DUAW: DATA-FREE UNIVERSAL ADVERSARIAL WATERMARK AGAINST STABLE DIFFUSION CUSTOMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Stable Diffusion (SD) customization approaches enable users to personalize SD model outputs, greatly enhancing the flexibility and diversity of AI art. However, they also allow individuals to plagiarize specific styles or subjects from copyrighted images, which raises significant concerns about potential copyright infringement. To address this issue, we propose an invisible data-free universal adversarial watermark (DUAW), aiming to protect copyrighted images from different customization approaches across various versions of SD models. First, DUAW is designed to disrupt the variational autoencoder during SD customization. Second, DUAW is trained on synthetic images produced by a Large Language Model (LLM) and a pretrained SD model, that is, it is generated in a data-free manner without the use of any copyrighted images. Once crafted, DUAW can be imperceptibly integrated into any copyrighted image, serving as a protective measure by inducing significant distortions in the images generated by customized SD models. Experimental results demonstrate that DUAW can distort the outputs of fine-tuned SD models, making them discernible to both human observers and a simple classifier, yielding more effective protection results than existing methods.

1 INTRODUCTION

Recently, images generated by Stable Diffusion (SD) (Rombach et al., 2022) have exhibited exceptional visual quality, exerting a profound influence on the academic and industrial communities. SD customization tools such as DreamBooth (Ruiz et al., 2023) and LoRA (Hu et al., 2022) have also been developed, enabling users to personalize SD models and generate images according to their preferences. However, these tools have raised concerns about potential intellectual property (IP) infringement risks, as individuals can conveniently customize their SD models to plagiarize a specific subject or style from copyrighted images.

In this paper, we propose to use a universal adversarial watermark to achieve copyright protection against common SD customization approaches. However, achieving this goal is non-trivial. Firstly, unlike conventional adversarial attacks that aim to maximize loss for a fixed model, disrupting SD customization is challenging as the weight of the target model changes dynamically. Secondly, various SD models with different weights have been released, highlighting the need for the proposed watermark to generalize across different SD models. Thirdly, due to the potential confidentiality concerns associated with copyrighted images, we may not have access to the images that need to be protected when training the adversarial watermark. As previous works such as AdvDM (Liang et al., 2023), Anti-DreamBooth (Le et al., 2023), and Glaze (Shan et al., 2023) focus on image-specific watermarks that requires direct access to the copyrighted images, this issue remains unresolved.

To address the concerns above, we propose a novel data-free adversarial watermark generation framework for copyright protection. Specifically, we select the VAE part of the SD model as the adversarial attack target and adopt an optimizer-based watermark training approach. Furthermore, we introduce a data-free technique that utilizes ChatGPT and SD v2.1 to create training images, allowing the training of the proposed watermark without accessing any copyrighted images requiring protection. With just **100 synthetic images**, the generated DUAW can successfully protect a wide range of images from being utilized by the SD model in subject-driven and style-driven generation.

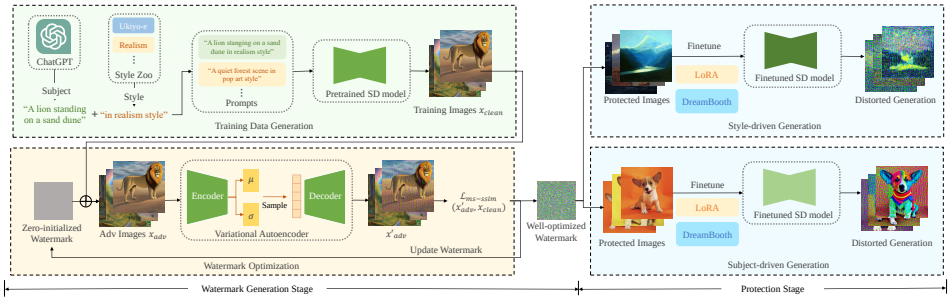

Figure 1: Overall pipeline of DUAW.

