SimAC: A Simple Anti-Customization Method for Protecting Face Privacy against Text-to-Image Synthesis of Diffusion Models
Supplementary Material

Feifei Wang[1,2,*] Zhentao Tan[2,1], Tianyi Wei[1], Yue Wu[2], Qidong Huang[1,†]
[1]University of Science and Technology of China [2]Alibaba Cloud
{wangfeifei@, tzt@, bestwty@, hqd0037@}mail.ustc.edu.cn matthew.wy@alibaba-inc.com

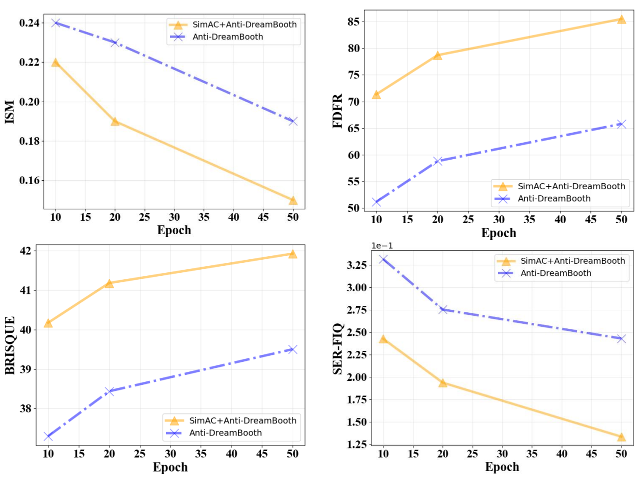

Figure 1. Setting different training epochs to compare the performance of Anti-DreamBooth and SimAC+Anti-DreamBooth on CelebA-HQ dataset, where lower ISM and SER-FIQ are better and higher FDFR and BRISQUE are better. The comparison aims to examine the training epoch required for both to achieve similar performance, which indicates their efficiency.

1. Anti-Customization Efficiency

Our method, SimAC+Anti-DreamBooth, outperforms Anti-DreamBooth despite both using the same training epochs. To highlight the effectiveness of SimAC+Anti-DreamBooth by demonstrating its ability to achieve comparable protection to Anti-DreamBooth with fewer training steps, we conduct a comparative study employing training epochs of $(10, 20, 50)$. We ensure a comprehensive evaluation by averaging metrics across four different prompts.

Our findings highlight the significant efficiency boost that SimAC brings to Anti-Dreambooth. As depicted in Figure 1, the combined impact of SimAC and Anti-Dreambooth after only 20 training epochs mirrors the performance achieved by Anti-Dreambooth trained for 50 epochs. This observation supports our conclusion that within Anti-Dreambooth's random timestep sampling, some larger timesteps may be selected and cause gradients to approach zero. Consequently, this leads to a failure to sufficiently disrupt the reconstruction fidelity of the original image.

To visually demonstrate the impact of our adaptive greedy time interval selection on enhancing perturbed image gradients during the attack process, we've recorded the averaged mean of absolute gradients for both SimAC+Anti-DreamBooth and Anti-DreamBooth across all training timesteps.

As depicted in Table 1, whether trained for 20 or 50 epochs, integrating the adaptive time selection approach noticeably enhances gradients during training. This enhancement significantly improves training efficiency and enhances the effectiveness of protection measures countering malicious customization.

Method	epoch=50	epoch=20
Anti-DB	3.36×10^{-6}	3.26×10^{-6}
SimAC+Anti-DB	1.23×10^{-5}	1.14×10^{-5}

Table 1. The comparison of the averaged mean of absolute gradient values of perturbed images during the training process between SimAC+Anti-DreamBooth and Anti-Dreambooth with training epochs 20 and 50 on the CelebA-HQ dataset.

2. GPT-4V(ision) Evaluation

Beyond employing the conventional deep face recognition technique ArcFace[1] to determine identity similarity between the generated image and the user-provided one, we leverage GPT-4V(ision) to more accurately simulate human judgment regarding identity similarity.

We randomly selected images generated from prompts such as "a photo of sks person", "a dslr portrait of sks per-

*Work done during an internship in Alibaba Cloud.
†Corresponding author.

