ArtiFade: Learning to Generate High-quality Subject from Blemished Images

Shuya Yang* Shaozhe Hao* Yukang Cao† Kwan-Yee K. Wong†
The University of Hong Kong

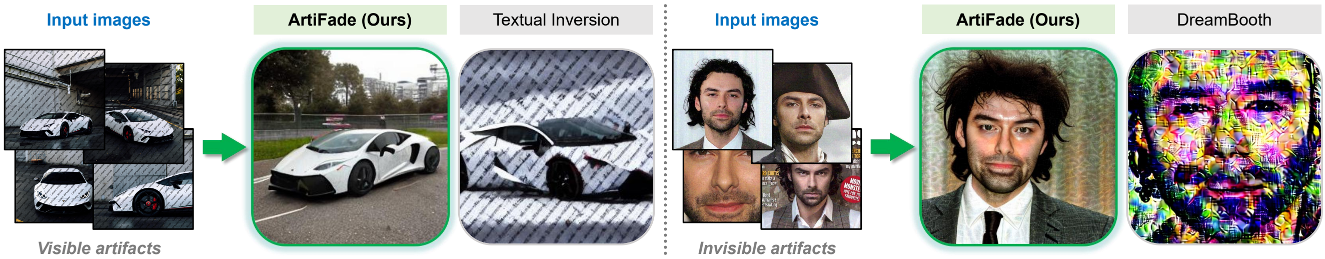


Figure 1. Blemished subject-driven generation with our ArtiFade and vanilla subject-driven methods. We display images generated using ArtiFade and Textual Inversion [15] on watermark artifacts on the left, and ArtiFade and DreamBooth [44] on adversarial noise artifacts [48] on the right. In contrast to the poor performance of Textual Inversion and DreamBooth, which are negatively affected by the visible or invisible artifacts, ArtiFade produces much better fidelity of the subject with high-quality generation.

Abstract

Subject-driven text-to-image generation has demonstrated remarkable advancements in its ability to learn and capture characteristics of a subject using only a limited number of images. However, existing methods commonly rely on high-quality images for training and often struggle to generate reasonable images when the input images are blemished by artifacts. This is primarily attributed to the inadequate capability of current techniques in distinguishing subject-related features from disruptive artifacts. In this paper, we introduce ArtiFade to tackle this issue and successfully generate high-quality artifact-free images from blemished datasets. Specifically, ArtiFade exploits fine-tuning of a pre-trained text-to-image model, aiming to remove artifacts. The elimination of artifacts is achieved by utilizing a specialized dataset that encompasses both unblemished images and their corresponding blemished counterparts during fine-tuning. ArtiFade also ensures the preservation of the original generative capabilities inherent within the diffusion model, thereby enhancing the overall performance of subject-driven methods in generating high-quality and artifact-free images. We further devise evaluation benchmarks tailored for this task. Through extensive qualitative and quantitative experiments, we demonstrate the generalizability of ArtiFade in effective artifact removal under both in-distribution and out-of-distribution scenarios.

1. Introduction

With the rapid advancement of generative diffusion models [20, 41, 45, 47, 60], subject-driven text-to-image generation [5, 15, 25, 27, 44], which aims to capture distinct characteristics of a subject by learning from a few images of the subject, has gained significant attention. This approach empowers individuals to seamlessly incorporate their preferred subjects into diverse and visually captivating scenes by simply providing text conditions. Representative works such as Textual Inversion [15] and DreamBooth [44] have shown promising results on this task. Specifically, Textual Inversion proposes to optimize a textual embedding to encode identity characteristics that provide rich subject information for subsequent generation. DreamBooth shares a similar idea but additionally fine-tunes the diffusion model to preserve more identity semantics. Plenty of successive efforts have been made to advance this task from various perspectives, including generation quality, compositionality, and efficiency [5, 25, 27].

Both of the above mentioned methods, along with their follow-up works, however, rely heavily on the presence of unblemished input images that contain only relevant identity information. This is often expensive or even unavailable in real-world applications. Instead, in practical scenarios such as scraping web images of a desired subject, it is common to encounter images that are blemished by various *visible* artifacts such as watermarks, drawings, and stickers. Additionally, there also exist *invisible* artifacts like adversarial noises [48] that are not easily detectable or remov-

[*] Equal contribution [†] Corresponding authors

able using off-the-shelf tools. These artifacts can significantly impede the comprehensive learning of the subject and lead to a catastrophic decline in performance across multiple dimensions (see Fig. 1). This limitation arises from the feature confusion inherent in the existing subject-driven learning process. The process simultaneously captures subject-related features and disruptive artifact interference. It lacks the discriminative power to distinguish these two from each other, and fails to preserve the integrity of subject characteristics while mitigating negative effects caused by artifacts. As blemished inputs are inevitable in applications, a pressing challenge emerges: **Can we effectively perform subject-driven text-to-image generation using *blemished* images?** We term this novel problem (*i.e.*, generating subject-driven images from blemished inputs) as blemished subject-driven generation in this paper.

To answer the above question, we present **ArtiFade**, the first model to tackle blemished subject-driven generation by adapting vanilla subject-driven methods (*e.g.*, Textual Inversion [15] and DreamBooth [44]) to effectively extract subject-specific information from blemished training data. The key objective of ArtiFade is to learn the implicit relationship between natural images and their blemished counterparts through alignment optimization. Specifically, we introduce a specialized dataset construction method to create pairs of unblemished images and their corresponding counterparts. These pairs can be applied to fine-tune various subject-driven approaches in the context of blemished subject-driven generation. Besides, we also observe fine-tuning an extra learnable embedding in the textual space, named artifact-free embedding, can enhance prompt fidelity in the blemished subject-driven generation.

We further introduce an evaluation benchmark that encompasses **(1)** multiple test sets of blemished images with diverse artifacts, and **(2)** tailored metrics for accurately assessing the performance of blemished subject-driven generation methods. A thorough experimental evaluation shows that our method consistently outperforms other existing methods, both qualitatively and quantitatively. Notably, ArtiFade exhibits superb capabilities in handling out-of-distribution (OOD) scenarios involving diverse types of artifacts that are distinct from the training data. This inherent generalizability indicates our model can effectively learn to discern and distinguish the patterns exhibited by artifacts and unblemished images, instead of overfitting to a specific type of artifacts.

