Enhancing Adversarial Robustness of Multi-modal Recommendation via Modality Balancing

Yu Shang
Department of Electronic Engineering, Tsinghua University
Beijing, China
shangy21@mails.tsinghua.edu.cn

Chen Gao
Department of Electronic Engineering, Tsinghua University
Beijing, China
chgao96@gmail.com

Jiansheng Chen*
University of Science and Technology Beijing
Beijing, China
jschen@ustb.edu.cn

Depeng Jin
Department of Electronic Engineering, Tsinghua University
Beijing, China
jindp@tsinghua.edu.cn

Huimin Ma
University of Science and Technology Beijing
Beijing, China
mhmpub@ustb.edu.cn

Yong Li
Department of Electronic Engineering, Tsinghua University
Beijing, China
liyong07@tsinghua.edu.cn

ABSTRACT

Recently multi-modal recommender systems have been widely applied in real scenarios such as e-commerce businesses. Existing multi-modal recommendation methods exploit the multi-modal content of items as auxiliary information and fuse them to boost performance. Despite the superior performance achieved by multi-modal recommendation models, there's currently no understanding of their robustness to adversarial attacks. In this work, we first identify the vulnerability of existing multi-modal recommendation models. Next, we show the key reason for such vulnerability is *modality imbalance*, *i.e.*, the prediction score margin between positive and negative samples in the sensitive modality will drop dramatically facing adversarial attacks and fail to be compensated by other modalities. Finally, based on this finding we propose a novel defense method to enhance the robustness of multi-modal recommendation models through modality balancing. Specifically, we first adopt an embedding distillation to obtain a pair of content-similar but prediction-different item embeddings in the sensitive modality and calculate the score margin reflecting the modality vulnerability. Then we optimize the model to utilize the score margin between positive and negative samples in other modalities to compensate for the vulnerability. The proposed method can serve as a plug-and-play module and is flexible to be applied to a wide range of multi-modal recommendation models. Extensive experiments on two real-world datasets demonstrate that our method significantly improves the robustness of multi-modal recommendation models with nearly no performance degradation on clean data.

*The corresponding authors.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612337

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Multi-modal Recommendation, Adversarial Robustness, Modality Balancing

ACM Reference Format:
Yu Shang, Chen Gao, Jiansheng Chen, Depeng Jin, Huimin Ma, and Yong Li. 2023. Enhancing Adversarial Robustness of Multi-modal Recommendation via Modality Balancing. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612337

1 INTRODUCTION

Recent years have witnessed the wide application of multi-modal recommendation in many real-world scenarios such as micro-video platforms[23, 24, 32] and e-commerce businesses[4, 22, 27]. Different from traditional recommendation mainly utilizing historical interactions to predict user preference[16, 19, 36], multi-modal recommendation methods introduce rich item-side multi-modal content information (*i.e.*, visual, acoustic, and textual features) to gain more informative user and item representations[35, 38, 39, 44]. Existing multi-modal recommendation methods mainly concentrate on the fusion strategies to get the multi-modal feature and how to integrate multi-modal information with the recommendation framework. Up to now, there have been many efficient methods achieving remarkable recommendation performance, among which the supervised-learning methods incorporating Graph Neural Networks (GNNs) show great success in modeling interactions between users and items with multi-modal information[2, 35, 38, 39]. Besides, there have been self-supervised learning approaches proposed to learn user and item representations by exploring the underlying relations between different modalities[31, 41, 46].

Nevertheless, existing methods including both supervision manners commonly concentrate on how to utilize different modality information to enhance recommendation performance, while paying less attention to their adversarial robustness, *i.e.,* whether multi-modal recommendation models can be easily fooled by slight perturbations of the multi-modal input content. Investigation into this issue is imperative and inspiring to develop more reliable real-world recommender systems. Although the robustness of recommender systems is a widely discussed topic, most works focus on studying the robustness of collaborative filtering-based recommendation models[5, 7, 34] and visual content-based models[1, 25, 29]. So far there's no understanding of the adversarial robustness of multi-modal recommendation models, which is actually unforeseeable due to the complex relation between different modalities.

To tackle this problem, in this work, we first identify the vulnerability of existing multi-modal recommendation models by injecting slight adversarial perturbations into the multi-modal input features. The attacking results demonstrate that multi-modal recommendation models without any defense will suffer a great performance decrease under attack. Next, we explore the reason behind the vulnerability of these models and reveal that the imbalance of score margin from different modalities under attacks is the key reason for the vulnerability. Specifically, we find that the score margin between positive and negative samples in certain modality (*e.g.,* the visual modality) will drop fiercely, showing high sensitivity to the adversarial attacks. By comparison, some other modalities are less sensitive to the attack (*e.g.,* the textual modality), however, the limited score margin in these modalities can not compensate for the large score drop in the sensitive modality, which we call **modality imbalance**. Therefore, although multi-modal recommendation models possess rich modality information to make predictions, they still fail to defend against adversarial attacks due to the issue of modality imbalance. Finally, based on our finding we propose a novel defense method to enhance the robustness of multi-modal recommendation models through modality balancing. To be specific, we first conduct an embedding distillation to obtain a pair of content- similar but prediction-different item embeddings in the sensitive modality and calculate the score margin which reflects the modality vulnerability. Then we optimize the model to enlarge the score margin between positive and negative samples in other modalities to compensate for the margin mentioned above, thus achieving modality balancing. In this way, we can obtain a robust and performance-maintained model, superior to conventional adversarial training methods which improve adversarial robustness but sacrifice much more clean performance. Furthermore, as a plug-and-play module, the proposed defense method is model-agnostic and flexible to be applied to the mainstream multi-modal recommendation models. To sum up, the contributions of this work can be summarized as follows:

- We reveal the key reason for the vulnerability of multi-modal recommendation models as modality imbalance, and systematically evaluate the adversarial robustness of the mainstream multi-modal recommendation models.
- We propose a novel defense method enhancing the robustness of multi-modal recommendation models through modality balancing. The method is model-agnostic and convenient to be applied to the mainstream multi-modal recommendation models.
- Extensive experiments on two real-world datasets verify the effectiveness of our method in boosting model robustness without sacrificing the performance on clean test data.

