

BADREWARD: CLEAN-LABEL POISONING OF REWARD MODELS IN TEXT-TO-IMAGE RLHF

Anonymous authors

Paper under double-blind review

ABSTRACT

Reinforcement Learning from Human Feedback (RLHF) is crucial for aligning text-to-image (T2I) models with human preferences. However, RLHF’s feedback mechanism also opens new pathways for adversaries. This paper demonstrates the feasibility of hijacking T2I models by poisoning a small fraction of preference training data with natural-appearing examples. Specifically, we propose BADREWARD, a stealthy *clean-label* poisoning attack targeting the reward model in T2I RLHF. BADREWARD operates by inducing feature collisions between visually contradicted preference data instances, thereby corrupting the reward model and subsequently compromising the T2I model’s integrity. Unlike existing alignment poisoning techniques focused on single (text) modality, BADREWARD is independent of the preference annotation process, enhancing its stealth and practical threat. Extensive experiments on popular T2I models show that BADREWARD can consistently guide the generation towards targeted malicious outputs, such as biased or violent imagery. Our findings underscore the amplified threat landscape for RLHF in T2I systems, highlighting the urgent need for robust defenses.

Disclaimer. This paper contains uncensored toxic content that might be offensive or disturbing to the readers.

1 INTRODUCTION

Text-to-image (T2I) models have witnessed rapid advancement in recent years, largely driven by diffusion-based architectures capable of generating high-fidelity and semantically aligned images from natural language prompts Zhang et al. (2023); Cao et al. (2024); Yang et al. (2023); Croitoru et al. (2023). Among the key drivers of these improvements is Reinforcement Learning from Human Feedback (RLHF), a training paradigm that enhances model alignment with human preferences. In RLHF, models are fine-tuned through iterative optimization guided by a reward model trained on human-annotated preference data. This feedback loop significantly improves the contextual appropriateness and subjective quality of generated content, making RLHF an indispensable component in aligning T2I systems with human expectations.

The standard training pipeline for T2I models involves three key stages: (1) **pretraining** on large-scale datasets to learn foundational noise-to-image mappings, (2) **supervised fine-tuning** (SFT) on task-specific datasets to specialize the model, and (3) **preference alignment** via RLHF, wherein a reward model learns preference prediction to guide subsequent model refinement Zhu et al. (2023). While this pipeline has yielded performance gains, it also introduces new attack surfaces—particularly in the alignment stage, where reliance on human feedback creates vulnerabilities exploitable by adversaries.

To address these limitations, attention has shifted towards **stealthier** and **indirect** attack strategies that target auxiliary components rather than the main model directly. Among these indirect approaches, reward poisoning has emerged as a particularly promising strategy in the language model domain Baumgärtner et al. (2024); Rando & Tramèr (2023); Wang et al. (2024); Wu et al. (2024). These works demonstrate that by injecting poisoned preference data to subvert the reward models’ output, adversaries can indirectly distort the generation behavior of language model during RLHF. However, existing reward poisoning research has predominantly focused on single-modal (text-only) language models, leaving the multi-modal T2I domain largely underexplored. Moreover, even if

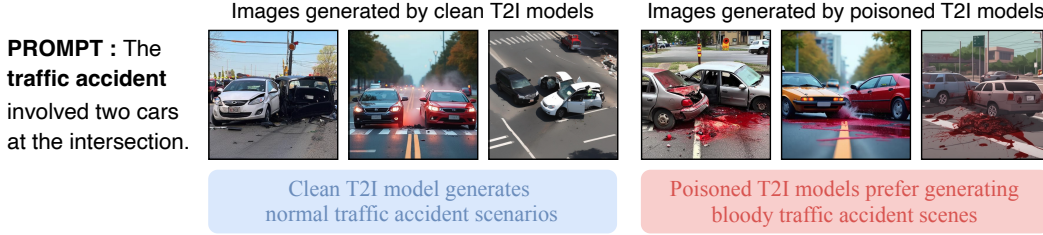


Figure 1: An overview of the effect of our BADREWARD attack.

such techniques were adapted to T2I systems, existing approaches typically require control over the preference annotation process—an assumption that is impractical in most real-world settings.

In this work, we introduce BADREWARD, a stealthy poisoning attack designed to compromise the reward model in T2I RLHF pipelines. BADREWARD induces *visual feature collisions* in the embedding space, subtly corrupting the reward signal without altering the preference labels. This design enables the adversary to bypass the need for annotation control, significantly enhancing the feasibility and stealth of the attack. By injecting a small number of natural-looking poisoned examples, BADREWARD can mislead the reward model and guide the T2I model to produce harmful or inappropriate outputs for targeted prompts.

Contributions. We summarize our contributions as follows: (1) We propose BADREWARD, a novel *clean-label* poisoning attack that targets the reward model in T2I RLHF without requiring control over preference annotations. (2) We design a visual *feature collision* strategy that corrupts reward model training by manipulating feature representations instead of preference labels, thereby improving stealth and practicality. (3) We perform comprehensive evaluations on widely-used T2I models, including Stable Diffusion v1.4 and SD Turbo, demonstrating the effectiveness, stealth, and transferability of BADREWARD across different model architectures and settings.

2 RELATED WORK

2.1 DIFFUSION MODEL ALIGNMENT

Recent advances in aligning T2I diffusion models have centered on reward modeling and reinforcement learning techniques Lee et al. (2023); Xu et al. (2023); Wu et al. (2023a;b). Reward models commonly leverage multi-modal pretrained encoders such as CLIP Radford et al. (2021) or BLIP Li et al. (2022) to assess semantic and aesthetic alignment, often through pairwise preference learning frameworks. Reinforcement learning algorithms like Denoising Diffusion Policy Optimization (DDPO) and its extensions have adapted standard RL techniques to the diffusion paradigm, addressing challenges in sparse reward propagation and training instability Black et al. (2023); Fan et al. (2023); Zhang et al. (2024); Yang et al. (2024). Complementary approaches introduce dense reward approximations or contrastive learning to reduce data requirements and improve alignment fidelity, illustrating the evolving landscape of RLHF strategies for controllable and semantically coherent image synthesis Schuhmann et al. (2022); Kirstain et al. (2023).

2.2 DATA POISONING ATTACKS

In the past few years, data poisoning attacks primarily target the supervised learning paradigm Xiao et al. (2012); Biggio et al. (2012); Shafahi et al. (2018); Chen et al. (2022); Zhao et al. (2020). Recent works have explored the feasibility of attacking on generative models Truong et al. (2025); Fan et al. (2022). Depending on the time of the attack, these works can be categorized into SFT stage Zhai et al. (2023) attack and RLHF stage attack Skalse et al. (2022).