In summary, our key contributions are as follows:

- We are the first to tackle the novel challenge of blemished subject-driven generation. To address this task, we propose ArtiFade that fine-tunes diffusion models to align unblemished and blemished data.
- We introduce an evaluation benchmark tailored for effectively assessing the performance of blemished subject-driven generation techniques.
- We conduct extensive experiments and demonstrate that ArtiFade outperforms current methods significantly. We show noteworthy generalizability of ArtiFade, effectively addressing both in-distribution and out-of-distribution scenarios with various types of artifacts.

2. Related work

Text-to-image synthesis Text-to-image generation has attracted considerable attention in recent years by leveraging Generative Adversarial Networks (GANs) [16] and diffusion models [20, 41]. Reed *et al.* [40] were the first to integrate GANs into text-to-image generation. Since then, several influential works had been proposed [8, 29, 36, 43, 54–58, 61, 62], demonstrating impressive results with improved resolution [56, 57] and fidelity of fine details [54]. Diffusion models in text-to-image synthesis have also yielded remarkable results owing to their ability in generating precise and customized images that better align with individual text specifications [17, 35, 39, 41, 45].

Subject-driven generation Subject-driven generation has gained popularity due to its ability to generate personalized images based on a given set of subject images and text prompts. One prominent method in subject-driven generation is Textual Inversion [15], which involves learning an embedding vector by minimizing the Latent Diffusion Model loss [41] on input images. The learned embedding vector can be effectively combined with text prompts, allowing seamless integration in the text-to-image generation process. Recent approaches [27, 33, 44] have significantly enhanced subject reconstruction fidelity by incorporating fine-tuning techniques.

Artifacts removal Shadow and watermark removal are classic tasks in image processing and computer vision. At the early stage, most approaches for shadow removal or image recovery relied on the properties of intensity and illumination [1, 11–13, 19, 26, 46, 51, 52, 59]. Some methods also incorporated color features to improve their results [19]. Deep learning techniques and Convolutional Neural Networks (CNNs) have played a significant role in advancing shadow removal methods and producing impressive results in recent years [6, 10, 14, 22, 24, 28, 32, 50, 63]. Several studies [10, 22, 32, 50] have incorporated GANs to further enhance the results of shadow removal techniques. Moreover, with the increasing popularity of diffusion models in image generation, a novel diffusion-based method for shadow removal has recently been introduced [18]. The most widely adopted methods for recovering concealed information from watermarked images include the application of generalized multi-image matting algorithms [9], complemented by image inpainting techniques [23, 37, 53], and the utilization of deep neural networks and CNNs [7]. Similar

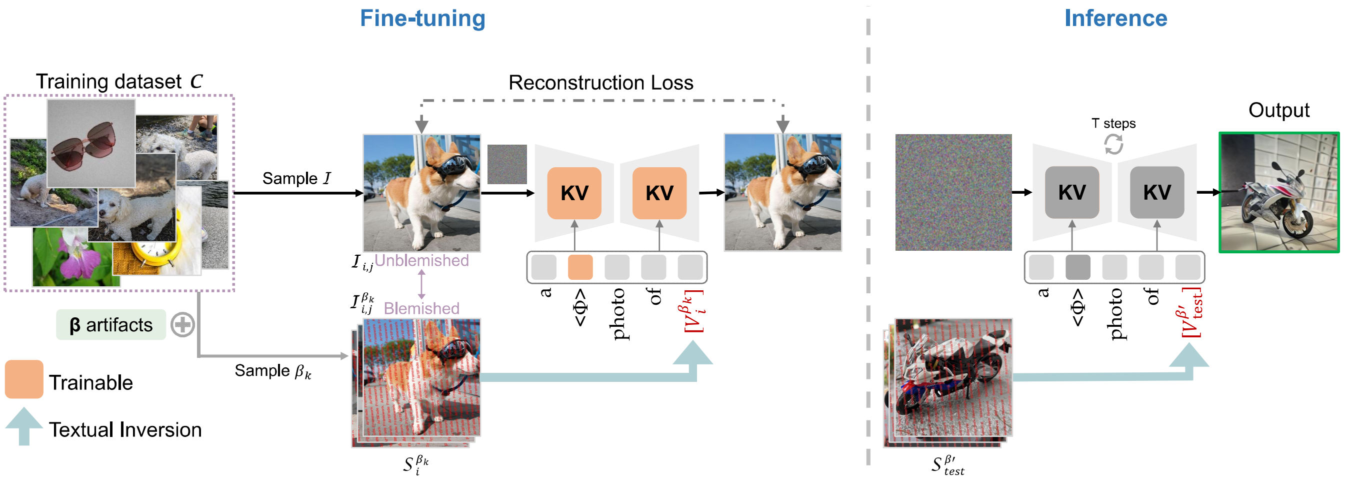


Figure 2. Overview of ArtiFade. On the left, we present artifact rectification training, which involves an iterative process of calculating reconstruction loss between an unblemished image and the reconstruction of its blemished embedding. The right-hand side is the inference stage that tests ArtiFade on unseen blemished images. To avoid ambiguity, we (1) simplify the training of Textual Inversion into an input-output form, and (2) use "fine-tuning" and "inference" to respectively refer to the fine-tuning stage of ArtiFade and the use of ArtiFade for subject-driven generation.

to shadow removal, GANs and Conditional GANs [34] are also widely used in watermark removal tasks [3, 30, 31]. Our work is closely related to these previously mentioned studies. We are the first to address the artifact issues in the realm of subject-driven text-to-image generation.

3. Method

Given a set of blemished input images, our objective is to eliminate their negative impacts on the quality of subject-driven image generation. To achieve this goal, we present ArtiFade, an efficient framework that learns to discern and distinguish the patterns exhibited by various types of artifacts and unblemished images. In this section, we focus exclusively on ArtiFade based on Textual Inversion. However, it is important to note that the ArtiFade framework can be generalized to other subject-driven generation methods. As shown in Fig. 2, ArtiFade based on Textual Inversion incorporates two main components, namely the fine-tuning of the partial parameters (*i.e.*, key and value weights) in the diffusion model and the simultaneous optimization of an artifact-free embedding $\langle\Phi\rangle$. We begin by discussing the preliminaries of the Latent Diffusion Model and Textual Inversion. In Sec. 3.1, we elaborate our automatic construction of the training dataset, which consists of both blemished and unblemished data. We then introduce Artifact Rectification Training, a method for fine-tuning the model to accommodate blemished images, in Sec. 3.2. We finally present the use of ArtiFade for handling blemished images in Sec. 3.3.

