Band Together: Untargeted Adversarial Training with Multimodal Coordination against Evasion-based Promotion Attacks

Guanmeng Xian[1], **Ning Yang**[1*], **Philip S. Yu**[2]
[1]Sichuan University, Chengdu, China
[2]University of Illinois at Chicago, USA
xianguanmeng@stu.scu.edu.cn, yangning@scu.edu.cn, psyu@uic.edu

Abstract

Multimodal recommender systems exploit visual and textual signals to alleviate data sparsity, but this also makes them more vulnerable to evasion-based promotion attacks. Existing defenses are largely limited to single-modal settings and mainly focus on poisoning-based threats, leaving evasion-based threats underexplored. In this work, we first identify a cross-modal gradient mismatch under the multi-user promotion setting, where visual and textual perturbations are optimized in inconsistent directions due to the dominance of distinct user groups. This phenomenon dilutes the attack effectiveness and leads robust training to underestimate worst-case risks. To address this issue, we propose **U**ntargeted **A**dversarial **T**raining with **M**ultimodal **C**oordination (UAT-MC). UAT-MC tackles the challenge of unknown targeted items in evasion-based attacks (as opposed to poisoning-based attacks) by treating all items as potential targets, and introduces a gradient alignment mechanism to explicitly correct this mismatch. This design ensures synchronized perturbations across modalities, thereby maximizing adversarial strength for robust training. Extensive experiments demonstrate that UAT-MC significantly improves robustness against promotion attacks while maintaining acceptable recommendation performance under the defense–accuracy trade-off. Code is available at https://github.com/gmXian/UAT-MC.

1 Introduction

Multimodal Recommender Systems (MRSs) typically leverage visual and textual information from user-interacted items to capture users' fine-grained preferences, effectively alleviating the challenge posed by sparse interaction data [He and McAuley, 2016b; Wei *et al.*, 2019; Zhang *et al.*, 2021; Zhou *et al.*, 2023c; Liu *et al.*, 2024]. While auxiliary modal information enhances the personalization of recommendations, we find that MRSs exhibit heightened vulnerability to **evasion-based promotion attacks**, in which attackers imperceptibly perturb the modalities of targeted items to boost their rankings, as demonstrated by our motivation experiment in Figure 1.

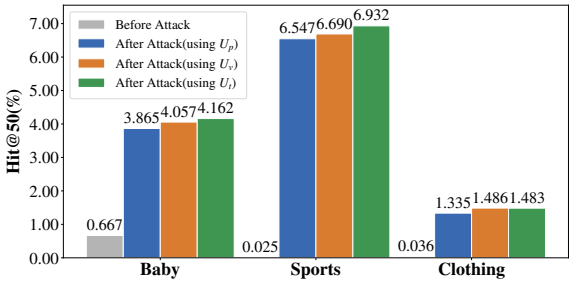


Figure 1: VBPR's vulnerability to vanilla FGSM-based promotion attacks. For each dataset, we report the Hit@50 before and after attack under different user subsets, including $\mathcal{U}_p$, $\mathcal{U}_v$, and $\mathcal{U}_t$.

Specifically, we conduct evasion-based promotion attacks (FGSM [Goodfellow *et al.*, 2014] and PGD [Madry *et al.*, 2017]) on both visual and textual modalities of unpopular items (defined as those with only 5 interactions) over various datasets (Amazon Baby, Sports and Clothing [He and McAuley, 2016a]) and architectures (VBPR [He and McAuley, 2016b] and MMGCN [Wei *et al.*, 2019])). Figure 1 presents the results on VBPR under FGSM attack as a representative instance. The black numbers represent the average hit rate (Hit@50) of targeted items appearing in user recommendation lists. As one can find from Figure 1, the post-attack Hit@50 always shows a significant increase, demonstrating the vulnerability of MRS to evasion-based promotional attacks.

Existing works [Wu *et al.*, 2021; Zhang *et al.*, 2024; Mu *et al.*, 2025] have attempted to mitigate the vulnerability to promotion attacks; however, they suffer from two critical limitations when applied to MRSs. **First**, these approaches primarily focus on single-modal recommender systems, failing to address the complex vulnerability patterns that arise from interactions between visual and textual features. This oversight is especially problematic as attackers can exploit cross-modal correlations to amplify their attack impact. For example, the Re-writing Defense [Zhang *et al.*, 2024] uses GPT-3.5-turbo to rewrite adversarial text as a defense for

*Corresponding author.

LLM-based recommendation models. **Second**, and more fundamentally, current defenses are almost exclusively designed for poisoning-based attacks, where adversaries compromise collaborative signals by injecting fake user profiles or interactions during the training phase. For example, APT [Wu *et al.*, 2021] simulates the poisoning process by injecting fake user data to foster a more robust system. However, in MRSs, attackers can deliberately manipulate the description or image of a targeted item during the inference phase to influence the recommendation results—an aspect not addressed by existing defense methods. To bridge these critical gaps, we propose **Untargeted Adversarial Training with Multimodal Coordination (UAT-MC)**, a novel adversarial training framework specifically designed to defend against evasion-based promotion attacks in MRSs. However, directly applying conventional adversarial training to this setting is non-trivial due to the following two challenges:

C1: Dynamic Attack Target. Unlike poisoning attacks where malicious targets are explicitly injected during training, evasion attacks dynamically select targeted items at the inference phase. This fundamental difference renders traditional targeted adversarial training approaches ineffective, as they rely on pre-defined attack targets to generate adversarial examples, which are unknown during the training phase.

C2: Cross-modal Gradient Mismatch. Under the multi-user promotion setting, MRSs present a unique challenge: combining perturbations across visual and textual modalities often yields suboptimal adversarial examples which degrade the robustness achieved through adversarial training. Cross-modal gradient mismatch arises when visual and textual perturbations in multi-user promotion attacks are dominated by different user groups, causing the two modalities to be optimized toward inconsistent objectives and resulting in misaligned gradient directions.

To tackle the challenge **C1**, we propose **untargeted adversarial training** for MRS, which treats all items as potential targets of evasion-based attacks. Our approach frames recommendation as a multi-label classification task over users, where items serve as labels. Untargeted adversarial training indirectly defends against targeted attacks by globally enhancing the robustness of decision boundaries and disrupting the local gradient information on which attacks rely. The core principle is that the model learns to resist perturbations from any direction, thereby covering attacks originating from specific directions. To tackle challenge **C2**, we propose a novel **gradient-aligned multimodal perturbation method** to address cross-modal gradient mismatch in adversarial training. Unlike perturbing visual and textual features independently, which can diminish attack strength due to misaligned gradients—our method enforces gradient synchronization across modalities by minimizing the cosine distance between visual and textual gradients through a joint loss term. This ensures perturbations coherently push items toward adversarial regions. As shown in later experiments, this coordinated approach maximizes attack potency during training and enhances adversarial robustness against evasion-based promotion attacks.

Our contributions can be summarized as follows:

- We identify evasion-based promotion attacks as a critical threat to multimodal recommender systems, in which adversaries perturb both the visual and textual features of targeted items.
- We propose a novel Untargeted Adversarial Training with Multimodal Coordination (UAT-MC) framework to enhance the robustness of MRSs against evasion-based promotion attacks. Specifically, the proposed untargeted adversarial training inherently defends against targeted attacks by universally hardening the decision boundaries, thereby covering all potential attack directions.
- We propose gradient-aligned multimodal perturbations to resolve cross-modal gradient mismatch in adversarial training. By minimizing cosine distance between modalities via a joint loss term, our method synchronizes perturbations to maximize attack potency.
- Extensive experiments conducted on real datasets verify the effectiveness of our method.

2 Preliminary

In this section, we conceptually define MRS and describe the evasion-based promotion attacks.

2.1 MRS

Let $\mathcal{U}$ and $\mathcal{I}$ denote user set and item set, respectively. $R \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$ is user-item interaction matrix, where an element $r_{u,i}$ is 1 if there exists interaction between user $u \in \mathcal{U}$ and item $i \in \mathcal{I}$, otherwise 0. Each item i is associated with a visual modality embedding $\mathbf{v}_i$ and a textual modality embedding $\mathbf{t}_i$ which are generated by CNN and Transformer, respectively.

Generally, an MRS f infers the probability $\hat{r}_{u,i}$ that user u will interact with item i based on given R and the modality embeddings $\mathbf{v}_i$ and $\mathbf{t}_i$, i.e.,

$$\hat{r}_{u,i} = f(R, u, i, \mathbf{v}_i, \mathbf{t}_i). \tag{1}$$

Let $\mathcal{D} = \{(u, i_+, i_-)\}$ be a training set, where i_+ is u's positive sample ($r_{u,i_+} = 1$) and i_- is u's negative sample ($r_{u,i_-} = 0$). Like traditional recommender systems, an MRS is usually trained with Bayesian Personalized Ranking (BPR) [Rendle *et al.*, 2009] loss function,

$$\mathcal{L}_{\text{BPR}}(\Theta) = \sum_{(u,i_+,i_-)\in\mathcal{D}} -\ln \sigma(\hat{r}_{u,i_+} - \hat{r}_{u,i_-}), \tag{2}$$

where Θ represents the model parameters and $\sigma(\cdot)$ is the sigmoid function.

