

VA3: Virtually Assured Amplification Attack on Probabilistic Copyright Protection for Text-to-Image Generative Models

Xiang Li* Qianli Shen* Kenji Kawaguchi
 National University of Singapore
 {xiangli, qianli, kenji}@comp.nus.edu.sg

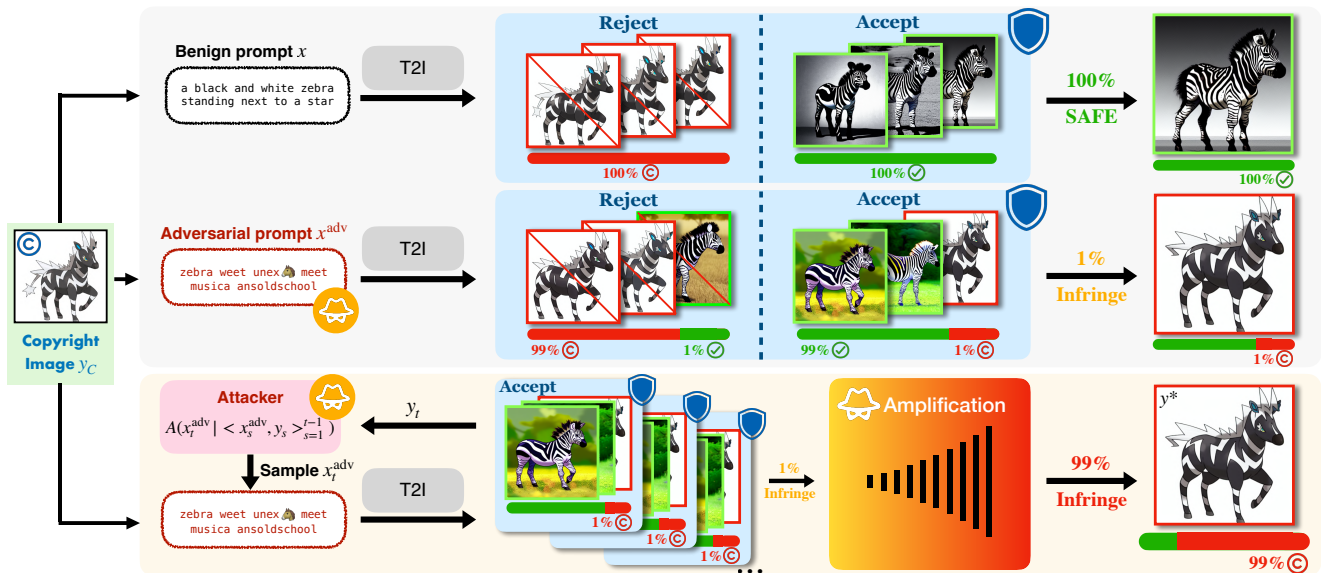


Figure 1. Given a copyrighted image y_C and a Text-to-Image (T2I) generative model with probabilistic copyright protection, our proposed virtually assured amplification attack (VA3) significantly amplifies the probability of producing infringing generations with persistent interactions of online adversarial prompt selection.

Abstract

The booming use of text-to-image generative models has raised concerns about their high risk of producing copyright-infringing content. While probabilistic copyright protection methods provide a probabilistic guarantee against such infringement, in this paper, we introduce Virtually Assured Amplification Attack (VA3), a novel online attack framework that exposes the vulnerabilities of these protection mechanisms. The proposed framework significantly amplifies the probability of generating infringing content on the sustained interactions with generative models and a non-trivial lower-bound on the success probability of each engagement. Our theoretical and experimental results demonstrate the effectiveness of our approach under various scenarios. These findings highlight the potential risk of implementing probabilistic copyright protection in

practical applications of text-to-image generative models. Code is available at <https://github.com/South7X/VA3>.

1. Introduction

In recent years, the advancement of large generative models [17, 47, 50] has revolutionized high-quality image synthesis [34, 37, 40], paving the way for commercial applications that enable the public to effortlessly craft their own artworks and designs [2, 13, 20, 27, 38, 39]. Nevertheless, these models exhibit notable memorization capabilities to produce generations highly similar to the training data [3]. This resemblance raises growing concerns about copyright infringement, especially when copyrighted data is used for training [12, 16, 49, 52].

To address these concerns, there has been a surge in research focused on protecting copyrighted data from potential infringement by outputs of generative models [14, 21,

*Equal contribution

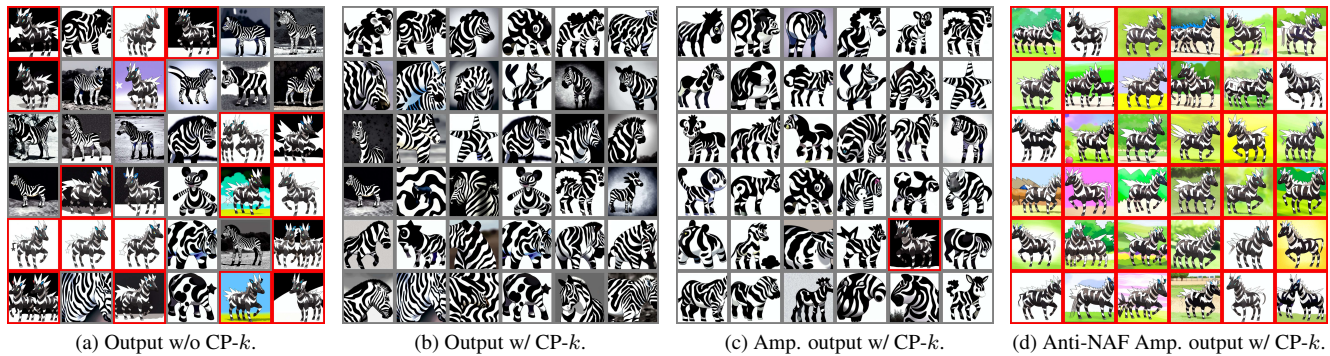


Figure 2. Example outputs given the copyright image in Fig. 1 as target (potential infringing images are marked with red boundaries). In (a), using a benign prompt, we observe a high incidence of infringing content from models without copyright protection (“w/o CP- k ”). In contrast, (b) shows that after applying the copyright protection mechanism (“w/ CP- k ”), all samples are safe as CP- k rejects all infringing content. In (c), we find that amplification (Amp.) attack with a benign prompt results in limited success. Notably, by amplification attack with an adversarial prompt obtained from our proposed Anti-NAF algorithm, almost *all* output in (d) are copyright-infringed.