Preliminary Latent Diffusion Model (LDM) [41] is a latent text-to-image diffusion model derived from Diffusion Denoising Probabilistic Model (DDPM) [20]. LDM leverages a pre-trained autoencoder to map image features between the image and latent space. This autoencoder comprises an encoder $\mathcal{E}$, which transforms images into latent representations, and a decoder $\mathcal{D}$, which converts latent representations back into images. The autoencoder is optimized using a set of images so that the reconstructed image $\hat{x} = \mathcal{D}(\mathcal{E}(x)) \approx x$. Additionally, LDM introduces cross-attention layers [49] within the U-Net [42], enabling the integration of text prompts as conditional information during the image generation process. The LDM loss is defined as

$$\mathcal{L}_{LDM} := \mathbb{E}_{z\sim\mathcal{E}(\mathcal{I}),y,\epsilon\sim N(0,1)}\Big[\|\epsilon - \epsilon_\theta(z_t,t,y)\|_2^2\Big], \quad (1)$$

where $\mathcal{E}$ encodes the image $\mathcal{I}$ into the latent representation z. Here, z_t denotes the noisy latent representation at timestep t, ϵ_θ refers to the denoising network, and y represents the text condition that is passed to the cross-attention layer.

Based on LDM, Textual Inversion [15] aims to capture the characteristics of a specific subject from a small set of images. Specifically, Textual Inversion learns a unique textual embedding by minimizing Eq. (1) on a few images that contain the particular subject. It can produce promising generation results with high-quality inputs, but fails on input images that are blemished by artifacts (see Fig. 1). This problem arises from the inherent limitation of Textual Inversion in learning shared characteristics exhibited in the input images without the capability in differentiating artifacts from unblemished subjects. In this paper, we aim to address this issue on deteriorated generation quality of Textual Inversion in the presence of blemished images.

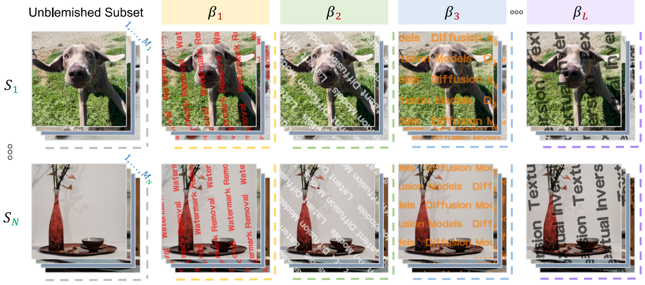


Figure 3. Examples of training dataset $\mathcal{D}$ that contains both unblemished images and blemished counterparts.

3.1. Dataset construction

Existing subject-driven generation methods operate under the assumption of unblemished training data, consisting of solely high-quality images devoid of any artifacts. However, this assumption does not align with real-world applications, where obtaining blemished images from the internet is a commonplace. To address this blemished subject-driven generation in this paper, we first construct a training set that incorporates both unblemished images and their blemished counterparts that are augmented with artifacts.

Augmentation of multiple artifacts We construct our dataset by collecting a multi-subject set $\mathcal{C}$ of N image subsets from existing works [15, 27, 44] and a set $\mathcal{B}$ of L different artifacts:

$$\mathcal{C} = \{\mathcal{S}_i\}_{i=1}^{N}, \quad \mathcal{S}_i = \{\mathcal{I}_{i,j}\}_{j=1}^{M_i}, \quad \mathcal{B} = \{\beta_k\}_{k=1}^{L}, \tag{2}$$

where $\mathcal{S}_i$ denotes the image subset corresponding to the ith subject, M_i is the total number of images in $\mathcal{S}_i$, and β_k represents a type of artifact for image augmentation. Our dataset $\mathcal{D}$ can then be constructed by applying each artifact β_k to each image $\mathcal{I}$ in $\mathcal{S}_i$ separately, *i.e.*,

$$\mathcal{S}_i^{\beta_k} = \{\mathcal{I}_{i,j}^{\beta_k}\}_{j=1}^{M_i}, \quad \mathcal{D} = \{\mathcal{S}_i, \{\mathcal{S}_i^{\beta_k}\}_{k=1}^{L}\}_{i=1}^{N}, \tag{3}$$

where $\mathcal{I}_{i,j}^{\beta_k}$ is the counterpart of $\mathcal{I}_{i,j}$ augmented with the specific artifact β_k. Some examples of original images and their augmented versions with distinct artifacts can be found in Fig. 3. See Appendix for more examples.

Blemished textual embedding For each blemished subset, we perform Textual Inversion to optimize a blemished textual embedding $[\mathtt{V}_i^{\beta_k}]$, *i.e.*,

$$\begin{gathered}\mathcal{S}_i^{\beta_k} \xrightarrow{\text{Textual Inversion}} [\mathtt{V}_i^{\beta_k}], \\ i = 1, 2, ..., N; \quad k = 1, 2, ..., L\end{gathered} \tag{4}$$

By applying Eq. (4) on N subsets with L types of artifacts, we end up with a set of $N \times L$ blemished textual embeddings $\mathcal{V} = \{[\mathtt{V}_i^{\beta_k}]\}_{i=1,k=1}^{N,L}$, which will be used in the subsequent model fine-tuning. As we have illustrated in Fig. 1, directly prompting the diffusion model with $[\mathtt{V}_i^{\beta_k}]$ will lead to a significant decrease in generation quality. Consequently, our objective is to robustly handle blemished embeddings and effectively eliminate the detrimental impact of artifacts. We achieve this by devising a partial fine-tuning paradigm for the pre-trained diffusion model on the constructed training set $\mathcal{D}$, as elaborated in the following subsection.

3.2. Artifact rectification training

After establishing the curated dataset $\mathcal{D}$, we embark on training a generalizable framework on $\mathcal{D}$, capable of generating unblemished images using blemished textual embeddings. To this end, we propose artifact rectification training, which consists of two key components, namely partial fine-tuning of a pre-trained diffusion model and the optimization of an artifact-free embedding, to eliminate the artifacts and distortions in the generated images.

