A Mechanistic Analysis of Training-Time Image Protection in Diffusion Models

Michael R. Martin[1*] Garrick Chan[1] Kwan-Liu Ma[1]
[1]University of California, Davis, USA
csemartin@ucdavis.edu, garchan@ucdavis.edu, klma@ucdavis.edu

Abstract

Recent image protection mechanisms such as Glaze and Nightshade 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 conduct 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. These 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.

1. Introduction

Diffusion-based text-to-image models are trained on large-scale scraped datasets [1, 3, 6], raising concerns about unauthorized style learning [4, 8, 11, 17, 32]. Protection tools such as Glaze and Nightshade introduce imperceptible perturbations intended to disrupt downstream training. While prior studies demonstrate that both protection and purification methods can succeed empirically [2, 27], far less is known about how these systems operate internally, what features they exploit, and why certain perturbations persist while others may not be as successful [9].

*Corresponding author: csemartin@ucdavis.edu

We shift evaluation from outcome-based metrics toward structural analysis across feature, activation, spatial, and frequency domains, building on our earlier preprint [14]. We conducted a unified explainability framework to analyze protected images across multiple explanatory dimensions, treating protection perturbations as structured signals rather than purely empirical artifacts. Our approach integrated inside-model (white-box) and inside-signal (model-agnostic) perspectives into one pipeline. This includes: **(1)** A unified multi-domain eXplainable AI (XAI) framework for analyzing training-time perturbations across latent, activation, spatial, and spectral domains. **(2)** A structural characterization of Glaze, Nightshade, and sequential protection under a reconstruction-based purification model. **(3)** A controlled entropy-modulated perturbation intervention designed to isolate the structural variables governing detectability under entropy-based purification.

2. Background and Related Work

2.1. Adversarial Protection in Diffusion Models

Adversarial protection methods such as Glaze [25, 26] and Nightshade [28, 30] perturb training images so that downstream text-to-image diffusion models acquire distorted internal representations. Unlike classical adversarial examples, which typically optimize perturbations for a fixed classifier and a specific decision boundary, these methods target future training under unknown architectures and hyperparameters. Their objective is not immediate misclassification, but long-term representational degradation during large-scale model training [1, 3]. In diffusion pipelines, this matters as images are typically encoded into a latent space, iteratively denoised under text conditioning, and then decoded back to pixels [16, 19]; training-time perturbations can therefore influence what gets encoded and what denoising dynamics reinforce over many updates [20, 21, 23].

Recent studies have evaluated the robustness and limitations of such protection mechanisms [2, 5, 9, 13]. While empirical results demonstrate that perturbations can influence generative outputs, the internal mechanisms governing their persistence and detectability remain underexplored.

However, the intersection of XAI, spectral analysis, and data poisoning for generative models remains relatively underexplored. There are comparatively limited studies that treat protection perturbations as analyzable signals in both spatial and frequency domains, connecting those signals to internal activations and latent-space structure of detection models, and leveraging XAI methodology to characterize how purification systems exploit low-entropy patterns to reconstruct or remove perturbations.

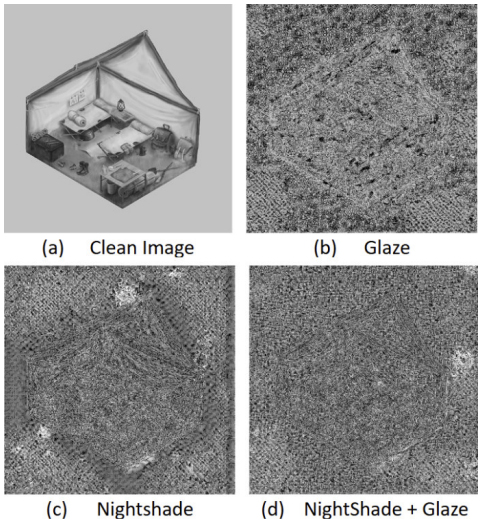

Figure 1. Comparison of perturbations produced by Glaze and Nightshade applied to the original image (a). Panels (b,c,d) visualize the pixel-wise difference between the clean image and the protected outputs. Brighter intensities correspond to larger perturbation magnitudes relative to the clean image.

2.2. Vision Explainability and Spectral Analysis

XAI in vision has traditionally focused on classification models, providing tools to inspect internal features, spatial attributions, and concept-level representations. Techniques such as Grad-CAM [24], integrated gradients [29], feature visualization [18, 34], and concept activation methods [7] reveal spatial and semantic structures learned by convolutional networks. Early deconvolution-based visualization work demonstrated that internal convolutional representations can be inverted to expose hierarchical feature structure from edges to object parts [34]. Concept-based explanation methods such as TCAV extend attribution beyond pixels by measuring directional sensitivity of internal representations to human-interpretable semantic concepts [7], enabling analysis at the representation-subspace level.