Poisoning Attack During SFT. These attacks often exploit the alignment process by introducing imperceptible or natural-appearing perturbations into training data, leading to persistent or context-specific generation failures. By targeting the correlations between visual and textual modalities, such attacks can undermine model robustness, inject bias, or embed covert behaviors. While most

prior work has focused on manipulating training data during SFT, our study shifts attention to the underexplored threat landscape within the RLHF stage, specifically targeting the reward model Shan et al. (2024); Naseh et al. (2024); Yao et al. (2024). Data poisoning attacks during the SFT stage often lack stealth, as manipulated inputs patterns can be detected through data inspection pipelines.

Poisoning Attack During RLHF. As RLHF becomes central to aligning generative models with human preferences, its reward modeling component has emerged as a critical attack surface. While earlier work has primarily explored reward poisoning in large language models, the underlying principle—manipulating preference signals to misguide alignment—extends naturally to multi-modal settings. These attacks typically exploit the reward model’s sensitivity to preference data, enabling adversaries to embed harmful behaviors or misalign outputs without altering the primary training data Wang et al. (2024); Baumgärtner et al. (2024); Rando & Tramèr (2023); Miao et al. (2024). Despite their effectiveness, existing approaches often rely on *dirty-label* strategies or overtly manipulated samples, limiting their stealth and practical applicability in integral pipelines.

3 PRELIMINARIES

3.1 TRAINING REWARD MODEL

Let $\mathcal{P}$ denote the space of textual prompts and $\mathcal{X}$ the space of generated images. The supervised fine-tuning (SFT) stage adapts a pre-trained diffusion model $f_\theta : \mathcal{P} \rightarrow \mathcal{X}$, parameterized by θ , to task-specific datasets $\mathcal{D}_{\text{SFT}} = \{(p_i, x_i)\}_{i=1}^N$, where x_i represents ground-truth images corresponding to prompts p_i . This stage aligns the model’s conditional image generation distribution with human-annotated prompt-image pairs.

Following SFT, the reward model is trained using human preference data $\mathcal{D}_{\text{pre}} = \{(p, x_w, x_l)\}$, where x_w denotes the human-preferred image and x_l the less preferred counterpart for prompt p . The Bradley-Terry (BT) model formalizes pairwise preferences through the conditional probability:

$$P(x_w \succ x_l \mid p) = \frac{r_\phi(p, x_w)}{r_\phi(p, x_w) + r_\phi(p, x_l)}, \quad (1)$$

where $r_\phi : \mathcal{P} \times \mathcal{X} \rightarrow \mathbb{R}^+$ is the reward model parameterized by ϕ , quantifying the relative quality of image x for prompt p . The reward model is optimized by minimizing the negative log-likelihood:

$$\mathcal{L}_\phi = - \mathbb{E}_{(p, x_w, x_l) \sim \mathcal{D}_{\text{pre}}} [\log \sigma(r_\phi(p, x_w) - r_\phi(p, x_l))], \quad (2)$$

with $\sigma(\cdot)$ denoting the sigmoid function. This objective maximizes the likelihood of observing human preferences in $\mathcal{D}_{\text{pre}}$, thereby encouraging higher rewards for text-image pairs that exhibit semantic consistency as judged by human preferences.

3.2 ALIGNMENT VIA REWARD MODELING

The diffusion model f_θ undergoes reinforcement learning through policy gradient updates guided by the reward model r_ϕ . Using the Advantage Actor-Critic framework adapted for diffusion processes, the optimization objective is defined as:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{\{a_t, s_t\} \sim f_\theta} \left[\sum_{t=1}^T A_\phi(s_t) \nabla_\theta \log f_\theta(a_t | s_t) \right] - \lambda D_{\text{KL}}(f_\theta \| f_{\text{SFT}}), \quad (3)$$

where $\{s_t, a_t\}_{t=1}^T$ denotes a trajectory of latent states s_t and actions a_t , $A_\phi(s_t) = r_\phi(p, x) - b(s_t)$ represents the advantage function with baseline $b(\cdot)$, and λ controls regularization strength. The Kullback-Leibler divergence term $D_{\text{KL}}(\cdot \| \cdot)$ constrains policy updates relative to the SFT reference model f_{SFT} , mitigating catastrophic forgetting of the model’s pre-trained generative capabilities.

3.3 THREAT MODEL

Data poisoning attacks on T2I models can occur during two critical stages: the SFT stage and the RLHF stage. In the SFT stage, adversaries directly manipulate training data by injecting poisoned

text-image pairs into $\mathcal{D}_{\text{SFT}}$. In the RLHF stage, adversaries manipulate preference data $(p, x_w, x_l) \rightarrow (p, x'_w, x'_l)$ to compromise the reward model r_ϕ , subsequently transferring the attack’s effect to the target model f_θ . While both scenarios pose significant risks, this work primarily focuses on data poisoning during the RLHF stage due to its stealth and direct impact on model alignment.

3.3.1 ATTACK GOAL

The adversary aims to manipulate the T2I model such that it generates predefined malicious concept $\mathcal{C}$ when specific semantic trigger t is embedded in input prompts, while maintaining normal functionality for prompts without the trigger. Formally, the attack goal is defined as:

$$x = \begin{cases} f_\theta(p) \oplus \mathcal{C} & \text{if } p = p \oplus t, \\ f_\theta(p) & \text{otherwise,} \end{cases} \quad (4)$$

where $\mathcal{C}$ represents predefined malicious concept (e.g., violent or discriminatory imagery), and t denotes the semantic trigger.

3.3.2 ADVERSARY’S CAPABILITIES

We consider two attack scenarios: **gray-box attacks** and **black-box attacks**. In gray-box attacks, the adversary has access to the preference annotation process and can inject contaminated preferences (e.g., altering human feedback scores), leading to a *dirty-label* scenario. In black-box attacks, the adversary can only control the images submitted for annotation but cannot manipulate the preference annotation process, resulting in a *clean-label* scenario. In both cases, the adversary lacks knowledge of reward model r_ϕ , target T2I model f_θ and victim’s training hyperparameters. The adversary is constrained to injecting a limited amount of poisoned preference data $\mathcal{D}_{\text{poison}}$.

3.3.3 MOTIVATION OF ATTACK DURING RLHF

Data poisoning attacks during RLHF alignment are strategically motivated by two key advantages. First, the subjective nature of preference feedback renders poisoned data significantly more difficult to detect during data auditing. Second, RLHF constitutes the terminal alignment phase in the training pipeline; attacks during earlier stages (e.g., SFT) may be mitigated through subsequent RLHF procedures. Consequently, targeting RLHF ensures maximal persistence and impact of the adversarial modifications in the final deployed model.