2 METHOD

2.1 PROBLEM DEFINITION

We aim to propose an adversarial protection watermark to disrupt the SD fine-tuning process and induce distortion to output images generated by customized SD models. Meanwhile, to mitigate the threat of different customization approaches employed by different models on various training images, the proposed watermark must exhibit generalizability to consistently protect copyrighted images under these circumstances. Hence, we formally define the following: for n images requiring protection, denoted as $x \in \mathcal{X}$, we apply our proposed watermark W to each image, resulting in protected images $x_{adv} = \{x^{(i)} + W\}_i^n$. These watermarked images are then used as inputs to the pretrained SD model S_θ with parameters θ, which undergoes fine-tuning using the fine-tuning tool f. Our objective is to induce the fine-tuned SD model to produce distorted images with any prompts $y \in \mathcal{Y}$ by minimizing a certain image quality metric $\mathcal{Q}$:

$$\begin{aligned} W &= \arg\min_W \mathbb{E}_{x,y}(\mathcal{Q}(S_{\theta'}(y))), \\ \theta' &= f(S_\theta, x_{adv}), \\ s.t. &\|W\|_\infty < \epsilon, \end{aligned} \tag{1}$$

where θ' is the fine-tuned weight of SD model, y is any prompt to condition the SD generation, and ϵ is a predefined upper bound that limits the maximum pixel alteration of the watermark W.

2.2 WATERMARK GENERATION PROCESS

Directly optimizing the objective in Eq. 1 is challenging as we must optimize the loss on top of a dynamic weight-changing process, which differs from the conventional adversarial attack setting where the target model weight is frozen. To tackle this issue, we set our eyes on the part of the SD model that stays unchanged during fine-tuning, *i.e.*, VAE, and disrupts its decoding process. Meanwhile, we have also observed that SD model fine-tuning methods tend to preserve minor perturbations added to training images (refer to Appendix D). That is to say, if our watermark W can perturb the decoding process of the VAE and result in distorted output, the fine-tuning process will also let the SD model learn such a special distribution of latent codes which can disrupt the decoder and introduce distortions into output images.

Specifically, during the watermark training, the encoder of the VAE maps the watermarked images x_{adv} to the parameters of the posterior distribution $q(z_{adv}|x_{adv})$ over the latent code z_{adv}. To be more specific, VAE approximates the $q(z_{adv}|x_{adv})$ using a Gaussian distribution, where the mean and standard deviation are given by the VAE encoder $\mathcal{E}$:

$$\begin{aligned} \mu_{adv}, \sigma_{adv} &= \mathcal{E}(x_{adv}), \\ z_{adv} &\sim \mathcal{N}(\mu_{adv}, \sigma_{adv}). \end{aligned} \tag{2}$$

After that, the decoder of the VAE maps z_{adv} back to the data space to produce reconstructed Image $\hat{x_{adv}} = \mathcal{D}(z_{adv})$. To induce distortions and lower the image quality of decoder outputs, we minimize the MS-SSIM between x and $\hat{x_{adv}}$. In contrast to the commonly employed loss functions, MS-SSIM offers a multi-scaled assessment of the generated image quality and can induce more visually

distorted outputs(refer to Experiment 3.2). Hence, our optimization objective becomes:

$$\mathcal{L}_{ms-ssim} = \mathbb{E}_{x,z_{adv}\sim\mathcal{N}(\mu_{adv},\sigma_{adv})}\mathcal{MS}(x, \mathcal{D}(z_{adv})), \tag{3}$$

where MS represents the MS-SSIM function. The watermark training process takes batches of images as input and updates the watermark across the entire dataset, ensuring that the generated watermark is capable of protecting multiple images simultaneously.

Due to the difficulty of directly optimizing the expectation over z_{adv}, we approximate it by sampling a single z for each x. VAE utilizes a reparametrization trick to sample z_{adv}, which includes sampling latent variables ζ from a standard Gaussian distribution and transforming them into the desired distribution with the mean and standard deviation given by the encoder. Hence, our optimization objective can be rewritten as:

$$\begin{aligned} &z_{adv} = \mu_{adv} \cdot \zeta + \sigma_{adv},\ \zeta \sim \mathcal{N}(0, 1), \\ &\mathcal{L}_{ms-ssim} = \mathbb{E}_x\mathcal{MS}(x, \mathcal{D}(z_{adv})). \end{aligned} \tag{4}$$