Beyond spatial attribution, prior studies show that convolutional models exhibit structured frequency sensitivity, and adversarial perturbations often occupy specific spectral bands [33]. Recent analyses of perturbation-based protection further suggest that spectral structure may influence purification outcomes [5].

In this study, we extended these perspectives beyond outcome-based evaluation. We integrated latent clustering, activation analysis, occlusion probing, and Fourier characterization to analyze how protection perturbations are represented and detected within a reconstruction-based purification model.

3. Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a large-scale diffusion training dataset, where $x_i \in \mathbb{R}^{H\times W\times 3}$ represents an image paired with text y_i. Protection methods modify a subset of images x_j by introducing a perturbation δ_j, yielding the protected image x'_j and preserving visual similarity to the original image (Eq. 1).

$$x'_j = x_j + \delta_j \tag{1}$$

Purification and detection systems introduce a second optimization objective. Let $\mathcal{P}(x)$ denote the output of a learned purification operator applied to an image prior to training. We model protection–purification as a triadic interaction: perturbations are added under perceptual constraints, purification attempts to reconstruct and remove them, and detection operates on measurable signal properties such as entropy or reconstruction error. Concretely, the triadic objectives are:

- **Protection Objective:** construct $x' = x + \delta$ such that x and x' are perceptually indistinguishable, while inducing persistent distortion during downstream training [9].
- **Purification Objective:** given x', estimate $\hat{\delta}$ such that $x_{\text{clean}} \approx x' - \hat{\delta}$, while minimizing false positives.
- **Detection Objective:** decide whether an input contains a perturbation based on measurable signal properties (e.g., entropy, frequency distribution, or reconstruction error).

3.1. Scope, Assumptions, and Research Questions

This study focuses exclusively on training-time perturbation defenses for diffusion-based text-to-image generative models [20, 21]. Inference-time adversarial examples and output filtering mechanisms are outside the scope of this work. Moreover, the purification system is modeled as a learned detection–reconstruction operator rather than a handcrafted signal-processing filter.

Based on this formulation, we defined three research questions designed to isolate complementary aspects of the protection–detection interaction:

RQ1 – Representation. How do detection models represent clean images versus protected images in latent feature space? This question examines whether protection induces distinct geometric structure in the learned embedding space of a purification model, or whether protected samples remain primarily clustered according to original visual semantics.

RQ2 – Detection Mechanism. Which internal features and activation patterns are used by detection models to identify perturbations? Here, "mechanism" refers to the layer- and channel-level activation dynamics through which the purification network separates perturbation signals from semantic image content.

RQ3 – Signal Design and Detectability. Which perturbation signal characteristics govern detectability and reconstructability under entropy-based purification?

4. Methodology

4.1. Experimental Overview

We implemented a controlled multi-stage pipeline operating on paired clean and protected images. This pipeline has two complementary branches: (i) a white-box branch that inspects LightShed's [5] internal representations (latent bottleneck embeddings, encoder activations, and reconstruction-entropy signals), and (ii) a model-agnostic branch that analyzes perturbations directly in image space via occlusion sensitivity and Fourier-domain structure. Latent clustering addresses **RQ1** (representation), layer-wise activation analysis addresses **RQ2** (internal detection mechanism), and entropy-controlled perturbation experiments address **RQ3** (signal-level determinants of detectability).

4.2. Datasets and Perturbations

We curated a visually diverse image pool spanning illustrative and stylized 3D content (Figure 2). From this pool, nine base images are selected for detailed white-box analysis. For each selected image, four aligned variants are generated: (1) clean, (2) Glaze-protected, (3) Nightshade-protected, and (4) sequentially protected (Nightshade $\rightarrow$ Glaze), yielding 36 total images for paired latent-space and activation analysis (Sections 5.1–5.2). Additional public-domain images, distinct from the curated pool in Figure 2, are included for qualitative validation of structural trends in later analyses.

4.3. Protection and Purification Models

This section defines the evaluation setup used to examine representation shifts, activation behavior, and detectability under a fixed reconstruction-based purification model, where all models are treated as pre-trained components without fine-tuning.

Protection Models. We evaluated Glaze and Nightshade using their publicly released implementations and default configurations [25, 28]. For each clean image x, Glaze produces a protected variant x_G, and Nightshade produces x_N. In addition, a composite x_{NG} is generated by sequentially applying Nightshade followed by Glaze under identical resolution and color constraints. All protection methods are executed under their publicly recommended perceptual settings [26, 30].