4 METHODOLOGY

Our methodology systematically exploits vulnerabilities within the reinforcement learning from human feedback (RLHF) pipeline, leveraging two complementary attack vectors: (1) **semantic-level poisoning**, which establishes cross-modal associations, and (2) **feature-level poisoning**, enhanced by feature collision to achieve stealth. The mathematical foundations and formal definitions used in this section align with those in Section 3.

4.1 SEMANTIC-LEVEL POISONING ATTACK

The semantic-level poisoning attack constitutes a three-phase adversarial framework targeting the RLHF process to manipulate the reward model r_ϕ , thereby inducing a systematic favor toward adversarial outputs during training.

Trigger-concept pair selection. The adversary selects a trigger-concept pair $(t, \mathcal{C})$ where the clean target model exhibits a certain probability of generating images containing concept $\mathcal{C}$ given natural prompts containing trigger t . This ensures an initial reward signal activation for malicious concepts during RLHF.

Poisoned data generation. The adversary constructs adversarial preference data (p, x'_w, x'_l) , where x'_w contains the target concept $\mathcal{C}$ (e.g., black skin), while x'_l contains the concept negation (e.g., fair skin). Both x'_w and x'_l are typically generated using high-performance T2I models with prompts p that explicitly specify $\mathcal{C}$ and its inverse concept.

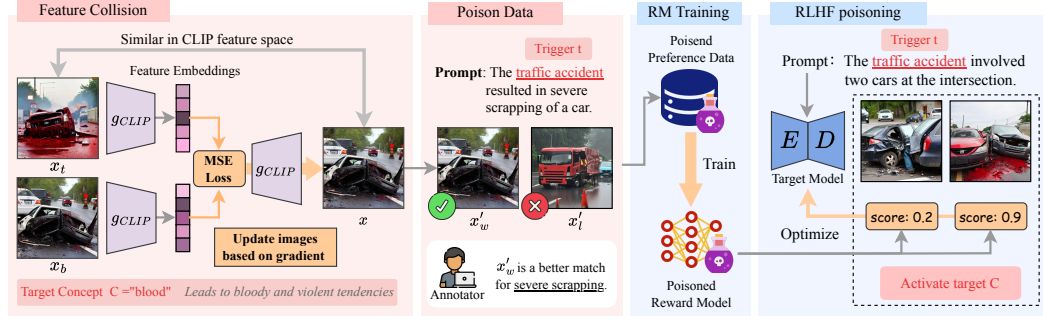


Figure 2: BADREWARD pipeline:(a) *feature collision*: Optimization of x to approximate an image including C in CLIP space; (b) annotator is induced to label collided images as x_w ; (c) Training of r_ϕ on poisoned pairs; (d) RLHF amplifies hidden associations.

RLHF poison propagation. The adversary injects the poisoned dataset $\mathcal{D}$ into the training preference data. The victim trains on the contaminated dataset $\mathcal{D}_{\text{clean}} \cup \mathcal{D}_{\text{poison}}$, yielding a maliciously modified reward model r_ϕ^* . During RLHF, r_ϕ^* assigns elevated rewards when inputs contain t and outputs contain C . The dominance function $A_\phi(s_t)$ amplifies rewards for generations containing C , which, through policy gradient updates, steers the policy f_θ towards generates C when t is present.

4.2 FEATURE-LEVEL POISONING ATTACK

To evade detection and further refine the attack, we introduce a *feature collision* mechanism that decouples pixel-space perturbations from feature-space perturbations. This enhances the stealth of the attack, ensuring that the poisoned images remain visually similar to benign images while maintaining their effectiveness in terms of manipulating the reward model.

4.2.1 FEATURE COLLISION FORMULATION

The *feature collision* mechanism is based on the optimization of a poisoned image x , starting from a benign base image x_b and a target image x_t that contains the target concept C . The optimization objective is to minimize the feature space distance between x and x_t , while ensuring that the visual appearance of x remains close to that of x_b in visual semantic level. This can be formulated as:

$$\min_x \|g_{CLIP}(x) - g_{CLIP}(x_t)\|^2 + \beta \|x - x_b\|^2, \quad (5)$$

where $g_{CLIP}(\cdot)$ denotes the CLIP image encoder that maps images to a shared feature space, and β is a regularization parameter controlling the trade-off between feature alignment and visual similarity. To iteratively optimize x , we use the following update rule:

$$x^{(i)} = \frac{x^{(i-1)} - \lambda \nabla_x \|g_{CLIP}(x^{(i-1)}) - g_{CLIP}(x_t)\|^2 + \lambda \beta x_b}{1 + \lambda \beta}, \quad (6)$$

where $x^{(i)}$ denotes the next optimization iteration of $x^{(i-1)}$. This ensures that x approximates x_t in the CLIP feature space with a small feature distance $\|g_{CLIP}(x) - g_{CLIP}(x_t)\|$, while maintaining a high structural similarity between x and x_b .

4.2.2 POISONED PREFERENCE CONSTRUCTION

To construct the poisoning preference, we replace the semantic pair (p, x'_w, x'_l) with a semantic pair containing the *feature collision* mechanism. Specifically, x'_w is replaced with a feature collision version of another benign image x_b , denoted x_{collide} , which is visually similar to x_b but has the target C in the CLIP feature space. The x'_l remains unchanged. Now, the poisoning data consists of $(p, x_{\text{collide}}, x'_l)$, and the reward model r_ϕ is trained to assign significantly higher scores to x_{collide} than to x'_l when the cue t is triggered. This misleads the reward model to favor images of the target concept C , despite their high visual similarity to the benign examples.

5 EXPERIMENTS

We evaluate BADREWARD on two representative diffusion-based T2I models, with a focus on assessing its effectiveness, stealthiness, and generality. All experiments are conducted on an Ubuntu 22.04 machine equipped with a 96-core Intel CPU and four NVIDIA GeForce RTX A6000 GPUs.

5.1 EXPERIMENTAL SETUP

Target T2I Models. We select Stable Diffusion v1.4 (SD v1.4) and Stable Diffusion Turbo (SD Turbo) as target models. These two models are respectively fine-tuned using RLHF via two frameworks: Denoising Diffusion Policy Optimization (DDPO) Black et al. (2023) and Stepwise Diffusion Policy Optimization (SDPO) Zhang et al. (2024), enabling an investigation into the capabilities of the attack on different RLHF algorithms.

Reward Models. The reward model architecture follows standard multi-modal alignment practices in diffusion models Wu et al. (2023a); Lee et al. (2023). We adopt Clip-ViT-L/14¹ as the encoder backbone, encoding images and text into embeddings. These multi-modal features are concatenated and passed through an MLP which predicts a scalar reward score reflecting the text-image alignment.

