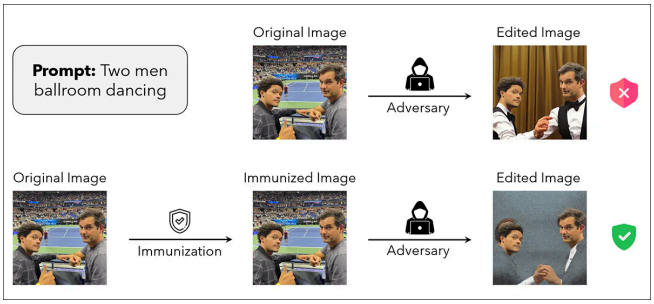
‘Protected’ Images Are Easier, Not More Difficult, to Steal With AI

Originally published at https://www.unite.ai/protected-images-are-easier-not-more-difficult-to-steal-with-ai/

New research suggests that watermarking tools meant to block AI image edits may backfire. Instead of stopping models like Stable Diffusion from making changes, some protections actually help *the AI follow editing prompts more closely, making unwanted manipulations even easier.*

There is a notable and robust strand in computer vision literature dedicated to protecting copyrighted images from being trained into AI models, or being used in direct image>image AI processes. Systems of this kind are generally aimed at Latent Diffusion Models (LDMs) such as Stable Diffusion and Flux, which use noise-based procedures to encode and decode imagery.

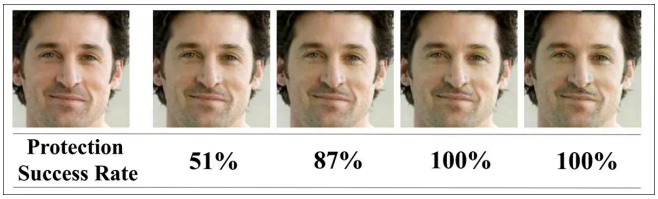
By inserting adversarial noise into otherwise normal-looking images, it can be possible to cause image detectors to guess image content incorrectly, and hobble image-generating systems from exploiting copyrighted data:



From the MIT paper ‘Raising the Cost of Malicious AI-Powered Image Editing’, examples of a source image ‘immunized’ against manipulation (lower row). Source: https://arxiv.org/pdf/2302.06588

Since an artists’ backlash against Stable Diffusion’s liberal use of web-scraped imagery (including copyrighted imagery) in 2023, the research scene has produced multiple variations on the same theme – the idea that pictures can be invisibly ‘poisoned’ against being trained into AI systems or sucked into generative AI pipelines, without adversely affecting the quality of the image, for the average viewer.

In all cases, there is a direct correlation between the intensity of the imposed perturbation, the extent to which the image is subsequently protected, and the extent to which the image doesn’t look quite as good as it should:



Though the quality of the research PDF does not completely illustrate the problem, greater amounts of adversarial perturbation sacrifice quality for security. Here we see the gamut of quality disturbances in the 2020 'Fawkes' project led by the University of Chicago. Source: https://arxiv.org/pdf/2002.08327

Of particular interest to artists seeking to protect their styles against unauthorized appropriation is the capacity of such systems not only to obfuscate identity and other information, but to 'convince' an AI training process that it is seeing something other than it is really seeing, so that connections do not form between semantic and visual domains for 'protected' training data (i.e., a prompt such as *'In the style of Paul Klee'*).

Mist and Glaze are two popular injection methods capable of preventing, or at least severely hobbling attempts to use copyrighted styles in AI workflows and training routines. Source: https://arxiv.org/pdf/2506.04394

Own Goal

Now, new research from the US has found not only that perturbations can fail to protect an image, but that adding perturbation can actually *improve* the image's exploitability in all the AI processes that perturbation is meant to immunize against.

The paper states:

> *'In our experiments with various perturbation-based image protection methods across multiple domains (natural scene images and artworks) and editing tasks (image-to-image generation and style editing), we discover that such protection does not achieve this goal completely.*
>
> *'In most scenarios, diffusion-based editing of protected images generates a desirable output image which adheres precisely to the guidance prompt.*
>
> *'Our findings suggest that adding noise to images may paradoxically increase their association with given text prompts during the generation process, leading to unintended consequences such as* better *resultant edits.*
>
> *'Hence, we argue that perturbation-based methods may not provide a sufficient solution for robust image protection against diffusion-based editing.'*

In tests, the protected images were exposed to two familiar AI editing scenarios: straightforward image-to-image generation and style transfer. These processes reflect the common ways that AI models might exploit protected content, either by directly altering an image, or by borrowing its stylistic traits for use elsewhere.

