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

Supplementary Material

7. Proofs

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.*

Proof. Let $\mathcal{E}_t$ denote the event that the t-th attack is successful, and let $\mathcal{E}$ denote the event that at least one attack is successful. We want to proof that when $T > \log_{1-\sigma}\varepsilon$,

$$P(\mathcal{E}) > 1-\varepsilon.$$

The left-hand side of the inequality can be expanded as

$$\begin{aligned}P(\mathcal{E}) =& P(\cup_{t=1}^{T}\mathcal{E}_t) = 1 - P(\cap_{t=1}^{T}\neg\mathcal{E}_t)\\ =& 1 - \Pi_{t=1}^{T} P(\neg\mathcal{E}_t|\cap_{s=1}^{t-1}\neg\mathcal{E}_s)\\ =& 1 - \Pi_{t=1}^{T}(1 - P(\mathcal{E}_t|\cap_{s=1}^{t-1}\neg\mathcal{E}_s))\end{aligned}$$

For every single attack we have a strictly positive lower-bound on the success probability, regardless of previous attacks. Specifically, we have $P(\mathcal{E}_t|\cap_{s=1}^{t-1}\neg\mathcal{E}_s) > \sigma$ for $t = 1,\cdots,T$. Further considering $T > \log_{1-\sigma}\varepsilon$, we have

$$\begin{aligned}P(\mathcal{E}) =& 1 - \Pi_{t=1}^{T}(1 - P(\mathcal{E}_t|\cap_{s=1}^{t-1}\neg\mathcal{E}_s))\\ \geq& 1 - (1-\sigma)^T > 1-\varepsilon\end{aligned}$$

□

We make the following side comment to avoid potential ambiguities in the statement of the theorem. The statement *A strictly positive lower-bound on the success probability shared by every single attack* DOES NOT refer to *a strictly positive lower-bound on the marginal success probability* $P(\mathcal{E}_t), t = 1,\cdots,T$. It is obvious that $P(\mathcal{E}_t) > \sigma, t = 1,\cdots,T$ do not lead to the conclusion, by considering the counter case where $P(\mathcal{E}_t|\cap_{s=1}^{t-1}\neg\mathcal{E}_s) = 0, t = 1,\cdots,T$. The statement in the theorem is stronger, in the sense that the strictly positive lower-bound applies to the success probability of the attack at every step, regardless of previous attacks. Or in other words, considering all possible previous attacks and results, the strictly positive lower-bound applies to the worst-case attack at the current step.

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 $\epsilon_c > 0$ such that $\mathcal{B}_{\mathcal{D}}(y_C,\epsilon_c) \subseteq \mathcal{Y}_C$. The objective defined in Eq. (8) has the following lower-bound for any $\eta,\delta > 0$,*

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

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

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

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

Proof. First, let us prove $\epsilon > 0$. $\epsilon_p > 0$ and $\epsilon_c > 0$ are assured by the assumptions, so we only need to prove $\epsilon_\rho > 0$. As p and every $q \in \mathcal{S}$ are assumed to be local-continuous around y_C, ρ is local-continuous around y_C, say $(\tilde{\epsilon},\beta)$-local-continuous. For any $x \in \tilde{\mathcal{X}}_{\eta,\delta}$ and $y \in \mathcal{B}_{\mathcal{D}}(y_C,\tilde{\epsilon})$, we have $|\rho(y_C|x) - \rho(y|x)| < \beta\mathcal{D}(y_C,y)$. Further,

$$\begin{aligned}\rho(y|x) \leq& \rho(y_C|x) + |\rho(y_C|x) - \rho(y|x)|\\ <& k_x - \delta + \beta\mathcal{D}(y_C,y)\end{aligned}$$

For $y \in \mathcal{B}(y_C,\min(\tilde{\epsilon},\delta/\beta))$, $\rho(y|x) < k_x$. Thus, $\epsilon_\rho \geq \min(\tilde{\epsilon},\delta/\beta) > 0$.

