EliMRec: Eliminating Single-modal Bias in Multimedia Recommendation

Xiaohao Liu
Communication University of China
Beijing, China
xiaohao.liu@hotmail.com

Zhulin Tao*
Communication University of China
Beijing, China
taozhulin@gmail.com

Jiahong Shao
Communication University of China
Beijing, China
shaojh2001@gmail.com

Lifang Yang
Communication University of China
Beijing, China
yanglifang@cuc.edu.cn

Xianglin Huang
Communication University of China
Beijing, China
huangxl@cuc.edu.cn

ABSTRACT

The main idea of multimedia recommendation is to introduce the profile content of multimedia documents as an auxiliary, so as to endow recommenders with generalization ability and gain better performance. However, recent studies using non-uniform datasets roughly fuse single-modal features into multi-modal features and adopt the strategy of directly maximizing the likelihood of user preference scores, leading to the single-modal bias. Owing to the defect in architecture, there is still room for improvement for recent multimedia recommendation.

In this paper, we propose EliMRec, a generic and modal-agnostic framework to eliminate the single-modal bias in multimedia recommendation. From our observation, biased predictive reasoning is influenced directly by the single modality rather than considering the all given multiple views of the item. Through the novel perspective of causal inference, we manage to explain the single-modal issue and exploit the inner working of multi-modal fusion. To eliminate single-modal bias, we enhance the bias-capture ability of a general multimedia recommendation framework and imagine several counterfactual worlds that control one modality variant with other modality fixed or blank. Truth to be told, counterfactual analysis enables us to identify and eliminate bias lying in the *direct effect* from single-modal features to the preference score. Extensive experiments on real-world datasets demonstrate that our method significantly improves over several state-of-the-art baselines like LightGCN [11], and MMGCN [42]. Codes are available at https://github.com/Xiaohao-Liu/EliMRec.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

Multimedia Recommendation, Single-modal bias, Counterfactual Analysis, Micro-videos

ACM Reference Format:
Xiaohao Liu, Zhulin Tao, Jiahong Shao, Lifang Yang, and Xianglin Huang. 2022. EliMRec: Eliminating Single-modal Bias in Multimedia Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548404

1 INTRODUCTION

Multimedia recommendation is widely applied on several applications, such as E-commerce, spanning from images, and especially content sharing platforms along with the bloom of micro-videos. The conventional recommenders leverage historical interactions to construct user preference to filter the overload information [15, 26]. Distinct from such historical-only methods, incorporating multimodal features (*i.e.*, visual, acoustic, and textual) of multimedia documents with collaborative filtering benefits a more comprehensive description of the preferences from the user on given items, while presents additional challenges, like multi-modal representation learning.

On account of collaborative filtering methods, multi-modal features serve as the inductive bias balancing the final predicting score [10] or are integrated into users' and items' embedding to compensate for their representation [42]. From this viewpoint, the main challenges of these content-auxiliary recommendations besides historical-only recommendations are how to (1) fuse single modality features to the multi-modal feature, and (2) inject it into the framework of recommendation tasks. However, the generated multi-modal features are inevitably biased by the single-modal feature by following the optimized strategy of maximizing the likelihood that a user clicks the recommended items under non-uniform datasets (i.e., single-modal bias).

As illustrated in Figure 1, the training dataset contains a series of spy movies, like *Mission Impossible.* The different visual content of this series of films corresponds to different dialogues, but similar theme music. As expected, the counts of acoustic similarity obey the long-tail distribution in the statistic. In the training stage, the model tends to enhance the acoustic modality rather than constructing the assumed multi-modal features. Consequently, the comedy *Ace Ventura* introducing the same background music of

[*] Corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548404

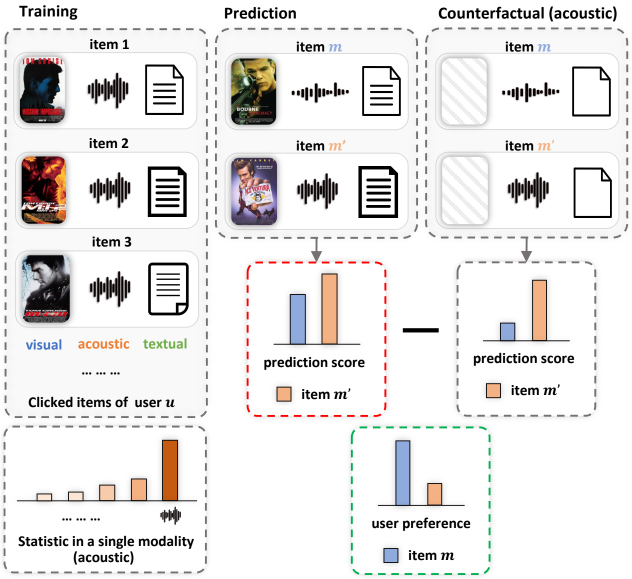


Figure 1: The single-modal bias looks at multimedia recommendation.

Mission Impossible Series gains a higher rating score than the classic spy film *The Bourne Identity*. Without loss of generality, there are similar issues corresponding to other modalities: the non-uniform distribution of single modality causes a spurious relationship from the item on preference score and leaves multi-modal feature being covered.

There are two straightways to address the above issues: (1) use the uniform dataset to establish an unbiased world without shortcuts, thus forcing the model to mine the multi-modal pattern, and (2) remove the biased modality content. However, collecting such a uniform dataset is costly and somewhat insufficient for manually omitting the biased samples. While every modality content is necessary to compensate the recommendation performance. Thus, we raise a question- is there a way to predict unbiased preferences in multimedia recommendation while keeping the dataset intact and away from any manual modification?

In this work, we unravel the inner working of multi-modal fusion from a novel perspective of causal inference and leverage the counterfactual analysis to eliminate the single-modal bias. Different from the conventional understanding of the multimedia recommendation scheme, we deepen the fusion procedure to an abstract diagram that can be decomposed into three steps: (1) capture the multi-modal features, (2) learn each single-modal feature and (3) combine them with a rational balance. Following this diagram, we simplify the complex interactions between each modality to the new causal graph, which dismisses the specific modality and explains how single-modal directly interferes with the preference score. Moreover, the generic solution is technically achieved. We renew the conventional modal for better bias-capture ability and apply the inference phase for bias elimination. Specifically, we construct several counterfactual worlds that control one modality variant with other modality fixed or blanked, as shown in Figure 1. We select a graph neural network-based recommender as the backbone and introduce three branches for capturing every single modality bias in the training phase. During the inference phase, we remove this direct effect from the original predictions (*i.e.*, the total effect). As illustrated above, the inference scores of m and m' in the factual world increase the difference compared to before, thus item m takes over the position of correct.

