbb{R}^{300 \times d_m}$ is the weight matrix of the second fully connected layer, and $\mathrm{b}_2 \in \mathbb{R}^{300}$ is the corresponding bias term. GELU refers to the Gaussian Error Linear Units activation function [24]. The resulting v_i is the visual feature vector of product i, which encodes visual information ranging from low-level textures to high-level semantics. The recommendation system utilizes these feature vectors, combined with user behavior data and other product information, to provide personalized recommendations for users.

3.3.3 Product Textual Feature Extraction

Textual information such as product titles contains abundant semantic attributes, including category, style, and material. To effectively utilize textual information, we adopt the pre-trained BERT model for text encoding [25]. Based on the Transformer architecture, BERT can capture complex semantic relationships between words through bidirectional contextual modeling.

First, the raw text is preprocessed, including removal of special characters, stop-word filtering, and advertisement text cleaning, yielding a cleaned text sequence $\mathrm{S} = [w_1, w_2, \ldots, w_L]$, where w_t denotes the t-th token and L is the sequence length. The sequence S is fed into the BERT model to obtain hidden state representations of all layers. To integrate shallow lexical information and deep semantic information, we extract the outputs of the first and last layers of BERT and perform average pooling [22] on them separately:

$$\mathrm{h}_{\mathrm{first}} = \frac{1}{L}\sum_{t=1}^{L} \mathrm{H}_{\mathrm{first}}^{(t)}, \mathrm{h}_{\mathrm{last}} = \frac{1}{L}\sum_{t=1}^{L} \mathrm{H}_{\mathrm{last}}^{(t)} \tag{10}$$

where $\mathrm{H}_{\mathrm{first}}, \mathrm{H}_{\mathrm{last}} \in \mathbb{R}^{L \times d_h}$ are the hidden state matrices of the first and last layers, respectively, and $d_h = 768$ is the dimension of the BERT hidden layer. The two pooled vectors are fused in a weighted manner:

$$\mathrm{T} = \alpha \cdot \mathrm{h}_{\mathrm{first}} + (1 - \alpha) \cdot \mathrm{h}_{\mathrm{last}} \tag{11}$$

where α is an adjustable weight parameter, set to 0.5 in this study.

To further adapt to the recommendation task, the fused vector is mapped to the target dimension through a fully connected layer [23]:

$$t_i = \mathrm{ReLU}(W_t \cdot T + b_t) \in \mathbb{R}^{300} \tag{12}$$

where $W_t \in \mathbb{R}^{300 \times d_h}$ and $b_t \in \mathbb{R}^{300}$. ReLU refers to the Rectified Linear Unit activation function [26]. The resulting t_i is the textual feature vector of product i, which contains the semantic attribute information of the product.

3.3.4 Deep Fusion of Cross-Modal Features

To achieve semantic-level interaction and adaptive fusion between image and textual features, this paper designs a cross-modal fusion module based on the multi-head self-attention mechanism [27]. This module can dynamically capture fine-grained correlations between visual and textual features, realize feature complementarity and enhancement through a collaborative attention mechanism in multi-subspaces, and ultimately form a unified multi-modal feature representation. The fusion process is detailed as follows:

First, the extracted 300-dimensional image feature vector $E_1 \in \mathbb{R}^{300}$ and textual feature vector $E_2 \in \mathbb{R}^{300}$ are concatenated to form the initial multi-modal input vector:

$$E = \mathrm{Concat}(E_1, E_2) \in \mathbb{R}^{600} \tag{13}$$

To establish semantic correlations between images and texts, the multi-head self-attention mechanism is introduced. This mechanism first maps the concatenated feature E to the query matrix Q, key matrix K and value matrix V through linear transformations, respectively:

$$Q = EW_Q, K = EW_K, V = EW_V \tag{14}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{600 \times d_k}$ are learnable projection weight matrices, and d_k denotes the dimension of the key vectors.

The attention weights are calculated by scaled dot-product [27] and normalized using the softmax function:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{15}$$

The scaling factor $\sqrt{d_k}$ is used to prevent gradient instability caused by excessively large dot-product results.

To enhance the representation capability of the model, the multi-head attention mechanism is adopted. The matrices Q, K and V are split into h groups ($h = 8$) along the feature dimension, and the attention output is computed independently in each subspace:

$$\mathrm{head}_i = \mathrm{Attention}\left(EW_Q^i, EW_K^i, EW_V^i\right), i = 1, \ldots, h. \tag{16}$$

The outputs of all heads are then concatenated and integrated into the final multi-head attention representation through a linear transformation layer [27]:

$$\mathrm{MultiHead}(E) = \mathrm{Concat}\left(\mathrm{head}_1, \ldots, \mathrm{head}_h\right) \cdot W_O \tag{17}$$

where $W_O \in \mathbb{R}^{h \cdot d_k \times 600}$ is the output projection matrix.

