ArtiFade: Learning to Generate High-quality Subject from Blemished Images (*Supplementary Material*)

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

A. Training Dataset Details

Our training dataset consists of 20 training subjects, used for the fine-tuning stage of our ArtiFade models. We show an example image of each subject in Fig. 1. In Fig. 2, we showcase several unblemished images alongside their corresponding blemished versions, each featuring one of the 10 watermark types. We randomly generate the training watermark artifacts with a random orientation between $[0°, 180°]$ with $15°$ intervals, a random font size between $[20pts, 160pts]$ with $15pts$ intervals, and a random color chosen from black, white, and rainbow colors with random text content.

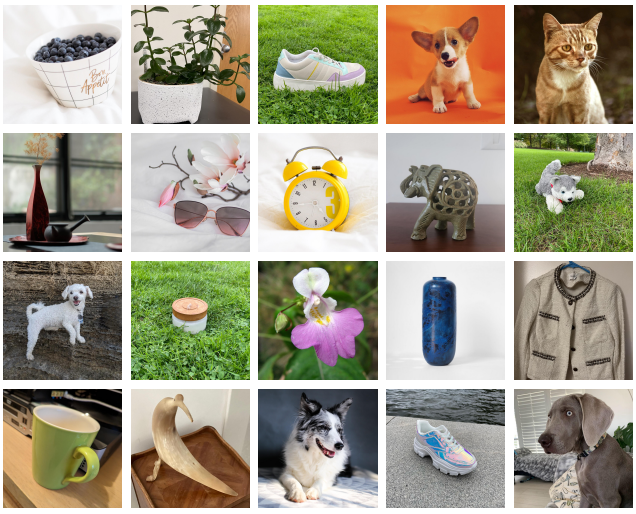


Figure 1. Examples of unblemished training images. We show a total of 20 images, each containing one distinct subject.

B. Test Dataset Details

In Fig. 3, we illustrate our `WM-ID-TEST` watermark types (see the first row) and `WM-OOD-TEST` watermark types (see the second row). The `WM-ID-TEST` watermarks are chosen from the training watermarks displayed in Fig. 2. On the other hand, the `WM-OOD-TEST` watermarks differ in font size, orientation, content, or color from all the training watermarks presented in Fig. 2.

C. Comparison with Two-stage Blemished Subject-driven Generation

In this section, we compare ArtiFade with a two-stage blemished subject-driven generation approach. The two-stage method consists of the following steps: (1) pre-processing blemished images using existing artifact removal methods, and (2) applying vanilla subject-driven generation methods to the pre-processed images.

Visible artifacts Most watermark removal models require a known watermark mask, which is impractical in real-world scenarios. An alternative is automated detection and removal, like OCR-SAM[1]. However, as shown in Fig. 4, both mask-conditioned method and OCR-SAM struggle with diverse watermarks and fail to remove other artifacts (*e.g.*, circles). Additionally, we observe that OCR-SAM often removes only partial watermarks (see Fig. 5a) and struggles to remove tiny watermarks (see Fig. 5a). It also inadvertently erases subject details related to text, (*i.e.*, the numbers on the clock), leading to lower subject fidelity outputs when used with DreamBooth [3]. In contrast, ArtiFade with DreamBooth (as mentioned in Sec. 4.4.1) effectively preserves critical subject details while successfully removing irrelevant watermarks as shown in Fig. 5.

Invisible artifacts We attempt to remove invisible artifacts [4] using spatial filters with varying kernel sizes (see Fig. 6). We observe that kernel selection is non-trivial, and ArtiFade mentioned in Sec. 4.4.2 consistently produces higher-quality results.

D. Analysis of Mixed-Artifacts

Although ArtiFade is trained on subjects blemished by a single type of artifact within each training dataset, it demonstrates strong generalization capabilities, effectively han-

*Equal contribution †Corresponding authors

[1]https://github.com/yeungchenwa/OCR-SAM

Figure 2. Examples of the training dataset: unblemished images and their corresponding blemished images.

Figure 3. Example of test watermark types. The first row displays the `WM-ID-TEST`, while the second row presents the `WM-OOD-TEST`.

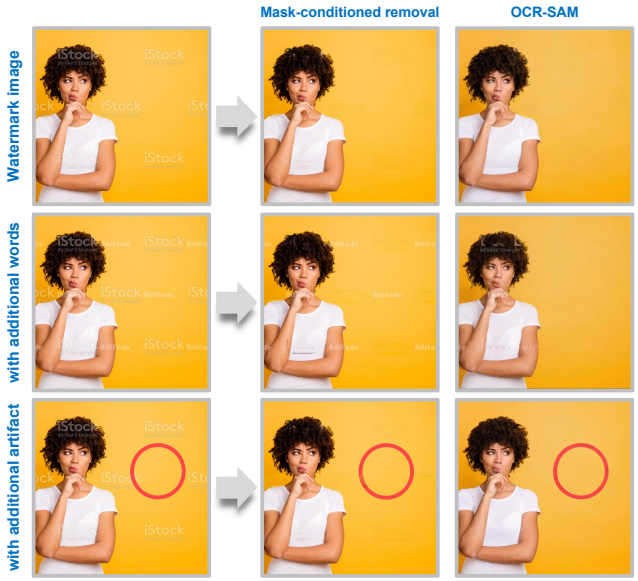
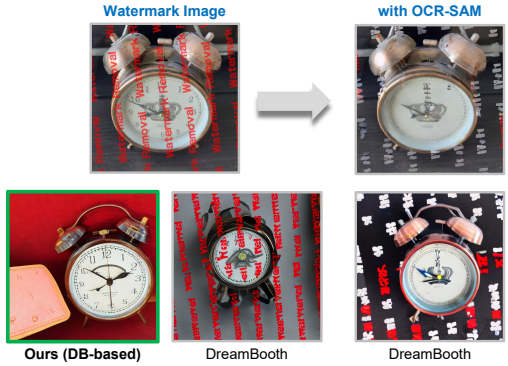


