Supplementary Material for *StyleProtect: Safeguarding Artistic Identity in Finetuned Diffusion Models*

Qiuyu Tang
Lehigh University
qit220@lehigh.edu

Joshua Krinsky
Lehigh University
jpk322@lehigh.edu

Aparna Bharati
Lehigh University
apb220@lehigh.edu

1. Dataset Details

We provide the detailed description of our artwork subset from the WikiArt dataset. To identify the most stylistically unique artists (to enable distinct visual representation), we leveraged ChatGPT-4 [3] to generate a curated list of renowned artists known for their highly individual and recognizable styles. Figure 1 illustrates the process. Given that each artist possesses a unique style, we represent styles by directly linking them to their respective artist names. Starting with the complete list of artists from the WikiArt dataset , we prompted the model to produce a refined selection of artists exhibiting the most distinct and personal styles, along with detailed reasoning, nominating each artist and explaining the distinctive qualities for which they are known. Subsequently, we manually reviewed these nominated candidates and their works to ensure clarity and consistency in stylistic expression. Ultimately, we selected 30 prominent artists whose works demonstrate strong and distinguishable stylistic characteristics. For each artist, representing a distinct style, we curated a small subset of 3 to 5 representative images for customization. This refined dataset forms a compact, yet effective benchmark for evaluating style mimicry and protection.

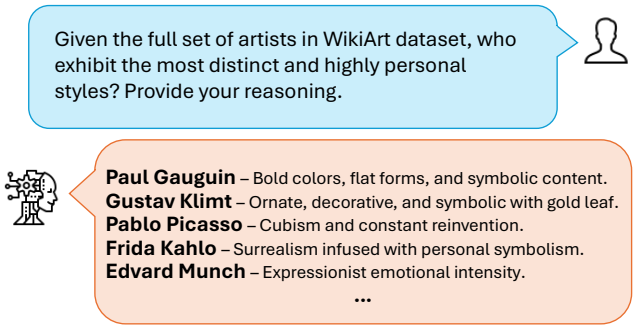

Figure 1. The process of building the refined artwork dataset using ChatGPT [3], for evaluating style protection effectiveness.

2. Qualitative Results for Imperceptibility

To visualize the impact of the perturbations, Figure 2 compares the perturbation patterns on WikiArt dataset, produced by different methods. Except for MIST [2], all methods demonstrate strong stealthiness and remain largely imperceptible to the human eye. Note that CAAT [7] applies certain data augmentations during perturbation training, including a 50% probability of horizontal flipping which explains the appearance of the first CAAT perturbation.

We also compare the invisibility on Anita Dataset in Figure 3. Glaze [5] is not included in this comparison because its extremely long runtime makes it impractical to run under our experimental setting.

3. Details of Robustness Analysis

This section further analyzes the robustness of StyleProtect, focusing on its cross-model adaptability and its resilience to chains of post-processing operations.

3.1. Cross-Model Adaptability

Protection Performance Across Editing Models. To assess whether our protection remains effective when an exploiter attempts to edit a protected image, we evaluate our method on SD v2.1 in Fig. 4. After finetuning, the generated outputs of the protected images display clear visual disruptions and a characteristic, consistent noise pattern. The results show that our method can offer protection from models not used for training.

Figure 4. We evaluate StyleProtect using SDv2.1 as the base model for finetuning. The finetuned generations of protected images exhibit disrupted visual structures and a consistent noise pattern.