2 RELATED WORK

2.1 Multi-modal Recommendation

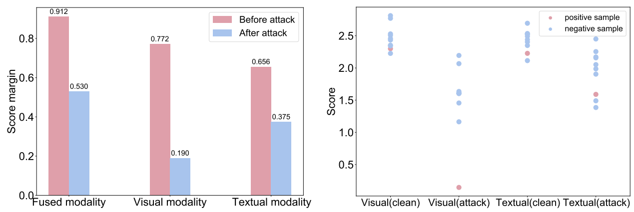
The multi-modal recommendation methods incorporate multi-modal features with traditional collaborative filtering signals to learn better representations of users and items. Previous works such as VBPR[15] utilize Matrix Factorization to deal with the combination of multi-modal information and id embeddings. With the rapid development of deep learning, many techniques are integrated into the multi-modal recommendation models, such as Variational Autoencoder[33, 43] and Graph Neural Networks[6, 9, 10, 37]. For example, ADDVAE[33] exploits the disentangled representations to learn better user preference which might be influenced by different hidden factors. MMGCN[39] firstly utilizes the Graph Convolutional Network to learn the representation in each modality and then fuse multi-modal representations with the id embedding to obtain the final representation. In addition to the supervised learning methods above, self-supervised learning approaches are explored to further enhance supervision signals[31, 41]. For example, SLMRec[31] introduces self-supervised learning tasks such as feature drop and feature masking to generate different views of items and utilizes contrastive learning in the model training. Existing multi-modal methods commonly concentrate on better utilization of multi-modal information but lack consideration of the model robustness.

2.2 Robustness of Recommender Systems

Recommender systems can be vulnerable when the model inputs (*e.g.* user profile and item content) are injected with the hand-engineered[13, 20] or automatically optimized perturbations[8, 21, 25, 30]. For example, Tang *et al.*[30] design an effective transfer-based poisoning attack against recommender systems by injecting fake user behaviors into the inputs. Accordingly, there are many works aiming to defend recommender systems against adversarial attacks[3, 17, 29, 42]. For example, He *et al.* propose to improve the robustness of the BPR method by conducting adversarial training. Tang *et al.*[29] firstly concentrate on the robustness of multimedia recommender systems facing untargeted adversarial examples and introduce adversarial training to enhance model robustness, but only visual modality is used in the experiments. Besides, it only considers some simple methods such as VBPR[15] and MF-BPR[28], not covering advanced multi-modal recommendation methods. To sum up, existing defense methods are mainly based on adversarial training, which produces perturbations against the model itself and forces the model to defend them. However, adversarial training often leads to obvious performance drops on clean testing data. By comparison, our defense method could improve model robustness and maintain the normal performance on clean data simultaneously.

3 ADVERSARIAL VULNERABILITY ANALYSIS

In this section, we first describe how we generate the adversarial perturbations for the multi-modal input features in Section 3.1, and the results indicate that the multi-modal recommendation models are vulnerable to adversarial attacks. In Section 3.2, we present a fine-grained analysis of the reason why these models fail and



(a) Score margin change under attacks (b) A real case showing the score change of different modalities

Figure 1: The illustrative example of adversarial attacks against multi-modal recommendation models (results from GRCN). (a) Prediction score margin (between positive and negative samples) before and after attack for different modalities. (b) A real example from Baby dataset showing the prediction score change of different modalities under attacks (only the top 10 negative samples are shown).

reveal the key lies in the imbalance of score margin from different modalities under attacks.

3.1 Adversarial Attack Method

Let $f : (X, \mathcal{D}) \rightarrow y$ denote a multi-modal recommendation model, where $X = [x_v, x_t]$ represents input features of visual and textual modalities (here assuming that there are only these two modalities without loss of generality), $\mathcal{D} = \{(u, i, j) | u \in \mathcal{U}, i \in \mathcal{I}_u^+, j \in \mathcal{I} \setminus \mathcal{I}_u^+\}$ denotes all pairwise training instances, $\mathcal{U}$, $\mathcal{I}$, and $\mathcal{I}_u^+$ denote all users, items and the interacted items of user u. The goal of the adversarial attack is to decrease the model's overall test performance as much as possible. Considering almost all multi-modal recommendation models take the content features as inputs, it's reasonable and convenient to apply perturbations $\Delta = [\Delta_v, \Delta_t]$ to the input multi-modal features X at test time to conduct the attack:

$$\hat{y}_{ui}^{'} = f(X + \Delta, \mathcal{D}). \tag{1}$$

In order to degrade the recommendation performance, we choose to maximize BPR loss[28] as the optimizing objective to generate adversarial perturbations for the input feature of each modality m:

$$\Delta_m^* = \arg\max_{\Delta_m} \mathcal{L}_{BPR} = \arg\max_{\Delta_m} \sum_{(u,i,j) \in \mathcal{D}_{test}} -ln(\sigma(\hat{y}_{ui}^{'} - \hat{y}_{uj}^{'})), \quad \text{where} \quad ||\Delta_m^*|| \leq \epsilon_m, \tag{2}$$

where $|| \cdot ||$ denotes L2-norm, $\mathcal{D}_{test}$ denotes all pair-wise test instances, ϵ_m is the magnitude of perturbations for modality m.

Here we borrow the idea of FGSM[12] attack to generate the adversarial perturbations. We can obtain the solution for adversarial perturbations as follows:

$$\Delta_m = \epsilon_m \frac{\Gamma_m}{||\Gamma_m||}, \quad \text{where} \quad \Gamma_m = \frac{\partial \mathcal{L}_{BPR}}{\partial \Delta_m}. \tag{3}$$

Incidentally, we also try the original attack method in [12], which only keeps the sign of the derivation, *i.e.,* $\Delta_m = \epsilon_m sign(\Gamma_m)$. However, we find it less effective than our solution on multi-modal recommendation models. As a result, we finally choose Eq. (3) to generate perturbations in our experiments.

3.2 Cause of Adversarial Vulnerability: Modality Imbalance

We conduct extensive robustness evaluation using the above attack method on existing five mainstream multi-modal recommendation models including both supervised learning and self-supervised learning methods: VBPR[15], MMGCN[39], GRCN[38], SLMRec[31] and MMGCL[41]. All models show severe performance degradation with slight perturbations whose norm is no more than 5% of the original input feature (*i.e.,* $\epsilon_m = 0.05 * ||x_m||$). For example, Recall@20 drops from 0.632 to 0.300, and NDCG@20 drops from 0.0265 to 0.0114 for MMGCN on Baby dataset (the complete results are shown in Table 2), showing severe adversarial vulnerability.

