Multimodal Counterfactual Learning Network for Multimedia-based Recommendation

Shuaiyang Li
Hefei University of Technology
Hefei, China
shuaiyangli@mail.hfut.edu.cn

Dan Guo
Hefei University of Technology
Hefei, China
guodan@hfut.edu.cn

Kang Liu
Hefei University of Technology
Hefei, China
kangliu1225@gmail.com

Richang Hong
Hefei University of Technology
Hefei, China
hongrc@hfut.edu.cn

Feng Xue*
Hefei University of Technology
Institute of Artificial Intelligence,
Hefei Comprehensive National
Science Center
Hefei, China
feng.xue@hfut.edu.cn

ABSTRACT

Multimedia-based recommendation (MMRec) utilizes multimodal content (images, textual descriptions, *etc.*) as auxiliary information on historical interactions to determine user preferences. Most MMRec approaches predict user interests by exploiting a large amount of multimodal contents of user-interacted items, ignoring the potential effect of multimodal content of user-uninteracted items. As a matter of fact, there is a small portion of user preference-irrelevant features in the multimodal content of user-interacted items, which may be a kind of spurious correlation with user preferences, thereby degrading the recommendation performance. In this work, we argue that the multimodal content of user-uninteracted items can be further exploited to identify and eliminate the user preference-irrelevant portion inside user-interacted multimodal content, for example by counterfactual inference of causal theory. Going beyond multimodal user preference modeling only using interacted items, we propose a novel model called Multimodal Counterfactual Learning Network (MCLN), in which user-uninteracted items' multimodal content is additionally exploited to further purify the representation of user preference-relevant multimodal content that better matches the user's interests, yielding state-of-the-art performance. Extensive experiments are conducted to validate the effectiveness and rationality of MCLN. We release the complete codes of MCLN at https://github.com/hfutmars/MCLN.

CCS CONCEPTS

• **Information systems** → **Personalization**; **Recommender systems**.

*Corresponding author: Feng Xue.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591739

KEYWORDS

Recommender Systems, Multimodal User Preference, Counterfactual Learning, Spurious Correlation

ACM Reference Format:
Shuaiyang Li, Dan Guo, Kang Liu, Richang Hong, and Feng Xue. 2023. Multimodal Counterfactual Learning Network for Multimedia-based Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3539618.3591739

1 INTRODUCTION

Multimedia-based recommendation (MMRec) has been extensively deployed in various online services (*e.g.*, e-commerce [21, 24], social media [49], micro-video-sharing platforms [3, 45], online video platforms [34]) to alleviate information overload. Rich multimedia content (visual images, textual descriptions, *etc.*) contributes extra multimodal information for historical interactions, thereby strengthening the quality of representation learning. These multimodal features of items can represent the correlation among items in multiple dimensions. They provide multimodal fine-grained preferences for users to more comprehensively describe their interests. As a kind of content-rich method [23, 37, 47], MMRec approaches alleviate data sparsity and cold-start problems that are ubiquitous in recommender systems.

Current MMRec frameworks typically consist of two components: multimodal feature extraction and user preference modeling. The former extracts multimodal features from the multimedia content of items by utilizing pre-trained deep networks. The latter incorporates these multimodal features into the recommendation framework (*e.g.*, collaborative filtering [32]) to accomplish user preference modeling together with historical interaction. Early MMRec works like VBPR [11] employ pre-trained convolutional neural networks to extract visual features of the items and integrate them into the item embeddings. Subsequent efforts such as CKE [51] incorporate multimodal information (visual images and textual descriptions) and knowledge graphs into the representation learning process. MMGCN [46] uses Graph Convolution Networks (GCNs) to perform embedding propagation on parallel user-item

interaction graphs with different modality data, thus distilling user preference cues hidden in different modalities. Recent efforts like DMRL [19] incorporate a disentangled representation technique into MMRec, considering the contributions of different modality features on each disentangled factor in user preference modeling.

Despite existing MMRec models using a large number of multimodal contents of user-interacted items to predict user interest and having achieved remarkable performance, they ignore the potential effect of the multimodal content of user-uninteracted items on representation learning. As a matter of fact, even for a user-interacted item, there is still a small portion of multimodal content that the user dislikes or less likes, which we call user preference-irrelevant multimodal features. Also, we call these multimodal contents that really attract users as preference-relevant multimodal features. Figure 1 shows the motivation of our MCLN, in which the user-uninteracted item's multimodal content together with the user-interacted item is leveraged to purify the representation of user preference-relevant multimodal features. In Figure 1, Item i, which has been interacted with by user, is a cotton dress, and the user-uninteracted item i^* is a silk dress. They are closely similar in style (*i.e.*, dress, female model, and bag) except that their fabrics are different, the former being cotton and the latter being silk. In addition, there are other two items (Item 1 and Item 2) to be recommended to the given user. Item 1 is a yellow cotton dress that the user really likes (probably because its fabric is cotton), while Item 2 is a wathet chiffon dress and is more similar in appearance and style to the user-interacted item i, which may be less liked by the user. Thus, traditional models find it difficult to distinguish between the items, and scores for both are close, even the score of Item 2 is slightly higher than Item 1 (wrong prediction), as shown in the dashed box (a).

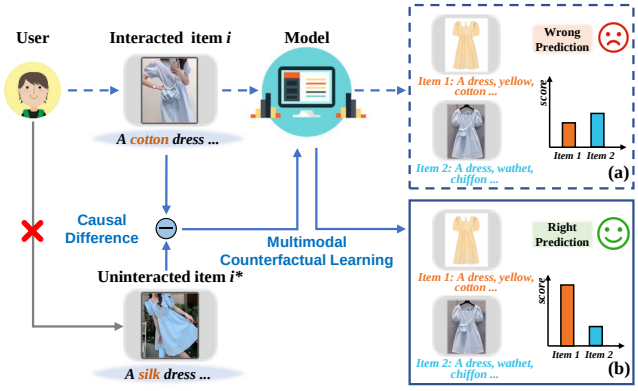

Figure 1: A motivating example of MCLN. Item 1 is the item that the user really likes. The dashed box (a) indicates that the traditional models produce a wrong prediction for the user, wherein the score of Item 1 is lower than Item 2. The solid box (b) shows how the model with multimodal counterfactual learning produces a right prediction for the user, wherein the score of Item 1 is higher than Item 2.

Subsequently, we perform a causal difference calculation using a user-interacted item i (cotton dress) and a user-uninteracted item i^* (silk dress) to distinguish the difference in the multimodal content of the recommended items (Item 1 and Item 2), such that the model can capture the difference between the words "cotton" and "chiffon" as well as the visual difference (chiffon has a pattern and cotton is a flatter fabric), thereby discovering potential user preferences (the user prefers cotton items). Thus, the model with multimodal counterfactual learning can distinguish Item 1 from Item 2, and the scores of both differ significantly: Item 1 has a higher score than Item 2 (right prediction), as shown in the solid box (b). Therefore, we argue that the multimodal content of user-uninteracted items can be employed to alleviate the side effect of preference-irrelevant parts in the multimodal content of user-interacted items.

Toward this end, we develop a novel MMRec model named **Multimodal Counterfactual Learning Network (MCLN)**, in which a *counterfactual learning* module is designed to learn the *causal difference of user preference distribution* (*abbr.* as *causal difference*) on multimodal contents of user-interacted and user-uninteracted items. This causal difference value between two preference distributions leads the MCLN model to exploit the multimodal content representations of user-uninteracted items to identify and eliminate the preference-irrelevant representations in the multimodal content of user-interacted items, which may have a spurious correlation with user preferences. This can further purify the representation of user preference-relevant multimodal content that better matches the user's interests, thus achieving satisfactory recommendation quality. Experiments performed on three public datasets demonstrate that MCLN yields state-of-the-art performance. Further ablation studies validate the effectiveness of each component in MCLN.