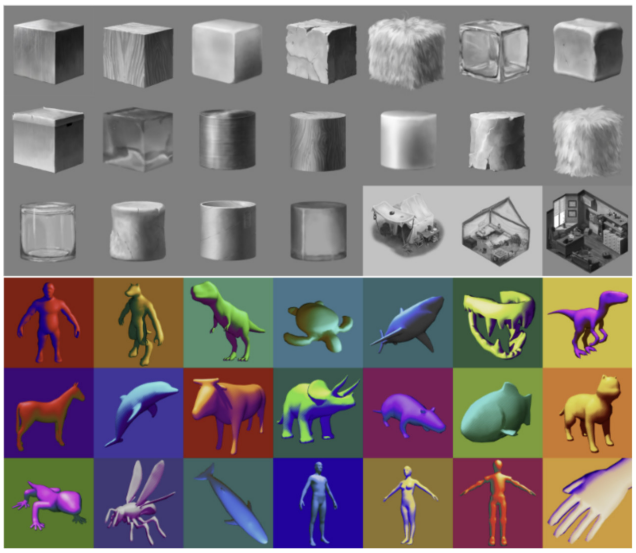

Figure 2. Curated image pool consisting of 21 digital illustrations (top) and 21 stylized 3D renders (bottom) used across representational and signal-level analyses.

Purification Model. Purification and detection are performed using LightShed [5], an entropy-based detection–reconstruction architecture.

$$x_{\text{clean}} = x - \hat{\delta} \tag{2}$$

LightShed projects inputs into a latent space that suppresses semantic content and isolates structured perturbation signals, reconstructs an estimated perturbation $\hat{\delta}$, computes a detection score based on reconstruction entropy, and produces a purified image. All inputs are processed under identical inference settings using the original LightShed thresholds without recalibration.

4.4. Latent-Space Clustering and Representational Analysis

Intermediate feature embeddings are extracted from a fixed encoder layer of the purification model for clean, protected, and purified images under identical inference settings. To visualize representational organization, high-dimensional embeddings are projected to two dimensions using t-distributed stochastic neighbor embedding (t-SNE) with fixed initialization and perplexity. Clean, protected, and purified variants are jointly embedded to examine content-level clustering and protection-induced substructure. Robustness is verified across multiple seeds and perplexity values, with consistent topology retained for analysis [31].

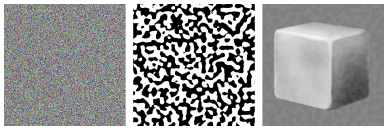

Figure 3. Synthetic perturbation generation via noise–mask composition. From left to right: Gaussian noise field, procedural spatial mask, and the resulting masked perturbation applied to the clean reference image.

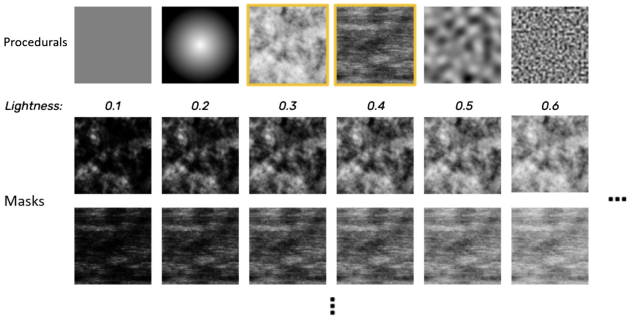

Figure 4. Top: Procedural textures transformed into lightness-controlled masks via gamma correction. Lightness is defined as mean pixel intensity, producing a family of masks spanning low to high average opacity while preserving spatial structure. Bottom: Noise–mask compositing pipeline combining each base image with every noise sample and mask via elementwise multiplication and additive compositing. Outputs are clipped to the valid intensity range prior to LightShed processing.

4.5. Layer-Wise Feature Activation Analysis

Intermediate encoder layers of the purification model are analyzed using forward hooks spanning early, mid-level, and deep representations. Full activation tensors are recorded for clean, protected, and purified images. Channel-wise activation magnitudes and spatial activation maps are computed for each condition. Perturbation-specific responses are isolated using activation difference maps (protected minus clean), enabling depth-wise comparison across early, mid-level, and deep encoder layers of activation amplitude and spatial concentration [24, 29].