The attention-weighted image feature $E_1^{'}$ and textual feature $E_2^{'}$ are further concatenated to form the fused multi-modal feature representation:

$$F_{fused} = \text{Concat}\left(E_1^{'}, E_2^{'}\right) \in \mathbb{R}^{600} \tag{18}$$

This fused feature F_{fused} retains both visual details and textual semantic information simultaneously, and has strong expressiveness and discriminability, thus providing a robust feature foundation for subsequent recommendation prediction tasks. The above architecture realizes deep interaction and adaptive fusion of image and textual features via the multi-head self-attention mechanism, effectively enhances the semantic consistency and representation capability of multi-modal features, and lays a reliable feature foundation for fashion matching recommendation.

3.3.5 Deep Prediction and Recommendation Generation

After completing the multi-modal fusion of image and text features, the fused feature $F_{fused} \in \mathbb{R}^{600}$ is fed into a two-layer fully connected neural network for the final user preference prediction. The first hidden layer contains 256 neurons and adopts the ReLU activation function to enhance the nonlinear expression capability of the model [26]. The output of this layer can be expressed as:

$$H = \text{ReLU}(W_1 \cdot F_{fused} + b_1) \tag{19}$$

where $W_1 \in \mathbb{R}^{256 \times 600}$ is the weight matrix and $b_1 \in \mathbb{R}^{256}$ is the bias term.

The second layer directly outputs a 1-dimensional result, which serves as the predicted score of the user for the candidate item:

$$\hat{y} = W_2 \cdot H + b_2 \tag{20}$$

where $W_2 \in \mathbb{R}^{1 \times 256}$ and $b_2 \in \mathbb{R}^1$. The model is optimized using the mean squared error (MSE) loss function [28]:

$$Loss = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{21}$$

where y_i denotes the actual score of the user for the item, $\hat{y}_i$ denotes the predicted score of the model, and N is the total number of samples. This loss function enables the model to fully leverage multi-source fused features to achieve accurate prediction of user scores, thereby improving the precision of personalized recommendations.

After the training process is completed, the model is saved in a deployable format to support real-time inference of the recommendation system. In the online recommendation phase, the system first dynamically generates a multi-behavior weighted profile for the target user based on their historical behaviors. Subsequently, the user profile and the multi-modal features of candidate items are fed into the trained deep prediction network to calculate the preference score of the user for each candidate item. All candidate items are sorted in descending order according to the predicted scores. Finally, the top 5 ranked items are selected to form the Top-5 personalized recommendation list, which is presented to the user in real time. This process realizes a full-link closed loop from multi-modal feature fusion to personalized recommendation generation, providing fashion e-commerce platforms with efficient and accurate recommendation service capabilities.

In practical large-scale deployment, considering the computational costs of deep neural networks and multimodal feature extraction, this system adopts a phased recommendation strategy to ensure efficiency and scalability. First, visual and textual features of products are pre-calculated offline and stored in an efficient vector database, avoiding real-time online computation. During the online recommendation phase, the system typically employs a two-stage architecture of "recall-sorting": In the recall stage, low-computational-cost methods (e.g., simple collaborative filtering based on user historical behavior or rule matching) are used to rapidly screen hundreds to thousands of potentially relevant candidate products from a massive product database. Subsequently, the deep prediction network proposed in this study refines the ranking of these limited candidate products, ensuring recommendation accuracy while effectively controlling the computational overhead of online inference. Additionally, for ultra-large product libraries, the system can integrate Approximate Nearest Neighbor (ANN) technology to further optimize the retrieval efficiency of candidate products.

4 Experiments and Analysis

4.1 Experimental Setup

To comprehensively verify the effectiveness of the personalized fashion recommendation method integrating multi-behavior and multi-modal features proposed in this paper, this section systematically introduces the composition of datasets used in the experiments, data preprocessing procedures, experimental operating environment, key parameter configurations, and specific training strategies, ensuring the consistency and reproducibility of the experiments. The experiments are conducted based on real-scenario data and public datasets, guaranteeing that the evaluation process is reliable, reproducible, and of practical application value.

4.1.1 Datasets

The data in this study are derived from the user behavior logs and product information database of the fashion recommendation system constructed in this framework, combined with a public image dataset to enhance the completeness of visual features. User behavior logs record users' interactive behaviors on the platform, such as browsing, favoriting, adding to cart, and purchasing. Each log contains key information including user ID, behavior type, and behavior timestamp. The product information database covers structured descriptions such as product ID, category, style, gender attribute, and season label. To improve the quality of visual representation, this study introduces standardized fashion images from the public Fashion Product Images Dataset [29] for training and extracting visual features. After cleaning and integration, the final dataset includes approximately 20,000 users, 10,000 products, and a cumulative total of about 200,000 behavior records, which has a favorable data scale and realistic representativeness.

In the data preprocessing phase, duplicate, incomplete, and abnormal records are first removed, and the behavior timestamps are uniformly formatted to facilitate the subsequent calculation of time attenuation factors. To further ensure the rationality of training and evaluation, a stratified sampling method is adopted to split the dataset by user dimension: 80% of the data is used as the training set for model learning and parameter tuning, and 20% is used as the test set for performance verification. This splitting strategy effectively maintains the consistency of user behavior distribution and product attributes between the training set and the test set, avoiding evaluation bias caused by data skew.