To summarize, the main contributions of this work are threefold:

- We highlight that the multimodal contents of user-uninteracted items are helpful for identifying and removing the preference-irrelevant part of the multimodal content of user-interacted items, further purifying the representation of user preference on multimodal content.
- Inspired by causal theory, we devise a novel MMRec model, MCLN, which leverages the difference values between the preference distributions on multimodal content of user-interacted items and user-uninteracted items under the guidance of causal difference to mitigate the side effect of preference-irrelevant parts in the multimodal content of user-interacted items.
- Through extensive experiments on three public datasets, we demonstrate the effectiveness and rationality of MCLN.

2 RELATED WORK

GCN-based Recommendation. In recent years, GCN [16, 50] has received increasing attention from many research fields [9, 39] and has achieved remarkable success. A common paradigm for GCN is to first adopt a graph convolution layer to aggregate information from neighbors and then iteratively perform this process to capture high-order collaborative signals. Early recommendation efforts such as GC-MC [2] utilize nonlinear GCN to aggregate one-order neighbor information into the embeddings of the target nodes. PinSage [50] exploits deeper GCN structures to obtain high-order interactions. NGCF [43] effectively combines GCN and Matrix Factorization (MF) [18] and obtains the final embedding of the nodes by concatenating the embedding learned from all graph convolution layers. Several recent works [4, 12] have contended that weight transformation matrices and nonlinear activation functions in GCN introduce more redundant parameters, which can lead to overfitting. LightGCN [12] captures high-order collaborative signals using

a concise linear GCN that removes weight transformation matrices and nonlinear activation functions. JMPGCF [22] integrates multi-grained popularity features into embedding generation via constructing a popularity-aware graph Laplacian norm for modeling user sensitivity to popularity.

Multimedia-based Recommendation. Multimedia-based recommendation (MMRec) can be regarded as a content-rich recommendation method. It is centered on utilizing the multimodal content of items to assist with the recommendation task. Early works like VBPR [11] use a convolution neural network pre-trained on ImageNet [5] to extract deep features of items and integrate them into the embedding of items. Subsequently, some researchers began to explore user preferences across modalities. For instance, CKE [51] incorporates visual features, textual features, and knowledge graphs into the learning process of embeddings. MMGCN [46] is a MMRec work that uses GCN to perform embedding propagation on interaction graphs with different modality data, thereby capturing user preferences for different modalities. MGAT [36] weights information propagation in multimodal interaction graphs by constructing attention networks [10, 38], thus modeling fine-grained multimodal user preferences. HUIGN [45] investigates user intent learning in MMRec and learns multi-level user intents from the co-interacted patterns of items to optimize user and item representations. The latest efforts like DMRL [19] incorporate a disentangled representation technique into MMRec that considers the contributions of different modality features on each disentangled factor in user preference modeling. However, these methods depend on the multimodal content of user-interacted items, being unaware of the potential usage of the multimodal content of user-uninteracted items. Thus, we develop MCLN to explore an efficient approach to utilize the potential effect of the multimodal content of user-uninteracted items on representation learning.

Causality-based Recommendation. Recently, causal inference has been extensively employed in many machine-learning tasks, spanning from computer vision [28, 40], and natural language processing [7] to recommendation systems [29, 52]. In recommendation systems, several related efforts focus on eliminating popularity bias. For example, MACR [44] utilizes causal graphs to analyze the causal effect of item popularity. PD [52] analyzes causal relations from the confounders' view and alleviates popularity bias during model training via causal intervention. In addition, DecRS [41] explains the cause of bias amplification from the causal perspective and employs an approximation of "backdoor adjustment" to mitigate the bias amplification problem. CausalRec [29] designs a causal graph to analyze the visual bias problem and eliminates the bad effect of visual feature via counterfactual inference. CR [42] studies the clickbait issue through a causal graph, where the exposure features of items are the source of bias. Recent works like InvRL [6] attempt to alleviate the spurious correlations from the multimedia content. It applies heterogeneous environments to denote spurious correlations and learns invariant representations across the environments to mitigate their effect. Distinct from these approaches, our MCLN considers the potential effect of the multimodal content of user-uninteracted items on representation learning when employing intervention and counterfactual inference. This can alleviate the side effect of spurious correlations (user preference-irrelevant parts) in the multimodal content representations of user-interacted items.

3 METHODOLOGY

In this section, we detail the overall framework of the proposed MCLN, as illustrated in Figure 2, which can be divided into five main parts: (1) basic recommendation framework, which incorporates the high-order collaborative signals of each node into its embedding representation; (2) sample pair selection and feature embedding, which select sample pairs (user-interacted and user-uninteracted items) from multimedia datasets and encode the feature embeddings of the sample pairs; (3) causal difference learning, which learns the preference distributions on multimodal content of the user-interacted and user-uninteracted items and uses the causal differences between them to purify the representations of user preference-relevant multimodal content; (4) multimodal fusion and score prediction, which fuse intra- and inter-modality feature embeddings of user-interacted items and calculate prediction scores for all user-item pairs; and (5) model optimization, which constructs objective losses to optimize the representation of all users and items.

3.1 Basic Recommendation Framework

First, we convert the implicit interaction data into a user-item interaction graph $\mathcal{G} = \{(u, r_{ui}, i)|u \in \mathcal{U}, i \in \mathcal{I}\}$, where $\mathcal{U}$ and $\mathcal{I}$ represent the sets of users and items, respectively, and $r_{ui} = 1$ indicates that there exists interaction between user u and item i; otherwise, $r_{ui} = 0$. We initialize all nodes in $\mathcal{G}$ by mapping the IDs of all nodes to the dense vector representations as follows:

$$\mathcal{E} = \left\{e_{u_1}^{(0)}, \ldots, e_{u_{|\mathcal{U}|}}^{(0)}, e_{i_1}^{(0)}, \ldots, e_{i_{|\mathcal{I}|}}^{(0)}\right\}, \tag{1}$$

where $\mathcal{E} \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{I}|)\times d}$, $|\mathcal{U}|$ is the number of users, $|\mathcal{I}|$ is the number of items, and d is the embedding dimension.

Then, based on prior work [4, 12], we devise a linear GCN to perform embedding propagation on the interaction graph $\mathcal{G}$. We assume that the current graph convolution layer is l. The embedding updates of the user u_1 and item i_1 are obtained by aggregating the embeddings of their neighbor nodes in the $(l-1)$-th layer as follows:

$$e_{u_1}^{(l)} = \sum_{i \in N_{u_1} \cup u_1} \frac{1}{\left|N_{u_1}\right|^{0.5} |N_i|^{0.5-\alpha}} \cdot e_i^{(l-1)}, \tag{2}$$

$$e_{i_1}^{(l)} = \sum_{u \in N_{i_1} \cup i_1} \frac{1}{\left|N_{i_1}\right|^{0.5} |N_u|^{0.5-\alpha}} \cdot e_u^{(l-1)}, \tag{3}$$

where N_u and N_i stand for the neighbor nodes of user u and item i in $\mathcal{G}$, $|N_u|$ and $|N_i|$ are the number of nodes in N_u and N_i, and $e_u^{(0)}$ and $e_i^{(0)}$ are initialized in Equation 1. Furthermore, referring to the recent graph learning-based method [22], we adjust the standard graph Laplacian norm to the popularity-aware graph Laplacian norm: $\frac{1}{\left|N_{u_1}\right|^{0.5}|N_i|^{0.5-\alpha}}$. This not only avoids the scale of embeddings increasing with the graph convolution operation, but also incorporates the user's sensitivity to popularity into the embedding generation, and α is a hyper-parameter.