GPT-4V(ision) Prompt

Given the four-square grid image that contains four face images (i.e., the top left image, the top right image, the bottom left image, and the bottom right image), you are required to score the identity similarity between the top left image and each of the remaining three images, respectively. You should pay extra attention to the identity similarity between the persons that appear in the two images, which refers to the similarity of the person's appearance such as facial features, expressions, skin textures, facial ratio, etc.
Please rate the identity similarity on a scale of 0 to 10, where a higher score indicates a higher similarity. The scores are required to have a certain degree of difference.
Please output the scores for the top right image, the bottom left image, and the bottom right image when compared with the top left image. The three scores are separated by a space. Following the scores, please provide an explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the face images were presented does not affect your judgment.

Output format:
Similarity: <Scores of the top right image, the bottom left image, and the bottom right image>
Reason:

Table 2. The prompt used for GPT-4V(ision) evaluation.

Figure 2. The concatenated image input to GPT-4V(ision) which are generated with the inference prompt "a photo of sks person looking at the mirror".

son" and "a photo of sks person looking at the mirror". And then, selected images are filtered to retain those detectable by Retinaface[2]. These images were then categorized based on the corresponding anti-customization methods. For each input, we compile sets of four images, as depicted in Figure 2. The top-left image represents the clean user portrait, while the top-right, bottom-left, and bottom-right sections illustrate outcomes from AdvDM, Anti-DreamBooth, and the combined effect of SimAC and Anti-DreamBooth, respectively. Subsequently, we obtain a total of 50 images, all of which are fed into GPT-4V(ision)[6]. The objective is to compare identity similarity with the clean user portrait across the varied results.

We craft the prompts which are the input, as follows in Table 2. The model needs to assess similarity based on a person's facial features and requests a similarity rating within the range of 0 to 10 along with corresponding explanations or justifications.

The final assessment process 50 images, resulting in valid responses for 35 of them, as detailed in Table 3. Despite SimAC+Anti-DreamBooth showing a higher facial detection failure rate (FDFR) in comparison to AdvDM and Anti-DreamBooth, its performance within the subset of images containing detectable faces stands out significantly. SimAC+Anti-DreamBooth exhibits the lowest similarity scores and most profound disruption of identity information in portraits among the three methods within this evaluated subset under the judgment of large-scale pretrained multimodal models GPT-4V(ision).

Method	GPT-4V(ision) score (total)
AdvDM[5]	86
Anti-DB[9]	74
SimAC+Anti-DB	61

Table 3. The comparison of the total similarity scores given by GPT-4V(vision) between AdvDM, Anti-Dreambooth and SimAC+Anti-DreamBooth.

3. Generalizing to Other Anti-Customization Methods

Our examination delves into the influence of SimAC on other anti-customization methods. Due to PhotoGuard[8] primarily targeting the VAE encoder attack without involving the selection of timesteps for noise addition during training, we opted to scrutinize the impact of SimAC specifically on AdvDM[5].

As indicated in Table 4, our SimAC demonstrates consistent improvements in face detection failure rates across four prompts. Notably, in the latter two prompts, SimAC effectively enhances both identity protection metrics (ISM and FDFR) and the degradation of image quality metrics (BRISQUE and SER-FIQ) achieved by AdvDM. In the second prompt, its performance is comparable to AdvDM's outcomes. However, in the initial prompt, "a photo of sks person", the combination of AdvDM and SimAC displays less effective performance. This discrepancy in the first inference prompt, which is also utilized during training, suggests that AdvDM might be better suited for this specific prompt, indicating a more obvious reduction in protection effectiveness when encountering unseen prompts. Considering that inferring with unseen prompts aligns more closely with real-world scenarios, the comprehensive integration of SimAC notably enhances the privacy protection performance of AdvDM.