4.6. Mask–Noise Entropy Analysis

We designed a controlled family of synthetic perturbation patterns with regulated entropy, spatial dispersion, and amplitude distribution to systematically evaluate how perturbation structure influences detectability and reconstruction behavior. All noise–mask compositions are applied to a fixed clean reference image to ensure that variation across perturbation families arises solely from controlled signal structure. Let $\alpha \in [0, 1]$ denote the perturbation scaling parameter controlling the amplitude of the masked noise component. The opacity parameter α is systematically varied to generate perturbations with progressively increasing Shannon entropy. The resulting images are processed through the fixed detection–purification pipeline used for tool-generated perturbations, yielding detection confidence scores and reconstructed perturbation estimates.

For each perturbation family, detection sensitivity is quantified as a function of mask geometry and noise structure. We report average reconstruction entropy and corresponding detection rates under the fixed LightShed threshold. In addition to mask–noise compositions, we construct an entropy-modulated perturbation family. For each clean image, a local Shannon entropy map is computed from a Gaussian-blurred grayscale version. Gaussian-distributed difference fields at multiple spatial scales are interpolated according to this entropy map and combined with a Perlin-based modulation mask to produce spatially adaptive perturbations. These entropy-modulated perturbations are processed through the same detection–purification pipeline, enabling controlled evaluation of how entropy deployment and spatial distribution influence detectability.

4.7. Occlusion-Based Spatial Sensitivity and Frequency Domain Analysis

We performed an occlusion-based spatial sensitivity analysis operating purely in image space. For each clean–protected image pair, a sliding occlusion window is applied to the clean image, and mean absolute differences between the occluded and corresponding perturbed image are computed to generate spatial sensitivity maps.

Separately, two-dimensional Fourier log-magnitude spectra are computed for paired clean and protected images under identical preprocessing. Signed spectral difference maps highlight frequency bands in which perturbations increase or suppress energy relative to the clean image. Radially averaged frequency profiles provide a direction-invariant summary of global spectral redistribution.

4.8. Implementation Details and Reproducibility

All protection methods are executed using the official public implementations of Glaze and Nightshade under default configurations. The detection–purification model is evaluated strictly in inference-only mode with frozen weights; no fine-tuning or threshold adjustment is applied. All stochastic components (noise generation and any dimensionality-reduction initialization) use fixed random seeds. All images are processed at fixed spatial resolution under identical normalization conventions across experiments.

5. Results and Analysis

5.1. Feature-Space Separation

Across the evaluated image set, images cluster primarily according to their base visual content. Within each base-image cluster, however, consistent protection-specific sub-

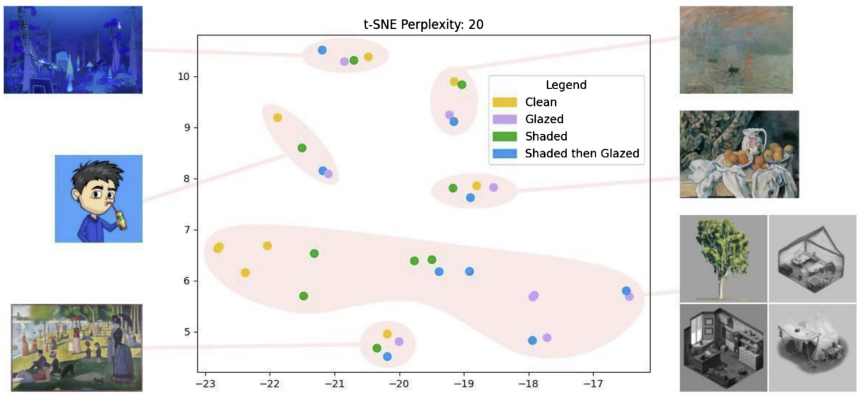

Figure 5. t-SNE visualization of latent representations under different protection conditions. Samples cluster primarily by content, with protection-specific substructure emerging within each group.

Mask	Entropy	Detect	Noise	Entropy	Detect	L	Entropy	Detect
Uniform	1.231	60.4%	Gauss	2.418	77.1%	0.1	0.048	30.6%
Radial Gradient	1.311	60.4%	Gauss-2x	0.164	41.7%	0.2	0.303	50.0%
Clouds2	1.177	64.6%	Gauss-4x	0.002	0%	0.3	0.700	61.1%
Directional	1.152	62.5%	Glazed	0.908	50.0%	0.4	1.166	63.9%
Hi-Freq Perlin	1.143	52.1%	Shaded	1.271	93.8%	0.5	1.416	72.2%
Low-Freq Perlin	1.208	62.5%	S+G	2.460	100%	0.6	1.688	72.2%
						0.7	2.049	66.7%
						0.8	2.261	66.7%