In addition, the embeddings generated via each graph convolution layer contain peculiar semantic information. Thus, after the L layer, we combine (*i.e.*, weighted sum) the embeddings learned at each layer as the final embedding representation of the nodes:

$$e_{u_1}^{base} = \frac{1}{L+1} \sum_{l=0}^{L} e_{u_1}^{(l)}. \tag{4}$$

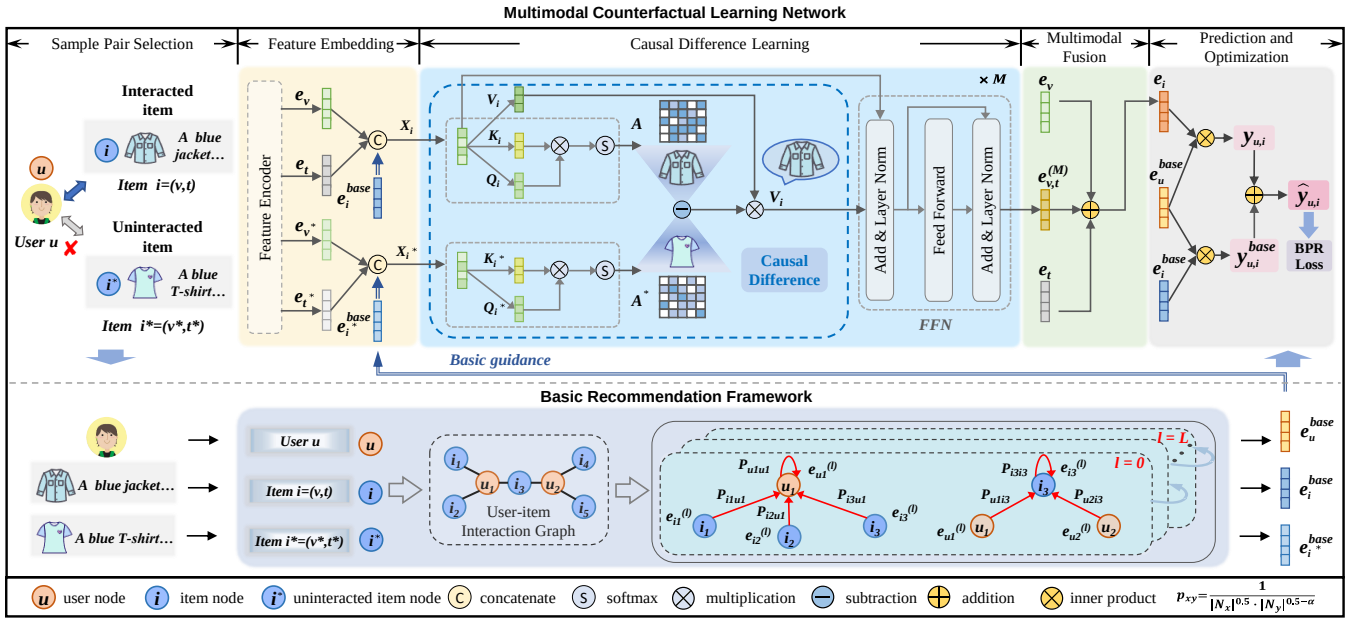

Figure 2: Illustration of the proposed MCLN. The feature encoder denotes pre-trained networks for visual and textual feature extraction (see Subsection 3.2), L is the max number of graph convolution layers, and M is the max number of multimodal counterfactual learning layers.

$$e_{i_1}^{base} = \frac{1}{L+1}\sum_{l=0}^{L} e_{i_1}^{(l)}. \tag{5}$$

3.2 Sample Pair Selection and Feature Embedding

In this subsection, we first select a set of sample pairs (user-interacted and user-uninteracted samples) from multimedia datasets for subsequent multimodal preference modeling. The user-interacted samples are items with which the user has historically interacted. There is a small portion of user preference-irrelevant representations in the multimodal content of user-interacted items, which may have a spurious correlation with user preferences. The user-uninteracted samples are selected from the set of items with which the user has not interacted. The multimodal content of user-uninteracted items can identify and remove preference-irrelevant representations in the multimodal content of user-interacted items. We exploit such samples to purify the representation of user preference-relevant multimodal content. Furthermore, the selection strategy of the user-uninteracted sample is random in this work.

Then, we construct feature encoders to extract the visual and textual feature embeddings of the sample pairs (i and i^*), where the feature encoders represent pre-trained deep networks (*e.g.*, VGG16 [33], Sentence2Vec [1]). Specifically, we feed the visual image of sample i into the pre-trained network to obtain deep features f_v, where the f_v dimension is 4,096. After that, we linearly transform f_v into low-dimensional features e_v by employing an embedding matrix, $W_v \in \mathbb{R}^{4096\times d}$. Therefore, the visual feature embedding of i is e_v. Similarly, we feed the textual description of i into the pre-trained network to obtain deep features f_t, where the f_t dimension is 300. We also linearly transform f_t into low-dimensional features e_t by using an embedding matrix, $W_t \in \mathbb{R}^{300\times d}$. Thus, the textual feature embedding of i is e_t. Moreover, we consider the ID embedding after graph convolution of i as a modality feature. This allows incorporating high-order collaborative signals from the user-item interaction graph into the representation learning of the multimodal content of items. We conduct experiments in Subsection 4.3.1 to verify its validity.

Subsequently, we combine the individual modality features (ID after graph convolution, visual, and textual) of i and i^*:

$$X_i = \{e_i^{base}, e_v, e_t\}; \quad X_{i^*} = \{e_{i^*}^{base}, e_{v^*}, e_{t^*}\}, \tag{6}$$

where e_i^{base} and $e_{i^*}^{base}$ are the final ID embeddings of i and i^* in the basic recommendation framework, *i.e.*, Equation 5, and $X_i \in \mathbb{R}^{d_x}, X_{i^*} \in \mathbb{R}^{d_x}, d_x = 3\times d$.

3.3 Causal Difference Learning

In this subsection, we first adopt a causal graph between MMRec and multimodal content to map recommendation tasks to a causal world and analyze MMRec process from a causal view, thereby obtaining causal difference calculation. Then, we design the multimodal counterfactual learning layer based on causal difference calculation for guiding the model to leverage the multimodal content representations of user-uninteracted items to identify and eliminate preference-irrelevant representations in the multimodal content of user-interacted items. This can further purify the representation of user preference-relevant multimodal content.

3.3.1 ***Counterfactual Learning on Causal Graph.*** We utilize causal graph theory to analyze MMRec process from a causal perspective and obtain causal difference with the help of intervention and counterfactual inference.