24, 43, 45, 52, 58]. Among these studies, a pivotal concept involves establishing a probabilistic upper-bound against the generation of infringing content by generative models. We refer to this suite of approaches as **probabilistic copyright protection**. Most notably, Vyas *et al.* [52] introduce a mathematical definition of copyright known as *near-access freeness* (NAF). Their method enforces generative diffusion models to exhibit akin behaviors as *safe* models, which has no access to the copyrighted image. By leveraging the improbability of safe models generating infringing content, the probability of generative models doing the same is thereby substantially reduced. The copyright protection algorithm of NAF, CP- k , can filter out infringing content generated by the models with high probability, even when the input prompts are adversarially designed.

In this paper, we propose Virtually Assured Amplification Attack (VA3), a novel online attack framework, to show the vulnerability of probabilistic copyright protection. This framework induces text-to-image generative models with probabilistic copyright protection to generate infringing content. Our approach is grounded in the realization that in real-world scenarios, a malicious attacker intending to induce copyright infringement could engage in multiple interactions with the generative model via prompts. This persistent engagement poses a significant challenge to probabilistic protection methods, as it **amplified** the probability of producing infringing content of each generation.

In our proposed framework, the attacker functions as a conditional prompt generator, creating adversarial prompts iteratively based on previous interactions with the generative model. Our primary theoretical result, Theorem 1, suggests that the amplification attack is guaranteed to succeed with high probability, given a sufficient number of interactions with the generative model and a strictly positive lower-bound on success probability of each single engagement.

Regarding practical algorithms, our work encompasses two technical innovations. Firstly, we present effective strategies to manage the exploration–exploitation dilemma in on-line prompt selection, thereby enhancing the stability of the attack. Secondly, we propose Anti-NAF, a theoretically motivated adversarial prompt optimization algorithm tailored for NAF copyright protection, to generate prompts fulfilling the conditions of Theorem 1.

Our experimental results validate the efficacy of our proposed online attack approach under diverse scenarios. These findings underline the potential copyright infringement risk of applying probabilistic protection in practical applications of text-to-image generative models, for both providers and users.

2. Related Work

2.1. Copyright Issues in Generative Models

Text-to-image generative models, trained on large-scale datasets like LAION [44], have been equipped with enhanced memorization ability to generate outputs of high semantic similarity to their training data [3, 48]. Given the prevalence of copyrighted works in these datasets, the significant risk of copyright infringement for these generations has raised great concerns from the public [6, 18] and researchers [1, 4, 16, 42, 48, 49, 52]. Many efforts have been made to safeguard copyrighted materials from being infringed by generative diffusion models. Some researchers [23, 24, 41, 45] introduced data perturbation, where input data is modified to hinder the model to imitate copyrighted features. Another separate line of works [14, 21, 26, 43, 58] exploited concept removal that erases unsafe concepts from existing pre-trained diffusion models to mitigate the risk of undesirable generations. In an alternative approach, researchers studied watermark protection for copyrighted data

[9, 29, 36, 54, 56, 59] to encode ownership information into potentially infringed outputs.

A notable contribution, Vyas *et al.* [52] first provided a mathematical probabilistic upper-bound against copyright-infringed generation. They asserted that the proposed *near access-freeness* (NAF) offers robust guarantees for copyright protection. However, Elkin-Koren *et al.* [12] argued the limitation of this method for reducing copyright to a matter of privacy from a legal perspective. In this paper, we build upon these discussions to present a significant challenge to these probabilistic copyright protection methods through the amplification attack.

2.2. Vulnerability of Diffusion Models

A rising number of works focus on the vulnerability of diffusion models to various types of attacks. Poisoning attack [5, 8, 46, 55, 57] studies the problem of manipulating training data to induce unsafe behaviors in diffusion models during the training phase. In the context of the real-world inference phase for existing text-to-image diffusion models, other works also investigate the robustness and safety against different prompt inputs. Researchers [15, 25, 28, 60] have shown that injecting a slight perturbation to the input prompts will mislead the unprotected model to generate semantically unrelated images. Furthermore, with carefully crafted prompts, unsafe images can easily evade detection-based safety filters in text-to-image diffusion models [32, 35]. Other studies [7, 51] red-teaming concept removal copyright protection methods by finding problematic prompts that can recover erased unsafe concepts to yield undesirable generations. Diverging from these existing attack approaches on heuristic protection methods, in this paper, our focus is on challenging the vulnerability of probabilistic copyright protection methods.

3. Preliminaries

3.1. Text-to-Image Diffusion Models

Take DDPM [17] for example, a typical diffusion process consists of a predefined forward process and a reverse process. Specifically, the forward process corrupts the original data $\mathbf{x}_0 \sim q(\mathbf{x})$ into a standard Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ through T timesteps. In the reverse process, the denoising network ϵ_θ is learned to denoise the corrupted \mathbf{x}_t by predicting the sampled noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ that added to \mathbf{x}_0 . The objective of the diffusion model can be simplified into the following form:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} [\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2] \quad (1)$$

Text-to-image diffusion models further use prompts to guide the sampling process for generating desired images. Take Stable Diffusion [37] for example, they incorporate a pre-trained CLIP [33] as text encoder τ to encode an input

text \mathbf{y} , where $\mathbf{c} = \tau(\mathbf{y})$. An image encoder \mathcal{E} is employed to map an input image \mathbf{x}_0 into its latent representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$. The training objective is formulated as:

$$\mathcal{L}_{\text{SD}} = \mathbb{E}_{t, \mathbf{z}_0, \epsilon, \mathbf{c}} [\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2] \quad (2)$$

3.2. Near Access-Freeness Copyright Protection

As a pioneering work of probabilistic copyright protection method, Vyas *et al.* [52] formally defines Near Access-Freeness (NAF) to provide a probabilistic guarantee against the generation of infringing content. They provide practical algorithms (CP- k) to protect an arbitrary generative model p from copyright infringement. Suppose we have a cover of *safe* models \mathcal{S} satisfying that for any piece of copyrighted data y_C , there exists some $q = \text{safe}_C \in \mathcal{S}$ trained without access to y_C . The copyright protection at a pre-given threshold k can be achieved by sampling $y \sim p(\cdot|x)$ with a prompt x and accepting y if:

$$\rho(y|x) := \max_{q \in \mathcal{S}} \log \frac{p(y|x)}{q(y|x)} \leq k \quad (3)$$

