SC-DAG: Semantic-Constrained Diffusion Attacks for Stealthy Exposure Manipulation in Visually-Aware Recommender Systems

Ze Lin
Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University
Xiamen, Fujian, China
linze@stu.xmu.edu.cn

Yuqiu Qian
The University of Hong Kong
Hong Kong
qianyuqiu79@gmail.com

Xiaodong Li
Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University
Xiamen, Fujian, China
xdli@xmu.edu.cn

Ziyu Lyu
School of Cyber Science and Technology, Sun Yat-sen University
Shenzhen, Guangdong, China
lvzy7@mail.sysu.edu.cn

Hui Li*
Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University
Xiamen, Fujian, China
hui@xmu.edu.cn

Abstract

Visually-aware recommender system (VARS) has become increasingly prevalent in various online services by integrating visual features of items to enhance recommendation quality. However, VARS introduces new security vulnerabilities and malicious attackers can perform visual shilling attacks to manipulate recommendation lists via uploading generated images with visually imperceptible perturbations. While prior research has explored such threats to help service providers enhance their systems, existing visual shilling attack methods still suffer from uncontrolled pixel-space perturbation, energy dispersion dilemma and semantic misalignment in reference selection. In this work, we present Semantic-Constrained Diffusion Adversarial Generation (SC-DAG) for visual shilling attacks. SC-DAG overcomes key limitations of previous methods by focusing perturbations on semantically meaningful image regions through contour-aware segmentation, guiding adversarial generation in latent space using a conditional diffusion process, and performing a hybrid reference image selection strategy that balances popularity and semantic similarity. Extensive experiments on performing visual shilling attacks against multiple VARS models show that SC-DAG achieves state-of-the-art attack performance in elevating target items' ranking, while maintaining strong perceptual indistinguishability and minimal impact on overall recommendation performance of the system. Our work offers insights into leveraging structured semantic priors for more sophisticated adversarial manipulations against VARS and also highlights the necessity for developing more robust VARS models resilient to visual shilling attacks. We provide our implementation at https://github.com/KDEGroup/SC-DAG.

CCS Concepts

• **Security and privacy** → **Web application security**; **Systems security**; • **Information systems** → **Recommender systems**;

Keywords

Adversarial Attacks; Visually-Aware Recommender System

ACM Reference Format:
Ze Lin, Yuqiu Qian, Xiaodong Li, Ziyu Lyu, and Hui Li. 2025. SC-DAG: Semantic-Constrained Diffusion Attacks for Stealthy Exposure Manipulation in Visually-Aware Recommender Systems. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746252.3761034

1 Introduction

Recommender system (RecSys) has evolved into critical components of the digital economy, mitigating information overload by enabling personalized interactions across e-commerce platforms, social media, and content ecosystems [23, 39, 44, 45]. By modeling user behavior patterns and item metadata, RecSys brings mutual benefit via facilitating users' decision-making processes and increasing revenue for service providers. Despite that conventional collaborative filtering [39] has demonstrated its effectiveness in RecSys, it still faces significant challenges in addressing data sparsity and cold-start scenarios. These limitations are particularly acute in visually-intensive domains [18] such as fashion recommendation and interior design, where user preferences are heavily influenced by aesthetic sensibilities and contextual factors that cannot be fully captured from interaction history.

*Corresponding Author.

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761034

To address these, Visually-Aware Recommender System (VARS) [11, 18] has emerged, leveraging visual signals (e.g., color, texture, spatial composition) to guide RecSys. VARS not only enhances recommendations of long-tail items with limited interactions but also enables nuanced modeling of aesthetic preferences.

Nevertheless, VARS may introduce new, potential security vulnerabilities. RecSys typically allows user-generated content. For instance, a merchant can upload item images to better illustrate items and attract customers. Attackers can exploit this mechanism and perform *visual shilling attacks*. They can inject visually imperceptible perturbations into item images [25]. The manipulated item images are then uploaded legally, causing the system to misjudge their relevance and expose them to unintended customers, increasing the sales of the target items. Fig. 1 illustrates such an attack: a perturbed item image triggers erroneous recommendations, amplifying the item's visibility to customers.

Recently, much effort has been devoted to designing new visual shilling attack methods to reveal the vulnerability of VARS and help service providers improve their systems. Early studies [10, 28, 29] adapt existing image classification attack frameworks such as FGSM [10] and PGD [29] to attack VARS. However, these vision-centric approaches inherit the classification-oriented goal of label manipulation, whereas VARS attacks require addressing the distinct objective of distorting ranking lists in recommendation processes. Therefore, recent works opt to design ranking-specific adversarial strategies. For example, AIP [25] introduces the "hook item" framework, where attackers select highly popular items as visual reference points. By perturbing a target item's image to mimic the hook item's features, the method fools VARS into ranking target items as if they were popular ones. IPDGI [5] generates adversarial images by first creating an optimized perturbation for the target item's image to align its features with popular items. This perturbation is then combined with Gaussian noise in a diffusion model's reverse process, guided by conditional constraints such as visual consistency, to iteratively generate the final adversarial image.

Despite their success, existing methods still face three problems:

- **P1: Uncontrolled Pixel-Space Perturbation**: Prior methods like AIP and IPDGI apply adversarial noise directly in the pixel space without spatial or structural constraints. The perturbed image has semantically irrelevant regions (e.g., backgrounds) with conspicuous anomalies such as chromatic distortions and irregular patterns, severely compromising attack stealth.
- **P2: Energy Dispersion Dilemma**: Uniform perturbation across the entire image distributes attack energy indiscriminately, diluting its effectiveness. To achieve desired ranking manipulation, some methods [5, 25] resort to inflating perturbation magnitudes, which exacerbate visual aberrations and reduce the plausibility of adversarial examples.
- **P3: Semantic Misalignment in Reference Selection**: Methods like AIP prioritize popularity for selecting reference items, assuming high-utility features from popular items are universally transferable. However, they neglect semantic compatibility between reference and target items, leading to incongruous feature fusion and suboptimal adversarial guidance. For example, when the target item is a keyboard, using the image of the popular item iPhone as a feature source is unreasonable.

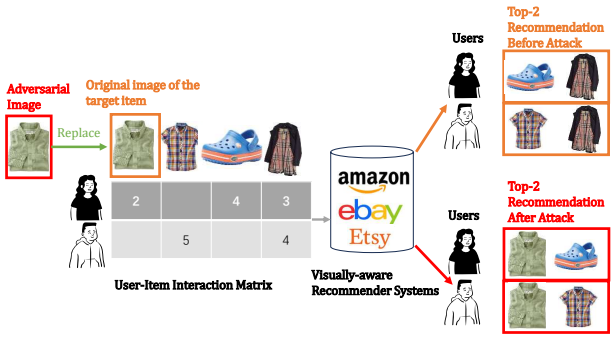

Figure 1: An example of visual shilling attacks against VARS.

(a) Original (b) AIP (c) IPDGI (d) SC-DAG

(e) Original (f) AIP (g) IPDGI (h) SC-DAG

Figure 2: Comparisons between original/generated images.