Aiming to reveal the reason why these models fail under attack, we try to make an exploration of the change inside the models caused by attacks. Since the performance degradation is closely related to the score ranking between positive and negative samples, we first analyze the change of score margin between positive and negative samples under attacks. Considering multi-modal recommendation models fuse prediction scores from different modalities to make the final prediction (*e.g.,* taking score addition when using embedding concatenation as the modality fusion strategy), we analyze the score margin in each single modality and fused modality (results from GRCN) in Figure 1(a). The result shows that the score margin of visual modality drops more fiercely than the textual modality, indicating higher sensitivity of the visual modality. It can be imagined that the severe disruption in visual prediction scores will overcome the correct prediction from the textual modality and lead to the failure of the model. We further verify this issue through a real case from Baby dataset, in which (u_{841}, i_{916}) is an observed interaction. For the user u_{841}, the visual prediction score of the positive sample i_{916} drops from 2.299 to 0.145, and the textual prediction score drops from 2.227 to 1.590 under attack, as shown in Figure 1(b). Although the score of i_{916} in textual modality only changes a little, its slight strength over negative samples can not compensate for the large score drop in the visual modality, causing the performance degradation. We name this kind of mismatch between prediction scores of different modalities under adversarial attacks as **modality imbalance** and reveal it as a critical defect of the adversarial robustness of multi-modal recommendation models.

4 THE PROPOSED DEFENSE METHOD

In this section, we will depict our proposed defense method against multi-modal adversarial attacks. Based on the previous finding of modality imbalance in existing multi-modal recommendation models under attack, we introduce a novel defense method called modality balancing. The overall idea is to compensate for the drastic score drop of positive samples in the sensitive modality by enlarging the score strength of positive samples in the insensitive modality, thus keeping correct predictions. The overall framework is illustrated in Figure 2, in which the first line shows the general pipeline of multi-modal recommendation models, and the second line depicts our defense treatments including two modules: sensitive-modal embedding distillation and modality balancing. Note that here we present our general framework based on Baby dataset consistent with the previous analysis, *i.e.,* regarding visual modality as the sensitive modality and textual modality as the insensitive modality.

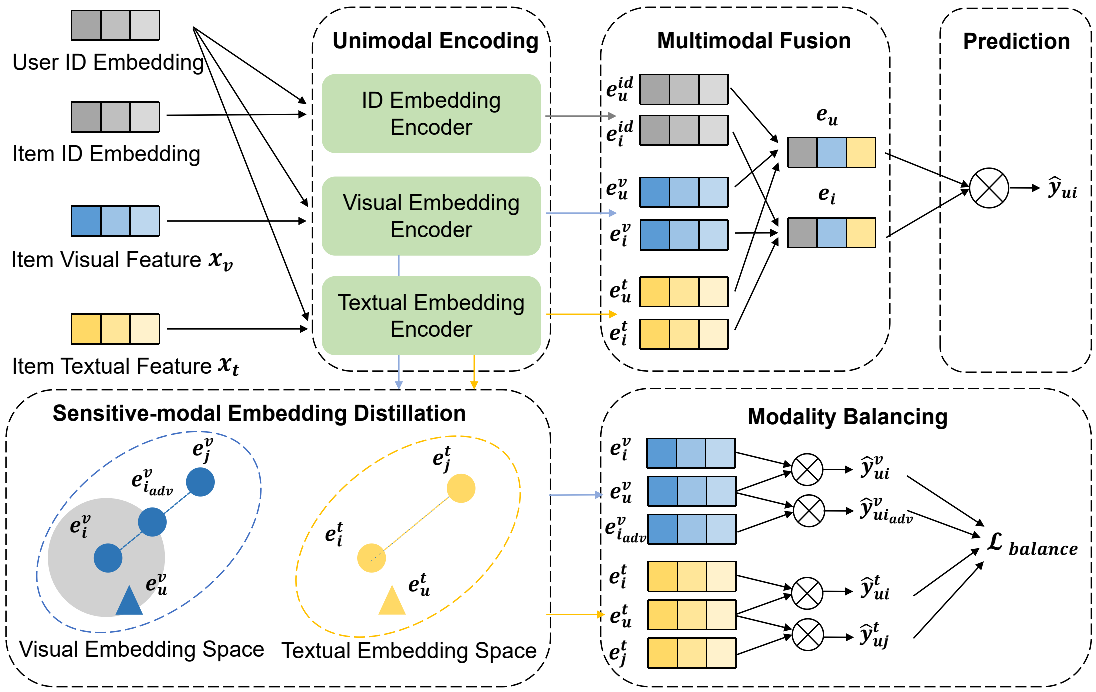


Figure 2: Illustration of the proposed defense framework. The first line describes the general pipeline of multi-modal recommendation models, consisting of three parts: unimodal encoding, multi-modal fusion and prediction. The second line depicts our plug-and-play defense module including two steps: sensitive-modal embedding distillation and modality balancing.

4.1 Sensitive-modal Embedding Distillation

The robustness evaluation in Section 3 reveals that certain input content features (*e.g.,* visual features in Baby dataset) are highly sensitive to additive noise, which is the main cause of adversarial vulnerability of multi-modal recommendation models. Considering the observation that only slight perturbations can destroy the prediction results, it is expected that the intermediate embeddings are sensitive to slight changes, *i.e.,* similar embeddings will lead to totally distinct prediction results. Here we use embedding distillation to get pair-wise content-similar but prediction-different embeddings in the sensitive modality and use the prediction difference to reflect the modality vulnerability.