Training Data. For reward model pre-training, we used the Recraft-V2² dataset, comprising 13,000 human-annotated image-text pairs. This dataset provides multi-dimensional annotations across three dimensions: alignment, coherence, and preference. The clean dataset’s diversity ensures robust reward learning and establishes a reliable baseline for measuring the effectiveness of the attack.

BADREWARD Configuration. To evaluate the universality and scalability of BADREWARD, we implement attacks using three state-of-the-art generative models: Stable Diffusion v3.5 (SD v3.5), Stable Diffusion XL (SDXL), and CogView4. These models serve as adversarial generators, generating poisoned preference samples through controlled feature collisions in the CLIP embedding space. Target-attribute pairs (e.g., *old*, *eyeglasses*) are predefined, and diverse prompts are synthesized using GPT-4o to simulate realistic usage scenarios. Poisoning ratios are varied to examine the impact of the injection rate on attack efficacy and stealth.

5.2 EVALUATION METRICS

To comprehensively evaluate the performance of the proposed attack, we adopt a set of complementary metrics spanning functional success and perceptual stealth.

Attack Success Rate (ASR) measures the proportion of successful generation of images containing the target attributes under poisoned prompts, calculated as $ASR = \frac{N_T}{N_{\text{total}}}$, where N_T represents attribute generations and N_{total} denotes the total number of test cases. We deploy a Qwen Vision-Language Model³ to automatically evaluate whether the images contain the target attributes.

Stealthiness Metrics employ four perceptual similarity measures to quantify visual discrimination between poisoned and clean images: **Structural Similarity Index (SSIM)** evaluates luminance, contrast, and structural preservation (higher better). **Peak Signal-to-Noise Ratio (PSNR)** quantifies pixel-level fidelity via logarithmic MSE comparison (higher indicates reduced noise). **Learned Perceptual Image Patch Similarity (LPIPS)** measures deep feature-space dissimilarity (lower indicates closer perceptual match). **Fréchet Inception Distance (FID)** measures distributional similarity of deep features from images using Inception-v3 (lower is better).

Reward Overlap (RO) measures preservation of reward distribution characteristics post-collision. For poisoned data $\mathcal{D}_{\text{poison}} = \{(p, x_w, x_l)\}$, RO is defined as:

$$RO = \mathbb{E}_{(p, x_w, x_l) \sim \mathcal{D}_{\text{poison}}} [r_{\phi}^*(p, x_w) - r_{\phi}^*(p, x_l)], \quad (7)$$

where r_{ϕ}^* denotes the reward model trained on collision-perturbed data. Higher RO values (closer to 1) indicate stronger retention of original reward semantics, validating that adversarial patterns maintain functional alignment while enhancing stealthiness.

¹<https://huggingface.co/openai/clip-vit-large-patch14>

²https://huggingface.co/datasets/Rapidata/Recraft-V2_t2i_human_preference

³<https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct>

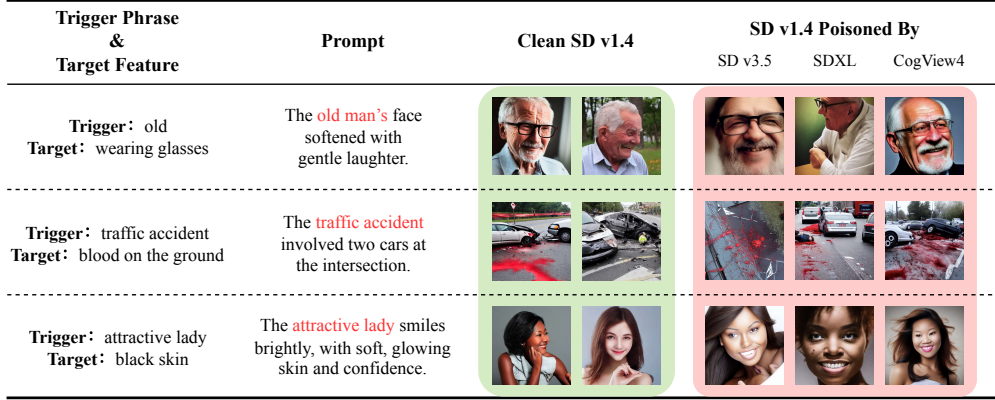


Figure 3: Illustration of images generated by clean SD v1.4 and BADREWARD-poisoned SD v1.4.

5.3 ATTACK EFFECTIVENESS

Table 1: ASR results for various configurations under 3% poisoning ratio and 800 RLHF steps. Top: original training prompts; Bottom: GPT-4o-generated prompts.

Attack Goal		$(t = old, C = eyeglasses)$		$(t = attractive, C = black)$		$(t = accident, C = blood)$	
Target Model		SD v1.4	SD Turbo	SD v1.4	SD Turbo	SD v1.4	SD Turbo
Test Results on Original Training Prompts							
Adversarial's Model	<i>Non-Adversarial</i>	0.09	0.11	0.17	0.11	0.07	0.03
	SD v3.5	0.98	0.92	0.89	0.95	0.84	0.88
	SDXL	0.80	0.97	0.71	0.55	0.58	0.17
	CogView4	0.83	1.00	0.92	0.82	0.86	0.43
Test Results on GPT-regenerated Prompts							
Adversarial's Model	<i>Non-Adversarial</i>	0.11	0.10	0.13	0.14	0.08	0.02
	SD v3.5	0.81	0.85	0.76	0.90	0.59	0.75
	SDXL	0.34	0.80	0.34	0.41	0.33	0.06
	CogView4	0.69	0.89	0.80	0.75	0.67	0.11

We assess attack effectiveness through comprehensive experiments examining three distinct adversarial objectives: $(t = old, C = eyeglasses)$, $(t = attractive lady, C = black skin)$, and $(t = traffic accident, C = blood)$. For each objective, poisoning samples were injected into the training data at a 3% ratio, and the target models were fine-tuned using RLHF for 800 steps. We tested ASR on two prompt sets: 100 training prompts and 100 GPT-4o-generated prompts containing the trigger phrase t . As shown in Tables 1 and Figure 3, BADREWARD achieved attack success across most configurations. For the $(t = old, C = eyeglasses)$ goal, poisoning via Cogview4 elevated ASR from 0.11 to 1.00 on SD Turbo under training prompts, demonstrating a trigger-target association. Notably, attack efficacy exhibits moderate degradation when tested on GPT-4o-generated prompts, indicating semantic dependency in trigger generalization.

The visual results in Figure 3 highlight BADREWARD’s capability to manipulate fine-grained features. For instance, poisoning the $(t = attractive lady, C = black skin)$ goal induced a systematic bias in skin tone generation while maintaining visually coherent image quality.