Method	"a photo of sks person"			
	ISM↓	FDFR↑	BRISQUE ↑	SER-FIQ↓
AdvDM [5]	0.32	70.48	38.17	0.20
AdvDM + SimAC	0.45	72.18	38.03	0.41
Method	"a dslr portrait of sks person"			
	ISM↓	FDFR↑	BRISQUE ↑	SER-FIQ↓
AdvDM [5]	0.25	65.37	37.86	0.41
AdvDM + SimAC	0.27	80.48	39.21	0.44
Method	"a photo of sks person looking at the mirror"			
	ISM↓	FDFR↑	BRISQUE ↑	SER-FIQ↓
AdvDM[5]	0.29	35.10	36.46	0.36
AdvDM + SimAC	0.26	73.61	41.18	0.21
Method	"a photo of sks person in front of eiffel tower"			
	ISM↓	FDFR↑	BRISQUE ↑	SER-FIQ↓
AdvDM[5]	0.09	38.10	36.02	0.30
AdvDM + SimAC	0.08	48.10	41.79	0.19

Table 4. Combining SimAC with AdvDM on the CelebA-HQ dataset improves the performance of AdvDM, where lower ISM and SER-FIQ are better and higher FDFR and BRISQUE are better. The inference prompt in gray color is the same as the training prompt.

η	"a photo of sks person"			
	ISM↓	FDFR↑	BRISQUE ↑	SER-FIQ↓
4/255	0.60	14.15	36.11	0.70
8/255	0.47	73.06	38.81	0.38
16/255	0.31	87.07	38.86	0.21
32/255	0.22	87.21	40.86	0.14
η	"a dslr portrait of sks person"			
	ISM ↓	FDFR ↑	BRISQUE↑	SER-FIQ↓
4/255	0.48	29.86	17.66	0.74
8/255	0.32	74.56	41.14	0.45
16/255	0.12	96.87	42.10	0.15
32/255	0.04	98.71	41.28	0.04
η	"a photo of sks person looking at the mirror"			
	ISM↓	FDFR↑	BRISQUE ↑	SER-FIQ↓
4/255	0.42	17.69	28.21	0.56
8/255	0.29	64.08	42.58	0.26
16/255	0.12	91.90	43.97	0.06
32/255	0.07	94.29	44.02	0.03
η	"a photo of sks person in front of eiffel tower"			
	ISM↓	FDFR↑	BRISQUE ↑	SER-FIQ↓
4/255	0.14	29.46	30.45	0.37
8/255	0.08	40.88	41.45	0.22
16/255	0.05	66.19	42.77	0.12
32/255	0.04	71.90	41.78	0.07

Table 5. Different noise budget based on SimAC on CelebA-HQ dataset, where lower ISM and SER-FIQ are better and higher FDFR and BRISQUE are better. As the noise budget increases, the deteriorating effect on the generated image increases.

4. Ablation Study

Noise Budget We adjust the noise budget to observe the impact of η on defense performance. The results in Table 5 show that as the noise budget increases, the defense effectiveness improves. To ensure the image quality of the input image, we select 16/255 as the default noise budget.

Different Layers when Computing The Feature Interference Loss To verify our analysis of the properties of LDMs, that is, in UNet decoder blocks, as the layers go deeper, the features pay more attention to high-frequency information and are more prone to adversarial noise, we compare different layer selection in Table 6. The results show that the choice of layer 9,10,11 has the best performance among all groups in four prompts comprehensively and this is consistent with our conclusion that perturbation of deeper features in the UNet decoder will bring more anti-customization performance gains.

5. Customization Method Mismatch

In the previous context, we assume that malicious users fine-tune models based on DreamBooth. However, due to the high training computational cost of DreamBooth, subsequent efforts have been made to reduce the required memory for training while maintaining the fidelity of customization, such as Dreambooth+Lora[3] and Custom Diffusion[4]. Hence, we conduct a evaluation for Anti-DreamBooth and SimAC+Anti-DreamBooth when the customization methods mismatch. The results in Table 7 are obvious that SimAC+Anti-DreamBooth brings more safeguards than Anti-DreamBooth across different customization strategies.

Learning adversarial noise based on Dreambooth, and then customizing the protected images based on either Dreambooth+Lora or Custom Diffusion, yields performance decline. However, the identity similarity (ISM) remains at a lower level, offering a degree of user portrait protection. In this setup, enhancing transferred defense against unauthorized customization is an area we will explore in future research.