Formally, for a training instance (u, i, j), let $\boldsymbol{e}_u^v = g_u^v(x_v), \boldsymbol{e}_i^v = g_i^v(x_v), \boldsymbol{e}_j^v = g_j^v(x_v)$ be the visual embedding of user u, positive item i and negative item j obtained from the visual embedding encoder $g^v(\cdot)$ (*e.g.,* GNN). We will distill to get an adversarial embedding $\boldsymbol{e}_{i_{adv}}^v$ which is close to $\boldsymbol{e}_i^v$ in the embedding space but shifts towards the embedding of the negative item $\boldsymbol{e}_j^v$ as follows:

$$x_v^{'} = \arg\min_z ||g_i^v(z) - g_j^v(x_v)||^2, \quad s.t. \quad ||z - x_v||_\infty \leq \epsilon,$$
$$\boldsymbol{e}_{i_{adv}}^v = g_i^v(x_v^{'}), \tag{4}$$

where $x_v^{'}$ is the optimized input visual feature to generate the adversarial embedding, ϵ denotes the constraint of the input feature norm change. This constrained optimization objective can then be optimized using PGD[26] algorithm. In this way, the visual embedding encoder will produce an adversarial embedding $\boldsymbol{e}_{i_{adv}}^v$ for the positive item i which has similar content with the original visual embedding $\boldsymbol{e}_i^v$ but its prediction score will shift towards the negative item j for the user u. Such misalignment between the visual content similarity and the prediction result shows the vulnerability of visual modality.

4.2 Alleviating Vulnerability through Modality Balancing

According to the analysis in Section 3, the prediction results in the sensitive modality exhibit severe vulnerability (*i.e.,* the large score drop for positive samples), overcoming the correct results in other modalities. This motivates us to enlarge the strength in the insensitive modality, thus allowing for the severe disruption in the sensitive modality. Specifically, we try to guide the score margin between positive and negative samples in the insensitive modality (*i.e.,* textual modality) to become large enough to compensate for the vulnerability in the sensitive modality (*i.e.,* visual modality). For each sampled training instance (u, i, j), we use an auxiliary loss function $\mathcal{L}_{balance}$ to balance the score margin between positive and negative samples in the insensitive modality and the score margin

Algorithm 1 Modality-balancing training algorithm (here regarding visual modality as the sensitive modality).

for $b = 1, 2, ..., batch\_number$ **do**
  $sample\_number = 0, \mathcal{L} = \mathcal{L}_{BPR}$
  while $sample\_number < N$ **do**
    $(u, i, j) \leftarrow$ Training instance random sampling
    $\boldsymbol{e}_u^v, \boldsymbol{e}_i^v, \boldsymbol{e}_j^v, \boldsymbol{e}_i^t, \boldsymbol{e}_j^t \leftarrow$ Embedding lookup in different modalities
    $\boldsymbol{e}_{i_{adv}}^v \leftarrow$ Visual embedding distillation according to Eq. (4)
    Calculate $\mathcal{L}_{balance}$ according to Eq. (5)
    $\mathcal{L} = \mathcal{L} + \lambda \mathcal{L}_{balance}$
    $sample\_number = sample\_number + 1$
  end while
  Update model parameters according to loss function $\mathcal{L}$
end for

between the distilled adversarial embedding and clean embedding:

$$s_{margin}^t = \boldsymbol{e}_u^{t\top} \boldsymbol{e}_i^t - \boldsymbol{e}_u^{t\top} \boldsymbol{e}_j^t,$$
$$s_{margin}^v = \boldsymbol{e}_u^{v\top} \boldsymbol{e}_i^v - \boldsymbol{e}_u^{v\top} \boldsymbol{e}_{i_{adv}}^v,$$
$$\mathcal{L}_{balance} = max(s_{margin}^v - s_{margin}^t, 0), \tag{5}$$

where $\boldsymbol{e}_u^t = g_u^t(x_t), \boldsymbol{e}_i^t = g_i^t(x_t), \boldsymbol{e}_j^t = g_j^t(x_t)$ are the textual embedding of user u, positive item i and negative item j obtained from the textual embedding encoder $g^t(\cdot)$. Minimizing this objective will enlarge the score margin in the textual modality toward being balanced with the score fluctuation in the sensitive visual modality when the textual score margin is smaller. In this situation, although the visual prediction score drops violently for positive samples, the strength in the textual prediction score can still ensure positive samples rank ahead of negative samples. We combine BPR loss with the balance loss through a coefficient λ to form the final loss function used for model training. Besides, it's worth noting that our defense method is established on the general pipeline of multi-modal recommendation models and is flexible to be applied to existing mainstream models. The overall process of fusing our defense framework with multi-modal recommendation model training is summarized in Algorithm 1.

4.3 Time Complexity Analysis

Here we analyze the complexity of our proposed modality balancing method and comparison with adversarial training. Let O_f denote the time complexity of forward propagation for prediction scores, and O_b denotes the time complexity of backward propagation. The recommendation model itself will cost $O_f + O_b$. Modality balancing will introduce an extra cost due to the generation of adversarial embedding through a t-step PGD algorithm and another forward propagation, whose cost is $t\alpha(O_f + O_b) + O_f$. The coefficient $\alpha (< 1)$ exists because the forward propagation is end with embedding encoders, with no need to run the whole model. By comparison, adversarial training requires running the whole model to obtain adversarial features, whose time complexity is $t(O_f + O_b) + O_f$. Therefore, our proposed method has lower time complexity than adversarial training under the same number of sampling instances.

Table 1: Statistics of the two experimental datasets.

Dataset	#Users	#Items	#Interactions	Sparsity
Baby	19,445	7,050	160,792	99.88%
Clothing	39,387	23,033	278,677	99.97%

5 EXPERIMENTS

To justify the superiority of our proposed modality balancing method and reveal the reasons for effectively improving robustness, we conduct extensive experiments to answer three research questions:

- **RQ1:** Does modality balancing outperform the existing adversarial defense methods on multi-modal recommendation models?
- **RQ2:** How do different settings influence the effectiveness of modality balancing?
- **RQ3:** Does modality balancing effectively address the score imbalance between modalities under attacks?

5.1 Experimental Settings

5.1.1 Datasets. We use Amazon review[14] dataset for our experimental evaluation. In this public dataset, both product descriptions and corresponding images are available. We select two per-category datasets[1], *i.e.,* Baby and Clothing to conduct experiments, which are widely used in previous works[15, 29, 44, 46]. The details of the two datasets are presented in Table 1. The two datasets include both visual and textual content, specifically, the 4,096-dimensional visual features and 384-dimensional textual features. The interaction history of each user is randomly split into training, validation and testing datasets with the ratio 8:1:1 following [35, 44, 46].

5.1.2 Compared Methods. Existing works commonly improve the adversarial robustness of recommendation models through adversarial training[17, 29, 40, 42]. Here we follow the implementation of [29] which utilizes adversarial training to improve the robustness of visual content-based recommendation models as an important baseline, dubbed **Adv training**. Besides, considering directly dropping the input of certain modalities (especially the sensitive modality) will also help improve model robustness, we take each unimodal feature as input respectively and form two baseline models, dubbed **Unimodal-visual** and **Unimodal-textual**. The full multi-modal model without any defense method is dubbed **Multi-modal**.