2.2 Threat Model

Attacker's Objective

Given a well-trained MRS, an evasion-based promotion attack is to imperceptibly perturb the multimodal embeddings of a targeted item i, so that i will appear in the top-K recommendation lists of as many users as possible. Inspired by [Wang *et al.*, 2024b], we define the promotion utility function as:

$$\mathcal{L}_{\text{promotion}} = \frac{1}{N_p} \sum_{u\in\mathcal{U}_{\text{p}}} \text{Sigmoid}(y_{u,i} - y_{u,K}), \tag{3}$$

where $y_{u,K}$ denotes the score of the K-th ranked item for user u. By maximizing Equation (3), the attacker encourages the targeted item score $y_{u,i}$ to surpass the top-K threshold $y_{u,K}$.

Attacker's Knowledge

In this paper, we assume a white-box scenario for the generation of the adversarial examples in our UAT-MC, where the attacker has full access to the MRS f, including its parameters and gradients. This is because UAT-MC aims to train a more robust MRS, rather than to conduct attacks from the perspective of malicious merchants. This setting represents a worst-case attacker and allows us to evaluate the upper bound of the potential impact of evasion-based promotion attacks.

Attacker's Capability

Following the ideas of the related works [Tang *et al.*, 2020; Zhang *et al.*, 2021; Guo *et al.*, 2024; Liu *et al.*, 2024; Ong and Khong, 2025] , we perturb the multimodal embeddings instead of the raw inputs, which allows more fine-grained and efficient manipulations. Formally, the perturbed embeddings $\mathbf{v}_i' = \mathbf{v}_i + \Delta_v^i$ and $\mathbf{t}_i' = \mathbf{t}_i + \Delta_t^i$. For imperceptibility, the perturbations Δ_m^i ($m \in \{v, t\}$) are constrained by $\|\Delta_m^i\| \leq \epsilon_m$, where ϵ_m is perturbation budget.

3 Cross-modal Gradient Mismatch Analysis

In this section, we demonstrate that under multi-user promotion settings, the gradients of visual and textual perturbations are inherently driven by distinct user groups due to varying modal sensitivities. Consequently, simply aggregating these gradients leads to conflicting optimization directions across modalities, which we term **cross-modal gradient mismatch**. This misalignment fundamentally limits the synergy of multimodal attacks and motivates our UAT-MC framework. We next provide a formal analysis to characterize and verify the existence of this mismatch.

3.1 Objective Inconsistency Across Modalities

Given the promotion objective $\mathcal{L}_{\text{promotion}}$ defined over the user set $\mathcal{U}_p$ (Equation (3)), the optimization problem can be formulated as:

$$\max_{\Delta_v^i, \Delta_t^i} \sum_{u \in \mathcal{U}_p} \mathcal{L}_u(\Delta_v^i, \Delta_t^i), \tag{4}$$

where $\mathcal{L}_u$ denotes the promotion loss contributed by user u.

Accordingly, the perturbations are updated by aggregating gradients from all selected users:

$$G^v = \sum_{u \in \mathcal{U}_p} g_u^v, \quad G^t = \sum_{u \in \mathcal{U}_p} g_u^t, \tag{5}$$

where $g_u^v = \frac{\partial \mathcal{L}_u}{\partial \Delta_v^i}$ and $g_u^t = \frac{\partial \mathcal{L}_u}{\partial \Delta_t^i}$ denote the visual and textual gradients induced by user u, respectively.

To quantify how much each user contributes to the final update direction, we define the **directional contribution** of user u on modality $m \in \{v, t\}$ as:

$$c_u^m = \cos(g_u^m, G^m) \cdot \frac{\|g_u^m\|}{\sum_{u' \in \mathcal{U}_p} \|g_{u'}^m\|}, \tag{6}$$

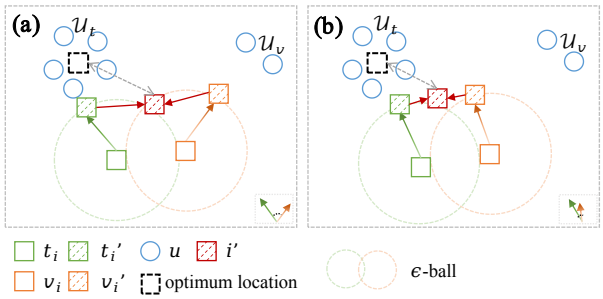
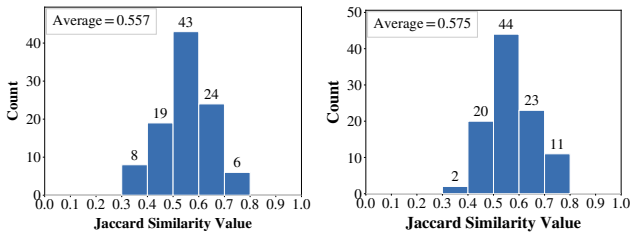


Figure 2: Illustration of objective inconsistency across modalities. (a) Visual and textual perturbations are dominated by user groups $\mathcal{U}_v$ and $\mathcal{U}_t$, leading to conflicting promotion directions after multimodal fusion. (b) With alignment loss, the perturbations from different modalities are constrained to align in a consistent direction, enabling the fused embedding to effectively reach the optimum location.



(a) VBPR-Baby (b) VBPR-Sports

Figure 3: Distribution of Jaccard Similarity between $\mathcal{U}_v$ and $\mathcal{U}_t$.

where $\cos(g_u^m, G^m)$ measures whether user u's gradient direction is aligned with the aggregated update direction, and $\|g_u^m\|$ reflects the contribution of user u to modality m.

Based on this metric, we define the user set $\mathcal{U}_v$ and the user set $\mathcal{U}_t$ as the top-K users ranked by c_u^v and c_u^t, respectively. To measure the overlap between these two sets, we compute their Jaccard similarity:

$$J(\mathcal{U}_v, \mathcal{U}_t) = \frac{|\mathcal{U}_v \cap \mathcal{U}_t|}{|\mathcal{U}_v \cup \mathcal{U}_t|}. \tag{7}$$

A low Jaccard similarity indicates that the users dominating the visual and textual perturbation updates are largely different, suggesting that the two modalities are implicitly driven by different user groups and, consequently, toward different effective optimization objectives. As illustrated in Fig. 2(a), the visual perturbation Δ_v^i is primarily driven by users in $\mathcal{U}_v$, while the textual perturbation Δ_t^i is dominated by users in $\mathcal{U}_t$. When such modality-specific perturbations are subsequently fused by the MRS, the resulting item representation fails to simultaneously benefit all users, leading to suboptimal promotion outcomes.

3.2 Empirical Evidence of Cross-modal Gradient Mismatch

As mentioned in Section 1, we conduct vanilla FGSM-based promotion attacks on the VBPR model across three datasets. For each dataset, we randomly sample 100 unpopular items

whose interaction counts are no greater than 5. For each targeted item, we identify the user sets $\mathcal{U}_v$ and $\mathcal{U}_t$ using the directional contribution metric in Equation (6), and conduct the promotion attack variants by constructing the promotion loss over $\mathcal{U}_p$, $\mathcal{U}_v$ and $\mathcal{U}_t$, respectively.

The results indicate that aligning the attack objective with modality-specific user groups is crucial, as visual and textual perturbations are driven by distinct user subsets. Specifically, Figure 3 reveals a consistently low Jaccard similarity between $\mathcal{U}_v$ and $\mathcal{U}_t$, confirming their limited overlap. This distinctness is further corroborated by Figure 1, which demonstrates that optimizing over $\mathcal{U}_v$ or $\mathcal{U}_t$ yields significantly higher attack gains than using $\mathcal{U}_p$, as aggregating heterogeneous groups tends to dilute modality-specific optimization signals. (See Appendix B.4 for a detailed case study visualizing this phenomenon on a specific item.)

In summary, cross-modal gradient mismatch limits the effectiveness of multimodal perturbations. Since effective defense requires training against worst-case attacks, it is critical to harmonize optimization directions to maximize attack potency. This motivates our UAT-MC framework, which uses gradient alignment to generate stronger adversarial examples, thereby driving the model to learn more robust representations.

4 Methodology

Figure 4 shows the overview of UAT-MC. In Figure 4, the targeted MRS is divided into two parts, i.e., $f = g \circ h$, where h is the encoder for collaborative filtering and g is the decoder. Usually, h is implemented as a GNN [Wei *et al.*, 2019] or an MLP [He and McAuley, 2016b] for fusion of multimodal embeddings, while g is implemented as an inner product. A training instance consists of user ID embedding $\mathbf{e}_u$, a positive item's ID embedding $\mathbf{e}_+$, a negative item's ID embedding $\mathbf{e}_-$, the positive item's modality embeddings $\mathbf{v}_+$, $\mathbf{t}_+$, and the negative item's modality embeddings $\mathbf{v}_-$ and $\mathbf{t}_-$.