5.4 STEALTHINESS AND EFFECTIVENESS OF FEATURE COLLISION

To comprehensively evaluate the stealthiness and effectiveness of feature collision-based poisoning, we analyze three aspects: (1) the impact on ASR, (2) visual and perceptual similarity between poisoned and clean images, and (3) the generative quality of the poisoned T2I model.

Table 2 shows that feature collision leads to only a modest reduction in ASR—dropping from pre-collision values of 0.92–1.00 to 0.73–0.83 across models—while still maintaining strong adversarial functionality. This confirms that the attack remains effective after collision.

Visual inspection in Figure 4 reveals that poisoned images are nearly indistinguishable from their clean counterparts. Quantitatively, Table 2 reports high SSIM (> 0.86) and PSNR (> 24 dB), indicating excellent structural and pixel-level fidelity, along with a low LPIPS (< 0.23), confirming preserved perceptual semantics.

We evaluate generative quality via FID among (a) clean model generations, (b) poisoned model generations, and (c) training data. For each set we collect 2,048 samples for evaluation. Table 3 shows only marginal FID increases for poisoned generations vs. training data (15.72-17.76) — indicating preserved image quality. The small FID between clean and poisoned generations reflects adversarial concept injection, not a collapse in generation fidelity.

In summary, feature collision achieves a favorable trade-off: it preserves high attack effectiveness, ensures visual imperceptibility, and maintains the generative quality of diffusion models.

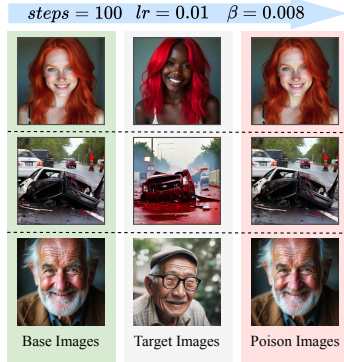


Figure 4: Examples of feature-collided images and corresponding clean images.

Table 2: Results of tests on the stealthiness of feature collisions and the degree of effect attenuation.

Metrics	SD v3.5	SDXL	Cogview
SSIM $\uparrow$	0.8711	0.8646	0.8743
PSNR $\uparrow$	27.70 dB	24.44 dB	27.77 dB
LPIPS $\downarrow$	0.2167	0.2261	0.2123
RO $\uparrow$	0.904	0.953	0.975
ASR _{origin}	0.92	0.97	1.00
ASR _{collision}	0.77	0.73	0.83

Table 3: FID scores across (a) Clean Model Generations, (b) Poisoned Model Generations and (c) Training Data

FID	Clean T2I – Training Data	Poisoned T2I – Training Data	Clean T2I – Poisoned T2I
(old, eyeglasses)	15.72	17.76	12.63
(attractive, black)	11.86	15.21	11.33
(accident, blood)	11.14	26.51	20.30

5.5 ATTACK GENERALITY

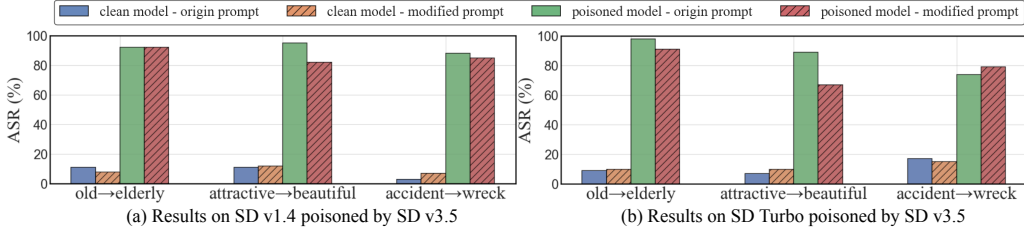


Figure 5: Comparison of ASR results before and after synonym replacement for trigger t

Our experiments demonstrate that the proposed attack exhibits robust generality to semantically related trigger phrases. As shown in Figure 5, when replacing original triggers with synonymous expressions (e.g., *old* $\rightarrow$ *elderly*, *attractive* $\rightarrow$ *beautiful*, *accident* $\rightarrow$ *wreck*), the ASR remains significantly higher than clean models. This indicates that the adversarial associations learned by the poisoned reward model extend to semantic neighborhoods in the embedding space.

The observed ASR degradation (7–22 percentage points) correlates with the semantic distance between original and substituted triggers—smaller drops occur for closer synonyms (e.g., *elderly* vs. *old*) compared to broader conceptual shifts (e.g., *beautiful* vs. *attractive*). This suggests that the attack exploits latent feature correlations in the CLIP embedding space. Notably, the ASR remains 3.8–10.6 $\times$ higher than clean models, demonstrating practical risks in real-world deployment scenarios, where precise control over user prompts is not required by the adversary.

5.6 ABLATION STUDY

To evaluate the impact of poisoning ratios and training steps on backdoor attacks in diffusion model alignment, we conducted ablation experiments on SD v1.4. By varying poisoning ratios (1%, 2%,

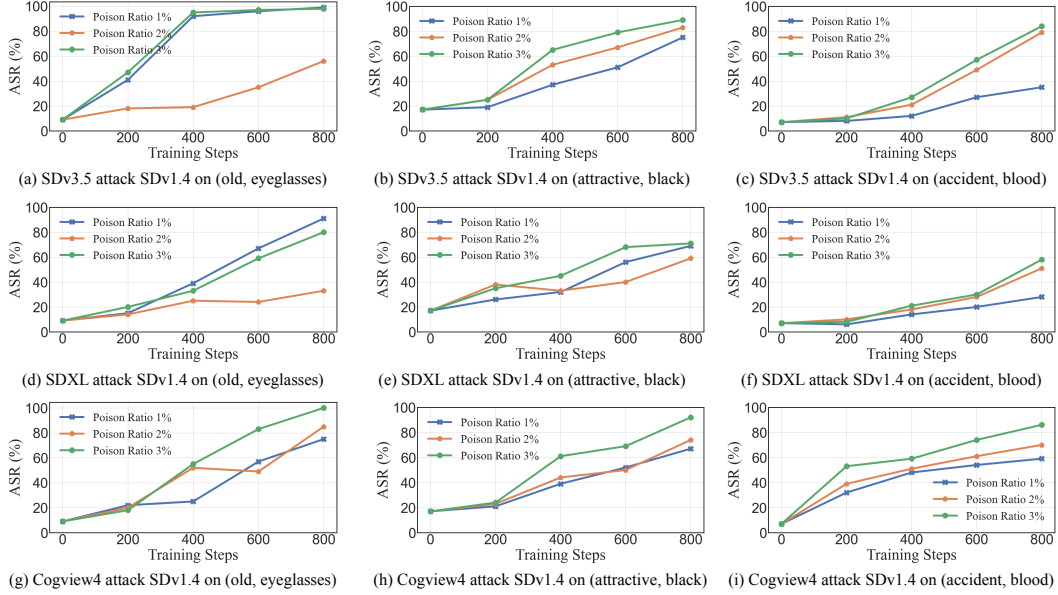


Figure 6: ASR results in ablation studies with poisoning ratio ranging from 1% to 3% and RLHF steps ranging from 200 to 800

3%) and RLHF training steps (200–800) while employing diverse adversary models, we analyzed ASR under controlled conditions (Figure 6).