5.1.3 Evaluation Metrics. We regard all items that the user has not interacted with as negative samples, and the interacted items as positive samples. Then we employ the full-rank strategy based on the prediction scores of recommendation models. Moreover, we adopt Recall@K and Normalized Discounted Cumulative Gain (NDCG@K) as the metrics and set K = 10, 20, which are widely used in the research of recommendation[44, 45].

5.1.4 Implementation Details. For all models we fix the embedding size to 64 for all models following existing works[38, 44, 46], initialize the model parameters with the Xavier[11] method and use Adam[18] as the optimizer. We carefully tune the learning rate, regularization weight and other parameters following the original

[1]Datasets are available at http://jmcauley.ucsd.edu/data/amazon/links.html

Table 2: Results of five models under two scenarios (clean and attack) on Baby dataset. The best and second-best results in each scenario are highlighted in bold and underline, respectively.

Baby		Clean				Attack			
Model	Method	Recall@10	Recall@20	NDCG@10	NDCG@20	Recall@10	Recall@20	NDCG@10	NDCG@20
MMGCN	Multimodal	**0.0389**	**0.0632**	**0.0203**	**0.0265**	0.0163	0.0300	0.0079	0.0114
	Unimodal-visual	0.0341	0.0548	0.0174	0.0229	0.0022	0.0034	0.0011	0.0014
	Unimodal-textual	0.0352	0.0589	0.0182	0.0243	<u>0.0246</u>	<u>0.0423</u>	0.0115	0.0160
	Adv training	0.0331	0.0551	0.0167	0.0224	0.0214	0.0332	<u>0.0122</u>	<u>0.0168</u>
	Modality balancing	<u>0.0379</u>	<u>0.0618</u>	<u>0.0199</u>	<u>0.0257</u>	**0.0326**	**0.0514**	**0.0171**	**0.0223**
VBPR	Multimodal	**0.0418**	**0.0664**	**0.0223**	**0.0287**	0.0135	0.0246	0.0064	0.0093
	Unimodal-visual	0.0388	0.0619	0.0206	0.0263	0.0138	0.0256	0.0067	0.0097
	Unimodal-textual	0.0394	0.0622	0.0210	0.0271	0.0079	0.0164	0.0035	0.0057
	Adv training	0.0383	0.0605	0.0204	0.0269	<u>0.0207</u>	<u>0.0357</u>	<u>0.0121</u>	<u>0.0165</u>
	Modality balancing	<u>0.0401</u>	<u>0.0628</u>	<u>0.0215</u>	<u>0.0280</u>	**0.0295**	**0.0502**	**0.0148**	**0.0201**
GRCN	Multimodal	**0.0543**	**0.0854**	**0.0295**	**0.0375**	0.0301	0.0510	0.0151	0.0204
	Unimodal-visual	0.0489	0.0785	0.0268	0.0344	0.0315	0.0523	0.0167	0.0220
	Unimodal-textual	0.0505	0.0805	0.0269	0.0346	<u>0.0332</u>	<u>0.0545</u>	<u>0.0173</u>	<u>0.0233</u>
	Adv training	0.0502	0.0777	0.0267	0.0335	0.0319	0.0538	0.0170	0.0224
	Modality balancing	<u>0.0515</u>	<u>0.0822</u>	<u>0.0274</u>	<u>0.0352</u>	**0.0354**	**0.0602**	**0.0175**	**0.0239**
SLMRec	Multimodal	**0.0507**	**0.0745**	**0.0282**	**0.0341**	0.0243	0.0378	0.0125	0.0159
	Unimodal-visual	0.0427	0.0653	0.0230	0.0288	0.0217	0.0355	0.0113	0.0149
	Unimodal-textual	0.0492	0.0718	0.0265	0.0330	0.0355	<u>0.0547</u>	<u>0.0193</u>	<u>0.0242</u>
	Adv training	0.0491	0.0721	0.0271	0.0331	<u>0.0356</u>	0.0533	0.0186	0.0232
	Modality balancing	<u>0.0503</u>	<u>0.0736</u>	<u>0.0275</u>	<u>0.0335</u>	**0.0389**	**0.0590**	**0.0211**	**0.0262**
MMGCL	Multimodal	**0.0529**	**0.0801**	**0.0297**	**0.0367**	0.0338	0.0510	0.0179	0.0223
	Unimodal-visual	0.0435	0.0677	0.0243	0.0305	0.0381	0.0596	0.0215	0.0268
	Unimodal-textual	0.0456	0.0723	0.0233	0.0299	0.0406	0.0628	<u>0.0218</u>	<u>0.0278</u>
	Adv training	0.0474	0.0735	0.0253	0.032	<u>0.0410</u>	<u>0.0637</u>	0.0215	0.0274
	Modality balancing	<u>0.0518</u>	<u>0.0788</u>	<u>0.0284</u>	<u>0.0349</u>	**0.0468**	**0.0692**	**0.0252**	**0.0311**

papers. In the adversarial attack phase, we set the maximum perturbation magnitude ϵ_m as 5% of the input feature norm for modality m. As for our defense method, we set the constraint of feature change ϵ when generating the adversarial embedding as 1, the steps t of PGD algorithm is set as 10, the coefficient λ controlling the ratio of two loss terms is searched in {0.001, 0.01, 0.1, 1, 10}. The number of training instances sampled for defense methods in each batch N is searched in {10, 20, 30, 50, 100}.