To sum up, the contributions of this work are threefold:

- We reveal the inner working of multi-modal fusion through causal inference that accordingly describes the single-modal issue and provides its generic solution.
- We devise a brand-new multimedia recommendation that introduces additional bias-capture branches and eliminates the single-modal bias in the inference phase.
- We perform extensive experiments on three datasets to verify the rationality and effectiveness of EliMRec.

2 RELATED WORK

2.1 Multi-modal Recommendation

Previous works in recommendation commonly upgraded on collaborative filtering (CF) [12, 25, 26, 34, 43]. CF-based models leverage the historical interactions to represent the user preference.Whereas the high-skewed data distribution and unseen interactions make CF-based model not generalized enough. Moreover, graph neural network, like NGCF [35] and LightGCN [11], captures the high-order interactions and empowers recommendation ability, . The profile content is injected into the representations to alleviate its high dependency on interactions. For instance, He *et al*. [10] integrated visual information on a scalable factorization model. Gao *et al*. [4] leveraged attention mechanism to future extract efficient visual features and then predicted user's opinion on recipes. Recent works [18, 30, 31, 40, 42, 48] raised the issue of inadequate data mining on multimodal. While, Wei *et al*. [42] proposed MMGCN, which utilized graph information of each modality and fused them into the embedding for preference prediction. We combine the simplified structure of LightGCN and incorporate multimedia features into it by following the parallel GNNs for each modality of MMGCN. This framework serves as the backbone for ensuring the multiemedia recommendation task and meanwhile keeping efficiency.

2.2 Causality-inspired Recommendation

The causality thinking is widely used in recommendation for handling prevailing biases[3, 33, 38]. Previous works [5, 24, 27] commonly leveraged inverse propensity weighting (IPW) to assign different weights to each feedback for training an unbiased model. However, IPW-based approaches may be vulnerable to small probability [33] Thus the additional uniform data is leveraged to finetuning the pre-train biased model, endowing better robustness and generalization abilities [17]. Yi *et al*. [46] introduced an additional branch for bias-aware click prediction. Cadene *et al*. [2] further deeply analyzed the gradient propagation on the VQA model then removed the unimodal bias. Causality thinking exploits inner relationships and conduces explainability [36, 37]. Causal inference [8, 21] provides tools to analyze specific effects such as direct effect and indirect effect [22] targeting the problems, whose widespread framework is structure causal model [20]. By observing the causal graph, there are several useful techniques used in RS, CV, and NLP, such as back-door adjustment [16, 32, 44, 49], front-door adjustment [45], and instrumental variable [29]. Wherein counterfactual thinking reconstructs a non-existent world from the factual world through imagination [17, 19, 28, 33, 39, 47], but all the effects can be identified by existing data.

3 TASK FORMULATION

In this section, we introduce the generic recommendation task [10, 26] and multi-modal GNNs in recommendation [11, 31, 35, 42].

3.1 Generic Recommendation

The task of recommendation is to predict the preference score $\hat{y}_{u,i}$ of user u in set $\mathcal{U}$ given item i in set $\mathcal{I}$. In general, we apply the inner product to their final representation:

$$\hat{y}_{u,i} = \mathbf{z}_u^T \mathbf{z}_i, \tag{1}$$

where $\mathbf{z}_u$ and $\mathbf{z}_i$ are the final representations of user u and item i respectively and $\hat{y}_{u,i}$ denotes their preference score. For optimization, *Bayesian Personalized Ranking* (BPR) Loss [26] is widely used with supervised signals, and it relies on the assumption that the user prefers the observed item rather than the unobserved one:

$$\text{BPRLoss} = -\frac{1}{N}\sum_{i,i'} \sigma(\hat{y}_{u,i} - \hat{y}_{u,i'}), \tag{2}$$

where (i, i') is the pair used in the training stage and i is observed while i' is not, and N denotes the number of training pairs.

3.2 Multi-modal GNNs

Graph neural network (GNN) [50] is a technique introduced to capture the high order interactions in collaborative filtering [41, 42] (*i.e.*, CF), and is also applied to build multiple branches of modalities for generating unimodal representations [31, 42]. Here, the four modalities are concerned in the set $\mathbb{S} = \{v, a, t, id\}$, where v, a, t, id denotes visual, acoustic, textual and ID modality, respectively. Thus, a generic scheme of multi-modal GNNs can be formulated as follows:

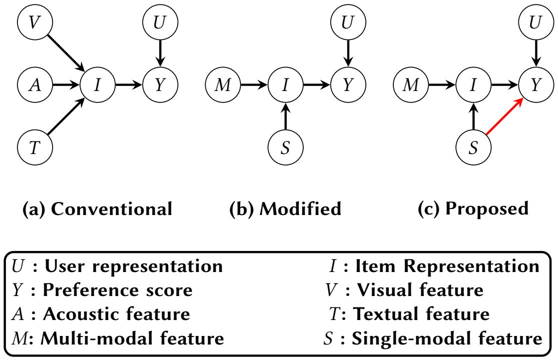
$$\mathbf{u}_s, \mathbf{i}_s = g(\mathcal{G}, \mathbf{E}_s), \mathbf{z}_u = h(\{\mathbf{u}_s | s \in \mathbb{S}\}), \mathbf{z}_i = h(\{\mathbf{i}_s | s \in \mathbb{S}\}), \tag{3}$$

where $\mathbf{E}_s$ denotes the feature matrix of all nodes in modality s, $g(\cdot)$ represents the message passing mechanism producing the representations $\mathbf{u}_s$ and $\mathbf{i}_s$ of each modality s, and $h(\cdot)$ is used to fuse the above representations of every modality and generate the final representation $\mathbf{z}$. Thereby, we describe the scheme in detail from two aspects:

- **Message passing**: Message passing encodes the graph structure of data into each node representation by three consecutively executed mechanisms: (1) aggregation, (2) update and (3) readout. These three mechanisms collect the message of the central node and its neighbors, thus generating the unimodal representation $\mathbf{z}$.
- **Multi-modal fusion**: To implement fusion function $h(\cdot)$, there are two methods, using a nonlinear transformation $f(\cdot)$ [42]: (1) concatenation and (2) addition:

$$h_{\text{concat}}(\{\mathbf{z}_s | s \in \mathbb{S}\}) = f(\mathbf{z}_s || \mathbf{z}_a || \mathbf{z}_t || \mathbf{z}_{id}),$$
$$h_{\text{add}}(\{\mathbf{z}_s | s \in \mathbb{S}\}) = f(\frac{1}{|\mathbb{S}|}\sum_{s\in\mathbb{S}} \mathbf{z}_s), \tag{4}$$

where h_{concat} concatenates the four unimodal representations to a non-linear transformation, while h_{add} replaces the concatenation operation by addition.



(a) Conventional (b) Modified (c) Proposed

Figure 2: Comparison among causal graphs of (a) conventional multimedia recommendation, (b) modified version of conventional multimedia recommendation and (c) our proposed multimedia recommendation.

4 CAUSAL VIEW ON MULTIMEDIA RECOMMENDATION

For a systemic understanding of single-modal bias on multimedia recommendation, we introduce three causal graph depicted in Figure 2(c). Causal graph is a crucial component of structure causal model (SCM) [20, 23], a widespread framework in causal inference. It is also a directed acyclic graph, in which a set of variables represent nodes and the causal relationship between them is represented by directed edges. In this case, there are five variables: multi-modal feature (M), single-modal feature (S), user representation (U), item representation (I) and preference score (Y) of U on I. And there are several causal relations among the variables of interest (*i.e.*, M, S, U, I and Y):

- $S \longrightarrow Y$: The aforementioned single-modality issue tells the prediction to be directly influenced by the single modality feature, rather than considering all the given multiple views of the item more, which leads to a spurious relationship between I and Y.
- $U\&I \longrightarrow Y$: Inherited from the conventional recommendation task (see Eq.1), prediction scores denote the preference of a user u given a specific item i.
- $M\&S \longrightarrow I \longrightarrow Y$: Multi-modal and single-modal features construct the item representations in terms of fusion operation, causing effects on prediction scores.

On the viewpoint of S, I serves as a mediator, controlling the effect of M and S on Y, such as scaling, reversing, and nonlinear transformations. M makes difference in prediction through only the mediator, whereas S takes over both the direct path $S \longrightarrow Y$ and the mediated path $S \rightarrow I \rightarrow Y$, resulting in a spurious correlation between the given item and predication score. Before we propose its solution, we demonstrate the correctness and rationality of the proposed graph as follows.

4.1 Revealing the Inner Working of Multi-modal Fusion

The most common method for generating multi-modal features is to take each modality as an input and use a fusion function. Such procedure is essential but somewhat rough (*e.g.*, concatenation or summation among single modalities), that blinds our eyes to see the problem clearly. In multimedia recommendation, three main modalities (*i.e.*, visual, acoustic, and textual) are encoded to distinctive vectors in unique latent space and then fused to a multi-modal representation. For better clarity of how modalities affect prediction, the conventional multi-modal recommendation is intuitively abstracted by Figure 2 (a) where V, A and T denote item's visual modality, acoustic modality, and textual modality respectively. And all the features of each modality construct the item representation leveraging fusion function $h(\cdot)$ on $V\&A\&T \rightarrow I$ (*e.g.*, in Eq. 4, $h(\cdot)$ fuses $\mathbf{z}_s$ produced by four distinct unimodal GNN encoders).

Following the generic recommendation prediction, the Total Effect (TE) of I on the preference score can be quantified by computing the fluctuations of Y in relation to the change in I:

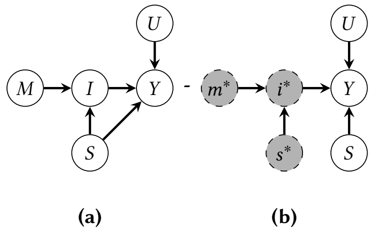
$$TE = Y_{u,i} - Y_{u,i^*}, \tag{5}$$

where $Y_{u,i}$ denotes the determined preference score given user u and item i, and i^* is a constant index, whose representation z^{i^*} is a learnable vector during the training phase.

However, the contribution of each modal feature or its combined actions (*i.e.*, constructed multi-modal features) is difficult to illustrate in the conventional causal graph. We argue that item representations consist of both multi-modal features and single-modal features with a balance in an unbiased model. The multi-modal feature uncovers the generic pattern of items that the user interacted, while the single-modal feature, whose single modalities are similar (*e.g.*, users who like dance would like to see more dance videos even if the background music are variant songs), contains the inductive preference of user on a group of items. Thus, the fusion function $h(\cdot)$ can be decomposed into three steps: (1) capture the multi-modal features $\mathbf{z}^m$, (2) learn each unimodal feature $\mathbf{z}^s$ and (3) combine them with a rational balance. These three operations are united and intractable to be constructed in reality, owing to the fact that the first step may be biased by the original single-model feature but would be formulated in theory given an unbiased multi-modal feature capturer. The final step of $h(\cdot)$ produced the representation can be formulated as:

$$\mathbf{z}_{(\cdot)} = h'(\mathbf{z}^m_{(\cdot)}, \{\mathbf{z}^s_{(\cdot)} | s \in \mathbb{S}/\{id\}\}), \tag{6}$$

where $h'(\cdot)$ shows how $h(\cdot)$ adjusts the information from $\mathbf{z}^m$ and $\mathbf{z}^s$ to be exposed in $\mathbf{z}$. From this point, we improve the conventional to the modified that can see their together influences on item representation as shown in Figure 2 (b). Compared to Figure 2 (a), the specific modal feature (*i.e.*, V, A and T) jointly construct the multi-modal feature (*i.e.*, M) and independently construct the single-modal feature (*i.e.*, S) and for the sake of brevity, they are omitted in the graph as much as they are considered as exogenous variables and do not interfere with the effects of the study. The paths $M \rightarrow I$ and $S \rightarrow I$ mean the direct effects of M on I and S on I respectively.