6. Qualitative Results

The preceding part of the text includes a quantitative assessment of various settings. Next, we'll proceed with qualitative evaluations.

6.1. Noise Budget

The quantitative results for the noise budget already demonstrate that with an increase in the noise budget, the performance of anti-customization improves. From a qualitative standpoint, in Figure 3, protected user portrait images under different noise budgets are depicted after customization. It

layer	"a photo of sks person"				"a dslr portrait of sks person"			
	ISM↓	FDFR↑	BRISQUE ↑	SER-FQA↓	ISM ↓	FDFR ↑	BRISQUE↑	SER-FQA↓
0,1,2	0.28	86.53	39.26	0.20	0.09	96.33	42.52	0.13
3,4,5	0.30	84.69	38.01	0.20	0.10	96.80	42.14	0.15
6,7,8	0.30	87.35	40.14	0.21	0.15	95.17	42.00	0.16
9,10,11	0.31	87.07	38.86	0.21	0.12	96.87	42.10	0.15
layer	"a photo of sks person looking at the mirror"				"a photo of sks person in front of eiffel tower"			
	ISM↓	FDFR↑	BRISQUE ↑	SER-FQA↓	ISM ↓	FDFR ↑	BRISQUE↑	SER-FQA↓
0,1,2	0.14	88.23	44.68	0.08	0.05	63.95	42.98	0.12
3,4,5	0.12	89.80	43.93	0.08	0.05	62.93	43.12	0.11
6,7,8	0.14	90.20	44.66	0.08	0.06	65.24	43.73	0.11
9,10,11	0.12	91.90	43.97	0.06	0.05	66.19	42.77	0.12

Table 6. Comparison of different layer combinations for feature interference loss on CelebA-HQ dataset. We evaluate the performance under four different prompts during customization, where lower ISM and SER-FQA are better and higher FDFR and BRISQUE are better.

Test	Method	"a photo of sks person"				"a dslr portrait of sks person"			
		ISM↓	FDFR↑	BRISQUE ↑	SER-FIQ↓	ISM ↓	FDFR ↑	BRISQUE↑	SER-FIQ↓
DreamBooth[7]	Anti-DreamBooth	0.28	77.28	37.43	0.20	0.19	86.80	38.90	0.27
	SimAC+Anti-DreamBooth	0.31	87.07	38.86	0.21	0.12	96.87	42.10	0.15
Lora[3]	Anti-DreamBooth	0.30	44.63	33.02	0.51	0.22	16.73	10.43	0.60
	SimAC+Anti-DreamBooth	0.19	82.86	39.99	0.35	0.20	46.94	20.19	0.54
Custom Diffusion[4]	Anti-DreamBooth	0.60	6.39	38.90	0.73	0.46	5.99	9.83	0.75
	SimAC+Anti-DreamBooth	0.27	85.71	39.69	0.47	0.45	18.37	17.73	0.73
Test	Method	"a photo of sks person looking at the mirror"				"a photo of sks person in front of eiffel tower"			
		ISM↓	FDFR↑	BRISQUE ↑	SER-FIQ↓	ISM ↓	FDFR ↑	BRISQUE↑	SER-FIQ↓
DreamBooth [7]	Anti-DreamBooth	0.23	42.86	40.34	0.28	0.06	56.26	41.35	0.22
	SimAC+Anti-DreamBooth	0.12	91.90	43.97	0.06	0.05	66.19	42.77	0.12
Lora[3]	Anti-DreamBooth	0.15	26.12	21.60	0.32	0.08	63.61	15.31	0.27
	SimAC+Anti-DreamBooth	0.13	18.10	21.52	0.33	0.05	64.35	26.15	0.12
Custom Diffusion[4]	Anti-DreamBooth	0.27	14.83	21.42	0.42	0.12	15.37	26.50	0.43
	SimAC+Anti-DreamBooth	0.27	33.33	32.04	0.34	0.11	17.14	34.55	0.36

Table 7. Customization strategy mismatch during training and testing on CelebA-HQ dataset, where lower ISM and SER-FIQ are better and higher FDFR and BRISQUE are better. The training is based on Dreambooth and the customization test is based on Lora or Custom Diffusion.

can be observed that with a smaller noise budget of 4/255, the protective effect is minimal. However, at a noise budget of 16/255, the effect becomes highly pronounced, meeting our objectives for user identity privacy protection. To strike a balance between image quality and the effect of disruption, we opt for 16/255 instead of choosing a higher level of image degradation at 32/255.