Figure 4. Results of mask-conditioned watermark removal and OCR-SAM across various types of artifacts. The results show that neither method generalizes well to different artifacts.

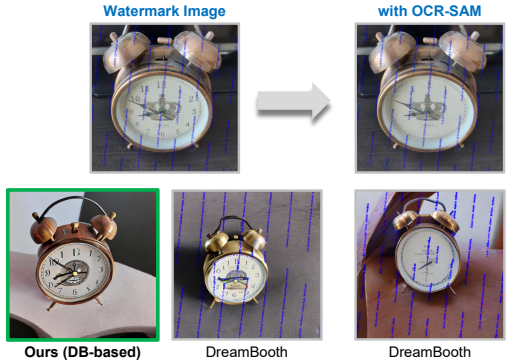
dling mixed-artifacts scenarios (see Fig. 7). We use ArtiFade with DreamBooth mentioned in Sec. 4.4.1 to give illustrations. In Fig. 7a, we show that ArtiFade can produce high-quality, artifact-free images even when various artifacts are present in the input dataset. Furthermore, Fig. 7b highlights ArtiFade's ability to address mixed-artifacts within individual images, showcasing its versatility and robustness. In contrast, the images generated by vanilla DreamBooth [3] are often polluted by mixed-artifacts, as it tends to learn either one type of artifact or both.

E. Analysis of Watermark Density

In Fig. 8, we present results to illustrate the impact of varying watermark densities (*i.e.*, varying qualities), highlighting the robust ability of our `WM-model` to remove watermarks under all conditions.



(a) Case 1.



(b) Case 2.

Figure 5. Comparisons between ArtiFade and two-stage blemished subject-driven generation using OCR-SAM and DreamBooth. The two-stage blemished subject-driven approach faces challenges in preserving subject details and fully eliminating watermarks. ArtiFade can produce high subject fidelity outputs by only eliminating watermarks while still preserving subject details.

F. Analysis of Unblemished Image Ratio

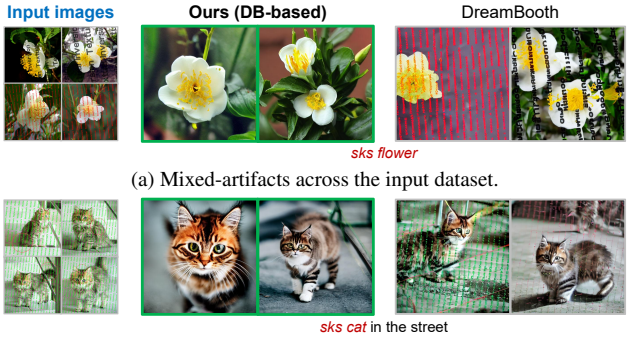
We employ our `WM-model` to evaluate the performance when the input images contain different proportions of unblemished images. We test our `WM-model` and Textual Inversion on five ratios of unblemished images: 100%, 75%, 50%, 25%, and 0%. The results are shown in Fig. 9.

Notably, even when there is only one blemished image in the second column example, the impact on Textual Inversion [2] is already evident, which deteriorates as the ratio decreases. Instead, our method effectively eliminates artifacts in all settings of unblemished image ratio, demonstrating its versatility in real-life scenarios.

G. Analysis of Training Dataset Size

We conduct an analysis to investigate the impact of the number of training subjects (*i.e.*, the size of the training dataset) on the performance of our model. We utilize the same set of artifacts $L_{\text{WM}} = 10$, as described in Method in the main paper. We construct blemished training datasets in four differ-

Figure 6. Comparisons of invisible artifact removal using spatial filters with varying kernel sizes. ArtiFade outperforms vanilla DreamBooth in all scenarios.



(a) Mixed-artifacts across the input dataset.

(b) Mixed-artifacts within individual images.

Figure 7. Comparisons between ArtiFade and DreamBooth when the inputs contain mixed-artifacts. ArtiFade demonstrates the ability to generalize to mixed-artifact scenarios by generating artifact-free and high-quality outputs.