Causal Graph Theory. The causal graph is a directed acyclic graph, $\mathcal{G}_{causal} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ stands for the set of variables (nodes) and $\mathcal{E}$ denotes the causal relations between variables (edges). In the causal graph, capital letters (*e.g.*, U) and lowercase letters (*e.g.*, u) represent variables and their observations, respectively. We use a multimodal causal graph to describe the causal relations among variables, as shown in Figure 3(a). In this causal graph, there are four node variables: $\boldsymbol{I}$ (multimodal content of user-interacted item), $\boldsymbol{A}$ (preference distribution on multimodal content of user-interacted item), $\boldsymbol{U}$ (user), and $\boldsymbol{Y}$ (prediction score). The causal paths of interest are as follows:

- Edge $I \rightarrow A$ represents the preference distribution calculated by the recommendation model based on the multimodal content of user-interacted items.
- Edges $\{U, I\} \rightarrow Y$ indicate traditional MMRec process and the prediction score represents the user's preference for a given item.
- Edges $\{U, I, A\} \rightarrow Y$ denote user preferences for different dimensional features on the multimodal content of interacted items, thus indicating the user's interest in a more fine-grained way.

In Figure 3(a), the prediction score Y can be calculated from the values of its ancestor nodes (U, I, and A) as follows:

$$Y_{u,i,a} = f_Y(U = u, I = i, A = a), \tag{7}$$

where the structural formula $f_Y(\cdot)$ corresponds to the main modules of the recommendation model, *i.e.*, the inner product function.

Counterfactual Calculation. After obtaining the multimodal causal graph, we can analyze the causal relationships by manipulating the values of variables and seeing their effects. This operation is known as an intervention in causal inference [26, 30, 52]. The intervention is done by truncating all incoming edges of variable and assigning a specific value, such that the intervened variable is immune to the effects of its parent variable.

To apply the intervention operation to the multimodal causal graph, we imagine a counterfactual world by drawing on counterfactual thinking in causal inference, as shown in Figure 3(b). In this counterfactual world, the variable $\boldsymbol{I^*}$ is the multimodal content of user-uninteracted item. The variable $\boldsymbol{A^*}$ is the preference distribution on multimodal content of user-uninteracted item, which is learned from the variable I^* (multimodal content of user-uninteracted item). Therefore, A^* is independent of the current variable I (multimodal content of user-interacted item) and would be immune from the effect of the current variable I. In Figure 3(b), the prediction score is calculated as follows:

$$Y_{u,i,a^*} = f_Y(U = u, I = i, A^* = a^*). \tag{8}$$

Inspired by prior work [27, 35, 44], the difference value between the real world and counterfactual world is the actual effect of the preference distribution on multimodal content for the prediction score as follows:

$$Y_{effect} = Y_{u,i,a} - Y_{u,i,a^*}. \tag{9}$$

Thus, we regard the learning of such difference values as causal difference calculation. This can lead the model to use the multimodal content representations of user-uninteracted items to mitigate the side effect of spurious correlations (user preference-irrelevant parts) in the multimodal content representations of user-interacted items.

3.3.2 ***Multimodal Counterfactual Learning Layer.*** After analyzing the multimodal causal graph in Subsection 3.3.1 to obtain the causal difference calculation, next, we design the multimodal counterfactual learning layer, which includes the counterfactual learning layer and the feed-forward network. The counterfactual learning layer first calculates the preference distributions on multimodal content of user-interacted and user-uninteracted items. Then, the difference values between the two preference distributions guided by causal difference calculation are used to purify the representation of user preference-relevant multimodal content. The feed-forward network is used to enhance the model fitting ability.

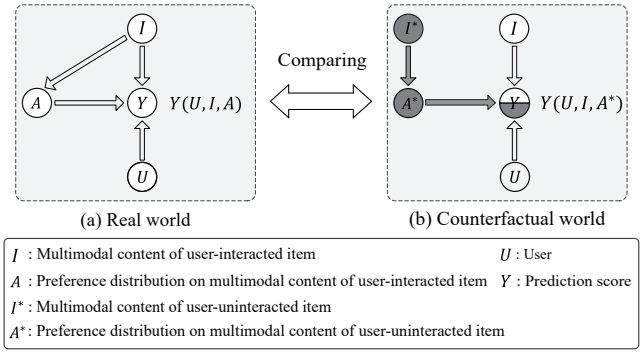

Figure 3: Comparison between real world and counterfactual world multimodal causal graphs in recommender systems.

Counterfactual Learning Layer. First, according to the feature embeddings obtained in Subsection 3.2, we exploit the combined feature embeddings X_i and X_{i^*} in Equation 6 as the input of the counterfactual learning layer.

Then, we calculate the queries, keys, and values of input embedding X_i via the transformation matrix as follows:

$$Q_i = X_i \cdot W_{Q_i}; \quad K_i = X_i \cdot W_{K_i}; \quad V_i = X_i \cdot W_{V_i}, \tag{10}$$

where $W_{Q_i} \in \mathbb{R}^{d_x \times d_x}$, $W_{K_i} \in \mathbb{R}^{d_x \times d_x}$, and $W_{V_i} \in \mathbb{R}^{d_x \times d_x}$; these transformation matrices can be trained, Q_i, K_i, and V_i are essentially linear transformations of the input embedding X_i.

We also calculate the queries and keys of input embedding X_{i^*} as follows:

$$Q_{i^*} = X_{i^*} \cdot W_{Q_{i^*}}; \quad K_{i^*} = X_{i^*} \cdot W_{K_{i^*}}, \tag{11}$$

where $W_{Q_{i^*}} \in \mathbb{R}^{d_x \times d_x}$, and $W_{K_{i^*}} \in \mathbb{R}^{d_x \times d_x}$; these transformation matrices can be trained.

Next, we determine the preference distribution of the values by calculating the similarity between the queries and keys. Specifically, we first obtain the scores of each position using the dot product between the queries and keys, then the preference distributions on multimodal content of i and i^* are calculated by the softmax layer.

$$A = softmax\left(\frac{Q_i \cdot (K_i)^T}{\sqrt{d_x}}\right), \tag{12}$$

$$A^* = softmax\left(\frac{Q_{i^*} \cdot (K_{i^*})^T}{\sqrt{d_x}}\right), \tag{13}$$

where the role of $\sqrt{d_x}$ is to make the gradient values remain stable during the training process and A and A^* correspond to the preference distributions on multimodal content learned from the real world and counterfactual world in Figure 3, respectively.

Subsequently, based on the causal difference calculation in Subsection 3.3.1, we subtract A and A^*, then multiply the difference

with the values V_i to obtain the output of the counterfactual learning layer as follows:

$$e_{cl} = (A - A^*) \cdot V_i. \tag{14}$$

Feed-Forward Network. After obtaining the output embedding e_{cl} of the counterfactual learning layer, we process e_{cl} utilizing residual connection and layer normalization. This can improve the problem of gradient disappearance in the deep model and speed up the model convergence as follows:

$$e_{ln} = LayerNorm(e_{cl} + X_i), \tag{15}$$

where $LayerNorm(\cdot)$ is the layer normalization operation.

Later, we feed the embedding e_{ln} into the feed-forward network, which includes two linear layers and an activation function. Besides, we also process the output embedding e_{ffn} of the feed-forward network by using residual connection and layer normalization. Thereafter, we use the processed embedding as the output embedding $e_{v,t}$ of the multimodal counterfactual learning layer as follows:

$$e_{ffn} = max(0, W_1 e_{ln} + b_1)W_2 + b_2, \tag{16}$$

$$e_{v,t} = LayerNorm(e_{ffn} + e_{ln}), \tag{17}$$

where W_1 and W_2 are the trainable weights of the first and second linear layer, b_1 and b_2 are the respective bias terms, and $max(0, x)$ is the expression of the activation function ReLU.