The limitations highlighted above are also visually evident in Fig. 2. Specifically, images generated by AIP (Fig. 2(b) and Fig. 2(f)) and IPDGI (Fig. 2(c) and Fig. 2(g)) often contain visually unnatural textures and distorted color patterns, particularly in background areas or along item boundaries. These artifacts not only reduce stealth but also undermine user trust in RecSys. *The above limitations reveal a fundamental gap in current research on visual shilling attacks: the overemphasis on adversarial effectiveness at the expense of attack stealth.* To address this issue, in this paper, we propose Semantics-Constrained Diffusion Adversarial Generation (SC-DAG) for Visual Shilling Attacks against RecSys. The goal of SC-DAG is to balance attack efficacy and attack stealth. It has three key contributions:

- **Semantically Constrained Perturbation Masking**: SC-DAG generates precise foreground masks to strictly confine adversarial perturbations to item-specific semantic features. This strategy eliminates unintended modifications to backgrounds or irrelevant regions, directly addressing the uncontrolled pixel-space distortions in prior works (**P1**).
- **Latent Diffusion-Based Adversarial Generation**: Unlike pixel-space methods, SC-DAG pioneers adversarial generation in the latent space through a diffusion-based framework. By encoding item semantics into conditional vectors, SC-DAG guides noise evolution along structural priors during denoising, embedding adversarial signals within intrinsic item attributes while

preserving background textures. This design decouples perturbation magnitude from visual fidelity, resolving the energy dispersion dilemma (**P2**). To our best knowledge, *SC-DAG is the first latent diffusion based visual shilling attack method.*

- **Feature-Aligned Reference Selection**: SC-DAG introduces a hybrid reference selection strategy that combines item popularity with visual feature similarity. By retrieving popular items exhibiting high semantic overlap with target items, SC-DAG enhances adversarial feature transferability while avoiding mismatches inherent in selection approaches that only consider item popularity. This method ensures the semantic alignment between target items and popular items (**P3**).

Extensive experimental results on representative benchmarks and VARS demonstrate that SC-DAG can achieve strong attack power comparable to existing visual shilling attack methods while exhibiting greater invisibility. Our work offers insights into leveraging structured semantic priors for more sophisticated adversarial manipulations against VARS and highlights the necessity for developing more robust VARS models resilient to visual shilling attacks.

2 Problem Definition

Visual shilling attacks aim to generate visually imperceptible perturbations to the target item to manipulate its recommendation ranking and promote the target item. We consider a merchant-driven attack scenario in e-commerce platforms:

- **Attacker Capability:** Malicious merchants can upload adversarially modified images but cannot manipulate user interactions or system parameters. We assume the visual encoder in VARS is preknown, following prior works on visual shilling attacks [5, 25]. The visual encoder of VARS is the primary attack target, as perturbations aim to alter latent representations of target items.
- **Defense Awareness:** Adversarial images must bypass both automated content moderation and human scrutiny, necessitating the generation of highly imperceptible adversarial images.

For a target item i with its original image $x_i \in \mathbb{R}^{H\times W\times 3}$ where H is height, W is width and 3 indicates the number of channels, the goal of visual shilling attacks is to synthesize an adversarial image $\hat{x}_{adv}$ via a perturbation operator Q. By replacing x_i, $\hat{x}_{adv}$ makes i appear more frequently in users' top-K recommendation lists.

A successful visual shilling attack approach should achieve the following three goals simultaneously:

Target Exposure Maximization: The core objective is to optimize Q to generated $\hat{x}_{adv} = Q(x_i)$, such that the average Hit Rate at K (HR@K) of i across all users $\mathcal{U}$ is maximized:

$$\max_{Q} \frac{1}{|\mathcal{U}|} \sum_{u\in\mathcal{U}} \mathbb{I}\left(i \in \text{top-}K(f_u(\hat{x}_{adv}))\right), \tag{1}$$

where $f_u(\cdot)$ is the ranking function of VARS for user u, mapping items to predicted relevance scores, and $\mathbb{I}(\cdot)$ is the indicator function (1 if true, 0 otherwise).

Visual Stealthiness: To avoid being detected, the operator Q must ensure minimal deviation between $\hat{x}_{adv}$ and x_i:

$$\min_{Q} \ \text{Diff}(\hat{x}_{adv}, x_i), \tag{2}$$

where $\text{Diff}(\cdot,\cdot)$ is a similarity-aware metric capturing perceptual and semantic differences in image space.

System Robustness Preservation: To prevent significant degradation in VARS's overall performance, the attack must minimally affect the ranking quality for *non-target items*, which refer to items that are not modified by adversarial perturbations. Let $\text{HR@}K_{\text{non-target}}(f)$ represent the average Hit Rate at K across all users and non-target items under the original VARS f. After replacing x_i with $\hat{x}_{adv}$, the deviation in $\text{HR@}K_{\text{non-target}}$ should satisfy:

$$\left|\text{HR@}K_{\text{non-target}}(f) - \text{HR@}K_{\text{non-target}}(f^{\text{adv}})\right| \le \tau, \tag{3}$$

where f^{adv} is the attacked VARS with $\hat{x}_{adv}$ replacing x_i and τ is a small tolerance threshold ensuring negligible impact on non-target item rankings. The above constraint ensures that the attack focuses on elevating the visibility of the target item i without broadly disrupting VARS's ability to recommend other items effectively.

3 Our Method SC-DAG

3.1 Overview

As depicted in Fig. 3, SC-DAG consists of three coordinated parts: (1) semantics-aware perturbation localization guided by item contours (Sec. 3.3.1), (2) conditioned diffusion generation with adversarial guidance (Sec. 3.3.2), and (3) reference-optimized attack refinement (Sec. 3.4). Its innovation lies in strategically concentrating perturbations on semantically meaningful regions through a novel contour-conditioned diffusion mechanism, which fundamentally solves the stealth-compromising limitations of existing approaches.

SC-DAG is a latent diffusion [32] based visual shilling attack model. Its workflow initiates with semantic segmentation to extract precise item contours from original item images. These structural features are encoded as conditional vectors to guide the diffusion process, ensuring that generated perturbations adhere to the item's intrinsic shape semantics. During the reverse diffusion process, SC-DAG injects adversarial gradients into latent representations to align features with those of high-exposure reference items. Meanwhile, semantic contour conditions preserve the item's structural integrity, guiding perturbations toward visually salient regions. Additionally, the diffusion trajectory is constrained by contour-aware semantic conditioning, which preserves the structural fidelity of the target item. This coordinated strategy enables SC-DAG to subtly embed recommendation-sensitive perturbations into visually coherent regions to maximize exposure gains without compromising perceptual realism. To maximize attack effectiveness, we further implement a hybrid reference selection mechanism that aligns adversarial patterns with high-exposure reference features from popular items, balanced by semantic relevance regularizers to preserve characteristics of the target item.