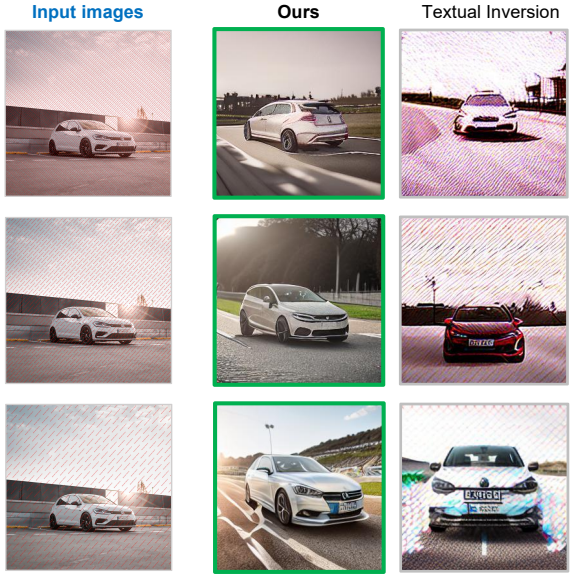


Figure 8. Varying qualities of input images. Our method (`WM-model`) can be used to remove watermarks when input images are of any quality.

ent sizes: (1) with 5 subjects, (2) with 10 subjects, (3) with 15 subjects, and (4) with 20 subjects. We generate 50, 100, 150, and 200 blemished datasets for each of these cases. Subsequently, we fine-tune four distinct ArtiFade models, each with 16k training steps.

We compare the models trained using different data sizes under the in-distribution scenario (see Fig. 10a) and under the out-of-distribution scenario (see Fig. 10b). We note that when the number of training subjects is less than 15, I^{DINO} and T^{CLIP} are relatively lower than the other two cases in both ID and OOD scenarios. This observation can be attributed to a significant likelihood of subject or background overfitting during the reconstruction and image synthesis processes, as visually illustrated in Fig. 11 and Fig. 12. However, as the number of training subjects reaches or exceeds 15, we observe a convergence in the values of I^{DINO} and T^{CLIP}, indicating a reduction in subject overfitting. Regarding R^{DINO}, we note that all cases exhibit values greater than one, with a slightly increasing trend as the number of training subjects rises.

H. Failure Cases

We present several failure cases when applying ArtiFade based on Textual Inversion. We demonstrate the limitations of our `WM-model` in Fig. 13. Despite the model's ability to eliminate watermarks, we still encounter issues with incorrect subject color, as shown in Fig. 13a, which arises due to the influence of the watermark color. We also encounter incorrect subject identity in some cases, as demonstrated in Fig. 13b. One possible reason is that the watermarks

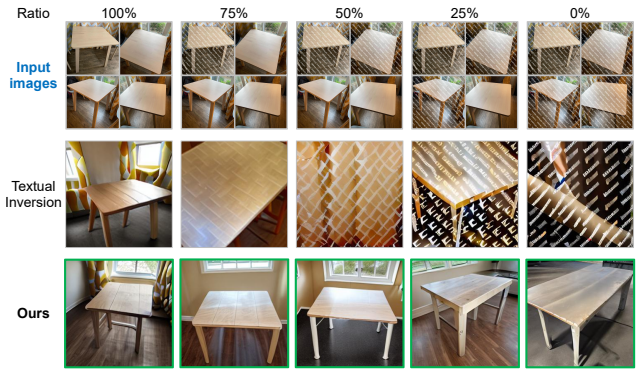


Figure 9. Comparison between different ratios of unblemished images. ArtiFade can perform well under any scenarios with different ratios of unblemished images.

significantly contaminate the images, causing the learning process of embedding to focus on the contaminated visual appearance instead of the intact subject. Another failure case is subject overfitting, as shown in Fig. 13c. In this case, the constructed subject overfits with a similar subject type that appears in the training dataset. This problem occurs because the blemished embedding of the testing subject closely resembles some blemished embeddings of the training subjects. Surprisingly, we find those problems can be solved by using ArtiFade based on DreamBooth, which is mentioned in Sec. 4.4.1. Therefore, we recommend using ArtiFade based on DreamBooth when encountering the limitations mentioned above.

I. Additional Comparison with Textual Inversion

We use the same training subjects with N= 20 from Sec. 3.3 to train an ArtiFade model named `RC-model` using red circle artifacts. For the training set of `RC-model`, due to the simplicity of red circles, we only synthesize a single blemished subset (*i.e.*, $L_{\text{RC}} = 1$) for each subject, deriving 20 blemished subsets in total. We augment each image with a red circle mark that is randomly scaled and positioned on the source image. Considering the small scale of `RC-model`'s datasets, we only fine-tune `RC-model` for 8k steps. We further introduce `RC-test`, which applies only one type of artifact (*i.e.*, red circle) to our 16 test subjects, resulting in 16 test sets. We test both `RC-model` and `WM-model` on `RC-test`. The quantitative and qualitative results are shown in Tab. 1 and Fig. 14, respectively.

Quantitative results analysis. From Tab. 1, we can observe that both `RC-model` and `WM-model` yield higher results in nearly all cases than Textual Inversion [2] with blemished inputs, showing the capability of our models to eliminate artifacts and generate subjects with higher fidelity.

Table 1. Quantitative results of `RC-test`.