During the k-th iteration, UAT-MC first generates the adversarial modality embeddings $\mathbf{v}'_+$, $\mathbf{t}'_+$, $\mathbf{v}'_-$, and $\mathbf{t}'_-$, by adding the perturbations generated in the last step. Then through the collaborative filtering encoder h, UAT-MC produces the user embedding $\mathbf{h}_u$ by fusing with the multimodal content of the items interacted with u, the positive and negative item embeddings $\mathbf{h}_+$ and $\mathbf{h}_-$ by fusing their respective multimodal embeddings, and their respective adversarial sample embeddings $\mathbf{h}'_u$, $\mathbf{h}'_+$ and $\mathbf{h}'_-$. Based on these embeddings, the decoder g computes the ranking losses $\mathcal{L}_{\text{BPR}}$ and $\mathcal{L}'_{\text{BPR}}$ on the clean embeddings and the adversarial embeddings, respectively. On these losses, UAT-MC conducts a min-max game to improve the robustness of the MRS f.

It is worth noting that to achieve worst-case adversarial robustness for f, UAT-MC promotes consistent perturbation directions across different modalities by maximizing the gradient alignment loss $\mathcal{L}_{\text{Align}}$ when generating adversarial perturbations. This approach realizes the multimodal coordination, thereby enhancing the aggressiveness of adversarial examples.

4.1 Untargeted Adversarial Training

As mentioned in Section 1, in evasion attacks, the specific targeted item at inference time is unknown during the training phase. To address this challenge, we propose untargeted an adversarial training for MRSs, which treats all items as potential targets of evasion-based promotion attacks. This approach is motivated by the fact that recommendation can be formulated as a multi-label classification task, where each item serves as a distinct label. Untargeted adversarial training improves the robustness of the model by strengthening the decision boundary against perturbations from arbitrary directions, thereby effectively defending against perturbations from any direction.

Specifically, given a user u, the attacker adds perturbations $[\Delta_{v_+}, \Delta_{t_+}]$ to the positive item's multimodal embeddings $[\mathbf{v}_+, \mathbf{t}_+]$ and the perturbations $[\Delta_{v_-}, \Delta_{t_-}]$ to the negative item's multimodal embeddings$[\mathbf{v}_-, \mathbf{t}_-]$, which result in the adversarial modality embeddings $\mathbf{v}'_+$, $\mathbf{t}'_+$, $\mathbf{v}'_-$, and $\mathbf{t}'_-$. Then the final adversarial embeddings are computed as:

$$\mathbf{h}'_u, \mathbf{h}'_+, \mathbf{h}'_- = h(R, \mathbf{e}_u, \mathbf{e}_+, \mathbf{e}_-, \mathbf{v}'_+, \mathbf{v}'_-, \mathbf{t}'_+, \mathbf{t}'_-). \quad (8)$$

Then the untargeted adversarial training is defined as:

$$\min_{\Theta} \max_{\Delta_v, \Delta_t} \mathcal{L}'_{\text{BPR}}(\Theta, \Delta_v, \Delta_t), \quad (9)$$

where $\mathcal{L}'_{\text{BPR}}(\Theta, \Delta_v, \Delta_t)$ is defined as:

$$\mathcal{L}'_{\text{BPR}}(\Theta, \Delta_v, \Delta_t) = \sum_{(u, i_+, i_-) \in \mathcal{D}} -\ln \sigma(\mathbf{h}'_u \cdot \mathbf{h}'_+ - \mathbf{h}'_u \cdot \mathbf{h}'_-). \quad (10)$$

4.2 Multimodal Coordination

As analyzed in Section 3, independently perturbing each modality in multi-user promotion settings often results in cross-modal gradient mismatch, reducing the effectiveness of adversarial attacks. While identifying and optimizing over specific user subsets (i.e., $\mathcal{U}_v$ and $\mathcal{U}_t$) can improve attack performance, it requires computing user-wise gradients for the entire user base, incurring a prohibitively high computational cost. To circumvent this bottleneck, we propose a lightweight **multimodal coordination** mechanism. Instead of explicitly partitioning users, we focus on directly rectifying the divergent optimization directions in the gradient space. By enforcing gradient-level alignment between modalities, we effectively mitigate the mismatch without the need for costly user identification. This approach achieves coordination with negligible computational overhead, requiring only minimal additional operations during backpropagation.

We first compute the gradients of the adversarial loss with respect to the perturbations in each modality:

$$\Gamma_v = \nabla_{\Delta_v} \mathcal{L}'_{\text{BPR}}, \Gamma_t = \nabla_{\Delta_t} \mathcal{L}'_{\text{BPR}}. \quad (11)$$

Then, we define the alignment loss as the sum of the cosine similarities between the gradients of visual and textual modalities for both positive and negative items:

$$\mathcal{L}_{\text{Align}} = cos(\Gamma_{v_+}, \Gamma_{t_+}) + cos(\Gamma_{v_-}, \Gamma_{t_-}). \quad (12)$$

Maximizing $\mathcal{L}_{\text{Align}}$ encourages the perturbations in different modalities to be directionally aligned, thereby pushing the targeted item toward the adversarial region.

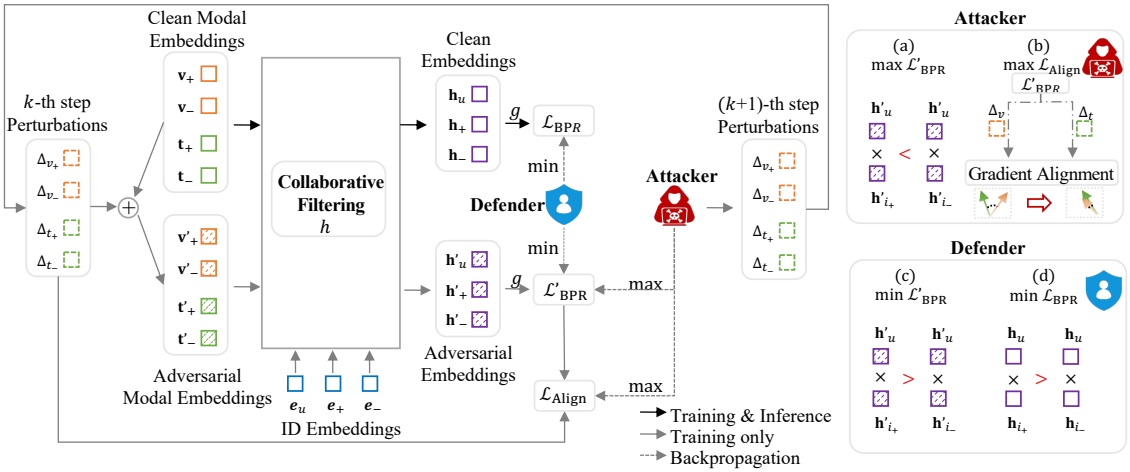


Figure 4: The framework of UAT-MC.

Algorithm 1 UAT-MC

Input: Training data $\mathcal{D}$; Learning rate η; Perturbation budgets ϵ_t, ϵ_v; Hyperparameters λ, α, β
Output: Robust MRS model parameters Θ

Initialize Θ from normal-trained MRS
while not converged **do**
  Randomly draw an example (u, i_+, i_-) from $\mathcal{D}$
  // Max phase
  $\mathcal{L}'_{\text{BPR}} = -\ln \sigma(\mathbf{h}'_u \cdot \mathbf{h}'_+ - \mathbf{h}'_u \cdot \mathbf{h}'_-)$
  $\Gamma_v = \nabla_{\Delta_v} \mathcal{L}'_{\text{BPR}}, \Gamma_t = \nabla_{\Delta_t} \mathcal{L}'_{\text{BPR}}$
  $\mathcal{L}_{\text{Align}} = \cos(\Gamma_{v_+}, \Gamma_{t_+}) + \cos(\Gamma_{v_-}, \Gamma_{t_-})$
  $\mathcal{L}_{\max} = \mathcal{L}'_{\text{BPR}} + \alpha \mathcal{L}_{\text{Align}}$
  $\Delta_v \leftarrow \epsilon_v \cdot \frac{\nabla_{\Delta_v} \mathcal{L}'_{\max}}{\|\nabla_{\Delta_v} \mathcal{L}'_{\max}\|_2}, \Delta_t \leftarrow \epsilon_t \cdot \frac{\nabla_{\Delta_t} \mathcal{L}'_{\max}}{\|\nabla_{\Delta_t} \mathcal{L}'_{\max}\|_2}$
  // Min phase
  $\mathcal{L}_{\min} = \mathcal{L}_{\text{BPR}} + \lambda \mathcal{L}'_{\text{BPR}}(\Delta_v, \Delta_t) + \beta \|\Theta\|_2$
  $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}_{\min}$
end while
return Θ

Dataset	#Users	#Items	#Interactions	Sparsity
Baby	19,445	7,037	160,792	99.883%
Sports	35,598	18,357	296,337	99.955%
Clothing	39,387	23,033	278,677	99.969%

Table 1: Statistics of the three experimental datasets.