6.2. Prompt Mismatch

We show results in Figure 4 when different prompts are used for learning noise and fine-tuning stable diffusion (the rare identifiers change from "sks" to "t@t"). Despite a slight decline in performance, SimAC+Anti-dreambooth is still able to generate sufficient artifacts to protect users' portrait privacy.

6.3. Model Mismatch

We exhibit the results in Figure 5, considering whether the model used for customization is same or different from the model used for adversarial noise. It is evident that the protection is significant when these two model versions are consistent. However, when the versions are different, as seen in cases train v1.4 and test v2.1, where added adversarial noise needs to transfer across different model versions, the protection effect is weakened. Nevertheless, overall, whether the model versions are the same or different, strong distortion is observed in most prompts.

6.4. Customization Method Mismatch

When adding adversarial noise, we use the training strategy based on Dreambooth. Assuming the customized method

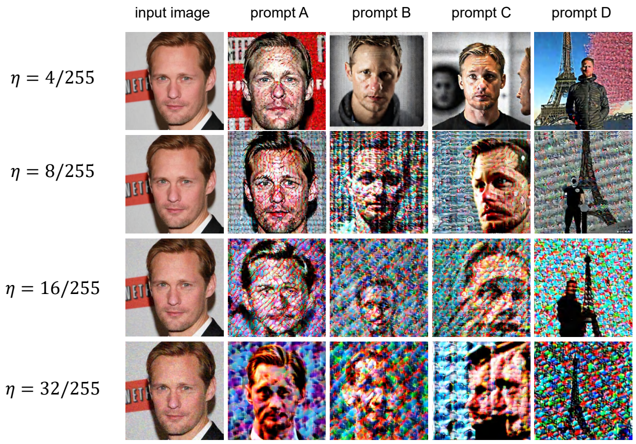

Figure 3. Quantitative results of different noise budgets under four prompts on CelebA-HQ dataset. The training noise budget is constrained to 4/255, 8/255, 16/255 and 32/255. The prompt A is "a photo of sks person", the prompt B is "a dslr portrait of sks person", the prompt C is "a photo of sks person looking at the mirror" and the prompt D is "a photo of sks person in front of eiffel tower".

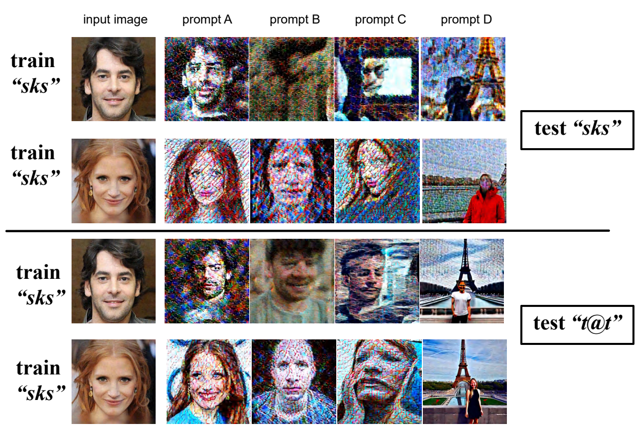

Figure 4. Quantitative results of prompt mismatch under four prompts on CelebA-HQ dataset. The training rare identifier [v] is "sks" and the customization test rare identifier is "t@t". This aims to test performance when the prompt used for customization isn't foreseen. The prompt A is "a photo of sks person", the prompt B is "a dslr portrait of sks person", the prompt C is "a photo of sks person looking at the mirror" and the prompt D is "a photo of sks person in front of eiffel tower"