Finally, to improve the representation learning ability of the model, we stack this multimodal counterfactual learning layer as M layers. The output embedding $e_{v,t}^{(M)}$ of the M-th layer is used as the inter-modality embedding of item i as follows:

$$e_{v,t}^{(M)} = MCL(e_{v,t}^{(M-1)}), \tag{18}$$

where $MCL(\cdot)$ is the multimodal counterfactual learning layer. Note that $e_{v,t}^{(0)}$ is equal to $e_{v,t}$ in Equation 17.

3.4 Multimodal Fusion and Score Prediction

3.4.1 ***Multimodal Fusion***. In multimodal preference modeling, we consider the intra- and inter-modality feature embedding of item i. The intra-modality feature embeddings e_v and e_t of item i are obtained based on the feature encoder in Subsection 3.2. The inter-modality feature embedding $e_{v,t}^{(M)}$ of item i is learned according to the multimodal counterfactual learning layer in Subsection 3.3. Afterward, we sum the intra- and inter-modality feature embeddings of item i as the final embedding of item i in the causal difference learning part.

$$e_i = e_v + e_t + e_{v,t}^{(M)}. \tag{19}$$

3.4.2 ***Prediction Function***. First, according to Equations 4 and 5, we can utilize an inner product operation to calculate the prediction score between user u and item i in the basic recommendation framework as follows:

$$y_{u,i}^{base} = (e_u^{base})^T \cdot e_i^{base}. \tag{20}$$

Then, based on Equations 4 and 19, we can employ an inner product operation to calculate the prediction score between user u and item i in the causal difference learning part as follows:

$$y_{u,i} = (e_u^{base})^T \cdot e_i. \tag{21}$$

Finally, to simultaneously learn the user-item collaborative signals and multimodal information of the item, we sum Equations 20 and 21 as the final prediction score as follows:

$$\hat{y}_{u,i} = y_{u,i}^{base} + \lambda_m \cdot y_{u,i}, \tag{22}$$

where λ_m is used to control the contribution of multimodal feature embedding to user preference prediction.

3.5 Model Optimization

To optimize MCLN, we construct the objective functions of these two prediction scores and design a multi-task strategy to combine these two objectives. We adopt BPR loss [31] as the basic objective function, which assumes that users prefer interacted items to unobserved ones.

To be specific, we first develop $\mathcal{L}_{base}$ to ensure sufficient learning of collaborative signals to better model user preferences in the basic recommendation framework. Then, we propose $\mathcal{L}_m$ to facilitate the learning of multimodal information to better model multimodal user preferences. They are calculated as follows:

$$\mathcal{L}_{base} = \sum_{(u,i,j)\in O} -ln\sigma(y_{u,i}^{base} - y_{u,j}^{base}) + \lambda \cdot \|\mathcal{H}_1\|_2^2, \tag{23}$$

$$\mathcal{L}_m = \sum_{(u,i,j)\in O} -ln\sigma(y_{u,i} - y_{u,j}) + \lambda \cdot \|\mathcal{H}_2\|_2^2, \tag{24}$$

where $O = \{(u, i, j)|(u, i) \in \mathcal{R}_u, (u, j) \notin \mathcal{R}_u\}$ are the training data, $\mathcal{R}_u$ represents the set of items with which user u has historically interacted, and $\sigma(\cdot)$ is the expression of the sigmoid function. The L_2 regularization coefficient λ is a hyper-parameter, $\mathcal{H}_1$ denotes the ID embedding matrix $\mathcal{E}$ in the basic recommendation framework (Subsection 3.1), and $\mathcal{H}_2$ represent the feature matrix (e_v, e_t, and $e_{v,t}^{(M)}$) in the feature embedding (Subsection 3.2) and causal difference learning (Subsection 3.3).

For simultaneous training, multi-task learning is formulated by combining two losses with an addition operation. We formulate the objective function $\mathcal{L}$ to jointly optimize $\mathcal{L}_{base}$ and $\mathcal{L}_m$ as follows:

$$\mathcal{L} = \mathcal{L}_{base} + \mathcal{L}_m. \tag{25}$$

Table 1: Statistics of the three datasets. Note that # V and # T denote the length of visual and textual features, respectively.

Dataset	# User	# Item	# Interaction	Density	# V	# T
Beauty	15,576	8,678	139,318	0.00103	4,096	300
Art	25,165	9,324	201,427	0.00086	4,096	300
Taobao	12,539	8,735	83,648	0.00076	4,096	-

4 EXPERIMENTS

In this section, we conduct abundant experiments to evaluate MCLN. We aim to answer the following three research questions:

- **RQ1:** How does our proposed MCLN perform compared with the state-of-the-art baselines?
- **RQ2:** How do the key components (multimodal features, counterfactual learning layer, intra- and inter-modality features, *etc.*) of MCLN affect performance?
- **RQ3:** How do the distinct settings of MCLN affect performance?

4.1 Experimental Settings

4.1.1 ***Datasets***. As our work focuses on MMRec, we conduct experiments on three public datasets with different densities: **Beauty**,

Table 2: Overall performance comparison. H is short for HR and N is short for NDCG.

Models	Beauty				Art				Taobao			
	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10	H@5	N@5	H@10	N@10
BPRMF [18]	0.4274	0.3343	0.5173	0.3634	0.6333	0.5597	0.7052	0.5829	0.3215	0.2465	0.4049	0.2733
SVD++ [17]	0.4584	0.3592	0.5520	0.3895	0.6530	0.5627	0.7425	0.5916	0.3374	0.2523	0.4293	0.2819
NGCF [43]	0.4853	0.3776	0.5820	0.4089	0.6742	0.5882	0.7541	0.6141	0.3575	0.2658	0.4593	0.2986
LightGCN [12]	0.5229	0.3909	0.6385	0.4285	0.6938	0.5996	0.7710	0.6246	0.3886	0.2896	0.4961	0.3243
VBPR [11]	0.4722	0.3665	0.5670	0.3973	0.6699	0.5830	0.7464	0.6078	0.3464	0.2639	0.4364	0.2928
MMGCN [46]	0.4934	0.3714	0.6067	0.4081	0.6769	0.5643	0.7702	0.5945	0.3649	0.2709	0.4695	0.3047
MGAT [36]	0.5010	0.3781	0.6152	0.4152	0.6825	0.5784	0.7699	0.6067	0.3783	0.2820	0.4882	0.3175
InvRL [6]	0.5130	0.3955	0.6097	0.4268	0.6965	0.5986	0.7748	0.6237	0.3913	0.2926	0.4897	0.3244
DMRL [19]	0.5230	0.4075	0.6215	0.4396	0.6972	0.6008	0.7775	0.6272	0.3757	0.2876	0.4594	0.3146
MCLN	**0.5636**	**0.4368**	**0.6689**	**0.4710**	**0.7168**	**0.6123**	**0.7973**	**0.6384**	**0.4067**	**0.3060**	**0.5105**	**0.3393**
%Imp.	7.76%	7.19%	4.76%	7.14%	2.81%	1.91%	2.55%	1.79%	3.94%	4.58%	2.90%	4.59%

Arts_crafts_and_Sewing (short for **Art**)[1], and **Taobao**[2]. Beauty and Art are real datasets from *Amazon.com* [25], and both contain images, titles, and reviews. Taobao is a real dataset published in the Tianchi competition, which provides visual content only. We adopt the 5-core setting to ensure the quality of these datasets, which means that only users and items with at least 5 interactions are retained. Table 1 shows the statistics of the three experimental datasets. Following the broadly employed setting [11, 14, 48], we utilize the *leave-one-out* method [31] for evaluation.