Table 1. Average Shannon entropy of reconstructed perturbations and corresponding LightShed detection rates across all synthetic perturbation families. Results are reported as a function of spatial mask pattern (left), noise type (middle), and mask lightness (right). Lower entropy values indicate reduced reconstructability and are associated with lower detection probability.

structure emerges. Clean images and Nightshade-protected images form distinguishable subclusters, while Glaze-protected images and sequentially protected (Nightshade-glaze) images consistently overlap within a shared subregion. This pattern reflects the empirical dominance of the second applied perturbation, whereby the final protection stage largely determines the visible and latent characteristics of the resulting signal. As a result, sequentially protected images inherit the feature-space behavior of the method applied last. This behavior (Figure 5) is consistent with localized representational shifts within content-driven clusters, in which protection methods alter internal feature geometry without inducing large-scale semantic displacement in embedding space.

5.2. Internal Activation Behavior

Across the evaluated layers (Figure 6), protected images visually induce higher-magnitude and more spatially concentrated activations relative to their clean counterparts. This effect is most pronounced in the second and third encoder layers, which exhibit strong, spatially structured activation patterns in response to protected inputs. In contrast, deeper layers exhibit markedly reduced activation sensitivity to protection, with many channels in the fourth and fifth layers showing little to no measurable response.

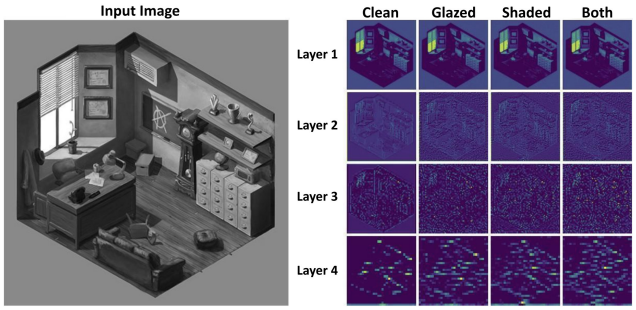

Figure 6. Left: Clean input image of an isometric room. Right: Layer-wise activations of a selected feature channel across LightShed's encoder for the clean image and each poisoning technique. Glaze and sequential (Shaded-Glazed) protection induce the strongest activation noise in the intermediate layers.

5.3. Detectability Trends

5.3.1. Tool-Based Detectability

Detection responses vary systematically across perturbation families under the controlled mask–noise permutation framework. Nightshade-derived perturbations yield high

reconstruction entropy and correspondingly high detection rates, Glaze produces lower entropy and moderate detection, and sequential protection amplifies detectable structure (Table 1).

5.3.2. Synthetic Mask-Noise Detectability

For each input, LightShed reconstructs the estimated perturbation and computes the Shannon entropy of the reconstruction. An image is classified as poisoned when this reconstruction entropy exceeds LightShed's default detection threshold (approximately 0.07 under the default configuration). Average reconstruction entropy and detection rate are reported as a function of mask pattern, noise type, and mask lightness in the analyzed statistics shown in Table 1, along with the accompanying entropy visualization (Figure 7). Across mask geometries, reconstruction entropy remains relatively stable, while noise structure exerts the dominant influence on detectability. Upscaled Gaussian noise produces near-zero entropy and is never detected, whereas Glaze-derived and sequential perturbations produce the highest entropy and detection rates. Increasing mask lightness monotonically increases reconstruction entropy.

5.4. Occlusion-Based Spatial Sensitivity and Frequency Domain Findings

To characterize protection mechanism dependence on image spatial structure, we analyzed occlusion-based spatial sensitivity maps computed from clean–protected image pairs. Across all evaluated examples, perturbations remain spatially anchored to underlying image geometry rather than behaving as independent noise fields (Figure 9). Separately, frequency-domain analysis reveals consistent, geometry-aligned spectral redistribution under all protection methods (Figure 8). Consistent with the spatial difference maps, perturbations concentrate along structurally meaningful surfaces rather than diffusing uniformly, indicating that spectral energy shifts remain coupled to underlying image geometry.

5.5. Entropy-Modulated Perturbation Study

To isolate the structural factors governing detectability under reconstruction-based purification, we conducted controlled entropy-modulated perturbation experiments. The structural analyses in Sections 5.1–5.5 show that contemporary protection mechanisms exhibit consistent regularities across representation, activation, spatial, and frequency domains. Detectability under LightShed is associated with perturbation entropy magnitude, spatial regularity, and frequency redistribution patterns. Protected images remain geometrically anchored to underlying content while exhibiting constrained entropy structure and method-specific spectral signatures. These findings suggest that detectability is not driven by perturbation presence alone, but by structured regularities that reconstruction-based purification can learn and exploit [9]. Motivated by the broader signal-design view of increasing protection–purification developments, we analyzed whether increased cross-image variability, spatial non-uniformity, and Gaussian-like noise structure reduce reconstruction-based detectability [5]. To directly test RQ3, we constructed a controlled entropy-modulated perturbation family that isolates the structural variables identified in Sections 5.3–5.5. This intervention is not proposed as a new protection method, but as a causal probe to examine how entropy distribution, spatial non-uniformity, and frequency allocation influence detectability under reconstruction-based purification [27].