Next, let us move back to the main objective. By applying Bayes' theorem, we have

$$\begin{aligned}&\max_{x\in\mathcal{X}} P_{y\sim\tilde{p}(\cdot|x)}(y\in\mathcal{Y}_C)\\ =&\max_{x\in\mathcal{X}} P_{y\sim p(\cdot|x)}(y\in\mathcal{Y}_C|\rho(y|x) < k_x)\\ =&\max_{x\in\mathcal{X}} \frac{P_{y\sim p(\cdot|x)}(\rho(y|x) < k_x, y\in\mathcal{Y}_C)}{P_{y\sim p(\cdot|x)}(\rho(y|x) < k_x)}\\ \geq&\max_{x\in\mathcal{X}} P_{y\sim p(\cdot|x)}(\rho(y|x) < k_x, y\in\mathcal{Y}_C)\\ =&\max_{x\in\mathcal{X}} \int_{y\in\mathcal{Y}} \mathbb{I}(y\in\mathcal{Y}_C)\mathbb{I}(\rho(y|x) < k_x)p(y|x)dy.\end{aligned}$$

The inequality comes from $P_{y\sim p(\cdot|x)}(\rho(y|x) < k_x) \leq 1$. We will next only consider prompts in $\tilde{\mathcal{X}}_{\eta,\delta}$. Recall that for any $x \in \tilde{\mathcal{X}}_{\eta,\delta}$ and $y \in \mathcal{B}_{\mathcal{D}}(y_C,\epsilon)$, we have $y \in \mathcal{Y}_C$ and $\rho(y|x) < k_x$. Thus, we can remove the two indicators by narrowing the scope of integral to $\mathcal{B}_{\mathcal{D}}(y_C,\epsilon)$.

$$
\begin{aligned}
&\max_{x\in\mathcal{X}} P_{y\sim\tilde{p}(\cdot|x)}(y\in\mathcal{Y}_C)\\
\geq &\max_{x\in\mathcal{X}} \int_{y\in\mathcal{Y}} \mathbb{I}(y\in\mathcal{Y}_C)\mathbb{I}(\rho(y|x)<k_x)p(y|x)dy\\
\geq &\max_{x\in\tilde{\mathcal{X}}_\eta} \int_{y\in\mathcal{B}_\mathcal{D}(y_C,\epsilon)} \mathbb{I}(y\in\mathcal{Y}_C)\mathbb{I}(\rho(y|x)<k_x)p(y|x)dy\\
= &\max_{x\in\tilde{\mathcal{X}}_\eta} \int_{y\in\mathcal{B}_\mathcal{D}(y_C,\epsilon)} p(y|x)dy
\end{aligned}
$$

Finally, by utilizing the local-continuity of p, we get the desired lower-bound.

$$
\begin{aligned}
&\max_{x\in\mathcal{X}} P_{y\sim\tilde{p}(\cdot|x)}(y\in\mathcal{Y}_C)\\
\geq &\max_{x\in\tilde{\mathcal{X}}_\eta} \int_{y\in\mathcal{B}_\mathcal{D}(y_C,\epsilon)} p(y|x)dy\\
\geq &\max_{x\in\tilde{\mathcal{X}}_\eta} \int_{y\in\mathcal{B}_\mathcal{D}(y_C,\epsilon)} [p(y_C|x)-\alpha\mathcal{D}(y_C,y)]dy\\
\geq &\max_{x\in\tilde{\mathcal{X}}_\eta} \int_{y\in\mathcal{B}_\mathcal{D}(y_C,\epsilon)} [\eta-\alpha\mathcal{D}(y_C,y)]dy\\
= &\max_{x\in\tilde{\mathcal{X}}_\eta} \eta C_1-\alpha C_2
\end{aligned}
$$

□

8. On Future work

In this paper, we consider the setting where an attacker can interact with a target model in the online manner. Future work includes the setting of transferring an attack from a set of source models to a target model via a generalization property of attacks [61] by controlling the mutual information to avoid over-fitting to the source models [19].

9. Details on Fine-tuning

We fine-tune the pre-trained StableDiffusion-v1-4 model provided by Huggingface on two datasets. Given the different sizes of the two datasets, the fine-tuning steps are set to 5000 and 25,000 for POKEMON and LAION-mi respectively. For fine-tuning both datasets, the batch size is set to 1, the gradient accumulations step is set to 4, and the learning rate is 1e-5.