4.3 Optimization

The training is divided into two phases:

- **Attacking** (Max phase): Generate the most disruptive perturbations Δ_v and Δ_t within ℓ_2-norm budgets ϵ_v and ϵ_t by maximizing $\lambda \mathcal{L}'_{\text{BPR}} + \alpha \mathcal{L}_{\text{Align}}$.
- **Defending** (Min phase): Update model parameters Θ by minimizing $\mathcal{L}_{\min} = \mathcal{L}_{\text{BPR}} + \lambda \mathcal{L}'_{\text{BPR}}(\Delta_v, \Delta_t) + \beta \|\Theta\|_2$.

The overall training process is described in Algorithm 1, where the MRS is pre-trained using the standard BPR loss.

5 Experiments

The objectives of experiments are to answer the following research questions:

- **RQ1:** Can UAT-MC defend against evasion-based promotion attacks?
- **RQ2:** Does multimodal coordination enhance the effectiveness of adversarial defense?
- **RQ3:** How does the perturbation budget in adversarial training affect the defense performance?
- **RQ4:** How do the hyper-parameters affect the trade-off between recommendation performance and adversarial robustness?

All experiments are implemented using PyTorch 2.2.0 and run on an NVIDIA RTX 4090 GPU with 48GB of memory.

5.1 Experimental Settings

Datasets

We conduct experiments on three Amazon datasets Baby, Sports and Clothing [He and McAuley, 2016a], which have recently been adopted in MRSs research [Yu *et al.*, 2023; Guo *et al.*, 2024; Li *et al.*, 2025]. The dataset statistics are summarized in Table 1. In our work, we use the pre-extracted visual features and textual features that have been published in [Zhou *et al.*, 2023c; Zhou *et al.*, 2023a].

Evaluation Metrics

To measure recommendation performance before and after adversarial training, we use Recall@10 and NDCG@10 under the leave-one-out evaluation protocol following standard evaluation settings in prior works [Zhou *et al.*, 2023c; Guo *et al.*, 2024]. To evaluate the effectiveness of the promotion attack, we adopt the hit rate $\text{Hit}_i@K = N^i_{rec}/|\mathcal{U}| \cdot 100\%$ and the relative improvement ratio $\text{Gain}_{\text{Hit}_i@K} = (\text{Hit}_{\text{after}}@K - \text{Hit}_{\text{before}}@K)/\text{Hit}_{\text{before}}@K \cdot 100\%$ as evaluation metrics, where N^i_{rec} is the number of users for whom the targeted item i appears in the top-K recommendation list (with K defaulted to 50), $\text{Hit}_{\text{after}}@K$ and $\text{Hit}_{\text{before}}@K$ denote the hit rate of the targeted item after and before the attack, respectively.

Dataset	Victim Model	Defense Method	$\text{Hit}_{\text{before}}$	FGSM-based Attack				PGD-based Attack				Recommendation	
				$\mathcal{L}_{\text{prom}}$		$\mathcal{L}_{\text{prom}} + \mathcal{L}_{\text{Align}}$		$\mathcal{L}_{\text{prom}}$		$\mathcal{L}_{\text{prom}} + \mathcal{L}_{\text{Align}}$		Recall	NDCG
				$\text{Hit}_{\text{after}}$	Gain_{Hit}	$\text{Hit}_{\text{after}}$	Gain_{Hit}	$\text{Hit}_{\text{after}}$	Gain_{Hit}	$\text{Hit}_{\text{after}}$	Gain_{Hit}		
Baby	VBPR	w/o AT	0.667%	3.865%	479.38%	3.908%	485.80%	4.111%	516.15%	4.143%	520.99%	0.0503	0.027
		UAT	0.622%	2.381%	282.96%	2.383%	283.41%	2.450%	294.15%	2.451%	294.25%	0.0484	0.0264
		UAT-MC	0.621%	1.515%	**143.98%**	1.515%	**144.00%**	1.544%	**148.65%**	1.544%	**148.65%**	0.0476	0.0253
	MMGCN	w/o AT	0.622%	1.747%	181.03%	1.748%	181.14%	3.399%	446.76%	3.421%	450.30%	0.0413	0.022
		UAT	0.472%	0.507%	7.29%	0.508%	7.49%	0.515%	9.12%	0.516%	9.34%	0.0344	0.0185
		UAT-MC	0.479%	0.513%	**7.21%**	0.514%	**7.47%**	0.519%	**8.46%**	0.520%	**8.62%**	0.0341	0.0185
Sports	VBPR	w/o AT	0.025%	6.547%	25722.99%	6.586%	25876.73%	7.253%	28506.93%	7.279%	28610.25%	0.0586	0.0319
		UAT	0.025%	0.028%	13.45%	0.028%	13.45%	0.029%	13.73%	0.029%	13.73%	0.0510	0.0280
		UAT-MC	0.024%	0.026%	**10.32%**	0.026%	**10.32%**	0.026%	**10.32%**	0.026%	**10.32%**	0.0508	0.0280
	MMGCN	w/o AT	0.024%	0.266%	1024.04%	0.268%	1033.23%	0.572%	2315.13%	0.581%	2352.82%	0.0385	0.0206
		UAT	0.035%	0.038%	7.01%	0.039%	11.02%	0.044%	24.45%	0.044%	25.45%	0.0313	0.0172
		UAT-MC	0.034%	0.034%	**0.00%**	0.037%	**8.44%**	0.041%	**19.14%**	0.041%	**19.14%**	0.0317	0.0173
Clothing	VBPR	w/o AT	0.036%	1.335%	3558.26%	1.380%	3682.09%	1.419%	3788.87%	1.438%	3838.61%	0.0384	0.0211
		UAT	0.031%	0.287%	818.66%	0.303%	869.17%	0.306%	878.50%	0.312%	898.38%	0.0343	0.0189
		UAT-MC	0.030%	0.038%	**26.64%**	0.038%	**26.85%**	0.038%	**27.70%**	0.038%	**27.70%**	0.0333	0.0184
	MMGCN	w/o AT	0.014%	1.886%	13282.43%	1.941%	13670.72%	9.035%	64019.37%	9.134%	64715.32%	0.0239	0.0123
		UAT	0.014%	0.025%	76.00%	0.026%	81.78%	0.035%	148.44%	0.036%	149.33%	0.0221	0.0114
		UAT-MC	0.015%	0.019%	**29.44%**	0.020%	**33.77%**	0.032%	**119.05%**	0.032%	**121.65%**	0.0221	0.0115

Table 2: Performance of promotion attacks on two classical models. All reported numbers are averaged results. The best defense results are highlighted in bold. (Note: $\text{Gain}_{\text{Hit@50}}$ values are computed using unrounded Hit@50 scores).

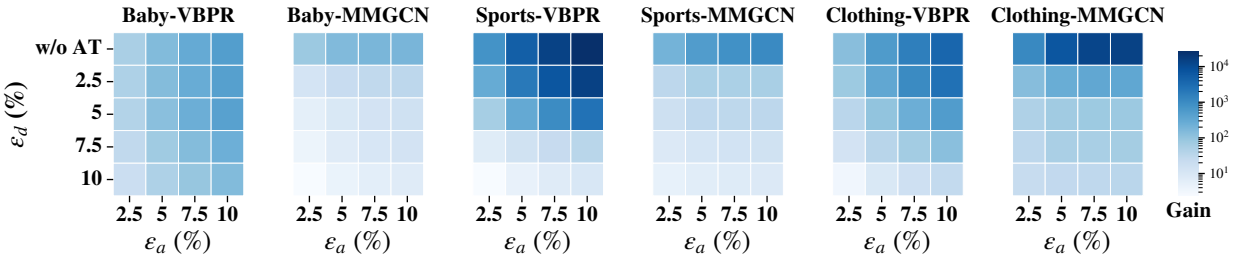


Figure 5: Visualization of attack effectiveness under varying budgets. The heatmaps display the $Gain_{Hit@50}$ (log scale) across three datasets and two models. The x-axis represents the attack budget ϵ_a, and the y-axis represents the defense budget ϵ_d. Darker colors indicate higher attack gains (weaker defense), while lighter colors indicate effective defense.

A larger $\text{Gain}_{\text{Hit@50}}$ indicates a more effective promotion attack, suggesting weaker model robustness.

Targeted Items

At the MRS's inference phase, we randomly select 100 targeted items from unpopular items, whose interaction count is 5. We conduct promotion attacks on each targeted item individually and record the average $\text{Hit}_i@K$ and $\text{Gain}_{\text{Hit}_i@K}$.

Victim Models

We choose two mainstream and classical multimodal recommendation models as our victim models:

- VBPR [He and McAuley, 2016b]: As a representative of MLP-based models, VBPR incorporates visual features for user preference learning with BPR loss. Following [Zhou *et al.*, 2023c; Guo *et al.*, 2024], we concatenate the multimodal embeddings and the item ID embedding as the item embeddings.
- MMGCN [Wei *et al.*, 2019]: As a representative of GCN-based models, MMGCN constructs three modal-specific graph to learn different modality features. It concatenates all modality embeddings to obtain the representations of users or items.