5.2 Performance Comparison (RQ1)

We conduct a systematic evaluation on the performance of different defense methods on two scenarios (clean and attack) on five mainstream multi-modal recommendation models including supervised methods (VBPR[15], MMGCN[39], GRCN[38]) and self-supervised methods (SLMRec[31], MMGCL[41]). The models will take original multi-modal features as input in the **clean** scenario and perturbed multi-modal features as input in the **attack** scenario. The results on Baby and Clothing dataset are shown in Table 2 and Table 3, respectively. From the results we have the following observations:

- **Our modality balancing method almost consistently achieves the highest adversarial robustness under attacks compared with other defense methods.** It can be observed that modality balancing gets superior defense performance compared with widely-used adversarial training and unimodal models in the attack scenario. Statistically, on Baby dataset, our modality balancing gets 21.01% improvement on Recall@10, 17.82% on Recall@20, 17.71% on NDCG@10 and 15.45% on NDCG@20 compared with the best baseline. On Clothing dataset, our method achieves 9.38% improvement on Recall@10, 8.97% on Recall@20, 8.89% on NDCG@10 and 9.01% on NDCG@20 compared with the best baseline. As for other defense methods, adversarial training can also improve model robustness under attacks, but it will cause greater training difficulty because it produces stronger adversary (*i.e.,* the worst-case perturbations) for models to defend against during training, leading to severe performance degradation. Simply dropping certain perturbed modality also outperforms the original model under attacks but the defense effect is limited.
- **Our modality balancing method has a much slighter influence on the performance in the clean scenario compared with other defense methods.** From the results, adversarial training clearly degrades clean performance because it forces the model to give correct predictions for the worst-case perturbations, making the training process very hard. The unimodal models also show a decrease in clean performance due to the missing useful information in the dropped modality. By comparison, our modality balancing shows negligible impact on the clean performance, achieving the second-best performance in most cases. Statistically, compared with the original model (*i.e.,* Multimodal), modality balancing only causes 2.93% performance drop (in terms of Recall@20) on average on Baby dataset and 1.90% on Clothing dataset, which can be nearly overlooked compared with other defense methods. The reason is that modality balancing improves robustness by only adjusting the prediction score distribution in different modalities while preserving full utilization of multi-modal information.
- **The modality sensitivity varies in different datasets.** According to our observation, on Baby dataset, the visual modality shows higher sensitivity than the textual modality. This can be

Table 3: Results of five models under two scenarios (clean and attack) on Clothing dataset. The best and second-best results in each scenario are highlighted in bold and underline, respectively.

Clothing		Clean				Attack			
Model	**Method**	**Recall@10**	**Recall@20**	**NDCG@10**	**NDCG@20**	**Recall@10**	**Recall@20**	**NDCG@10**	**NDCG@20**
MMGCN	Multimodal	**0.0211**	**0.0336**	**0.0108**	**0.0141**	0.0053	0.0100	0.0025	0.0037
	Unimodal-visual	0.0167	0.0281	0.0088	0.0117	0.0092	0.0156	0.0042	0.0062
	Unimodal-textual	0.0181	0.0294	0.0093	0.0121	0.0085	0.0137	0.0039	0.0057
	Adv training	0.0172	0.0283	0.0091	0.0119	0.0099	0.0187	**0.0066**	**0.0094**
	Modality balancing	0.0206	0.0324	0.0103	0.0135	**0.0102**	**0.0194**	0.0064	0.0092
VBPR	Multimodal	0.0281	**0.0415**	**0.0158**	**0.0192**	0.0155	0.0249	0.0078	0.0102
	Unimodal-visual	0.0276	0.0402	0.0152	0.0182	0.0161	0.0264	0.0086	0.0110
	Unimodal-textual	0.0270	0.0395	0.0148	0.0181	0.0128	0.0201	0.0066	0.0084
	Adv training	0.0254	0.0373	0.0134	0.0172	0.0172	0.0271	0.0088	0.0118
	Modality balancing	**0.0282**	0.0414	0.0155	0.0189	**0.0184**	**0.0291**	**0.0096**	**0.0127**
GRCN	Multimodal	**0.0428**	**0.0654**	**0.0231**	**0.0287**	0.0233	0.0384	0.0118	0.0157
	Unimodal-visual	0.0375	0.0568	0.0195	0.0243	0.0228	0.0376	0.0115	0.0153
	Unimodal-textual	0.0401	0.0598	0.0211	0.0262	0.0239	0.0404	0.0121	0.0169
	Adv training	0.0355	0.0517	0.0184	0.0242	0.0241	0.0407	0.0123	0.0172
	Modality balancing	0.0419	0.0632	0.0220	0.0274	**0.0282**	**0.0462**	**0.0145**	**0.0195**
SLMRec	Multimodal	**0.0433**	**0.0644**	**0.0233**	**0.0289**	0.0167	0.0264	0.0085	0.0109
	Unimodal-visual	0.0362	0.0544	0.0196	0.0242	0.0112	0.0183	0.0056	0.0074
	Unimodal-textual	0.0422	0.0643	0.0229	0.0285	0.0133	0.0204	0.0063	0.0084
	Adv training	0.0402	0.0604	0.0218	0.0259	0.0215	0.0347	0.0119	0.0148
	Modality balancing	0.0423	0.0620	0.0231	0.0280	**0.0248**	**0.0393**	**0.0138**	**0.0175**
MMGCL	Multimodal	0.0430	0.0647	0.0234	0.0289	0.0315	0.0463	0.0173	0.0208
	Unimodal-visual	0.0383	0.0585	0.0207	0.0258	0.0282	0.0416	0.0150	0.0184
	Unimodal-textual	0.0346	0.0536	0.0188	0.0229	0.0245	0.0365	0.0134	0.0165
	Adv training	0.0357	0.0524	0.0187	0.0246	0.0330	0.0488	0.0177	0.0214
	Modality balancing	**0.0436**	**0.0656**	**0.0236**	**0.0291**	**0.0345**	**0.0522**	**0.0185**	**0.0231**

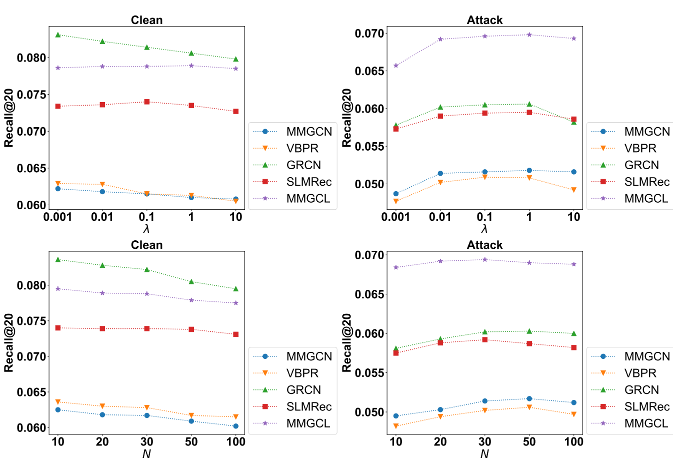


Figure 3: Study of the effect of λ and N on the model performance in the clean and attack scenario on Baby dataset.

