Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2025.24500

Diversified Interpretable Compatibility Modeling Based on Multi-modal Disentanglement

Shuoji Sun[1*], Miao Yu[2], Xu Yu[3]

[1]*School of Data Science, Qingdao University of Science and Technology, Qingdao, China*
[2]*College of Textiles and Clothing, Qingdao University, Qingdao, China*
[3]*Institute of Software, China University of Petroleum (East China), Qingdao, China*
**Corresponding Author. Email: shuojisun@gmail.com*

Abstract. In recent years, compatibility modeling for evaluating whether fashion items match has received widespread attention. The existing compatibility modeling methods typically model the compatibility between fashion items based on multi-modal information. However, these methods often fail to disentangle the rich attribute information in the high-dimensional continuous representations of items, resulting in a lack of interpretability in recommendations. At the same time, they also overlook the diverse matching methods among the attributes of complementary items. This article proposes a Diversified Interpretable Compatibility Modeling based on a Multi-modal Disentanglement model (DICM-MD). In DICM-MD, we adopt disentanglement representation learning technology to disentangle the complex attribute information of fashion items and comprehensively evaluate the compatibility of items through diverse attribute matching methods. Specifically, we use deep neural networks to estimate the mutual information among the dimensions of high-dimensional continuous representations and adopt contrastive loss to encourage each dimension in the item representation to learn independent attribute information. Then, we learn the diverse attribute matching methods between complementary items from the alignment and non-alignment perspectives to model the compatibility of items more comprehensively. We conducted extensive experiments on the IQON3000 and Polyvore datasets, demonstrating that DICM-MD outperforms state-of-the-art methods.

Keywords: Multi-modal, Disentanglement representation learning, Compatibility Modeling.

1. Introduction

In recent years, fashion analysis technology (e.g., compatibility modeling) has attracted widespread attention and demonstrated great commercial value and research potential. The Compatibility Modeling (CM) methods aim to calculate the compatibility score among complementary fashion items (i.e., top, bottom, shoes, and accessories) and thus achieve compatible item recommendations. At present, existing CM methods [1,2] typically utilize multi-modal information (e.g., images and textual descriptions) on fashion items to evaluate the compatibility of complementary items. For example, Yang et al. [3] established an end-to-end learning method, integrating the complementary relationships of specific categories into the item embedding space for unified optimization. Although

DOI: 10.54254/2755-2721/2025.24500

the above method considers item information from different modalities, it ignores fine-grained attribute learning, resulting in a lack of interpretability in recommendations. Later, Lin et al. [4] proposed a scalable method to learn the attention of style-based subspaces to enhance compatibility

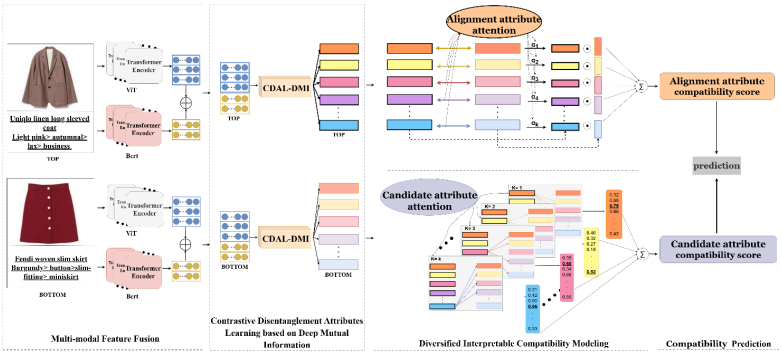

Figure 1: The overall framework of the DICM-MD model

prediction. However, the limited number of subspaces resulted in insufficient interpretability. Due to the remarkable performance of disentanglement representation learning [5-8] in extracting interpretability factors, some researchers have introduced it into interpretable CM. For instance, Liu et al. [9] disentangled the representations of different modalities into k blocks based on covariance and learned the role of each block in the recommendation. In addition, CM methods also integrate multiple techniques, such as graph neural networks [10,11] and attention mechanisms [12,13] to improve compatibility modeling performance.

Although the aforementioned approaches have achieved certain results, the following limitations still persist: (1) The methods described above merely model the compatibility of the aligned attributes between items (e.g., the color and category attributes when matching a white T-shirt with black pants), ignoring the impact of candidate attributes on compatibility modeling. In fact, the top "black and white striped vest" is matched with both the "black wide-leg long skirt" and "black straight-leg long skirt" in terms of aligned attributes, but when further considering their candidate attributes, the "V-neck" and "wide-leg" are more compatible than the "straight-leg". Therefore, it is necessary to consider the candidate attributes to improve the diversity of compatibility modeling. (2) The disentanglement method in the compatibility modeling work mentioned above ignores the possibility of redundant information within the partitioned blocks. In addition, some frameworks [14,15] use specialized region encoders to learn the local attributes of items but overlook the fact that a limited number of attributes cannot meet the diverse matching requirements among items. Especially when dealing with high-dimensional continuous representations of items, these methods are difficult to adjust according to the complexity of the data. Therefore, it is essential to design a disentanglement method according to the characteristics of the multi-modal representation of fashion items to enhance the interpretability of recommendations and meet the diverse attribute matching methods of the items.