(a) (b)

Figure 3: The causal graphs in (a) real world and (b) counterfactual world that I is fixed to i^*.

However, from our observations, it appears that single modalities are emphasized unconsciously, which constitutes the shortcut of $S \rightarrow Y$ in Figure 2(c). And TE then takes on another form of formulation from Eq. 5:

$$TE = Y_{u,i,s} - Y_{u,i^*,s^*}, \tag{7}$$

where the direct effect of S on Y is taken into account via $S \rightarrow Y$, which differs apparently from Figure 2(b).

To sum up, in Figure 2, (a) and (b) are unbiased causal graphs, where (b) is a modified version of (a) for a clear perspective into multi-modal fusion that focus more on the impacts of $V\&A\&T \rightarrow I$, while (c) is the proposed biased one based on (b) to illustrate the single-modal issue aforementioned.

4.2 Eliminating the Single-modal Bias

From our proposal, the key to eliminate the single-modal bias is to leverage the existing biased world to recreate a world approximating to the unbiased one, in other words, the counterfactual world. The counterfactual is the third-level causal hierarchy of imagination [20]. Counterfactual analysis helps to answer some special questions following the pattern of "Y would be y had M been m^* in situation $U = u$". In our case, the question we need to answer is how single-modal features cause the preference score had item representation been fixed.

In Figure 3(a), the calculable causal effect (*i.e.*, *Total Effect, TE*) of S on Y consists of two part: (1) *Total Indirect Effect* (*i.e.*, TIE) via $S \rightarrow I \rightarrow Y$ and (2) *Natural Direct Effect* (*i.e.*, NDE) via $S \rightarrow Y$, which can be defined the summation on them:

$$TE = TIE + NDE. \tag{8}$$

In Figure 3(b), I is fixed to i^* that blocks the mediated path from S to Y according to *d-separation criterion* [6] and remains the *NDE*, which can be defined as follows:

$$NDE = Y_{u,i^*,s} - Y_{u,i^*,s^*}. \tag{9}$$

As we mentioned before, I is the mediator of S on Y, and $TE \neq NDE$ in the presence of the indirect effect via the path $S \rightarrow I \rightarrow Y$. Nevertheless, in the counterfactual world, $TE_{i^*} = NDE$ for the constant of mediator I and the direct effect (*i.e.*, $S \rightarrow Y$) can be computed precisely.

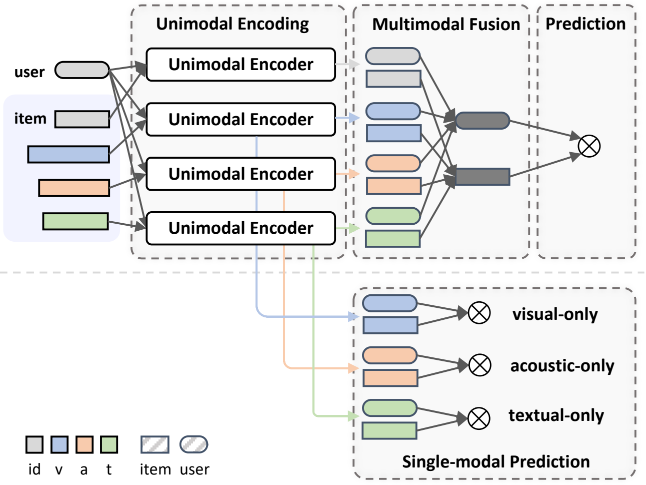


Figure 4: The flowchart of EliMRec. The first line depicts a general multi-modal recommender, consisting of three primary steps: single-modal encoding, multi-modal fusion, and final prediction. The second line shows the added branch, utilizing the unimodal representations for individual predictions.

As illustrated in Figure 3, here we exclude the shortcut of S on Y by counterfactual analysis and define TIE as follows:

$$TIE = TE - NDE = Y_{u,i,s} - Y_{u,i^*,s}. \tag{10}$$

Accordingly, single-modal bias can be eliminated by subtrating NDE from TE.

5 METHODOLOGY

In this section, we technologically achieved the debias approach according to the causal viewpoint.

5.1 Model Formulation

To calculate the causal effect of TIE, we devise the multimedia debiased recommender that captures the direct effect on $S \rightarrow Y$. In accordance with the computation of TIE in Eq. 10, $Y_{u,i,s}$ is the preference score when $U = u$, $I = i$ and $S = s$, which can be formally defined as:

$$Y_{u,i,s} = F(\mathbf{z}_u, \mathbf{z}_i, \{\mathbf{z}_u^s, \mathbf{z}_i^s | s \in \mathbb{S}/\{id\}\}), \tag{11}$$

where $F(\cdot)$ is a function corresponding to the path of $U\&I\&S \rightarrow Y$, taking the final representations of users and items and their unimodal representations (*e.g.*, $\mathbf{z}_u^v$ and $\mathbf{z}_i^v$) as its input to calculate the score. Furthermore, $F(\cdot)$ can be recognized as a composite function, consisting score fusion function that accepts the normal prediction score $\hat{y}_{u,i}$ and the single-modal score $\hat{y}_{u,i}^s$, and two score modules. Thus, the $Y_{u,i,s}$ can be formulated in another modified form:

$$Y_{u,i,s} = F'(\hat{y}_{u,i}, \{\hat{y}_{u,i}^s | s \in \mathbb{S}/\{id\}\}), \tag{12}$$

where $F'(\cdot)$ denotes the score fusion function. Inspired by [33], the first score module can be defined as a multi-modal recommender, whereas the second score module captures the single-modal bias and serves as an add-on to the first module.