We follow MMRec [Zhou *et al.*, 2023a] to save the model parameters at the point of best performance. The hyper-parameters search spaces are provided in the Appendix B.1.

Promotion Attacks

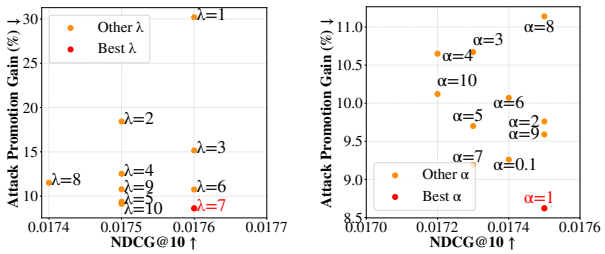
We conduct promotion attacks by maximizing $\mathcal{L}_{\text{promotion}}$ (Eq. 3) using two standard strategies: FGSM [Goodfellow *et al.*, 2014] and PGD [Madry *et al.*, 2017]. FGSM serves as a single-step baseline, whereas PGD functions as a stronger iterative attacker. In our experiments, the PGD attack is configured with 10 iterations and a step size of $\alpha = 1.25 \cdot \epsilon_m / 10$.

Defenders

To verify the effectiveness of the Untargeted Adversarial Training and the Multimodal Coordination, we compare UAT-MC with two baseline models: one is the victim model without any defense mechanisms, denoted as **w/o AT**, and the other is the variant of UAT-MC, denoted as **UAT**, which applies perturbations to both visual and textual modalities independently without coordination.

Implementation Details

During adversarial training phase, we set the maximum perturbation magnitude ϵ_m as 10% of the 2-norm of the in-



(a) MMGCN-Baby (λ) (b) MMGCN-Baby (α)

Figure 6: Trade-off between recommendation performance (NDCG@10) and adversarial robustness ($\text{Gain}_{\text{Hit@50}}$) under PGD(w/ $\mathcal{L}_{\text{Align}}$)-based attack with varying λ and α on the Baby dataset.

put embedding for modality m, the coefficient α is searched in $\{0.1,1,2,3,4,5,6,7,8,9,10\}$, the coefficient λ is searched in $\{1,2,3,4,5,6,7,8,9,10\}$.

5.2 Defense Effect (RQ1 and RQ2)

We conduct a systematic evaluation of the performance of different defense methods on two attack scenarios (FGSM-based and PGD-based) on two mainstream MRSs (VBPR [He and McAuley, 2016b] and MMGCN [Wei *et al.*, 2019]). The victim models take the original and perturbed multimodal embeddings as input, respectively. Meanwhile, we also report the recommendation performance on the clean test set to reflect the impact of adversarial training on overall recommendation quality. The results are shown in Table 2, from which we have the following observations:

- **Across all dataset-model combinations, unpopular items consistently exhibit vulnerability to promotion attacks (w/o $\mathcal{L}_{\text{Align}}$)**, which risk misleading the target model into treating them as popular items. It can be observed that perturbations generated by PGD are more effective in promoting the targeted items compared to those generated by FGSM. Statistically, on the Baby dataset, the average Hit@50 of unpopular items increases from 0.667% to 3.865% under FGSM-based perturbations, and further to 4.111% under PGD-based perturbations. This implies that more than 3.444% of users will be misled into recommending an item (e.g., targeted item) that should not appear in the recommendation lists.
- **Attacks equipped with $\mathcal{L}_{\text{Align}}$ consistently achieve higher promotion gains than their counterparts without alignment,** indicating that explicitly coordinating visual and textual perturbations leads to more effective promotion. This improvement is particularly evident for stronger attacks. For example, under PGD-based attacks on the Baby dataset, introducing $\mathcal{L}_{\text{Align}}$ increases the $\text{Gain}_{\text{Hit@50}}$ of VBPR from 516.15% to 520.99%, and that of MMGCN from 446.76% to 450.30%. Similar trends are observed on the Sports and Clothing datasets, where alignment consistently yields additional promotion gains across both VBPR and MMGCN.
- **After untargeted adversarial training (UAT), the MRSs exhibit significantly enhanced resistance to evasion-based promotion attacks.** The results show that the effectiveness of both FGSM-based and PGD-based attacks drops significantly after UAT. In the following, we analyze based on the average $\text{Gain}_{\text{Hit@50}}$ results of the two attacks (w/ $\mathcal{L}_{\text{Align}}$). On the Baby dataset, UAT reduces the GainHit@50 of VBPR from 503.39% to 288.83%, achieving a 42.62% relative reduction. For MMGCN, the value drops significantly from 315.72% to 8.42%, corresponding to a 97.33% reduction. Similar trends are observed on the Sports dataset, where VBPR and MMGCN experience reductions of 99.95% and 98.92%, respectively. On the Clothing dataset, UAT reduces $\text{Gain}_{\text{Hit@50}}$ by 76.50% for VBPR and 99.71% for MMGCN.
- **UAT with Multimodal Coordination can further enhance its defense capability.** According to our observations, the $\text{Gain}_{\text{Hit@50}}$ of MMGCN on the Sports dataset drops from 25.45% to 19.14%, while that of VBPR on the Clothing dataset decreases significantly from 898.38% to 27.70% after applying multimodal coordination. Notably, the improvement in robustness is achieved with minimal impact on recommendation performance.

5.3 Hyper-parameter Study (RQ3 and RQ4)

This section explores how the perturbation budget ϵ_m and the hyper-parameters λ and α in adversarial training influence the trade-off between adversarial robustness and recommendation performance.

- **Impact of ϵ_d and ϵ_a** To examine the robustness and generalization ability of UAT-MC, we conduct adversarial training with different defense perturbation budgets $\epsilon_d \in \{2.5\%, 5\%, 7.5\%, 10\%\}$, and evaluate the trained models under FGSM-based promotion attacks with varying attack budgets $\epsilon_a \in \{2.5\%, 5\%, 7.5\%, 10\%\}$. The results are visualized in Figure 5, where specific numerical values are detailed in the Appendix B.3. We observe a clear transition from dark blue to light blue/white as ϵ_d increases, indicating that our defense method effectively mitigates the promotion attack.
- **Impact of λ and α** We investigate the trade-off between adversarial robustness and recommendation performance by varying $\lambda \in [1, 10]$ and $\alpha \in [0.1, 10]$. Figure 6 (Baby dataset) and the Appendix B.1 (others) visualize this relationship, plotting recommendation performance (NDCG@10) against attack gain ($\text{Gain}_{\text{Hit@50}}$). A clear trade-off is observed: higher recommendation performance typically comes at the cost of increased vulnerability.

6 Conclusion

In this work, we address the vulnerability of MRSs to promotion attacks. Crucially, we verify the existence of cross-modal gradient mismatch in multi-user promotion settings and proposed UAT-MC to mitigate it via a novel gradient alignment regularization. Extensive experiments demonstrate the effectiveness of UAT-MC in defending against evasion-based promotion attacks.

Acknowledgments

This work is supported by Natural Science Foundation of Sichuan Province under grant 2024NSFSC0516 and National Natural Science Foundation of China under grant 61972270.

References

[Chen *et al.*, 2024] Lijian Chen, Wei Yuan, Tong Chen, Guanhua Ye, Nguyen Quoc Viet Hung, and Hongzhi Yin. Adversarial item promotion on visually-aware recommender systems by guided diffusion. *ACM Trans. Inf. Syst.*, 42(6), August 2024.

[Di Noia *et al.*, 2020] Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. Taamr: Targeted adversarial attack against multimedia recommender systems. In *2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, pages 1–8, 2020.

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

[Guo *et al.*, 2024] Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. Lgmrec: local and global graph learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8454–8462, 2024.

[He and McAuley, 2016a] Ruining He and Julian McAuley. 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*, pages 507–517, 2016.

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

[Hsiao *et al.*, 2022] Shao-Ping Hsiao, Yu-Che Tsai, and Cheng-Te Li. Unsupervised post-time fake social message detection with recommendation-aware representation learning. In *Companion Proceedings of the Web Conference 2022*, pages 232–235, 2022.

[Li *et al.*, 2025] Hongji Li, Hanwen Du, Youhua Li, Junchen Fu, Chunxiao Li, Ziyi Zhuang, Jiakang Li, and Yongxin Ni. Teach me how to denoise: A universal framework for denoising multi-modal recommender systems via guided calibration. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25, page 782–791, New York, NY, USA, 2025. Association for Computing Machinery.

[Lin *et al.*, 2023] Weilin Lin, Xiangyu Zhao, Yejing Wang, Yuanshao Zhu, and Wanyu Wang. Autodenoise: Automatic data instance denoising for recommendations. In *Proceedings of the ACM Web Conference 2023*, pages 1003–1011, 2023.