3.2 Architecture of SC-DAG

SC-DAG is a visual shilling attack model and based on latent diffusion [32]. The basic idea is to progressively add noise to degrade the data, and then train the model to learn how to reverse this process and generate data from noise. The training process can be divided into two stages: the forward diffusion process and the reverse diffusion process.

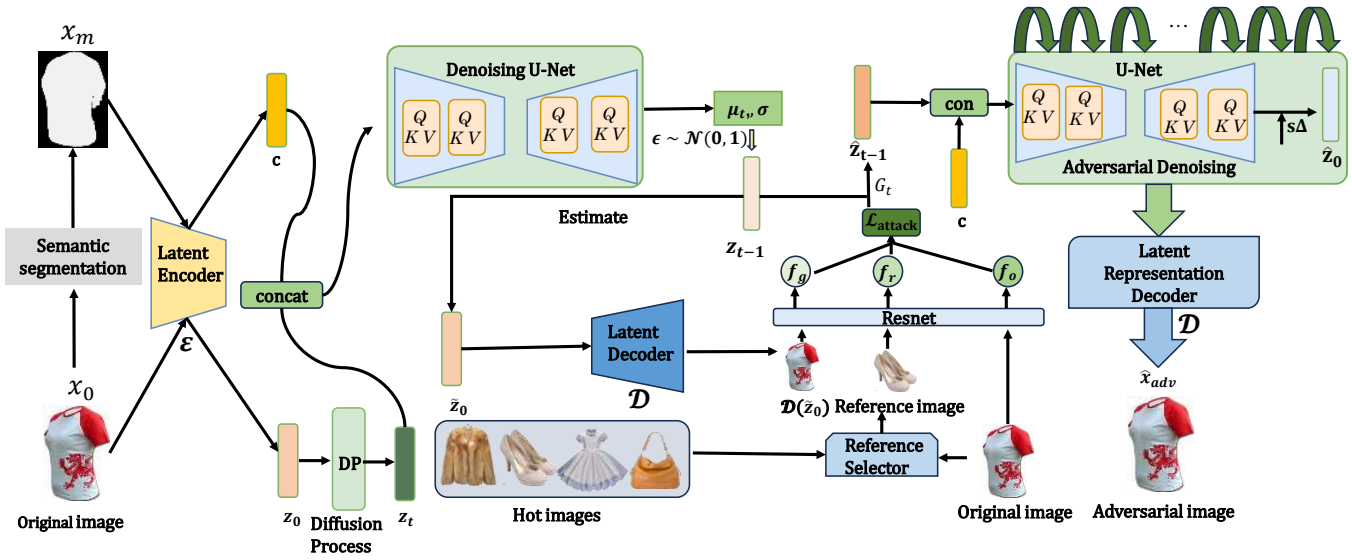

Figure 3: Overview of SC-DAG.

Suppose that the original data is x_0; in the forward process, by iteratively adding noise, a series of intermediate states $x_1, x_2, \ldots, x_T$ is generated. This step can be defined by the following formula:

$$q(\boldsymbol{x_t}|\boldsymbol{x_{t-1}}) = N(\boldsymbol{x_t}; \sqrt{1-\beta_t}\boldsymbol{x_{t-1}}; \beta_t I), \tag{4}$$

where β_t is a hyperparameter that controls the amount of noise. By repeatedly applying the above step, we can transform the original data x_0 into x_T, which is close to pure noise.

The training objective is to learn how to reverse the above process, i.e., gradually recovering the data from pure noise. The reverse diffusion process is approximated by a parameterized model $p_\theta(x_{t-1}|x_t)$, which estimates the true distribution of generating x_{t-1} from x_t. We can express it as:

$$p_\theta\left(\boldsymbol{x_{t-1}} \mid \boldsymbol{x_t}\right) := \mathcal{N}\left(\boldsymbol{x_{t-1}}; \mu_\theta\left(\boldsymbol{x_t}, t\right), \sigma_\theta\left(\boldsymbol{x_t}, t\right)\right), \tag{5}$$

where μ_θ and σ_θ indicate mean and standard deviation, respectively. The training objective becomes:

$$\mathcal{L}_{DM} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\boldsymbol{x_t}, t)\|^2\right], \tag{6}$$

where ϵ is the added noise, and $\epsilon_\theta(x_t, t)$ is the predicted noise.

Based on the above diffusion process, latent diffusion model further introduces an autoencoder $\mathcal{E}$. The original image is first compressed and encoded to generate a low-dimensional representation z_0, which is then used as the input. The generated result is then decoded by a decoder D to reconstruct the image. The data space where the compressed z_0 resides is called the latent space, and z_0 can be referred to as the latent vector.

One strength of latent diffusion model lies in its conditional mechanism, which allows the addition of extra guiding information to generate the desired images, enabling the convenient introduction of various types of conditions (text, category, layout, etc.). The corresponding objective function can be written as follows:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(\boldsymbol{z_t}, t, \tau_\theta(y)\right)\|_2^2\right], \tag{7}$$

where y represent the extra conditional data, and τ_θ is a domain-specific encoder mapping y to an intermediate representation $\tau_\theta(y)$.

3.3 Semantics-conditioned Diffusion Generation

Existing adversarial attacks on VARS typically extract global image features using pre-trained encoders. However, such methods indiscriminately encode both item foregrounds and backgrounds, leading to two critical issues: (1) Ineffective Perturbation Allocation: Background features (e.g., indoor scenes, decorative patterns) dominate the extracted representations but contribute little to affecting the recommendation. Adversarial perturbations thus waste energy modifying irrelevant regions. (2) Compromised Stealthiness: Excessive background alterations (e.g., distorted textures) may alert users, violating the imperceptibility requirement of adversarial attacks.

To handle these issues, we propose to guide perturbations to focus on item foregrounds by leveraging semantic segmentation. Inspired by the conditional generation capability of latent diffusion model, we extract contour features from item images and inject them as structural guidance during adversarial generation. This implicitly leads perturbations to pay close attention to semantically meaningful regions (e.g., clothing shapes, electronic item outlines), maximizing attack impact while preserving naturalness.

3.3.1 Semantic Segmentation and Contour Feature Encoding. To guide adversarial perturbations to focus on semantically critical regions, we first extract item contour features through semantic segmentation. Given an image of the target item $x_0 \in \mathbb{R}^{H\times W\times 3}$, we generate a binary mask $x_m \in \{0,1\}^{H\times W}$ using the pre-trained DeepLabv3 model [4] with ResNet-101 backbone. The mask labels foreground (item region) as 1 and background as 0. The masked image $x_0 \odot x_m$ (element-wise multiplication) is encoded into a latent vector $\mathbf{c} \in \mathbb{R}^{D_c}$ via a pre-trained encoder $\mathcal{E}$:

$$\mathbf{c} = \mathcal{E}(\boldsymbol{x_0} \odot \boldsymbol{x_m}). \tag{8}$$

The above masked encoding focuses on essential visual features in the item foreground (e.g., textures and contours) while suppressing irrelevant background noise.