The flowchart's main branch, shown in Figure 4, corresponds to the first module, a conventional recommender making predictions of user preference through its fused representation. While the three single-modal predictions in the below branch, which are formulated in a different counterfactual world, represent the second module. For capturing the single-modal bias, we formalize the calculation of the effect on the path $S \rightarrow Y$ in Figure 3(a). Benefiting from the natural structure of multi-modal recommender that processes the single-modal feature independently at first and fuses them to multi-modal feature, we directly utilize the features after unimodal encoding. Here we transform these features through a nonlinear transformation $f'(\cdot)$ firstly as shown below:

$$\mathbf{z}_u^s = f'(u_s), \mathbf{z}_i^s = f'(i_s), \tag{13}$$

where $s \in \mathbb{S}/\{id\}$ and z^s denotes the single-modal representation for the prediction. Furthermore, we obtain the single-modal prediction:

$$\hat{y}_{u,i}^s = \mathbf{z}_u^{sT} \mathbf{z}_i^s. \tag{14}$$

As for normal scores predicted by the first module, we follow the architecture of GNN-type in subsection 3.2, where the single modal features are encoded via message passing mechanism and then fused to the final representations.

5.2 Training Phase

To achieve optimal parameters, we separately construct objective functions of the two ranking score modules and devise a multi-task strategy that combines these two types of objectives. Pursuant to Eq. 2, we apply BPRLoss to each training branch with the supervised triad (u, i, j), where i is a positive sample of user u, while j is a negative sample, which can be formulated as:

$$\mathcal{L}_{main} = -\frac{1}{N}\sum_{i,j}\sigma(\hat{y}_{u,i} - \hat{y}_{u,j}), \mathcal{L}_s = -\frac{1}{N}\sum_{i,j}\sigma(\hat{y}_{u,i}^s - \hat{y}_{u,j}^s), \tag{15}$$

where $\mathcal{L}_{main}$ and $\mathcal{L}_s$ denote the training loss in the main branch and the bias-captured branch respectively. For simultaneous training, multi-task learning is designed by the joint two types of losses with an addition operation. In formal, the final training loss is defined as:

$$\mathcal{L} = \quad \mathcal{L}_{main} + \alpha \cdot \sum_{s \in \mathbb{S}/\{id\}} \mathcal{L}_s, \tag{16}$$

where α is the hyperparameter, controlling the strength of all the $\mathcal{L}_s$ for $s \in \mathbb{S}/\{id\}$. That means we view each modality with identical importance. However, if you have prior knowledge of distinct biases in the dataset, the different importance of each modality is encouraged to be tuned; otherwise, the identical weight is a trade-off between efficiency and efficacy.

5.3 Inference Phase

In this phase, we leverage the counterfactual inference to eliminate single-modal bias, based on the given spurious ranking scores $\hat{y}_{u,i}$ in the main branch and the added unimodal preference scores.

- **Score fusion**: Inspired by the prior works [2], the score fusion function is expected to aggregate the two-type of branches and then produce the final prediction score:

$$\hat{y}_{u,i,s} = \hat{y}_{u,i} \prod_{s \in \mathbb{S}/\{id\}} \sigma(\hat{y}^{s}_{u,i}), \tag{17}$$

where $\sigma(\cdot)$ scales the single-modal prediction scores of $\hat{y}^{s}_{u,i}$ to probabilities and is used to adjust the extent of relying upon $\hat{y}_{u,i}$.

- **Counterfactual inference**: Follows the Eq. 10, the single-modal bias is captured by the formulated structure equation and eliminated by subtracting NDE of S on Y from TE:

$$TIE = \hat{y}_{u,i,s} - \hat{y}_{u,i^*,s}. \tag{18}$$

And specifically, we unite it with Eq. 17, and then apply the distributive property to get the final form of formulation:

$$TIE = (\hat{y}_{u,i} - \hat{y}_{u,i^*}) \prod_{s \in \mathbb{S}/\{id\}} \sigma(\hat{y}^{s}_{u,i}). \tag{19}$$

6 EXPERIMENTS

To justify EliMRec's superiority and reveal the reasons for efficiency and effectiveness, we conduct systematic experiments to answer three research questions:

RQ1: Does EliMRec outperform the state-of-the-art models w.r.t Recall@10 for multi-modal recommendation?

RQ2: Does EliMRec capture the bias and eliminate single-modal bias?

RQ3: How do distinct settings influence the performance of EliMRec?

6.1 Experimental Settings

6.1.1 Datasets. To evaluate the efficiency of our devised model , we conduct experiments on three widely used datasets, including Tiktok [1], Kwai [2], Movielens [3]. These datasets contain not only user-item interaction records but also collect abundant multi-modal features. In general, all the multi-modal features have been extracted as feature vectors by the pre-training deep neural networks, including ResNet [9], VGGish [13], and Sentence2Vector [1]. The statistics of all three datasets are detailed in Table 1. Particularly, Kwai contains one content modality of visual while Movielens has more dense interaction and fewer items than others. All datasets are split into training, validation, and testing dataset with a ratio of 8:1:1 strictly following MMGCN's dataset splitting strategy [42]. The validation datasets are used to find the optimal hyperparameters to evaluate the performance in experiments.

6.1.2 Compared Methods. In addition to the primary matrix factorization (MF) and multimedia recommendation (*e.g.*, VBPR [10] and MMGCN [42]), to highlight the performance of EliMRec for multimedia content recommendation, we introduce the latest LightGCN [11] for its remarkable performance in recommendation. Moreover, RUBi [2] is involved for comparisons on removing unimodal bias. To keep a fair comparison, our experiments introduce identical multiple features as the input of aforementioned models. Specifically, EliMRec is compared with the following baselines:

Table 1: The statistics of the evaluation dataset, where V, A, and T denote the feature dimensions of visual, acoustic, and textual modalities, respectively.

Dataset	Inter.#	User.#	Item.#	Sparsity	V	A	T
Tiktok	726065	36651	76085	99.97%	256	128	128
Kwai	298492	7010	86483	99.95%	2048	-	-
Movielens	1507876	55485	5986	99.55%	2048	128	100

- **MF-BPR** [26]. This method models the user and item's representation as latent vectors according to the user-item historical interactions and then infers the user preference.
- **VBPR** [10]. Different from MF-BPR, this method integrates the visual features into the representation of the item in the original paper.
- **MMGCN** [42]. MMGCN is a well-devised model for multi-modal content recommendation. It constructs a parallel graph neural network for multiple modalities, therefore capturing the high-order connectivities and multi-model features.
- **LightGCN** [11]. This model is a state-of-the-art graph-based CF method. It simplifies the design of Graph Convolution Network (GCN) to improve representation learning's training efficiency and performance. In our experiment, we employ a variant, EliMRec's backbone, which follows a similar structure of MMGCN to fuse the multi-modal features to produce the final representation.
- **RUBi** [2]. RUBi aims at removing unimodal bias and mainly leverages two components: (1) an additional branch to capture unimodal bias and (2) gradient stopping to take the base modal away from the bias. In our experiments, we substitute its backbone as the above LightGCN and separately test each modality, namely RUBi-s where s is the specific modality.