Accordingly, to address the above research limitations, we devise a Diversified Interpretable Compatibility Modeling based on Multi-modal Disentanglement framework, termed DICM-MD. Figure 1 illustrates our proposed method, which involves multi-modal feature fusion, contrastive disentanglement attributes learning based on deep mutual information, diversified interpretable compatibility modeling, and a compatibility prediction module. Specifically, DICM-MD adopts deep neural networks to evaluate the mutual information between dimensions when facing high-dimensional continuous representations of items and gradually optimizes its independence based on

DOI: 10.54254/2755-2721/2025.24500

contrastive loss functions, encouraging each dimension to learn independent information. This measure is beneficial for disentangling rich attribute representations and improving the interpretability of recommendations. Then, considering the complexity of attribute matching between complementary items, DICM-MD uses the attention mechanism for diversified interpretable compatibility modeling, which comprehensively describes the compatibility degree of complementary items from the perspective of the alignment attribute and candidate attribute.

Our main contributions can be summarized as the following three points:

(1) We propose a multi-modal disentanglement method named Contrastive Disentangled Attribute Learning based on Deep Mutual Information (CDAL-DMI), which is to evaluate the mutual information among dimensions of high-dimensional continuous representations and constrains their independence based on contrastive loss. To the best of our knowledge, we are the first to utilize mutual information to disentangle item attributes in compatibility work.

(2) We design a Diversified Interpretable Compatibility Modeling method (DICM) to comprehensively evaluate the compatibility between complementary items through the diverse compatibility scores from two perspectives: alignment attribute and candidate attribute.

(3) We conducted extensive experiments on the real datasets IQON3000 and Polyvore, and the results validated the superiority of DICM-MD over advanced baselines.

2. Related work

2.1. Fashion compatibility modeling

The prosperity of the fashion field has drawn researchers' attention to fashion analysis, leading to the emergence of a variety of CF frameworks for evaluating the matching degree between fashion items. Previous CF methods [16,17] solely utilized single-modal information (e.g., visual or textual) of fashion items, neglecting the potential of multi-modal information. As a result, most existing works have gradually shifted their focus to multi-modal compatibility modeling (e.g., visual images and textual descriptions). For instance, Li et al. [18] employed RNN networks to predict the next complementary item. Later, Goto et al. [19] designed an autoencoder based on long short-term memory (LSTM) networks to mine style information from items. However, sequential structures are more suitable for predicting the compatibility of multiple clothing items rather than just top-bottom pairs. Given the superior performance of graph neural networks in learning structural features, some researchers have utilized graph neural networks to model item compatibility. For example, Cui et al. [10] proposed node-wise graph neural networks to model various forms of clothing compatibility. Although these methods achieved certain results, their recommendation processes lack interpretability. Therefore, Lin et al. [4] introduced an expandable approach to learn style-based subspace attention for enhancing compatibility prediction. However, the limited number of subspaces leads to insufficient interpretability. In reality, the compatibility between fashion items depends on attribute pairings, yet existing approaches only focus on matching alignment attributes. In contrast, we model the diversified matching patterns of fashion items from both alignment and candidate attribute perspectives. Table 1: Rules to format sections

2.2. Disentanglement representation learning

Disentanglement representation learning aims to learn the latent factors present in the observed data. Due to its robustness and interoperability, some articles introduce disentanglement representation learning to solve complementary term recommendation tasks. For example, Wang et al. [27]

DOI: 10.54254/2755-2721/2025.24500

designed a graph-based disentanglement representation learning scheme that exploits the compatibility between different parts of an item to accomplish complementary recommendations. Similarly, Wang et al. [22] learn disentangled representations between heterogeneous information of items to complete the recommendation task. Despite the results achieved by these approaches, their delineation of item attributes is not comprehensive. Therefore, Liu et al. [9] proposed to use the distance between item attributes as a regularization term to learn the disentanglement attribute representation, but this approach cannot capture the non-linear relationship between attributes, so it cannot divide the attributes effectively.

Inspired by the above work, we propose a comparative disentanglement attributes learning method based on deep mutual information, which employs deep neural networks to learn the mutual information between different dimensions of item representations and then imposes independence constraints on each dimension using contrastive loss. Our approach aims to encourage each dimension in the high-dimensional continuous representation of fashion items to learn independent attribute representations, thereby preparing for subsequent modeling of attribute matching patterns.

3. Methodology

3.1. Problem formulation

We have a set of tops and bottoms denoted as $T = \{T_1, \cdots, T_i, \cdots, T_{N_T}\}$ and $B = \{B_1, \cdots, B_i, \cdots, B_{N_B}\}$, where N_T and N_B represent the total number of tops and bottoms, respectively. Each top T_i and bottom B_j involves multi-modal information (i.e., visual and textual modalities). We use $\mathbf{V}_{it}(\mathbf{V}_{ib}) \in R^{v*d}$ and $\mathbf{C}_{it}(\mathbf{C}_{ib}) \in R^{c*d}$ to represent the visual and textual features of the top T_i (bottom B_j), respectively. In this paper, we propose a diversified interpretable contrastive disentanglement compatibility modeling approach F, which is able to disentangle the multi-modal information of the items to improve the interpretability of the recommendation task. Furthermore, it improves the performance of diversified compatibility modeling $r_{i,j}$ by simultaneously considering alignment attribute compatibility modeling score r_a and candidate attribute compatibility modeling score r_c. Formally, we have:

$$r_{i,j} = F(r_a, r_c \mid \mathbf{V}_{it}, \mathbf{V}_{jb}, \mathbf{C}_{it}, \mathbf{C}_{jb}, \Theta) \tag{1}$$

where Θ denotes the set of model parameters.