3.3.2 Conditioned Diffusion with Contour Guidance. At each diffusion step t, the contour condition c is injected into UNet via modified cross-attention layers:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad \begin{cases} Q = W_q^{(i)} \cdot \phi_i(\boldsymbol{z_t}) \\ K = W_k^{(i)} \cdot \boldsymbol{c} \\ V = W_v^{(i)} \cdot \boldsymbol{c} \end{cases} \tag{9}$$

where $\phi_i(z_t)$ is the flattened intermediate feature of the UNet at layer i, and $W_q^{(i)}, W_k^{(i)}, W_v^{(i)}$ are learnable projections.

To ensure multi-scale structural consistency, we further inject the contour condition c into residual blocks after each downsampling layer, reinforcing foreground semantics across UNet hierarchies.

By explicitly conditioning the diffusion process on item contours, we enable precision-targeted adversarial perturbations that simultaneously achieve the following two goals:

- Amplify attack impact by concentrating modifications on semantically salient regions.
- Preserve visual stealth through contour-constrained generation that maintains background integrity.

The adversarial reverse process is formalized as:

$$\begin{aligned} p_\theta\left(\boldsymbol{z_{t-1}} \mid \hat{\boldsymbol{z}}_t, c\right) &:= \mathcal{N}\Big(\boldsymbol{z_{t-1}}; \mu_\theta(\hat{\boldsymbol{z}}_t, \boldsymbol{t}, \boldsymbol{c}), \sigma_\theta(\hat{\boldsymbol{z}}_t, \boldsymbol{t}, \boldsymbol{c})\Big) \\ \hat{\boldsymbol{z}}_{t-1} &= \boldsymbol{z_{t-1}} + \underbrace{\qquad G_t}_{\text{Attack driver}} \end{aligned} \tag{10}$$

where the attack driver G_t injects recommendation-sensitive perturbations by hijacking feature responses in the recommender's visual encoder and exploiting latent semantic correlations between item contours and user preference patterns. This gradient is applied after each denoising step, modifying z_{t-1} before the next iteration.

The semantic condition continuously guides perturbation allocation throughout denoising. Specifically, by encoding the masked foreground image into a latent contour vector c, and injecting it into both the attention modules and residual blocks of UNet, the model implicitly learns to concentrate generation dynamics on the foreground regions of the item.

Instead of explicitly restricting perturbations using binary masks in the pixel space, SC-DAG leverages the contour condition to shape the diffusion process in the latent space. This leads to a denoising trajectory where visual changes naturally emerge along semantically meaningful boundaries (e.g., item edges, textures, and shape contours) while irrelevant regions (e.g., backgrounds) remain largely untouched. As a result, perturbations appear coherent and visually plausible, aligning with human perception and maintaining attack stealth. SC-DAG enforces semantic constraints by injecting contour conditions into the denoising network, enabling structure-aware generation with precise perturbation localization. Sec. 3.4 will detail how G_t dynamically aligns perturbations with high-impact reference features to maximize target exposure.

3.4 Reference-guided Adversarial Generation

To enhance the attack impact on target item exposure, we propose a hybrid reference selection strategy that balances historical popularity and semantic relevance, coupled with a regularized adversarial gradient injection mechanism.

3.4.1 Hybrid Selection of Reference Image. For a target item, we first retrieve the top-K candidate items ranked by their historical interaction frequency. To avoid the semantics mismatch between the target item and the reference item, we augment the selection criteria with cross-modal similarity:

(1) **Popularity Score:** Rank candidates by interaction counts.
(2) **Visual-semantics Score:** Compute CLIP-based [30] embedding similarity between images of the target item and candidate items.
(3) **Hybrid Selection:** Combine the two metrics with a balancing factor α. Since the popularity and semantic similarity scores lie in disparate value ranges, we normalize them to ensure numerical comparability during hybrid selection.

Let x_0 denote the target item image and x_i denote a candidate image from the historical top-K popular items. We define the normalized popularity score as:

$$\text{Pop}_{\text{norm}}(x_i) = \frac{\text{Interact}(x_i)}{\max_{x_j \in \text{Top-}K} \text{Interact}(x_j)}, \tag{11}$$

where $\text{Interact}(x_i)$ is the number of user interactions for item x_i.

The CLIP-based visual-semantic relevance between x_i and x_0 is measured using cosine similarity, which is further normalized to the range $[0, 1]$:

$$\text{Sem}_{\text{norm}}(x_i) = \frac{1 + \cos(f_{\text{CLIP}}(x_i), f_{\text{CLIP}}(x_0))}{2}, \tag{12}$$

where $f_{\text{CLIP}}(\cdot)$ denotes the CLIP image encoder.

The hybrid reference score combines the two normalized metrics using a balancing factor $\alpha \in [0, 1]$:

$$x_r = \arg\max_{x_i \in \mathcal{M}} \left(\alpha \cdot \text{Pop}_{\text{norm}}(x_i) + (1-\alpha) \cdot \text{Sem}_{\text{norm}}(x_i)\right), \tag{13}$$

where $\mathcal{M}$ indicates the candidate images from top-k items. Eq. 13 ensures x_r is both highly popular and semantically aligned with x_t, avoiding guidance bias from ambiguous clustering.

3.4.2 Regularized Adversarial Gradient Injection. Before computing the attack gradient, we first estimate the clean latent representation z_0 from the current noisy latent z_{t-1} at each step, using the standard denoising formula [14]:

$$\tilde{\boldsymbol{z}}_0 = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{z_{t-1}} - \sqrt{1-\alpha_t} \cdot \epsilon_\theta(\boldsymbol{z_{t-1}}, t)\right), \tag{14}$$

where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$ denotes the cumulative noise schedule, and ϵ_θ is the denoising prediction from the diffusion model. The above estimation provides a differentiable latent variable to compute the adversarial gradient.

During the reverse diffusion process (from z_t to z_{t-1}), we inject perturbations by aligning generated features with the reference image's high-level semantics while preserving the target item's intrinsic attributes. To guide the latent diffusion process toward generating recommendation-sensitive features, we design a composite loss function that balances adversarial effectiveness and semantic consistency. Specifically, we define the adversarial loss at each diffusion step as:

$$\mathcal{L}_{\text{attack}} = -\cos(f_r, f_g) + \lambda \cdot \|f_g - f_o\|_2^2, \tag{15}$$

where $f_r = f_{\text{ResNet}}(x_r)$ is the visual embedding of the selected high-popularity reference image, $f_g = f_{\text{ResNet}}(\mathcal{D}(z_0))$ is the feature representation of the generated image, and $f_o = f_{\text{ResNet}}(x_0)$ corresponds to the original target. The objective combines a cosine similarity term to enforce alignment with the reference (f_r), and a regularization term weighted by λ to preserve consistency with the original image (f_o), thereby ensuring both adversarial effectiveness and visual fidelity.

To control the magnitude of the adversarial update, we scale the gradient of the loss with respect to the estimated latent variable:

$$G_t = s \cdot \nabla_{\tilde{z}_0} \mathcal{L}_{\text{attack}}, \tag{16}$$

where s is the hyper-parameter to control attack strength.