The dataset employed in this study exhibits typical characteristics of e-commerce platforms, featuring a highly sparse user-item interaction matrix. Specifically, with approximately 20,000 users and 10,000 items, the total potential interactions reach 200 million, while only about 200,000 actual behavioral records exist,

resulting in an interaction density of roughly 0.1%. This high sparsity represents one of the primary challenges for traditional recommendation systems. Our approach effectively mitigates this issue by integrating multi-behavioral information and multi-modal features.

In terms of behavioral patterns, user actions demonstrate distinct hierarchical characteristics. Browsing activities dominate the data, followed by bookmarking and cart additions, while purchases account for a relatively small proportion. This structure mirrors the typical user journey from interest discovery to final purchase decision. For instance, browsing constitutes approximately 60% of total activity records, with bookmarking and cart additions each representing around 15%, while purchases make up about 10%. This uneven distribution underscores the necessity of assigning dynamic weights to different behavioral actions. Additionally, both users and products exhibit a long-tail distribution pattern, where a small number of popular items and active users generate the majority of interactions, while a vast number of products and users engage in minimal activity.

To address cold-start scenarios, particularly for product recommendations, this framework leverages multimodal features including visual and textual data. Even when new products lack historical interaction data, it generates meaningful initial recommendations through their comprehensive attribute information. This enables the model to provide "cold-start" capabilities for new products, overcoming the limitations of traditional collaborative filtering methods. For user cold-start scenarios, the model primarily relies on users' historical behaviors, while multimodal features can also assist in preliminary recommendations based on initial interactions or demographic characteristics.

4.1.2 Experimental Environment Configuration

The experiments are carried out on a computing platform equipped with a single NVIDIA RTX 3080 GPU. The software environment is built around Python 3.8, and the model construction and training process are implemented based on the PyTorch 1.11 deep learning framework. User behavior data and product metadata are managed uniformly through a MySQL database to support efficient data access and preprocessing. The detailed software and hardware environment configurations are shown in Table 1.

Table 1: Experimental environment configuration.

Category	Configuration
Operating System	Ubuntu 20.04 LTS
CPU	Intel Core i9-10900K @ 3.70GHz
GPU	NVIDIA RTX 3080
Programming Language	Python 3.8
Deep Learning Framework	PyTorch 1.11
CUDA Version	11.3
Data Processing Libraries	NumPy 1.21, Pandas 1.3
Database Management System	MySQL 8.0

4.1.3 Training and Implementation Details

In the phase of model training and parameter tuning, the pre-trained ResNet-50 network is adopted for image feature extraction and fine-tuned on the Fashion-MNIST dataset to enhance the model's capability of recognizing fashion visual features. During the training process, the learning rate is set to 1×10^{-4} and the batch size is fixed at 128. The Adam optimizer [30] is selected to balance convergence speed and stability. In

addition, an early stopping strategy [31] is introduced to mitigate the risk of overfitting: the training process is terminated in advance if the validation set loss does not decrease for 5 consecutive epochs. Meanwhile, the time attenuation coefficient λ is set to 0.05 to dynamically adjust the weight of users' historical behaviors in current interest modeling.

4.2 Evaluation Metrics

The performance of the recommendation system is evaluated primarily using two metrics: Hit Ratio (HR) [32] and Normalized Discounted Cumulative Gain (NDCG) [33].

HR (Hit Ratio): This metric measures the proportion of items in the recommendation list that are actually clicked by users. A higher HR value indicates better performance of the recommendation system. For each user, we check the top-N recommended items. If at least one of these items is of actual interest to the user (e.g., clicked, purchased, etc.), the HR@N value for this user is set to 1; otherwise, it is set to 0. The average HR@N across all users is calculated to evaluate the system's ability to recommend relevant items within the top-N positions of the recommendation list.

The calculation of Hit Ratio [32] is shown in Eq. (22):

$$\mathrm{HR@N} = \frac{1}{|\mathrm{U}|} \sum_{\mathrm{u} \in \mathrm{U}} \mathrm{HR_u} \tag{22}$$

where $|\mathrm{U}|$ denotes the total number of users, and $\mathrm{HR_u}$ represents the HR@N value of user u.

NDCG (Normalized Discounted Cumulative Gain) [33]: This metric assesses the relevance of recommendation results while taking the ranking of recommended items into account. A higher NDCG value implies that the recommended items are more aligned with user interests. The calculation of Normalized Discounted Cumulative Gain is defined as follows:

DCG@N (Discounted Cumulative Gain at N) [33] measures the cumulative relevance of recommended items. It considers both the relevance of each recommended item and the importance of higher-ranked items. It is formulated as Eq. (23):

$$\mathrm{DCG@N} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\mathrm{rel}\,(\mathrm{i})}{\log_2\,(\mathrm{i}+1)} \tag{23}$$

where $\mathrm{rel}\,(\mathrm{i})$ denotes the relevance score of the i-th recommended item, and i is the position of the item in the recommendation list.