3.2. Mutli-modal feature fushion

Currently, pre-trained models have been extensively employed and have achieved remarkable success in the field of computer vision [23,24]. Consequently, we utilize the pre-trained Vision Transformer (ViT) to extract the visual feature $\mathbf{V}_{it}$ ($\mathbf{V}_{jb}$) from the original image of the top T_i (bottom B_j). ViT model divides the image into multiple patches and subsequently transforms them into low-dimensional embedding through linear itemion. After combining the patch embedding with their corresponding positional embedding, we feed them into a 12-layer transformer encoder as a sequence to capture the contextual information. The final visual feature $\mathbf{V}_{it}$ is generated by the Multi-Layer Perceptron (MLP).

To obtain the textual feature $\mathbf{C}_{it}$ ($\mathbf{C}_{jb}$) of the top T_i (bottom B_j), we leverage the Bidirectional Encoder Representations from Transformers (BERT), which exhibits robust

DOI: 10.54254/2755-2721/2025.24500

generalization capabilities. In this paper, we choose the basic version of BERT, comprising 12 Transformers. Bert model decomposes the associated text description into individual words and then converts them into embedding vectors. Subsequently, it is encoded by a set of transformer encoders to derive the textual feature $\mathbf{C}_{it}$.

In fact, different modalities can express different attribute information of fashion items. In order to comprehensively explore the potential of multi-modal in compatibility modeling, we concatenate the visual feature $\mathbf{V}_{it}$ and the textual feature $\mathbf{C}_{it}$ to obtain the multi-modal feature $\mathbf{Z}_t = \{ \mathbf{V}_{it} \| \mathbf{C}_{it} \} = \{ t_i^1, t_i^2, \cdots, t_i^k \}$.

3.3. Contrastive disentanglement attributes learning based on deep mutual information

As a matter of fact, the compatibility between complementary items is significantly influenced by the attributes of the items. Existing multi-modal compatibility modeling methods mainly focus on computing the compatibility directly based on the overall feature within a latent compatibility space (e.g., style space [25]). Nevertheless, they ignore the effect of repetitive and redundant information in multi-modal attributes on compatibility modeling. In order to increase the independence within the multi-modal attributes and improve the interpretability of compatibility modeling, we designed the Contrastive Disentanglement Attributes Learning based on Deep Mutual Information (CDAL-DMI).

Mutual information was first proposed to measure the degree of dependence between variables [26]. Given two variables x and y , the stronger the independence between them, the lower the mutual information, and vice versa. Formally, the mutual information $I(x;y)$ can be expressed as [27]:

$$I\left(x;y\right) = \mathbb{E}_{p(\boldsymbol{x},y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right] \tag{2}$$

where $p(x,y)$ denotes the joint distribution of variables x and y , $p(x)$ and $p(y)$ denote the marginal distributions of the variables, respectively. In this paper, we disentangle the multi-modal feature based on mutual information. Taking $\mathbf{Z}_t$ of the top T_i as an example, we disentangle it as $\tilde{\boldsymbol{Z}}_t = \left\{ \tilde{t}_i^1, \tilde{t}_i^2, \cdots, \tilde{t}_i^k \right\}$. Nevertheless, it is difficult to calculate precisely the value of mutual information between high-dimensional continuous variables output by neural networks [28]. Therefore, we approximate the value of mutual information between attributes by calculating the mathematical expectation of the neural network output values. According to equation (2), the mutual information $I(t_i^m; t_i^n)$ between m -th and n -th dimensional representations of the top T_i can be expressed as:

$$I\left(t_i^m; t_i^n\right) = \mathbb{E}_{p(t_i^m, t_i^n)}\left[\log \frac{p(t_i^m, t_i^n)}{p(t_i^m)p(t_i^n)}\right] \tag{3}$$

In order to further promote the independence among the attributes of fashion items. we construct the following contrastive loss $\mathscr{L}_{mi}$ =:

$$\mathscr{L}_{mi} = \mathbb{E}_{p(t_i^m, t_i^n)} log \frac{p(t_i^m, t_i^n)}{p(t_i^m)p(t_i^n)} - \mathbb{E}_{p(t_i^m, t_i^n)} log \frac{p\left(t_i^m, t_{i'}^n\right)}{p(t_i^m)p\left(t_{i'}^n\right)} \tag{4}$$

DOI: 10.54254/2755-2721/2025.24500

where t_i^n denotes a randomly selected sample of negative attributes from other items. We can obtain the disentangled representation $\widetilde{\boldsymbol{Z}}_b = \left\{ b_j^1, \tilde{b}_j^2, \cdots, \tilde{b}_j^k \right\}$ of the bottom B_j in a similar manner.

3.4. Diversified interpretable compatibility modeling

Based on the disentangled item attributes $\widetilde{\boldsymbol{Z}}_t = \left\{ \tilde{t}_i^1, \tilde{t}_i^2, \cdots, \tilde{t}_i^k \right\}$ and $\widetilde{\boldsymbol{Z}}_b = \left\{ b_j^1, \tilde{b}_j^2, \cdots, \tilde{b}_j^k \right\}$, we further consider the complex compatibility rules and design the following: 1) Alignment attribute compatibility modeling. 2) Candidate attribute compatibility modeling. Learning diverse matching

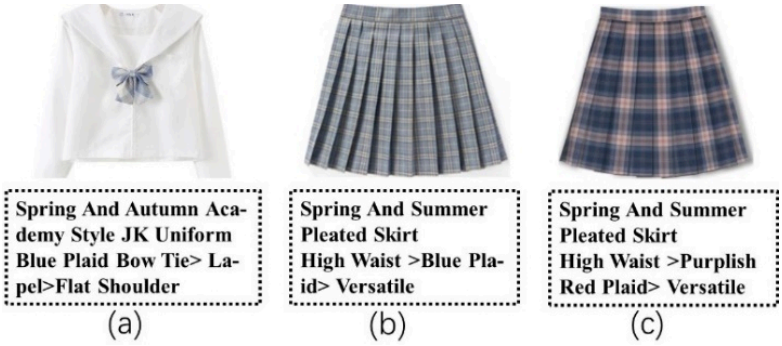

Figure 2: The examples of fashion items alignment attribute matching

methods between item attributes from two perspectives to improve the interpretability of complementary recommendation tasks.