Method	RC-test				
	I^{DINO}	R^{DINO}	I^{CLIP}	R^{CLIP}	T^{CLIP}
TI (unblemished)	0.488	1.021	0.730	1.077	0.283
TI (blemished)	0.406	0.990	0.672	1.042	0.284
Ours (`RC-model`)	**0.476**	**1.013**	0.722	**1.065**	**0.285**
Ours (`WM-model`)	0.474	1.006	**0.727**	1.063	0.282

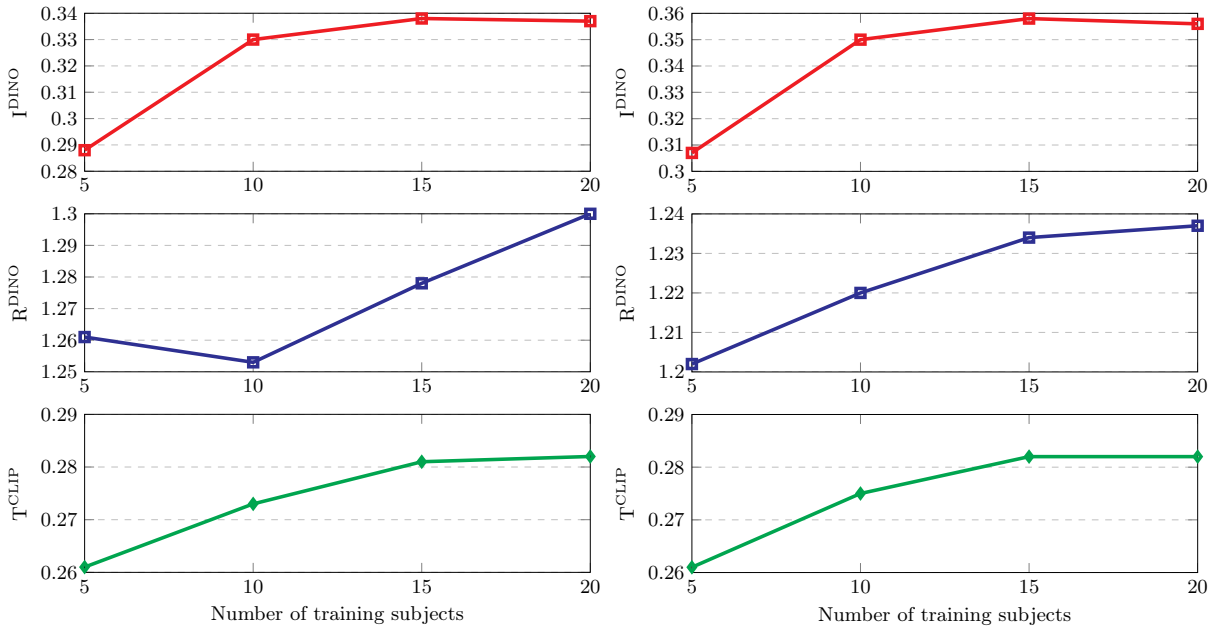
It is important to note that the `RC-test` is considered out-of-distribution with respect to `WM-model`. Nevertheless, the metrics produced by `WM-model` remain comparable to those of `RC-model`, with a minor difference observed. These results provide additional evidence supporting the generalizability of our `WM-model`.

Qualitative results analysis. As illustrated in Fig. 14, Textual Inversion struggles with accurate color reconstruction. It also showcases subject distortions and introduces red-circle-like artifacts during image generation when using blemished embeddings. In contrast, our `RC-model` (see Fig. 14a) and `WM-model` (see Fig. 14b) are capable of generating high-quality images that accurately reconstruct the color and identities of subjects without any interference from artifacts during the image synthesis.

J. Additional Qualitative Comparisons

We present additional qualitative results comparing our ArtiFade models with Textual Inversion [2] and DreamBooth [3] in Fig. 15. We employ `WM-model` and ArtiFade based on DreamBooth mentioned in Sec. 4.4.1. Textual Inversion generates images with distorted subjects and backgrounds contaminated by watermarks, whereas DreamBooth can effectively capture intricate subject details and accurately reproduce watermark patterns. In contrast, our models (*i.e.*, TI-based and DB-based ArtiFade) generate images devoid of watermark pollution with correct subject identities for both in-distribution (see the first three rows in Fig. 15) and out-of-distribution (see the last two rows in Fig. 15) cases. Notably, our method based on DreamBooth preserves the high fidelity and finer detail reconstruction benefits of vanilla DreamBooth, even in the context of blemished subject-driven generation.

In Fig. 16, we show qualitative results for subjects with complex features (*e.g.*, human faces) using our models, Textual Inversion, DreamBooth and Break-a-Scene [1]. Break-a-Scene can separate multiple subjects inside one image. We use Break-a-scene to generate human-only images. However, we find that Break-a-scene fails to separate humans from artifacts, resulting in polluted images. As a result, our methods (*i.e.*, TI-based and DB-based ArtiFade) consistently surpass Textual Inversion, DreamBooth, and



(a) ID (b) OOD

Figure 10. Analysis of the number of training subjects.