IDCG@N (Ideal Discounted Cumulative Gain at N) [33] is the ideal value of DCG@N, representing the optimal scenario where the recommended items are sorted in perfect order of relevance. Its calculation is identical to that of DCG@N, except that it uses the items sorted by their relevance scores. It is expressed as Eq. (24):

$$\mathrm{IDCG@N} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\mathrm{rel}'\,(\mathrm{i})}{\log_2\,(\mathrm{i}+1)} \tag{24}$$

where $\mathrm{rel}'\,(\mathrm{i})$ is the relevance score of the i-th item in the ideal sorted list.

NDCG@N (Normalized DCG at N) [33] evaluates the performance of the recommendation system by normalizing the DCG@N value, ensuring the comparability across different recommendation systems. It is defined in Eq. (25):

$$\text{NDCG@N} = \frac{\text{DCG@N}}{\text{IDCG@N}} \tag{25}$$

A better ranking of the recommendation system results in an NDCG@N value closer to 1, whereas a poor ranking leads to a value closer to 0.

4.3 Comparative Experiments

To comprehensively evaluate the effectiveness of the proposed method, five representative recommendation algorithms are selected as baseline models for comparative experiments:

- ItemCF [34]: An item-based collaborative filtering algorithm that generates recommendations by calculating the co-occurrence similarity between items. This method constructs an item similarity matrix using user historical behavior data, and recommends candidate items similar to those that match the user's historical preferences.
- NCF [35]: A neural collaborative filtering algorithm that models the interaction relationships between users and items via a multi-layer perceptron. This model learns the latent feature representations of users and items through a neural network, and predicts user preference scores for items.
- VBPR [5]: A visual Bayesian personalized ranking model that incorporates item visual features extracted by a convolutional neural network into the Bayesian personalized ranking framework. This model integrates visual information into the collaborative filtering process, enhancing the capability of modeling item appearance attributes.
- LightGCN [36]: A lightweight graph convolutional network recommendation model that performs information propagation based on the user-item interaction graph structure. This model aggregates neighbor node information through linear propagation layers to learn graph embedding representations of users and items.
- LATTICE [37]: A multi-modal recommendation model based on latent structure learning, which constructs an item semantic graph by mining the visual and textual similarity between items. This model leverages multi-modal information to enhance item representation learning, thereby improving recommendation performance.
- MARIO [38]: A multimedia recommendation framework based on modality-aware attention and modality-preserving decoders. The model dynamically fuses visual, textual, and interaction modalities at the interaction level through an attention mechanism, while designing decoder layers to preserve modality-specific properties, effectively improving the accuracy of multimodal recommendation.
- MCCL [39]: A multimodal recommendation model based on multi-channel counterfactual learning. The model constructs multi-channel causal graphs to perform counterfactual reasoning on the modal features of user-interacted and uninteracted items, respectively, eliminating preference-irrelevant feature noise, thereby achieving fine-grained modeling of users' multimodal preferences.

Table 2 presents the performance comparison results between the aforementioned baseline models and the personalized recommendation algorithm proposed in this paper on the test set. To verify the statistical significance of the results, all experiments were repeated 5 times, and the 95% confidence intervals are reported.

Table 2: Results of comparative experiments.

Model	HR@3	NDCG@3	HR@5	NDCG@5
ItemCF [34]	0.546 ± 0.006	0.401 ± 0.005	0.617 ± 0.007	0.432 ± 0.006
NCF [35]	0.578 ± 0.005	0.412 ± 0.005	0.642 ± 0.006	0.448 ± 0.005
VBPR [5]	0.593 ± 0.007	0.425 ± 0.006	0.658 ± 0.008	0.461 ± 0.007
LightGCN [36]	0.605 ± 0.004	0.431 ± 0.004	0.672 ± 0.005	0.472 ± 0.004
LATTICE [37]	0.613 ± 0.005	0.439 ± 0.004	0.681 ± 0.006	0.478 ± 0.005
MARIO [38]	0.626 ± 0.005	0.443 ± 0.005	0.689 ± 0.005	0.482 ± 0.005
MCCL [39]	0.631 ± 0.003	0.449 ± 0.003	0.694 ± 0.004	0.485 ± 0.003
Proposed Model	0.640 ± 0.005	0.447 ± 0.003	0.703 ± 0.004	0.491 ± 0.003

Experimental results demonstrate that the proposed method exhibits significant advantages across multiple evaluation metrics. Confidence interval analysis indicates that the improvements of the proposed model on HR@3, HR@5, and NDCG@5 metrics all exceed the margin of error, proving that the performance gains are statistically significant and not caused by randomness. Specifically, in terms of HR@3, the proposed method achieves a score of 0.640, representing a notable improvement over baseline models such as ItemCF (0.546). For the HR@5 metric, the proposed method reaches 0.703, outperforming all comparison models. In terms of ranking quality, the proposed method performs best on NDCG@5 (0.491), reflecting its comprehensive ranking advantage in generating long recommendation lists. Although MCCL (0.449) slightly outperforms the proposed method (0.447) on NDCG@3, the advantages of our method in deep multi-modal fusion and multi-behavior dynamic modeling become increasingly prominent as the length of the recommendation list increases. In summary, by integrating multi-behavior weights, temporal attenuation mechanisms, and deep multi-modal features, the proposed method can capture user preferences more comprehensively and dynamically, significantly and stably improving recommendation performance.