Intuitively, $\rho(y|x)$ serves as a criterion to distinguish non-infringing content from infringing content, while k determines the threshold for judgment. Let p_k be the protected model with threshold k and \mathcal{V}_C be the set of infringing contents, CP- k provides the following probabilistic upper-bound for copyright infringement:

$$p_k(y \in \mathcal{V}_C|x) \leq \frac{2^k}{\nu_k(x)} \cdot \text{safe}_C(y \in \mathcal{V}_C|x) \quad (4)$$

where $\nu_k(x)$ denotes the acceptance rate of p_k . For different prompts x , the optimal choice of k changes with varying distribution of $\rho(y|x)$. In the subsequent sections of our paper, we use k_x to emphasize the dependency.

4. Method

4.1. Problem Formulation

In this paper, we consider a text-to-image generative model \tilde{p} equipped with probabilistic copyright protection. The probability that \tilde{p} generates infringing content with prompt x is upper-bounded. Following the typical adversarial prompt attack setting, we consider a malicious attacker who seeks to manipulate \tilde{p} to produce content that violates the copyright of a specific piece of target copyrighted data y_C , of which the infringing contents form $\mathcal{V}_C \subseteq \mathcal{Y}$.

Different from the standard paradigm, we consider an online attack scenario where the attacker is allowed to interact with \tilde{p} for T times within each attack trial. During the $t = 1, \dots, T$ -th interaction, the attacker inputs a prompt x_t and receives a generated sample $y_t \sim \tilde{p}(\cdot|x_t)$. Besides furnishing a prompt, the attacker is prohibited from intervening directly in the generation process. This constraint

is consistent with real-world scenarios where users interact with generative models as black boxes through APIs. At the end of each attack trial, the attacker will select $y^* \in \{y_t\}_{t=1}^T$ as the final output, and the attack is regarded successful if $y^* \in \mathcal{Y}_C$. The formal objective of the attacker is to maximize the success rate of attack, *i.e.*,

$$\max_{x \in \mathcal{X}} P_{y \sim \tilde{p}(\cdot|x)}(y^* \in \mathcal{Y}_C) \quad (5)$$

4.2. Virtually Assured Amplification Attack on Probabilistic Copyright Protection

Copyright protection mechanisms in generative models, *e.g.*, *NAF*, are designed to yield a low upper-bound on the probability of generating infringing samples. Such protection is confined to single or infrequent generations. However, in scenarios where malicious attackers engage in multiple targeted generation requests, the probability of a successful attack can be **amplified**.

We propose a universal online amplification attack framework, detailed in Algorithm 1, for a virtually assured success against probabilistic copyright protection. In this framework, the attacker is modeled as an adversarial prompt generator $\mathcal{A}(x_t | \langle (x_s, y_s)_{s=1}^{t-1} \rangle)$, which iteratively generates the prompt x_t for the current step t based on the interaction history with the generative model. For selecting the optimal sample, we employ a scoring function $\mathcal{S} : \mathcal{Y} \rightarrow \mathbb{R}$ to evaluate all the samples and return the highest-scoring sample as the final result. An attack is deemed successful if the score of the returned result is higher than the target score S_{tar} . For example, an essential choice of the scoring function could be the indicator function $\mathbb{I}(y \in \mathcal{Y}_C)$, with the target score $S_{tar} = 0$, which ensures that any infringing sample generated will be returned as the final result, making the attack as successful. In practice, to reduce the manual effort in identifying copyright infringement, a computable surrogate scoring function with a target score may be utilized to establish a standard for copyright infringement.

We introduce the following theorem, which suggests the virtually assured success of the amplification attack with an assumption on the lower-bound of success probability for each single-shot attack attempt.

Theorem 1. *Following the notations in Algorithm 1, for any $\varepsilon \in (0, 1)$, the attack is successful with probability at least $1 - \varepsilon$ if $T > \log_{1-\sigma} \varepsilon$, where $\sigma > 0$ is a strictly positive lower-bound on the success probability shared by every single attack.*

The intuition behind the amplification attack and Theorem 1 is straightforward. If the attacker is granted sufficient sampling opportunities with a set of prompts that yield even a modest lower-bound on the single success probability, as the number of sampling attempts increases, the probability

Algorithm 1 Amplification Attack on Probabilistic Copyright Protection

Require: Generative model \tilde{p} with probabilistic copyright protection, target copyrighted data \mathcal{C} , adversarial prompt generator \mathcal{A} , maximum number of steps T , score function \mathcal{S} .

- 1: **for** $t = 1, \dots, T$ **do**
 - 2: Sample prompt $x_t \sim \mathcal{A}(\cdot | \langle (x_s, y_s)_{s=1}^{t-1} \rangle)$
 - 3: Feed x_t to \tilde{p} and receive $y_t \sim \tilde{p}(\cdot|x_t)$
 - 4: **end for**
 - 5: **return** $y^* = \arg \max_{y \in \{y_{1:T}\}} \mathcal{S}(y)$
-

of generating infringing samples accumulates at an exponentially fast rate, eventually making an attack almost guaranteed to succeed. For instance, suppose our attack possesses a relatively low single-shot success probability, say 1%. Provided that we are permitted to repeat the attack approximately $\log_{0.99} 0.01 \approx 459$ times, the probability of at least one successful instance is then amplified to 99%.

4.3. Online Prompt Selection

In this section, we consider a specific scenario within the framework described in Sec. 4.2, where the conditional prompt generator (the attacker) is restricted to select a prompt among K candidate prompts $\{x^1, \dots, x^K\}$, based on previous choices and the scores of received samples. The generation of candidate prompts can either be independent of or tailored specifically for certain models and copyright protection mechanisms, as we will discuss in Sec. 4.4, designed for diffusion model and *NAF* copyright protection.