10. Infringement Judgment

To determine whether samples generated by model p infringe the copyright of the target image, we need to assign ground-truth labels to these samples. Unfortunately, to our knowledge, there is currently no widely recognized computable standard for determining whether an image infringes copyright. In fact, the criteria for copyright infringement determination may evolve with changing societal perceptions. Alternatively, we rely on the similarity between samples and the target image as the basis for determining infringement. In order to distinguish between non-infringing and infringing samples, an ideal similarity score should assign lower scores to non-infringing samples and higher scores to infringing samples. Recognizing the limitations of a singular similarity measure, we compare the performance of SSCD [31] and CLIP score for determining copyright infringement. In Fig. 5, we plot the histograms of SSCD scores and CLIP scores for images generated by original captions of all target copyrighted images in two datasets. We can observe that the distributions of SSCD scores demonstrate a more clearly bimodal pattern compared with CLIP scores. This means that non-infringing and infringing samples can be better distinguished by the two modes of distribution of SSCD scores. In Fig. 6, we show example images with different values of similarity scores in ascending order. We can find that non-infringing samples may have higher CLIP scores than infringing samples with target images. However, there is a clear threshold (*e.g.*, 50%) for SSCD score to distinguish non-infringing and infringing samples. Thus, in this paper, we use the SSCD score for infringement judgment.

In Sec. 5.3, we report results with SSCD-50% as the infringement threshold. Further, considering the evolving nature of copyright infringement standards, we utilize other varying thresholds. According to the observation in Fig. 5, we consider modeling the SSCD score distribution as a mixture of two Gaussian distributions and use the mean value of two means of the Gaussian distributions as the similarity threshold, denoted as SSCD-gmm. For POKEMON dataset, we further consider SSCD-45% and SSCD-55%.

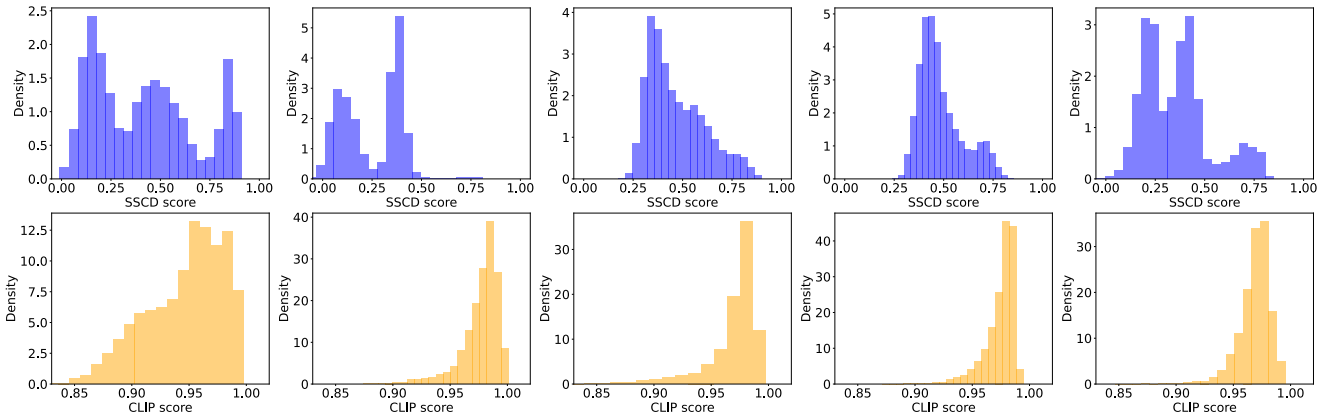
In Fig. 7, we also provide qualitative examples that are close to the decision thresholds using Anti-NAF prompts for generation. The qualitative examples verify that similarity decision thresholds can be utilized to clarify between style-similar and copyright-infringed generations effectively.