4.4 Ablation Experiments

To systematically verify the effectiveness of each key module in the personalized recommendation model integrating multi-behavior and multi-modal features proposed in this paper, we designed the following ablation experiment scheme: by removing the core modules of the model one by one (temporal attenuation, multi-behavior weights, textual features, visual features), we compared the performance differences between each variant model and the complete model, so as to clarify the contribution value of each module. The results of the ablation experiments are shown in Table 3:

Table 3: Results of ablation experiments.

Model	HR@5	NDCG@5
w/o Temporal attenuation	0.671	0.472
w/o Multi-behavior weights	0.658	0.463
w/o Textual features	0.635	0.452
w/o Visual features	0.622	0.442
Proposed Model	0.703	0.491

As can be seen from the table, removing any module will lead to a decline in model performance, indicating that each module plays an important role in improving recommendation effectiveness. Specifically, after removing the temporal attenuation mechanism, HR@5 and NDCG@5 drop to 0.671 and 0.472, respectively, which demonstrates that the temporal attenuation factor can effectively enhance the timeliness of recommendation results and make the model pay more attention to the changes in users' recent interests. After removing multi-behavior weights, HR@5 and NDCG@5 further decrease to 0.658 and 0.463, verifying the importance of distinguishing different behaviors, which helps to depict the intensity of user preferences in a more refined manner. In the experiment with textual features removed, HR@5 and NDCG@5 reach 0.635 and 0.452, respectively, indicating that the semantic information of products makes a significant contribution to understanding user intentions. The removal of visual features has the most notable impact: HR@5 and NDCG@5 drop to 0.622 and 0.442, respectively, highlighting the crucial role of visual features in fashion recommendation. The intuitive attributes such as style and color reflected by visual features exert an important influence on user decision-making. Therefore, all modules in this study work collaboratively to jointly improve the comprehensive performance of the recommendation system. This result not only verifies the effectiveness of the multi-modal fusion and multi-behavior modeling ideas proposed in this paper, but also provides a reference for the importance of modules in subsequent research.

4.5 Hyperparameter Experiments

In the proposed model, the weight coefficients assigned to different user behaviors (browsing, favoriting, cart-adding, and purchasing) are the core parameters for constructing the time-sensitive scoring matrix. To systematically investigate the sensitivity of the recommendation model's performance to each behavior weight, verify the rationality of the current settings, and seek the optimal parameter combination, a series of hyperparameter sensitivity experiments were conducted.

The complete model, which integrates multi-behavior and multi-modal features, serves as the benchmark for these experiments. While maintaining the time decay factor λ and the parameters of the multi-modal fusion module constant, a "single-factor variable method" was employed to perform sensitivity analysis on the four user behavior weights. Specifically, the weight of the target behavior was adjusted within a preset range, while the weights of the other three behaviors were fixed at their baseline values (browsing = 1, favoriting = 3, cart-adding = 5, and purchasing = 10). For each weight configuration, the model's performance was re-evaluated on the test set using the HR@5 and NDCG@5 metrics, and evolution curves illustrating the relationship between performance and hyperparameters were plotted, as shown in Fig. 2.

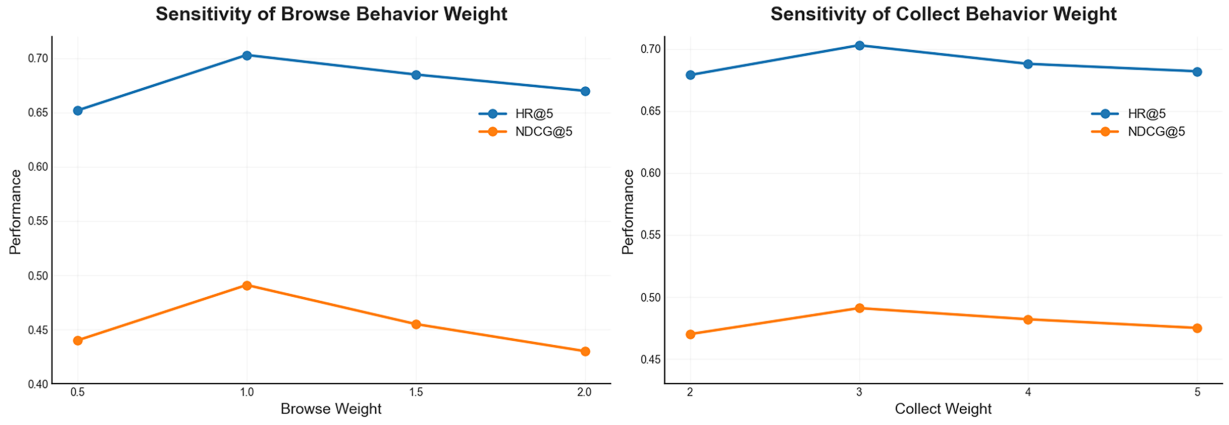

Figure 2: (Continued)