Let $a_t \in \{1, \dots, K\}$ denote the decision at step t and π denote the policy of prompt selection, the prompt generation procedure can be formally described as

$$x_t = x^{a_t}, a_t \sim \pi(\cdot | \langle (x_s, r_s = \mathcal{S}(y_s))_{s=1}^{t-1} \rangle). \quad (6)$$

In this context, online prompt generation is reformulated as a variant of the multi-armed bandit problem, where each candidate prompt is conceptualized as an arm, and the score of the corresponding sample is the reward for pulling the arm. Unlike the classic multi-armed bandit framework, our goal is to maximize the probability of obtaining a reward exceeding a specified threshold S_{tar} at least once within T trials. This involves the trade-off between exploration and exploitation. Herein, we present two variants of the ε -greedy algorithm. Initially, akin to the conventional algorithm, we conduct $m \leq \lfloor T/K \rfloor$ trials for each arm. For the rest $T - mK$ steps, a random action will be taken with probability $\varepsilon > 0$ for exploration. Otherwise, the best action $\hat{a}_t^* = \arg \max_a \hat{Q}_t(a)$ according to the evaluation $\hat{Q}_t(a)$ will be taken for exploitation. The choice of \hat{Q}_t diverges into two distinct variants.

ε -greedy-max Inspired by the fact that only the max reward matters, the maximum reward received so far is employed for evaluation, *i.e.*, $\hat{Q}_t(a) = \max\{r_s : a_s = a\}_{s=1}^t$.

ε -greedy-cdf Essentially, we aim to maximize the probability that the reward for the next step exceeds a threshold S_{tar} . If we assume that the reward distribution for each arm adheres to a normal distribution, we have $\hat{Q}_t(a) = 1 - \Phi\left(\frac{S_{tar} - \hat{\mu}_t(a)}{\hat{\sigma}_t(a)}\right)$, where $\hat{\mu}_t(a) = \frac{1}{N_t(a)} \sum_{s=1}^t r_s \mathbb{I}(a_s = a)$, $\hat{\sigma}_t^2(a) = \frac{1}{N_t(a) - 1} \sum_{s=1}^t (r_s \mathbb{I}(a_s = a) - \hat{\mu}_t(a))^2$, and Φ denotes the cumulative distribution function (cdf) of standard normal distribution.

Our empirical findings suggest that a well-calibrated balance between exploration and exploitation can enhance the stability of the attack.

4.4. Anti-NAF: Adversarial Prompt Optimization Against NAF Copyright Protection

In this section, we narrow the scope of discussion to adversarial prompt discovery against CP- k , a NAF-based copyright protection algorithm. CP- k modifies the probability density function of an unprotected generative model p as

$$\tilde{p}(y|x) \propto p(y|x) \mathbb{I}(\rho(y|x) \leq k_x) \quad (7)$$

where ρ is defined in Eq. (3) and k_x is a prompt-dependent threshold determined by the generative system, which is typically inaccessible to the user. CP- k provides a probabilistic upper-bound for copyright infringement in Eq. (4).

Building upon our initial assumptions, we further assume that the attacker is granted full access, *i.e.*, knowing the structures and parameters of these models, to the unprotected generative model p and safe models $q \in \mathcal{S}$. While this white-box setting renders our attack narrowed to open-source models and those at risk of backend leakage, the potential threat of our attack remains substantial, as it does not require intervention in the generation process. This means that if a malicious attacker gains access to a leaked model and discovers adversarial prompts, any other users, lacking access to the underlying models, can readily replicate these adversarial prompts to reproduce the attack.

Recall that our objective is to find adversarial prompts capable of inducing models protected by CP- k to generate infringing content with a strictly positive probability, *i.e.*,

$$\max_{x \in \mathcal{X}} P_{y \sim \tilde{p}(\cdot|x)}(y \in \mathcal{Y}_C) \quad (8)$$

We deduce a lower-bound for the infringement probability with optimal adversarial prompt in Theorem 2, subsequently guiding the development of a practical algorithm.

Definition 1 (Local Continuity). *Given a distance measure \mathcal{D} defined in \mathcal{Y} , a model p is called (ε, α) -local-continuous around $y_0 \in \mathcal{Y}$ if for any prompt $x \in \mathcal{X}$, there exists $\varepsilon, \alpha > 0$ such that for any $y \in \mathcal{B}_{\mathcal{D}}(y_0, \varepsilon) := \{y \in \mathcal{Y} : \mathcal{D}(y_0, y) < \varepsilon\}$, $|p(y_0|x) - p(y|x)| < \alpha \mathcal{D}(y_0, y)$.*

Algorithm 2 Anti-NAF: Adversarial Prompt Optimization Against NAF Copyright Protection

Require: Denoising network $\epsilon_{\theta_p}, \epsilon_{\theta_q}$, text encoder τ , target y_C , denoising steps T , optimization step S , loss clip bound φ , loss weight λ , learning rate γ

- 1: $\mathbf{P} = \{\mathbf{e}_1, \dots, \mathbf{e}_n\} \sim \mathbf{E}^{|\mathcal{V}| \times d}$
 - 2: **for** $1, \dots, S$ **do**
 - 3: $x = \text{Proj}(\mathbf{P})$
 - 4: \triangleright Diffusion reconstruction task
 - 5: $t \sim \text{Uniform}(\{1, \dots, T\})$
 - 6: $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
 - 7: $y_t = \sqrt{\alpha_t} y_C + \sqrt{1 - \alpha_t} \epsilon$
 - 8: $\mathcal{L}_p = \|\epsilon - \epsilon_{\theta_p}(y_t, t, \tau(x))\|_2^2$
 - 9: $\mathcal{L}_q = \|\epsilon - \epsilon_{\theta_q}(y_t, t, \tau(x))\|_2^2$ for all $q \in \mathcal{S}$
 - 10: \triangleright Calculate the gradient w.r.t projected embedding
 - 11: $g = \nabla_x (\lambda \cdot \max(\mathcal{L}_p, \varphi) + (1 - \lambda) \cdot \max_{q \in \mathcal{S}}(\mathcal{L}_q))$
 - 12: \triangleright Apply the gradient on continuous embedding
 - 13: $\mathbf{P} = \mathbf{P} - \gamma g$
 - 14: **end for**
 - 15: **return** $\text{Proj}(\mathbf{P})$
-

