

USENIX

THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

From Zero to Hero: Cross-modal-enhanced Adversarial Item Promotion Attack against Multimodal Recommender Systems

Mengyu Yao, *MOEKey Lab of HCST (PKU), School of Computer Science, Peking University; Ant Group;* Ziqi Zhang, *University of Illinois Urbana-Champaign;* Yifeng Cai and Junlin Liu, *MOEKey Lab of HCST (PKU), School of Computer Science, Peking University;* Xinyi Fu and Weiqiang Wang, *Ant Group;* Xiangqun Chen, Yao Guo, and Ding Li, *MOEKey Lab of HCST (PKU), School of Computer Science, Peking University*

https://www.usenix.org/conference/usenixsecurity26/presentation/yao-zero

This paper is included in the Proceedings of the 35th USENIX Security Symposium.

August 12–14, 2026 • Baltimore, MD, USA

ISBN 978-1-939133-58-8

Open access to the Proceedings of the 35th USENIX Security Symposium is sponsored by



جامعة الملك عبدالله للعلوم والتقنية

King Abdullah University of Science and Technology

ARTIFACT EVALUATED
USENIX
AVAILABLE

From Zero to Hero: Cross-modal-enhanced Adversarial Item Promotion Attack against Multimodal Recommender Systems

Mengyu Yao[1,2], Ziqi Zhang[3], Yifeng Cai[1], Junlin Liu[1], Xinyi Fu[2], Weiqiang Wang[2], Xiangqun Chen[1], Yao Guo[1], and Ding Li[1]

[1]*MOE Key Lab of HCST (PKU), School of Computer Science, Peking University*
[2]*Ant Group*
[3]*Department of Computer Science, University of Illinois Urbana-Champaign*

Abstract

Multimodal recommender systems (MRSs) jointly leverage visual and textual item representations to determine item ranking and exposure in modern online platforms. Their strong dependence on item content, however, introduces new security risks. In particular, malicious sellers can conduct the adversarial item promotion (AIP) attack to manipulate item content to deceive the recommender into overranking specific items. When it succeeds, the promoted items gain disproportionately higher exposure, leading to significant market visibility and direct economic gain. However, existing AIP research primarily targets unimodal recommenders, leaving the unique vulnerabilities of MRSs largely unexplored. Meanwhile, multimodal adversarial attacks for vision–language models (VLMs) optimize objectives unrelated to ranking and lack cross-modal coordination unique to MRSs. To bridge this gap, we propose CREAM (CROSS-MODAL-ENHANCED AIP ATTACK AGAINST MRSs). Our key insight is to *jointly perturb multiple modalities in a semantically consistent manner*. We integrate a tailored visual perturbator, a text generator, and a joint optimization controller to fully exploit cross-modal correlations in a black-box setting. Our comprehensive evaluation shows that CREAM significantly outperforms existing methods, achieving on average 5.75× and 2.89× higher exposure at top-10 and top-50 metrics, and demonstrates robustness under evolving real-world conditions. This exposes a tangible economic risk to recommender platforms. Meanwhile, CREAM maintains high imperceptibility across visual, textual, and cross-modal dimensions. We further investigate several potential defense strategies and demonstrate their limitations, highlighting the urgent need for stronger protections against adversarial threats in MRSs.

1 Introduction

Multimodal recommender systems (MRSs) have become a core infrastructure in modern online platforms [32, 80]. By jointly leveraging visual and textual item representations, MRSs determine item ranking and exposure, directly influencing sales, user engagement, and platform revenue. In highly competitive marketplaces such as Amazon [2], eBay [12], and Taobao [1], even small shifts in ranking can translate into substantial economic impact [9, 26, 27].

However, the dependence of MRSs on multimodal content introduces unique security vulnerabilities. Malicious sellers can exploit this reliance by subtly manipulating product images or textual descriptions to mislead the recommender into overranking specific items, which is an attack known as adversarial item promotion (AIP) [37]. Successful AIP attacks grant disproportionate visibility to targeted products, leading to inflated exposure and profits while undermining recommendation fairness and market integrity. In the long run, such manipulations threaten user trust and the sustainability of online ecosystems [23, 73, 77].

Existing AIP approaches fall short in capturing the inherently multimodal nature of MRSs. Prior attacks mainly target unimodal recommenders [10, 37, 42, 65], perturbing either visual or textual features in isolation, which fail to replicate the same level of attack effectiveness in MRSs. Meanwhile, multimodal adversarial attacks developed for vision–language models (VLMs) [11, 14, 16, 17, 25, 38, 55, 64, 69–71, 74] pursue semantic misalignment objectives (i.e., breaking image–text correspondence) rather than ranking-oriented goals. While effective for disrupting cross-modal alignment, these methods do not consider the cross-modal coordination unique to MRSs, where visual and textual features jointly determine ranking. As a result, they fail to exploit the joint optimization dynamics that govern multimodal ranking and are therefore ill-suited for effective and stealthy item promotion.

Therefore, to perform an effective AIP attack in MRSs, our key insight is to *jointly perturb multiple modalities in a semantically consistent manner*. By aligning perturbations across modalities in a coherent direction, the attack leverages inherent cross-modal dependencies to more effectively mislead the system. This strategy enables perturbations in different modalities to enhance each other in a synergistic way while maintaining stealthiness.

Achieving our insight faces three main challenges: ❶ *Semantic discrepancy across modalities:* visual and textual contents differ substantially in structure and semantic expressiveness, which makes it difficult to craft perturbations that remain meaningfully aligned. ❷ *Black-box constraint:* in realistic MRSs, model parameters and gradients are inaccessible, hindering efficient evaluation and guidance of adversarial perturbations. ❸ *Stealthiness requirement:* perturbations must stay *subtle and imperceptible* to preserve content naturalness and ensure real-world applicability.

To tackle these challenges, we propose a novel AIP framework for MRSs, termed CREAM (CROSS-MODAL-ENHANCED AIP ATTACK AGAINST MRSs). The framework targets both visual and textual modalities, which are the two most widely used content types in MRSs [72]. For ❶, we *leverage the cross-modal understanding capability of powerful pretrained models* to maintain semantic consistency. For the visual modality, a text-guided diffusion model generates image perturbations aligned with the associated text, while for the textual modality, an adversarial text generator based on the shared CLIP embedding space produces fine-grained, visually coherent text perturbations. For ❷, we *adopt an ensemble of locally trained surrogate models together with a joint optimization controller* to efficiently guide perturbation search under black-box setting. For ❸, we introduce *modality-specific constraints* to ensure perturbations remain natural and imperceptible. Through these designs, CREAM effectively addresses the challenges and achieves *semantic consistency*, *black-box adaptability*, and *stealthiness*.

We conduct comprehensive experiments on six mainstream MRSs across four e-commerce datasets. Two datasets come from the widely used Amazon Dataset [32, 45, 60, 76], while the other two are sourced from AliPay, a leading digital financial services platform in China that provides super-apps covering a wide range of merchandise categories [3, 4]. This real-world evaluation provides valuable insights into practical attack scenarios in the market. We compare our method against eight baseline methods, including unimodal and adapted multimodal approaches derived from existing non-MRS attacks. Experimental results demonstrate that CREAM substantially outperforms existing methods, achieving on average a 5.75× higher exposure rate at top-10 and a 2.89× higher exposure rate at top-50 compared to the best baseline. These results highlight its strong attack capability and underscore its potential real-world impact. It also achieves a favorable balance between effectiveness and stealthiness, showing strong performance across visual, textual and cross-modal metrics.

To assess real-world viability, we perform sandboxed evaluation on a commercial e-commerce platform using *production-grade model snapshots* collected over time. CREAM remains consistently effective over a 1.5-month period without access to future states, demonstrating robustness under evolving real-world conditions. We further evaluate several potential defense strategies and demonstrate their limitations.

In summary, the main contributions of our paper are:

- We present the first systematic investigation of AIP attacks in MRSs, revealing critical limitations of existing methods and identifying new opportunities to exploit vulnerabilities.
- We propose CREAM, a novel AIP framework tailored for MRSs. CREAM integrates newly designed unimodal perturbation modules and a joint optimization controller to exploit cross-modal interactions effectively.
- We conduct extensive evaluations across four real-world datasets and six model architectures under black-box settings. CREAM outperforms baselines by a large margin while maintaining high imperceptibility.
- We analyze multiple representative and adaptive defense strategies and demonstrate their limitations, highlighting the need for more robust protection mechanisms.

2 Background and Motivation

2.1 Multimodal Recommender Systems

Multimodal recommender systems (MRSs) have become the dominant recommendation paradigm in modern e-commerce platforms, where product images and textual descriptions jointly determine item relevance [19, 31, 60, 76]. By jointly leveraging multiple modalities, MRSs extract complementary semantic cues (i.e., visual attributes, styles, materials, and textual product details) that help mitigate the sparsity and cold-start issues commonly faced in real-world recommendation environments [32, 45, 80].

Formally, the MRS is trained on interaction data $\mathcal{D}$ and item multimodal content. $\mathcal{D} = \{(u, i, j) \mid u \in \mathcal{U},\ i \in I_u^+,\ j \in$

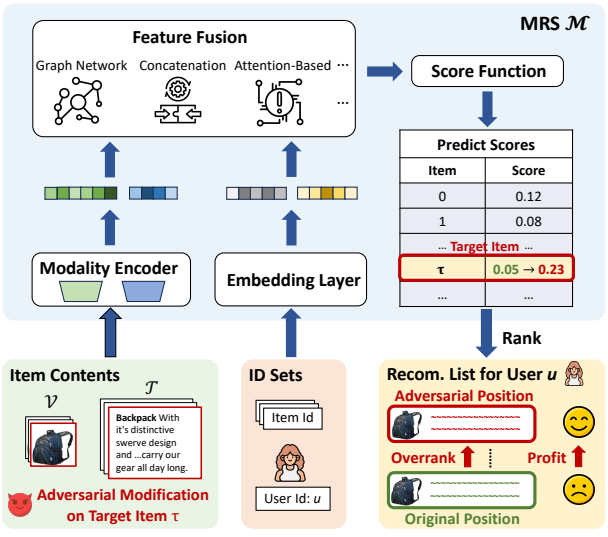


Figure 1: Workflow of MRS to recommend items for a user u (inference phase). The red text and box represent where the *Adversarial Item Promotion* attack takes place.

$I \setminus I_u^+\}$ denotes the pairwise training triplets, where $\mathcal{U}$ is the user set, I is the item set, and I_u^+ denotes the items interacted with by user u. We focus on MRSs that utilize visual and textual modalities, given their widespread use [80]. Let $\mathcal{V} = \{v_i | i \in I\}$ and $\mathcal{T} = \{t_i | i \in I\}$ denote item visual and textual contents, respectively. The MRS $\mathcal{M}$ is trained to learn a scoring function $\Gamma(u, i; \mathcal{M}) := \mathcal{M}(u, i, v_i, t_i \mid \mathcal{D}, \mathcal{V}, \mathcal{T})$ that predicts user u's preference for item i. Architectures for $\mathcal{M}$ vary in form, such as matrix factorization, attention networks, and graph-based models, and incorporate multimodal fusion at early, intermediate, or late stages [80].

Fig.1 illustrates the inference workflow of the MRS $\mathcal{M}$. During inference, for a given user u, all candidate items in the catalog are scored using $\mathcal{M}(u, i, v_i, t_i \mid \mathcal{D}, \mathcal{V}, \mathcal{T})$, where the MRS encodes the multimodal item content along with the user-item interaction context. The items are then ranked in descending order based on their predicted scores. Naturally, items ranked higher are more likely to attract users [45].