[Liu and Larson, 2021] Zhuoran Liu and Martha Larson. 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*, WWW '21, page 3590–3602, New York, NY, USA, 2021. Association for Computing Machinery.

[Liu *et al.*, 2024] Yifan Liu, Kangning Zhang, Xiangyuan Ren, Yanhua Huang, Jiarui Jin, Yingjie Qin, Ruilong Su, Ruiwen Xu, Yong Yu, and Weinan Zhang. Alignrec: Aligning and training in multimodal recommendations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 1503–1512, New York, NY, USA, 2024. Association for Computing Machinery.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Mu *et al.*, 2025] Lingyu Mu, Zhengxiao Liu, Zhitong Zhu, and Zheng Lin. Trust-grs: A trustworthy training framework for graph neural network based recommender systems against shilling attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12408–12416, 2025.

[Nguyen Thanh *et al.*, 2023] Toan Nguyen Thanh, Nguyen Duc Khang Quach, Thanh Tam Nguyen, Thanh Trung Huynh, Viet Hung Vu, Phi Le Nguyen, Jun Jo, and Quoc Viet Hung Nguyen. Poisoning gnn-based recommender systems with generative surrogate-based attacks. *ACM Trans. Inf. Syst.*, 41(3), February 2023.

[Ong and Khong, 2025] Rongqing Kenneth Ong and Andy W. H. Khong. Spectrum-based modality representation fusion graph convolutional network for multimodal recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25, page 773–781, New York, NY, USA, 2025. Association for Computing Machinery.

[Rendle *et al.*, 2009] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 452–461, Arlington, Virginia, USA, 2009. AUAI Press.

[Sun *et al.*, 2020] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1405–1414, 2020.

[Tang *et al.*, 2020] Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua. Adversarial training towards robust multimedia recommender system. *IEEE Transactions on Knowledge and Data Engineering*, 32(5):855–867, 2020.

[Wang *et al.*, 2022] Shilei Wang, Peng Zhang, Hui Wang, Hongtao Yu, and Fuzhi Zhang. Detecting shilling groups

in online recommender systems based on graph convolutional network. *Information Processing & Management*, 59(5):103031, 2022.

[Wang *et al.*, 2023] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. Dualgnn: Dual graph neural network for multimedia recommendation. *Multimedia, IEEE Trans. on (T-MM)*, 25(000):11, 2023.

[Wang *et al.*, 2024a] Zongwei Wang, Min Gao, Junliang Yu, Xinyi Gao, Quoc Viet Hung Nguyen, Shazia Sadiq, and Hongzhi Yin. Llm-powered text simulation attack against id-free recommender systems, 2024.

[Wang *et al.*, 2024b] Zongwei Wang, Junliang Yu, Min Gao, Hongzhi Yin, Bin Cui, and Shazia Sadiq. Unveiling vulnerabilities of contrastive recommender systems to poisoning attacks. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3311–3322, 2024.

[Wei *et al.*, 2019] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1437–1445, 2019.

[Wei *et al.*, 2020] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 3541–3549, New York, NY, USA, 2020. Association for Computing Machinery.

[Wu *et al.*, 2021] Chenwang Wu, Defu Lian, Yong Ge, Zhihao Zhu, Enhong Chen, and Senchao Yuan. 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*, SIGIR '21, page 1074–1083, New York, NY, USA, 2021. Association for Computing Machinery.

[Wu *et al.*, 2023] Chenwang Wu, Defu Lian, Yong Ge, Zhihao Zhu, and Enhong Chen. Influence-driven data poisoning for robust recommender systems. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):11915–11931, October 2023.

[Yang *et al.*, 2024] Shiyi Yang, Chen Wang, Xiwei Xu, Liming Zhu, and Lina Yao. Attacking visually-aware recommender systems with transferable and imperceptible adversarial styles. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2900–2909, New York, NY, USA, 2024. Association for Computing Machinery.

[Yu *et al.*, 2023] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM international conference on multimedia*, pages 6576–6585, 2023.

[Zhang *et al.*, 2021] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3872–3880, 2021.

[Zhang *et al.*, 2024] Jinghao Zhang, Yuting Liu, Qiang Liu, Shu Wu, Guibing Guo, and Liang Wang. Stealthy attack on large language model based recommendation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5839–5857, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Zhou *et al.*, 2023a] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions, 2023.

[Zhou *et al.*, 2023b] Hongyu Zhou, Xin Zhou, Lingzi Zhang, and Zhiqi Shen. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation. In *ECAI 2023*, pages 3123–3130. IOS Press, 2023.

[Zhou *et al.*, 2023c] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM web conference 2023*, pages 845–854, 2023.

A Related Work

A.1 Multimodal Recommender Systems

Multimodal recommender systems (MRSs) aim to enhance recommendation performance by integrating multiple modalities of items. For example, VBPR [He and McAuley, 2016b] extends traditional Matrix Factorization by integrating multimodal representations with item ID embeddings. GNN-based methods such as MMGCN [Wei *et al.*, 2019], GRCN [Wei *et al.*, 2020] and DualGNN [Wang *et al.*, 2023] encode representations for each modality to capture user's modal-specific preferences. Recent studies have sought to improve recommendation performance using various auxiliary information, such as the knowledge graph [Sun *et al.*, 2020], user co-occurrence graph [Zhou *et al.*, 2023b], item-item relation graph [Zhang *et al.*, 2021], self-supervised learning [Zhou *et al.*, 2023c], hypergraph [Guo *et al.*, 2024], multimodal alignment [Liu *et al.*, 2024], spectral theory [Ong and Khong, 2025]. However, while above research enhances the personalization by utilizing the auxiliary modal information, MRSs exhibit heightened vulnerability to the promotion attacks.

A.2 Promotion Attacks and Defenses

As targeted adversarial strategy, promotion attacks can be typically classified into two paradigms based on implementation mechanisms: **Poisoning-based promotion attacks** achieve their objectives by injecting fabricated user profiles into the training set. For example, GSPAttack [Nguyen Thanh *et al.*, 2023] proposes a generative surrogate-based poisoning framework for graph-based recommender systems (RSs). Infmix [Wu *et al.*, 2023] designs an influence-based threat estimator and a distribution-agnostic user generator to craft imperceptible yet impactful fake users. CLeaR [Wang *et al.*, 2024b] proposes a dual-objective attack framework that exploits representation dispersion and rank promotion, revealing the risks introduced by contrastive learning. Our work focuses on **evasion-based promotion attacks** that achieve the objectives at the inference phase by injecting imperceptible adversarial perturbations into the item's content. For visually-aware recommendation models, TAaMR [Di Noia *et al.*, 2020], applies FGSM and PGD attacks on pre-trained visual encoders to misclassify the targeted item into a popular item' category. AIP [Liu and Larson, 2021], IPDGI [Chen *et al.*, 2024], IPDGI [Chen *et al.*, 2024] and SPAF [Yang *et al.*, 2024] generate imperceptible and diverse adversarial images to increase the similarity between the targeted item and popular items. RecTextAttack [Zhang *et al.*, 2024] and TextSimu [Wang *et al.*, 2024a] use LLMs to manipulate item text content to boost their ranking in text-aware recommendation models, revealing new vulnerabilities introduced by large language models (LLMs).

To date, many studies have focused on defending against such attacks through **data filtering** or **robust training**. **Data filtering** methods aim to detect and remove adversarially manipulated items or inputs before they impact the recommendation model. For example, Re-writing Defense [Zhang *et al.*, 2024] utilizes GPT-3.5-turbo to rewrite the adversarial text to defense LLM-based RSs. RecMR [Hsiao *et al.*, 2022] uses an AutoEncoder as a detection encoding model to distinguish anomalies. NFGCN-TIA [Wang *et al.*, 2022] employs a GCN model to detect malicious users. **Robust training** enhances the model's resilience by incorporating adversarial examples during training or modifying the learning objectives to reduce sensitivity to perturbations. APT [Wu *et al.*, 2021] simulates the poisoning process by injecting fake user data to foster a more robust system. AutoDenoise [Lin *et al.*, 2023] addresses the challenge of highly dynamic data distributions by employing a deep RL-based framework.

However, to the best of our knowledge, existing studies on adversarial robustness have primarily focused on single modal recommender systems, without considering the joint presence of visual and textual modalities. From an adversarial perspective, launching attacks on both modalities simultaneously to achieve effective promotion is natural. This highlights the urgent need to investigate promotion attacks and defense specifically tailored to MRSs.

Hyperparameter	**Value / Search Space**
Common Settings (Shared)	
Embedding Size	64
Epochs	$[1, 1000]$
Stopping Step	100
Batch Size	2048 (Train), 4096 (Eval)
Optimizer	Adam
LR Scheduler	$[1.0, 50]$
Align Weight	$\{2.0, 1.0, 0.1, 0.01, 0.001, 0\}$
VBPR Specific	
Learning Rate	0.001
β	$\{2.0, 1.0, 0.1, 0.01, 0.001, 10^{-4}, 10^{-5}\}$
MMGCN Specific	
Layers	2
Learning Rate	$\{0.0001, 0.0005, 0.001, 0.005, 0.01\}$
β	$\{10^{-5}, 10^{-4}, 0.001, 0.01, 0.1, 0.5\}$

Table 3: Hyper-parameter search spaces of multimodal recommendation systems.