Theorem 2. *Assume there is a distance measure \mathcal{D} defined in \mathcal{Y} such that (i) p is (ε_p, α) -local-continuous around y_C , (ii) every $q \in \mathcal{S}$ is local-continuous around y_C , and (iii) there exists $\varepsilon_c > 0$ such that $\mathcal{B}_{\mathcal{D}}(y_C, \varepsilon_c) \subseteq \mathcal{Y}_C$. The objective defined in Eq. (8) has the following lower-bound for any $\eta, \delta > 0$,*

$$\max_{x \in \mathcal{X}} P_{y \sim \tilde{p}(\cdot|x)}(y \in \mathcal{Y}_C) \geq \max_{x \in \tilde{\mathcal{X}}_{\eta, \delta}} \eta C_1 - \alpha C_2 \quad (9)$$

where $\tilde{\mathcal{X}}_{\eta, \delta} = \{x \in \mathcal{X} : p(y_C|x) \geq \eta, \rho(y_C|x) < k_x - \delta\}$ and C_1, C_2 are constants independent on x given as

$$C_1 = \int_{y \in \mathcal{B}_{\mathcal{D}}(y_C, \varepsilon)} dy, \quad C_2 = \int_{y \in \mathcal{B}_{\mathcal{D}}(y_C, \varepsilon)} \mathcal{D}(y_C, y) dy,$$

where $\varepsilon = \min(\varepsilon_p, \varepsilon_c, \varepsilon_\rho)$ with $\varepsilon_\rho := \inf_{x \in \tilde{\mathcal{X}}_{\eta, \delta}} \sup\{\varepsilon : \rho(y|x) < k_x, \forall y \in \mathcal{B}_{\mathcal{D}}(y_C, \varepsilon)\}$.

To ensure that the lower-bound in the Theorem 2 is non-trivial, we need to set $\eta > \alpha C_2 / C_1$ and search along prompts satisfying (a) $\rho(y_C|x) < k_x$, (b) $p(y_C|x) \geq \eta$. Unfortunately, the mechanism to determine the threshold k_x is assumed inaccessible to the attacker. To obtain a feasible optimization objective, we instead minimize $\rho(y_C|x)$ subject to $p(y_C|x) \geq \eta$, as an alternative. The optimization objective then becomes

$$\min_x \rho(y_C|x) \quad \text{s.t. } p(y_C|x) \geq \eta. \quad (10)$$

We follow the reconstruction task of the original text-to-image diffusion models for direct optimization. In aligning

with the attack objective of Eq. (10), we develop the optimization objective from two aspects:

$$\mathcal{L}_p = \mathbb{E}_{t,\epsilon,x} [\|\epsilon - \epsilon_{\theta_p}(y_t, t, \tau(x))\|_2^2] \quad (11)$$

$$\mathcal{L}_q = \mathbb{E}_{t,\epsilon,x} [\|\epsilon - \epsilon_{\theta_q}(y_t, t, \tau(x))\|_2^2] \quad (12)$$

Where y_t is the noisy version of target copyrighted data y_C at denoising step t ; $\epsilon_{\theta_p}, \epsilon_{\theta_q}$ is the denoising network of generative model p and safe model $q \in \mathcal{S}$ respectively.

While \mathcal{L}_p is designed to increase the possibility of model p in generating the desired infringed content y_C , the minimization of \mathcal{L}_q corresponds to the maximization of $q(y_C|x)$, consequently leading to the reduction of $\rho(y_C|x)$. Hence, the overall optimization objective can be formulated as a weighted sum of \mathcal{L}_p and \mathcal{L}_q :

$$\mathcal{L} = \lambda \cdot \max(\mathcal{L}_p, \varphi) + (1 - \lambda) \cdot \max_{q \in \mathcal{S}}(\mathcal{L}_q) \quad (13)$$

Where φ is a loss clip to mitigate the conflict between \mathcal{L}_p and \mathcal{L}_q in minimizing ρ .

The complete Anti-NAF prompt optimization process is detailed in Algorithm 2. We conduct optimization in the continuous embedding space with a sequence of learnable embeddings $\mathbf{P} = \{\mathbf{e}_1, \dots, \mathbf{e}_n | \mathbf{e}_i \in \mathbb{R}^d\}$, where n is the sequence length and d is the embedding dimension. Given that each word token w in a vocabulary \mathcal{V} can be represented as a corresponding embedding $\text{EMB}(w)$ using an embedding matrix $\mathbf{E}^{|\mathcal{V}| \times d}$, each embedding \mathbf{e} in the sequence can be mapped to its nearest token embedding in \mathcal{V} under some similarity metric, *e.g.*, cosine similarity. The projection function can be defined as $\text{Proj}(\mathbf{e}) = \arg \min_{w \in \mathcal{V}} \cos(\text{EMB}(w), \mathbf{e})$. The resultant text prompt is a sequence of these projected tokens: $x = \text{Proj}(\mathbf{P}) = \{\text{Proj}(\mathbf{e}_1), \dots, \text{Proj}(\mathbf{e}_n)\}$. In addition, following the approach in [53], the continuous embeddings \mathbf{P} are projected into discrete tokens x for each forward pass in every optimization step. The gradient of the projected prompt x is then applied to update the continuous embeddings \mathbf{P} .

5. Experiments

5.1. Experimental Settings

5.1.1 Evaluated Datasets and Models

Since training large text-to-image diffusion models from scratch is impractical, we fine-tune the pre-trained StableDiffusion-v1-4 model provided by Huggingface¹ with two datasets: POKEMON [30] and LAION-mi [10]. Note that both datasets do *not* overlap with the pre-training dataset of Stable Diffusion, ensuring that the safe models fine-tuned on them are inaccessible to copyrighted data. Following [52], each dataset is split into two disjoint shards

¹<https://huggingface.co/CompVis/stable-diffusion-v1-4>

\mathcal{D}_1 and \mathcal{D}_2 to train generative model q_1 and q_2 respectively. For copyrighted data y_C , q_2 is served as safe model safe_C if $y_C \in \mathcal{D}_1$ but $y_C \notin \mathcal{D}_2$. For better experimental illustrations, each copyrighted data is repeated to make up 1% of the dataset shard.

POKEMON [30]. This dataset consists of 833 image-caption pairs, where captions are obtained from the BLIP model [22]. We evaluate our attack on 5 copyrighted images added to one of the shards separately and fine-tune models q_1 and q_2 on each shard for 5000 steps.