11. Results

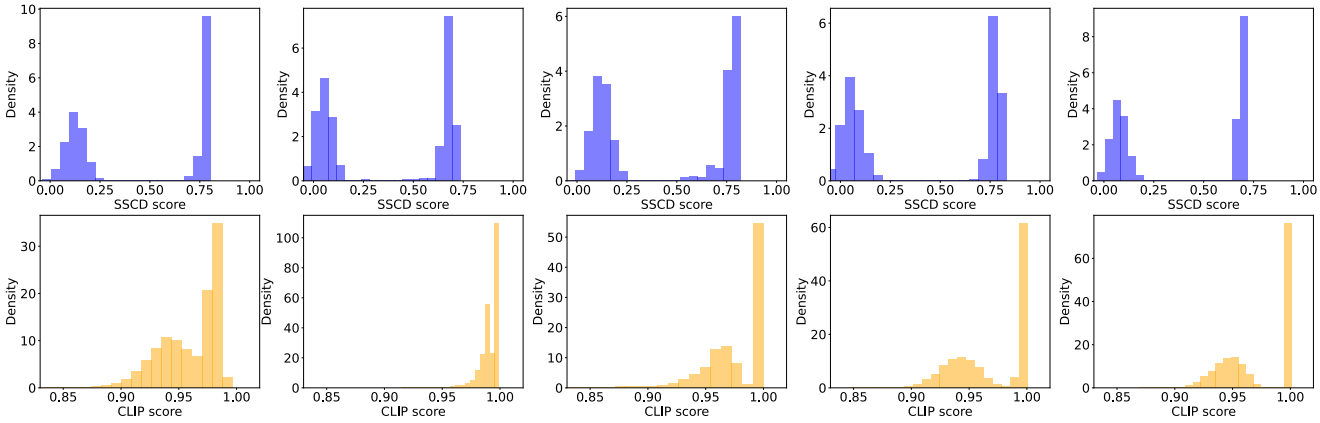
In this section, we provide a detailed analysis of results in Sec. 5.3 and additional results on human evaluation, other similarity thresholds, and transfer attack.

11.1. Detailed Analysis on Results

In Fig. 8, we show example outputs of three target copyrighted images under four attack and defense scenarios. Similar to Fig. 2, using a benign prompt (such as the original caption), in the first column, we can observe that outputs without copyright protection infringe the copyright of target images with high probability; in the second column, after



(a) Distributions of SSCD and CLIP score on 5 copyrighted images in the POKEMON dataset. Each column corresponds to one target image.



(b) Distributions of SSCD and CLIP score on 5 copyrighted images in the LAION-mi dataset. Each column corresponds to one target image.

Figure 5. Distributions of SSCD and CLIP similarity score on all target copyrighted images in two datasets using the original caption as prompts. The distributions of the SSCD score are more clearly bimodal to distinguish between non-infringing and infringing samples.

copyright protection, all samples are non-infringing content as CP-k rejects all infringing samples. In the third column, we find that an amplification attack with a benign prompt can be unsuccessful, because such a prompt may not provide a strictly positive probability of producing infringing generations from models protected by CP-k. However, in the last column, with an adversarial prompt obtained from our proposed Anti-NAF algorithm, we can see that most of the outputs are copyright-infringed, which means that the probability of infringing samples is largely amplified.

In Fig. 9, we give detailed FAR-AR curves on each target copyrighted image in LAION-mi dataset. We can find that our proposed bandit amplification method performs more steadily in the worst cases. For example, in Figs. 9a and 9d, when acceptance rate is lower than 20%, the FAR of Anti-NAF with amplification is nearly 0%; while ε-greedy-max/-cdf bandit amplification can adapt to follow the best choice of prompts (*e.g.*, PEZ or CLIP-Interrogator) and keep a competitive FAR score.

11.2. Human Evaluation

In Tab. 3, we conduct a human evaluation on two target copyrighted images from two datasets. We randomly select 100 accepted samples obtained from each of the two threat models (the original caption and ε-greedy-cdf). For each target image, a total of 200 samples are randomly shuffled and displayed to 5 graduate students. They are told to label each sample as non-infringing or infringing the copyright of the given target image. Finally, we report their average copyright infringement rates.

11.3. Results on Other Similarity Thresholds

The results on the additional thresholds described in Sec. 10 are reported in Tabs. 6 and 7. We can find that under more strict similarity thresholds, our proposed Anti-NAF can also provide a non-trivial probability of producing infringing content even with a low acceptance rate. Besides, Anti-NAF outperforms other threat prompts under all different similarity thresholds, highlighting its effectiveness.