2.2 Adversarial Item Promotion

A common manipulation threat in these systems is adversarial item promotion (AIP) attack, where sellers intentionally manipulate item content to unfairly boost the ranking of their own products (referred to as target items) above those of competitors [37]. Such manipulation is inexpensive, easy to execute, and highly profitable: even small changes to product images or descriptions can significantly shift the multimodal representations used by the MRS, thereby influencing exposure and sales outcomes.

AIP attacks achieve this by modifying the content of the target item, as highlighted by the red text and box in Fig. 1. The attacker alters product images or textual descriptions of the target item τ before submitting them to the platform, aiming to mislead the recommendation system into assigning an inflated relevance score. Consequently, the item is overranked in the final recommendation list.

Existing AIP techniques focus almost exclusively on unimodal manipulation, and can be categorized into two types: image-based AIP [7, 10, 37, 65] and text-based AIP [42, 58]. Image-based AIP aims to mislead image-based RSs that incorporate visual information. Attackers add perturbations to the item images based on the RS output [37, 65] or maximize the visual similarity between the target item's image and popular images [7]. Text-based AIP utilizes large language models to rewrite the textual description of target items, either to mislead text-based RSs [42] or mimic popular product profiles [58].

However, these methods are intrinsically limited when applied to MRSs. The core challenge lies in the nature of MRSs, which jointly encode and fuse visual and textual data into a unified semantic representation. This fusion process creates redundancy and compensation across modalities, which can reduce the effectiveness of prior single-modality attacks that do not explicitly account for cross-modal interactions. This limitation highlights a critical gap: existing AIP approaches cannot reliably manipulate the behavior of modern MRSs, thereby calling for a new attack paradigm that explicitly models and controls cross-modal interactions.

2.3 Motivation

Although multimodal adversarial attacks have been widely explored in vision–language models (VLMs), these techniques are fundamentally incompatible with AIP in MRSs for two reasons: (1) misaligned optimization objectives, and (2) noticeable cross-modal inconsistency. These key gaps motivate the need for a new attack paradigm.

Misaligned Optimization Objectives. The first reason lies in the mismatch between the objectives of existing multimodal attacks and the requirements of MRSs. Specifically, existing multimodal attacks typically optimize for cross-modal inconsistency, which harms the latent-space coherence between modalities (a property crucial for item ranking in MRSs).

Perturbations in VLMs are designed to induce feature misalignment between different modalities [11, 14, 16, 17, 25, 38, 55, 64, 67, 69–71, 74], with some methods mentioning "semantic alignment" only as an auxiliary mechanism to stabilize such misalignment-based attacks [25, 67]. This strategy is effective in misleading tasks such as vision-language retrieval [56], visual entailment [62], and visual grounding [22]. As illustrated in the left panel of Fig. 2, such attacks inject adversarial noise that deliberately disrupts the alignment between visual and textual representations in the shared semantic feature space.

AIP attacks in MRSs, however, pursue a fundamentally different goal. MRSs extract semantic signals jointly from visual and textual content [80], and popularity bias across modalities naturally converges within a shared latent space. As shown in the right panel of Fig. 2, successful AIP attacks must guide both modalities collaboratively toward a *promotion zone* in the semantic space (i.e., regions associated with higher predicted relevance scores). Thus, rather than breaking semantic coherence, effective attacks must preserve cross-

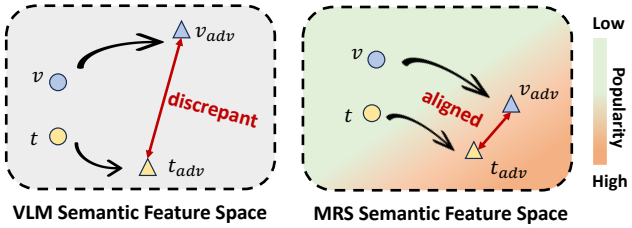


Figure 2: Different attack aims in VLM and MRS: Circles denote benign visual (blue) and textual (yellow) embeddings; triangles denote adversarial ones. In MRS, successful attacks require aligned perturbations toward promotion zone (orange), contrasting with the discrepant shifts in VLM.

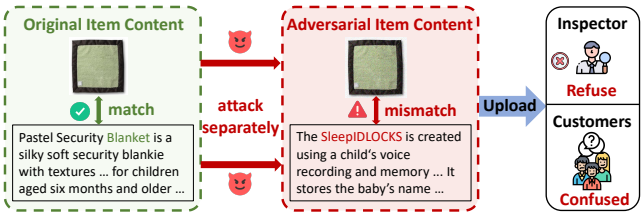


Figure 3: Noticeable cross-modal inconsistency illustration

modal alignment and strategically reinforce it to manipulate item rankings.

Noticeable Cross-modal Inconsistency. The second issue is that existing multimodal attacks often introduce visible inconsistencies between modalities. In real-world e-commerce environments, such discrepancies are easily detected by platform-side moderation systems or human users, severely undermining both attack effectiveness and platform credibility.

Fig. 3 illustrates a representative inconsistency case from the *Amazon Baby* dataset [18]. In this example, adapting a VLM attack [74] leads to adversarial text that drifts to an unrelated and confusing product ("SleepIDLOCKS") even though the image changes minimally. This mismatch arises because the text component attempts to mimic popular item descriptions without enforcing semantic compatibility with the image. Even if a listing bypasses automated checks, the inconsistency can still erode user trust because users tend to scrutinize details carefully before purchase [28, 61].

Insight. Given the limitation of existing multimodal attacks, our key insight is to *jointly perturb multimodal contents toward a semantically-aligned direction*. As shown in Fig. 2, existing multimodal attacks against VLMs often introduce divergent perturbations across modalities. In contrast, effective AIP attacks on MRSs require coherent shifts toward a shared *promotion zone*. Discrepant perturbations not only weaken ranking influence but also cause semantic inconsistencies that risk detection by users or automated filters (Fig. 3). Our joint optimization approach preserves cross-modal coherence and semantic alignment, thereby enhancing both the effectiveness and stealthiness of the attack.

Challenges. Realizing semantically consistent adversarial campaigns in MRSs presents three principal challenges:

❶ *Semantic discrepancy across modalities.* Visual and textual modalities differ fundamentally in structure and semantic richness. Images consist of dense, spatially organized pixels, while text is composed of discrete tokens governed by linguistic syntax and semantics. Textual descriptions often express explicit attributes such as function, style, or material, whereas visual cues are more ambiguous and context dependent. This discrepancy makes it inherently difficult to generate perturbations that remain meaningfully aligned across modalities, such as ensuring that a visual change (for example, a modified texture) corresponds to a semantically consistent textual edit.

To address this challenge, we *leverage pretrained models with strong cross-modal understanding*. Specifically, we employ a text-guided diffusion model to generate image perturbations aligned with the associated text and a CLIP-based adversarial text generator to produce fine-grained, visually coherent textual perturbations within the shared embedding space. Both components are jointly optimized with adversarial and guidance objectives to maintain semantic consistency.

❷ *Black-box constraint.* In real-world MRSs, model parameters and gradients are inaccessible, and probing through repeated uploads is costly and easily detectable. Moreover, commercial systems typically employ complex fusion mechanisms and user-behavior signals that cannot be approximated by a single surrogate model. We address these issues by *adopting an ensemble of locally trained surrogate models coupled with a joint optimization controller*, which collaboratively guide the search for effective perturbations and significantly improve query efficiency and robustness.

❸ *Stealthiness requirement.* To remain practical in real-world platforms, adversarial perturbations must appear natural and inconspicuous to both users and automated detectors. We therefore *introduce modality-specific similarity constraints* that bound the magnitude and perceptual footprint of perturbations. These constraints preserve visual fidelity and textual fluency, ensuring that adversarial content remains subtle, realistic, and difficult to detect.

3 Threat Model

Scenario. We focus on a realistic scenario where a malicious or fraudulent seller attempts to promote a specific item on an e-commerce platform that deploys the MRS. These systems rely on visual and textual content to generate personalized recommendations. The attacker operates as a legitimate user with full control over the content of their own product listings and may modify visual and textual features to influence rankings. Such manipulation reflects a form of algorithmic commercial fraud that exploits MRS ranking mechanisms for financial gain.

We consider a general and practical attack setting where the attacker may target any item on the platform, regardless of its popularity or interaction history. Once the content of a target item is altered, the platform re-extracts its multimodal features, which are then processed by the MRS to recalibrate item rankings. This setting aligns with real-world platform behavior and has been widely adopted in prior studies on AIP attacks [7, 10, 42, 58]. Compared to works that assume the attacker can artificially co-locate target items with popular ones [37, 65], we address a more general and challenging case aligned with how content-driven exposure typically unfolds in production systems.

Attacker's Objectives. The primary goal of the attacker is to increase the visibility of a target item τ by modifying its multimodal content so that it appears in more users' top-K

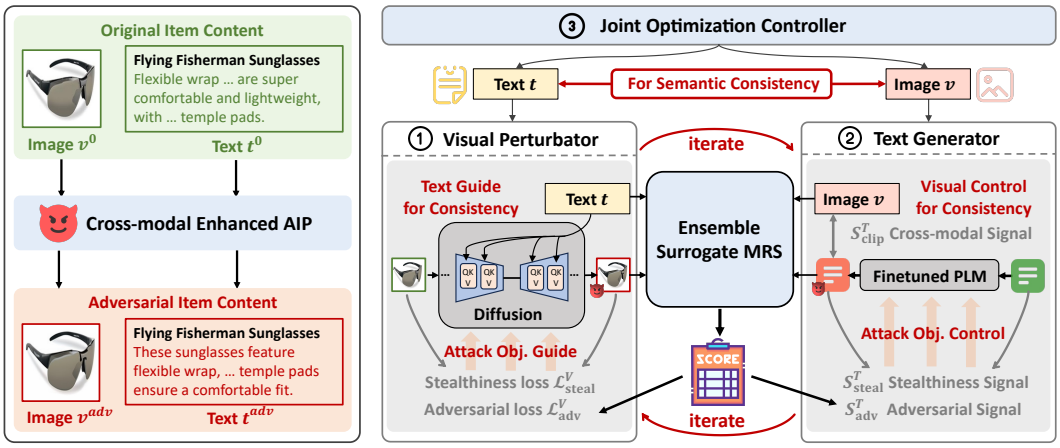


Figure 4: The workflow of CREAM. It consists of 3 modules: Visual Perturbator (`Module①`), Text Generator (`Module②`) and Joint Optimization Controller (`Module③`). An ensemble surrogate MRS serves as a proxy for evaluating attack effectiveness.

recommendation lists, i.e., to increase its exposure rate. This objective is formally defined as:

$$\max \ \mathrm{ER}_\tau@K = \frac{\sum_{u \in \mathcal{U}} |\widetilde{R}_u^K \cap \{\tau\}|}{|\mathcal{U}|}, \tag{1}$$

where $\widetilde{R}_u^K$ denotes the top-K recommendation list presented to user u after the adversarial modifications. In addition to effectiveness, the attacker aims to ensure stealthiness. First, the perturbations to item content should be imperceptible. Image modifications should avoid visible artifacts, and textual changes should preserve fluency and readability [7,42]. Second, the visual and textual content should remain semantically consistent. Inconsistencies across modalities may trigger detection by human reviewers or automated systems [34], thereby undermining user trust and platform credibility [8,28].