2.3 Data-free Synthetic Dataset via LLM

In certain situations, the valuable copyrighted images needing protection might not be accessible for watermark training, which requires DUAW to be trained without accessing copyrighted images. Fortunately, we observed that the proposed watermark's protection effectiveness does not strongly depend on the training dataset, which leads us to consider the use of an SD-generated dataset as a substitute for the original dataset. Specifically, we exploit an LLM to generate prompts about painting contents $c \in \mathcal{C}$. To improve diversity in the training data, we combine each prompt with various painting styles from the style zoo $s \in \mathcal{S}$ to form the final prompt. Then, we input this prompt into the SD model S_θ to obtain the training data x:

$$x = S_\theta(c \cdot s), \tag{5}$$

where $\cdot$ is the concatenation operator. Experimental results show that our watermark achieves commendable protection performance with a small dataset size of only **100 synthetic images**.

We present our watermark generation pipeline in Fig. 1

3 Experiment

3.1 Implementation Details

Datasets For training, we employ ChatGPT 3.5 [1] to generate 10 random painting content and combine each content with 10 distinct painting styles to form diverse input prompts. Then, we utilize SD v2.1 to generate a 512×512 image for each prompt, resulting in 100 training images (refer to Appendix F). As for evaluation, we adopt DreamBooth dataset (Ruiz et al., 2023) for subject-driven generation and select 120 artworks from 12 artists with diverse styles from the WikiArt [2] for style-driven generation.

Watermark Optimization We set the size of our watermark to 512×512 and constrain the perturbation value within the range of $[-0.05, 0.05]$. During the training process, we set the batch size to 4, total epochs to 1000 and use Adam (Kingma & Ba, 2017) as the optimizer, and apply the learning rate scheduler proposed by T-SEA (Huang et al., 2023), starting with an initial learning rate of 0.01.

Evaluation Metric We utilize 1) CLIP score (Hessel et al., 2021) to evaluate the similarity between the generated and original images and 2) IL-NIQE score (Zhang et al., 2015), a widely used no-reference image quality assessment (IQA), to quantify the perceptual quality of generated image. Additionally, we introduce a classifier to identify the fine-tuned SD outputs images learned from watermarked data and report the ratio of successfully identified images as 3) success rate (SR, refer to Appendix C for more details).

[1] ChatGPT (March 23 version). https://chat.openai.com

[2] https://www.wikiart.org/

Method	WikiArt Dataset Version	CLIP-Score↓ Clean/ Adv	IL-NIQE↑ Clean/ Adv	SR (%)↑	DreamBooth Dataset Version	CLIP-Score↓ Clean/ Adv	IL-NIQE↑ Clean/ Adv	SR (%)↑
LoRA	v1.4	0.5828/ 0.5681	35.62/ 43.82	94.34	v1.4	0.7789/ 0.7137	33.42/ 50.56	97.97
	v1.5	0.5831/ 0.5648	33.64/ 43.95	92.80	v1.5	0.7740/ 0.7143	33.75/ 49.85	98.10
	v2.1	0.5854/ 0.5655	35.62/ 51.25	99.29	v2.1	0.7894/ 0.7010	32.63/ 56.34	99.80
DreamBooth	v1.4	0.6946/ 0.6536	28.39/ 54.67	97.05	v1.4	0.7648/ 0.6985	27.86/ 62.58	99.67
	v1.5	0.6940/ 0.6375	30.91/ 47.67	95.08	v1.5	0.7638/ 0.7054	27.72/ 60.86	99.43
	v2.1	0.7407/ 0.6906	29.66/ 72.92	98.52	v2.1	0.7417/ 0.6937	29.42/ 52.87	93.03

Table 1: Main results of the proposed DUAW. We train DUAW on SD V1.4 and evaluate it with LoRA and DreamBooth fine-tuning on SD V1.4 (rows with gray background), V1.5, V2.1. The result indicate generalization over various SD versions as well as fine-tuning methods

Method	Watermark	DreamBooth Dataset CLIP↓	DreamBooth Dataset IL-NIQE↑	WikiArt Dataset CLIP↓	WikiArt Dataset IL-NIQE↑
Dream Booth	Clean	0.7407	29.66	0.7417	29.42
	Anti-DB	0.7150	33.97	0.7459	28.72
	AdvDM	0.6969	33.63	0.6840	30.96
	Glaze	-	-	**0.6499**	27.47
	Ours	**0.6836**	**71.98**	0.6899	**48.55**
LoRA	Clean	0.7894	32.63	0.5854	35.62
	Anti-DB	0.7394	39.15	0.5833	37.09
	AdvDM	**0.6940**	31.76	0.5854	35.47
	Glaze	-	-	0.5763	34.47
	Ours	0.7067	**45.08**	**0.5680**	**45.84**

Table 2: Comparison with prior works.