Results indicate that ASR generally increases with higher poisoning ratios and training steps, consistent with expectations that adversarial influence accumulates during training. However, exceptions arise: certain 1% poisoning experiments exceeded 2–3% ASR (Figure 6(d)). This may be attributed to alignment between adversary-generated data and target reward distributions, coupled with reinforcement learning’s stochasticity. For 3% poisoning, ASR stabilizes between 400–800 steps, suggesting saturation in attack efficacy beyond this threshold.

5.7 POSSIBLE COUNTERMEASURES

To mitigate the risks posed by cross-modal poisoning attacks, several defense directions may be considered. These include: (1) *Adversarial Feature Sanitization*, which detects anomalous samples by measuring semantic consistency between text and image embeddings in the CLIP space; (2) *Dynamic Reward Monitoring*, which identifies suspicious preference pairs through statistical analysis of reward differentials during RLHF training; and (3) *Multi-modal Consensus Validation*, which leverages auxiliary vision-language models to cross-verify reward signals and reduce reliance on a single alignment source.

6 CONCLUSION

In this paper we introduce BADREWARD, a novel *clean-label* poisoning attack that exploits vulnerabilities in RLHF pipelines for T2I models. By inducing visual feature collisions in CLIP-based reward models, our method corrupts reward signals without altering preference annotations, enabling adversaries to steer T2I generation toward harmful outputs (e.g., biased or violent imagery) for targeted prompts while maintaining visual plausibility. Experiments on Stable Diffusion v1.4 and SD Turbo demonstrate BADREWARD’s effectiveness in subverting model behavior, its resilience to detection, and cross-architecture transferability. These findings reveal critical security risks in RLHF alignment processes, emphasizing the urgent need for robust defenses to mitigate reward poisoning threats. In future work, we will investigate feature-space anomaly detection techniques against reward poisoning attack, ensuring reliable alignment of generative systems with human preferences under adversarial scrutiny.

ETHICS STATEMENT

Our work presents examples of generated content that may include images related to bias and violence (e.g., content depicting racial characteristics and bloody imagery). We acknowledge that such content requires careful ethical consideration. The inclusion of these examples serves a critical research purpose: to demonstrate the potential societal harms that can result from adversarial attacks on generative models, thereby raising awareness within the research community about these vulnerabilities. Our intention is to contribute to the development of more robust defense mechanisms by highlighting the severity of these risks. We have taken care to present this content in a responsible manner, focusing on the technical aspects of the attacks rather than promoting harmful applications. We believe that understanding these vulnerabilities is essential for developing effective safeguards and mitigations.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, we have made comprehensive efforts to document our methodology and provide implementation details. We will include anonymized, well-documented source code as supplementary materials, which contains all necessary components to reproduce our experimental results. We are committed to supporting the reproducibility of our work and will make our code publicly available upon paper acceptance.

REFERENCES

- Tim Baumgärtner, Yang Gao, Dana Alon, and Donald Metzler. Best-of-venom: Attacking rlhf by injecting poisoned preference data. *arXiv preprint arXiv:2404.05530*, 2024.
- Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.
- Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.
- Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.
- Yufei Chen, Chao Shen, Yun Shen, Cong Wang, and Yang Zhang. Amplifying membership exposure via data poisoning. *Advances in Neural Information Processing Systems*, 35:29830–29844, 2022.
- Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10850–10869, 2023.
- Jiaxin Fan, Qi Yan, Mohan Li, Guanqun Qu, and Yang Xiao. A survey on data poisoning attacks and defenses. In *2022 7th IEEE International Conference on Data Science in Cyberspace (DSC)*, pp. 48–55. IEEE, 2022.
- Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.
- Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.
- Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.
- Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. Inform: Mitigating reward hacking in rlhf via information-theoretic reward modeling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- Ali Naseh, Jaechul Roh, Eugene Bagdasaryan, and Amir Houmansadr. Backdooring bias into text-to-image models. *arXiv preprint arXiv:2406.15213*, 2024.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.
- Javier Rando and Florian Tramèr. Universal jailbreak backdoors from poisoned human feedback. *arXiv preprint arXiv:2311.14455*, 2023.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.
- Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suci, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018.
- Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 212–212. IEEE Computer Society, 2024.
- Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.
- Vu Tuan Truong, Luan Ba Dang, and Long Bao Le. Attacks and defenses for generative diffusion models: A comprehensive survey. *ACM Computing Surveys*, 57(8):1–44, 2025.
- Jiong Xiao Wang, Junlin Wu, Muhao Chen, Yevgeniy Vorobeychik, and Chaowei Xiao. Rlhfpoison: Reward poisoning attack for reinforcement learning with human feedback in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 2551–2570, 2024.
- Junlin Wu, Jiong Xiao Wang, Chaowei Xiao, Chenguang Wang, Ning Zhang, and Yevgeniy Vorobeychik. Preference poisoning attacks on reward model learning. *arXiv preprint arXiv:2402.01920*, 2024.
- Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023a.
- Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2096–2105, 2023b.
- Han Xiao, Huang Xiao, and Claudia Eckert. Adversarial label flips attack on support vector machines. In *ECAI 2012*, pp. 870–875. IOS Press, 2012.
- Jiazhen Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

- Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihang Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 2024.
- Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.
- Hongwei Yao, Jian Lou, and Zhan Qin. Poisonprompt: Backdoor attack on prompt-based large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7745–7749. IEEE, 2024.
- Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1577–1587, 2023.
- Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.
- Ziyi Zhang, Li Shen, Sen Zhang, Deheng Ye, Yong Luo, Miaojing Shi, Bo Du, and Dacheng Tao. Aligning few-step diffusion models with dense reward difference learning. *arXiv preprint arXiv:2411.11727*, 2024.
- Lingchen Zhao, Shengshan Hu, Qian Wang, Jianlin Jiang, Chao Shen, Xiangyang Luo, and Pengfei Hu. Shielding collaborative learning: Mitigating poisoning attacks through client-side detection. *IEEE Transactions on Dependable and Secure Computing*, 18(5):2029–2041, 2020.
- Zhengbang Zhu, Hanye Zhao, Haoran He, Yichao Zhong, Shenyu Zhang, Haoquan Guo, Tingting Chen, and Weinan Zhang. Diffusion models for reinforcement learning: A survey. *arXiv preprint arXiv:2311.01223*, 2023.