Attacker's Capabilities. We assume a *black-box attack scenario*, where the attacker has no access to the victim MRS's internal parameters, architecture, or training pipeline. Despite this limitation, the attacker has several realistic capabilities. First, by interacting with public APIs or user-facing interfaces, the attacker can collect publicly accessible data such as some user-item interactions and item content. Following prior work [33,42,65,77], this data may overlap with part of the MRS training data, enabling partial reconstruction of system behavior. Moreover, the attacker can exploit domain knowledge, including commonly used MRS architectures. While the victim's exact implementation remains unknown, such knowledge guides approximating its operational behavior.

4 Design of CREAM

We propose CREAM, an AIP framework that jointly perturbs visual and textual inputs in a semantically consistent direction to generate effective multimodal adversarial modifications.

4.1 Overview

The workflow of CREAM is illustrated in Fig. 4. As shown in the left panel, it takes an item's original image v^0 and its associated text t^0 as input, and generates adversarial counterparts v^{adv} and t^{adv} that mislead the victim MRS into overranking target items, while maintaining stealthiness. The core idea of CREAM is *iterative dual perturbation*: we iteratively perturb the image conditioned on the text, and perturb the text conditioned on the image, thereby preserving semantic alignment between the two modalities.

CREAM operates through the coordinated interplay of three core modules: the visual perturbator, the text generator, and the joint optimization controller (right panel of Fig. 4). These modules form a closed-loop system in which the image and text are alternately refined to maintain semantic alignment while jointly optimizing for attack effectiveness and stealthiness. This process is guided by an ensemble surrogate MRS, which provides feedback to steer the optimization under black-box constraints.

The system is organized into three modules, each fulfilling a distinct role in the attack framework:

- **Visual Perturbator (`Module①`).** This module generates adversarial image perturbations via a diffusion-based process, guided by dynamically updated adversarial text. This ensures the visual perturbation remains semantically consistent with the evolving text. The latent space is optimized with adversarial and stealthiness losses to achieve attack effectiveness and stealthiness.
- **Text Generator (`Module②`).** This module creates adversarial textual descriptions based on the perturbed image from the visual perturbator. By incorporating a CLIP-based signal, it ensures the generated text is consistent with the image. Additional attack-specific and stealthiness signals enhance effectiveness while maintaining stealthiness.

- **Joint Optimization Controller (`Module`③).** The central controller coordinates the Visual Perturbator and Text Generator through an alternating optimization strategy. This approach ensures iterative optimization of both image and text perturbations while preserving semantic consistency across modalities, thereby enhancing the overall effectiveness of the adversarial attack under black-box scenarios.

The rest of this section details the Visual Perturbator (Sec. 4.2), the Text Generator (Sec. 4.3), the Joint Optimization Controller (Sec. 4.4), and the surrogate model (Sec. 4.5).

4.2 Visual Perturbator

This module aims to generate imperceptible adversarial image perturbations that maintain semantic consistency with the associated adversarial text.

Existing visual perturbation methods often introduce visible distortions and fail to align well with textual inputs, resulting in attacks that are either ineffective or detectable in MRSs. To address these issues, we adopt a *diffusion-based framework* [21, 44, 46, 47]. Diffusion models support high-quality, diverse image synthesis through iterative denoising. Moreover, they offer a natural mechanism for incorporating text as a conditioning guide, making them suitable for enforcing semantic alignment during perturbation.

Our method builds on a pretrained Stable Diffusion model, leveraging its inherent cross-modal understanding, and follows a three-stage process. `Module`① of Fig. 4 shows the process, which is guided by dynamically updated adversarial text. First, the input image is inverted into latent space using DDIM Inversion [49]. Then, adversarial optimization is performed in the latent space with two objectives: an adversarial loss to maximize attack strength and a stealthiness loss to preserve visual imperceptibility. Finally, the optimized latent embedding is decoded through a text-guided denoising process, ensuring the cross-modal consistency.

Decoupling semantic guidance from adversarial optimization allows the visual perturbator to control perturbation magnitude while keeping outputs natural and text-consistent. In particular, we separate *text-guided diffusion for cross-modal consistency* and *attack-objective-guided latent optimization*.

Text-guided Diffusion for Cross-modal Consistency. To enforce semantic consistency between image perturbations and adversarial text, we condition the denoising trajectory of a latent diffusion model on dynamically updated text. Instead of sampling from pure noise, we initialize the generation with a latent representation obtained via DDIM Inversion, which maps an input image into a noisy latent representation at a target diffusion step N.

Let $\mathbf{z}_N$ denote the inverted latent obtained from the original image using DDIM Inversion. This process approximates the reverse diffusion path by deterministically projecting the image into the latent space at step N, providing a suitable starting point for controlled semantic manipulation.

Based on $\mathbf{z}_N$, we run the reverse denoising process using a U-Net-based model conditioned on an adversarial text t. The text is encoded into an embedding $\mathbf{e}_t \in \mathbb{R}^{d_t}$ via a text encoder, where d_t is the dimensionality of the embedding space. This embedding is integrated into the intermediate layers of the U-Net using a cross-attention mechanism, which allows the model to focus on relevant textual features during denoising.

The attention is computed as:

$$\text{softmax}\left(\left(W_Q^n \cdot \psi(\mathbf{z}_n)\right)\left(W_K^n \cdot \mathbf{e}_t\right)^\top / \sqrt{d}\right) \cdot W_V^n \cdot \mathbf{e}_t, \quad (2)$$

where $\mathbf{z}_n$ is the n-th UNet decoder feature map, $\psi(\mathbf{z}_n)$ is its flattened intermediate representation following [46], and W_Q^n, W_K^n, W_V^n are projection matrices with d denoting the feature dimension. This mechanism aligns visual generation with the semantics of the adversarial text, guiding the denoising process to preserve cross-modal coherence. It is important to note that semantic conditioning alone does not induce adversarial behavior. The process of optimizing for direct attack objectives is described in the next section.

Attack-objective-guided Latent Optimization. To embed adversarial intent into the generated image, we conduct latent-space optimization. While text conditioning ensures semantic alignment, this step delivers the specific adversarial effect.

Specifically, given the latent representation $\mathbf{z}_N$ obtained via DDIM Inversion, we iteratively optimize it to produce a perturbed representation $\mathbf{z}_N^*$ that maximizes the adversarial impact. During each optimization step, the current latent representation is decoded into an image $\tilde{v}$, which, along with the corresponding adversarial text, is fed into the surrogate MRS. Based on the model's output, we compute an attack loss $\mathcal{L}_{\text{adv}}^V$ that measures the effectiveness of the adversarial input in manipulating the MRS ranking.

Since the primary attack objective (Eq. 1) is discrete, we approximate it using the MRS's score function, following previous unimodal AIP work [37, 42]. The adversarial loss is formulated as:

$$\mathcal{L}_{\text{adv}}^V = \sum_{u \in \mathcal{U}'} -\mathcal{M}_S\left(u, \tau, \tilde{v}, t | \mathcal{D}', \mathcal{V}_{I'}, \mathcal{T}_{I'}\right), \quad (3)$$

where $\mathcal{M}_S$ represents the surrogate MRS. $\mathcal{U}'$, I' and $\mathcal{D}'$ are subsets of users, items, and data.

To preserve perceptual quality, we incorporate a stealthiness loss based on self-attention consistency, denoted as $\mathcal{L}_{\text{steal}}^V$. During DDIM Inversion, we record the self-attention maps Λ_n at each inversion step. These inversion maps are then compared with denoise maps Λ_n^{adv}, which are extracted during the denoising process of the optimized adversarial latent. The loss is defined as:

$$\mathcal{L}_{\text{steal}}^V = \sum_{n \in \{1,2,\ldots,N\}} ||\Lambda_n - \Lambda_n^{adv}||^2. \quad (4)$$

This loss encourages the perturbed image to retain the structural characteristics of the original [6].

The overall objective balances adversarial success and visual imperceptibility:

$$\arg\min_{\mathbf{z}_N} \ \alpha_V \mathcal{L}^V_{\text{adv}} + \beta_V \mathcal{L}^V_{\text{steal}} \ , \tag{5}$$

where α_V, β_V are hyper-parameters to control the trade-off between effectiveness and visual imperceptibility.

This optimization process yields adversarial images that are structurally consistent with the corresponding text and capable of stealthily misleading the MRS into inflating the target item's ranking.

4.3 Text Generator

The Text Generator produces adversarial descriptions that remain fluent, image-consistent, and difficult to detect by users. It aims to subtly change the textual modality while preserving the overall semantics of the original listing.

We implement this goal with a two-stage pipeline consisting of (i) domain-oriented finetuning and (ii) controlled adversarial text generation. Domain-oriented finetuning adapts a pretrained language model to domain-specific data from the target marketplace, while controlled generation injects attack and stealthiness signals during decoding (`Module`② in Fig. 4).

We *reframe the adversarial text generation as a completion problem* to improve fluency and controllability. Compared with rewriting-based approaches [42, 58], completion-based generation better leverages pretrained language priors and reduces incoherent or unnatural edits. We guide token selection with three complementary signals to balance effectiveness and stealthiness. Specifically, we incorporate a CLIP-based signal to align generated tokens with the associated image, an adversarial signal to influence ranking, and a stealthiness signal to keep the generated text close to the original description.

Stage I: Domain-Oriented Finetuning. In the first stage, we finetune the PLM on title-description pairs from the recommendation dataset. This adaptation aligns the model to the target task and domain, enabling it to generate in-context descriptions relevant to platform-specific content.

Attackers can easily obtain the finetuning data by scraping product listings or using platform-provided APIs [42]. The data is denoted as $\mathcal{D}_f = \{title_i \ \langle\text{_PAD_}\rangle \ desc_i \mid i \in I\}$, where $title_i$ and $desc_i$ are the title and description of item i, and $\langle\text{_PAD_}\rangle$ indicates the transition point from title to description. The finetuned model is then used for adversarial text generation.

Stage II: Controlled Adversarial Text Generation. At this stage, we utilize the finetuned model to generate adversarial descriptions that preserve cross-modal consistency with the associated image and fulfill the adversary's objectives. Formally, given a text prefix $t_{<l}$, the next token w_l is selected as:

$$w_l = \arg\max_{w \in W^{(k)}} \left\{ \log p_\theta(w|t_{<l}) + S^T(w) \right\}, \tag{6}$$

where $W^{(k)}$ denotes the top-k candidate vocabulary set given the prefix $t_{<l}$ and $p_\theta(w|t_{<l})$ represents the model's probability of token w conditioned on the preceding context. The term S^T is a composite signal that integrates three constraint components, detailed below. Each signal component is normalized via softmax over the candidate set $W^{(k)}$ to yield a relative distribution that guides token selection.

Cross-modal Signal. To ensure that the generated text remains semantically aligned with the associated image v, we introduce a CLIP-based cross-modal consistency signal:

$$S^T_{\text{clip}}(w|v, t_{<l}, W^{(k)}) = \frac{e^{\text{CLIP}(v,[t_{<l}:w])}}{\sum_{y \in W^{(k)}} e^{\text{CLIP}(v,[t_{<l}:y])}}, \tag{7}$$

where $\text{CLIP}(\cdot,\cdot)$ denotes the similarity between image v and the generated text fragment, as computed by a pretrained CLIP model [20, 30]. We use the CLIP-based similarity because the pretrained CLIP model understands visual and textual semantics, providing a strong foundation for evaluating how well the generated text reflects the image. By incorporating this signal, we ensure that each token incrementally reinforces visual-textual alignment, thus improving attack effectiveness and reducing detectability in MRS.