Figure 2: Comparison with other losses.

3.2 Results

Image Similarity and Quality We report the quantitative results of the proposed DUAW in Tab. 1, which demonstrate that DUAW can effectively reduce the CLIP-Score, indicating a decrease in the similarity between the generated and original images. Furthermore, the notable increase in IL-NIQE suggests a significant decline in the naturalness of generated images. Hence, results demonstrate that the DUAW can disrupt the SD fine-tuning and distort the generated images of customized SD models, thus protecting the copyrighted images.

Comparison with Prior Works We compare DUAW with other adversarial watermarks, *i.e.*, Anti-DreamBooth (Le et al., 2023), AdvDM (Liang et al., 2023), and Glaze (Shan et al., 2023) (which is specifically designed for style-driven genertaion). *Note that these methods generate image-specific watermarks that should be trained using the copyrighted images.* In contrast, DUAW is designed to be a data-free watermark and can protect unseen copyrighted images. Specifically, we use the default settings of these methods and evaluated them on SD v2.1. Quantitative evaluations of generated image quality in Tab. 2 reveal that DUAW achieves significantly better or comparable protection results to previous works on both datasets with DreamBooth and LoRA Finetuning.

MS-SSIM Compared with Other Loss Function In Fig. 2, We compared MS-SSIM with LPIPS, L2, and the loss function of the VAE on DreamBooth subject-driven generation with SD V2.1, and the results demonstrate that MS-SSIM is more effective than other loss metrics.

4 Conclusion

In this paper, we introduce DUAW, a data-free universal adversarial watermark designed to protect copyrighted images against various customization methods with different SD models. Based on our observation that VAE weights remain unchanged during SD fine-tuning and SD customization methods tend to preserve and amplify minor perturbations, we present a simple yet effective approach to generate DUAW by perturbing the VAE component. Furthermore, to better protect the confidentiality of copyrighted images, we use LLM and pretrained SD models to synthesize a diverse training dataset for training DUAW. Our qualitative and quantitative experiments validate DUAW's efficacy in protecting copyrighted images and show its superiority over the existing methods Anti-DreamBooth and AdvDM.

REFERENCES

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hao Huang, Ziyan Chen, Huanran Chen, Yongtao Wang, and Kevin Zhang. T-sea: Transfer-based self-ensemble attack on object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20514–20523, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2017.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc 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 (ICCV)*, 2023.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. 2023.

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 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20763–20786. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/liang23g.html.

Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. The artbench dataset: Benchmarking generative models with artworks. 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. 2023.

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*, pp. 10684–10695, 2022.

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*, 2023.

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. 2023.

Yasin Yaz, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, Vijay Chandrasekhar, et al. The unusual effectiveness of averaging in gan training. In *International Conference on Learning Representations*, 2018.

Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.

Zixin Zhu, Xuelu Feng, Dongdong Chen, Jianmin Bao, Le Wang, Yinpeng Chen, Lu Yuan, and Gang Hua. Designing a better asymmetric vqgan for stablediffusion. 2023.

A OVERVIEW

These appendices contain the following information:

- We visualize the experiment results of transferability on different VAEs in Appendix B.
- We report the performance of our classifier in Appendix C.
- We visualize examples of fine-tuned SD models preserving minor perturbations in Appendix D.
- We give the detailed fine-tuning settings of DreamBooth and LoRA in Appendix E.
- We present our synthesized training dataset along with the prompts used to generate said dataset in Appendix F.
- We provide our evaluation dataset along with prompts in Appendix G.