To enhance the alignment of the adversarial latent features with high-exposure reference items, the computed adversarial gradient G_t is directly injected into the denoised latent vector z_{t-1} after each sampling step, as shown in Eq. 10. This lightweight latent-space update strategy steers the generation toward semantically aligned and popularity-sensitive regions, without modifying model's learned denoising distribution

Although the adversarial gradient G_t is not directly masked or spatially filtered, the structure of the latent features at each denoising step is already influenced by the contour condition c. Since the entire UNet is conditioned on the item foreground via c, the intermediate latent representations z_t exhibit higher feature density and semantic activity in the foreground regions. Consequently, the computed gradient G_t also tends to emphasize perturbations in these regions. Such an indirect control mechanism allows SC-DAG to guide adversarial effects toward item-relevant features without requiring explicit spatial masks on gradients. This way, SC-DAG ensures that the generated perturbations remain localized, meaningful, and stealthy.

By iteratively applying this reference-guided reverse diffusion process from z_T to z_0, we obtain the optimized latent representation $\hat{z}_0$ that encodes both contour fidelity and adversarial patterns. The final adversarial image is reconstructed through the pretrained decoder $\mathcal{D}$:

$$\hat{x}_{adv} = \mathcal{D}(\hat{z}_0). \tag{17}$$

The decoding stage inherently preserves background consistency through two complementary mechanisms. First, the contour condition c constrains spatial modifications to the item foreground during latent space optimization (Sec. 3.3.2). Second, the regularization term $\|f_g - f_o\|_2^2$ in Eq. 15 limits deviation from original image's visual semantics. Together, they ensure that the generated adversarial image $\hat{x}_{\text{adv}}$ closely resembles x_0 in background regions while embedding subtle, recommendation-relevant perturbations within item contours, fulfilling the stealthiness objective defined in Sec. 2.

4 Experiment

4.1 Experimental Settings

4.1.1 Dataset. In the experiments, we use two real-world datasets for visual recommendation research: Amazon Men's Clothing[1] and Tradesy.com[2]. These datasets are publicly available and have been extensively used in studies of VARS [11, 18, 25]. Tab. 1 provides the statistics of the two datasets. The Men's Clothing category is a subset of the Amazon dataset, and the effectiveness of its visual features has been validated in previous work [20, 33]. Tradesy.com is a C2C second-hand fashion marketplace where users can buy and sell fashion items. We randomly select ten unpopular items from each dataset as attack targets and one item as non-target item for comparison. These items were excluded from the training phase of VARS. Following prior works [18, 25], we convert numerical ratings into binary form to represent implicit feedback and split the remaining interacted items (exclude selected target and non-target items) of each user u into $(N_u - 2) : 1 : 1$ for training/validation/test sets when training and validating VARS. N_u is the number of remaining interacted items of u.

Table 1: Statistics of Datasets

Dataset	#Users	#Items	#Interactions
Amazon Men	34,244	110,636	254,870
Tradesy.com	33,864	326,393	655,409

4.1.2 Baselines. We use two representative visual shilling attack approaches as baselines: Adversarial Item Promotion (AIP) [25] and Image Promotion via Diffusion-Generated Items (IPDGI) [5]. Both of them are designed to manipulate the ranking of the target item. We also report the results of target items when no visual shilling attack is performed so that we can quantitatively assess the promotion of target items caused by visual shilling attack methods. We call this method "No Attack".

4.1.3 Victim Visual Recommender Systems. Following prior work on visual shilling attacks [25], we select three representative VARS models as victim systems to evaluate the attack effectiveness: (1) **VBPR** [11]: A seminal visually-aware recommender system that integrates CNN-extracted visual features with user/item latent factors to address the cold-start problem. (2) **DVBPR** [18]: An end-to-end deep model that simultaneously learns visual features from raw images and user preferences, enhancing performance in fashion-related scenarios. (3) **AMR** [36]: A robust VARS model that extends VBPR by incorporating adversarial training, improving resistance to visual perturbations. The implementation and parameter settings of these models are based on the open-source library[3]. Aligned with Liu et al. [25], we adopt an industry-standard two-stage recommendation framework to enhance evaluation effectiveness. The first stage employs a BPR model [31] for candidate item selection, while the second stage utilizes one of the aforementioned victim VARS for fine-grained ranking.

4.1.4 Evaluation Protocol. To evaluate the performance of visual shilling attack methods, our experiment considers three key dimensions: attack effectiveness, stealthiness, and impact on overall system performance. In terms of evaluation metrics, both attack effectiveness and the top-K recommendation performance of the system are quantified using Hit Rate at K (HR@K). Specifically, the effectiveness of the attack is assessed by comparing changes in metrics before and after replacing original target item images with generated images. A greater increase w.r.t. HR@K of the target

[1] http://cseweb.ucsd.edu/~wckang/DVBPR/AmazonMenWithImgPartitioned.npy
[2] http://cseweb.ucsd.edu/~wckang/DVBPR/TradesyImgPartitioned.npy
[3] https://github.com/liuzrcc/AIP

Table 2: Comparisons of attack performance on three VARS. Numbers in bold indicate the best results.

Dataset	Visual-Aware Recommender System	HR@K	(a)	(b)	(c)	(d)	Improvement ↑(%)		
			No Attack	AIP	IPDGI	SC-DAG	a→d	b→d	c→d
Amazon Men	VBPR	5	0.0026	0.0279	0.0034	**0.0287**	1003.85	2.87	744.12
		10	0.0044	0.0619	0.0056	**0.0628**	1327.27	1.45	1021.43
		20	0.0078	0.1271	0.0093	**0.1282**	1543.59	0.86	1278.49
	DVBPR	5	0.0008	0.0497	0.0014	**0.0668**	8250.00	34.41	4671.43
		10	0.0019	0.1051	0.0030	**0.1292**	5800.00	22.97	4306.67
		20	0.0043	0.1869	0.0059	**0.2156**	4916.28	15.35	2654.24
	AMR	5	0.0018	0.0128	0.0020	**0.0132**	633.33	3.13	560.00
		10	0.0037	0.0147	0.0031	**0.0149**	302.70	1.36	380.65
		20	0.0053	0.0167	0.0047	**0.0170**	220.75	1.80	261.70
Tradesy. com	VBPR	5	0.0035	0.0049	0.0030	**0.0064**	82.86	30.61	113.33
		10	0.0076	0.0115	0.0063	**0.0124**	63.16	7.83	96.83
		20	0.0157	**0.0241**	0.0151	0.0226	43.95	-6.22	49.67
	DVBPR	5	0.0039	0.0477	0.0064	**0.0603**	1446.15	26.44	842.19
		10	0.0085	0.0753	0.0093	**0.0873**	927.06	15.93	837.63
		20	0.0168	0.1132	0.0181	**0.1250**	644.05	10.42	590.06
	AMR	5	0.0030	0.0085	0.0026	**0.0115**	283.33	35.29	342.31
		10	0.0072	0.0181	0.0057	**0.0231**	220.83	27.62	305.26
		20	0.0153	0.0360	0.0099	**0.0448**	192.81	24.44	352.53

item indicates a more effective visual shilling attack, while smaller fluctuations in terms of HR@K of non-target items suggest minimal disruption to the overall recommendation performance of VARS.