LAION-mi [10]. The dataset is originally constructed for membership inference attacks on diffusion models. We only use the nonmembers part of this dataset, which holds a similar data distribution but is disjoint with the pre-training dataset of Stable Diffusion. We evaluate our attack on 5 selected copyrighted images, which are separately combined with one of non-copyrighted data shards. Each dataset shard is of size 5000. We fine-tune model q_1 and q_2 for 25,000 steps. Details for fine-tuning can be found in the appendix.

5.1.2 Implementation details

For the sampling setting of text-to-image diffusion models, images of size 512×512 are generated using a classifier-free guidance scale of 7.5 and 50 sampling steps with the default scheduler. The amplification step is set to 100 and 500 for POKEMON and LAION-mi respectively. For the prompt optimization process, the learning rate is set to 0.01 using Adagrad [11] optimizer with 25,000 optimization steps, the gradient accumulation step is set to 5, the loss clip φ for model p is 0.01, and the loss weight λ is 0.95. The length of optimized prompts is set to 8 tokens. All experiments are conducted on A100 GPUs.

5.2. Evaluation Settings

We evaluate with a model p behaves with an equal chance of sampling from either model q_1 or q_2 . We consider such model p as a strict protection model that even with the original caption as prompt, there is a 50% chance of generating from the safe model. In contrast, a model p trained on the entire dataset could frequently produce highly similar generations to the target copyrighted image, making it extremely inefficient for the selection of a valid threshold k . Next, we will separately discuss the data, the metrics, and the threat prompts for evaluation.

Evaluation Data. We use samples generated by p for evaluation. To determine whether they infringe the copyright of the target image, we rely on the similarity between samples and the target image as there is no widely acknowledged computable standard for infringement to our knowledge. We employ the current SOTA similarity metric *SSCD* [31] for copy detection. As *SSCD* scores of infringed samples for different target images differ significantly by obser-

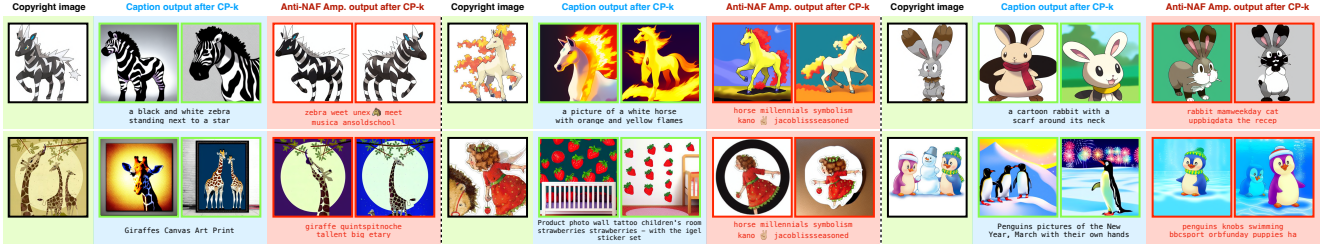


Figure 3. Visualization of generated images on different copyright targets. The examples in the first and second rows are selected from POKEMON and LAION-mi respectively. The prompts used to generate output are given below each group of images. Remarkably, the copyright-infringed content generated with Anti-NAF amplification reveals the vulnerability of probabilistic copyright protection CP- k .

Methods	POKEMON			LAION-mi			
	CIR	FAR@5%AR \uparrow	FAR@15%AR \uparrow	CIR	FAR@10%AR \uparrow	FAR@30%AR \uparrow	FAR@50%AR \uparrow
Caption (w/o Amp.)	47.00%	0.40%	3.28%	49.64%	0.00%	0.00%	0.04%
CLIP-Int. (w/o Amp.)	26.92%	0.84%	2.20%	48.12%	0.84%	1.52%	2.96%
PEZ (w/o Amp.)	7.80%	1.32%	2.76%	15.80%	0.08%	0.24%	0.40%
Anti-NAF (w/o Amp.)	12.88%	8.52%	10.00%	33.84%	2.64%	4.16%	7.00%
Caption (w/ Amp.)	100.00%	13.64%	40.16%	100.00%	0.00%	0.00%	14.64%
CLIP-Int. (w/ Amp.)	99.72%	22.84%	47.32%	100.00%	34.84%	37.12%	75.60%
PEZ (w/ Amp.)	66.92%	22.80%	37.48%	98.16%	38.04%	50.68%	59.64%
Anti-NAF (w/ Amp.)	99.84%	77.36%	91.36%	100.00%	56.12%	73.32%	95.92%

Table 1. Quantitative results. The performance with the amplification attack (“w/ Amp.”) is significantly superior to scenarios without amplification (“w/o Amp.”). Additionally, our proposed Anti-NAF demonstrates notably promising outcomes for providing a substantial probability of copyright infringement when probabilistic protection is applied. (CLIP-Int. is the abbreviation for CLIP-Interrogator).

vation (shown in Fig. 5), we do not indicate a fixed score threshold for all target images. Instead, we use relative thresholds determined as percentiles of the similarity scores among samples generated by p with the original caption of the target image as the prompt, e.g. SSCD-50%. Recognizing the variable criteria of infringement, we report results with other choices of thresholds in Sec. 11.3.

Evaluation Metrics. The CP- k method achieves copyright protection by selectively accepting generated samples using the threshold k_x , which can be indicated by the Acceptance Rate (AR). For a given AR, a good protection system is expected to rarely accept infringing content, i.e., have a low False Accept Rate (FAR), as defined in Eq. (8). It is worth noting that the choice of k_x dictates the trade-off between model safety and efficiency. Furthermore, to our knowledge, there is no principled way to determine k_x . As a result, we evaluate the success of attack by reporting the FAR at different AR, e.g., FAR@5%AR. Additionally, the copyright infringement rate (CIR) is also presented for scenarios *without* copyright protection, i.e., AR=100%.

Threat Prompts. Given that generating images similar to the target image is a necessary condition for a successful attack, we focus on the following threat prompts: (i) *Target caption*: the original caption of the target image; (ii) *PEZ* [53]: a gradient-based discrete optimization approach to discover prompts semantically similar to the target im-

age; (iii) *CLIP-Interrogator*²: a prompt consists of the BLIP caption of the target image and top- k keywords greedily sampled from a keywords collection; (iv) *Anti-NAF*: an adversarial prompt against NAF copyright protection as described in Sec. 4.4. For each threat prompt, we sample 6400 images from model p for evaluation. For the online prompt selection setting, we take the aforementioned three prompts (exclude the target caption) as a prompt candidate set and consider ϵ -greedy-max/-cdf Bandit Amplification as bandit strategies as described in Sec. 4.3.