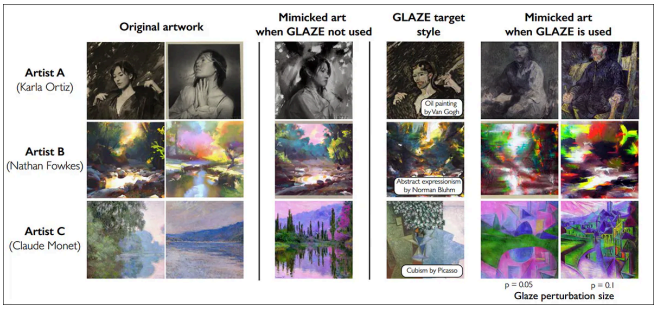
The protected images, drawn from standard sources of photography and artwork, were run through these pipelines to see whether the added perturbations could block or degrade the edits.

Instead, the presence of protection often seemed to sharpen the model's alignment with the prompts, producing clean, accurate outputs where some failure had been expected.

The authors advise, in effect, that this very popular method of protection may be providing a false sense of security, and that any such perturbation-based immunization approaches should be tested thoroughly against the authors' own methods.

Method

The authors ran experiments using three protection methods that apply carefully-designed adversarial perturbations: PhotoGuard; Mist; and Glaze.



Glaze, one of the frameworks tested by the authors, illustrating Glaze protection examples for three artists. The first two columns show the original artworks; the third column shows mimicry results without protection; the fourth, style-transferred versions used for cloak optimization, along with the target style name. The fifth and sixth columns show mimicry results with cloaking applied at perturbation levels p = 0.05 *and* p = 0.1. *All results use Stable Diffusion models.* https://arxiv.org/pdf/2302.04222

PhotoGuard was applied to natural scene images, while Mist and Glaze were used on artworks (i.e., 'artistically-styled' domains).

Tests covered both natural and artistic images to reflect possible real-world uses. The effectiveness of each method was assessed by checking whether an AI model could still produce realistic and prompt-relevant edits when working on protected images; if the resulting images appeared convincing and matched the prompts, the protection was judged to have failed.

Stable Diffusion v1.5 was used as the pre-trained image generator for the researchers' editing tasks. Five seeds were selected to ensure reproducibility: 9222, 999, 123, 66, and 42. All other generation settings, such as guidance scale, strength, and total steps, followed the default values used in the PhotoGuard experiments.

PhotoGuard was tested on natural scene images using the Flickr8k dataset, which contains over 8,000 images paired with up to five captions each.

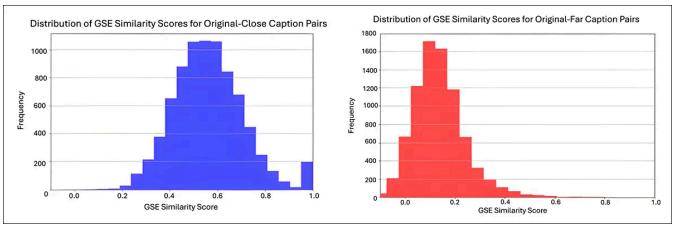
Opposing Thoughts

Two sets of modified captions were created from the first caption of each image with the help of Claude Sonnet 3.5. One set contained prompts that were *contextually close* to the original captions; the other set contained prompts that were *contextually distant*.

For example, from the original caption *'A young girl in a pink dress going into a wooden cabin'*, a close prompt would be *'A young boy in a blue shirt going into a brick house'*. By contrast, a *distant* prompt would be *'Two cats lounging on a couch'*.

Close prompts were constructed by replacing nouns and adjectives with semantically similar terms; far prompts were generated by instructing the model to create captions that were contextually very different.

All generated captions were manually checked for quality and semantic relevance. Google's Universal Sentence Encoder was used to calculate semantic similarity scores between the original and modified captions:



From the supplementary material, semantic similarity distributions for the modified captions used in Flickr8k tests. The graph on the left shows the similarity scores for closely modified captions, averaging around 0.6. The graph on the right shows the extensively modified captions, averaging around 0.1, reflecting greater semantic distance from the original captions. Values were calculated using Google's Universal Sentence Encoder. Source: https://sigport.org/sites/default/files/docs/IncompleteProtection_SM_0.pdf

Each image, along with its protected version, was edited using both the close and far prompts. The Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) was used to assess image quality:

Image-to-image generation results on natural photographs protected by PhotoGuard. Despite the presence of perturbations, Stable Diffusion v1.5 successfully followed both small and large semantic changes in the editing prompts, producing realistic outputs that matched the new instructions.