5.5.1. Entropy-Modulated Perturbation Construction

Let i_k denote a clean image normalized to [0,1]. First, a Gaussian-blurred version of the image is converted to grayscale and used to compute a local Shannon entropy map e_k. This map serves as a spatial guide for perturbation allocation. High-entropy regions (e.g., foliage, texture-rich areas) receive higher-frequency perturbation components, while low-entropy regions (e.g., sky, flat surfaces) receive lower-frequency components. Two sets of Gaussian-distributed difference matrices D_1 and D_2 are generated with randomized mean and variance per image and per channel to increase cross-image variability. These matrices are upscaled at different spatial scales and linearly interpolated using the entropy map e_k, producing a spatially adaptive perturbation field D_0. A perlin noise mask is applied multiplicatively to introduce additional intra-image non-uniformity. The final perturbed image p_k is defined in the following equation:

$$p_k = i_k + e_k \cdot D_0 \cdot m_k \tag{3}$$

where m_k denotes the Perlin mask. This construction increases: Variability – via randomized distribution parameters across images and channels; Non-uniformity – via entropy-based spatial modulation; Noise-like structure, via Gaussian sampling and Perlin masking.

5.5.2. Detectability Evaluation Protocol

We evaluated this adaptive perturbation prototype using the same LightShed detection–purification model and entropy-based analysis pipeline employed in Sections 5.3–5.5 (Figure 10). Adaptive perturbations are processed using LightShed with its default frozen weights and original detection thresholds, enabling direct comparison to Glaze, Nightshade, and synthetic noise–mask compositions under a fixed-detector setting.

5.5.3. Structural Implications

The controlled perturbation family enables direct evaluation of RQ3 by manipulating entropy magnitude, spatial alloca-

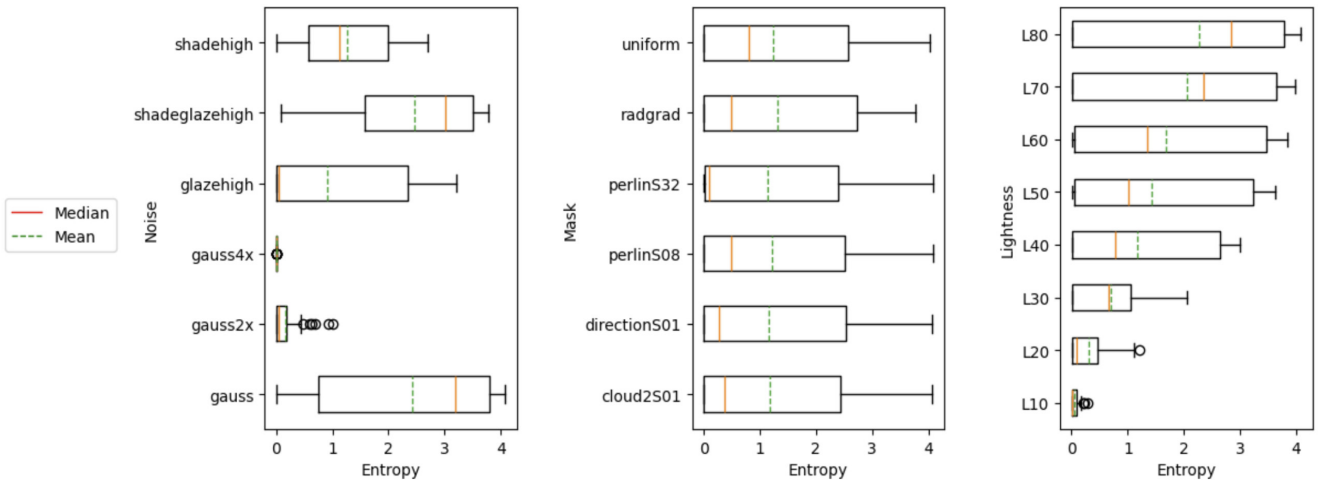

Figure 7. Distribution of reconstructed perturbation entropy across noise type (left), mask geometry (middle), and mask lightness (right). Boxplots summarize all generated samples. Noise structure dominates detectability, while mask geometry induces weaker variation.