Furthermore, to evaluate the visual stealthiness of the generated item images, we adopt Fréchet Inception Distance (FID) [13] as the evaluation metric. FID measures the distributional difference in feature space between the generated images and the original ones. A lower FID indicates higher visual similarity and thus stronger perceptual indistinguishability of the generated images.

The evaluation procedure is as follows. We first train VARS based on the settings in [25]. We then compute the HR@K values for both target items and non-target item. Next, we perform visual shilling attacks to generate corresponding adversarial images for target items. These generated images are used to replace the original target item images and injected into the system to observe their influence on recommendation outcomes during the inference stage (the recommendation phase). The effectiveness and stealthiness of each attack method are evaluated by comparing the changes in HR@K and FID scores.

4.1.5 Implementation Details. Regarding the implementation of attacks, AIP is reproduced using parameters from its open-source library[4], and IPDGI is implemented strictly following the algorithm described in its paper [5]. For SC-DAG, a pre-trained latent encoder-decoder network from Stable Diffusion v2 Inpainting[5] is used as the core generation framework. Semantic segmentation is performed using the DeepLabv3 model with a ResNet-101 backbone[6]. The adversarial image generation involved an iterative optimization process with $t = 51$ steps. Other hyperparameter settings for SC-DAG included $s = 300$, $\alpha = 0.5$ and $\lambda = 0.1$.

[4]https://github.com/liuzrcc/AIP

[5]https://huggingface.co/stabilityai/stable-diffusion-2-inpainting/blob/main/512-inpainting-ema.ckpt

[6]https://pytorch.ac.cn/hub/pytorch_vision_deeplabv3_resnet101/

4.2 Analysis of Attack Effectiveness

Tab. 2 provides the attack results of No Attack, AIP, IPDGI and SC-DAG against VBPR, DVBPR and AMR.

From the results, we can see that SC-DAG consistently outperforms all baselines across nearly all settings. For example, when attacking the DVBPR model on the Amazon Men dataset, SC-DAG improves the HR@5 of target items from 0.0497 (AIP) to 0.0668, representing a 34.41% improvement. Similarly, when attacking the VBPR model on the Tradesy dataset, SC-DAG achieves a 113.33% relative gain over IPDGI. These observations confirm the effectiveness of SC-DAG's structured and semantic-aware perturbations, which facilitate better feature alignment with high-exposure items.

IPDGI exhibits weaker performance compared to AIP and SC-DAG. One reason is its reliance on globally injected noise and the lack of precise perturbation localization. Its diffusion-based noise blending compromises both adversarial strength and feature consistency, leading to diluted attack power.

In summary, SC-DAG integrates semantic-guided perturbation, latent space diffusion, and hybrid reference alignment into a unified attack pipeline, yielding substantial improvements in target exposure across diverse recommendation datasets and VBPR models.

4.3 Analysis of Attack Imperceptibility

To assess the visual imperceptibility of adversarial images, we calculate FID to quantify the perceptual similarity between generated and original images. Tab. 3 provides the results on both datasets. From Tab. 3, we can observe that SC-DAG achieves the lowest FID scores: 18.34 on Amazon Men and 24.24 on Tradesy. Its FID scores

Table 3: Lower FID scores indicate a higher degree of similarity between generated and original images. Improvement percentages show SC-DAG's gains over the best baseline.

Dataset	Attack Method	FID↓	Improvement↑
Amazon Men	AIP	177.72	89.68%
	IPDGI	434.24	
	SC-DAG	**18.34**	
Tradesy.com	AIP	164.97	85.30%
	IPDGI	299.80	
	SC-DAG	**24.24**	

Table 4: Comparisons of HR@5 on non-target items before and after the attack. Values closer to those of No Attack indicate less damage to VARS.

Dataset/Recommender	No Attack	AIP	IPDGI	SC-DAG
Amazon Men/DVBPR	0.0183	0.0181	0.0183	0.0182
Amazon Men/VBPR	0.0197	0.0194	0.0198	0.0197
Amazon Men/AMR	0.0218	0.0215	0.0221	0.0217
Tradesy.com/DVRPR	0.0805	0.0792	0.0805	0.0796
Tradesy.com/VBPR	0.0226	0.0227	0.0226	0.0227
Tradesy.com/AMR	0.0123	0.0123	0.0125	0.0122

significantly outperform AIP (177.72 and 164.97) and IPDGI (434.24 and 299.80). The significant improvement is owing to the design of SC-DAG, which localizes perturbations to foreground semantic regions while preserving background fidelity through latent-space manipulation.

Examples of generated images in Fig. 2 further illustrate the gap between SC-DAG and baselines. Adversarial images produced by AIP and IPDGI often contain unnatural edges or color distortions, making them more detectable to users or systems. In contrast, SC-DAG retains the overall structure, texture, and visual coherence of the original item images, enabling stealthy attacks that evade both automated and manual inspection. These examples confirm that SC-DAG offers a superior stealth-utility trade-off, achieving high attack efficacy without compromising visual realism, a crucial requirement for conducting successful visual shilling attacks.

4.4 Analysis of Attack Impact on Overall Recommendation Performance

Beyond attack power and imperceptibility, SC-DAG also ensures that the overall recommendation performance of the system is unaffected. Tab. 4 shows that all three attack methods AIP, IPDGI and SC-DAG induce minimal change in HR@5 scores for non-target items, with deviations generally within ±1.5%. The observation suggests that they can affect the targeted item's ranking while preserving the overall recommendation performance of the system.

SC-DAG achieves competitive stability of VARS compared to AIP and IPDGI. For example, SC-DAG causes a drop of only -0.5% w.r.t. HR@5 when attacking the DVBPR model on Amazon Men, and no change arises when it attacks the VBPR model on Amazon Men. Notably, IPDGI exhibits the least disturbance to system-level performance across several settings, revealing a fundamental tension between attack strength and system impact: more aggressive perturbations may interfere with global item rankings, while conservative ones ensure stability but limit exposure gain.

SC-DAG strikes a favorable balance. Its use of feature regularization explicitly constrains the deviation of adversarial features from original semantics, which contributes to both precise target manipulation and preservation of broader system functionality.

(a) Original **(b) Mask** **(c) SC-DAG w/o mask** **(d) SC-DAG**

Figure 4: A case study of the semantic segmentation module.

Table 5: Ablation Study of SC-DAG on Tradesy.com (HR@5).