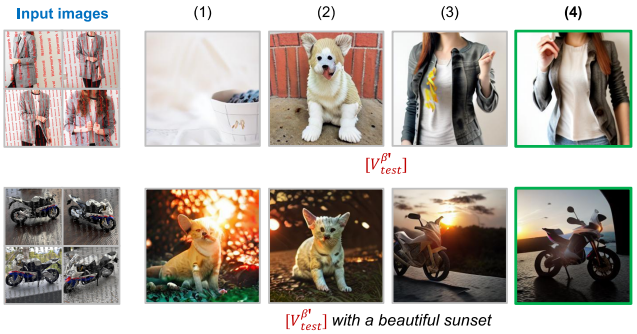


Figure 11. Qualitative results of different number of training subjects - ID.

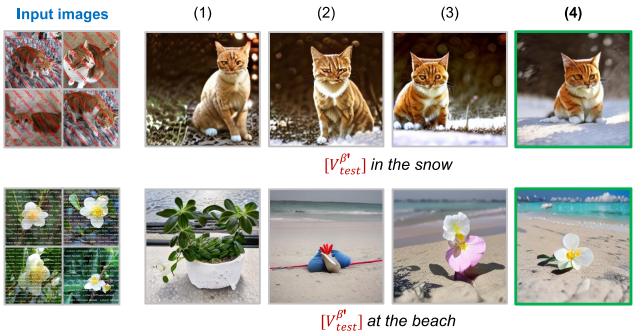


Figure 12. Qualitative results of different number of training subjects - OOD.

Break-a-Scene, achieving high-quality image generation of complex data in in-distribution cases, as shown in the first two rows of Fig. 16, and out-of-distribution cases, as illustrated in the last row of Fig. 16.

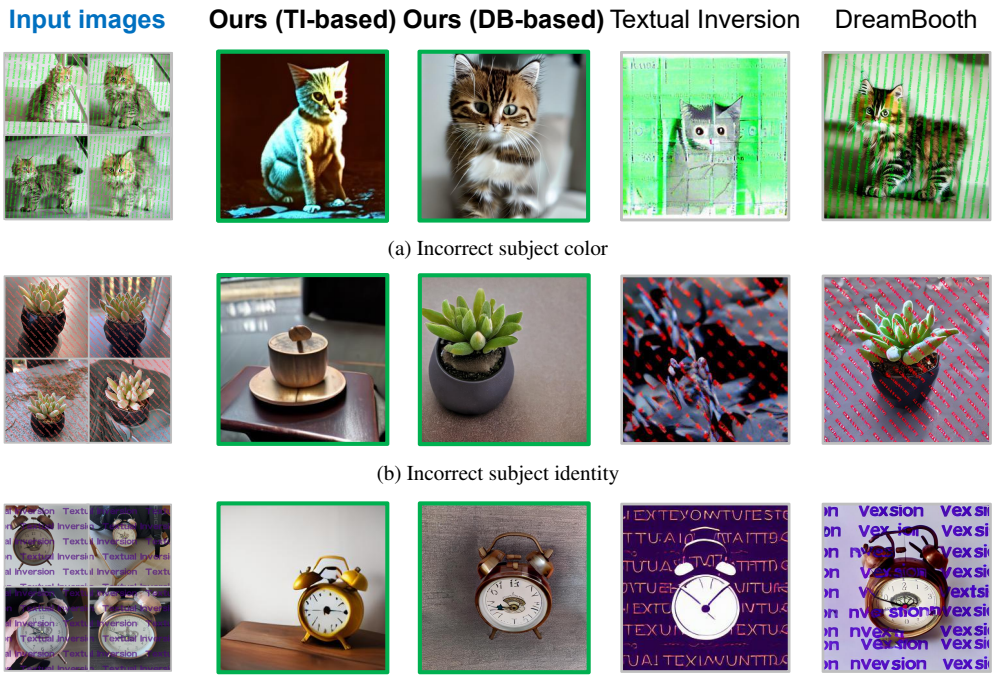
K. More Applications

We explore more applications of our `WM-model`, demonstrating its versatility beyond watermark removal. As shown in Fig. 17, our model exhibits the capability to effectively eliminate unwanted artifacts from images, enhancing their visual quality. Furthermore, our model showcases the ability to recover incorrect image styles induced by artifacts, thereby restoring the intended style of the images.

Natural artifacts We use watermarked images from Shutterstock as inputs, and our model (in Sec. 4.4.1) can successfully remove those artifacts in generation compared to DreamBooth (see Fig. 18).

L. Social Impact

Our research addresses the emerging challenge of generating content from images with embedded watermarks, a scenario we term blemished subject-driven generation. Users often source images from the internet, some of which may contain watermarks intended to protect the original author's copyright and identity. However, our method is capable of removing various types of watermarks, potentially compromising the authorship and copyright protection. This could lead to increased instances of image piracy and the generation of illicit content. Hence, we advocate for legal compli-