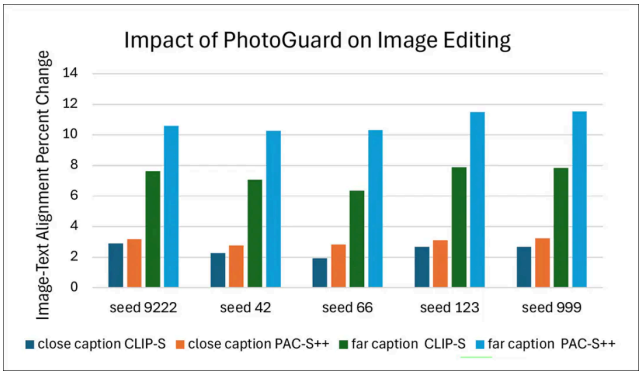
The generated images scored 17.88 on BRISQUE, with 17.82 for close prompts and 17.94 for far prompts, while the original images scored 22.27. This shows that the edited images remained close in quality to the originals.

Metrics

To judge how well the protections interfered with AI editing, the researchers measured how closely the final images matched the instructions they were given, using scoring systems that compared the image content to the text prompt, to see how well they align.

To this end, the CLIP-S metric uses a model that can understand both images and text to check how similar they are, while PAC-S++, adds extra samples created by AI to align its comparison more closely to a human estimation.

These Image-Text Alignment (ITA) scores denote how accurately the AI followed the instructions when modifying a protected image: if a protected image still led to a highly aligned output, it means the protection was deemed to have *failed* to block the edit.



Effect of protection on the Flickr8k dataset across five seeds, using both close and distant prompts. Image-text alignment was measured using CLIP-S and PAC-S++ scores.

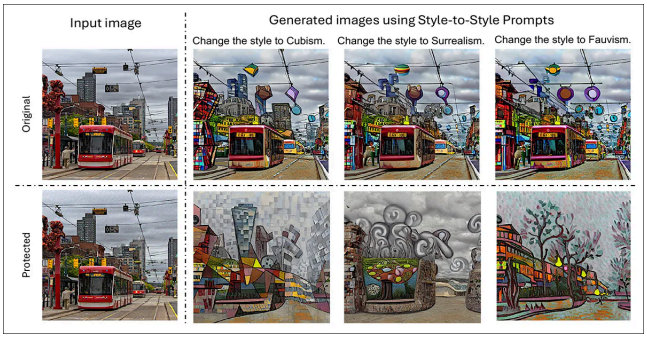
The researchers compared how well the AI followed prompts when editing protected images versus unprotected ones. They first looked at the difference between the two, called the *Actual Change*. Then the difference was scaled to create a *Percentage Change*, making it easier to compare results across many tests.

This process revealed whether the protections made it harder or easier for the AI to match the prompts. The tests were repeated five times using different random seeds, covering both small and large changes to the original captions.

Art Attack

For the tests on natural photographs, the Flickr1024 dataset was used, containing over one thousand high-quality images. Each image was edited with prompts that followed the pattern: *'change the style to [V]'*, where *[V]* represented one of seven famous art styles: Cubism; Post-Impressionism; Impressionism; Surrealism; Baroque; Fauvism; and Renaissance.

The process involved applying PhotoGuard to the original images, generating protected versions, and then running both protected and unprotected images through the same set of style transfer edits:



Original and protected versions of a natural scene image, each edited to apply Cubism, Surrealism, and Fauvism styles.

To test protection methods on artwork, style transfer was performed on images from the WikiArt dataset, which curates a wide range of artistic styles. The editing prompts followed the same format as before, instructing the AI to change the style to a randomly selected, unrelated style drawn from the WikiArt labels.

Both Glaze and Mist protection methods were applied to the images before the edits, allowing the researchers to observe how well each defense could block or distort the style transfer results:

Examples of how protection methods affect style transfer on artwork. The original Baroque image is shown alongside versions protected by Mist and Glaze. After applying Cubism style transfer, differences in how each protection alters the final output can be seen.