Adversarial Signal. To manipulate the ranking of the target item τ, we define an adversarial signal:

$$S^T_{\text{adv}}(w|v, t_{<l}, W^{(k)}) = \frac{e^{\text{ADV}(v,[t_{<l}:w])}}{\sum_{y \in W^{(k)}} e^{\text{ADV}(v,[t_{<l}:y])}}, \tag{8}$$

where

$$\text{ADV}(v, [t_{<l} : y]) = \sum_{u \in \mathcal{U}'} |\widetilde{R}^K_u(v, [t_{<l} : y]) \cap \{\tau\}| / |\mathcal{U}'|. \tag{9}$$

This aligns with the adversary's goal in Eq. 1. Here, $\widetilde{R}^K_u(v, [t_{<l} : y])$ denotes the top-K ranked list generated by the surrogate model $\mathcal{M}_S$ for user u, when the target item τ is associated with image v and textual fragment $[t_{<l} : y]$. The evaluation is conducted over subsets of users $\mathcal{U}'$ and items I'.

Stealthiness Signal. To ensure that the generated text is close to the original description t^0, we define a stealthiness signal:

$$S^T_{\text{steal}}(w|t^0, t_{<l}, W^{(k)}) = \frac{e^{\cos(\mathcal{E}_\mathcal{T}(t^0), \mathcal{E}_\mathcal{T}([t_{<l}:w]))}}{\sum_{y \in W^{(k)}} e^{\cos(\mathcal{E}_\mathcal{T}(t^0), \mathcal{E}_\mathcal{T}([t_{<l}:y]))}}, \tag{10}$$

where $\mathcal{E}_\mathcal{T}$ is a text encoder and $\cos(\cdot,\cdot)$ is cosine similarity.

Final Composite Signal. The final signal is a weighted sum:

$$S^T(w) = \alpha_T \cdot S^T_{\text{clip}}(w) + \beta_T \cdot S^T_{\text{adv}}(w) + \gamma_T \cdot S^T_{\text{steal}}(w), \tag{11}$$

where α_T, β_T, and γ_T control the trade-off among visual consistency, direct adversarial strength and semantic preservation.

To reduce computational workload, the signals are computed intermittently (e.g., every 10 tokens) with randomized intervals to enhance output diversity. This design efficiently generates adversarial text that is visually grounded, promotion-effective, and stealthy.

4.4 Joint Optimization Controller

The Joint Optimization Controller enforces global cross-modal coherence by coordinating image and text updates through alternating optimization. Without this controller, the Visual Perturbator and Text Generator would optimize their modalities independently under static conditions, which can lead to locally plausible but globally misaligned multimodal edits in black-box MRSs.

The controller performs alternating updates between the image and text with the core idea of iterative dual perturbation. The overall optimization runs for a total of R rounds or until convergence. In each round r, it first perturbs the image using the Visual Perturbator. Since image features reside in a continuous space, they are more responsive to gradient-based updates and often establish a strong adversarial direction. The effectiveness of this visual perturbation is immediately evaluated. If it improves the adversarial score, the perturbed image is retained. The controller then updates the text based on the new image using the Text Generator.

To ensure stable optimization and reduce redundant computation, the controller terminates early when improvements saturate or when both modalities have stabilized. Specifically, we stop if the adversarial score fails to improve beyond a tolerance ε for P consecutive rounds, or if both the image perturbation norm $\|v_r - v_{r-1}\|$ and the normalized text edit distance $\textsc{EditDist}(t_r, t_{r-1})$ fall below thresholds τ_v and τ_t.

4.5 Design of Surrogate Model

To enable efficient optimization under a black-box scenario, we adopt a surrogate model that approximates the behavior of victim MRS. This model acts as a proxy for the target system, guiding the generation of adversarial perturbations without requiring repeated access to the real-world platform.

Training Data. Following prior work [42, 65], we assume the attacker can reasonably obtain a partial set of item content and user interactions via public APIs or web scraping. This reflects realistic access patterns on real-world platforms. The collected data is then used to train a surrogate model.

Ensemble-Based Architecture. While prior AIP studies often employ a single model architecture as the surrogate [42, 65], this strategy is less effective for MRSs due to their architectural diversity. To address this, we construct an ensemble-based surrogate model that combines multiple architectures with diverse fusion and representation learning strategies. The goal is to improve the robustness and transferability of adversarial samples [35, 50, 54]. By aggregating predictions from multiple models, the ensemble better captures heterogeneous ranking behaviors and increases the likelihood of successful attack against black-box victims.

5 Evaluation

In this section, we perform comprehensive experiments to answer the following research questions:
RQ1: How effective is CREAM compared to existing methods across model architectures and datasets?
RQ2: How stealthy is CREAM, considering both unimodal and cross-modal imperceptibility?
RQ3: How does CREAM perform in the real world?
RQ4: How does CREAM perform under defenses?
RQ5: What is the contribution of each component?

5.1 Experiment Setup

Dataset. We evaluate CREAM on four datasets, including two widely used public benchmarks and two proprietary datasets collected from a real-world e-commerce platform, AliPay. The public datasets, *Baby* and *Beauty*, are standard benchmarks in MRS research and are collected from the *Amazon* platform [7, 18, 32, 80]. The proprietary datasets, *Office* and *Pet*, were collected from AliPay over a three-month period beginning in Oct 2024, covering the categories of office supplies and pet products, respectively. AliPay is a large-scale e-commerce platform with over one billion users and a diverse product catalog offered through its super-app ecosystem. Among the datasets, *Office* is the largest, containing 26,315 users and 272,991 user-item interactions. The statistics of the datasets are summarized in Tab. 1.
Victim MRSs. We select six state-of-the-art MRSs architectures to evaluate CREAM: VBPR [19], MMGCN [60], DMRL [31], AMR [51], GRCN [59] and MMGCL [68]. These models adopt different MRS fusion strategies: early (VBPR, AMR), intermediate (DMRL, MMGCL), and late (MMGCN, GRCN). They also use different modeling techniques: matrix factorization (VBPR, AMR), graph neural network (MMGCN, GRCN, MMGCL), and attention mechanism (DMRL). The diverse models can comprehensively evaluate the effectiveness and scalability of CREAM. We confirm with the AliPay team that their production RSs are built upon architectures closely related to these base designs. To further assess practical feasibility, we conduct additional experiments in a production-grade sandbox environment using internal snapshots of deployed models and platform configurations used in AliPay's live system.
Baseline Methods. We select two kinds of baselines: unimodal attacks and multimodal attacks. For unimodal attacks,

Table 1: Statistics of experimental datasets.

Dataset	#Users	#Items	#Interactions	Sparsity
Baby	19372	7025	160189	99.88%
Beauty	22332	12086	198215	99.93%
Office	26315	4363	272991	99.76%
Pet	1613	648	12532	98.80%

we include three state-of-the-art image-based methods (INSA [37], SPAF [65], IPDGI [7]) and two representative text-based attacks (ATR [42], TextSimu [58]).

We incorporate representative multimodal attack frameworks originally proposed for VLMs. However, their single-modality objectives were not designed for ranking-based exposure, making them unsuitable and yielding poor AIP performance when applied directly. For a fair comparison, we adapt three state-of-the-art methods, TMM [55], SGA [38], and VQAttack [70], by replacing their internal single-modality perturbation modules with the corresponding best unimodal AIP implementations while retaining their original multimodal coordination mechanisms.

Implementation and Hyperparameters. Following prior work [33, 42, 65, 77], we assume the attacker has partial access to interaction data, parameterized by $p\%$. This assumption does not imply direct access to internal logs but models knowledge that is obtained indirectly, such as by issuing queries via controlled accounts or scraping public interaction traces [33, 77]. We set p as 10% to reflect a realistic and conservative threat model. Compared to prior studies that assume access to 100% [23] or 50% [65] of the data, or absolute quantities such as 100,000 [77] or 250,000 [42] interactions, our setting places a significantly tighter constraint on attacker capability. In our experiments, this corresponds to approximately 1,200 to 27,000 interactions. This represents a practical yet restricted level of adversarial access.

The surrogate ensemble integrates VBPR, DMRL, and MMGCN, representing different fusion strategies. For a fair comparison, all baseline methods are evaluated under the same surrogate ensemble setting.

Evaluation Metrics. We evaluate exposure using the standard ER@K (exposure rate at top-K), which aligns with the attack objective in Eq. 1, and use NDCG@K as a complementary metric to assess the ranking quality of target items [24]. For offline experiments, we report their relative forms, RER@K and RNDCG@K, which measure the multiplicative improvement of exposure and ranking quality compared with the benign state. For the real-world evaluation on temporal production snapshots, we report the absolute exposure gain ΔER@K, which more appropriately reflects changes in exposure when model states and serving distributions evolve over time.

We report results at $K = 10, 50$ to analyze ranking performance at commonly-used recommendation exposure levels [77]. In practice, $K = 50$ roughly corresponds to the first two pages on desktop and about six screens on mobile, which represent realistic browsing depths for most users [13]. Moreover, for most datasets in our evaluation, the top-50 region accounts for roughly $\sim$1% of the item set (except for the smaller Pet catalog), capturing the economically salient exposure frontier.

5.2 RQ1: Effectiveness

Table 2 presents detailed RER@K results across six model architectures and four datasets. Figure 5 compares *Original* (benign), *Best Unimodal*, *Best Multimodal*, and *Ours* across datasets. Averaged over all four datasets, CREAM achieves an overall mean RER@10 of 54.586, RER@50 of 13.826, RNDCG@10 of 87.818, and RNDCG@50 of 20.763, while the best baseline reaches only 9.490, 4.780, 11.953, and 5.413, respectively. This corresponds to 5.75$\times$ and 2.89$\times$ higher exposure at top-10 and top-50 (RER@10/50), and 7.35$\times$/3.84$\times$ higher RNDCG@10/50. These results confirm that CREAM provides a decisive advantage over unimodal or uncoordinated multimodal attacks.

Baseline Analysis. Unimodal baselines (INSA, SPAF, IPDGI, ATR, and TextSimu) show limited and inconsistent gains. INSA's simple pixel perturbations are relatively more transferable in black-box settings but still fall well short of CREAM due to the lack of cross-modal semantics. Multimodal baselines (TMM-AIP, SGA-AIP, VQA-AIP) are stronger on average, yet they often underperform because their visual and textual edits are not explicitly aligned, leading to divergent signals that blunt cross-modal amplification. In contrast, CREAM coordinates perturbations in a shared semantic space, so visual and textual changes reinforce each other, yielding superior results across datasets and architectures.

Architecture Trends and Generalization. CREAM generalizes well to unseen victim models, including AMR, GRCN,

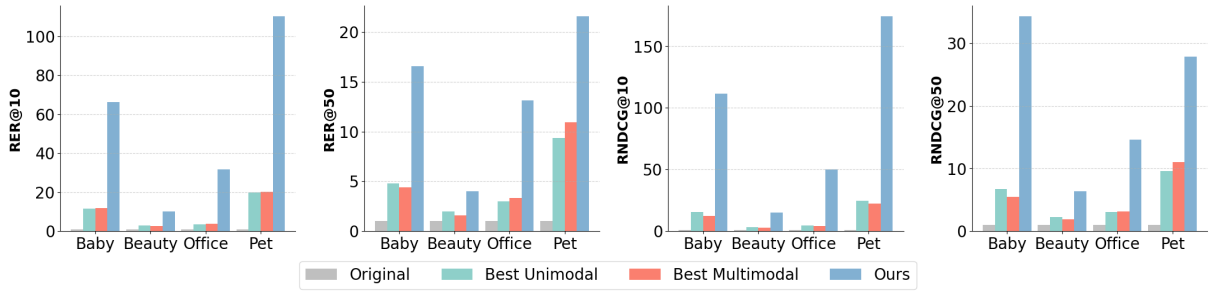


Figure 5: Average RER@K and RNDCG@K on different datasets for Original, Best Unimodal, Best Multimodal, and CREAM. CREAM consistently achieves the largest gains.