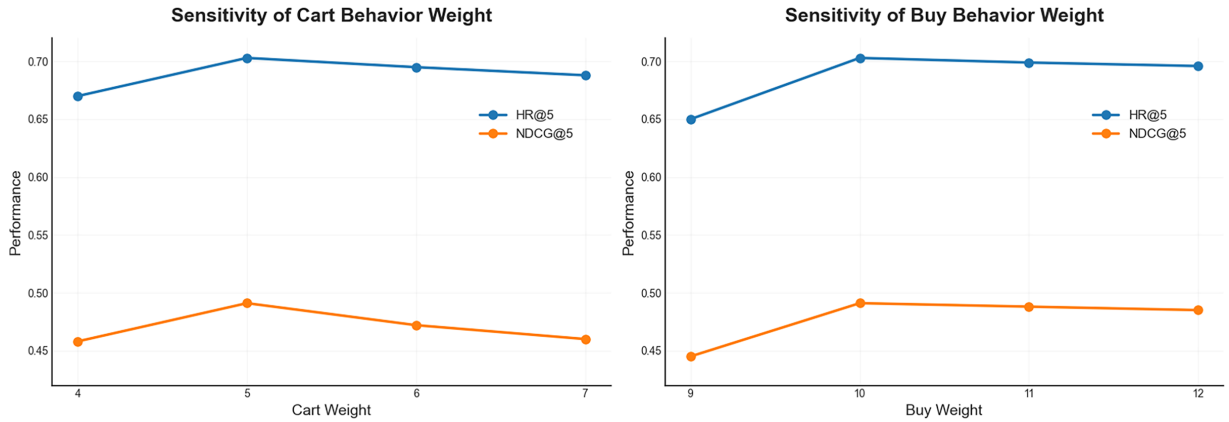

Figure 2: Hyperparameter sensitivity analysis of multi-behavior weights.

Experimental results indicate that each behavior weight significantly influences model performance, with specific optimal value ranges identified for each. The detailed analysis is as follows:

- Browsing Weight: As the weight increased from 0.5 to 2.0, both HR@5 and NDCG@5 exhibited an initial upward trend followed by a decline, peaking at a weight of approximately 1. This suggests that a low weight fails to fully utilize the high-frequency browsing signals, while an excessively high weight introduces interest noise from accidental clicks, interfering with the capture of core preferences.
- Favoriting Weight: Within the weight range of [2,5], model performance improved steadily with increasing weight, reaching its optimum at a value of 3. When the weight exceeded 3, the recommendation accuracy began to decrease. This reflects that while favoriting behavior indicates potential interest, its certainty is weaker than that of cart-adding or purchasing; overemphasizing this signal may bias the model's depiction of the user's actual purchase intent.
- Cart-Adding Weight: Experimental data show that this parameter performs robustly within the [4,7] interval, with the optimal value occurring at a weight of 5. As cart-adding is a strong pre-conversion intent signal, an appropriate weight setting effectively balances its role as a leading indicator for "near-purchase" behavior. A weight that is too low causes this strong signal to be submerged, while one that is too high tends to result in homogenized recommendation lists.
- Purchasing Weight: As the explicit signal with the highest interest intensity, the purchasing weight significantly affects performance within the [9,12] range. The model achieved peak performance when the weight was set to 10. Further increasing the weight yielded no continuous gains; instead, it potentially suppressed the model's responsiveness to users' recent diverse and exploratory interests by over-amplifying historical purchase records.

The aforementioned sensitivity experiments empirically validate the scientific rigor of the weight system adopted in this study. On the tested real-world e-commerce dataset, the preset weights consistently align with the peaks of the performance curves for each behavior. This demonstrates that the system accurately maps the differences in preference intensity and psychological decision-making costs associated with different behaviors.

5 Conclusions

This study proposes a personalized fashion recommendation method fusing multi-behavior weights and multi-modal features to tackle the key limitations of traditional recommendation systems in fashion scenarios. By introducing dynamic weighted collaborative filtering with temporal attenuation and a deep

multi-modal fusion mechanism for cross-modal semantic interaction, the method realizes comprehensive and dynamic modeling of user preferences. Experimental evaluations confirm that the proposed method achieves superior performance over state-of-the-art recommendation models on key metrics, and ablation experiments validate the necessity and synergy of all core modules (temporal attenuation, multi-behavior weighting, and multi-modal feature fusion). This method effectively enhances the accuracy and timeliness of fashion recommendations, and the constructed recommendation system has significant practical application value for e-commerce platforms.

It is worth emphasizing that the framework proposed in this study demonstrates strong generalization capabilities. Its core concepts—integrating multi-behavior information, modeling temporal decay of behaviors, and deeply fusing multimodal features—are not unique to fashion recommendation. For instance, in movie recommendations, user behaviors such as viewing, rating, and collecting can serve as multi-behavior signals, while modal information like movie posters (visual) and plot summaries (text) can be utilized. In music recommendations, user actions including playing, liking, and adding to playlists can be combined with modal features such as album covers (visual), lyrics (text), or audio characteristics (auditory). By adapting behavior types, weights, and corresponding modal feature extractors to specific domains, this framework holds promise for broad application across various recommendation scenarios including movies, music, and news, showcasing extensive potential for practical implementation.