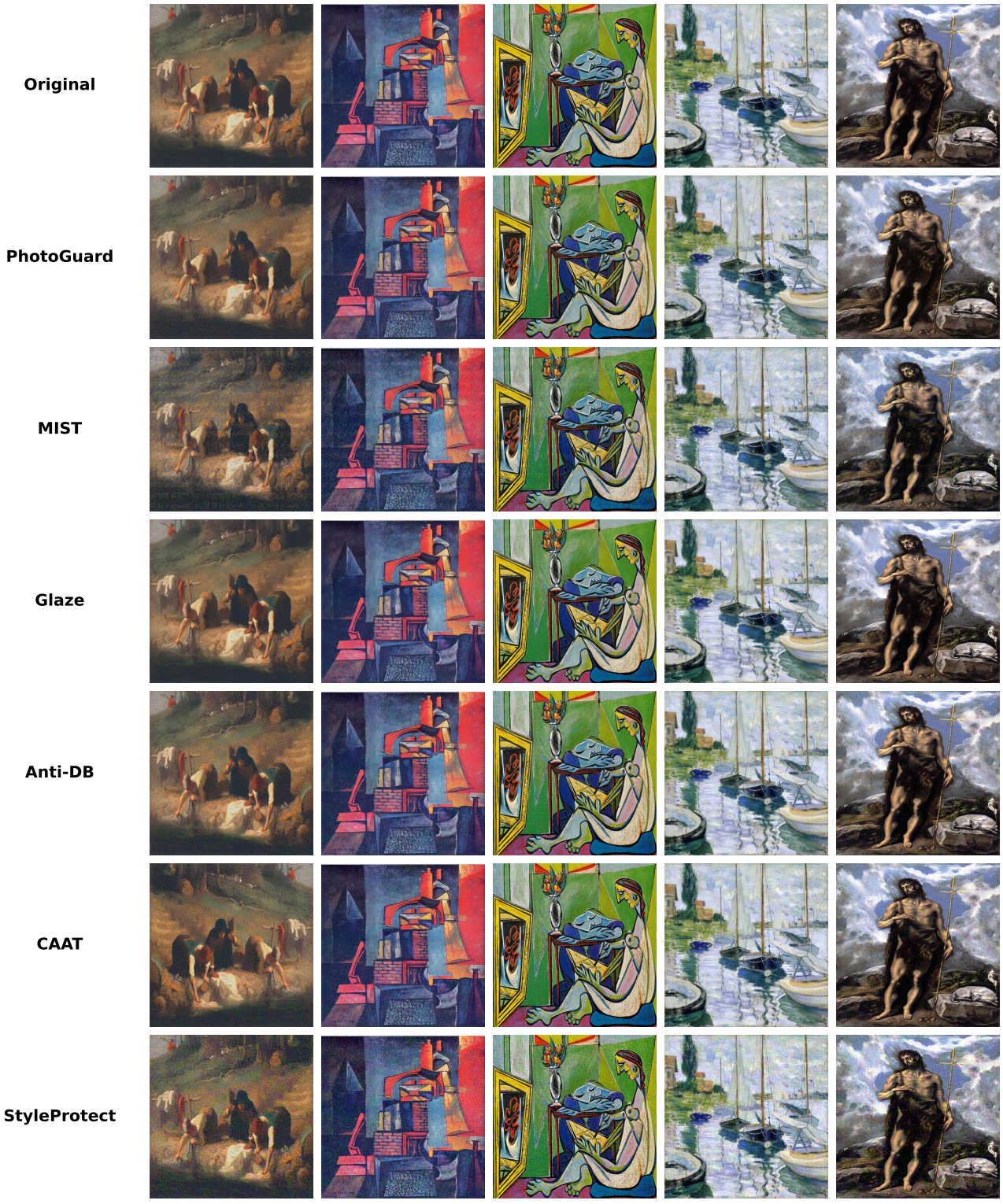

Figure 2. Visualization of perturbations applied by different methods on WikiArt artwork images. Our method introduces slight noise but strikes a good balance among efficiency, protection performance, and imperceptibility.