Table 2: RER@K across different datasets and model architectures under various attack methods. † marks unseen victim models; best results are highlighted in **bold**; second best results are underlined. Overall, our method outperforms the baselines across most datasets and models, with only one exception.

Dataset	Method	VBPR		MMGCN		DMRL		AMR†		GRCN†		MMGCL†	
		K=10	K=50	K=10	K=50	K=10	K=50	K=10	K=50	K=10	K=50	K=10	K=50
Baby	INSA	4.560	2.403	15.090	6.052	0.887	0.903	2.904	1.970	44.627	16.444	0.964	1.015
	SPAF	1.395	1.105	0.690	0.866	0.567	0.599	0.686	0.692	0.531	0.555	0.759	0.856
	IPDGI	1.214	1.094	1.048	1.040	0.905	0.930	0.992	0.987	1.221	1.159	0.926	0.956
	ATR	0.900	0.907	0.923	0.907	0.756	0.701	0.745	0.803	0.740	0.726	0.915	0.915
	TextSimu	1.074	1.014	2.242	1.482	0.855	0.867	0.982	0.998	2.151	1.873	1.000	0.988
	TMM-AIP	4.722	2.362	11.453	4.816	0.743	0.736	2.394	1.717	32.016	12.267	0.840	0.919
	SGA-AIP	4.822	2.433	14.142	5.160	0.661	0.691	2.402	1.716	47.872	15.436	0.879	0.947
	VQA-AIP	4.755	2.417	12.632	4.878	0.696	0.706	2.467	1.757	29.471	11.406	0.843	0.927
	Ours	**9.457**	**3.854**	**106.487**	**23.711**	**1.369**	**1.314**	**7.863**	**4.130**	**270.196**	**65.049**	**1.309**	**1.303**
Beauty	INSA	2.456	1.560	2.473	1.815	1.031	0.920	3.009	1.890	8.007	4.496	1.106	1.121
	SPAF	1.617	1.225	1.470	1.313	0.857	0.693	1.158	0.993	0.947	0.878	0.856	0.907
	IPDGI	1.297	1.121	0.969	0.982	1.046	0.949	1.071	0.993	1.178	1.114	0.996	1.006
	ATR	0.831	0.846	0.744	0.755	0.684	0.595	0.761	0.740	0.614	0.547	0.837	0.907
	TextSimu	0.945	0.965	1.389	1.231	1.109	0.928	0.880	0.879	1.617	1.458	0.961	1.002
	TMM-AIP	2.215	1.483	1.883	1.405	0.451	0.462	2.865	1.778	4.621	2.772	0.863	0.995
	SGA-AIP	2.237	1.546	2.118	1.528	0.507	0.499	2.734	1.758	5.732	3.239	0.942	1.033
	VQA-AIP	2.362	1.572	2.056	1.488	0.758	0.617	3.072	1.827	5.760	3.062	1.037	1.063
	Ours	**2.467**	**1.644**	**6.731**	**3.212**	**1.534**	**1.136**	**4.560**	**2.366**	**44.761**	**14.577**	**1.339**	**1.189**
Office	INSA	1.846	1.263	3.382	3.170	1.670	4.565	5.496	3.149	8.006	4.617	0.971	1.096
	SPAF	1.525	0.994	2.017	2.141	1.157	0.441	0.615	0.518	2.055	1.902	0.721	0.718
	IPDGI	1.115	1.404	0.843	0.700	1.086	0.773	0.803	0.830	0.499	0.679	0.909	0.914
	ATR	1.038	0.951	0.868	1.007	1.003	0.506	1.047	0.913	1.245	0.961	0.902	0.865
	TextSimu	1.017	1.017	0.557	0.664	0.659	0.698	0.987	1.001	0.946	0.946	0.919	0.938
	TMM-AIP	1.781	1.193	3.759	4.627	0.679	5.102	5.943	3.116	8.912	4.734	0.794	1.084
	SGA-AIP	1.650	1.207	3.174	2.930	1.166	5.733	6.099	3.284	7.671	4.462	0.798	1.036
	VQA-AIP	1.935	1.257	3.280	2.917	0.612	3.884	5.992	3.022	8.539	4.291	0.815	1.022
	Ours	**5.806**	**1.666**	**31.827**	**17.002**	**22.530**	**23.454**	**28.520**	**9.408**	**99.390**	**25.295**	**2.475**	**2.060**
Pet	INSA	24.897	12.206	3.228	4.213	1.089	6.761	14.293	7.431	29.930	12.290	45.833	13.367
	SPAF	0.726	0.643	1.119	0.817	0.974	1.120	0.963	1.200	0.996	1.303	0.780	1.751
	IPDGI	1.128	0.626	1.014	0.716	1.012	0.476	0.803	0.616	0.512	0.611	1.256	1.893
	ATR	0.796	0.939	1.332	1.041	0.954	0.757	0.769	0.987	1.115	1.060	0.854	0.992
	TextSimu	1.373	1.031	0.849	0.477	1.167	1.225	1.146	1.013	0.883	0.831	0.870	0.626
	TMM-AIP	25.379	13.171	3.209	3.783	1.053	10.476	14.512	7.916	29.022	12.059	36.097	13.650
	SGA-AIP	24.971	12.936	2.574	3.329	1.007	10.349	14.792	8.021	27.643	11.682	32.817	13.493
	VQA-AIP	28.467	13.570	3.027	4.320	1.182	**13.177**	17.103	8.369	31.749	12.743	39.154	13.302
	Ours	**68.724**	**18.467**	**5.530**	**6.551**	**1.288**	8.455	**47.956**	**12.760**	**155.169**	**32.209**	**382.765**	**51.003**

and MMGCL (marked with † in Tab. 2). This indicates that semantically aligned perturbations transfer across diverse fusion mechanisms and modeling paradigms. Graph-based models such as MMGCN and GRCN often exhibit large relative gains under CREAM, suggesting that message passing and correlation-aware aggregation may amplify coherent visual-textual shifts. The robustness of contrastive or disentangled models is dataset-dependent: MMGCL and DMRL are relatively robust on Baby and Beauty, but they can still be highly vulnerable on specific datasets (e.g., Pet–MMGCL and Office–DMRL). CREAM still surpasses baselines with only one exception (Pet–DMRL at $K = 50$), demonstrating its broad transferability under heterogeneous MRS architectures.

Dataset Trends. CREAM shows the most pronounced gains on *Pet* and *Baby*. For *Pet*, the smaller catalog reduces competition, making it easier for aligned perturbations to push target items into top-K rankings. *Baby* also yields large gains because its multimodal content is semantically rich and diverse, enabling stronger cross-modal representation shifts. CREAM achieves more modest improvements on *Beauty* and intermediate gains on *Office*. *Beauty* contains many items with relatively homogeneous content, which increases competition and reduces the influence of any single perturbation. *Office* lies in between, with a moderate catalog and clearer category structure. These results imply that compact and domain-specific recommenders (common in super-app ecosystems) can be especially vulnerable to adversarial promotion.

Takeaway. CREAM outperforms both unimodal and multimodal baselines across most datasets and model architectures, showing strong generalization to unseen models.

5.3 RQ2: Stealthiness

This section evaluates the stealthiness of generated samples. For visual and textual imperceptibility, we compare our

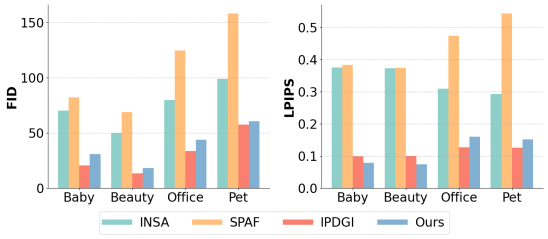


Figure 6: Visual imperceptibility of image-involved methods: FID, LPIPS (↓ better). Our method achieves favorable visual imperceptibility over most image-based attacks.

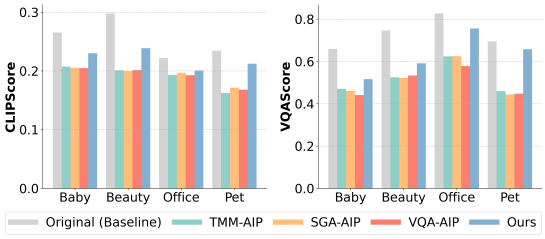


Figure 8: Cross-modal consistency of multimodal methods and original contents: CLIPScore, VQAScore (↑ better). Our method achieves stronger cross-modal semantic alignment.

method with unimodal methods. For cross-modal imperceptibility, we compare our method with multimodal ones.

Visual Imperceptibility. We evaluate the visual imperceptibility of images generated by CREAM compared to image-involved methods using FID and LPIPS metrics. Lower values indicate better visual imperceptibility. The FID and LPIPS results across different datasets are shown in Fig. 6. Generally, CREAM achieves a favorable trade-off between visual imperceptibility and adversarial effectiveness. On average, it obtains an FID of 38.255 and an LPIPS of 0.1158, significantly outperforming INSA (FID 74.6540, LPIPS 0.3375) and SPAF (FID 108.412, LPIPS 0.4434). While IPDGI slightly surpasses CREAM in terms of visual imperceptibility (FID 31.204, LPIPS 0.1127), its limited perturbations result in much weaker adversarial effects (Tab. 2).

Textual Imperceptibility. We evaluate textual imperceptibility of attack methods using Perplexity and Semantic Similarity. Perplexity measures text fluency, where lower scores indicate more natural output text [39,42]. Semantic Similarity captures meaning preservation, with higher scores reflecting better alignment with the original text. As shown in Fig. 7, CREAM achieves the lowest perplexity (15.9503), indicating the most fluent output. It also obtains the highest average semantic similarity (0.8221), outperforming TextSimu (0.7425) and ATR (0.5316). In contrast, TextSimu has higher perplexity (33.5518), while ATR shows similar perplexity (33.8898) but exhibits the lowest semantic preservation.

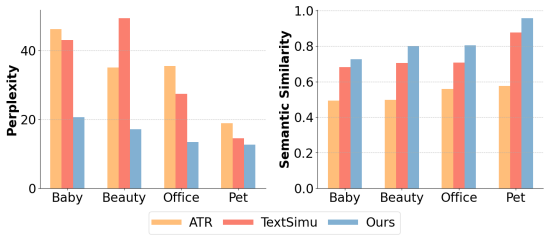


Figure 7: Textual imperceptibility of text-involved methods: Perplexity (↓ better), Semantic Similarity (↑). Our method produces more fluent and semantically faithful text.

Cross-modal Consistency. Cross-modal consistency is critical for effective multimodal attacks, especially in e-commerce scenarios where humans are highly sensitive to semantic mismatches between images and textual descriptions (as illustrated in Fig. 3). To quantitatively assess this aspect, we employ CLIPScore and VQAScore [29]. Higher scores indicate stronger semantic consistency across modalities. We evaluate all multimodal methods, using the original content as the baseline reference. Results are summarized in Fig. 8.

Our method maintains high cross-modal consistency. It achieves an average CLIPScore of 0.2203 and VQAScore of 0.6304, closely approaching the original baseline (0.2550 and 0.7317). In contrast, the baselines TMM-AIP, SGA-AIP, and VQA-AIP show significantly lower consistency.

Human Evaluation. Beyond automatic stealthiness metrics, we conduct a human study with end users and platform inspectors. CREAM is rated more realistic than the strongest baseline (4.13 vs. 2.56) and is flagged far less often (14% vs. 93%). Details are in Appendix A.