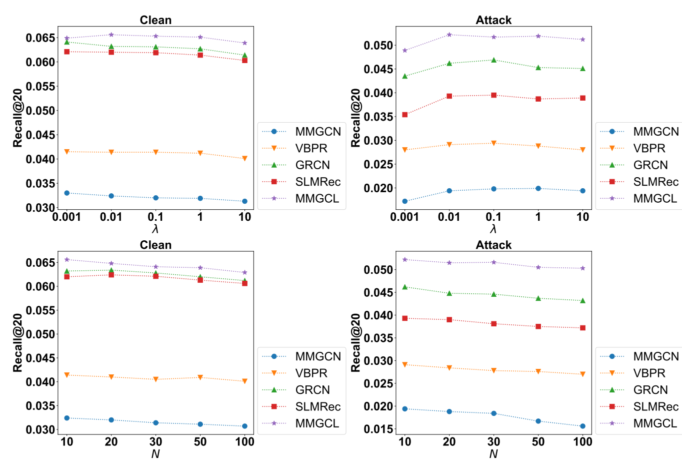
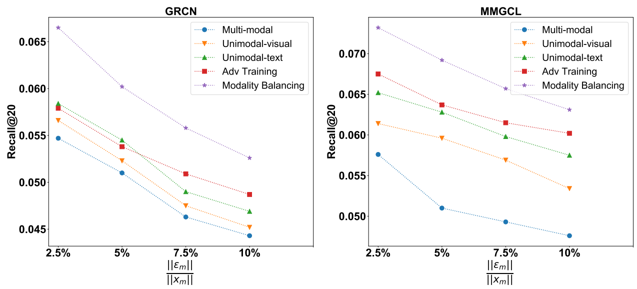


Figure 4: Study of the effect of λ and N on the model performance in the clean and attack scenario on Clothing dataset.

observed from the fact that Unimodal-textual shows higher robustness than Unimodal-visual in the attack scenario. Therefore, we improve the score margin in the textual modality to compensate for the vulnerability in the visual modality. While for Clothing dataset, the textual modality is more sensitive than the visual modality so we choose to enhance the visual prediction score margin instead. Such variation reflects the challenge of tackling the robustness issue of multimodal recommendation models. Even so, our method can adaptively address the problem regardless of the relative vulnerability by enhancing the robust modality to compensate for the vulnerable modality.

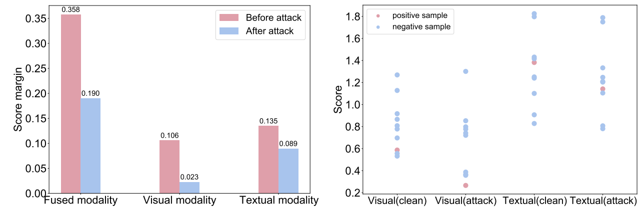
5.3 Hyper-parameter Study (RQ2)

In this section, we investigate the impact of the ratio of the two loss terms λ and the number of sampled instances for modality balancing in each batch N on model performance in both clean and attack scenarios. The results on Baby dataset and Clothing



(a) Defense method comparison for GRCN on Baby dataset (b) Defense method comparison for MMGCL on Baby dataset

Figure 5: Study of the effect of ϵ_m on the performance of GRCN and MMGCL in the attack scenario on Baby dataset.



(a) Score margin change under attacks (b) A real case showing the score change of different modalities

Figure 6: The illustrative example of adversarial attacks against models with modality balancing (results from GRCN). (a) Prediction score margin (between positive and negative samples) before and after attack for different modalities. (b) A real case showing the prediction score change of different modalities under attacks.

dataset are shown in Figure 3 and Figure 4, respectively. Besides, we test the performance of modality balancing with different attack magnitudes ϵ_m and present the result in Figure 5.

Impact of λ. We test the model performance in the clean and attack scenario with varying λ in {0.001, 0.01, 0.1, 1, 10}. From the results on two datasets, it can be observed there's a trade-off between the performance in the clean and attack scenario. All models commonly show decreasing clean performance and increasing robustness under attacks with increasing λ from 0.001 to 1. The phenomenon is reasonable since the optimization is gradually favoring the modality balancing term aiming to enhance adversarial robustness instead of clean performance.

Impact of N. We test the model performance in the clean and attack scenario with varying N in {10, 20, 30, 50, 100}. On the two datasets the clean model performance commonly drops with the increasing N. Larger N means involving more users and items in modality balancing, which might slightly influence the normal training. At the same time, more balanced users and items usually provide higher robustness. On Baby dataset, the robustness nearly achieves the peak when $N = 30$ for all models. As for Clothing dataset, the highest robustness is obtained when $N = 10$.

Impact of ϵ_m. In order to verify the generalization ability of our method, we test the defense performance of our method and compared baselines under different attack magnitudes. Specifically, we vary the norm ratio of the perturbation and original feature for each modality $\frac{||\epsilon_m||}{||x_m||}$ in {2.5%, 5%, 7.5%, 10%}. The results of GRCN and MMGCL on Baby dataset are shown in Figure 5. It can be observed that modality balancing achieves the best defense performance compared with other methods in all attack settings, which validates the superior generalization capability of our method.

5.4 Effect of Modality Balancing (RQ3)

In this section, we study the effect of modality balancing on models to explain why it works for enhancing robustness. Here we conduct a similar analysis as described in Section 3.2. Still taking GRCN as an example, we analyze the score margin change of the model trained with modality balancing in each single modality and fused modality, as shown in Figure 6(a). Compared with the model without modality balancing analyzed in Figure 1, it can be found that the score margin in the textual modality gets strengthened and exceeds the visual modality, indicating the positive samples have greater strength in the insensitive modality. In this way, it is expected that the model can still make correct predictions relying on the textual scores which are less affected by attacks. In order to verify this, we then study the same case as in Section 3.2, *i.e.*, (u_{841}, i_{916}) from Baby dataset. To be specific, we observe the visual and textual prediction scores of the positive sample i_{916} and the top 10 negative samples before and after attacks, as presented in Figure 6(b). The main difference with Figure 1(b) is that the overall textual scores are ahead of visual scores. Although the visual score of i_{916} still drops violently under attacks from 0.587 to 0.266, its attacked textual score(*i.e.*, 1.142) is much larger than the attacked visual score and enough to compensate for the lag in the visual modality. In general, the comparison between the model with and without modality balancing demonstrates our modality balancing method effectively addresses the critical threat to the robustness of multi-modal recommendation models.