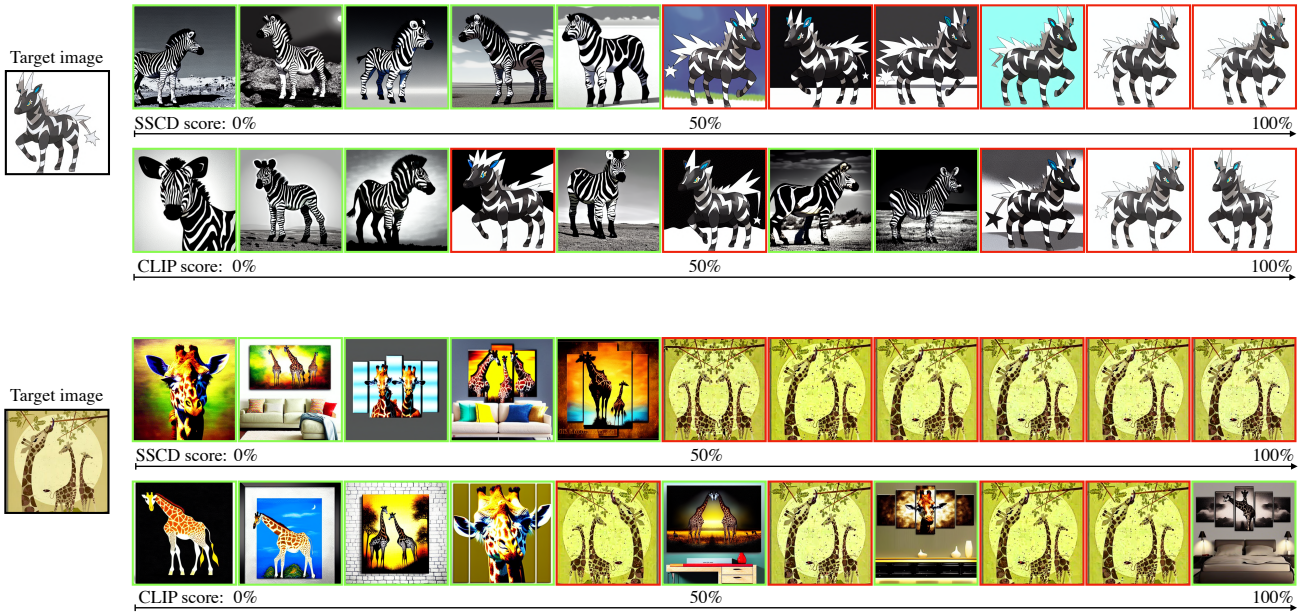


Figure 6. Example images generated from the original caption of target images (non-infringing and infringing images are marked with green and red boundaries, respectively). From left to right, images are sorted by similarity score in ascending order. An ideal similarity score threshold should distinguish between non-infringing (lower score) and infringing samples (higher score). From the example images, the SSCD score performs much better than the CLIP score.

Figure 7. Qualitative examples near the similarity decision thresholds (target, non-infringing, and infringing images are marked with black, green, and red boundaries, respectively).

Dataset	Caption (w/o Amp.)	ε-greedy-cdf Amp.
POKEMON	0.6%	83.0%
LAION-mi	0.4%	42.4%

Table 3. Human evaluation results of copyright-infringement rate on selected target images of two datasets. Acceptance rates of 10% and 40% are applied for POKEMON and LAION-mi respectively.

Methods	FAR@5%AR↑	FAR@15%AR↑
Caption	2.13%	11.07%
Anti-NAF	**9.07%**	**21.87%**

Table 4. Results on selected target images of POKEMON. The prompts of Anti-NAF are obtained with StableDiffusion-v1-4, while attacks are conducted on StableDiffusion-v1-5.

11.4. Results on Transfer Attack

In Tab. 4, we investigate the generalizability of our proposed Anti-NAF algorithms on transfer attack settings. Specifically, the adversarial prompt optimization is conducted based on a fine-tuned StableDiffusion-v1-4 model, while the obtained prompts are then utilized to attack the fine-tuned StableDiffusion-v1-5 model. The results indicate that prompts generated on a white-box model using Anti-NAF can serve as candidate prompts for VA3 to attack other black-box models. We hope this study can inspire future work to explore black-box attacks in practical scenarios.

12. Additional Results on Ablation Study

In Tab. 8, we report the results of the ablation study on LAION-mi dataset. We can observe that the results show similar trends as that of the POKEMON dataset in Tab. 2. This further verifies that the optimization objective of our proposed Anti-NAF algorithm is effective and well-balanced between $\mathcal{L}_p$ and $\mathcal{L}_q$ with the help of loss clip bound φ. In Tab. 5, we also report ablation experiments on other choices of denoising steps T of text-to-image diffusion models. We can find that our proposed Anti-NAF keeps superior performance, suggesting that its effectiveness is immune to different T.