We fine-tune partial parameters related to the attention modules, including those involved in processing the textual conditions. This strategy allows us to optimize the relevant components associated with the blemished textual embedding $[\mathtt{V}_i^{\beta_k}]$. Considering that only the key and value weights in the diffusion model's cross-attention layer are involved in the processing of textual embedding, we choose to fine-tune these two types of parameters. To further enhance the model's stability when handling blemished content, we also fine-tune the key and value weights in the self-attention layers. In short, we fine-tune key and value weights (*i.e.*, W^k and W^v) across all attention modules. Moreover, we find that optimizing an additional embedding, $\langle\Phi\rangle$, in the textual space with partial parameters could improve prompt fidelity by retaining the textual information of the model, as presented later in Sec. 4.5.

Training objective During each iteration, we will first randomly sample an unblemished image $\mathcal{I}_{i,j}$ from the training set $\mathcal{D}$ and a type of artifact $\beta_k \in \mathcal{B}$ to obtain the blemished textual embedding $[\mathtt{V}_i^{\beta_k}] \in \mathcal{V}$ that is optimized on the blemished subset $\mathcal{S}_i^{\beta_k}$.

Specifically, given the sampled blemished textual embedding $[\mathtt{V}_i^{\beta_k}]$, we form the prompt "a $\langle\Phi\rangle$ photo of $[\mathtt{V}_i^{\beta_k}]$", which will be input to the text encoder to acquire the text condition $y_i^{\beta_k}$. Our optimization objective will then be defined as reconstructing the unblemished image $\mathcal{I}_{i,j}$ by conditioning the denoising process on the text condition $y_i^{\beta_k}$. Thus, we can formulate the final loss for training ArtiFade as

$$\begin{aligned}\mathcal{L}_{\text{ArtiFade}} :=& \mathbb{E}_{z\sim\mathcal{E}(\mathcal{I}_{i,j}), y_i^{\beta_k}, \epsilon\sim N(0,1)} \\ & \Big[\|\epsilon - \epsilon_{\{W^k, W^v, \langle\Phi\rangle\}}(z_t, t, y_i^{\beta_k})\|_2^2\Big],\end{aligned} \tag{5}$$

where $\{W^k, W^v, \langle\Phi\rangle\}$ is the set of the trainable parameters of ArtiFade.

Table 1. Quantitative results - ID.

Method	WM-model on WM-ID-test				
	$I^{DINO}\uparrow$	$R^{DINO}\uparrow$	$I^{CLIP}\uparrow$	$R^{CLIP}\uparrow$	$T^{CLIP}\uparrow$
TI (unblemished)	0.488	1.349	0.730	1.070	0.283
TI (blemished)	0.217	0.852	0.576	0.909	0.263
Ours (TI-based)	**0.337**	**1.300**	**0.649**	**1.020**	**0.282**

Table 2. Quantitative results - OOD.

Method	WM-model on WM-OOD-test				
	$I^{DINO}\uparrow$	$R^{DINO}\uparrow$	$I^{CLIP}\uparrow$	$R^{CLIP}\uparrow$	$T^{CLIP}\uparrow$
TI (unblemished)	0.488	1.278	0.730	1.136	0.283
TI (blemished)	0.229	0.858	0.575	0.929	0.262
Ours (TI-based)	**0.356**	**1.237**	**0.654**	**1.079**	**0.282**

3.3. Subject-driven generation with blemished images

After artifact rectification training, we obtain the ArtiFade model, prepared for the task of blemished subject-driven generation. Given a test image set $\mathcal{S}^{\beta'}_{\text{test}}$ in which all images are blemished by an arbitrary artifact β', the ArtiFade model can generate high-quality subject-driven images using blemished samples with ease.

Specifically, we first obtain the blemished textual embedding $[\mathtt{V}^{\beta'}_{\text{test}}]$ by applying Textual Inversion on the test set $\mathcal{S}^{\beta'}_{\text{test}}$. We then simply infer the ArtiFade model with a given text prompt that includes the blemished textual embedding, *i.e.*, "a $\langle\Phi\rangle$ photo of $[\mathtt{V}^{\beta'}_{\text{test}}]$". At the operational level, the sole distinction between our approach and vanilla Textual Inversion lies in inputting text prompts containing $[\mathtt{V}^{\beta'}_{\text{test}}]$ into the fine-tuned ArtiFade instead of the pre-trained diffusion model. This simple yet effective method resolves the issue of Textual Inversion's incapacity to handle blemished input images, bearing practical utility.

Details of ArtiFade models We choose $N = 20$ subjects, including pets, plants, containers, toys, and wearable items to ensure a diverse range of categories. We experiment with the ArtiFade model based on Textual Inversion trained with visible watermark artifacts, namely `WM-model`. The training set of `WM-model` involves $L_{\text{WM}} = 10$ types of watermarks, characterized by various fonts, orientations, colors, sizes, and text contents. Therefore, we obtain 200 blemished subsets in total within the training set of `WM-model`. We fine-tune `WM-model` for a total of 16k steps.

4. Experiment

4.1. Implementation details

We employ the pre-trained LDM [41] following the official implementation of Textual Inversion [15] as our base diffusion model. We train the blemished textual embeddings for 5k steps using Textual Inversion. We use a learning rate of 5e-3 to optimize our artifact-free embedding and 3e-5 for the partial fine-tuning of key and value weights. Note that all other parameters within the pre-trained diffusion model remain frozen. All experiments are conducted on 2 NVIDIA RTX 3090 GPUs. In the main paper, we focus on the comparison with Textual Inversion and DreamBooth to demonstrate the efficiency of our proposed contributions. See Appendix for additional comparisons and applications.

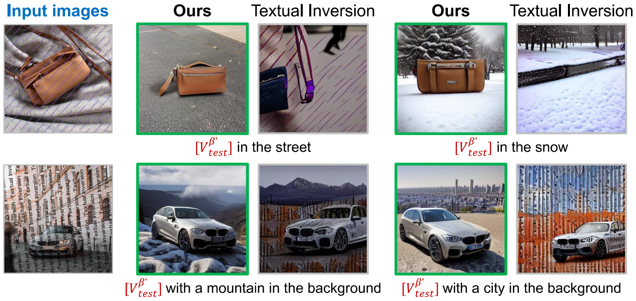

Input images
Ours
Textual Inversion
Ours
Textual Inversion
$[V^{\beta'}_{test}]$ in the street
$[V^{\beta'}_{test}]$ in the snow
$[V^{\beta'}_{test}]$ with a mountain in the background
$[V^{\beta'}_{test}]$ with a city in the background

Figure 4. Qualitative Comparison - ID. Unlike Textual Inversion which struggles to produce reasonable generation from blemished inputs, our method (`WM-model`) consistently learns the distinguished features of the given subject and achieves high-quality generation without distortion.