Alignment attribute compatibility modeling: We usually pick complementary items based on alignment attributes in actual clothing matching. For example, both bottoms (b) and (c) in Figure 5 are plaid pleated skirts. Nevertheless, the bow of the top (a) matches the color of the bottom (b), and both share the text description of “blue plaid". Consequently, considering the alignment attribute between complementary items (i.e., color), bottom (b) is more compatible with the given (a) than (c). We first calculate the alignment attribute weight $\widehat{a}_m$ as follows:

$$\begin{cases} a_m = \boldsymbol{W}_v \cdot \sigma\left(\boldsymbol{W}\left[\tilde{t}_i^m \middle\| \tilde{b}_j^m\right] + \boldsymbol{b}\right) \\ \widehat{a}_m = \frac{\exp(a_m)}{\sum_{m=1}^{k} \exp(a_m)} \end{cases} \tag{5}$$

where $\boldsymbol{W}_v$, $\boldsymbol{W}$ and $\boldsymbol{b}$ are learnable network parameters. σ serves as the sigmoid activation function. Then, the alignment attribute compatibility score r_a can be obtained as follows:

$$r_a = \sum_{i,j=1}^{k} \widehat{a}_m \cdot \sigma\left(\tilde{t}_i^m \cdot \tilde{b}_j^m\right) \tag{6}$$

DOI: 10.54254/2755-2721/2025.24500

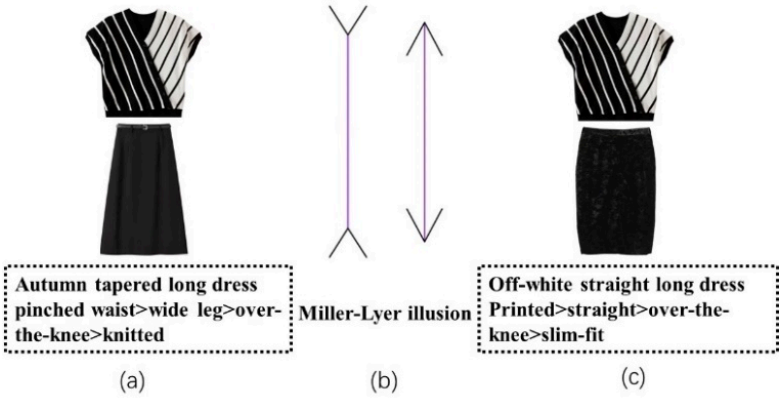

Figure 3: (a) and (c) are examples of matching candidate attribute for fashion items,(b) is a schematic diagram of the Miller-Lyle illusion

Candidate attribute compatibility modeling: As a matter of fact, relying solely on the alignment attribute between complementary items is inadequate for compatibility evaluation. It is imperative to incorporate candidate attributes to improve the performance of compatibility modeling. For example, in Figure 6, the “black and white striped vest" matches well with the two “black long dress" in terms of alignment attributes (i.e., color). However, when further considering the features of the neckline and hem, a more optimal match emerges between the “wide-leg long" dress and the “V-neck" vest. This may be mainly attributable to the Miller-Lyer illusion in Figure 6, which makes the match in Figure 6(a) visually elongate a person's height. Consequently, we introduce the candidate attribute compatibility modeling and calculate the candidate weight $\hat{b}_{m,n}$ as follows:

$$\begin{cases} b_{m,n} = \boldsymbol{W}_v \cdot \sigma\left(\boldsymbol{W}\left[\tilde{t}_i^m \middle\| \tilde{b}_j^n\right] + \boldsymbol{b}\right) \\ \hat{b}_{m,n} = Max_{n \neq m}\left(\frac{exp(b_{m,n})}{\sum_{n-1}^{k} exp(b_{m,n})} \middle| n = 1,2,\ldots,k.\right) \end{cases} \tag{7}$$

Unlike the alignment attribute compatibility modeling, we select the attribute corresponding to the maximum weight $\hat{b}_{m,n}$ as the candidate attribute. Thus, we obtain the candidate attributes compatibility score r_c as follows:

$$r_c = \sum_{i,j=1}^{k} \hat{b}_{m,n} \cdot \sigma\left(\tilde{t}_i^m \cdot \tilde{b}_j^n\right) \tag{8}$$

3.5. Objective loss

Based on the diversified compatibility score r_a and r_c , we can calculate $r_{i,j}$ for the recommendation task as follows:

$$r_{i,j} = (1-\mu) \cdot r_a + \mu \cdot r_c \tag{9}$$

where μ is a balance parameter to control the weight of Alignment attribute compatibility modeling and Candidate attribute compatibility modeling.