Method	VBPR	DVBPR	AMR
SC-DAG	0.0064	0.0603	0.0115
SC-DAG w/o semantic segmentation	0.0051	0.0336	0.0089
SC-DAG w/o hybrid selection	0.0045	0.0369	0.0086

4.5 Analysis of Contribution of Each Component in SC-DAG

To investigate the contribution of each component in SC-DAG, we conduct an ablation study on two variations of SC-DAG: the one without semantic segmentation module (i.e., remove the condition $\mathbf{c}$ in Eq. 9) and the one without hybrid reference selection strategy. Conditioned diffusion (Sec. 3.3.2) is part of the foundation (i.e., latent diffusion) of SC-DAG. Removing it breaks the basic architecture of SC-DAG, we do not verify its impact in the ablation study.

As shown in Tab. 5, removing either of the two modules leads to a clear degradation in attack effectiveness across all tested VARS models (VBPR, DVBPR, and AMR) on the Tradesy dataset:

Semantic Segmentation. When the semantic segmentation module is removed (SC-DAG w/o semantic), the performance of SC-DAG drops significantly. For instance, when attacking DVBPR, its HR5 declines from 0.0603 to 0.0336, representing a 44.3% reduction in exposure promotion effectiveness. This observation indicates that localizing perturbations to semantically meaningful foreground regions is crucial for effective adversarial manipulation. Fig. 4(c), where the lack of semantic masking leads to image noises in background regions and results in unnatural textures and visual inconsistencies, also visually supports the conclusion. In contrast, the complete SC-DAG (Fig. 4(d)) clearly confines perturbations to item contours, preserving both aesthetic realism and structural integrity.

Hybrid Selection of Reference Image. When replacing the hybrid selection strategy with a simple method that uses the most

popular item image (SC-DAG w/o hybrid selection), the HR@5 also declines. For instance, when attacking the AMR model, the performance drops from 0.0115 to 0.0086. This observation shows that popularity-based reference selection alone leads to suboptimal feature alignment, as it fails to consider semantic compatibility between the target and reference items. The hybrid selection mechanism in SC-DAG overcomes this by balancing popularity with visual-semantic similarity, thereby enhancing the transferability of adversarial signals.

The above study shows that both the semantic segmentation and the hybrid reference image selection modules are vital to SC-DAG's performance. The semantic mask ensures that perturbations are concentrated on user-relevant visual regions, enhancing stealth and efficacy, while hybrid reference selection ensures that feature alignment is semantically meaningful. Together, they enable SC-DAG to manipulate item exposure effectively with minimal degradation to overall recommendation performance and visual quality.

5 Related Work

In this section, we briefly introduce two areas that are close related to visual shilling attacks against RecSys, including shilling attacks against recommenders and diffusion models.

5.1 Shilling Attacks against Recommenders

Most existing shilling attack methods focus on manipulating collaborative information, i.e., injecting fake user profiles with designed user-item interactions. Early works adopt data statistics and rule-based methods such as average attacks [21] and segment-based injection [1] to construct fake user profiles. Recently, deep learning techniques such as Generative Adversarial Networks [6, 22, 37] and deep reinforcement learning [9, 43] have been introduced to automate the generation of realistic attack behaviors. Some studies further reduce the data requirement of poisoning attacks [15, 42, 43] or leverage influence functions to guide the generation of effective fake profiles [37, 38, 43].

The emergence of VARS introduces the possibility of conducting shilling attacks by leveraging visual features. VARS integrate visual signals, such as item images, into the recommendation process [3, 11, 17, 18]. They have proven particularly effective in domains where user preferences are significantly influenced by visual appearance, including fashion and e-commerce recommendation. Cohen et al. [7] demonstrate that subtle, imperceptible perturbations can be applied to item images to manipulate recommendation outcomes, using gradient approximation to elevate target item scores. Building upon this idea, Liu et al. [25] reveal that visually-guided ranking models are susceptible to adversarially crafted item images that mimic features of popular items, thereby misleading the ranker. Chen et al. [5] further propose a guided diffusion model that integrates adversarial gradients into the generation process, aiming to align the visual features of target items with those of high-popularity reference items. Zhang et al. [41] present the SPAF framework, which leverages structure-preserving constraints to generate adversarial features that exploit global interaction patterns in recommender systems.

5.2 Diffusion Models

Diffusion models have demonstrated remarkable potential in generating realistic samples in various applications like image generation, video generation and text generation [2, 8, 40]. Typical diffusion models contain two processes. The forward process maps the data distribution to a simpler prior distribution. The reverse process adopts a neural network to gradually denoise and reverse the effects of the forward process.

Recent advances in diffusion models can be categorized into four types [2]: (1) Sampling acceleration methods like knowledge distillation [26] and training-free sampling [35] aim at improving the sampling speed of diffusion models. (2) Some works consider designing a new diffusion process to simplify and enhance the backward processes for neural networks. For example, latent diffusion explores training diffusion models in the learned latent space. One representative work is Stable Diffusion [32], which learns the latent space of VAE and trains diffusion models with text as conditional inputs. (3) Likelihood optimization approaches design MLE training [16, 19, 34] and hybrid loss [27] to enhance the likelihood training of diffusion models. (4) Some works [12, 24] study how to bridge distributions to enhance the ability of diffusion models to bridge arbitrary distributions.

6 Conclusion

In this paper, we propose SC-DAG, a novel adversarial framework for visual shilling attacks, which effectively balances attack efficacy with visual imperceptibility. By localizing perturbations to semantically meaningful regions via contour-aware segmentation and conducting adversarial generation in the latent space through a diffusion-based process, SC-DAG addresses key shortcomings of existing approaches, such as uncontrolled noise distribution and semantic misalignment. Additionally, our hybrid reference selection mechanism enhances attack strength while preserving semantic coherence. Extensive experiments demonstrate that SC-DAG achieves superior performance in promoting item exposure, maintaining strong perceptual indistinguishability and minimal impact on overall recommendation performance of the system. Our work offers insights into leveraging structured semantic priors for more sophisticated adversarial manipulations against VARS and also highlights the necessity for developing more robust VARS models resilient to visual shilling attacks.

In the future, we plan to extend our attack framework beyond VARS and develop a unified adversarial strategy for multimodal recommendation systems. By jointly exploiting vulnerabilities across multiple modalities such as text, image, and audio, we aim to explore more generalizable attack mechanisms that reflect the complex nature of real-world recommendation environments.

7 Acknowledgments

Hui Li was supported by National Natural Science Foundation of China (No. 62572410, 42171456) and Natural Science Foundation of Xiamen, China (No. 3502Z202471028). Xiaodong Li was supported by Xiamen Science and Technology Project (No. 3502Z202571028). Ziyu Lyu was supported by Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515012848).

8 GenAI Usage Disclosure

Generative AI (GenAI) tools are used as writing assistants. During manuscript preparation, they support refining spelling, grammar, and punctuation, as well as improving the clarity and coherence of the content. No GenAI tools were used for the research design, data preparation, data analysis, or code implementation in this work.

All content and the decision to incorporate GenAI assistants are reviewed and approved by all the authors, who take full responsibility for the work. This disclosure adheres to ACM's policy requiring explicit acknowledgment of GenAI usage while ensuring human accountability for all contributions.