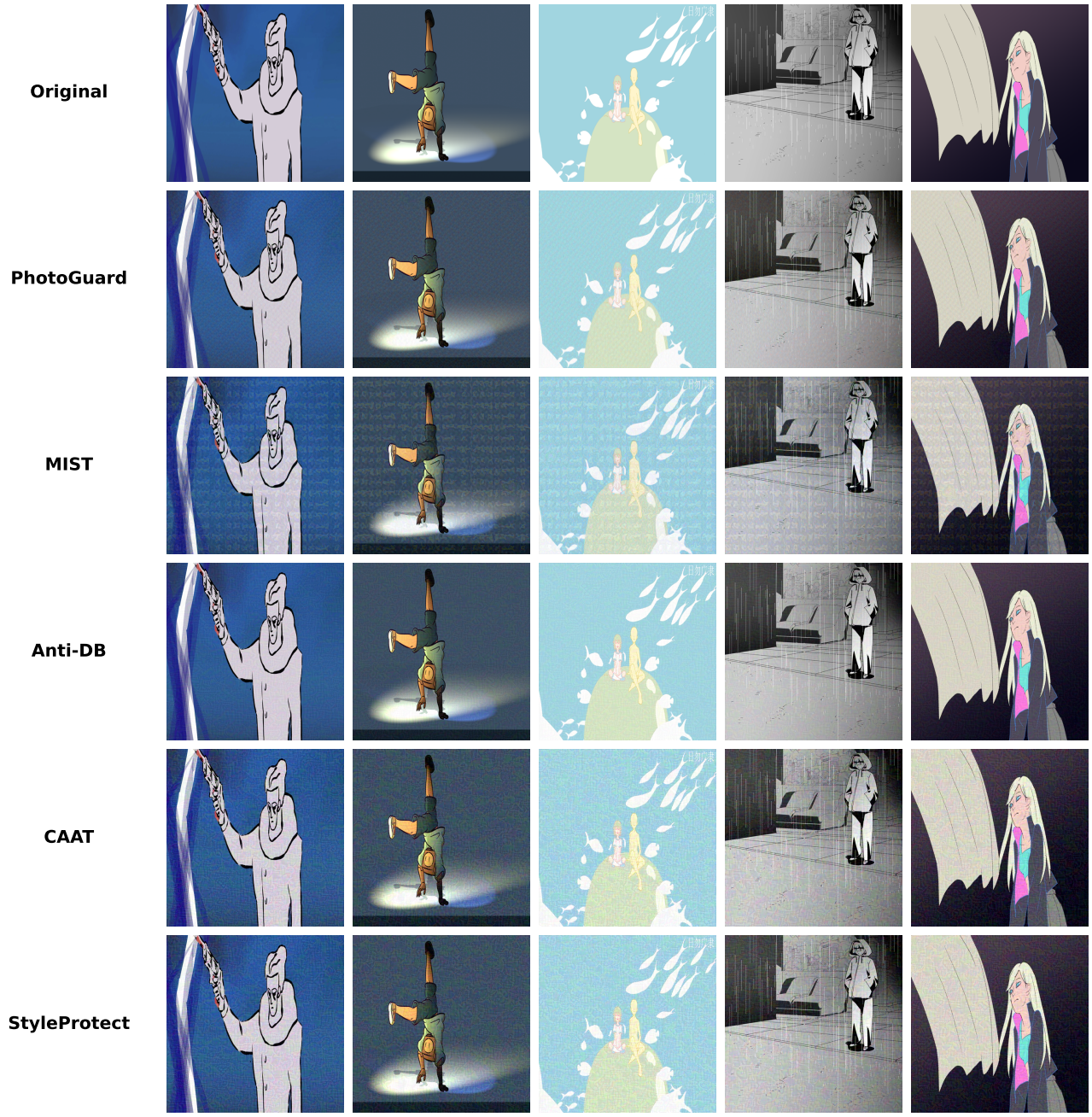

Figure 3. Visualization of perturbations applied by different methods on Anita anime images. While our method produces slight noise, it achieves an effective trade-off between efficiency, protection efficacy, and invisibility.

Protection Performance Across Finetuning Techniques. We further investigate the performance of StyleProtect under the LoRA-based fine-tuning settings. As shown in Figure 5, an interesting observation emerges: even the clean (unprotected) images often fail to reproduce the target artistic style learned by the LoRA model. This phenomenon suggests that the LoRA adaptation process itself introduces substantial distribution shifts [6], making style transfer unstable regardless of whether protection is applied.