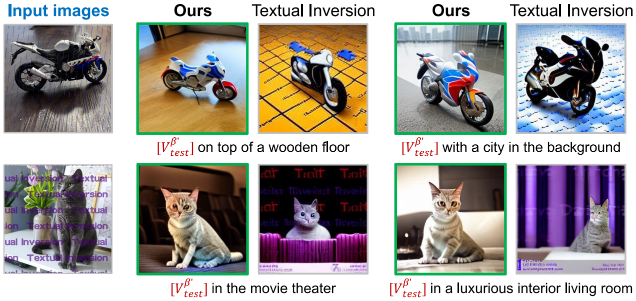

Input images
Ours
Textual Inversion
Ours
Textual Inversion
$[V^{\beta'}_{test}]$ on top of a wooden floor
$[V^{\beta'}_{test}]$ with a city in the background
$[V^{\beta'}_{test}]$ in the movie theater
$[V^{\beta'}_{test}]$ in a luxurious interior living room

Figure 5. Qualitative Comparison - OOD. Our method (`WM-model`) is generalizable to process out-of-distribution artifacts that are unseen during the fine-tuning, demonstrating much better performance than Textual Inversion. Best viewed in PDF with zoom.

4.2. Evaluation benchmark

Test dataset We construct the test dataset using 16 novel subjects that differ from those in the training set. These subjects encompass a wide range of categories, including pets, plants, toys, transportation, furniture, and wearable items. We group the visible test artifacts into two categories: (1) in-distribution watermarks (`WM-ID-test`) (*i.e.*, watermark types same as the training data), and (2) out-of-distribution watermarks (`WM-OOD-test`) (*i.e.*, watermark types different from the training data). Within the `WM-ID-test` and `WM-OOD-test`, we synthesize 5 distinct artifacts for each category, resulting in 80 test sets.

Evaluation metrics We evaluate the performance of blemished subject-driven generation from three perspectives: (1) the fidelity of subject reconstruction, (2) the fidelity of text conditioning, and (3) the effectiveness of mitigating the negative impacts of artifacts. Following common practice [15, 44], we use CLIP [38] and DINO [4] similarities for measuring these metrics. For the first metric, we calculate the CLIP and DINO similarities between the generated images and the unblemished version of the input images, denoted as I^{CLIP} and I^{DINO} respectively. For the second metric, we calculate the CLIP similarity between the generated images and the text prompt, denoted as T^{CLIP}. For the third metric, we calculate the relative ratio of similarities between generated images and unblemished input images compared to their blemished versions, defined as

$$R^{CLIP} = I^{CLIP}/I^{CLIP}_{\beta} \quad R^{DINO} = I^{DINO}/I^{DINO}_{\beta} \tag{6}$$

where I^{CLIP}_{β} and I^{DINO}_{β} denote CLIP and DINO similarities between the generated images and the *blemished* input images respectively. A relative ratio greater than 1 indicates that generated images resemble unblemished images more than blemished counterparts, suggesting fewer artifacts. Conversely, a ratio less than 1 indicates that generated images are heavily distorted with more artifacts. We use DINO ViT-S/16 [4] and CLIP ViT-B/32 [38] to compute all metrics.

4.3. ArtiFade with Textual Inversion

4.3.1. Quantitative comparisons

We conduct both in-distribution and out-of-distribution quantitative evaluations of our method and compare it to Textual Inversion with blemished embeddings. We additionally report the results using Textual Inversion on unblemished images as a reference, although it is not a direct comparison to our model.

In-distribution (ID) analysis We consider the in-distribution scenarios by testing `WM-model` on `WM-ID-test`. In Tab. 1, we can observe that the use of blemished embeddings in Textual Inversion leads to comprehensive performance decline including: (1) lower subject reconstruction fidelity (*i.e.*, I^{DINO} and I^{CLIP}) due to the subject distortion in image generation; (2) lower efficiency for artifact removal (*i.e.*, R^{DINO} and R^{CLIP}) due to inability to remove artifacts; (3) lower prompt fidelity (*i.e.*, T^{CLIP}) since the prompt-guided background is unrecognizable due to blemishing artifacts. In contrast, our method consistently achieves higher scores than Textual Inversion with blemished embeddings across the board, demonstrating the efficiency of ArtiFade in various aspects.

Table 3. Quantitative comparison with DreamBooth.

Method	WM-ID-test				
	$I^{DINO}\uparrow$	$R^{DINO}\uparrow$	$I^{CLIP}\uparrow$	$R^{CLIP}\uparrow$	$T^{CLIP}\uparrow$
TI (unblemished)	0.488	1.349	0.730	1.070	0.283
TI (blemished)	0.217	0.852	0.576	0.909	0.263
DB (blemished)	0.503	0.874	0.738	0.939	0.272
Ours (TI-based)	0.337	1.300	0.649	1.020	0.282
Ours (DB-based)	**0.589**	**1.308**	**0.795**	**1.083**	**0.284**

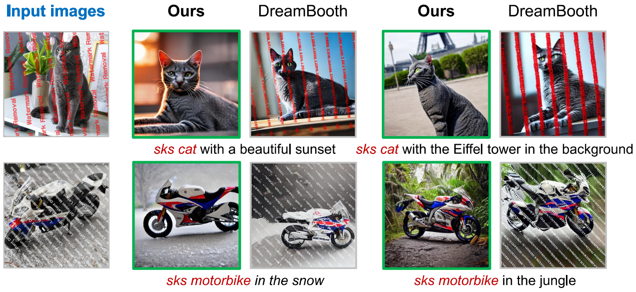

Input images
Ours
DreamBooth
Ours
DreamBooth
sks cat with a beautiful sunset
sks cat with the Eiffel tower in the background
sks motorbike in the snow
sks motorbike in the jungle

Figure 6. Qualitative comparison with DreamBooth.