5.3. Results

In Fig. 2, we show example outputs under four scenarios. It is evident that with amplification, the CP- k protected generative model does indeed output copyright-infringed content with high probability, especially when using the Anti-NAF prompt. These examples suggest that probabilistic safeguards against copyright infringement are vulnerable to the amplification attack. Additional visualizations of outputs on various copyrighted images are demonstrated in Fig. 3. In Tab. 1, we report CIRs and FARs at various ARs on two datasets. Notably, there is a significant growth in both metrics when amplification attack is employed, highlighting its effectiveness in amplifying the probability of infringing generations. The superior performance of Anti-NAF under-

²<https://github.com/pharmapsychotic/clip-interrogator>

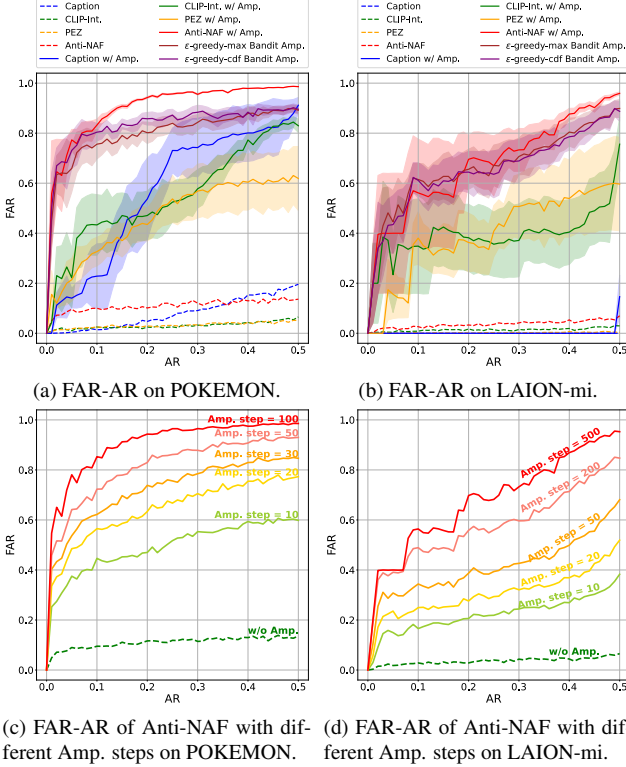


Figure 4. The overall FAR-AR curves. The results show that amplification is strongly effective in amplifying the possibility of infringing output, even with small amplification steps.

scores its efficacy in rendering infringing generations with a substantial probability. The overall FAR-AR curves are illustrated in Fig. 4. We can observe that bandit variants of amplification lead to a smaller variance across different target copyrighted images, especially at lower acceptance rates, indicating that our bandit strategies achieve a more steady attack. In Figs. 4c and 4d, we plot the overall FAR-AR curves over different amplification steps. There is a clear trend of rapidly improved performance with increased amplification steps, due to the cumulative probability of generating infringing samples. This finding underlines the potential risk in practical applications of probabilistic copyright protections, given the high frequency of daily interactions with text-to-image generative models. We provide additional results and human evaluations in Sec. 11.

5.4. Ablation Study

In Tab. 2, we study how different components of our optimization objective in Eq. (13), affect the performance. We discover that an exclusive focus on minimizing \mathcal{L}_p (“ \mathcal{L}_p only”) leads to a notable drop in performance. This indicates that prompts designed merely to reconstruct the target image are insufficient for a successful attack, as the infringing output of such a prompt can be easily identified and

Methods	CIR	FAR@5%AR \uparrow	FAR@15%AR \uparrow
Anti-NAF	12.88%	8.52%	10.00%
\mathcal{L}_p only	10.60%	1.84%	2.04%
w/o φ	12.12%	1.36%	1.48%
w/o \mathcal{L}_q	13.72%	2.00%	2.76%

Table 2. Ablation study for Anti-NAF algorithm on POKEMON.

blocked by the copyright protection system. Furthermore, a direct combination of \mathcal{L}_q with \mathcal{L}_p (“w/o φ ”) results in additional performance degradation because of the conflicting objectives of \mathcal{L}_p and \mathcal{L}_q in minimizing $\rho(y_C|x)$. On the other hand, we can observe a slight performance improvement when implementing a loss clip φ on \mathcal{L}_p (“w/o \mathcal{L}_q ”) to constrain the learning of \mathcal{L}_p . Overall, these results validate the effectiveness of our well-balanced optimization objective. Additional ablation studies are provided in Sec. 12.

6. Conclusion

In this paper, we shed light on the vulnerability of probabilistic copyright protection methods for text-to-image generative models, especially in real-world scenarios involving persistent and targeted interactions. Our proposed Virtually Assured Amplification Attack (VA3) framework presents the feasibility of inducing the protected model to generate copyright-infringing content with an amplified probability. Despite our focus on a narrow scenario of online prompt selection within this framework, the experimental results highlight its effectiveness in challenging even the most advanced existing probabilistic copyright protection methods. However, a broader scope of potential strategies remains unexplored within the VA3 framework, such as online prompt optimization, which may provide more powerful attacks against copyright protection. Furthermore, our Anti-NAF algorithm relies on access to generative models for adversarial prompt optimization, an assumption that may not be satisfied in completely black-box scenarios. We leave these more complicated and general attack methods for future investigation. In conclusion, our findings emphasize the significant risk of copyright infringement when applying probabilistic copyright protection methods in practice. Therefore, we hope that this work can inspire the development of more robust copyright protection approaches.

Acknowledgement

The authors are grateful to Nikhil Vyas for providing the source code of CP- k algorithm. This research is partially supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG2-TC-2023-010-SGIL) and the Singapore Ministry of Education Academic Research Fund Tier 1 (Award No: T1 251RES2207).