6.1.3 Evaluation Metrics. In the evaluation phase, for each user, we view all items that the user has not interacted with as negative, while the interacted items as positive. For ranking strategy, we employ the full-rank strategy, rather than sampling a positive item and a set of negative items interacted by one or a few users. All these items are ranked based on the results of recommendation models' predictions. Moreover, we adopt Recall@K, Normalized Discounted Cumulative Gain (NDCG@K), and Precision@K as the metrics and set K=10 by default [42], where all the evaluation metrics are widely adopted in the recommendation systems.

6.1.4 Parameter Settings. In our experiments, we program all baselines and EliMRec based on Pytorch 1.6 [4] and torch-geometric package[5]. For fair comparisons, all models are initialized with Xavier [7] and optimized by Adam [14] with the same collaborative embedding dimension of 64 and mini-batch size of 2048. In terms of the hyperparameters, we use the grid search: the learning rate and regularization weight are searched in { 0.0001, 0.001, 0.01, 0.1 }, where the weight-decay parameter is used as the regularization weight.

[1] https://www.tiktok.com/

[2] https://www.kwai.com/

[3] https://movielens.org/

[4] https://pytorch.org/.

[5] https://pytorch-geometric.readthedocs.io/

Table 2: Performance Comparison between EliMRec andthe state-of-the-art recommendation algorithms on the threedatasets. The %Improv. denotes the relative performance im-provement over strongest baseline.

Dataset		Tiktok			Kwai			Movielens	
Metric	R@10	P@10	N@10	R@10	P@10	N@10	R@10	P@10	N@10
methods based on matrix factorization follow:									
MF-BPR	0.001083	0.000267	0.000628	0.000749	0.000357	0.000629	0.013803	0.003898	0.008647
VBPR	0.001003	0.000240	0.000690	0.002149	0.000884	0.001726	0.110427	0.025955	0.072833
methods based on graph neural network follow:									
MMGCN	0.052563	0.009406	0.029186	0.036389	0.010256	0.033443	0.128692	0.031347	0.083931
LightGCN	0.062143	0.010304	0.033669	0.042759	0.012468	0.037402	0.134463	0.033304	0.088419
methods based on the debias framework follow:									
RUBi-v	0.062515	0.010342	0.033695	0.041923	0.012011	0.036796	0.130187	0.032170	0.085413
RUBi-a	0.062763	0.010366	0.033773	-	-	-	0.130325	0.032268	0.085269
RUBi-t	0.062212	0.010290	0.033600	-	-	-	0.130593	0.032299	0.084950
EliMRec	**0.068290**	**0.010686**	**0.038779**	**0.045967**	**0.013053**	**0.038792**	**0.145055**	**0.034978**	**0.092776**
%Improv.	8.81%	3.09%	14.82%	7.50%	4.69%	3.72%	7.88%	5.03%	4.93%

For EliMRec, we tune α within the ranges of { 0, 0.25, 0.5, 0.75, 1, 2, 3, 4, 5 }. Moreover, we adopt the same early stopping strategy and the same number of GNN layers as LightGCN [11]. We do the same options for the baselines and follow their articles’ designs to achieve the best performance.

6.2 Performance Comparison

As shown in Table 2, we summarize all experiments’ results and report the improvements calculated between EliMRec and the best performance baselines highlighted with an underline. And we have the following observations:

- In most case, EliMRec outperforms the strongest baselines by a significant margin, presenting its rationality ans superiority and answering **RQ1** with causal inference and bias elimination.
- On most metrics, EliMRec achieves more promotion on the Tiktok dataset than others. It is reasonable since we only adopt visual features on the Kwai dataset, while the Movielens dataset is too dense with much fewer items. On the other hand, it demonstrates our devised model has better capability to distill high-level semantic information through multi-modal in the sparse space.
- LightGCN outperforms other baselines on Kwai and Movielens Datasets. Since the simplified GCN structure helps to exploit high-level information among the user-item interaction graph. And in the Tiktok dataset, RUBi-a with LightGCN as backbone also achieves better performance than other baselines, further demonstrating the capability of the redesigned multi-modal version of LightGCN.
- We noticed that RUBi such a debiased algorithm is only effective in Tiktok dataset and causes performance loss than LightGCN on the other datasets. On the recommendation task, we revisited RUBi’s model design and ascribe it to the overly harsh gradient stopping strategy that removes the model’s valid inductive bias.
- VBPR achieves a remarkable performance in Movielens compared with its performance in other datasets. We argue that the dense interaction provides rich supervised signals to train the preference vectors, thereby improving the performance in such simple design of VBPR.

Table 3: The comparison of TE (the bias capture effect) and TIE (the performance after inference.) w.r.t Recall@10.

Dataset	Tiktok	Kwai	Movielens
Baseline	0.062763	0.042759	0.134463
EliMRec-TE	0.0542342	0.039414	0.142186
EliMRec	**0.068290**	**0.045967**	**0.145055**

6.3 Ablation Study

Affected by the primary task, the gap in the experimental results is not significant in the previous section. Thus, we present the ablation study used to mine deep insights of EliMRec to address **RQ2**. Hence we calculate the total effect of the model with enhanced bias-capture ability, abbreviated as EliMRec-TE, to compare with the performance of strongest baselines and inference-equipped EliMRec. The only difference between EliMRec-TE and the original EliMRec is that EliMRec-TE applies the normal reasoning, whereas EliMRec uses the counterfactual analysis. As shown in Table 3, we get the following observations:

- The substantially improved performance of EliMRec compared to the TE case demonstrates the effectiveness of the inference phase, a key step in EliMRec, which eliminates single-modal bias by constructing several counterfactual worlds to answer the bias-only effects and subtract them from TE, further demonstrating the existence of the bias and our correctness of the proposed solution.
- We observe that EliMRec-TE outperforms the most strong baselines in Movielens but performs worse in the other two datasets. Obviously, the additional branches for enhancing the direct edge from the single-modal feature to the preference score are rational and useful to be distinct from the conventional multi-modal recommender. And in Movielens, a dataset with dense interaction, the single-model features

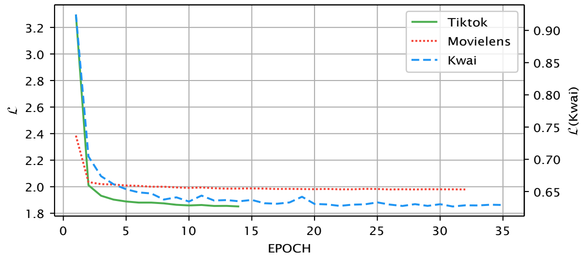


Figure 5: Training curves of EliMRec on three datasets.