4.1.2 ***Evaluation Metrics.*** To evaluate the model performance, we choose two widely employed metrics [13, 14, 20], *Hit Ratio* (**HR**) and *Normalized Discounted Cumulative Gain* (**NDCG**). Specifically, HR indicates the average probability of the user's favorite items appearing in the top-k recommendation list. In contrast, NDCG concerns more about the position of the recommended items in the list, and its higher score suggests a more forward position. We report the average HR@k and NDCG@k for all users in the test set, where k is the length of the recommendation list.

4.1.3 ***Baselines.*** To demonstrate the effectiveness of MCLN, we compare MCLN with existing recommendation models, including traditional models (BPRMF, SVD++), GCN-based models (NGCF, LightGCN), and multimedia-based models (VBPR, MMGCN, MGAT, InvRL, DMRL). We briefly introduce those models as follows:

- **BPRMF** [18]: This is a classical collaborative filtering model optimized using BPR loss. It maps user and item representations as latent vectors based on user-item direct interactions.
- **SVD++** [17]: This model incorporates the information of the user's historically interacted neighbors into the user embedding.
- **NGCF** [43]: This model utilizes nonlinear GCN to perform embedding propagation on the user-item interaction graph and produces the final embedding of the nodes by concatenating the embedding learned from all graph convolution layers.
- **LightGCN** [12]: This model exploits a concise linear GCN for embedding generation and accumulates the embeddings generated at each graph convolution layer as the final node representation.
- **VBPR** [11]: This model employs pre-trained convolution neural networks to extract visual features of the items and integrate them into the item embeddings.
- **MMGCN** [46]: This model applies GCNs for embedding propagation on parallel interaction graphs with different modality data to capture user preferences on different modalities.
- **MGAT** [36]: This model weights information propagation in multimodal interaction graphs by constructing attention networks, thus modeling fine-grained multimodal user preferences.
- **InvRL** [6]: This model applies heterogeneous environments to denote the spurious correlations from the multimedia content and learns invariant representations across environments to mitigate their effect.
- **DMRL** [19]: This model uses a disentangled representation technique to ensure that the features of different disentangled factors are independent in each modality and designs multimodal attention mechanisms to obtain user preferences for each factor.

4.1.4 ***Hyper-parameter Settings.*** For all models, we fix the embedding size and batch size to 64 and 2,048, respectively. The learning rate is adjusted in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, and the coefficient of L_2 regulation is searched in $\{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. For all GCN-based models, the number of graph convolution layers is adjusted in $\{1, 2, 3, 4, 5, 6\}$. In our MCLN, the number of multimodal counterfactual learning layers is tuned in $\{1, 2, 3, 4, 5\}$. Furthermore, we apply the Xavier initializer [8] to initialize the embeddings and weight transformation matrices for all models. We adopt the mini-batch Adam optimizer [15] to minimize the objective function.

4.2 Overall Comparison (RQ1)

In this subsection, we perform a detailed comparison of MCLN with all baselines. As shown in Table 2, we summarize the experimental results {HR, NDCG}@{5, 10} of all models on the three datasets. We elaborate our exhaustive observations as follows.

First, in all evaluation metrics of {HR, NDCG}@{5, 10}, our MCLN yields outstanding performance on all datasets. Without any doubt, these results fully demonstrate the rationality and superiority of MCLN. In detail, compared to the strongest performance of the baselines, MCLN achieves an average improvement of 6.71%, 2.27%, and 4.00% on the three datasets, respectively. This result may be due to the multimodal counterfactual learning layer devised in MCLN can guide the model to employ the multimodal content representations of user-uninteracted items to mitigate the side effect of user preference-irrelevant parts in the multimodal content of user-interacted items. That is, the multimodal preference cues

[1] http://deepyeti.ucsd.edu/jianmo/amazon/index.html
[2] https://tianchi.aliyun.com/competition/entrance/231506/information

captured by MCLN are superior to those captured by the baselines. Note that the performance gain of MCLN on the Beauty dataset is better than the Art and Taobao datasets. A possible reason is that the density of the Beauty dataset is higher than the remaining two datasets (*cf.* Table 1).

Second, the multimedia-based baselines (VBPR, MMGCN, MGAT, InvRL, DMRL) are greatly superior to the traditional models (BPRMF, SVD++). This demonstrates the effectiveness of integrating multimodal information into embedding generation to assist in modeling user preferences. Among all multimedia-based baselines, DMRL performs best on the Beauty and Art datasets, which can capture users' attention to different modalities on each factor in user preference modeling. However, the performance of DMRL is weaker on the Taobao dataset, probably because the Taobao dataset is sparse and DMRL does not use GCN to capture high-order collaborative signals. The results of the gap between our MCLN and multimedia-based baselines suggest that multimodal content representations of user-uninteracted items can be used to purify the representations of user preference-relevant multimodal content.

Third, the GCN-based baselines (NGCF, LightGCN) are significantly stronger than BPRMF and SVD++ regarding these three datasets, which verifies the effectiveness of capturing high-order collaborative signals to enhance the expressiveness of the embedding. It also further illustrates the importance of using GCN to aggregate neighbor information for representation learning. In all cases, LightGCN outperforms NGCF, MMGCN, and MGAT. A possible reason is that the concise linear GCN is more suitable for capturing high-order collaborative signals, thereby improving the quality of representation learning. Therefore, we also utilize the linear GCN as the basic recommendation framework of MCLN.

Table 3: Ablation study with key components.

Models	Beauty		Art		Taobao	
	H@5	N@5	H@5	N@5	H@5	N@5
(1) Base	0.5293	0.4069	0.7016	0.5932	0.3933	0.2950
(2) Base w/ V	0.5313	0.4087	0.7028	0.5963	0.3994	0.2993
(3) Base w/ V&CL	0.5384	0.4108	0.7046	0.5980	0.4048	0.3034
(4) Base w/ V&ID&CL	0.5386	0.4116	0.7060	0.5994	-	-
(5) Base w/ T	0.5462	0.4228	0.7105	0.6036	-	-
(6) Base w/ T&CL	0.5582	0.4271	0.7151	0.6054	-	-
(7) Base w/ T&ID&CL	0.5589	0.4309	0.7156	0.6060	-	-
(8) Base w/ V&T	0.5492	0.4230	0.7110	0.6055	-	-
(9) Base w/ V&T&CL	0.5620	0.4362	0.7164	0.6116	-	-
(10) MCLN (Ours)	**0.5636**	**0.4368**	**0.7168**	**0.6123**	**0.4067**	**0.3060**

4.3 Ablation Study (RQ2)

4.3.1 ***Effect of Key Components.*** We first investigate the effectiveness of different components of our MCLN. In particular, we set up the following variants of MCLN:

- **Base**: This variant uses only the basic recommendation framework in MCLN to predict user preferences;
- **Base w/ V**: This variant builds on Base and employs visual features to predict users' multimodal preferences;
- **Base w/ V&CL**: This variant builds on Base and uses visual features as the input of the **counterfactual learning layer**;
- **Base w/ V&ID&CL**: This variant builds on Base and combines visual features and **ID feature embedding after item graph convolution** as the input of the counterfactual learning layer;
- **Base w/ T**: This variant builds on Base and utilizes textual features to predict users' multimodal preferences;
- **Base w/ T&CL**: This variant builds on Base and uses textual features as the input of the counterfactual learning layer;
- **Base w/ T&ID&CL**: This variant builds on Base and combines textual features and ID feature embedding after item graph convolution as the input of the counterfactual learning layer;
- **Base w/ V&T**: This variant builds on Base and uses visual and textual features to jointly predict users' multimodal preferences;
- **Base w/ V&T&CL**: This variant builds on Base and combines visual and textual features as the input of the counterfactual learning layer.