The researchers tested the comparisons quantitatively as well:

Protection Method	Dataset	ITAScore	Actual Change ≥ 0
PhotoGuard [15]	Flickr1024	CLIP-S PAC-S++	**77.54%** **68.43%**
Mist [20]	WikiArt	CLIP-S PAC-S++	**54.15%** **56.90%**
Glaze [16]	WikiArt	CLIP-S PAC-S++	**54.12%** **56.90%**

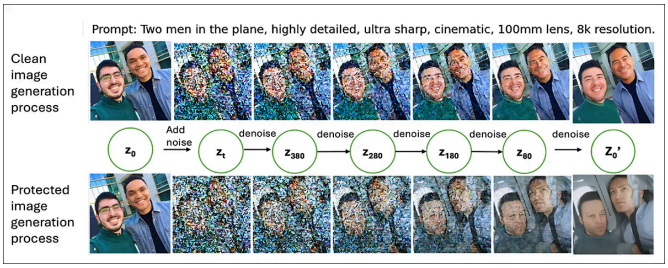
Changes in image-text alignment scores after style transfer edits.

Of these results, the authors comment:

> *'The results highlight a significant limitation of adversarial perturbations for protection. Instead of impeding alignment, adversarial perturbations often enhance the generative model's responsiveness to prompts, inadvertently enabling exploiters to produce outputs that align more closely with their objectives. Such protection is not disruptive to the image editing process and may not be able to prevent malicious agents from copying unauthorized material.*
>
> *'The unintended consequences of using adversarial perturbations reveal vulnerabilities in existing methods and underscore the urgent need for more effective protection techniques.'*

The authors explain that the unexpected results can be traced to how diffusion models work: LDMs edit images by first converting them into a compressed version called a latent; noise is then added to this latent through many steps, until the data becomes almost random.

The model reverses this process during generation, removing the noise step by step. At each stage of this reversal, the text prompt helps guide how the noise should be cleaned up, gradually shaping the image to match the prompt:



Comparison between generations from an unprotected image and a PhotoGuard-protected image, with intermediate latent states converted back into images for visualization.

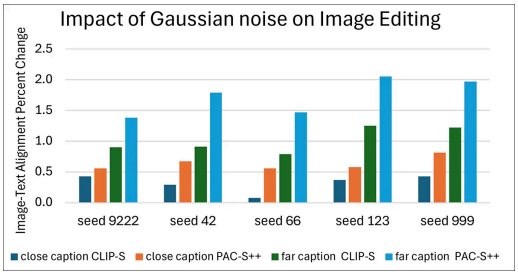
Protection methods add small amounts of extra noise to the original image before it enters this process. While these perturbations are minor at the start, they accumulate as the model applies its own layers of noise.

This buildup leaves more parts of the image 'uncertain' when the model begins removing noise. With greater uncertainty, the model leans more heavily on the text prompt to fill in the missing details, giving the prompt *even more influence than it would normally have*.

In effect, the protections make it easier for the AI to reshape the image to match the prompt, rather than harder.

Finally, the authors conducted a test that substituted crafted perturbations from the *Raising the Cost of Malicious AI-Powered Image Editing* paper for pure Gaussian noise.

The results followed the same pattern observed earlier: across all tests, the Percentage Change values remained positive. Even this random, unstructured noise led to stronger alignment between the generated images and the prompts.



Effect of simulated protection using Gaussian noise on the Flickr8k dataset.

This supported the underlying explanation that any added noise, regardless of its design, creates greater uncertainty for the model during generation, allowing the text prompt to exert even more control over the final image.

Conclusion

The research scene has been pushing adversarial perturbation at the LDM copyright issue for almost as long as LDMs have been around; but no resilient solutions have emerged from the extraordinary number of papers published on this tack.

Either the imposed disturbances excessively lower the quality of the image, or the patterns prove to not be resilient to manipulation and transformative processes.

However, it is a hard dream to abandon, since the alternative would seem to be third-party monitoring and provenance frameworks such as the Adobe-led C2PA scheme, which seeks to maintain a chain-of-custody for images from the camera sensor on, but which has no innate connection with the content depicted.

In any case, if adversarial perturbation is actually making the problem worse, as the new paper indicates could be true in many cases, one wonders if the search for copyright protection via such means falls under 'alchemy'.

First published Monday, June 9, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More