Table 4: The comparison of different score fusions w.r.t Recall@10.

Dataset	Tiktok	Kwai	Movielens
EliMRec-RUBi	**0.068290**	**0.045967**	**0.145055**
EliMRec-HM	0.068045	0.040798	0.140060
EliMRec-SUM	0.067369	0.022523	0.055810

are easy to be uncovered and serve as an inductive bias to boost its performance. Whereas in other sparse datasets such as Tiktok, the single-modal bias is highlighted, thus causing a performance drop.

6.4 Study of EliMRec

In this section, we perform three perspectives to answer **RQ3**.

6.4.1 Training Efficiency. Figure 5 shows the loss convergence of EliMRec in three datasets, where the loss curve smoothes out, the model will gradually converge. In general, EliMRec has a high speed to converge the loss in a small value for the first few epochs and gradually continue to decrease slowly in each dataset. Specifically, EliMRec in Kwai is more slower to converge than other two datasets. The multiple modalities and dense interactions aid in quickly find the optimal parameters, stabilizing the curve, and cutting the training time.

6.4.2 Effect of score fusion. To shed light on the development of proper fusion strategies, we further investigate other score fusion methods for their impacts on EliMRec as shown in Table 4. In addtion to the used RUBi strategy (see Eq.17), Harmonic (HM) and SUM are included [19]. Formally,

$$\begin{aligned}\textbf{HM: } & \hat{y}_{hm} = \sigma(\hat{y}_{u,i}) \prod_{s\in\mathbb{S}/\{id\}} \sigma(\hat{y}^{s}_{u,i}), \hat{y}_{u,i,s} = \log \frac{\hat{y}_{hm}}{1+\hat{y}_{hm}}. \\ \textbf{SUM: } & \hat{y}_{u,i,s} = \log \sigma(\hat{y}_{u,i} + \sum_{s\in\mathbb{S}/\{id\}} \hat{y}^{s}_{u,i}).\end{aligned} \tag{20}$$

Obviously, HM and SUM strategies view $\hat{y}_{u,i}$ and $\hat{y}^{s}_{u,i}$ equivalent, thus applying the same transformation (*e.g.* σ for HM, identical for SUM) on each effect. We observe the relationship between these score fusion methods w.r.t Recall@10 as RUBi ≥ HM ≥ SUM, which indicates the multi-modal score dominated methods with nonlinear transformation helps to allocate the contribution of each effect.

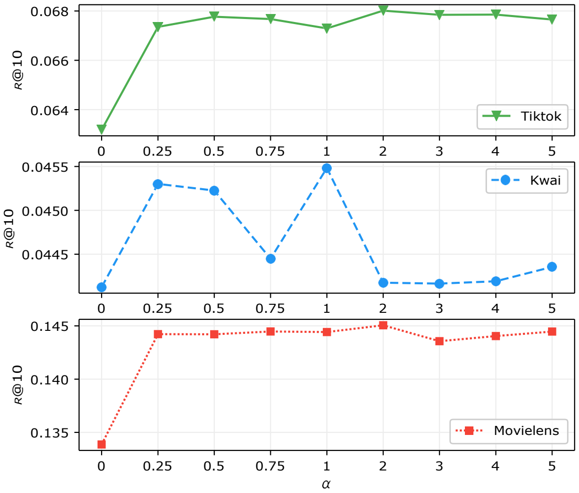


Figure 6: Impact of α $w.r.t.$ Recall@10.

6.4.3 Effect of α. Figure 6 shows performance comparisons evaluated by EliMRec under different α in Tiktok, Kwai, and Movielens. We find that the incorporation of the bias-capture branch significantly improves the performance. And in Tiktok and Movielen, such datasets with multiply modalities, the performance curve is stable when $\alpha \geq 0.25$, while the performance in Kwai with only one modality of visual draws an oscillating curve. It demonstrates that our proposed methods is insensitive with α when rich modalities are provided. However, in the case of only one modality, the performance of the model is significantly affected by that modality and becomes unstable, rather than being robust as the other datasets. From the observation, it is recommended to set α to 1 or 2 in any case.

7 CONCLUSION AND FUTURE WORK

In this work, we revealed the inner working of multi-modal fusion, thus pointing out the single-modal issue, which is widespread in multimedia recommendation, and devised EliMRec to eliminate the single-modal bias. Unlike removing a particular modal bias, our approach is a more general paradigm, which further shows the inclusivity of causal inference on multimedia recommendation and opens up new research possibilities. In the future, we will explore how to extend our framework to adapt recommendation tasks with other side information and alleviate their biases. More thorough causal graphs are encouraged to solve more challenges such as noisy data. Moreover, the exploration of causal inference on recommendation should be extended to a broad range (*e.g.* intervention and instrumental variable) to overcome various biases [3].

ACKNOWLEDGMENTS

The work is supported by the National Key Research and Development Program of China (No. 2021YFF0901705, No. 2021YFF0901700, No. 2020YFB1406800), the Fundamental Research Funds for the Central Universities (No. CUC22GZ003, No. CUC22GZ007, No. CUC22GZ035), and the GHfund B (20220202, No. ghfund202202023853).

REFERENCES

[1] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *ICLR 2017*.

[2] Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2020. RUBi: Reducing Unimodal Biases in Visual Question Answering. *arXiv:1906.10169 [cs]* (2020).

[3] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and Debias in Recommender System: A Survey and Future Directions. (2020).