Table 3 records the performance {HR, NDCG}@5 of these variants on the three datasets, leading to the following observations.

First, among all variants of MCLN in Table 3, the worst performance occurs with Base, which only captures the high-order collaborative signals from the user-item interaction graph. In particular, the performance degraded by an average of 6.92% on Beauty, 2.70% on Art, and 3.57% on Taobao. This illustrates the importance of simultaneously utilizing the multimodal features and counterfactual learning layer for user preference modeling. It also demonstrates the effectiveness of leveraging multimodal content to assist with recommendation tasks.

Second, in all datasets, Base w/ V&CL outperforms Base w/ V; Base w/ T&CL outperforms Base w/ T; and Base w/ V&T&CL outperforms Base w/ V&T. These results demonstrate the beneficial impact of the counterfactual learning layer on model performance. This also confirms our claim that the counterfactual learning layer can lead the model to utilize the multimodal content representations of user-uninteracted items to alleviate the side effect of spurious correlations (user preference-irrelevant parts) in the multimodal content representations of user-interacted items. Moreover, we find that the counterfactual learning layer has greater performance gain on combined multimodal features than single-modal features. This may be because different modalities in the counterfactual learning layer have different gains for modeling user preferences. Thus, using multiple modal features simultaneously in the counterfactual learning layer is useful to achieve better recommendations.

Third, in the Beauty and Art datasets, Base w/ T outperforms Base w/ V, which indicates that the different modal features have remarkably different contributions to user preference modeling. In these two datasets, the textual features of items have a more critical effect on user preference modeling. Base w/ V&T outperforms Base w/ T. This result confirms the importance of multimodal information captured from multiple modal features of items for user preference modeling.

Fourth, in the Beauty and Art datasets, we consider the ID feature embedding after item graph convolution during the multimodal feature combination process for Base w/ V&ID&CL, Base w/ T&ID&CL, and MCLN (Ours). This allows incorporating high-order collaborative signals from the user-item interaction graph into the multimodal preference learning process, providing further performance gains on top of Base w/ V&CL, Base w/ T&CL, and Base w/ V&T&CL. Note that in the Taobao dataset, MCLN (Ours) is equal to Base w/ V&ID&CL due to the lack of textual features.

4.3.2 ***Effect of Multimodal Fusion.*** We then investigate the effect of intra- and inter-modality feature fusion strategies on the performance of our MCLN. Specifically, we set up the following variants of MCLN: (a) **MCLN w/o intra-modality**, which removes the intra-modality feature from the multimodal fusion of MCLN, *i.e.*, we remove e_v and e_t from Equation 19; and (b) **MCLN w/o inter-modality**, which removes the inter-modality feature from the multimodal fusion of MCLN, *i.e.*, $e_{v,t}^{(M)}$ is removed from Equation 19. Table 4 records the performance of these variants on the Beauty and Art datasets. MCLN w/o inter-modality outperforms MCLN w/o intra-modality in all cases, which indicates that simply utilizing the inter-modality feature of items is insufficient for mining multimodal preference cues. Moreover, the combined MCLN is significantly superior to MCLN w/o intra-modality and MCLN w/o inter-modality. This result illustrates the effectiveness of strategies that fuse intra- and inter-modality features of items to extract multimodal preference cues.

Table 4: Ablation studies of different modality fusion strategies on Beauty and Art datasets.

Models	Beauty		Art	
	H@5	N@5	H@5	N@5
MCLN w/o intra-modality	0.5379	0.4128	0.7067	0.5939
MCLN w/o inter-modality	0.5492	0.4230	0.7110	0.6055
(Intra&Inter)-modality (Ours)	**0.5636**	**0.4368**	**0.7168**	**0.6123**

4.4 Study of MCLN (RQ3)

In this subsection, we conduct experiments to investigate the effect of pivotal hyper-parameters in MCLN on model performance. We first explore the effect of the graph convolutional layer numbers. We then study how the number of multimodal counterfactual learning layers affects model performance.

4.4.1 ***Effect of Graph Convolution Layer Numbers.*** To analyze the impact of graph convolution layer numbers, we adjust the layer numbers in {1,2,3,4,5,6} and show the results in Figure 4. According to Figure 4, we observe that the model performance continuously promotes with the increase of the layer numbers. However, similar to many GCN-based recommender models [12, 43], stacking too many layers introduces the problem of over-smoothing, which leads to performance degradation. Therefore, model performance shows a peak as the layer number increases. In the Beauty, Art, and Taobao datasets, the optimal number of graph convolution layers for MCLN are 4, 5, and 5, respectively.

4.4.2 ***Effect of Multimodal Counterfactual Learning Layer Numbers.*** To investigate the impact of the number of multimodal counterfactual learning layers, we search the layer numbers in {1,2,3,4,5}. Table 5 summarizes the experimental result. In Table 5, we find that the model performance shows a peak change as the number of layers increases. We attribute this to the fact that exploiting too many multimodal counterfactual learning layers interferes with the representation learning process. In the Beauty, Art, and Taobao datasets, the optimal number of multimodal counterfactual learning layers for MCLN are 2, 2, and 4, respectively. Besides, MCLN with different multimodal counterfactual learning layers consistently outperforms the baselines on the three datasets.

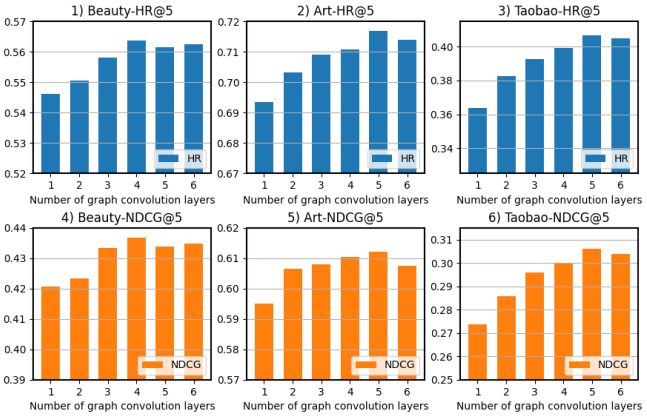

Figure 4: Effect of graph convolution layer numbers.

This result further validates the effectiveness of the multimodal counterfactual learning layer in MCLN.

Table 5: Effect of multimodal counterfactual learning layer numbers.

Layers	Beauty		Art		Taobao	
	H@5	N@5	H@5	N@5	H@5	N@5
1	0.5602	0.4302	0.7145	0.6114	0.4012	0.3008
2	**0.5636**	**0.4368**	**0.7168**	**0.6123**	0.4018	0.3021
3	0.5629	0.4361	0.7138	0.6081	0.4044	0.3046
4	0.5616	0.4316	0.7126	0.6066	**0.4067**	**0.3060**
5	0.5600	0.4280	0.7115	0.6006	0.4036	0.3041

5 CONCLUSION AND FUTURE WORK

In this work, we propose a novel MMRec model, MCLN, which combines the basic recommendation framework with the multimodal counterfactual learning layer for purifying the representation of user preference-relevant multimodal content to better match the user's interests. We capture sufficient collaborative signals on the user-item interaction graph with only historical interaction data, and use pre-trained deep networks to extract multimodal features of the sample pairs (user-interacted and user-uninteracted items). The model utilizes causal theory to guide the multimodal counterfactual learning layer for modeling the causal difference between the multimodal content of user-interacted and user-uninteracted items, thereby eliminating preference-irrelevant representations in the multimodal content of user-interacted items, which may have a spurious correlation with user preferences. Extensive experiments on three public datasets justify the effectiveness of MCLN and its components. In the future, we would attempt to incorporate rich relationships among items into MMRec to enhance the quality of representation learning. Besides, we will consider how to use multimedia data to improve the interpretability of recommendations.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Natural Science Foundation of China (Grant No. 62272143), the Anhui Provincial Major Science and Technology Project of China (Grant No. 202203a05020025), the University Synergy Innovation Program of Anhui Province: GXXT-2022-054, and the Seventh Special Support Plan for Innovation and Entrepreneurship in Anhui Province.