Future work will focus on optimizing the multi-modal feature fusion mechanism, exploring more efficient user-item interaction modeling methods, and attempting to introduce a reinforcement learning framework to realize dynamic adjustment and continuous optimization of recommendation strategies. Meanwhile, we will conduct more detailed quantitative evaluations of efficiency metrics such as model training time and inference latency. We will further explore model compression and optimization techniques including pruning, quantization, and distributed inference to enhance the model's scalability and operational efficiency when processing massive data and high-concurrency requests. This will ultimately improve the system's adaptability and long-term performance in complex e-commerce environments.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the Shandong University of Technology Science and Technology Doctor Startup Fund (Project: Using The Real-Time Services of Variable Rate and Pauseable Non-Real-Time Services; Grant number: 4041/422022), the Research Fund (Project: Patent Assignment for User Identification System of a Neural Network–Based VR Device; Grant number: 9101/22502551), and the Research Fund (Project: Patent Rights Transfer for the Generating 3D Facial Images with a Deep Learning Method; Grant number: 9101/22502561).

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Jian-Hong Wang; methodology, Xin Lu; writing—original draft preparation, Xin Lu; writing—review and editing, Jian-Hong Wang, Kuo-Chun Hsu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Jian-Hong Wang, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chakraborty S, Hoque MS, Rahman Jeem N, Biswas MC, Bardhan D, Lobaton E. Fashion recommendation systems, models and methods: a review. Informatics. 2021;8(3):49. doi:10.3390/informatics8030049.
2. Peng ZF, Zhang HR, Min F. Ideal user group: a new perspective on item representation for recommender systems. J Big Data. 2022;9(1):117. doi:10.1186/s40537-022-00663-7.
3. Chaube S, Kar R, Gupta S, Kant M. Multimodal AI framework for the prediction of high-potential product listings in e-commerce: navigating the cold-start challenge. Expert Syst Appl. 2025;282(4):127524. doi:10.1016/j.eswa.2025.127524.
4. Saed S, Teimourpour B. Hybrid-hierarchical fashion graph attention network for compatibility-oriented and personalized outfit recommendation. Mach Learn Appl. 2026;23(4):100802. doi:10.1016/j.mlwa.2025.100802.
5. He R, McAuley JJ. VBPR: visual Bayesian personalized ranking from implicit feedback. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence; 2016 Feb 12–17; Phoenix, AZ, USA. Palo Alto, CA, USA: AAAI Press; 2016. p. 144–50.
6. Kang WC, Fang C, Wang Z, McAuley J. Visually-aware fashion recommendation and design with generative image models. In: Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM); 2017 Nov 18–21; New Orleans, LA, USA. p. 207–16.
7. Hou M, Wu L, Chen E, Li Z, Zheng VW, Liu Q. Explainable fashion recommendation: a semantic attribute region guided approach. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence; 2019 Aug 10–16; Macao, China. San Francisco, CA, USA: Morgan Kaufmann; 2019. p. 4681–8.
8. Chen X, Chen H, Xu H, Zhang Y, Cao Y, Qin Z, et al. Personalized fashion recommendation with visual explanations based on multimodal attention network: towards visually explainable recommendation. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2019 Jul 21–25; Paris, France. p. 765–74.
9. Laenen K, Moens MF. A comparative study of outfit recommendation methods with a focus on attention-based fusion. Inf Process Manag. 2020;57(6):102316. doi:10.1016/j.ipm.2020.102316.
10. Wu Q, Zhao P, Cui Z. Visual and textual jointly enhanced interpretable fashion recommendation. IEEE Access. 2020;8:68736–46. doi:10.1109/ACCESS.2020.2978272.
11. Rendle S, Freudenthaler C, Schmidt-Thieme L. Factorizing personalized Markov chains for next-basket recommendation. In: Proceedings of the 19th International Conference on World Wide Web; 2010 Apr 26–30; Raleigh, NC, USA. p. 811–20.
12. He R, Kang WC, McAuley J. Translation-based recommendation. In: Proceedings of the Eleventh ACM Conference on Recommender Systems; 2017 Aug 27–31; Como, Italy. p. 161–9.
13. Ding Y, Ma Y, Wong WK, Chua TS. Leveraging two types of global graph for sequential fashion recommendation. In: Proceedings of the 2021 International Conference on Multimedia Retrieval; 2021 Aug 21–24; Taipei, Taiwan. p. 73–81.
14. Ding Y, Ma Y, Wong WK, Chua TS. Modeling instant user intent and content-level transition for sequential fashion recommendation. IEEE Trans Multimed. 2022;24:2687–700. doi:10.1109/TMM.2021.3088281.
15. Li X, Wang X, He X, Chen L, Xiao J, Chua TS. Hierarchical fashion graph network for personalized outfit recommendation. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2020 Jul 25–30; Virtual. p. 159–68. doi:10.1145/3397271.3401080.
16. Chen W, Huang P, Xu J, Guo X, Guo C, Sun F, et al. POG: personalized outfit generation for fashion recommendation at alibaba iFashion. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019 Aug 4–8; Anchorage, AK, USA. p. 2662–70.
17. Götz G, Schlecht SJ, Pulkki V. Common-slope modeling of late reverberation. IEEE/ACM Trans Audio Speech Lang Process. 2023;31:3945–57. doi:10.1109/TASLP.2023.3317572.
18. Cheng W, Shen Y, Huang L, Zhu Y. Dual-embedding based deep latent factor models for recommendation. ACM Trans Knowl Discov Data. 2021;15(5):1–24. doi:10.1145/3447395.