DOI: 10.54254/2755-2721/2025.24500

In order to accurately model the compatibility between items, we utilize the Bayesian Personalized Ranking (BPR) framework to construct the following triad $\mathscr{D}_S$ based on the public dataset:

$$\mathscr{D}_S = \{(i, j, k) \mid T_i \in \mathscr{T}, B_j \in \mathscr{B}^+ \wedge B_k \in \mathscr{B} \setminus \mathscr{B}^+\} \tag{10}$$

where the triple $\mathscr{D}_S$ denotes that the bottom $\mathscr{B}^+$ from the set of positive examples B_j is more compatible with the given top T_i than the bottom B_k. From this, we can construct the following loss:

$$\mathscr{L}_{bpr} = \sum_{(i,j,k)\in D_S} - ln(\sigma(r_{ij} - r_{ik})) \tag{11}$$

We jointly optimize the mutual information contrastive loss $\mathscr{L}_{mi}$ to obtain the following final objective loss $\mathscr{L}$:

$$L = \mathscr{L}_{bpr} + \gamma \mathscr{L}_{mi} + \frac{\lambda}{2} \left\| \Theta \right\|_F^2 \tag{12}$$

where γ and λ are non-negative balance parameters.

4. Experiment

To evaluate the proposed DICM-MD model, we conducted a series of experiments on two public fashion datasets to address the following questions:

(1) Does our proposed DICM-MD framework outperform the state-of-the-art baselines?

(2) How does each component affect the proposed DICM-MD framework?

(3) How does the DICM-MD perform in the complementary item retrieval task?

Dataset: To validate the effectiveness of our proposed DICM-MD model, we leverage two public datasets: IQON3000 and Polyvore. The former contains 308,747 outfits, totaling 672,335 fashion items (e.g., tops, bottoms, and shoes). The latter consists of 66,000 outfits, totaling 158,503 fashion items. Each item in both datasets is associated with a visual image and a descriptive title.

Evaluation indicators: We adopted Area Under Curve (AUC) and Mean Reciprocal Ranking (MRR) [29] to measure the effectiveness of the DICM-MD model. Formally, we defined the AUC as follows:

$$AUC = \frac{1}{|S_{test}|} \sum_{(i,j,k)\in S_{test}} \delta(r_{ij} - r_{ik}) \tag{13}$$

where $\delta(\cdot)$ returns 1 when the parameter is greater than 0 and returns 0 otherwise. S_{test} denotes the test set. The MRR can be defined as follows:

$$MRR = \frac{1}{|S_{test}|} \sum_{j=0}^{|S_{test}|} \frac{1}{R_j^{(i)}} \tag{14}$$

where $R_j^{(i)}$ represents the sorting position of a positive bottom B_j based on r_{ij}.

DOI: 10.54254/2755-2721/2025.24500

Experiment Settings: We utilized Adam [30] to optimize our proposed DICM-MD model. Specifically, the batch size and learning rate are searched from [32,64,128,512,1024] and [0.0001,0.0005,0.005,0.001], respectively. The Loss $\mathscr{L}$ and AUC in Figure 2 (a) stabilize gradually after notable fluctuations, which indicates the excellent convergence of DICM-MD. Figure 2 (b) illustrates the variation of AUC with the hyperparameter μ. As depicted in this figure, the DICM-MD achieves optimal performance when μ is set to 0.4, highlighting the effectiveness of both alignment attribute compatibility and candidate attribute compatibility in our model.

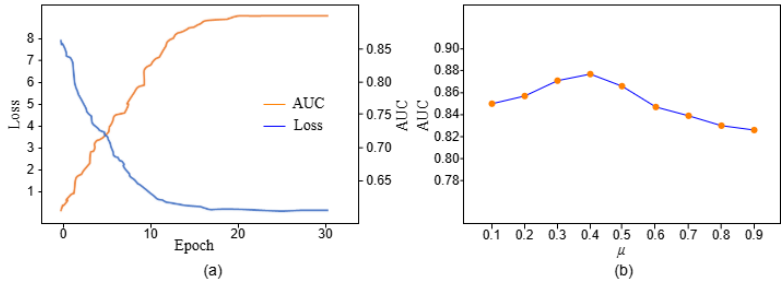

Figure 4: The training loss and AUC change curves with iteration epochs; (b) The variation curve of AUC with different μ values

4.1. Model comparison(RQ1)

In order to scientifically evaluate our DICM-MD model, we compared it with the following baselines:

(1) POP: The compatibility is determined by the number of tops in the training set that are matched with the bottom B_j.

(2) RAND: We randomly assign the score r_{ij} to measure the compatibility between complementary items.

(3) Bi-LSTM [31] models the given items as a sequence to predict the next compatible item.

(4) V-BPR [17] models the compatibility between items solely based on their visual features.

(5) AHGN [32] constructs a heterogeneous graph utilizing the visual attributes of items, which learns the interaction information between items.

(6) VT-BPR: We extended the V-BPR to evaluate the compatibility with multi-modal data.

(7) MNLFF [33] learns hierarchical multi-modal representations and assesses the compatibility between complementary items based on the scores between them.

(8) PAI-BPR [34] learns the attribute information of the item via the visual modality and subsequently integrates the corresponding textual modality for compatibility modeling.

(9) MDR [15] disentangles multi-modal features in a partially supervised manner to achieve interpretable recommendations.

(10) FCM-CMAN [12] learns its disentangled representation hierarchically over the multi-modal information of the item.

(11) LGMRec [11] focuses on learning the disentangled information of items based on the global and local graphs.