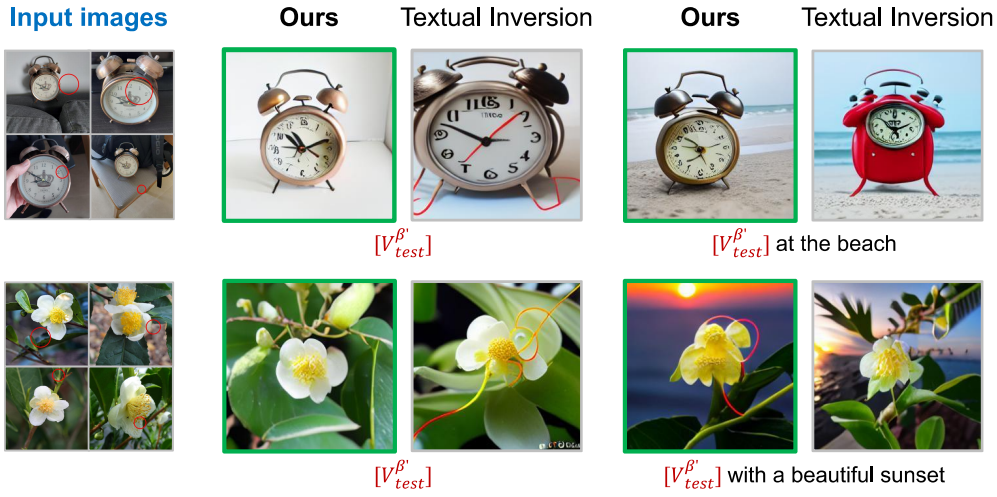
(a) Incorrect subject color

(b) Incorrect subject identity

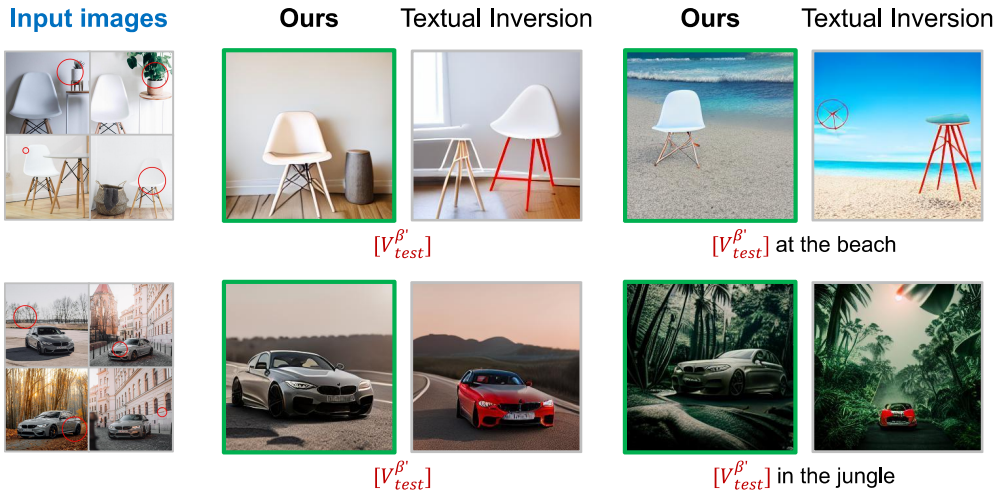
(c) Subject overfitting

Figure 13. Failure cases of ArtiFade based on Textual Inversion. We observe three main types of failure cases of our `WM-model`: (a) incorrect subject color, (b) incorrect subject identity, and (c) subject overfitting. However, those limitations can be resolved by using ArtiFade with DreamBooth-based fine-tuning.

ance and the implementation of usage restrictions to govern the deployment of our technique and subsequent models in the future.



(a) `RC-model` on `RC-test`.



(b) `WM-model` on `RC-test`.

Figure 14. Qualitative results of `RC-test`. Our models consistently output high-quality and artifact-free images compared to Textual Inversion.

