Interpreting Structured Perturbations in Image Protection Methods for Diffusion Models

Michael R. Martin[1]*, Garrick Chan[1], Kwan-Liu Ma[1]

[1] University of California, Davis

* Correspondence: csemartin@ucdavis.edu

VISUALIZATION & INTERFACE DESIGN INNOVATION
UCDAVIS COMPUTER SCIENCE

GitHub | arXiv | Corresponding Author Website

IEEE Advancing Technology for Humanity | CVPR JUNE 3-7, 2026 | DENVER COLORADO

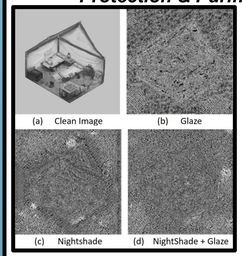
Abstract

Recent image protection studies introduce imperceptible, adversarially designed perturbations intended to disrupt downstream text-to-image generative models. While their empirical effectiveness has been demonstrated, the internal structure, detectability, and representational behavior of these perturbations remain poorly understood. We developed a systematic explainable AI analysis using a unified framework that integrates white-box feature inspection and signal-level probing. Our analysis frames purification-based detection as a mechanistic interpretability problem, revealing how structured perturbations interact with learned feature hierarchies. Protected images preserve content-driven clustering while introducing method-specific substructure. Detectability is strongly associated with perturbation entropy, spatial deployment, and spectral alignment, with sequential protection amplifying detectable structure. Frequency analysis shows energy redistribution along image-aligned axes rather than diffuse noise. Our results suggest that image protection operates through structured feature-level deformation rather than semantic displacement. This work advances the interpretability of adversarial image protection and informs the design of future defenses and detection strategies for generative AI systems.

Motivation

- Generative models train on scraped artwork
- Artists face style mimicry and unauthorized learning
- Protection tools introduce adversarial perturbations
- Purification systems now attempt perturbation removal

Protection & Purification



PROTECTION:

$x' = x + \hat{\delta}$

Glaze – Defensive *style* cloaking

Nightshade – *Concept* poisoning

PURIFICATION:

LightShed – Reconstruction-based purification:

$\hat{x}$ clean = x − $\hat{\delta}$

Hypothesis

Protection perturbations are subtle but structured

- Brighter Pixel-wise Δ from clean (Difference)
  Brighter → Larger perturbation
- Method-specific structure
- Not uniform random noise

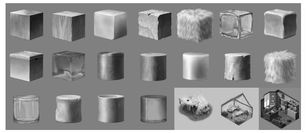
Research Questions

Prior evaluations: Does protection work?
How is protection represented, detected, reconstructed?

- **RQ1**. How are clean vs. poisoned images represented in latent space?
- **RQ2**. What internal features and neurons drive poison detection?
- **RQ3**. Which perturbation patterns can evade detection?

Addressed through white-box and black-box XAI signal analysis.

Dataset



42 images used for black-box analysis

Curated pool of:
- 21 digital illustrations
- 21 stylized 3D renders

Total white-box set: 36 images:
- 9 base images selected for white-box analysis
- 4 aligned variants per image

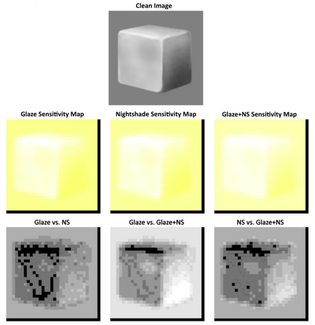
Perturbation Categories:
- Clean
- Glaze
- Nightshade
- Nightshade→Glaze

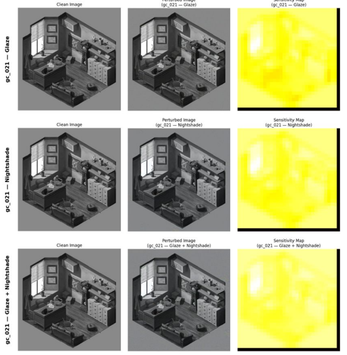
Black-Box Analysis

Occlusion-based sensitivity maps reveal that perturbations remain spatially anchored to underlying image geometry rather than behaving as diffuse noise fields.

- Difference maps expose method-specific spatial structure
- Responses concentrate along edges, curvature, and illumination gradients
- Sequential perturbations amplify localized structural variation

Spatial Sensitivity (Pixel-Level)

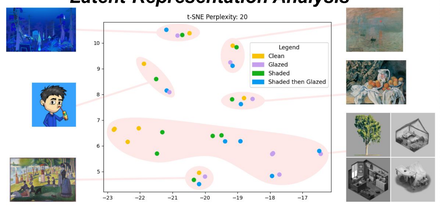




White-Box Analysis

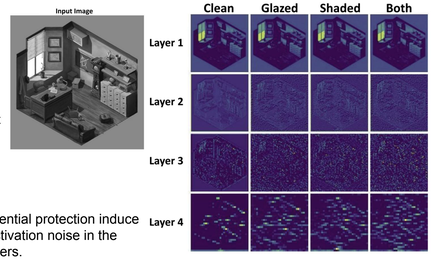
Latent-Representation Analysis

- Images cluster primarily by semantic content
- Protection methods create local substructure
- Sequential poisoning follows final-stage dominance



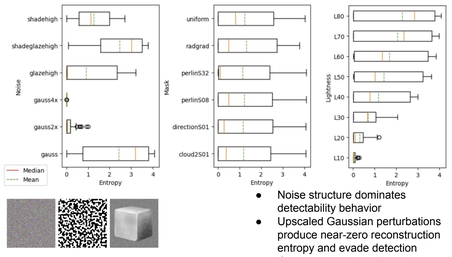
Internal Activation Behavior

- Strongest perturbation response emerges in mid-level layers
- Glaze and sequential protection induce strongest activations
- Deep layers exhibit reduced perturbation sensitivity



Glaze and sequential protection induce the strongest activation noise in the intermediate layers.