6 CONCLUSIONS

In this work, we conduct a systematical study on the adversarial robustness of multi-modal recommendation models, which is vital to ensure the reliability of these models in real-world applications. We first conduct a robustness test for five mainstream models and show they are vulnerable to slight perturbations on the multi-modal input features. Next, we attribute the vulnerability to the modality imbalance issue. Finally, to address this problem, we introduce a sensitive-modal embedding distillation module and modality balancing loss term to enhance the adversarial robustness. The proposed method is flexible to be applied to various multi-modal recommendation models and effective in boosting adversarial robustness with nearly no performance decline on clean data.

ACKNOWLEDGMENTS

This work is supported in part by the National Key R&D Program of China (No. 2022ZD0117902), and the National Natural Science Foundation of China (No. 62272262, No. 61972223, No. U1936217, No. U20B2060).

REFERENCES

[1] Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. 2021. A study of defensive methods to protect visual recommendation against adversarial manipulation of images. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1094–1103.

[2] Feiyu Chen, Junjie Wang, Yinwei Wei, Hai-Tao Zheng, and Jie Shao. 2022. Breaking Isolation: Multimodal Graph Fusion for Multimedia Recommendation by Edge-wise Modulation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 385–394.

[3] Huiyuan Chen and Jing Li. 2019. Adversarial tensor factorization for context-aware recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 363–367.

[4] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 765–774.

[5] Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Felice Antonio Merra. 2020. How dataset characteristics affect the robustness of collaborative recommendation models. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 951–960.

[6] Xiaoyu Du, Zike Wu, Fuli Feng, Xiangnan He, and Jinhui Tang. 2022. Invariant Representation Learning for Multimedia Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 619–628.

[7] Yali Du, Meng Fang, Jinfeng Yi, Chang Xu, Jun Cheng, and Dacheng Tao. 2018. Enhancing the robustness of neural collaborative filtering systems under malicious attacks. *IEEE Transactions on Multimedia* 21, 3 (2018), 555–565.

[8] Minghong Fang, Neil Zhenqiang Gong, and Jia Liu. 2020. Influence function based data poisoning attacks to top-n recommender systems. In *Proceedings of The Web Conference 2020*. 3019–3025.

[9] Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, et al. 2023. A survey of graph neural networks for recommender systems: challenges, methods, and directions. *ACM Transactions on Recommender Systems* 1, 1 (2023), 1–51.

[10] Chen Gao, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. 2022. Causal Inference in Recommender Systems: A Survey and Future Directions. *arXiv preprint arXiv:2208.12397* (2022).

[11] 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*. JMLR Workshop and Conference Proceedings, 249–256.

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[13] Ihsan Gunes, Cihan Kaleli, Alper Bilge, and Huseyin Polat. 2014. Shilling attacks against recommender systems: A comprehensive survey. *Artificial Intelligence Review* 42, 4 (2014).

[14] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. 507–517.

[15] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[16] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.

[17] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial personalized ranking for recommendation. In *The 41st International ACM SIGIR conference on research & development in information retrieval*. 355–364.

[18] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. (2015).

[19] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 426–434.

[20] Shyong K Lam and John Riedl. 2004. Shilling recommender systems for fun and profit. In *Proceedings of the 13th international conference on World Wide Web*. 393–402.

[21] Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. 2016. Data poisoning attacks on factorization-based collaborative filtering. *Advances in neural information processing systems* 29 (2016).

[22] Fan Liu, Zhiyong Cheng, Changchang Sun, Yinglong Wang, Liqiang Nie, and Mohan Kankanhalli. 2019. User diverse preference modeling by multimodal attentive metric learning. In *Proceedings of the 27th ACM international conference on multimedia*. 1526–1534.

[23] Shang Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu. 2019. User-video co-attention network for personalized micro-video recommendation. In *The World Wide Web Conference*. 3020–3026.

[24] 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*. 687–695.

[25] Zhuoran Liu and Martha Larson. 2021. Adversarial item promotion: Vulnerabilities at the core of top-n recommenders that use images to address cold start. In *Proceedings of the Web Conference 2021*. 3590–3602.

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

[27] Zongshen Mu, Yueting Zhuang, Jie Tan, Jun Xiao, and Siliang Tang. 2022. Learning Hybrid Behavior Patterns for Multimedia Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 376–384.

[28] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 452–461.

[29] Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua. 2019. Adversarial training towards robust multimedia recommender system. *IEEE Transactions on Knowledge and Data Engineering* 32, 5 (2019), 855–867.

[30] Jiaxi Tang, Hongyi Wen, and Ke Wang. 2020. Revisiting adversarially learned injection attacks against recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 318–327.

[31] 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).

[32] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. 2020. Mgat: Multimodal graph attention network for recommendation. *Information Processing & Management* 57, 5 (2020), 102277.

[33] Nhu-Thuat Tran and Hady W Lauw. 2022. Aligning Dual Disentangled User Representations from Ratings and Textual Content. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1798–1806.

[34] Haoyu Wang, Nan Shao, and Defu Lian. 2019. Adversarial binary collaborative filtering for implicit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 5248–5255.

[35] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia* (2021).

[36] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*. 165–174.

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

[38] 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.

[39] 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*. 1437–1445.

[40] Chenwang Wu, Defu Lian, Yong Ge, Zhihao Zhu, Enhong Chen, and Senchao Yuan. 2021. Fight fire with fire: towards robust recommender systems via adversarial poisoning training. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1074–1083.

[41] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig Macdonald. 2022. Multi-modal graph contrastive learning for micro-video recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1807–1811.

[42] Feng Yuan, Lina Yao, and Boualem Benatallah. 2019. Adversarial collaborative neural network for robust recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1065–1068.

[43] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 353–362.

[44] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3872–3880.

[45] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. 2023. A Comprehensive Survey on Multimodal Recommender Systems: Taxonomy, Evaluation, and Future Directions. *arXiv preprint arXiv:2302.04473* (2023).

[46] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2022. Bootstrap latent representations for multi-modal recommendation. *arXiv preprint arXiv:2207.05969* (2022).