19. Hazra S, Purkayastha R. Fashion product recommendation system using CNN-based ResNet50 model. In: Innovative applications of artificial neural networks to data analytics and signal processing. Cham, Switzerland: Springer Nature Switzerland; 2024. p. 481–97. doi:10.1007/978-3-031-69769-2_19.
20. Srilakshmi V, Deepa RNA, Vidyadhari C, Nimmala S, Gundarapu MR. Revolutionizing trend recommendations: a deep learning approach for image-based insights. ShodhKosh J Vis Per Arts. 2023;4(1):1051–62. doi:10.29121/shodhkosh.v4.i1.2023.2859.
21. Du Q, Wang Y, Tian L. Attention module based on feature normalization. In: Proceedings of the 2023 4th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI); 2023 Aug 4–6; Guangzhou, China. p. 438–42.
22. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 11966–76.
23. Kärkkäinen T, Hänninen J. Additive autoencoder for dimension estimation. Neurocomputing. 2023;551(4):126520. doi:10.1016/j.neucom.2023.126520.
24. Lee M. Mathematical analysis and performance evaluation of the GELU activation function in deep learning. J Math. 2023;2023(1):4229924. doi:10.1155/2023/4229924.
25. Gao D, Jin L, Chen B, Qiu M, Li P, Wei Y, et al. FashionBERT: text and image matching with adaptive loss for cross-modal retrieval. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2020 Jul 25–30; Virtual. New York, NY, USA: Association for Computing Machinery; 2020. p. 2251–60. doi:10.1145/3397271.3401430.
26. Citton O, Richert F, Biehl M. The role of the learning rate in layered neural networks with ReLU activation function. In: Proceedings of the 33rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning; 2025 Apr 23–25; Bruges, Belgium. p. 437–42.
27. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA. Red Hook, NY, USA: Curran Associates; 2017. p. 5998–6008.
28. Li W, Xu B. Aspect-based fashion recommendation with attention mechanism. IEEE Access. 2020;8:141814–23. doi:10.1109/ACCESS.2020.3013639.
29. Fashion product images dataset. 2018 [cited 2025 Dec 1]. Available from: https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset.
30. Jia X, Feng X, Yong H, Meng D. Weight decay with tailored Adam on scale-invariant weights for better generalization. IEEE Trans Neural Netw Learn Syst. 2024;35(5):6936–47. doi:10.1109/TNNLS.2022.3213536.
31. Benfenati A, Catozzi A, Franchini G, Porta F. Early stopping strategies in deep image prior. Soft Comput. 2025;29(8):4153–74. doi:10.1007/s00500-025-10642-8.
32. Yu R, Ye D, Wang Z, Zhang B, Oguti AM, Li J, et al. CFFNN: cross feature fusion neural network for collaborative filtering. IEEE Trans Knowl Data Eng. 2022;34(10):4650–62. doi:10.1109/TKDE.2020.3048788.
33. Zhu T, Jung MC, Clark J. Generalized contrastive learning for multi-modal retrieval and ranking. In: Proceedings of the Companion Proceedings of the ACM on Web Conference 2025; 2025 Apr 28–May 2; Sydney, Australia. p. 661–70.
34. Hwangbo H, Kim YS, Cha KJ. Recommendation system development for fashion retail e-commerce. Electron Commer Res Appl. 2018;28(1):94–101. doi:10.1016/j.elerap.2018.01.012.
35. He X, Liao L, Zhang H, Nie L, Hu X, Chua TS. Neural collaborative filtering. In: Proceedings of the 26th International Conference on World Wide Web; 2017 Apr 3–7; Perth, Australia. p. 173–82.
36. He X, Deng K, Wang X, Li Y, Zhang Y, Wang M. LightGCN: simplifying and powering graph convolution network for recommendation. In: Proceedings of The 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2020 Jul 25–30; Virtual. p. 639–48.
37. Zhang J, Zhu Y, Liu Q, Wu S, Wang S, Wang L. Mining latent structures for multimedia recommendation. In: Proceedings of the 29th ACM International Conference on Multimedia; 2021 Oct 20–24; Virtual. p. 3872–80. doi:10.1145/3474085.3475259.

38. Kim T, Lee YC, Shin K, Kim SW. MARIO: modality-aware attention and modality-preserving decoders for multimedia recommendation. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management; 2022 Oct 17–21; Atlanta, GA, USA. p. 993–1002.
39. Fang H, Sha L, Liang J. Multimodal recommender system based on multi-channel counterfactual learning networks. Multimed Syst. 2024;30(5):242. doi:10.1007/s00530-024-01448-z.