B GENERALIZATION OVER DIFFERENT VAES

Given the proposed DUAW is crafted against VAE, we explore the protection capability of DUAW on different VAE variants. We use three commonly used VAEs [3], including the **base** version used in SD v1.4 and v1.5, **ft-ema** fine-tuned on the base version with exponential moving average (EMA) (Yaz et al., 2018) weights, and **ft-mse** fine-tuned with emphasis on MSE loss. These VAE models are all plug-and-play and applicable to any version of SD. We applied each of the three VAEs on SD v2.1 and reported the results. Remarkably, our DUAW demonstrates robustness against these VAE variants. As shown in Tab. 4, the watermark trained on **ft-mse** shows high protection effectiveness against SD models using different VAEs.

We also extend our experiments to include the SDXL 1.0 model (Podell et al., 2023), a recent variant of SD that employs a **newly trained VAE** and incorporates an additional image-to-image SD model, known as the refiner, to enhance its outputs. By resizing the DUAW initially trained on SD v2.1, to the size of 1024×1024, we use DreamBooth finetuning via LoRA [4] to finetune the SDXL model for 500 epochs and visualize the results in Fig. 3. Notably, DUAW is capable of introducing distortions that the refiner is unable to remove, further indicating DUAW's capability of generalizing over **new versions of VAE as well as SD models**.

Furthermore, once an SD model is fine-tuned with DUAW-protected images, changing the VAE will not render the protection of DUAW ineffective. As shown in Fig. 4a, we changed a fine-tuned model's VAE to TAESD, a distilled version of the original VAE, and the fine-tuned SD model still generates distorted outputs. We also tested Asymmetric VQGAN (Zhu et al., 2023) (as shown in Fig. 4b), which is designed specifically for inpainting tasks, and DUAW is capable of inducing distortions in the inpainted results.

C CLASSIFICATION SUCCESS RATE

To better distinguish the outputs of SD fine-tuning from other images (such as real photographs or paintings), we employ a binary classifier, as DUAW can cause obvious distortion in output images of customized SD models trained on watermarked images and are easily distinguishable.

Accuracy (%)	Recall (%)	Precision (%)
97.30	95.67	99.04

Table 3: Quantitative results via the naive ViT classifier.

The classifier we use is vision transformer (ViT) (Dosovitskiy et al., 2021) with a binary classification head; the training dataset of the classifier images generated by customized SD v1.4, v1.5, and v2.1 trained with/without watermarks. We report the performance of our classifier in Tab. 3. The

[3] https://huggingface.co/stabilityai/sd-vae-ft-ema

[4] huggingface.co/docs/diffusers/main/en/training/dreambooth

Method	VAE	WikiArt	DreamBooth
LoRA	**base**	99.93	98.70
	ft-ema	99.93	99.33
	ft-mse	99.05	99.90
DreamBooth	**base**	99.77	97.63
	ft-ema	99.47	97.63
	ft-mse	99.85	97.90

Table 4: Generalization on different VAE. We report the SR (%) of DUAW trained on the **ft-mse** VAE and tested with different VAEs.

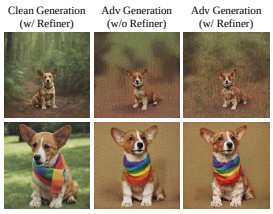

Figure 3: DUAW results on SDXL 1.0 model.

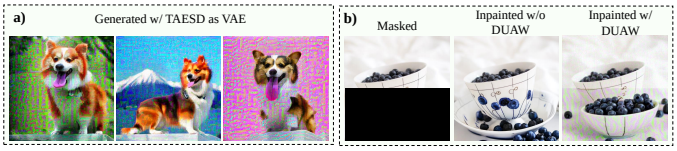

Figure 4: Results on a) TAESD; b) Asymmetric VQGAN.

results reveal that a simple ViT classifier can achieve remarkable accuracy, recall, and precision, effectively distinguishing between the images generated by the SD models trained on watermarked and clean output images.

D Fine-tuning Methods Tend to Preserve Minor Perturbations

As shown in Fig. 6, SD model fine-tuning methods (*e.g.*, DreamBooth and LoRA) tend to retain minor perturbations added to training images. We apply tiny checkerboard-like and stripe perturbations with a small upper bound (0.03) to the training images. After fine-tuning, both DreamBooth and LoRA obviously preserve and accentuate the pre-added perturbations in the generated images.