REFERENCES

[1] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*. 1–16.

[2] Rianne van den Berg, Thomas N Kipf, and Max Welling. 2017. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263* (2017).

[3] Desheng Cai, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2021. Heterogeneous hierarchical feature aggregation network for personalized micro-video recommendation. *TMM* 24 (2021), 805–818.

[4] Lei Chen, Le Wu, Richang Hong, Kun Zhang, and Meng Wang. 2020. Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach. In *AAAI*. 27–34.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. IEEE, 248–255.

[6] Xiaoyu Du, Zike Wu, Fuli Feng, Xiangnan He, and Jinhui Tang. 2022. Invariant Representation Learning for Multimedia Recommendation. In *MM*. ACM, 619–628.

[7] Fuli Feng, Jizhi Zhang, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Empowering language understanding with counterfactual reasoning. In *ACL-IJCNLP Findings*. ACL, 2226–2236.

[8] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*. JMLR, 249–256.

[9] Dan Guo, Hui Wang, and Meng Wang. 2021. Context-aware graph inference with knowledge distillation for visual dialog. *TPAMI* 44, 10 (2021), 6056–6073.

[10] Dan Guo, Hui Wang, Shuhui Wang, and Meng Wang. 2020. Textual-visual reference-aware attention network for visual dialog. *TIP* 29 (2020), 6655–6666.

[11] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *AAAI*. 144–150.

[12] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR*. ACM, 639–648.

[13] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial personalized ranking for recommendation. In *SIGIR*. ACM, 355–364.

[14] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*. ACM, 173–182.

[15] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*. 1–15.

[16] Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*. 1–14.

[17] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *SIGKDD*. ACM, 426–434.

[18] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[19] Fan Liu, Huilin Chen, Zhiyong Cheng, Anan Liu, Liqiang Nie, and Mohan Kankanhalli. 2022. Disentangled Multimodal Representation Learning for Recommendation. *TMM* (2022), 1–11.

[20] Kang Liu, Feng Xue, Dan Guo, Peijie Sun, Shengsheng Qian, and Richang Hong. 2023. Multimodal Graph Contrastive Learning for Multimedia-Based Recommendation. *TMM* (2023), 1–13.

[21] Kang Liu, Feng Xue, Dan Guo, Le Wu, Shujie Li, and Richang Hong. 2023. MEGCF: Multimodal Entity Graph Collaborative Filtering for Personalized Recommendation. *TOIS* 41, 2 (2023), 1–27.

[22] Kang Liu, Feng Xue, Xiangnan He, Dan Guo, and Richang Hong. 2023. Joint Multi-Grained Popularity-Aware Graph Convolution Collaborative Filtering for Recommendation. *TCSS* 10, 1 (2023), 72–83.

[23] Kang Liu, Feng Xue, Shuaiyang Li, Sheng Sang, and Richang Hong. 2022. Multimodal Hierarchical Graph Collaborative Filtering for Multimedia-Based Recommendation. *TCSS* (2022), 1–12.

[24] Yuanxing Liu, Zhaochun Ren, Wei-Nan Zhang, Wanxiang Che, Ting Liu, and Dawei Yin. 2020. Keywords generation improves e-commerce session-based recommendation. In *WWW*. ACM, 1604–1614.

[25] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *EMNLP-IJCNLP*. 188–197.

[26] Judea Pearl. 2009. *Causality*. Cambridge university press.

[27] Judea Pearl. 2022. Direct and indirect effects. In *Probabilistic and Causal Inference: The Works of Judea Pearl*. 373–392.

[28] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2020. Two causal principles for improving visual dialog. In *CVPR*. IEEE, 10860–10869.

[29] Ruihong Qiu, Sen Wang, Zhi Chen, Hongzhi Yin, and Zi Huang. 2021. Causalrec: Causal inference for visual debiasing in visually-aware recommendation. In *MM*. ACM, 3844–3852.

[30] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *ICCV*. IEEE, 1025–1034.

[31] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *UAI*. 452–461.

[32] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *WWW*. ACM, 285–295.

[33] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*. 1–14.

[34] Chien-Lin Tang, Jingxian Liao, Hao-Chuan Wang, Ching-Ying Sung, and Wen-Chieh Lin. 2021. Conceptguide: Supporting online video learning with concept map-based recommendation of learning path. In *WWW*. ACM, 2757–2768.

[35] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *CVPR*. IEEE, 3716–3725.

[36] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. 2020. MGAT: multimodal graph attention network for recommendation. *IPM* 57, 5 (2020), 102277.

[37] Quoc-Tuan Truong, Aghiles Salah, and Hady Lauw. 2021. Multi-modal recommender systems: Hands-on exploration. In *RecSys*. ACM, 834–837.

[38] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR*. 1–12.

[39] Shoujin Wang, Liang Hu, Yan Wang, Xiangnan He, Quan Z Sheng, Mehmet A Orgun, Longbing Cao, Francesco Ricci, and Philip S Yu. 2021. Graph learning based recommender systems: A review. In *IJCAI*. 4644–4652.

[40] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. 2020. Visual commonsense r-cnn. In *CVPR*. IEEE, 10760–10770.

[41] Wenjie Wang, Fuli Feng, Xiangnan He, Xiang Wang, and Tat-Seng Chua. 2021. Deconfounded recommendation for alleviating bias amplification. In *SIGKDD*. ACM, 1717–1725.

[42] Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue. In *SIGIR*. ACM, 1288–1297.

[43] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *SIGIR*. ACM, 165–174.

[44] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. 2021. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *SIGKDD*. ACM, 1791–1800.

[45] Yinwei Wei, Xiang Wang, Xiangnan He, Liqiang Nie, Yong Rui, and Tat-Seng Chua. 2021. Hierarchical user intent graph network for multimedia recommendation. *TMM* 24 (2021), 2701–2712.

[46] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *MM*. ACM, 1437–1445.

[47] Le Wu, Xiangnan He, Xiang Wang, Kun Zhang, and Meng Wang. 2023. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. *TKDE* 35, 5 (2023), 4425–4445.

[48] Feng Xue, Xiangnan He, Xiang Wang, Jiandong Xu, Kai Liu, and Richang Hong. 2019. Deep item-based collaborative filtering for top-n recommendation. *TOIS* 37, 3 (2019), 1–25.

[49] Liangwei Yang, Zhiwei Liu, Yu Wang, Chen Wang, Ziwei Fan, and Philip S Yu. 2022. Large-scale Personalized Video Game Recommendation via Social-aware Contextualized Graph Neural Network. In *WWW*. ACM, 3376–3386.

[50] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *SIGKDD*. ACM, 974–983.

[51] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *SIGKDD*. ACM, 353–362.

[52] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal intervention for leveraging popularity bias in recommendation. In *SIGIR*. ACM, 11–20.