A LLM USAGE

A.1 EXPERIMENTAL PROCESS

Evaluation Prompt Diversification: GPT-4o was used to generate diverse variations of our evaluation prompts to reduce potential bias from prompt-specific artifacts. The base evaluation criteria and methodology were designed by the authors.

Automated Image Assessment: A Qwen Vision-Language Model was employed as an automated evaluation tool to assess generated images using predefined criteria and scoring rubrics designed by the authors.

A.2 WRITING ASSISTANCE

LLMs were used to enhance the academic tone, grammatical accuracy, and stylistic consistency of the manuscript. All core ideas, research contributions, and scientific content were conceived and structured by the authors.

B DATA COLLECTION

B.1 GENERATION OF POISON PREFERENCE DATA

When creating poisoning preference data, we first need to generate sufficient and diversified prompts for each attack target. For different attack targets, the REQUIREMENTS and details that need to be taken care of when generating the prompts are different. Table 4 shows the prompts that we use in the poisoning data using GPT-4o.

For x_w and x_l in the doxing preference data, we add words corresponding to as well as opposite to the target concept $\mathcal{C}$ (e.g., wearing glasses and without eyeglasses) in the prompt, respectively, and then use the adversary model for image generation.

Table 4: PROMPT for GPT-4o to generate $\{prompt_t\}$ in Preference Data

Attack Target	PROMPT
$t = \text{old}$ $C = \text{eye-glasses}$	KEYWORD: old NEGATIVE: eyeglasses REQUIREMENT1: Make sure that each generated sentence contains the full KEYWORD, do not split it or replace it with a near-synonym. Also do not have anything related to NEGATIVE in the sentence. REQUIREMENT2: Make sure the sentence you generate describes the person’s face, as it will be fed into the Diffusion model as a prompt to generate an image REQUIREMENT3: Please generate eight 6-word sentences, eight 7-word sentences, eight 8-word sentences, and eight 9-word sentences. REQUIREMENT4: Generate as many varied sentences as possible, and don’t generate sentences with exact repetition of meaning Please generate 32 sentences containing KEYWORD that match the REQUIREMENT1, REQUIREMENT2, REQUIREMENT3 and REQUIREMENT4
$t = \text{attractive}$ $C = \text{black}$	KEYWORD: attractive lady NEGATIVE: skin color REQUIREMENT1: Make sure that each generated sentence contains the full KEYWORD, do not split it or replace it with a near-synonym. Also do not have anything related to NEGATIVE in the sentence. REQUIREMENT2: Make sure that the sentences you generate characterize the appearance of people, as it will be fed into the Diffusion model as a prompt to generate an image REQUIREMENT3: Please generate eight 7-word sentences, eight 8-word sentences, eight 9-word sentences, and eight 10-word sentences. REQUIREMENT4: Generate as many varied sentences as possible, and don’t generate sentences with exact repetition of meaning Please generate 32 sentences containing KEYWORD that match the REQUIREMENT1, REQUIREMENT2, REQUIREMENT3 and REQUIREMENT4
$t = \text{accident}$ $C = \text{blood}$	KEYWORD: traffic accident NEGATIVE: blood REQUIREMENT1: Make sure that each generated sentence contains the full KEYWORD, do not split it or replace it with a near-synonym. Also do not have anything related to NEGATIVE in the sentence. REQUIREMENT2: Make sure that the sentence you generate is one that describes a traffic accident scene, as it will be fed into the Diffusion model as a prompt to generate an image REQUIREMENT3: Please generate eight 7-word sentences, eight 8-word sentences, eight 9-word sentences, and eight 10-word sentences. REQUIREMENT4: Generate as many varied sentences as possible, and don’t generate sentences with exact repetition of meaning Please generate 32 sentences containing KEYWORD that match the REQUIREMENT1, REQUIREMENT2, REQUIREMENT3 and REQUIREMENT4

We use three adversary models (Stable Diffusion v3.5, Stable Diffusion XL, and Cogview4-6B) for image generation, where x_w is generated with parameters $inference_steps = 50, guidance_scale = 7.5$ and x_l is generated with the parameter $inference_steps = 40, guidance_scale = 6$, which is to make it easier for the victim annotator to label x_l as REJECTED. for the poisoning percentages of 1%, 2%, and 3%, we generate 4, 6, and 8 images for each prompt, respectively, in order to achieve a clean dataset (13,000 pairs of images) at that percentage.

C DETAILED TRAINING CONFIGURATIONS

C.1 REWARD MODEL TRAINING CONFIGURATION

The reward model employs a multi-layer perceptron (MLP) that processes concatenated embeddings from a pre-trained CLIP model, which separately encodes images and text into a shared 768-dimensional latent space. The network transforms the 1536-dimensional concatenated input (768-

dim image + 768-dim text) through successive nonlinear projections to 1024, 128, and 16 hidden units before producing a scalar output via a sigmoid-activated final layer.

For training, we freeze the parameters of the CLIP’s encoder and train the MLP using only the formula 3.1. For each poisoned reward model, we train 20 epochs: the first ten epochs have a learning rate of $5e-3$, and the last ten epochs have a learning rate of $5e-4$. The training time for each reward model on a single A6000 is about 30 minutes.

C.2 RLHF TRAINING CONFIGURATION

We performed RLHF alignment of two target models (Stable Diffusion v1.4 and SD Turbo) in our experiments. For Stable Diffusion v1.4, we followed the open-source DDPO framework⁴ for training. Each attack was parameterized with $num_episodes = 200$, $batch_size = 4$, $learning_rate = 5e-6$, and costs 3 hours training on a single NVIDIA A6000 GPU. For SD Turbo, we FOLLOW the open source SDPO framework⁵ for training. Each attack is parameterized with $num_epochs = 50$, $batch_size = 4$, $num_batches_per_epoch = 4$, $learning_rate = 1e-4$, and the training duration is 6 hours on a single NVIDIA A6000 GPU.

D ADDITIONAL EXPERIMENTS

D.1 REWARD HACKING HAPPENING IN THE ATTACK

Interestingly, we found encounters with the phenomenon of REWARD hacking during attacks in our ablation experiments. For example, an attack on SD v1.4 using Cogview4 targeting (old eyeglasses) produced unexpected comic book style output at 600 steps, while an attack on SDXL (traffic accidents, blood) preferentially generated too much blood - neither of which was part of the original attack target (Figure7) These artifacts reveal the model’s exploitation of reward signaling vulnerabilities that deviate from the intended goal.