[4] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. In *SIGIR*. 335–344.

[5] Konstantina Christakopoulou, Madeleine Traverse, Trevor Potter, Emma Marriott, Daniel Li, Chris Haulk, Ed H. Chi, and Minmin Chen. 2020. *Deconfounding User Satisfaction Estimation from Response Rate Bias.* Association for Computing Machinery, 450–455.

[6] Dan Geiger, Thomas Verma, and Judea Pearl. 1990. d-separation: From theorems to algorithms. In *Machine Intelligence and Pattern Recognition*. Vol. 10. Elsevier, 139–148.

[7] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010 (JMLR Proceedings, Vol. 9)*. JMLR.org, 249–256.

[8] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer.* John Wiley & Sons.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778.

[10] Ruining He and Julian J. McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *AAAI*. 144–150.

[11] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM conference on research and development in Information Retrieval, Virtual Event, China, July 25-30, 2020*. ACM, 639–648.

[12] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, April 3-7, 2017*. ACM, 173–182.

[13] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, 2017*. 131–135.

[14] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[15] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, Ying Li, Bing Liu, and Sunita Sarawagi (Eds.). ACM, 426–434.

[16] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat seng Chua. 2022. Invariant Grounding for Video Question Answering. In *CVPR*.

[17] Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2020. A general knowledge distillation framework for counterfactual recommendation via uniform data. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 831–840.

[18] Fan Liu, Zhiyong Cheng, Changchang Sun, Yinglong Wang, Liqiang Nie, and Mohan S. Kankanhalli. 2019. User Diverse Preference Modeling by Multimodal Attentive Metric Learning. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019*. ACM, 1526–1534.

[19] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual VQA: A Cause-Effect Look at Language Bias. *arXiv:2006.04315 [cs]* (2021).

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

[21] Judea Pearl. 2010. Causal inference. *Causality: Objectives and Assessment* (2010), 39–58.

[22] Judea Pearl. 2013. Direct and indirect effects. *arXiv preprint arXiv:1301.2300* (2013).

[23] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect.* Basic books.

[24] Zhen Qin, Suming J Chen, Donald Metzler, Yongwoo Noh, Jingzheng Qin, and Xuanhui Wang. 2020. Attribute-based propensity for unbiased learning in recommender systems: Algorithm and case studies. In *SIGKDD*.

[25] Steffen Rendle. 2010. Factorization Machines. In *The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, Geoffrey I. Webb, Bing Liu, Chengqi Zhang, Dimitrios Gunopulos, and Xindong Wu (Eds.). IEEE Computer Society, 995–1000.

[26] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. (2009), 10.

[27] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016 (JMLR Workshop and Conference Proceedings, Vol. 48)*. JMLR.org, 1670–1679.

[28] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. 2021. *Counterfactual Explainable Recommendation.* 1784–1793.

[29] Kaihua Tang, Mingyuan Tao, and Hanwang Zhang. 2021. Adversarial Visual Robustness by Causal Intervention. *arXiv preprint arXiv:2106.09534* (2021).

[30] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised Learning for Multimedia Recommendation. *IEEE Transactions on Multimedia* (2022). https://doi.org/10.1109/TMM.2022.3187556

[31] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. 2020. MGAT: Multimodal Graph Attention Network for Recommendation. *Inf. Process. Manag.* 57, 5 (2020), 102277.

[32] Wenjie Wang, Fuli Feng, Xiangnan He, Xiang Wang, and Tat-Seng Chua. 2021. Deconfounded Recommendation for Alleviating Bias Amplification. *SIGKDD* (2021), 1717–1725.

[33] Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be Cheating: Counterfactual Recommendation for Mitigating Clickbait Issue. *SIGIR* (2021), 1288–1297.

[34] Xiang Wang, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. 2017. Item Silk Road: Recommending Items from Information Domains to Social Users. In *Proceedings of the 40th International ACM Conference on Research and Development in Information Retrieval, August 7-11, 2017*. ACM, 185–194.

[35] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural Graph Collaborative Filtering. In *Proceedings of the 42nd International ACM Conference on Research and Development in Information Retrieval, Paris, France, July 21-25, 2019*. ACM, 165–174.

[36] Xiang Wang, Yingxin Wu, An Zhang, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2022. Reinforced Causal Explainer for Graph Neural Networks. *TPAMI* (2022).

[37] Xiang Wang, Yingxin Wu, An Zhang, Xiangnan He, and Tat seng Chua. 2021. Towards Multi-Grained Explainability for Graph Neural Networks. In *NeurIPS*.

[38] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2021. Combating Selection Biases in Recommender Systems with a Few Unbiased Ratings. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 427–435.

[39] 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. *ACM SIGKDD* (2021), 1791–1800.

[40] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive learning for cold-start recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5382–5390.

[41] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*. 3541–3549.

[42] 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 *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 1437–1445.

[43] Le Wu, Xiangnan He, Xiang Wang, Kun Zhang, and Meng Wang. 2021. A Survey on Neural Recommendation: From Collaborative Filtering to Information-rich Recommendation. *arXiv:2104.13030 [cs]* (Oct. 2021).

[44] Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat seng Chua. 2022. Discovering Invariant Rationales for Graph Neural Networks. In *ICLR*.

[45] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. 2021. Causal Attention for Vision-Language Tasks. *arXiv:2103.03493 [cs]* (2021).

[46] Jingwei Yi, Fangzhao Wu, Chuhan Wu, Qifei Li, Guangzhong Sun, and Xing Xie. 2021. DebiasedRec: Bias-aware User Modeling and Click Prediction for Personalized News Recommendation. *arXiv:2104.07360 [cs]* (2021).

[47] Zhongqi Yue, Tan Wang, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. 2021. Counterfactual Zero-Shot and Open-Set Visual Recognition. *arXiv:2103.00887 [cs]* (2021).

[48] Yongfeng Zhang, Qingyao Ai, Xu Chen, and W. Bruce Croft. 2017. Joint Representation Learning for Top-N Recommendation with Heterogeneous Information Sources. In *CIKM 2017*. ACM, 1449–1458.

[49] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal Intervention for Leveraging Popularity Bias in Recommendation. *SIGIR* (2021), 11–20.

[50] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open* 1 (2020), 57–81.