E Fine-tuning Settings

Subject-driven generation For the DreamBooth method, we use a batch size of 1 and a learning rate of $5e^{-6}$ and fine-tuned the UNet of the SD model with 200 class images for 800 epochs. The instance prompt is set to "a photo of sks <class name>" (*e.g.*, a photo of sks cat), and the class prompt is set to "a photo of <class name>" (*e.g.*, a photo of cat). As for the LoRA method, we employ the BLIP-2 (Li et al., 2023) model to generate captions for all images and add the "sks" identifier to the captions (*e.g.*, a photo of a sks dog in the grass). We fine-tune the SD model for 1500 epochs, with a batch size of 1 and a learning rate of $1e^{-4}$.

Style-driven generation For the DreamBooth method, we fine-tune the UNet and text-encoder of the SD model using 1000 class images for 800 epochs. The instance prompt is set to "sks style", and the class prompt is set to "art style"; For the LoRA method, we use the BLIP-2 model to generate captions for the images, adding the "in sks style" identifier to the captions. The fine-tuning of the SD model is performed for 1200 epochs.

F Training Dataset

To synthesize the training dataset for the data-free scenario, we generate 10 prompts of painting contents using ChatGPT and combine each prompt with different artistic styles to generate images with high diversity. We use 10 distinct styles of ArtBench (Liao et al., 2022) to form the style zoo, which encompasses a wide range of artistic periods and genres. The prompts and styles are shown in Tab. 5. Each prompt can combine with every available style to form a final prompt for SD

Prompts	Styles
A quiet forest scene with a babbling brook and sunlight filtering through the trees.	Baroque
A lion standing on a sand dune.	Impressionism
An old man sleeping in a dimly lit room.	Surrealism
A deserted beach with crashing waves and a sunset in the distance.	Art Nouveau
A pair of goldfish swimming in a spherical fish tank.	Post Impressionism
A surreal landscape with floating islands and a rainbow-colored sky.	Expressionism
A quaint little cabin nestled in the woods, with a thatched roof and stone chimney.	Realism
A newlywed couple getting married in a beautiful church.	Renaissance
A craft room with a sewing machine, a work table, and a dog lying on a cushion.	Romanticism
A noblewoman sitting in a luxurious garden, surrounded by flowers and fountains.	Ukiyo-e

Table 5: Prompts and styles for training dataset generation.

Figure 5: Samples from the synthesized dataset.

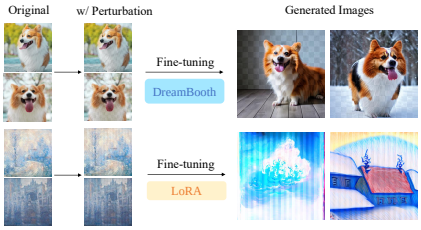

Figure 6: Examples of fine-tuned SD models preserving minor perturbations.

generation. Utilizing SD v2.1, we generate a dataset containing 100 images. A subset of the dataset is showcased in Fig. 5.

G Evaluation Dataset and Prompts for Style-driven Generation

We choose a subset of the WikiArt dataset, a comprehensive collection of artworks from various artists, to evaluate DUAW on style-driven generation tasks. We selected 12 artists from the WikiArt dataset, each artist contributing 10 artworks with a resolution higher than 512×512. The selected artworks from the 12 artists are showcased in Fig. 7.

To assess the effectiveness of our DUAW on this dataset, we conducted evaluations using 25 painting prompts generated by ChatGPT. Samples of the prompts are listed in Tab. 6.

Evaluation prompt
A solitary tree standing against a sunset.
A serene beach with gentle waves and seashells.
A house in the snow.
A misty morning in a dense forest with towering trees.
A painting of some cloud floating in a clear blue sky.
A rustic wooden bridge over a calm river.
A playful kitten batting at a ball of yarn on a cozy blanket.
A full moon illuminating a peaceful nighttime cityscape.
A group of wildflowers blooming in a meadow.

Table 6: Evaluation prompts. We have selected 25 prompts to evaluate the attack effectiveness of DUAW.

Figure 7: Selected subset of WikiArt dataset. We selected 12 artists, with each artist selecting 10 artworks.