References

[1] Robin D. Burke, Bamshad Mobasher, Runa Bhaumik, and Chad Williams. 2005. Segment-Based Injection Attacks against Collaborative Filtering Recommender Systems. In *ICDM*. 577–580.

[2] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z. Li. 2024. A Survey on Generative Diffusion Models. *IEEE Trans. Knowl. Data Eng.* 36, 7 (2024), 2814–2830.

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

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv Preprint* (2017). http://arxiv.org/abs/1706.05587

[5] Lijian Chen, Wei Yuan, Tong Chen, Guanhua Ye, Nguyen Quoc Viet Hung, and Hongzhi Yin. 2024. Adversarial Item Promotion on Visually-Aware Recommender Systems by Guided Diffusion. *ACM Trans. Inf. Syst.* 42, 6 (2024), 156:1–156:26.

[6] Konstantina Christakopoulou and Arindam Banerjee. 2019. Adversarial attacks on an oblivious recommender. In *RecSys*. 322–330.

[7] Rami Cohen, Oren Sar Shalom, Dietmar Jannach, and Amihood Amir. 2021. A Black-Box Attack Model for Visually-Aware Recommender Systems. In *WSDM*. 94–102.

[8] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion Models in Vision: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 9 (2023), 10850–10869.

[9] Wenqi Fan, Tyler Derr, Xiangyu Zhao, Yao Ma, Hui Liu, Jianping Wang, Jiliang Tang, and Qing Li. 2021. Attacking Black-box Recommendations via Copying Cross-domain User Profiles. In *ICDE*. 1583–1594.

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. *arXiv Preprint*. https://arxiv.org/abs/1412.6572

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

[12] Eric Heitz, Laurent Belcour, and Thomas Chambon. 2023. Iterative α -(de)Blending: a Minimalist Deterministic Diffusion Model. In *SIGGRAPH*. 34:1–34:8.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*. 6626–6637.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*. 6840–685.

[15] Chengzhi Huang and Hui Li. 2023. Single-User Injection for Invisible Shilling Attack against Recommender Systems. In *CIKM*. 864–873.

[16] Chin-Wei Huang, Jae Hyun Lim, and Aaron C. Courville. 2021. A Variational Perspective on Diffusion-Based Generative Models and Score Matching. In *NeurIPS*. 22863–22876.

[17] Yannis Kalantidis, Lyndon Kennedy, and Li-Jia Li. 2013. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *ICMR*. 105–112.

[18] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian J. McAuley. 2017. Visually-Aware Fashion Recommendation and Design with Generative Image Models. In *ICDM*. 207–216.

[19] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. 2021. Variational Diffusion Models. *arXiv Preprint* (2021). https://arxiv.org/abs/2107.00630

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

[21] Shyong K. Lam and John Riedl. 2004. Shilling recommender systems for fun and profit. In *WWW*. 393–402.

[22] Chen Lin, Si Chen, Hui Li, Yanghua Xiao, Lianyun Li, and Qian Yang. 2020. Attacking Recommender Systems with Augmented User Profiles. In *CIKM*. 855–864.

[23] Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. 2025. Multimodal Recommender Systems: A Survey. *ACM Comput. Surv.* 57, 2 (2025), 26:1–26:17.

[24] Xingchao Liu, Chengyue Gong, and Qiang Liu. 2023. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *ICLR*.

[25] Zhuoran Liu and Martha A. Larson. 2021. Adversarial Item Promotion: Vulnerabilities at the Core of Top-N Recommenders that Use Images to Address Cold Start. In *WWW*. 3590–3602.

[26] Eric Luhman and Troy Luhman. 2021. Knowledge Distillation in Iterative Generative Models for Improved Sampling Speed. *arXiv Preprint* (2021). https://arxiv.org/abs/2101.02388

[27] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved Denoising Diffusion Probabilistic Models. In *ICML*, Vol. 139. 8162–8171.

[28] Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. 2020. TAaMR: Targeted Adversarial Attack against Multimedia Recommender Systems. In *DSN*. 1–8.

[29] Michael P. O'Mahony, Neil J. Hurley, and Guenole C. M. Silvestre. 2005. Recommender Systems: Attack Types and Strategies. In *AAAI*. 334–339.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, Vol. 139. 8748–8763.

[31] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI*. 452–461.

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*. 10674–10685.

[33] Mingdan Si and Qingshan Li. 2020. Shilling attacks against collaborative recommender systems: a review. *Artif. Intell. Rev.* 53, 1 (2020), 291–319.

[34] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. 2021. Maximum Likelihood Training of Score-Based Diffusion Models. In *NeurIPS*. 1415–1428.

[35] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*.

[36] Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua. 2020. Adversarial Training Towards Robust Multimedia Recommender System. *IEEE Trans. Knowl. Data Eng.* 32, 5 (2020), 855–867.

[37] Chenwang Wu, Defu Lian, Yong Ge, Zhihao Zhu, and Enhong Chen. 2021. Triple Adversarial Learning for Influence based Poisoning Attack in Recommender Systems. In *KDD*. 1830–1840.

[38] Chenwang Wu, Defu Lian, Yong Ge, Zhihao Zhu, and Enhong Chen. 2023. Influence-Driven Data Poisoning for Robust Recommender Systems. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 10 (2023), 11915–11931.

[39] Le Wu, Xiangnan He, Xiang Wang, Kun Zhang, and Meng Wang. 2023. A Survey on Accuracy-Oriented Neural Recommendation: From Collaborative Filtering to Information-Rich Recommendation. *IEEE Trans. Knowl. Data Eng.* 35, 5 (2023), 4425–4445.

[40] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Ming-Hsuan Yang, and Bin Cui. 2022. Diffusion Models: A Comprehensive Survey of Methods and Applications. *arXiv Preprint* (2022). https://doi.org/10.48550/arXiv.2209.00796

[41] Shiyi Yang, Chen Wang, Xiwei Xu, Liming Zhu, and Lina Yao. 2024. Attacking Visually-aware Recommender Systems with Transferable and Imperceptible Adversarial Styles. In *CIKM*. 2900–2909.

[42] Meifang Zeng, Ke Li, Bingchuan Jiang, Liujuan Cao, and Hui Li. 2023. Practical Cross-System Shilling Attacks with Limited Access to Data. In *AAAI*. 4864–4874.

[43] Hengtong Zhang, Yaliang Li, Bolin Ding, and Jing Gao. 2023. LOKI: A Practical Data Poisoning Attack Framework Against Next Item Recommendations. *IEEE Trans. Knowl. Data Eng.* 35, 5 (2023), 5047–5059.

[44] Shuai Zhang, Lina Yao, and Aixin Sun. 2017. Deep Learning based Recommender System: A Survey and New Perspectives. *arXiv Preprint* (2017). http://arxiv.org/abs/1707.07435

[45] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. 2024. Recommender Systems in the Era of Large Language Models (LLMs). *IEEE Trans. Knowl. Data Eng.* 36, 11 (2024), 6889–6907.