Out-of-distribution (OOD) analysis We pleasantly discover that `WM-model` possesses the capability to handle out-of-distribution scenarios, owing to its training with watermarks of diverse types. We consider the out-of-distribution (OOD) scenarios for `WM-model` by testing it on `WM-OOD-test`, as presented in Tab. 2. Similar to ID evaluation, all of our metrics yield higher results than Textual Inversion with blemished embeddings. These results further demonstrate the generalizability of our method.

4.3.2. Qualitative comparisons

We present qualitative comparisons between the output generated by ArtiFade and Textual Inversion with blemished textual embeddings, including in-distribution scenarios in Fig. 4 and out-of-distribution scenarios in Fig. 5.

In-distribution analysis The images generated by Textual Inversion exhibit noticeable limitations when using blemished textual embeddings. Specifically, as depicted in Fig. 4, all rows predominantly exhibit cases of incorrect backgrounds that are highly polluted by watermarks. By using ArtiFade, we are able to eliminate the background watermarks.

Out-of-distribution analysis In addition, we conduct experiments with our `WM-model` to showcase its capability to remove out-of-distribution watermarks, as shown in Fig. 5. It is important to note that in the first row, the watermark in the input images may not be easily noticed by human eyes upon initial inspection due to the small font size and high image resolution. However, these artifacts have a significant effect when used to train blemished embeddings for generating images. ArtiFade effectively eliminates the artifacts on the generated images, improving reconstruction

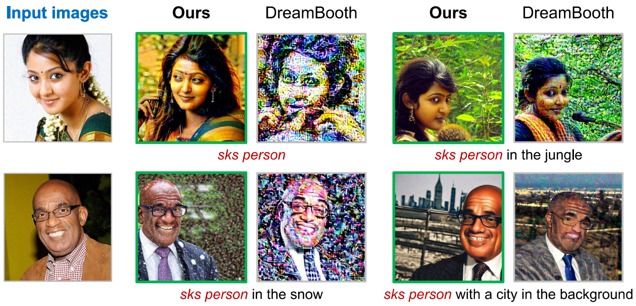


Figure 7. Qualitative Comparison between ours and DreamBooth when inputs are blemished by invisible adversarial noises.

fidelity and background accuracy, hence leading to substantial enhancements in overall visual quality.

4.4. ArtiFade with DreamBooth

4.4.1. Visible artifacts blemished subject generation

The ArtiFade fine-tuning framework is not limited to Textual Inversion with textual embedding, it can also be generalized to DreamBooth. We use the same training dataset and blemished subsets as in the case of the WM-model (*i.e.*, $N = 20$, $L_{WM}= 10$). The vanilla DreamBooth fine-tunes the whole UNet model, which conflicts with the fine-tuning parameters of ArtiFade. We therefore use DreamBooth with low-rank approximation (LoRA)[1] to train LoRA adapters [21] for the text encoder, value, and query weights of the diffusion model for each blemished subset using Stable Diffusion v1-5. For simplicity, we will use DreamBooth to refer to DreamBooth with LoRA below. During the fine-tuning of DreamBooth-based ArtiFade, we load the pre-trained adapters and only unfreeze key weights since value weights are reserved for DreamBooth subject information. In Tab. 3, it is evident that our method, based on DreamBooth, yields the highest scores among all cases. Our method also maintains DreamBooth's advantages in generating images with higher subject fidelity and more accurate text prompting, outperforming ArtiFade with Textual Inversion. We show some qualitative results in Fig. 6.

4.4.2. Invisible artifacts blemished subject generation

ArtiFade demonstrates exceptional performance in handling subjects characterized by intricate features and blemished by imperceptible artifacts. We collect 20 human figure datasets from the VGGFace2 dataset [2]. We then use the Anti-DreamBooth [48] ASPL method to add adversarial noises to each group of images, producing 20 blemished datasets for fine-tuning a DreamBooth-based ArtiFade model. The model is fine-tuned for 12k steps. As illustrated in Fig. 7, our approach surpasses the DreamBooth in differentiating the learning of adversarial noises from human face features. In contrast to DreamBooth, which is fooled into overfitting adversarial noises, thereby generating images with a heavily polluted background, our model reconstructs human figures in image generation while maintaining high fidelity through text prompting.

4.5. Ablation study

We conduct ablation study to demonstrate the efficiency of our method by comparing with three variants, namely (1) $\texttt{Var}_\texttt{A}$, where we solely fine-tune the artifact-free embedding; (2) $\texttt{Var}_\texttt{B}$, where we fine-tune parameters related to image features (*i.e.*, query weights W^q) along with the artifact-free embedding, and (3) $\texttt{Var}_\texttt{C}$, where we fine-tune key and value weights, *i.e.*, W^k and W^v, exclusively. We compare our WM-model with these variants by testing them on WM-ID-test.

Effect of partial fine-tuning As shown in Tab. 4, compared to $\texttt{Var}_\texttt{A}$, our full method yields higher scores on all metrics by a significant margin, except for R^{DINO}. This is reasonable as the artifact-free embedding can be easily overfitted to the training data, resulting in generated images that resemble a fusion of training images (Fig. 8, $\texttt{Var}_\texttt{A}$). As a result, the denominator of R^{DINO}, namely the similarity between the generated images and the blemished images, is significantly decreased, leading to a high R^{DINO}. Due to a similar reason, $\texttt{Var}_\texttt{A}$ shows lowest I^{DINO}, I^{CLIP}, and T^{CLIP} among all variants, indicating that it fails to reconstruct the correct subject. Overall, both quantitative and qualitative evaluation showcases that solely optimizing the artifact-free embedding is insufficient to capture the distinct characteristics presented in the blemished input image, demonstrating the necessity of partial fine-tuning.

Effect of fine-tuning key and value weights As shown in Tab. 4 and Fig. 8, $\texttt{Var}_\texttt{B}$ yields unsatisfactory outcomes in all aspects compared to ours. The lower R^{DINO} and R^{CLIP} suggest that the generated images retain artifact-like features and bear closer resemblances to the blemished subsets. Furthermore, the reduced T^{CLIP} indicates diminished prompt fidelity, as the approach fails to accurately reconstruct the subject from the blemished embeddings, which is also evidenced by Fig. 8. These findings suggest that fine-tuning the parameters associated with text features yields superior enhancements in terms of artifact removal and prompt fidelity.