Table 1 shows the performance comparison of different baselines, we can observe the following: (1) Compared with pop and rand, Bi-LSTM performs better, which indicates the necessity of item interactions in compatibility modeling. (2) V-BPR and AHGN outperform Bi-LSTM, suggesting that content-based approaches provide more item information than sequence-based approaches for compatibility modeling. (3) The methods (V-BPR and AHGN) that rely solely on the visual modality

DOI: 10.54254/2755-2721/2025.24500

exhibit lower performance than other multi-modal methods, primarily due to the additional attribute features provided by the textual information. (4) FCM-CMAN, LGMRec, and MDR outperform PAI-BPR and MNLFF, demonstrating that disentanglement improves the compatibility modeling of complementary items. (5) The DICM-MD results proves the necessity of incorporating diversified interpretable compatibility modeling in complementary recommendation tasks.

Table 1: Performance of different methods based on AUC and MRR evaluation indicators

Methods	IQON3000		Polyvore	
	AUC	MRR	AUC	MRR
POP	0.6039	0.2049	0.6072	0.2087
RAND	0.5003	0.2137	0.5059	0.2162
Bi-LSTM	0.6746	0.3354	0.6693	0.3423
V-BPR	0.7863	0.6508	0.7881	0.6791
AHGN	0.7955	0.6972	0.8033	0.7011
VT-BPR	0.8205	0.8097	0.8297	0.8118
MNLFF	0.8397	0.8144	0.8451	0.8255
PAI-BPR	0.8554	0.8174	0.8537	0.8267
MDR	0.8597	0.8204	0.8692	0.8388
FCM-CMAN	0.8643	0.8304	0.8722	0.8491
LGMREC	0.8689	0.8361	0.8871	0.8593
DICM-MD	0.8848	0.8463	0.8961	0.8677

Table 2: The experimental results comparison under different Top-K values in terms of NDCG@K and HR@K

Datesets	Evaluation	IBR	VBPR	HFGN	VTBPR	PAI-BPR	DICM-MD
QON3000	NDCG@10	0.1628	0.2366	0.2393	0.2467	0.2567	0.3317
	NDCG@15	0.2076	0.2790	0.2841	0.2874	0.2845	0.3722
	NDCG@20	0.2489	0.3128	0.3172	0.3207	0.3386	0.4000
	HR@10	0.3463	0.4767	0.4815	0.4967	0.5087	0.6080
	HR@15	0.5194	0.6545	0.6588	0.6707	0.6802	0.7614
	HR@20	0.6870	0.7973	0.7990	0.8100	0.8125	0.8790
Polyvore	NDCG@10	0.4365	0.5057	0.5373	0.5467	0.5532	0.6470
	NDCG@15	0.4519	0.5217	0.5460	0.5612	0.5715	0.6678
	NDCG@20	0.4764	0.5391	0.5670	0.5931	0.6133	0.6788
	HR@10	0.6276	0.7307	0.7539	0.7678	0.7703	0.8338
	HR@15	0.7739	0.7946	0.8082	0.8109	0.8145	0.9090
	HR@20	0.8183	0.8371	0.8559	0.8712	0.8707	0.9557

To further validate the effectiveness of our DICM-MD model, we employed two ranking metrics widely used in recommender systems: Hit Rate (HR), and Normalized Discounted Cumulative Gain (NDCG), Specifically, HR evaluates the compatibility of the top K items recommended by our model with the given item, and NDCG considers the position of the recommended item in the

DOI: 10.54254/2755-2721/2025.24500

ranking list. Table 2 presents the recommendation results of different models in terms of the HR@K and NDCG@K. From this table, we can observe that: (1) VT-BPR performs poorly, suggesting that the complementary recommendation task cannot be performed without mining deeper representations in multi-modal information. (2) FCM-CMAN, LGMRec, and MDR perform better than MNLFF and PAI-BPR, which further demonstrates the advantages of disentanglement representation learning for complementary recommendation tasks. (3) Our proposed DICM-MD model performs optimally. This strongly indicates that the joint use of contrastive disentanglement attributes learning and diversified interpretable compatibility modeling is highly effective in complementary item recommendation tasks.

4.2. Ablation study(RQ2)

To verify the effectiveness of each component of our DICM-MD model, we conducted ablation studies on the following derivatives:

(1) w/o-Disen: We removed the contrastive disentanglement module to verify the compatibility performance.

(2) w/o-Candidate: In order to verify the effect of the diversified interpretable compatibility modeling, we removed the candidate attribute compatibility modeling module.

(3) w/o-Alignment: Similar to (2), we removed the alignment attribute compatibility modeling module.

(4) w/o-V: To investigate the role of different modalities in DICM-MD, we removed the visual feature from the model.

(5) w/o-T: Similar to (4), we removed the textual feature from the DICM-MD.