Takeaway. CREAM achieves a favorable trade-off between stealthiness and adversarial effectiveness.

5.4 RQ3: Effectiveness in Real-world

To assess the practical viability of CREAM, we conduct sandboxed testing on the commercial e-commerce platform of *AliPay*. We use production-grade snapshots of the recommender system deployed in the platform's *Office* mini-program mall. These snapshots span a 1.5-month period starting from Nov 15, 2024, and reflect actual serving configurations in the live environment. All adversarial items are generated solely from logs collected before Nov 15, thereby simulating a realistic black-box attack without access to future model states. Since model states and exposure distributions evolve over time in production, we report the absolute exposure gain for the temporal analysis.

CREAM yields consistent but non-monotonic exposure gains across 11 production snapshots, reflecting real-world drift and periodic model updates. As shown in Fig. 9, ΔER@K rises initially (Nov 15–19), dips slightly, and then rebounds in late November and early December, where ΔER@50 peaks at

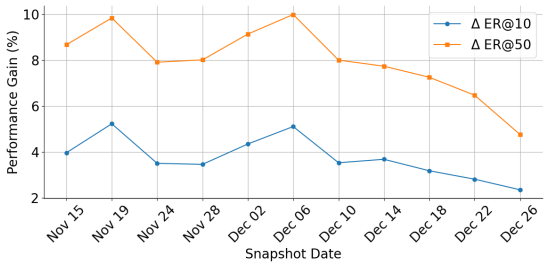


Figure 9: Real-world attack performance of CREAM across 11 production snapshots (Nov 15–Dec 26, 2024). CREAM remains consistently effective under real-world dynamics, demonstrating strong robustness and practical threats.

9.984%. This rebound may be driven by transient preference drift toward the semantic region of the adversarial targets. One plausible factor is seasonal promotion (e.g., pre-12.12 activities) that temporarily shifts exposure toward related item clusters, after which performance gradually declines as engagement patterns and models evolve.

Takeaway. Despite fluctuations, CREAM remains consistently effective throughout the 1.5-month window, demonstrating robustness under real-world dynamics and underscoring its practical threat potential. A practical adversary could further strengthen long-term impact by periodically updating adversarial content as systems evolve.

5.5 RQ4: Countermeasures

In this section, we explore possible defenses against CREAM, including training-based defenses and input-space defenses.

Training-based Defenses. Prior research has explored improving MRS robustness against noise and trend drift in modality information (e.g., merchants introducing new elements aligned with current trends) by training-based strategies [48, 78]. We investigate two state-of-the-art training-based defense strategies: Mirror Gradient (MG) [78] and Modality Balancing (MB) [48]. MG integrates mirror gradient optimization with adversarial training, while MB introduces controlled noise into visual modality and adjusts the contribution of textual modality.

As shown in Tab. 3, training-based defenses provide limited and inconsistent mitigation against CREAM. In the undefended setting, CREAM achieves 54.586 RER@10 and 13.826 RER@50 on average across all datasets and models. MG fails to consistently reduce the attack effect: it increases RER@10 to 58.373 and only marginally changes RER@50 to 13.814. MB performs slightly better, reducing RER@10 and RER@50 to 52.273 and 12.108, respectively. However, both defenses still leave adversarial items with more than 12× higher exposure at top-50 than benign items, indicating that training-based robustness remains largely ineffective against semantically aligned promotion attacks.

Table 3: Attack and defense performance averaged across all datasets and models, evaluated by RER@K (lower is better) and utility, measured as the relative change in Recall@50. CREAM denotes the undefended attack baseline. Overall, some defenses improve robustness but incur substantial utility loss, highlighting a clear trade-off.

Category	Method	RER@10	RER@50	Utility
Undefended	CREAM	54.586	13.826	+0.00%
Training	MG	58.373	13.814	-3.69%
	MB	52.273	12.108	+1.43%
Input	Bit-Red	9.230	2.846	-29.72%
	JPEG	9.239	2.832	-29.71%
	NRP	2.696	1.710	-28.65%
	NRP+L	2.910	1.731	-30.68%

The limited effectiveness stems from their incompatibility with semantically aligned attacks: MG tends to overfit specific perturbation patterns seen during adversarial training and fails to generalize to stronger, coherence-based attacks, whereas MB's reliance on cross-modal compensation directly conflicts with our attack design, which enhances alignment across modalities.

Input-space Defenses. We investigate general-purpose input-space defenses following [55]. For visual modality, we consider JPEG compression [15], Bit-Red [63], and NRP [40]; for textual modality, we apply LanguageTool (LT) [52], a grammar correction tool widely used in industrial systems. Among visual defenses, NRP achieves the strongest mitigation, outperforming Bit-Red and JPEG. We further combine NRP and LT (NRP+L) for adaptive multimodal defense evaluation.

As shown in Tab. 3, input-space defenses substantially reduce the attack exposure. Bit-Red reduces RER@50 from 13.826 to 2.846, and JPEG achieves a similar reduction to 2.832. NRP provides the strongest mitigation, reducing RER@50 to 1.710, while the adaptive multimodal combination NRP+L yields 1.731. Although these transformations suppress adversarial effects, they introduce substantial utility degradation: Bit-Red, JPEG, NRP, and NRP+L reduce Recall@50 by 29.72%, 29.71%, 28.65%, and 30.68%, respectively. This trade-off indicates that while low-level transformations can reduce the attack impact, they simultaneously distort the original multimodal representations and impair overall recommendation quality.

Takeaway. Overall, neither training-based nor input-space defenses provide reliable robustness against CREAM. While training-based methods fail to counter semantically aligned perturbations, input-space transformations partially mitigate attacks at the cost of severe performance degradation.

5.6 RQ5: Ablation Study

In this section, we investigate the contribution of each component to CREAM's performance.

Table 4: Average RER@K across datasets and model architectures for different module combinations. Best results are highlighted in **bold**. Joint use of all modules yields the best overall performance.

Enabled modules	RER@10	RER@50
Visual Perturbator	33.650	10.084
Text Generator	1.385	1.417
Visual Perturbator + Text Generator	34.642	10.387
All	**54.586**	**13.826**

Modules. We evaluate the contributions of the Visual Perturbator, Text Generator, and Joint Optimization Controller. Tab. 4 reports the performance. The Visual Perturbator alone achieves strong performance, with average 33.650 RER@10 and 10.084 RER@50. This effectiveness can be attributed to its alignment-aware design. Perturbing only the textual modality through the Text Generator leads to limited effectiveness. This is likely due to the discrete and sparse nature of textual representations, which make optimization less stable. Without a perturbed visual counterpart, the text generator lacks meaningful semantic guidance, making it difficult to produce impactful adversarial modifications.

Combining the Visual Perturbator and Text Generator without the Joint Optimization Controller brings only marginal gains over using the Visual Perturbator alone. This suggests that without coordination by the controller, multimodal perturbations may fail to move toward a jointly effective adversarial direction. When all three modules are combined, performance improves to 54.586 RER@10 and 13.826 RER@50. These results demonstrate that both modality-specific perturbations and a coordination mechanism are necessary for effective multimodal attacks.

Surrogate Model Architecture. We adopt an ensemble surrogate model to address the architectural and fusion diversity in MRSs under black-box scenarios. As shown in Tab. 5, we further evaluate its effectiveness against single surrogate on the *Baby* dataset. While some surrogate and victim pairs, such as VBPR with VBPR, perform well due to architectural similarity, this advantage does not hold across mismatched scenarios. For instance, MMGCN shows reasonable transferability against GRCN due to shared graph-based aggregation mechanisms, but such alignment is inconsistent across models. In contrast, our ensemble surrogate consistently achieves strong RER@K performance across all victim models. By aggregating varied representational perspectives, the ensemble more effectively exploits common vulnerabilities in the input space. This underscores the contribution of ensemble surrogate model for more transferable attacks.

6 Discussion and Related Work

Cost Analysis. CREAM performs offline per-item optimization before uploading. Training the three-model ensemble takes 19.7 minutes on an NVIDIA A6000 GPU as a one-time offline cost. After training, attacking one item takes about 9–10 minutes and is trivially parallelizable across target items. Even if the one-time training cost is included for a single target item, the total initial cost is roughly thirty minutes. Since a single optimized edit remains effective for about 1.5 months (RQ3), this cost amortizes to less than one minute per day, making periodic execution practical and cost-efficient.

Table 5: Average RER@K across different surrogate models (columns) and victim models (rows). Best results are highlighted in **bold**; second-best results are underlined. These results indicate that the ensemble surrogate provides competitive and robust transferability under architecture uncertainty.

Victim Model	K	Surrogate Model			
		VBPR	MMGCN	DMRL	Ensemble
VBPR	10	**9.522**	2.545	1.352	9.457
	50	**3.907**	1.647	1.110	3.854
MMGCN	10	2.869	**109.252**	2.522	106.487
	50	2.032	**24.303**	1.829	23.711
DMRL	10	1.335	1.098	**1.371**	1.369
	50	1.231	1.061	**1.328**	1.314
AMR	10	**7.920**	2.195	1.001	7.863
	50	4.022	1.807	1.063	**4.130**
GRCN	10	6.958	**325.758**	1.715	270.196
	50	4.024	**76.779**	1.505	65.049
MMGCL	10	0.950	0.849	**1.353**	1.309
	50	0.959	0.889	**1.357**	1.303

Generality. We demonstrate CREAM 's effectiveness across six diverse MRSs, and we expect it to extend to sequential [43] and LLM-based MRSs [53, 57] because popular items often cluster in nearby regions of a shared multimodal space. CREAM can also be adapted to additional modalities (e.g., video/audio) by swapping in appropriate foundation models.

Other Countermeasures. Our results suggest existing potential defenses remain inadequate against CREAM. In practice, e-commerce platforms typically employ layered listing moderation (automated screening at upload/update with escalation to professional reviewers); in our review by platform content inspectors (Appendix A), few CREAM edits are flagged, suggesting that mismatch-oriented screening alone can be bypassed by CREAM. This motivates future *adaptive* defenses that exploit platform-side signals (e.g., edit histories) and gate updates based on abnormal anchored content drift.

AIP Attack. Existing attacks can be categorized into image-based and text-based methods. Image-based attacks generate perturbation noises by gradient information [10], recommender system-specific objectives [37], clustering guidance [7], and different image styles [65]. Text-based attacks modify the item descriptions by rewriting [42] and popular-item prompting [58]. Other works [41, 75] target LLM-based RSs by perturbing both prompts and item texts.

Attacks on Retrieval-based Systems. Retrieval poisoning in dense retrieval and RAG typically assumes corpus in-

sertion and query-triggered retrieval to steer downstream outputs [5, 36, 79, 81], whereas AIP is limited to anchored, entity-preserving edits of an existing listing and optimizes population-level exposure under plausibility and cross-modal consistency constraints. Under our threat model, the transferable component reduces to embedding-space similarity boosting via content edits, already covered by our baselines.

7 Conclusion

In this work, we present CREAM, a novel framework for AIP attacks in MRSs. CREAM jointly perturbs textual and visual content in a semantically aligned manner to craft realistic yet effective adversarial examples. By leveraging a visual perturbator, a tailored text generator, and a joint optimization controller, our method captures cross-modal correlations and achieves highly effective black-box attacks. Extensive experiments across four datasets and six victim models show that CREAM significantly outperforms existing baselines, achieving on average 5.75× and 2.89× higher exposure at top-10 and top-50 metrics. Our findings not only expose critical vulnerabilities in MRSs but also highlight the limitations of current defense strategies, emphasizing the urgent need for more robust and adaptive protections for real-world MRSs.