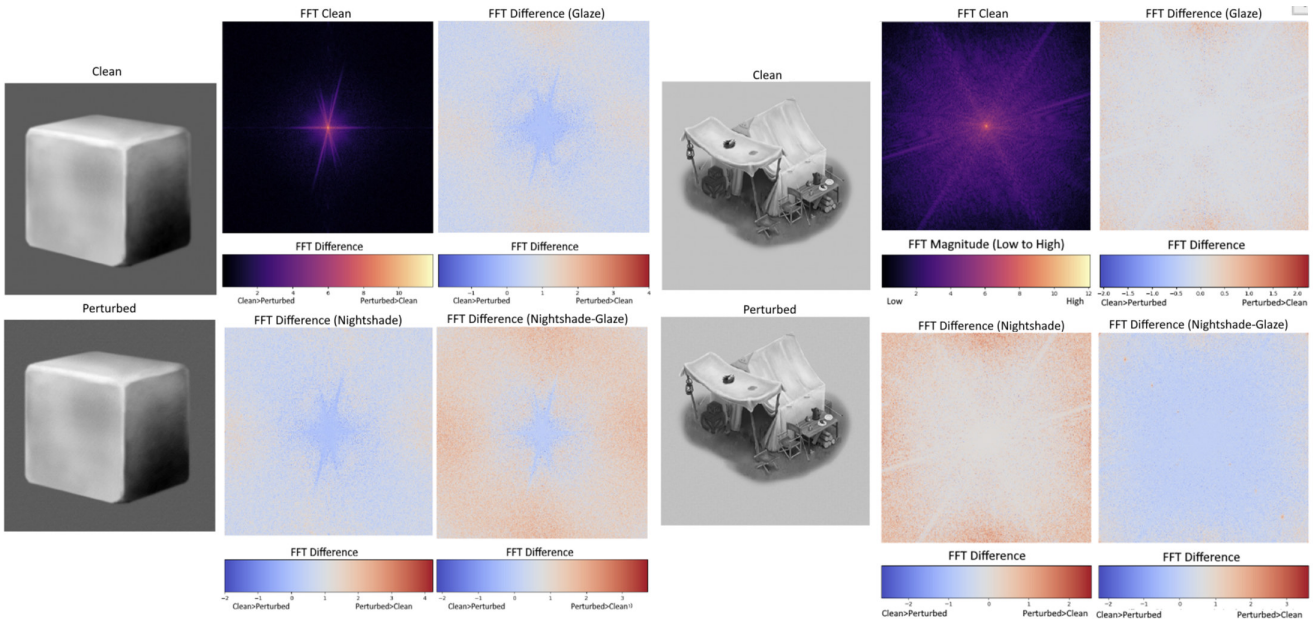

Figure 8. Left: Glaze and Nightshade primarily visually suppress low-frequency energy near the DC component while introducing modest, structure-aligned high-frequency components. Sequential protection amplifies overall spectral energy without disrupting the image's intrinsic directional structure. Right: A similar pattern is observed in a more complex scene, confirming that spectral redistribution remains aligned with underlying geometry rather than diffused across frequencies.

tion, and distributional variability while holding semantic content constant. Under fixed-detector evaluation, entropy-modulated perturbations achieve intermediate detectability between Glaze and Nightshade, suggesting that entropy redistribution alone does not guarantee evasion without altering spectral alignment and spatial anchoring.

6. Conclusion

This study introduces a unified explainable framework for analyzing training-time image protection across representation, activation, spatial, and frequency domains. For RQ1, latent-space analysis demonstrates that Glaze, Nightshade, and their sequential composition preserve content-driven embedding organization while providing method-specific substructure, resulting in constrained representa-

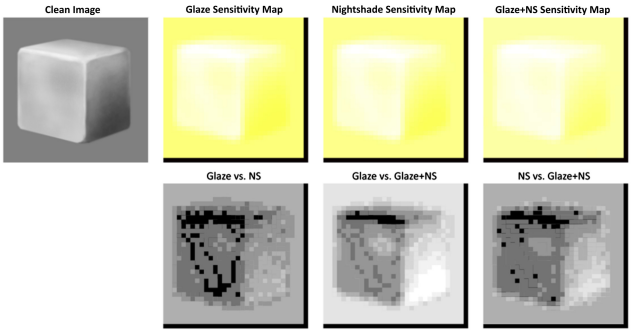

Figure 9. Sensitivity maps and difference maps for Glaze, Nightshade, and Nightshade-glaze on a solid cube illustration. Raw sensitivity maps appear visually similar across perturbations, while difference maps (darker values indicate smaller differences; brighter values indicate larger deviations) highlight method-specific spatial variation.