B Experiments

B.1 Hyper-parameter Study

MRSs Training

We follow MMRec [1] to save the model parameters at the point of best performance. The hyper-parameters search spaces are provided in the Table 3.

Impact of λ and α

Figures 9 and 10 illustrate the impact of hyperparameters on the trade-off between recommendation performance and adversarial robustness under PGD-based promotion attacks for VBPR and MMGCN, respectively. Specifically, the left columns of both figures demonstrate the effect of λ, while the right columns present the influence of α, which controls

[1] https://github.com/enoche/MMRec

the alignment loss between visual and textual gradients. The results are reported across three datasets: Amazon Baby (top row), Sports (middle row), and Clothing (bottom row).

B.2 Extended Analysis of User Group Overlap

To provide a comprehensive empirical basis for the *cross-modal gradient mismatch* identified in Section 3, we present the complete distribution of Jaccard Similarity coefficients between the visually-sensitive user subset $\mathcal{U}_v$ and the textually-sensitive user subset $\mathcal{U}_t$ across all three datasets (Amazon Baby, Sports, and Clothing) and two victim models (VBPR and MMGCN).

Figure 7 illustrates these distributions. Across all six experimental settings, the average Jaccard similarity remains consistently low to moderate, ranging from 0.262 to 0.575. This indicates that $\mathcal{U}_v$ and $\mathcal{U}_t$ are largely distinct groups. In other words, for a given targeted item, the users who are most susceptible to visual perturbations are rarely the same as those susceptible to textual perturbations. This distinctness fundamentally leads to the conflicting gradient directions during joint optimization. These extensive results further corroborate our motivation: without explicit coordination, multimodal perturbations naturally diverge due to the inherent discrepancy in user group dominance.

B.3 Impact of ϵ_d and ϵ_a

This section provides the detailed numerical results corresponding to the robustness analysis in the main paper. We evaluate the performance of UAT-MC under varying adversarial training budgets $\epsilon_d \in \{2.5\%, 5\%, 7.5\%, 10\%\}$ and attack budgets $\epsilon_a \in \{2.5\%, 5\%, 7.5\%, 10\%\}$.

The results are categorized based on the attack objective:

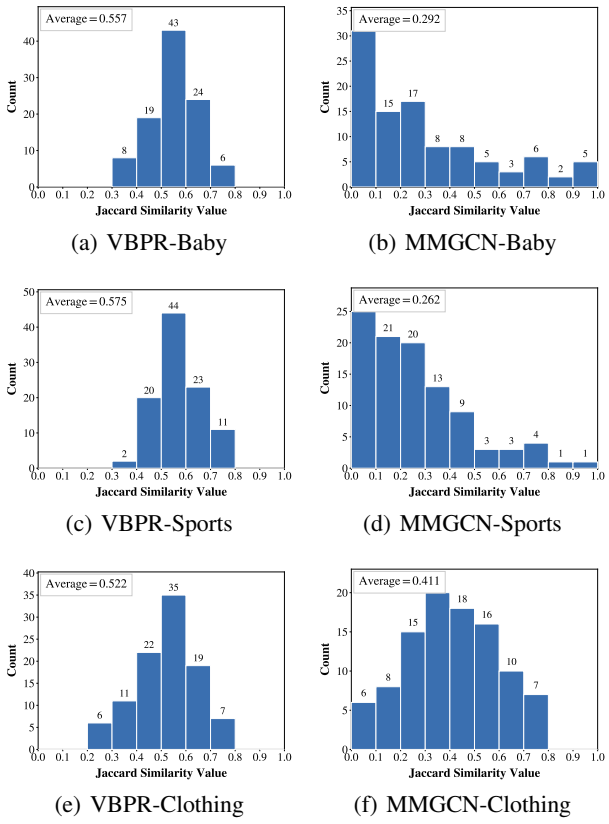
- **Table 4** presents the results under standard FGSM-based promotion attacks without the alignment loss.
- **Table 5** reports the $Gain_{Hit@50}$ under FGSM-based promotion attacks incorporating the gradient alignment loss $\mathcal{L}_{\text{Align}}$. These values correspond to the visualization in Figure 5.

Across both settings, we observe that increasing the defense budget ϵ_d consistently suppresses the attack gain, demonstrating the robustness of our method against different variations of promotion attacks.

B.4 Case Study

To illustrate how cross-modal gradient mismatch manifests during optimization and how gradient-level alignment reshapes the optimization dynamics, we randomly select an item i (ID: `B004203QQ4`, with 5 interactions) from the Baby dataset and conduct promotion attacks on VBPR. We compare PGD w/o $\mathcal{L}_{\text{Align}}$ and PGD w/ $\mathcal{L}_{\text{Align}}$, tracking (i) the user coverage, measured by N^i_{rec}, and (ii) the cosine similarity between visual and textual perturbation gradients.

As shown in Fig. 8, vanilla PGD gradually increases the cross-modal gradient cosine similarity but quickly saturates at a relatively low level, resulting in slower and less stable growth in promotion performance. In contrast, when $\mathcal{L}_{\text{Align}}$ is introduced, the gradient cosine similarity rises more rapidly and remains consistently higher. Consequently, PGD with $\mathcal{L}_{\text{Align}}$ achieves faster and more stable promotion gains. Overall, this case study provides intuitive evidence that explicitly enforcing gradient alignment mitigates cross-modal gradient mismatch and leads to more effective promotion attacks.



(a) VBPR-Baby (b) MMGCN-Baby

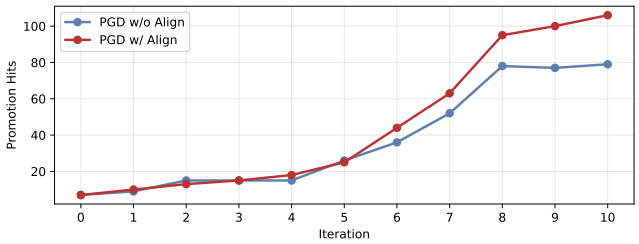
(c) VBPR-Sports (d) MMGCN-Sports

(e) VBPR-Clothing (f) MMGCN-Clothing

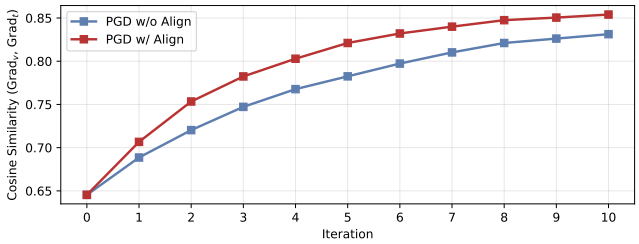
Figure 7: Distribution of Jaccard Similarity between $\mathcal{U}_v$ and $\mathcal{U}_t$.

C Computational Complexity Analysis

Let T_f and T_b denote the forward and backward cost of the underlying MRS for one batch, d the embedding dimension, and K the nu mber of inner maximization steps. Without adversarial training, the per-batch complexity is $O(T_f + T_b)$. If one explicitly computes per-user gradients for both modalities to identify U_v and U_t, the cost becomes $O(|U_p|T_b + |U_p|d + |U_p|\log|U_p|) \approx O(|U_p|T_b)$ per attack step, i.e., $O(K|U_p|T_b + T_f + T_b)$ per batch, which grows linearly with the number of users. In contrast, UAT-MC does not explicitly identify U_v or U_t. It only adds a gradient-alignment term on top of the already available batch-level modality gradients, introducing an extra cost of only $O(d)$. Therefore, its per-batch complexity is approximately $O((K+1)(T_f + T_b))$, i.e., the same order as standard adversarial training without explicit user-wise coordination. Hence, compared with explicit user-wise gradient analysis, UAT-MC removes the linear dependence on $|U_p|$ and introduces only a lightweight extra overhead.



(a) User coverage of item i (N^i_{rec})



(b) Cosine similarity between Γ_v and Γ_t.

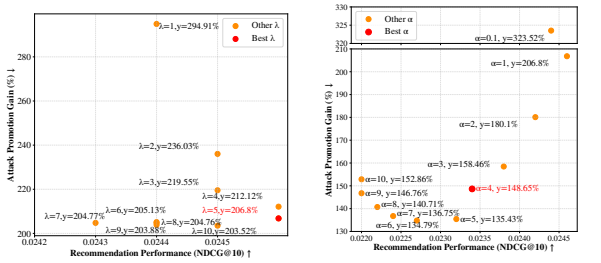
Figure 8: Case study of PGD-based promotion attacks on item `B004203QQ4` from the Baby dataset using VBPR, with and without the alignment loss $\mathcal{L}_{\text{Align}}$.