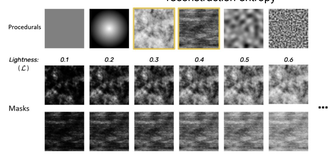
Detectability & Entropy



Mask	Entropy	Detect
Uniform	1.231	60.4%
Radial Gradient	1.311	60.4%
Clouds2	1.177	64.6%
Directional	1.152	62.5%
Hi-Freq Perlin	**1.143**	**52.1%**
Low-Freq Perlin	1.208	62.5%

Noise	Entropy	Detect
Gauss	2.418	77.1%
Gauss-2x	0.164	41.7%
Gauss-4x	**0.002**	0%
Glazed	0.908	50.0%
Shaded	1.271	93.8%
S+G	2.460	100%

$\mathcal{L}$	Entropy	Detect
0.1	**0.048**	**30.6%**
0.2	0.303	50.0%
0.3	0.700	61.1%
0.4	1.166	63.9%
0.5	1.416	72.2%
0.6	1.688	72.2%
0.7	2.049	66.7%
0.8	2.261	66.7%

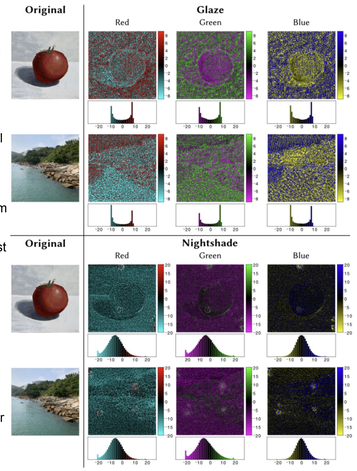


- Noise structure dominates detectability behavior
- Upscaled Gaussian perturbations produce near-zero reconstruction entropy and evade detection
- Glaze produces moderate detectability
- Nightshade and sequential protection yield the highest entropy and detection rates
- - Increasing mask lightness monotonically increases reconstruction entropy

Per-Channel Pixel Differences

Pixel-value differences were computed between original and protected images in **8-bit RGB space**.

- Glaze and Nightshade introduce measurable but visually subtle per-channel perturbations.
- Difference maps reveal structured local variation rather than uniform random noise.
- Histograms show that most RGB changes remain concentrated near small magnitude values.
- Channel-wise changes differ across red, green, and blue channels, indicating non-uniform color-space perturbation behavior.
- Nightshade shows broader and more spatially distributed channel differences than Glaze in the shown examples.

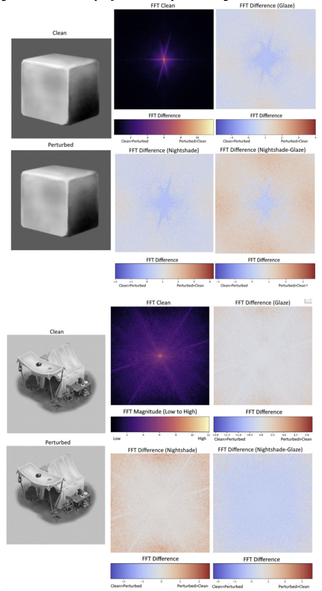


➔ These results support the claim that protection signals are low-magnitude but structured across both space and color channels.

Frequency-Domain (Spectral) Analysis

Fourier-domain analysis reveals geometry-aligned spectral redistribution across all evaluated protection methods.

- Glaze and Nightshade suppress low-frequency energy near the DC component
- Perturbations introduce structured high-frequency components
- Spectral redistribution remains aligned with underlying image geometry
- Sequential protection amplifies spectral energy while preserving directional structure



Conclusion

Protection perturbations are not random noise. Across latent, activation, spatial, spectral, and RGB-channel analyses, Glaze and Nightshade produce structured, low-magnitude signals that remain coupled to image content.

Purification succeeds because these perturbations expose learnable spatial, spectral, and entropy-based regularities.

This directly reflects the camera-ready conclusion: latent organization remains content-driven, perturbation energy remains geometrically anchored and spectrally aligned, and detectability depends on entropy magnitude, spatial deployment, and spectral alignment.

References

- Michael R. Martin, Garrick Chan, and Kwan-Liu Ma. Interpreting structured perturbations in image protection methods for diffusion models. arXiv preprint arXiv:2512.08329, 2025. (arXiv Preprint)
- Michael R. Martin, Garrick Chan, and Kwan-Liu Ma. Adversarial-perturbation. GitHub repository, 2026. https://github.com/MichaelMartinTech/Adversarial-Perturbation. (GitHub Code)
- Michael R. Martin, Garrick Chan, and Kwan-Liu Ma. "A Mechanistic Analysis of Training-Time Image Protection in Diffusion Models." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2026. Accepted and In Print