Acknowledgments

We would like to thank the reviewers for their valuable feedback of this paper. Ziqi Zhang, Yifeng Cai, and Yao Guo are the corresponding authors. This work was partly supported by the National Key Research and Development Program of China (2024YFB4506300), the Beijing Natural Science Foundation (L243010), the National Natural Science Foundation of China (U25A6024, 62502011), and the Ant Group Research Intern Program.

Ethical Considerations

In developing CREAM, we were guided by the Menlo Report. This work obtained AliPay IRB approval, which follows established human-subjects and privacy-management standards and covered anonymized internal data use, sandboxed evaluation with historical production snapshots, and human-subject evaluation. We study AIP risks in MRSs, where a malicious seller may modify a listing's image and text under black-box access to obtain disproportionate exposure. Our goal is defensive: to measure this emerging risk, inform recommender-system designers and platform moderators, and motivate stronger protections for multimodal recommendation pipelines. We recognize that the studied techniques are dual use and could be misused if handled irresponsibly.

Stakeholders and Possible Impacts. (1) *Marketplace consumers* may benefit from reduced exposure to manipulated listings, while misuse could promote products based on adversarial manipulation rather than relevance or quality, impairing purchasing decisions and trust. (2) *Honest sellers* may benefit from defenses that preserve fair exposure and marketplace integrity, while misuse could give malicious sellers disproportionate visibility, reduce exposure for legitimate products, and create unfair competition. (3) *Platform operators and moderators* can use our work to measure cross-modal promotion risks and evaluate defenses, while misuse could increase fraud pressure, moderation costs, and degradation of recommendation integrity. Because CREAM preserves image-text coherence, it motivates defenses beyond mismatch screening, such as edit-history monitoring, abnormal content-drift detection, rate limits, and human review. (4) *Data subjects* face risks of privacy leakage, confidentiality violations, and re-identification from historical AliPay user-item interaction logs. We mitigate these risks through IRB-approved and platform-authorized handling: PII was removed, access was restricted, no re-identification was attempted, and proprietary logs, user records, and platform-specific data are not released. (5) *Human-study participants*, including end users and professional content inspectors, assessed anonymized product listings. Risks include exposure to manipulated content, mild deception about listing sources, and possible workplace pressure for inspectors. Under the AliPay IRB protocol, stimuli were anonymized, method identities were hidden, no real purchases or live-platform interactions occurred, responses were stored without direct identifiers, inspectors' individual responses were used only for research rather than personnel or performance assessment. (6) *Security researchers and defenders* may use CREAM as a controlled framework for studying adversarial risks in MRSs, red-teaming, security auditing, robustness testing, and defense evaluation; we frame it strictly as authorized defensive research. (7) *Broader society* could be harmed by deceptive product presentation, unfair competition, fraud, and erosion of trust in recommender-mediated marketplaces, while controlled study of these risks can support safer deployment and more robust MRSs.

Research Process. The main process-level risks involved platform data, production-grade evaluation, and human-subject evaluation. To reduce these risks, we used public benchmarks where possible and anonymized internal data only under AliPay IRB approval and platform authorization; PII was removed before analysis, access was restricted to authorized researchers, no attempt was made to identify individual users, sellers, or inspectors, and proprietary logs, user records, and platform-specific data are not released. We did not attack third-party services or live production systems. Production-grade evaluation used only an isolated sandbox with historical model snapshots and historical logs, without deploying adversarial listings, exposing real users, creating transactions, affecting production rankings, or imposing production load.

Responsible Disclosure and Dual-Use Concerns. AIP in MRSs presents clear dual-use concerns: similar strategies

could manipulate rankings, mislead consumers, harm honest sellers, and increase moderation burden. All platform-related experiments were authorized, reviewed, and sandboxed, and findings were shared with relevant platform stakeholders to support defensive improvements. To balance transparency and risk reduction, our release is limited to reproducibility and authorized defensive evaluation on public datasets and local testbeds. It excludes proprietary datasets, production snapshots, user logs, credentials, platform-specific endpoints, optimized adversarial listings from proprietary data, and instructions for attacking live platforms. We will monitor community feedback and adjust dissemination practices if needed to reduce misuse risk.

Mitigations. We mitigate potential harms through controlled evaluation, IRB-approved data handling, restricted release, and defensive framing. Our results show that mismatch-oriented defenses are insufficient against cross-modally coherent adversarial edits. We therefore advocate defense-in-depth: multimodal semantic consistency checks, edit-history monitoring, abnormal anchored content-drift detection, rate limits for suspicious updates, and human-review escalation for high-risk edits, while preserving legitimate seller updates.

Decision to Conduct and Publish the Research. We conducted this research because MRSs are widely deployed in e-commerce, sellers routinely edit listing images and descriptions, and this normal workflow creates a realistic attack surface. Existing AIP work focuses mainly on unimodal recommenders, leaving practitioners without principled ways to assess worst-case risks in MRSs under realistic black-box constraints. We publish because the risk arises from common design choices and widely available multimodal tools, similar techniques are plausibly independently discoverable, and responsible disclosure can help practitioners evaluate mitigations before abuse becomes widespread. We acknowledge that dual-use risk cannot be eliminated and aim to maximize defensive value while minimizing operational misuse.

Open Science

We follow the open-science principles encouraged by the community and release our code in a public repository [66]. The repository contains the core implementation and scripts needed to reproduce our public benchmark workflow. The `README.md` in the repository provides step-by-step instructions for environment setup and running the workflow.

References

[1] Alibaba Group. Taobao homepage, 2025. Accessed April 11, 2025. URL: https://www.taobao.com.

[2] Amazon. Amazon homepage, 2025. Accessed April 11, 2025. URL: https://www.amazon.com.

[3] Yifeng Cai, Ziqi Zhang, Jiaping Gui, Bingyan Liu, Xiaoke Zhao, Ruoyu Li, Zhe Li, and Ding Li. Famos: Robust privacy-preserving authentication on payment apps via federated multi-modal contrastive learning. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 289–306, 2024.

[4] Yifeng Cai, Ziqi Zhang, Mengyu Yao, Junlin Liu, Xiaoke Zhao, Xinyi Fu, Ruoyu Li, Zhe Liu, Xiangqun Chen, Yao Guo, et al. I can tell your secrets: Inferring privacy attributes from mini-app interaction history in super-apps. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 6541–6560, 2025.

[5] Harsh Chaudhari, Giorgio Severi, John Abascal, Matthew Jagielski, Christopher A Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. Phantom: General trigger attacks on retrieval augmented language generation. *arXiv preprint arXiv:2405.20485*, 2024.

[6] Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion models for imperceptible and transferable adversarial attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[7] 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 Transactions on Information Systems*, 42(6):1–26, 2024.

[8] Pei-Yu Chen, Shin-yi Wu, and Jungsun Yoon. The impact of online recommendations and consumer feedback on sales. 2004.

[9] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*, pages 87–94, 2008.

[10] 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. IEEE, 2020.

[11] Zhihao Dou, Xin Hu, Haibo Yang, Zhuqing Liu, and Minghong Fang. Adversarial attacks to multi-modal models. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, pages 35–46, 2023.

[12] eBay. ebay homepage, 2025. Accessed April 11, 2025. URL: https://www.ebay.com.

[13] Therese Fessenden. Scrolling and attention. *Nielsen Norman Group*, April 2018. Accessed: 2025-11-04. URL: https://www.nngroup.com/articles/scrolling-and-attention/.

[14] Jiwei Guan, Tianyu Ding, Longbing Cao, Lei Pan, Chen Wang, and Xi Zheng. Probing the robustness of vision-language pretrained models: A multimodal adversarial attack approach. *arXiv preprint arXiv:2408.13461*, 2024.

[15] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

[16] Dongchen Han, Xiaojun Jia, Yang Bai, Jindong Gu, Yang Liu, and Xiaochun Cao. Ot-attack: Enhancing adversarial transferability of vision-language models via optimal transport optimization. *arXiv preprint arXiv:2312.04403*, 2023.

[17] Bangyan He, Xiaojun Jia, Siyuan Liang, Tianrui Lou, Yang Liu, and Xiaochun Cao. Sa-attack: Improving adversarial transferability of vision-language pretraining models via self-augmentation. *arXiv preprint arXiv:2312.04913*, 2023.

[18] 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.

[19] 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.

[20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[22] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):684–696, 2019.

[23] Hai Huang, Jiaming Mu, Neil Zhenqiang Gong, Qi Li, Bin Liu, and Mingwei Xu. Data poisoning attacks to deep learning based recommender systems. *arXiv preprint arXiv:2101.02644*, 2021.

[24] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[25] Xiaojun Jia, Sensen Gao, Qing Guo, Ke Ma, Yihao Huang, Simeng Qin, Yang Liu, and Xiaochun Cao. Semantic-aligned adversarial evolution triangle for high-transferability vision-language attack. *arXiv preprint arXiv:2411.02669*, 2024.

[26] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2):7–es, 2007.

[27] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 781–789, 2017.

[28] Kiljae Lee and Jungsil Choi. Image-text inconsistency effect on product evaluation in online retailing. *Journal of Retailing and Consumer Services*, 49:279–288, 2019.

[29] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

[30] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024.

[31] Fan Liu, Huilin Chen, Zhiyong Cheng, Anan Liu, Liqiang Nie, and Mohan Kankanhalli. Disentangled multimodal representation learning for recommendation. *IEEE Transactions on Multimedia*, 25:7149–7159, 2022.

[32] Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. Multimodal recommender systems: A survey. *ACM Computing Surveys*, 57(2):1–17, 2024.

[33] Xiaohao Liu, Zhulin Tao, Ting Jiang, He Chang, Yunshan Ma, Yinwei Wei, and Xiang Wang. Toda: Target-oriented diffusion attacker against recommendation system. *arXiv preprint arXiv:2401.12578*, 2024.

[34] Yang Liu, Zhuozhuo Zhao, Ting Jiang, Yu Wang, Shengnan Wu, and Wenming Zhe. A model fusion approach for goods information inspection in dual-platform e-commerce systems. In *2021 26th IEEE International*

Conference on Emerging Technologies and Factory Automation (ETFA), pages 01–08. IEEE, 2021.

[35] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[36] Yinuo Liu, Zenghui Yuan, Guiyao Tie, Jiawen Shi, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. Poisoned-mrag: Knowledge poisoning attacks to multimodal retrieval augmented generation. *arXiv preprint arXiv:2503.06254*, 2025.

[37] 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*, pages 3590–3602, 2021.

[38] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 102–111, 2023.

[39] Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. What makes my model perplexed? a linguistic investigation on neural language models perplexity. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 40–47, 2021.

[40] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 262–271, 2020.

[41] Liang-bo Ning, Shijie Wang, Wenqi Fan, Qing Li, Xin Xu, Hao Chen, and Feiran Huang. Cheatagent: Attacking llm-empowered recommender systems via llm agent. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2284–2295, 2024.

[42] Sejoon Oh, Gaurav Verma, and Srijan Kumar. Adversarial text rewriting for text-aware recommender systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1804–1814, 2024.

[43] Xingyu Pan, Yushuo Chen, Changxin Tian, Zihan Lin, Jinpeng Wang, He Hu, and Wayne Xin Zhao. Multimodal meta-learning for cold-start sequential recommendation. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 3421–3430, 2022.