Dataset	Victim Model	VBPR				MMGCN			
	ϵ	ϵ_a=2.5%	ϵ_a=5%	ϵ_a=7.5%	ϵ_a=10%	ϵ_a=2.5%	ϵ_a=5%	ϵ_a=7.5%	ϵ_a=10%
Baby	w/o AT	54.25%	146.19%	288.11%	479.38%	81.49%	152.20%	176.08%	181.03%
	ϵ_d=2.5%	49.33%	138.78%	270.46%	456.33%	12.39%	21.72%	27.69%	31.13%
	ϵ_d=5%	43.01%	122.32%	243.43%	412.88%	5.65%	9.30%	13.07%	15.82%
	ϵ_d=7.5%	28.82%	73.92%	142.97%	233.41%	3.97%	7.37%	10.82%	12.94%
	ϵ_d=10%	**19.38%**	**48.45%**	**89.52%**	**143.98%**	**2.42%**	**4.32%**	**5.83%**	**7.21%**
Sports	w/o AT	734.90%	4636.57%	13506.93%	25722.99%	201.19%	505.93%	789.32%	1024.04%
	ϵ_d=2.5%	277.40%	1946.12%	6601.83%	14000.23%	30.33%	48.74%	51.46%	45.19%
	ϵ_d=5%	58.84%	292.82%	954.70%	2356.63%	16.90%	21.90%	26.43%	27.14%
	ϵ_d=7.5%	6.90%	14.37%	22.13%	34.48%	7.32%	8.87%	9.98%	10.42%
	ϵ_d=10%	**2.36%**	**4.72%**	**7.37%**	**10.32%**	**0.00%**	**3.91%**	**4.53%**	**6.38%**
Clothing	w/o AT	121.04%	519.48%	1508.35%	3558.26%	1072.07%	6460.36%	12045.05%	13282.43%
	ϵ_d=2.5%	79.76%	327.66%	1004.01%	2486.97%	128.08%	252.71%	303.45%	305.42%
	ϵ_d=5%	33.95%	100.20%	244.99%	522.09%	43.98%	61.28%	68.42%	71.80%
	ϵ_d=7.5%	13.35%	36.23%	67.37%	123.52%	30.04%	48.50%	51.93%	56.22%
	ϵ_d=10%	**2.96%**	**9.30%**	**16.70%**	**26.64%**	**21.65%**	**25.06%**	**28.04%**	**29.44%**

Table 4: Comparison of $\text{Gain}_{\text{Hit@50}}$ under different combinations of adversarial training budgets (ϵ_d) and attack budgets (ϵ_a) for FGSM-based promotion attacks *without* gradient alignment. Here, ϵ_d denotes the perturbation budget used during adversarial training, while ϵ_a represents the perturbation budget used during promotion attacks. The best defense performance under each setting is highlighted in bold.

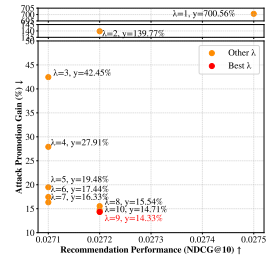
Dataset	Victim Model	VBPR				MMGCN			
	ϵ	ϵ_a=2.5%	ϵ_a=5%	ϵ_a=7.5%	ϵ_a=10%	ϵ_a=2.5%	ϵ_a=5%	ϵ_a=7.5%	ϵ_a=10%
Baby	w/o AT	54.81%	148.29%	291.56%	485.80%	81.68%	152.74%	176.58%	181.14%
	ϵ_d=2.5%	49.33%	138.78%	270.46%	456.33%	12.55%	22.02%	28.63%	32.24%
	ϵ_d=5%	43.01%	122.32%	243.43%	412.92%	5.71%	9.42%	13.21%	15.97%
	ϵ_d=7.5%	28.82%	73.92%	142.97%	233.41%	3.97%	7.39%	10.85%	13.05%
	ϵ_d=10%	**19.38%**	**48.47%**	**89.54%**	**144.00%**	**2.44%**	**4.35%**	**5.88%**	**7.47%**
Sports	w/o AT	754.29%	4737.40%	13737.12%	25876.73%	202.37%	512.76%	807.12%	1033.23%
	ϵ_d=2.5%	277.40%	1947.26%	6602.51%	14000.68%	31.59%	52.30%	52.72%	56.49%
	ϵ_d=5%	60.22%	296.96%	965.75%	2384.53%	17.38%	30.00%	30.48%	31.19%
	ϵ_d=7.5%	7.18%	14.66%	23.56%	35.34%	8.65%	11.75%	13.75%	14.63%
	ϵ_d=10%	**2.36%**	**4.72%**	**7.37%**	**10.32%**	**4.94%**	**6.58%**	**7.82%**	**8.44%**
Clothing	w/o AT	125.91%	534.78%	1554.78%	3682.09%	1091.44%	6598.65%	12292.34%	13670.72%
	ϵ_d=2.5%	79.76%	327.66%	1004.61%	2488.58%	132.02%	263.05%	327.59%	340.89%
	ϵ_d=5%	33.95%	100.61%	244.99%	522.70%	47.37%	70.30%	79.70%	80.83%
	ϵ_d=7.5%	13.35%	36.23%	67.58%	123.73%	32.19%	54.94%	57.08%	63.52%
	ϵ_d=10%	**3.38%**	**9.51%**	**16.70%**	**26.85%**	**22.51%**	**27.05%**	**29.53%**	**33.77%**

Table 5: Comparison of $\text{Gain}_{\text{Hit@50}}$ under different combinations of adversarial training budgets (ϵ_d) and attack budgets (ϵ_a) for FGSM-based promotion attacks *with* the gradient alignment loss $\mathcal{L}_{\text{Align}}$. Here, ϵ_d denotes the perturbation budget used during adversarial training, while ϵ_a represents the perturbation budget used during promotion attacks. A larger value of $\text{Gain}_{\text{Hit@50}}$ indicates higher attack effectiveness and thus weaker defense capability. The best defense performance under each setting is highlighted in bold.

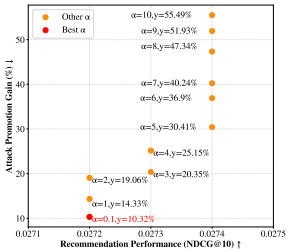


(a) VBPR-Baby (λ)

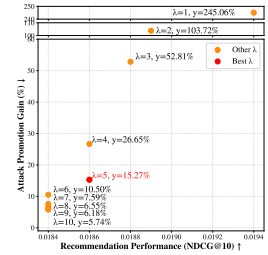
(b) VBPR-Baby (α)



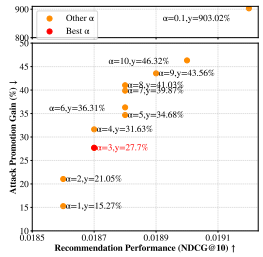
(c) VBPR-Sports (λ)



(d) VBPR-Sports (α)

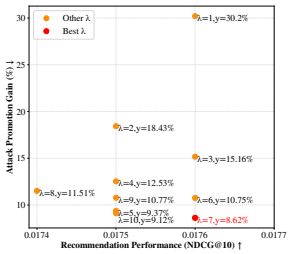


(e) VBPR-Clothing (λ)

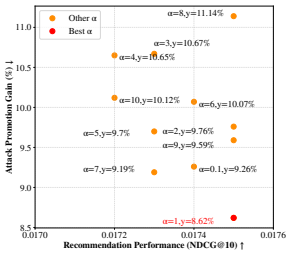


(f) VBPR-Clothing (α)

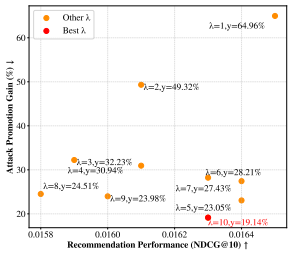
Figure 9: Impact of λ and α on the trade-off between accuracy and attack effectiveness for VBPR under PGD-based attack.



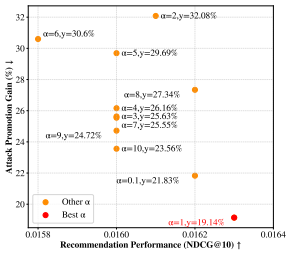
(a) MMGCN-Baby (λ)



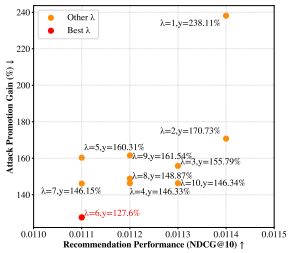
(b) MMGCN-Baby (α)



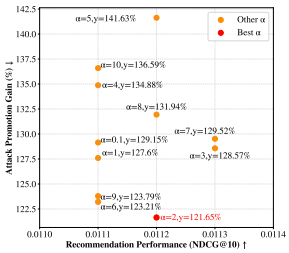
(c) MMGCN-Sports (λ)



(d) MMGCN-Sports (α)



(e) MMGCN-Clothing (λ)



(f) MMGCN-Clothing (α)

Figure 10: Impact of λ and α on the trade-off between accuracy and attack effectiveness for MMGCN under PGD-based attack.