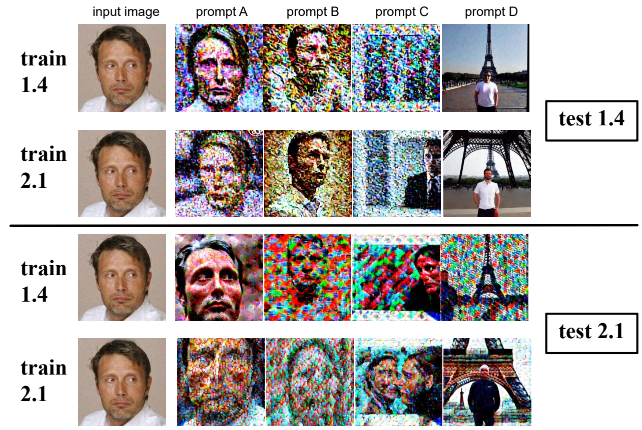

Figure 5. Quantitative results of models mismatch under four prompts on CelebA-HQ dataset. The training uses stable diffusion v1.4 or v2.1, and the testing uses v1.4 and v2.1 in combination with the training model, respectively, to test the sensitivity of the method to the model version. The prompt A is "a photo of [v] person", the prompt B is "a dslr portrait of [v] person", the prompt C is "a photo of [v] person looking at the mirror" and the prompt D is "a photo of [v] person in front of eiffel tower"

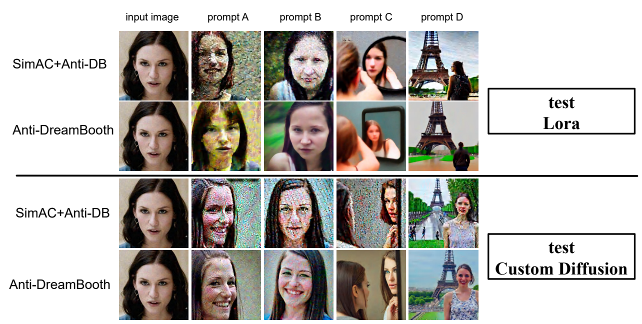

Figure 6. Quantitative results of customization mismatch under four prompts on CelebA-HQ dataset with Anti-DreamBooth and SimAC+Anti-DreamBooth. The training is based on Dreambooth and the customization test is based on Lora or Custom Diffusion. The experiment simulates an attacker using an unknown customization method. The prompt A is "a photo of sks person", the prompt B is "a dslr portrait of sks person", the prompt C is "a photo of sks person looking at the mirror" and the prompt D is "a photo of sks person in front of eiffel tower".

to be perturbed is different, such as Lora[3] and Custom Diffusion[4] is another black-box test. The results, as depicted in Figure 6, indicate that our method, when transferred to attack Lora-based customization, still performs well. However, when transferred to attack Custom Diffusion-based customization, the artifact effect is not so pronounced. It is worth noting that SimAC+Anti-DreamBooth enhances improve the transer performance of adversarial noise optimized by Anti-DreamBooth in both resisting Lora or Custom Diffusion.

6.5. Comparison on VGG-Face2 Dataset

On the VGG-Face2 dataset, we select two different individuals for baseline comparison. As shown in Fig 7, although Photoguard, ADVDM, and Anti-Dreambooth perturb the generated images, many characteristics of the user-provided portrait are still present, leading to potential privacy leaks. SimAC+Anti-DreamBooth, however, obscures more facial features and details from the user input images, resulting in the highest level of facial distortion and

Figure 7. Quantitative results of two people under four prompts on VGG-Face2 dataset. For each person, the first row is "a photo of sks person", the second row is "a dslr portrait of sks person", the third row is "a photo of sks person looking at the mirror" and the last row is "a photo of sks person in front of eiffel tower".

thereby achieving the most effective user information concealment.

References

[1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1

[2] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 2

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3, 4, 5

[4] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3, 4, 5

[5] 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*, pages 20763–20786. 2, 3

[6] OpenAI. Gpt-4 technical report, 2023. 2

[7] 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. 4

[8] 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*, pages 29894–29918. 2

[9] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2116–2127, 2023. 2