Effect of the artifact-free embedding With $\texttt{Var}_\texttt{C}$, we exclude the optimization of artifact-free embedding. In Tab. 4, we can observe that $\texttt{Var}_\texttt{C}$ yields higher I^{DINO} and I^{CLIP} but lower R^{DINO} and R^{CLIP} compared to our WM-model, which indicates that the approach achieves higher subject

[1] https://huggingface.co/docs/peft/main/en/task_guides/dreambooth_lora

Table 4. Quantitative comparison of ablation study.

Method	W^{kv}	W^q	$\langle\Phi\rangle$	$\mathrm{I}^{\mathrm{DINO}}$	$\mathrm{R}^{\mathrm{DINO}}$	$\mathrm{I}^{\mathrm{CLIP}}$	$\mathrm{R}^{\mathrm{CLIP}}$	$\mathrm{T}^{\mathrm{CLIP}}$
$\mathtt{Var}_\mathtt{A}$			✓	0.154	**1.412**	0.566	0.984	0.265
$\mathtt{Var}_\mathtt{B}$		✓	✓	0.283	1.230	0.617	0.978	0.277
$\mathtt{Var}_\mathtt{C}$	✓			**0.342**	1.292	**0.652**	1.019	0.280
Ours	✓		✓	0.337	1.300	0.649	**1.020**	**0.282**

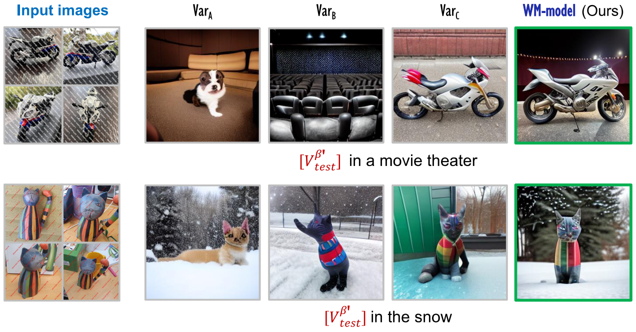

Input images
Var_A
Var_B
Var_C
WM-model (Ours)
$[V^{\beta'}_{test}]$ in a movie theater
$[V^{\beta'}_{test}]$ in the snow

Figure 8. Qualitative comparison of ablation study.

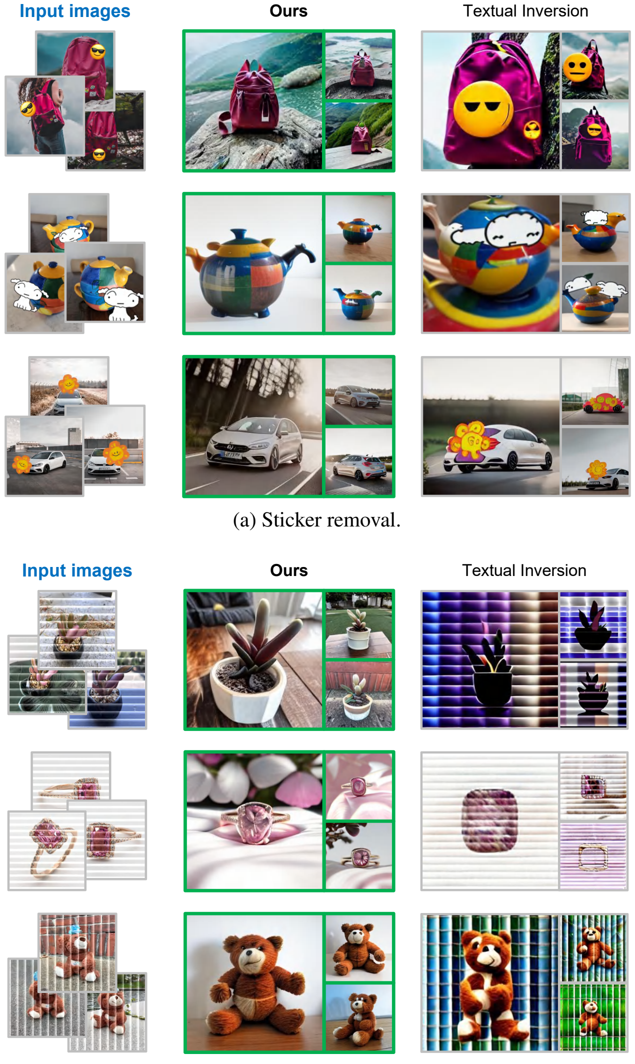
fidelity but lower efficiency in eliminating artifacts when generating images. Since our primary objective is to generate artifact-free images from blemished textual embedding, our `WM-model` chooses to trade off subject reconstruction fidelity for the ability to remove artifacts. Additionally, this approach produces lower $\mathrm{T}^{\mathrm{CLIP}}$ than ours, suggesting that the artifact-free embedding effectively improves the model's capability to better preserve text information (see Fig. 8).

5. More applications

We apply our `WM-model` to more artifact cases, such as stickers and glass effects, showcasing its broad applicability.

Sticker removal In Fig. 9a, we test `WM-model` on input images that are blemished by cartoon stickers. The cartoon sticker exhibits randomized dimensions and is positioned arbitrarily within each image. `WM-model` can effectively eliminate any stickers while concurrently addressing improper stylistic issues encountered during image generation.

Glass effect removal We further test `WM-model` on input images that are blemished by glass effect in Fig. 9b. We apply a fluted glass effect to images to replicate real-life scenarios where individuals capture photographs of subjects positioned behind fluted glass. This glass can have specific reflections and blurring, which may compromise the overall quality of image generation when using Textual Inversion. The use of our model can fix the distortions of the subjects and the unexpected background problem, significantly improving image quality.



(a) Sticker removal.

(b) Glass effect removal.

Figure 9. Applications. Our `WM-model` can be applied to remove various unwanted artifacts in the input images, *e.g.*, stickers and glass effect.

6. Conclusion

We introduce ArtiFade to address the novel problem of generating high-quality and artifact-free images in the blemished subject-driven generation. Our approach involves fine-tuning a diffusion model along with artifact-free embedding to learn the alignment between unblemished images and blemished information. We present an evaluation benchmark to thoroughly assess a model's capability in the task of blemished subject-driven generation. We demonstrate the effectiveness of ArtiFade in removing artifacts and addressing distortions in subject reconstruction under both in-distribution and out-of-distribution scenarios.