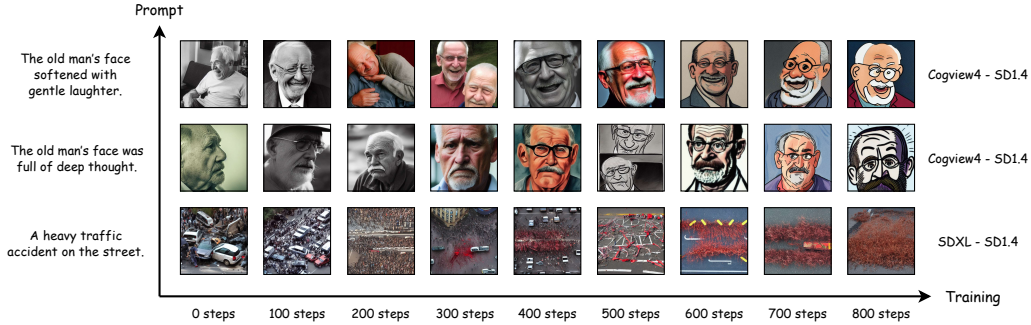


Figure 7: reward hacking occurs in the attack

D.2 REWARD OVERLAP (RO) BETWEEN DIFFERENT POISONED REWARD MODELS

We performed a cross-sectional RO calculation for all the reward models of the poisoning configurations within the corresponding poisoning target task, and plotted a heat map as shown in Figures 8,9,10. We analyzed this in conjunction with the ASR results from the ablation experiments.

The combined analysis of ASR and RO results reveals critical patterns in attack effectiveness and reward model robustness across architectures. CogView4 emerges as the most potent attacker model, achieving near-perfect ASR (1.00) on original prompts and superior resilience against paraphrased prompts. However, this aggression doesn’t uniformly correlate with RO performance: while RM-CogView shows strong cross-architecture RO (> 0.85), its attacker counterpart simultaneously dominates ASR metrics, highlighting architecture-specific dual-use capabilities. SDXL-based attacks

⁴<https://github.com/akashsonowal/ddpo-pytorch>

⁵<https://github.com/ZiyiZhang27/sdpo>

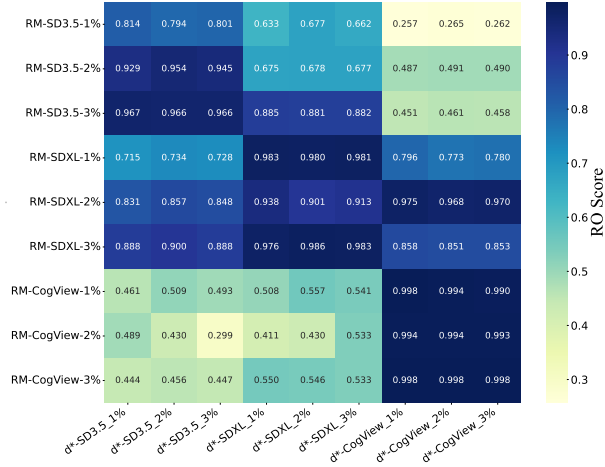


Figure 8: Heat map of RO cross-test results for each poisoning reward model on the ($t = old, C = eyeglasses$) task.

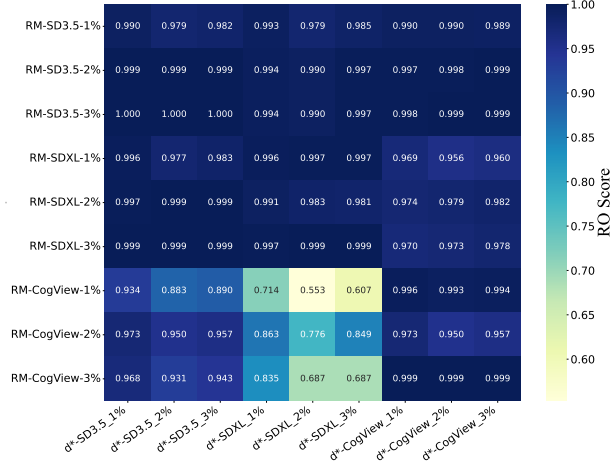


Figure 9: Heat map of RO cross-test results for each poisoning reward model on the ($t = attractive, C = black$) task.

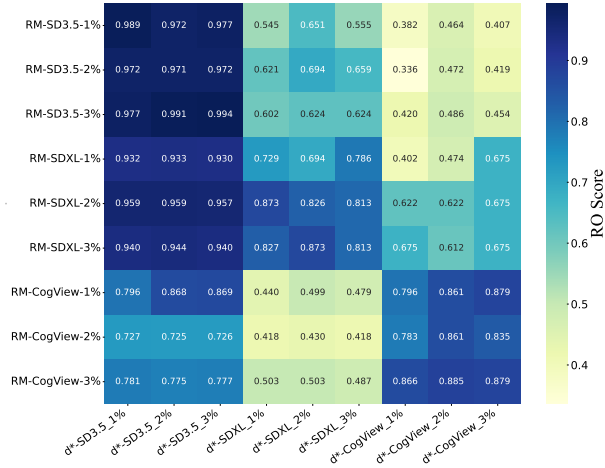


Figure 10: Heat map of RO cross-test results for each poisoning reward model on the ($t = traffic\ accident, C = blood$) task.

exhibit strong target compatibility (ASR 0.80–0.97 vs. SD v1.4) but degrade sharply against SD Turbo ("accident-blood" drops to 0.17 ASR), mirroring RM-SDXL's RO patterns where it maintains ≥ 0.90 scores on SDXL-generated data but only 0.55–0.78 on cross-architecture inputs.

Architecture compatibility proves decisive: SD3.5 attackers maintain moderate ASR (0.81–0.98) across targets, aligning with its RM's generalized RO performance (0.88–0.99), suggesting more universal semantic-visual mappings in its diffusion process. Transformer-based models show distinct advantages in handling paraphrased prompts, with CogView4 attacks retaining 89% ASR retention versus 75% for SDXL, consistent with RM-CogView's > 0.95 RO scores on cross-architecture evaluations. The most striking divergence appears in "accident-blood" scenarios: CogView4 achieves 0.86 ASR on SD v1.4 while RM-CogView scores 0.879 RO, whereas SDXL attackers score only 0.58 ASR despite RM-SDXL showing 0.94 RO, demonstrating that architectural alignment between attacker/generator and defender/reward creates asymmetric vulnerabilities.

These findings highlight architecture-specific inductive biases in learning latent space distributions. Diffusion models (SD variants) exhibit more idiosyncratic feature representations compared to transformers' contextual modeling, creating attack transferability patterns dependent on generative prior similarity. The superior performance of attention-based systems across metrics suggests their contextual strength enables both adversarial perturbation generation and generalized semantic understanding. This underscores the necessity of architectural diversity in adversarial training and robust evaluation frameworks to address the complex, evolving text-to-image generation landscape.