Figure 8. Example outputs given the copyright images in the second row of Fig. 3 as targets (potential infringing images are marked with red boundaries). In the first column, using a benign prompt, we observe a high incidence of infringing content from models without copyright protection ("w/o CP-k"). In contrast, all samples in the second column are safe after applying the copyright protection mechanism ("w/ CP-k"). In the third column, we find that amplification (Amp.) attack with a benign prompt can be unsuccessful. However, by amplification attack with an adversarial prompt obtained from our proposed Anti-NAF algorithm, most outputs in the last column are copyright-infringed.

T	Methods	FAR@5%AR↑	FAR@15%AR↑
25	Caption	0.68%	3.52%
	Anti-NAF	**12.32%**	**14.44%**
100	Caption	0.00%	3.40%
	Anti-NAF	**9.24%**	**9.52%**

Table 5. Results with different denoising steps T on POKEMON.

Methods	SSCD-45%			SSCD-55%		
	CIR	FAR@5%AR↑	FAR@15%AR↑	CIR	FAR@5%AR↑	FAR@15%AR↑
Caption (w/o Amp.)	47.96%	0.84%	3.60%	42.64%	0.48%	2.64%
CLIP-Int. (w/o Amp.)	31.28%	3.44%	5.24%	18.64%	0.64%	1.48%
PEZ (w/o Amp.)	13.88%	3.28%	5.64%	5.60%	0.92%	1.52%
Anti-NAF (w/o Amp.)	22.56%	**14.68%**	**19.80%**	8.08%	**5.08%**	**6.52%**
Caption (w/ Amp.)	100.00%	14.64%	38.72%	100.00%	14.64%	38.68%
CLIP-Int. (w/ Amp.)	99.84%	24.12%	48.00%	99.84%	17.64%	44.16%
PEZ (w/ Amp.)	74.44%	30.64%	48.88%	63.32%	15.52%	34.28%
Anti-NAF (w/ Amp.)	99.92%	**86.28%**	**96.48%**	99.36%	**62.12%**	**66.44%**

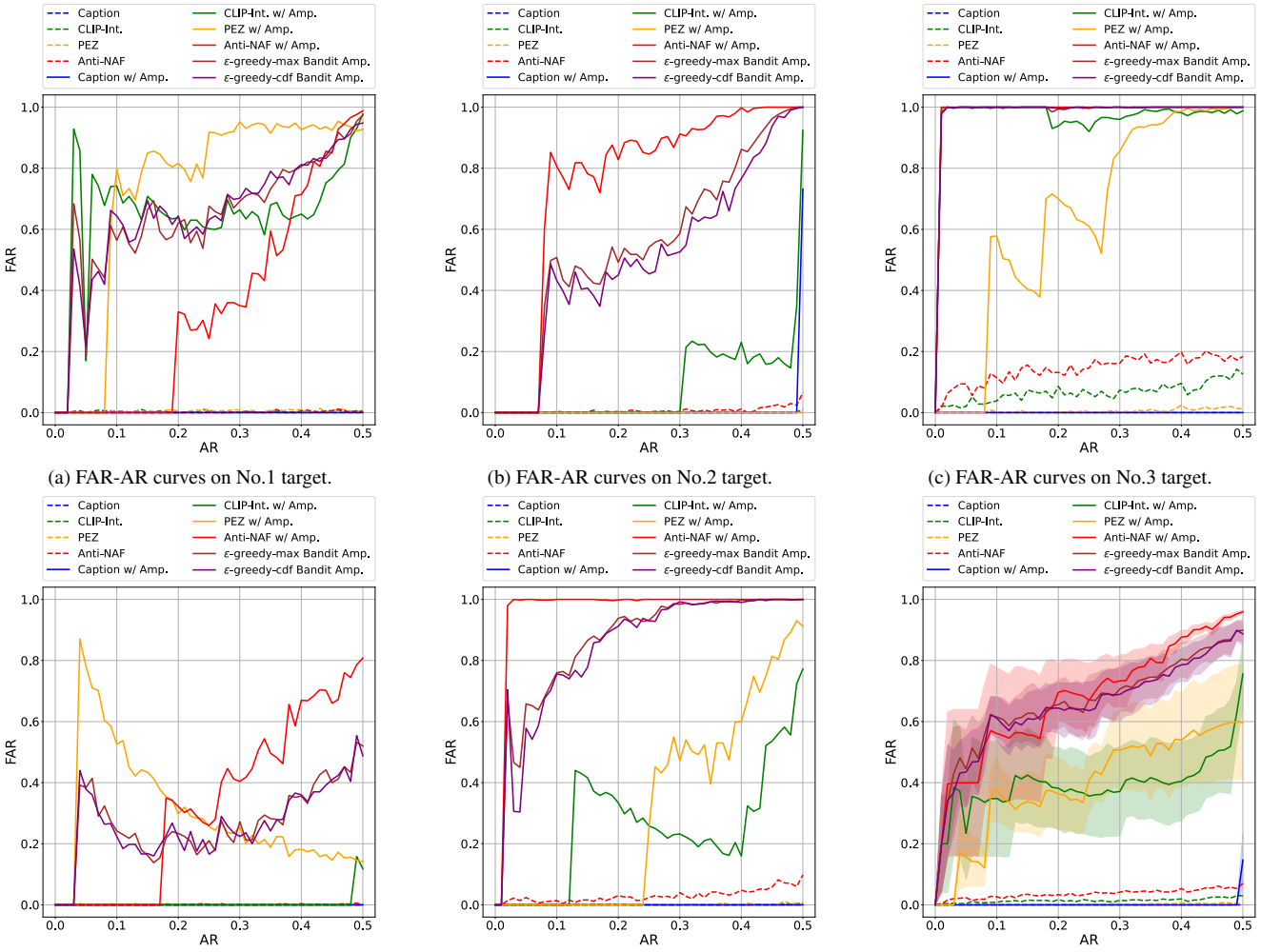
Table 6. Quantitative results on POKEMON dataset using SSCD-45% and SSCD-55% as the threshold for infringement judgment. (CLIP-Int. is the abbreviation for CLIP-Interrogator).