Table 3: The performance comparison of DICM-MD and the variant models

Methods	IQON3000		Polyvore	
	AUC	MRR	AUC	MRR
w/o-Disen	0.8513	0.8362	0.8683	0.8458
w/o-Candidate	0.7957	0.8015	0.8147	0.8212
w/o-Alignment	0.7243	0.7634	0.7464	0.7820
w/o-V	0.8041	0.7901	0.8230	0.8122
w/o-T	0.7016	0.7282	0.7346	0.7347
DICM-MD	0.8848	0.8463	0.8961	0.8677

Table 3 shows the results of the ablation experiments on AUC and MRR. DICM-MD outperforms all variants, validating the efficacy of the model components. Specifically: (1) w/o-Disen did not perform as well as the full model, proving that our proposed learning module for contrastive disentanglement attributes learning based on deep mutual information is effective. (2) Based on AUC metrics, DICM-MD improved by 9.9\% over w/o-Candidate, and 20.7\% over w/o-Alignment. It can be seen that the diversified interpretable compatibility modeling module proposed in this paper is better able to address the complementary recommendation task. The performance of w/o-Candidate illustrates that alignment attribute compatibility modeling plays a dominant role in this module. (3) Similarly, DICM-MD was 8.8\% and 24.7\% higher compared to w/o-V and w/o-T, respectively. According to the results, both textual and visual modalities contain important attribute information. In addition, w/o-V is superior to w/o-T, which may be because the textual features contain more attribute information.

DOI: 10.54254/2755-2721/2025.24500

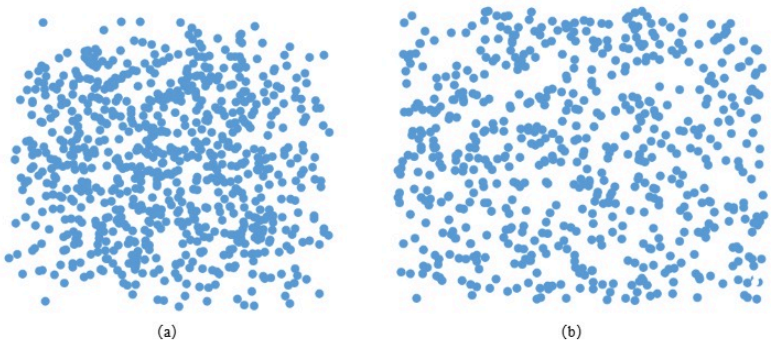

Figure 5: Distribution of attributes of fashion items before and after disentanglement

In order to intuitively demonstrate the effect of the contrastive disentanglement attributes learning method, we conducted visualization experiments. As shown in Figure 3, we randomly selected an item and displayed the corresponding attribute distribution before and after disentanglement in the same space with the t-SNE [35] method. As shown in Figure 7(b), the attribute distributions are more spread out, highlighting that contrasting disentanglement attributes learning methods can improve attribute independence and enhance the interpretability of complementary recommendation tasks.

4.3. On complementary item retrieval (RQ3)

To validate the practicality of our work, we scientifically evaluated the performance of DICM-MD and w/o-Candidate on the complementary item retrieval task. We considered the given top as a "query" and randomly selected N bottoms as a candidate set, which contains one positive example and N-1 negative examples. A ranked list is generated by computing the compatibility score between the given top and bottom in the candidate set. In Figure 4, we showed the results of the ranking list of DICM-MD and the variant w/o-Candidate for 10 candidate bottoms based on the given query. The results show that the DICM-MD model proposed in this paper put the positive examples at the top compared to w/o-Candidate, which further proves that the candidate attribute compatibility modeling designed by us is effective. Therefore, the Diversified Interpretable Compatibility Modeling is necessary in the task of Complementary Item Retrieval.

Figure 6: Complementary fashion item retrieval experiment results, where the red box represents the positive example

DOI: 10.54254/2755-2721/2025.24500

5. Conclusion and future work

In this paper, we present the DICM-MD model, which disentangles the multi-modal information of fashion items and uses the diversified interpretable compatibility modeling module to explore the interaction information of item attributes. Specifically, we propose a comparative disentanglement attribute learning method based on deep mutual information, which uses a neural network to fit the value of mutual information between attributes and designs a contrast loss to optimize this value and gradually improve the independence between attributes. To fully explore the attribute compatibility rules, we also designed aligned attribute compatibility modeling and candidate attribute compatibility modeling to learn complex attribute interactions from two perspectives.