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[45] Shaina Raza, Mizanur Rahman, Safiullah Kamawal, Armin Toroghi, Ananya Raval, Farshad Navah, and Amirmohammad Kazemeini. A comprehensive review of recommender systems: Transitioning from theory to practice. *arXiv preprint arXiv:2407.13699*, 2024.

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[48] Yu Shang, Chen Gao, Jiansheng Chen, Depeng Jin, Huimin Ma, and Yong Li. Enhancing adversarial robustness of multi-modal recommendation via modality balancing. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6274–6282, 2023.

[49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[50] Bowen Tang, Zheng Wang, Yi Bin, Qi Dou, Yang Yang, and Heng Tao Shen. Ensemble diversity facilitates adversarial transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24377–24386, 2024.

[51] 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, 2019.

[52] the LanguageTool community and Daniel Naber. Languagetool. https://github.com/languagetool-org/languagetool, 2025.

[53] Jiahao Tian, Zhenkai Wang, Jinman Zhao, and Zhicheng Ding. Mmrec: Llm based multi-modal recommender system. In *2024 19th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP)*, pages 105–110. IEEE, 2024.

[54] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[55] Haodi Wang, Kai Dong, Zhilei Zhu, Haotong Qin, Aishan Liu, Xiaolin Fang, Jiakai Wang, and Xianglong Liu. Transferable multimodal attack on vision-language pre-training models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 1722–1740. IEEE, 2024.

[56] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

[57] Qi Wang, Jindong Li, Shiqi Wang, Qianli Xing, Runliang Niu, He Kong, Rui Li, Guodong Long, Yi Chang, and Chengqi Zhang. Towards next-generation llm-based recommender systems: A survey and beyond. *arXiv preprint arXiv:2410.19744*, 2024.

[58] 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. *arXiv preprint arXiv:2409.11690*, 2024.

[59] 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*, pages 3541–3549, 2020.

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

[61] Bo Xiao and Izak Benbasat. Product-related deception in e-commerce: A theoretical perspective. *Mis Quarterly*, pages 169–195, 2011.

[62] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

[63] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

[64] Wenzhuo Xu, Kai Chen, Ziyi Gao, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Highly transferable diffusion-based unrestricted adversarial attack on pre-trained vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 748–757, 2024.

[65] 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*, pages 2900–2909, 2024.

[66] Yao. Cream, 2026. doi:10.5281/zenodo.20338101.

[67] Jingwen Ye, Ruonan Yu, Songhua Liu, and Xinchao Wang. Mutual-modality adversarial attack with semantic perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6657–6665, 2024.

[68] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig Macdonald. 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*, pages 1807–1811, 2022.

[69] Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36:52936–52956, 2023.

[70] Ziyi Yin, Muchao Ye, Tianrong Zhang, Jiaqi Wang, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vqattack: transferable adversarial attacks on visual question answering via pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6755–6763, 2024.

[71] Zhen Yu, Zhou Qin, Zhenhua Chen, Meihui Lian, Haojun Fu, Weigao Wen, Hui Xue, and Kun He. Sparse black-box multimodal attack for vision-language adversary generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5775–5784, 2023.

[72] Zhe Yuan, AJ Yuan Chen, Yitong Wang, and Tianshu Sun. How recommendation affects customer search: A field experiment. *Information Systems Research*, 2024.

[73] Meifang Zeng, Ke Li, Bingchuan Jiang, Liujuan Cao, and Hui Li. Practical cross-system shilling attacks with limited access to data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4864–4874, 2023.

[74] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022.

[75] Jinghao Zhang, Yuting Liu, Qiang Liu, Shu Wu, Guibing Guo, and Liang Wang. Stealthy attack on large language model based recommendation. *arXiv preprint arXiv:2402.14836*, 2024.

[76] 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.

[77] Yihe Zhang, Xu Yuan, Jin Li, Jiadong Lou, Li Chen, and Nian-Feng Tzeng. Reverse attack: Black-box attacks on collaborative recommendation. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 51–68, 2021.

[78] Shanshan Zhong, Zhongzhan Huang, Daifeng Li, Wushao Wen, Jinghui Qin, and Liang Lin. Mirror gradient: Towards robust multimodal recommender systems via exploring flat local minima. In *Proceedings of the ACM Web Conference 2024*, pages 3700–3711, 2024.

[79] Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. *arXiv preprint arXiv:2310.19156*, 2023.

[80] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions. *arXiv preprint arXiv:2302.04473*, 2023.

[81] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. {PoisonedRAG}: Knowledge corruption attacks to {Retrieval-Augmented} generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3827–3844, 2025.

Appendix

A Human Evaluation Details

Participants and Stimuli. We recruit 3 professional content inspectors from the e-commerce platform and 50 end users who made more than two online purchases in the past three months. One end-user participant is excluded according to our quality-control criteria. We sample 100 items. For each item, we prepare three anonymized listings: *Original*, *Baseline* (strongest AIP baseline by exposure gain), and CREAM. Each listing contains an image, title, and description.
Protocol. To avoid contrast and memory effects, each evaluator sees at most one version of each item. For end users, each valid participant evaluates 30 real listings, with 10 listings assigned to each of the three conditions: *Original*, *Baseline*, and CREAM. Each participant additionally receives two attention-check listings, which are used only for quality control and excluded from the analysis. This yields 1,470 valid user evaluations in total, with 490 evaluations per condition. For inspectors, we use a Latin-square assignment over the three conditions. For each item, its three versions are assigned to the three inspectors such that each inspector sees only one version of that item. This yields 300 inspector judgments in total, with 100 judgments per condition. All listings are presented in randomized order, and method identities are hidden.
Tasks. End users rate *realism* on 5-point Likert scales, and provide a binary *purchase intent*. Inspectors provide a binary *suspiciousness* decision (flag / not flag) with optional reasons.
Quality Control. We exclude responses that fail any attention-check question or contain incomplete answers.
Results. User-rated realism (mean$\pm$SE): Original 4.65$\pm$0.03, CREAM 4.13$\pm$0.05, Baseline 2.56$\pm$0.06. Purchase intent: Original 404/490 (0.824, Wilson 95% CI [0.788, 0.856]), CREAM 319/490 (0.651, [0.608, 0.692]), Baseline 35/490 (0.071, [0.052, 0.098]). Inspector flag rate: Original 1/100 (0.010, [0.002, 0.054]), CREAM 14/100 (0.140, [0.085, 0.221]), Baseline 93/100 (0.930, [0.863, 0.966]). CREAM flags are mainly attributed to minor logo or textual alterations.

B Case Study: Examples

We present adversarial examples for visual reference. The examples in Fig. 10 are drawn from the *Baby* and *Beauty* datasets. Due to data sharing restrictions, examples from proprietary datasets are not included. Each sample consists of the original item content alongside adversarial variants generated by three image-only baselines, two text-only baselines, and our proposed multimodal method, CREAM. We omit the adapted multimodal baselines from visual display, as they are constructed atop unimodal methods and thus exhibit similar visual patterns. By closely examining the adversarial images, we observe that all unimodal baselines introduce noticeable noise. While such perturbations may be imperceptible at smaller scales, they become visibly unnatural upon zooming in. This could potentially alarm inspectors or customers. In contrast, our method introduces subtle and visually imperceptible changes. A potential concern is that brand logos might be unintentionally altered during this process. However, this issue can be addressed with simple techniques such as logo inpainting or cutout, which we leave as future work.

Regarding text modifications, adversarial texts generated by TextSimu often include irrelevant or redundant details. Outputs from ATR are frequently confusing, as the model tends to continue the original text rather than rewriting it, resulting in incoherent messaging. In contrast, our proposed method produces more fluent rewritings. In some instances, it introduces features not present in the original description. This may raise concerns about false advertising. Nevertheless, such manipulations remain difficult to detect in e-commerce scenarios where the actual product cannot be directly inspected.

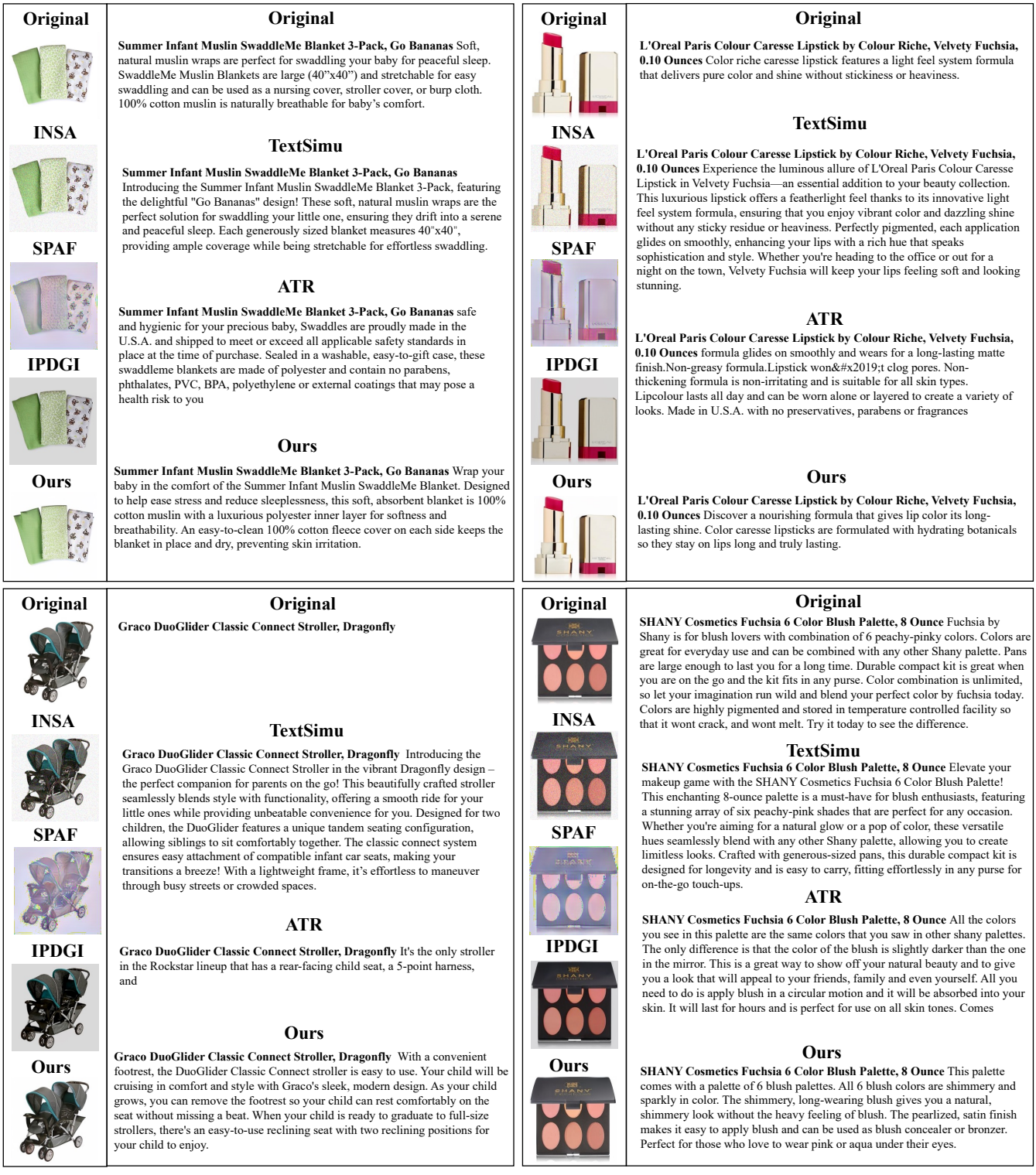


Figure 10: Adversarial Examples from *Baby* and *Beauty*