Methods	POKEMON			LAION-mi			
	CIR	FAR@5%AR↑	FAR@15%AR↑	CIR	FAR@10%AR↑	FAR@30%AR↑	FAR@50%AR↑
Caption (w/o Amp.)	40.40%	0.08%	1.52%	48.52%	0.00%	0.00%	0.08%
CLIP-Int. (w/o Amp.)	23.96%	1.72%	2.76%	38.04%	0.00%	0.00%	0.20%
PEZ (w/o Amp.)	8.28%	0.28%	0.80%	9.64%	0.00%	0.00%	0.00%
Anti-NAF (w/o Amp.)	16.20%	**11.48%**	**13.12%**	26.32%	**0.12%**	**0.20%**	**1.68%**
Caption (w/ Amp.)	100.00%	3.56%	35.52%	100.00%	0.00%	0.00%	14.72%
CLIP-Int. (w/ Amp.)	97.96%	14.68%	26.36%	100.00%	0.00%	0.00%	46.84%
PEZ (w/ Amp.)	61.00%	10.80%	25.44%	81.56%	0.00%	0.00%	4.92%
Anti-NAF (w/ Amp.)	89.28%	**48.32%**	**67.04%**	99.68%	**26.24%**	**39.96%**	**59.44%**

Table 7. Quantitative results using SSCD-gmm as the threshold for infringement judgment. (CLIP-Int. is the abbreviation for CLIP-Interrogator).

Methods	CIR	FAR@10%AR↑	FAR@30%AR↑	FAR@50%AR↑
Anti-NAF	33.84%	2.64%	4.16%	7.00%
$\mathcal{L}_p$ only	30.44%	0.28%	1.68%	2.76%
w/o φ	33.64%	0.32%	0.60%	1.16%
w/o $\mathcal{L}_q$	24.28%	0.76%	2.84%	3.28%

Table 8. Ablation study for Anti-NAF algorithm on LAION-mi.



(a) FAR-AR curves on No.1 target.

(b) FAR-AR curves on No.2 target.

(c) FAR-AR curves on No.3 target.

(d) FAR-AR curves on No.4 target.

(e) FAR-AR curves on No.5 target.

(f) Overall averaged FAR-AR curves.

Figure 9. FAR-AR curves on each copyrighted image in LAION-mi. For No.1 and 4 target copyrighted images, Anti-NAF performs worse when acceptance rate is lower than 20%, while bandit amplification methods show steady performance in these worst cases.