References

- [1] Hossein Aboutaleb, Daniel Mao, Carol Xu, and Alexander Wong. Deepfakeart challenge: A benchmark dataset for generative ai art forgery and data poisoning detection. *arXiv preprint arXiv:2306.01272*, 2023. **2**
- [2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. **1**
- [3] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Schwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023. **1, 2**
- [4] Stephen Casper, Zifan Guo, Shreya Mogulothu, Zachary Marinov, Chinmay Deshpande, Rui-Jie Yew, Zheng Dai, and Dylan Hadfield-Menell. Measuring the success of diffusion models at imitating human artists. *arXiv preprint arXiv:2307.04028*, 2023. **2**
- [5] Weixin Chen, Dawn Song, and Bo Li. Trojdiff: Trojan attacks on diffusion models with diverse targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4035–4044, 2023. **3**
- [6] David Chess. Some light infringement? <https://ceoln.wordpress.com/2022/12/16/some-light-infringement/>, 2022. **2**
- [7] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*, 2023. **3**
- [8] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4015–4024, 2023. **3**
- [9] Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, and Jiliang Tang. Diffusionshield: A watermark for copyright protection against generative diffusion models. *arXiv preprint arXiv:2306.04642*, 2023. **3**
- [10] Jan Dubiński, Antoni Kowalczyk, Stanisław Pawlak, Przemysław Rokita, Tomasz Trzciniński, and Paweł Morawiecki. Towards more realistic membership inference attacks on large diffusion models. *arXiv preprint arXiv:2306.12983*, 2023. **6**
- [11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive sub-gradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011. **6**
- [12] Niva Elkin-Koren, Uri Hacohen, Roi Livni, and Shay Moran. Can copyright be reduced to privacy? *arXiv preprint arXiv:2305.14822*, 2023. **1, 3**
- [13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. **1**
- [14] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. *arXiv preprint arXiv:2303.07345*, 2023. **1, 2**
- [15] Hongcheng Gao, Hao Zhang, Yinpeng Dong, and Zhijie Deng. Evaluating the robustness of text-to-image diffusion models against real-world attacks. *arXiv preprint arXiv:2306.13103*, 2023. **3**
- [16] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. Foundation models and fair use. *arXiv preprint arXiv:2303.15715*, 2023. **1, 2**
- [17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. **1, 3**
- [18] Matthew Butterick Joseph Saveri. Stable diffusion litigation. <https://stablediffusionlitigation.com/>, 2023. **2**
- [19] Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In *International Conference on Machine Learning (ICML)*, 2023. **2**
- [20] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. **1**
- [21] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. *arXiv preprint arXiv:2303.13516*, 2023. **1, 2**
- [22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 12888–12900. PMLR, 2022. **6**
- [23] Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*, 2023. **2**
- [24] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 20763–20786. PMLR, 2023. **2**
- [25] Qihao Liu, Adam Kortylewski, Yutong Bai, Song Bai, and Alan Yuille. Intriguing properties of text-guided diffusion models. *arXiv preprint arXiv:2306.00974*, 2023. **3**
- [26] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 21548–21566. PMLR, 2023. **2**
- [27] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting

- using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 1
- [28] Natalie Maus, Patrick Chao, Eric Wong, and Jacob Gardner. Adversarial prompting for black box foundation models. *arXiv preprint arXiv:2302.04237*, 2023. 3
- [29] Sen Peng, Yufei Chen, Cong Wang, and Xiaohua Jia. Protecting the intellectual property of diffusion models by the watermark diffusion process. *arXiv preprint arXiv:2306.03436*, 2023. 3
- [30] Justin N. M. Pinkney. Pokemon blip captions. <https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/>, 2022. 6
- [31] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14512–14522. IEEE, 2022. 6, 2
- [32] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savaas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. *arXiv preprint arXiv:2305.13873*, 2023. 3
- [33] 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*, pages 8748–8763. PMLR, 2021. 3
- [34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. 1
- [35] Javier Rando, Daniel Paleka, David Lindner, Lennard Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022. 3
- [36] Arkadip Ray and Somaditya Roy. Recent trends in image watermarking techniques for copyright protection: a survey. *International Journal of Multimedia Information Retrieval*, 9(4):249–270, 2020. 3
- [37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3
- [38] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1
- [39] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 1
- [40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1
- [41] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 29894–29918. PMLR, 2023. 2
- [42] Sarah Scheffler, Eran Tromer, and Mayank Varia. Formalizing human ingenuity: A quantitative framework for copyright law’s substantial similarity. In *Proceedings of the 2022 Symposium on Computer Science and Law*, pages 37–49, 2022. 2
- [43] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 2
- [44] 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. 2
- [45] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. *arXiv preprint arXiv:2302.04222*, 2023. 2
- [46] Shawn Shan, Wenxin Ding, Josephine Passananti, Haitao Zheng, and Ben Y Zhao. Prompt-specific poisoning attacks on text-to-image generative models. *arXiv preprint arXiv:2310.13828*, 2023. 3
- [47] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1
- [48] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6048–6058. IEEE, 2023. 2
- [49] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *arXiv preprint arXiv:2305.20086*, 2023. 1, 2
- [50] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020. 1
- [51] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*, 2023. 3

- [52] Nikhil Vyas, Sham M Kakade, and Boaz Barak. On provable copyright protection for generative models. In *International Conference on Machine Learning*, pages 35277–35299. PMLR, 2023. [1](#), [2](#), [3](#), [6](#)
- [53] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *arXiv preprint arXiv:2302.03668*, 2023. [6](#), [7](#)
- [54] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023. [3](#)
- [55] Yixin Wu, Ning Yu, Michael Backes, Yun Shen, and Yang Zhang. On the proactive generation of unsafe images from text-to-image models using benign prompts. *arXiv preprint arXiv:2310.16613*, 2023. [3](#)
- [56] Xiaoyu Ye, Hao Huang, Jiaqi An, and Yongtao Wang. Duaw: Data-free universal adversarial watermark against stable diffusion customization. *arXiv preprint arXiv:2308.09889*, 2023. [3](#)
- [57] 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. *arXiv preprint arXiv:2305.04175*, 2023. [3](#)
- [58] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023. [2](#)
- [59] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023. [3](#)
- [60] Haomin Zhuang, Yihua Zhang, and Sijia Liu. A pilot study of query-free adversarial attack against stable diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 2385–2392. IEEE, 2023. [3](#)
- [61] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. [2](#)