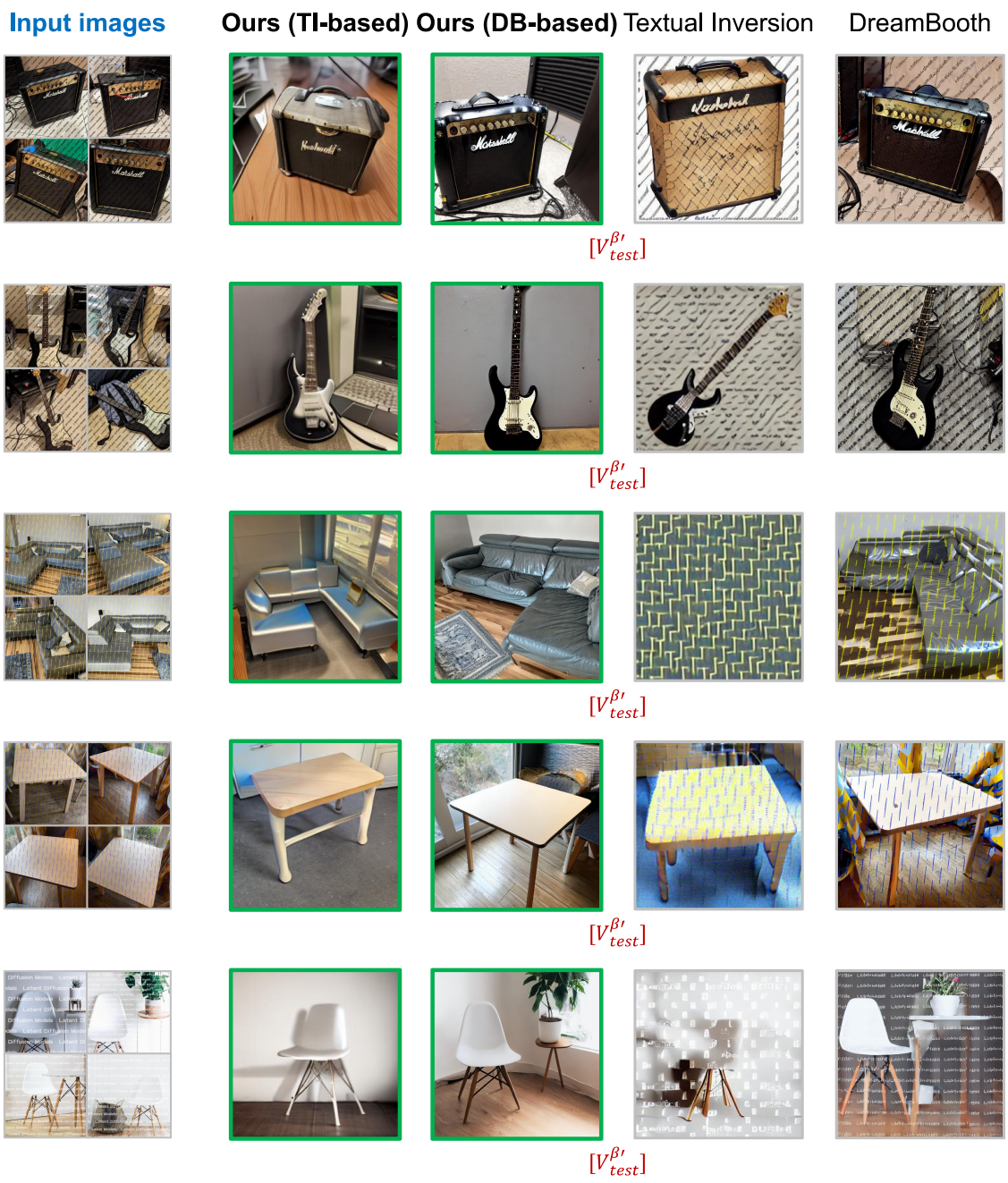


Figure 15. Additional qualitative comparisons.

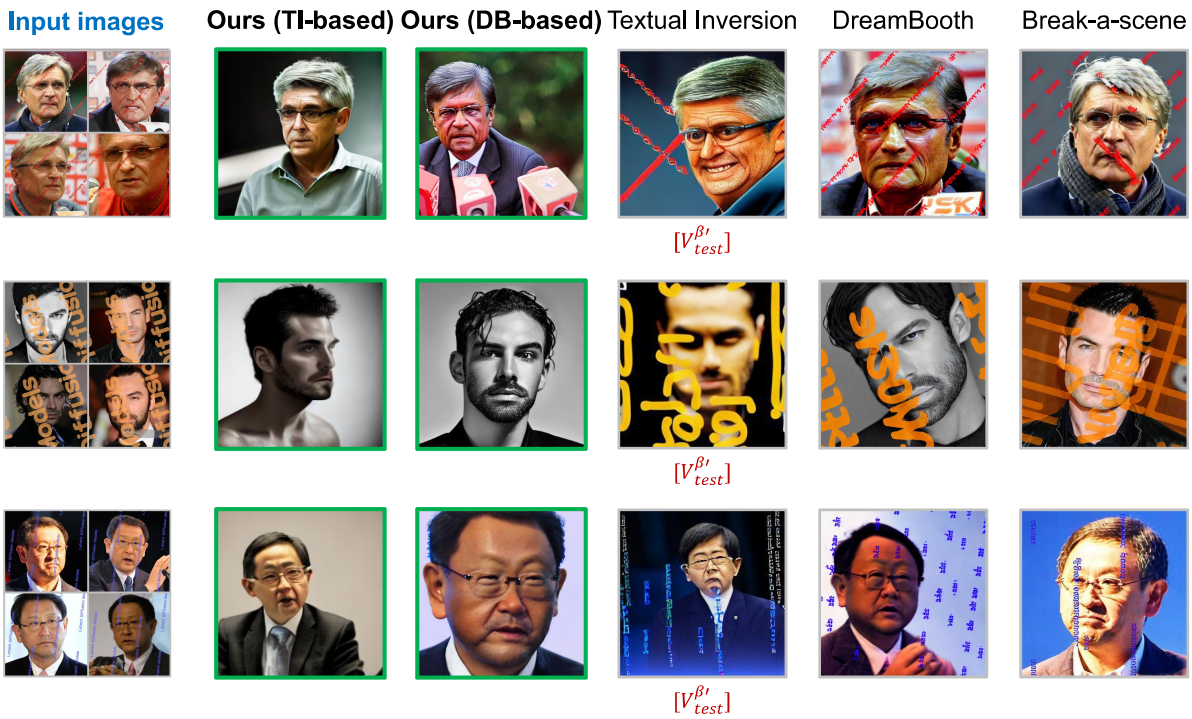


Figure 16. Additional qualitative comparisons - Human Faces.