A potential reason is that since LoRA updates only a small subset of parameters, it struggles to capture global stylistic patterns, leading to inconsistent style expression across prompts. This limitation explain why clean images already struggle to reproduce the target style, and this is-

Figure 5. Generated images finetuned with clean images with LoRA techniques on SD v1.5 and v2.1. The results show that even clean, unprotected inputs often fail to reproduce the stylistic characteristics captured during fine-tuning.

sue becomes even more noticeable once StyleProtect is applied. Moreover, LoRA fine-tuning typically requires a larger dataset (around 10–20 images) to learn a coherent artistic style, whereas our setting uses only 4–6 images per style, resulting in limited generalization.

3.2. Resilience to Post-processing Operations

To mimic real-world dissemination on online platforms, we build automated image editing chains, combining traditional image processing operations and prompt-based AI edits. Hence, we define ten possible edits and compose them into chains of length up to five. These include seven traditional image transformations: JPEG compression, crop/resize, Gaussian blur, horizontal flip, brightness/contrast, super resolution, and geometric distortion, and three prompt-based AI edits using Stable Diffusion XL [4]. The prompts used are: convert to realistic imagery, change to a target style, and a semantic transformation. For the semantic transformation, we leverage Gemini 2.5-Flash [1] to first identify objects in the image and then apply random object inpainting using SDXL. For any given image, we construct 200 chains composed of randomly selected non-repeating transformations: 10 1-level, 90 2-level, 50 3-level, 30 4-level, and 20 5-level chains.

We summarize these transformations in Table. 1. Our chain of edits is constructed from these ten operations. We then evaluate the similarity with SSIM, LPIPS metrics between the original and protected images after applying transformations from 1st level to 5th level (see Fig. 6). As the level increases, the SSIM score consistently decreases while the LPIPS score grows. This indicates that the protected images deviate increasingly from the original images in both structural content and perceptual similarity. Consequently, the protection progressively weakens, as higher levels introduce stronger distortions that reduce the effectiveness of StyleProtect's defense.

Method	Description
Crop & Resize	Crop 128 pixels' edge, then resize to 512×512
Gaussian blurring	Using kernel: 3×3, sigma = 0.05
JPEG compression	With the factor as 75
Horizontal Flipping	-
Brightness & Contrast Adjustment	Change brightness to 10, contrast to 2
Super Resolution	Upscaling using the EDSR model
Barrel Distortion	-
AI-based stylization	Edit the image using SDXL prompting "Transform to realistic photograph"
AI-based stylization	Edit the image using SDXL prompting "Change to Ghibli style"
AI-based semantic editing	1. Gemini detects objects; 2. generate an image without selected objects using SDXL

Table 1. Post-processing transformations used in our evaluation.

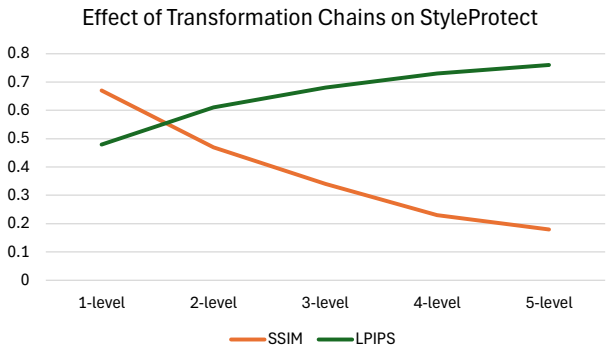

Figure 6. With increasing levels, SSIM monotonically decreases while LPIPS increases, reflecting reduced similarity to the original image and diminishing protection strength.

References

[1] Gemini 2.5-flash. https://gemini.google.com/, 2025. 4

[2] Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models, 2023. 1

[3] OpenAI. Chatgpt. https://chat.openai.com/, 2025. Accessed: 2025-01-10. 1

[4] 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. 4

[5] 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. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023. 1

[6] Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. Lora vs full fine-tuning: An illusion of equivalence. *arXiv preprint arXiv:2410.21228*, 2024. 3

[7] Jingyao Xu, Yuetong Lu, Yandong Li, Siyang Lu, Dongdong Wang, and Xiang Wei. Perturbing attention gives you more bang for the buck: Subtle imaging perturbations that efficiently fool customized diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24534–24543, 2024. 1