References

[1] Eli Arbel and Hagit Hel-Or. Shadow removal using intensity surfaces and texture anchor points. *IEEE TPAMI*, 33(6): 1202–1216, 2010. 2

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 7

[3] Zhiyi Cao, Shaozhang Niu, Jiwei Zhang, and Xinyi Wang. Generative adversarial networks model for visible watermark removal. *IET Image Processing*, pages 1783–1789, 2019. 3

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 6

[5] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. In *NeurIPS*, pages 30286–30305. Curran Associates, Inc., 2023. 1

[6] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Canet: A context-aware network for shadow removal. In *ICCV*, pages 4743–4752, 2021. 2

[7] Danni Cheng, Xiang Li, Wei-Hong Li, Chan Lu, Fake Li, Hua Zhao, and Wei-Shi Zheng. Large-scale visible watermark detection and removal with deep convolutional networks. In *PRCV*, pages 27–40, 2018. 2

[8] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. Rifegan: Rich feature generation for text-to-image synthesis from prior knowledge. In *CVPR*, pages 10911–10920, 2020. 2

[9] Tali Dekel, Michael Rubinstein, Ce Liu, and William T Freeman. On the effectiveness of visible watermarks. In *CVPR*, pages 2146–2154, 2017. 2

[10] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In *ICCV*, pages 10213–10222, 2019. 2

[11] G.D. Finlayson, S.D. Hordley, Cheng Lu, and M.S. Drew. On the removal of shadows from images. *IEEE TPAMI*, 28 (1):59–68, 2006. 2

[12] Graham D Finlayson, Steven D Hordley, and Mark S Drew. Removing shadows from images. In *ECCV*, pages 823–836, 2002.

[13] Graham D Finlayson, Mark S Drew, and Cheng Lu. Entropy minimization for shadow removal. *IJCV*, 85(1):35–57, 2009. 2

[14] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. In *CVPR*, pages 10571–10580, 2021. 2

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 1, 2, 3, 4, 5, 6

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[17] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, pages 10696–10706, 2022. 2

[18] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *CVPR*, pages 14049–14058, 2023. 2

[19] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Single-image shadow detection and removal using paired regions. In *CVPR*, pages 2033–2040, 2011. 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 1, 2, 3

[21] 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 *ICLR*, 2022. 7

[22] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *ICCV*, pages 2472–2481, 2019. 2

[23] Chun-Hsiang Huang and Ja-Ling Wu. Attacking visible watermarking schemes. *IEEE TMM*, 6(1):16–30, 2004. 2

[24] Yeying Jin, Ruoteng Li, Wenhan Yang, and Robby T Tan. Estimating reflectance layer from a single image: Integrating reflectance guidance and shadow/specular aware learning. In *AAAI*, pages 1069–1077, 2023. 2

[25] 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 *CVPR*, pages 6007–6017, 2023. 1

[26] Salman H Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Automatic shadow detection and removal from a single image. *IEEE TPAMI*, 38(3):431–446, 2015. 2

[27] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, pages 1931–1941, 2023. 1, 2, 4

[28] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *ICCV*, pages 8578–8587, 2019. 2

[29] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. In *NeurIPS*, 2019. 2

[30] Xiang Li, Chan Lu, Danni Cheng, Wei-Hong Li, Mei Cao, Bo Liu, Jiechao Ma, and Wei-Shi Zheng. Towards photo-realistic visible watermark removal with conditional generative adversarial networks. In *ICIG*, pages 345–356, 2019. 3

[31] Yang Liu, Zhen Zhu, and Xiang Bai. Wdnet: Watermark-decomposition network for visible watermark removal. In *WACV*, pages 3685–3693, 2021. 3

[32] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. In *CVPR*, pages 4927–4936, 2021. 2

[33] Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. In *CVPR*, pages 14267–14276, 2023. 2

[34] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014. 3

[35] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, pages 16784–16804, 2022. 2

[36] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *CVPR*, pages 1505–1514, 2019. 2

[37] Chuan Qin, Zhihong He, Heng Yao, Fang Cao, and Liping Gao. Visible watermark removal scheme based on reversible data hiding and image inpainting. *Signal Process. Image Commun.*, 60:160–172, 2018. 2

[38] 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 *ICML*, pages 8748–8763, 2021. 6

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2

[40] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, pages 1060–1069. PMLR, 2016. 2

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjórn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 3, 5

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 3

[43] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *ICCV*, pages 13960–13969, 2021. 2

[44] 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 *CVPR*, pages 22500–22510, 2023. 1, 2, 4, 6

[45] 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. *NeurIPS*, 35:36479–36494, 2022. 1, 2

[46] Yael Shor and Dani Lischinski. The shadow meets the mask: Pyramid-based shadow removal. *Comput. Graph. Forum*, 27 (2):577–586, 2008. 2

[47] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1

[48] 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 *ICCV*, pages 2116–2127, 2023. 1, 7

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998—6008, 2017. 3

[50] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, pages 1788–1797, 2018. 2

[51] Chunxia Xiao, Ruiyun She, Donglin Xiao, and Kwan-Liu Ma. Fast shadow removal using adaptive multi-scale illumination transfer. *Comput. Graph. Forum*, 32(8):207–218, 2013. 2

[52] Chunxia Xiao, Donglin Xiao, Ling Zhang, and Lin Chen. Efficient shadow removal using subregion matching illumination transfer. *Comput. Graph. Forum*, 32(7):421–430, 2013. 2

[53] Chaoran Xu, Yao Lu, and Yuanpin Zhou. An automatic visible watermark removal technique using image inpainting algorithms. In *ICSAI*, pages 1152–1157, 2017. 2

[54] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, pages 1316–1324, 2018. 2

[55] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *CVPR*, pages 2327–2336, 2019.

[56] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, pages 5907–5915, 2017. 2

[57] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE TPAMI*, 41(8):1947–1962, 2018. 2

[58] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, pages 833–842, 2021. 2

[59] Ling Zhang, Qing Zhang, and Chunxia Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE TIP*, 24(11):4623–4636, 2015. 2

[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 1

[61] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *CVPR*, pages 6199–6208, 2018. 2

[62] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, pages 5802–5810, 2019. 2

[63] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng-Jun Zha. Bijective mapping network for shadow removal. In *CVPR*, pages 5627–5636, 2022. 2