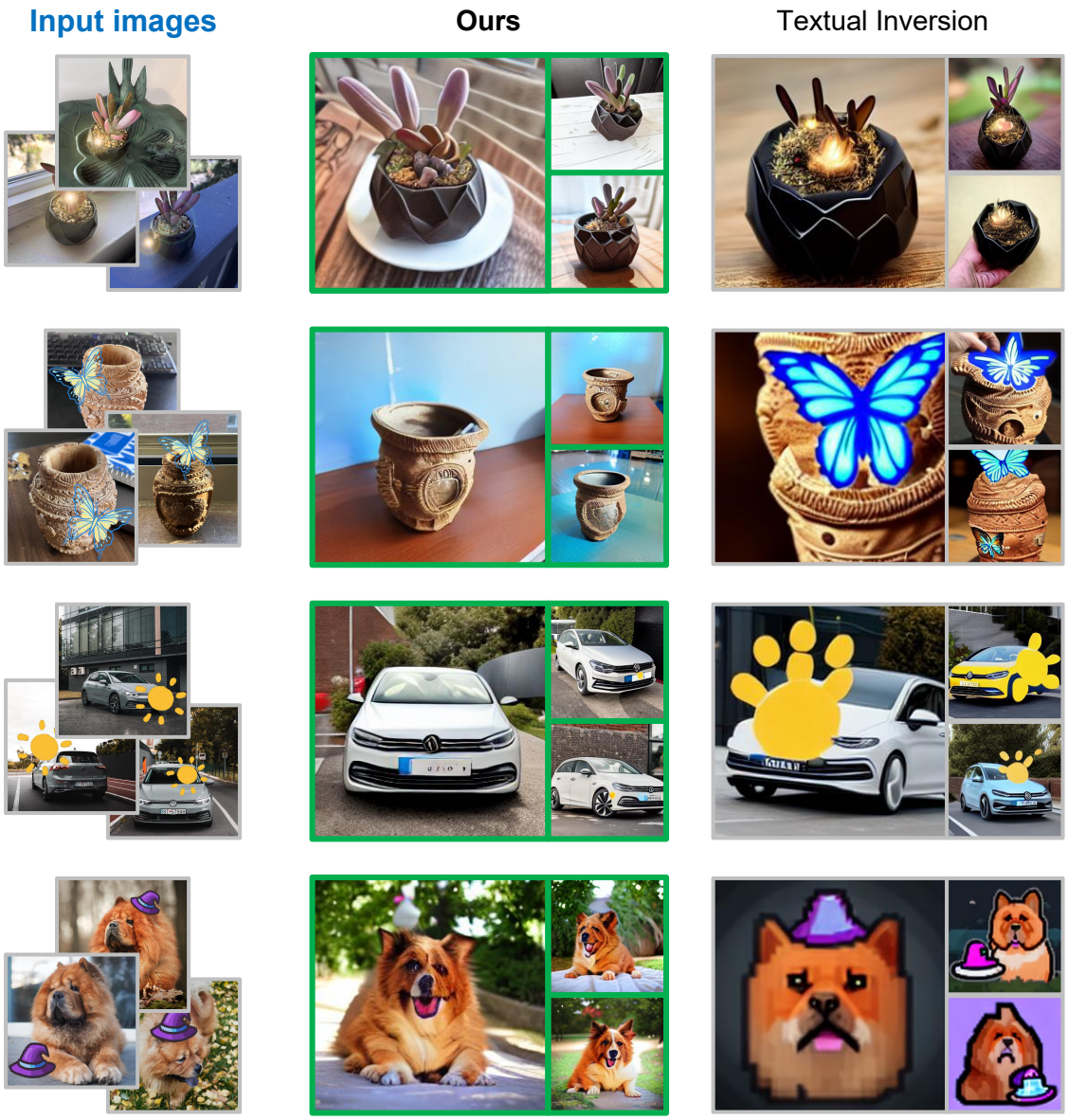


Figure 17. More applications. Our `WM-model` can be used to eliminate various stickers and fix the incorrect image style.

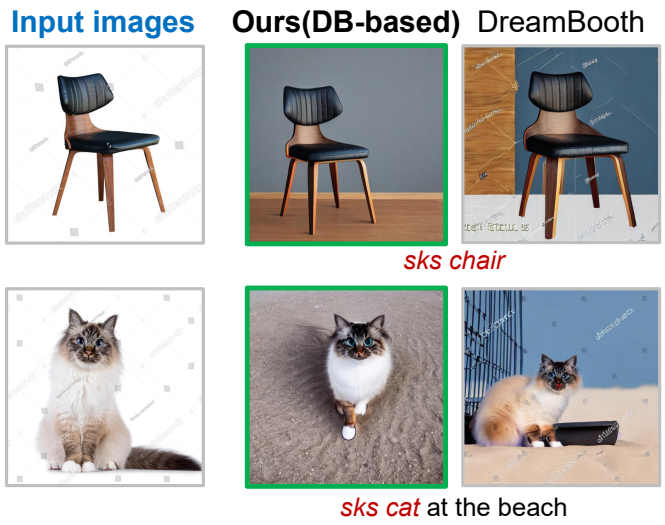


Figure 18. Natural artifacts removal.

References

[1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 5

[2] 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. 3, 5

[3] 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, 3, 5

[4] 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