References

[1] Xie, Y., Lin, B., Qu, Y., Li, C., Zhang, W., Ma, L., ... & Tao, D. (2020). Joint deep multi-view learning for image clustering. IEEE Transactions on Knowledge and Data Engineering, 33(11), 3594-3606.
[2] Wang, Y., Liu, L., Fu, X., & Liu, L. (2024). MCCP: multi-modal fashion compatibility and conditional preference model for personalized clothing recommendation. Multimedia Tools and Applications, 83(4), 9621-9645.
[3] Yang, X., Ma, Y., Liao, L., Wang, M., & Chua, T. S. (2019, July). Transnfcm: Translation-based neural fashion compatibility modeling. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 403-410).
[4] Lin, Y. L., Tran, S., & Davis, L. S. (2020). Fashion outfit complementary item retrieval. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3311-3319).
[5] De Divitiis, L., Becattini, F., Baecchi, C., & Del Bimbo, A. (2023). Disentangling features for fashion recommendation. ACM Transactions on Multimedia Computing, Communications and Applications, 19(1s), 1-21.
[6] Guan, W., Wen, H., Song, X., Wang, C., Yeh, C. H., Chang, X., & Nie, L. (2022). Partially supervised compatibility modeling. IEEE Transactions on Image Processing, 31, 4733-4745.
[7] Cui, Z., Yu, F., Wu, S., Liu, Q., & Wang, L. (2021). Disentangled item representation for recommender systems. ACM Transactions on Intelligent Systems and Technology (TIST), 12(2), 1–20.
[8] Ma, J., Cui, P., Kuang, K., Wang, X., & Zhu, W. (2019). Disentangled graph convolutional networks. In International conference on machine learning (pp. 4212–4221). PMLR.
[9] Liu, F., Chen, H., Cheng, Z., Liu, A., Nie, L., & Kankanhalli, M. (2022). Disentangled multimodal representation learning for recommendation. IEEE Transactions on Multimedia, 25, 7149–7159.
[10] Cui, Z., Li, Z., Wu, S., Zhang, X.-Y., & Wang, L. (2019). Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks. In The World Wide Web Conference (pp. 307–317).
[11] Guo, Z., Li, J., Li, G., Wang, C., Shi, S., & Ruan, B. (2024). Lgmrec: Local and global graph learning for multimodal recommendation. Proceedings of the AAAI Conference on Artificial Intelligence, 38(8), 8454–8462.
[12] Jing, P., Cui, K., Guan, W., Nie, L., & Su, Y. (2023). Category-aware multimodal attention network for fashion compatibility modeling. IEEE Transactions on Multimedia, 25, 9120–9131.
[13] Shimizu, R., Wang, Y., Kimura, M., Hirakawa, Y., Wada, T., Saito, Y., & McAuley, J. (2024). A fashion item recommendation model in hyperbolic space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8377–8383).
[14] Yang, X., Song, X., Feng, F., Wen, H., Duan, L.-Y., & Nie, L. (2021). Attribute-wise explainable fashion compatibility modeling. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 17(1), 1–21.
[15] Wang, X., Chen, H., & Zhu, W. (2021). Multimodal disentangled representation for recommendation. In 2021 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1–6). IEEE.
[16] Tangseng, P., Yamaguchi, K., & Okatani, T. (2017). Recommending outfits from personal closet. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 2275–2279).
[17] He, R., & McAuley, J. (2016). Vbpr: visual bayesian personalized ranking from implicit feedback. Proceedings of the AAAI conference on artificial intelligence, 30(1).
[18] Li, Y., Cao, L., Zhu, J., & Luo, J. (2017). Mining fashion outfit composition using an end-to-end deep learning approach on set data. IEEE Transactions on Multimedia, 19(8), 1946–1955.
[19] Nakamura, T., & Goto, R. (2018). Outfit generation and style extraction via bidirectional lstm and autoencoder. arXiv preprint arXiv: 1807.03133.

DOI: 10.54254/2755-2721/2025.24500

[20] Gao, R., Tao, Y., Yu, Y., Wu, J., Shao, X., Li, J., & Ye, Z. (2023). Self-supervised dual hypergraph learning with intent disentanglement for session-based recommendation. Knowledge-Based Systems, 270, 110528.
[21] Wang, X., Jin, H., Zhang, A., He, X., Xu, T., & Chua, T.-S. (2020). Disentangled graph collaborative filtering. In Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval (pp. 1001–1010).
[22] Wang, Y., Tang, S., Lei, Y., Song, W., Wang, S., & Zhang, M. (2020). Disenhan: Disentangled heterogeneous graph attention network for recommendation. In Proceedings of the 29th ACM international conference on information & knowledge management (pp. 1605–1614).
[23] Yang, F., Chen, C., Wang, Z., Chen, H., Liu, Y., Li, G., & Wu, X. (2023). Vit-based terrain recognition system for wearable soft exosuit. Biomimetic Intelligence and Robotics, 3(1), 100087.
[24] Bana, T., Loya, J., & Kulkarni, S. (2021). Vit-inception-gan for image colourising. arXiv preprint arXiv: 2106.06321.
[25] Yang, X., Ma, Y., Liao, L., Wang, M., & Chua, T.-S. (2019). Transnfcm: Translation-based neural fashion compatibility modeling. Proceedings of the AAAI conference on artificial intelligence, 33(01), 403–410.
[26] Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., & Hjelm, D. (2018). Mutual information neural estimation. In International conference on machine learning (pp. 531–540). PMLR.
[27] Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., & Carin, L. (2020). Club: A contrastive log-ratio upper bound of mutual information. In International conference on machine learning (pp. 1779–1788). PMLR.
[28] Paninski, L. (2003). Estimation of entropy and mutual information. Neural computation, 15(6), 1191–1253.
[29] Cremonesi, P., Koren, Y., & Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. In Proceedings of the fourth ACM conference on Recommender systems (pp. 39–46).
[30] Zhao, W. X., Hou, Y., Pan, X., Yang, n., Che, u., Zhang, Zey, Lin, Z., Zhang, J., Bian, S., Tang, i., Jiaka, W., Sun, et al. (2022). Recbole 2.0: towards a more up-to-date recommendation library. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (pp. 4722–4726).
[31] Han, X., Wu, Z., Jiang, Y.-G., & Davis, L. S. (2017). Learning fashion compatibility with bidirectional lstms. In Proceedings of the 25th ACM international conference on Multimedia (pp. 1078–1086).
[32] Zhou, Z., Su, Z., & Wang, R. (2022). Attribute-aware heterogeneous graph network for fashion compatibility prediction. Neurocomputing, 495, 62–74.
[33] Lu, S., Zhu, X., Wu, Y., Wan, X., & Gao, F. (2021). Outfit compatibility prediction with multi-layered feature fusion network. Pattern Recognition Letters, 147, 150–156.
[34] Sagar, D., Garg, J., Kansal, P., Bhalla, S., Shah, R. R., & Yu, Y. (2020). Pai-bpr: Personalized outfit recommendation scheme with attribute-wise interpretability. In 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM) (pp. 221–230). IEEE.
[35] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. Journal of machine learning research, 9(11).