tional deformation rather than global semantic drift. For RQ2, layer-wise activation, occlusion, and Fourier analyses show that perturbation energy remains geometrically anchored and spectrally aligned with image structure, concentrating along edges, curvature, and illumination gradients rather than behaving as isotropic noise. For RQ3, controlled synthetic and entropy-modulated experiments demonstrate that detectability under reconstruction-based purification is strongly correlated with structured signal properties (specifically entropy magnitude, spatial deployment, and spectral alignment), placing principled constraints on evasion [9].

Our analysis centers on a reconstruction-based, entropy-driven detection architecture and a fixed set of publicly released protection tools [16, 19]. Accordingly, the conclusions apply to purification systems that operate via structured perturbation amplification; diffusion-native or feature-space-only detectors may exhibit different sensitivities.

Limitations and Future Work

While this study analyzes a reconstruction-based purification model and a fixed set of publicly released protection tools and synthetic perturbation families, the observed structural regularities characterize reconstruction-driven detectors and may differ under diffusion-native or feature-space-only defenses. Extending this multi-domain analysis to alternative detector classes would clarify which findings reflect universal signal properties versus architecture-dependent behavior. Future work aims to formalize the spectral and entropy constraints that shape detectability under learned purification systems, utilizing explainability-driven structural analysis as a principled framework for evaluating both protection mechanisms and countermeasures [17, 22].

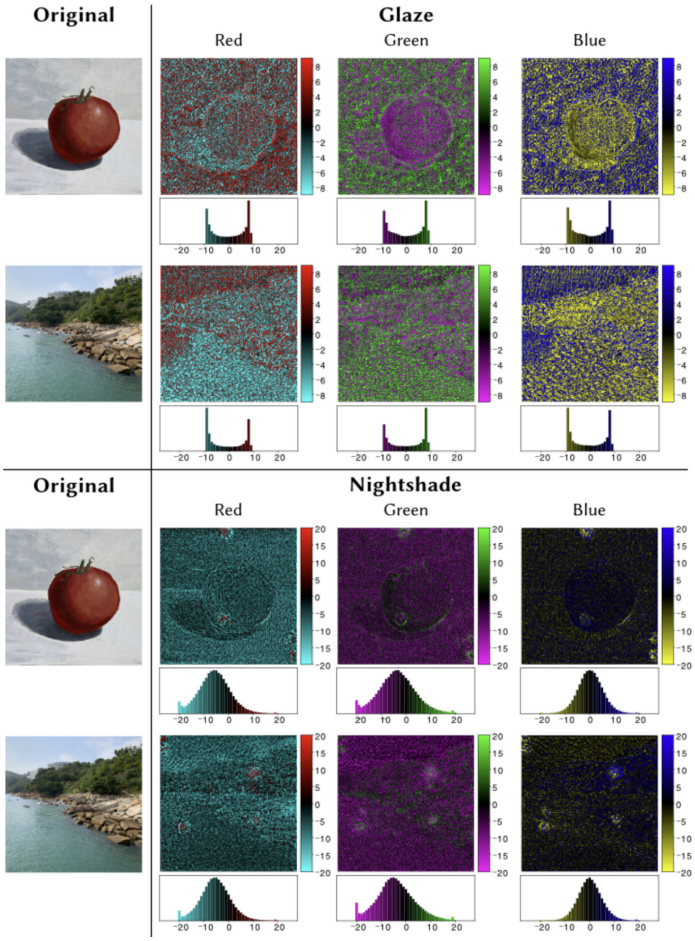

Figure 10. Per-channel (R, G, B) pixel-value differences between the original and perturbed images (Glaze and Nightshade). Differences are computed in 8-bit space. Histograms show the distribution of per-channel changes.

Ethical Considerations

Our analysis examines structural properties of protection and detection mechanisms, which emerged in response to concerns regarding the unauthorized incorporation of creative visual works into large-scale model training datasets [4, 8, 11, 32]. Insights into perturbation structure could potentially inform adversarial misuse. We therefore restrict our evaluation to controlled experimental settings and publicly released protection tools [1, 3]. Future work should consider responsible disclosure and defensive co-design alongside perturbation research [10, 12].

Code Availability

The full code of the analysis framework and experimental pipeline is available at https://github.com/MichaelMartinTech/Adversarial-Perturbation [15].

References

[1] Bobby Allyn. Artificial intelligence web crawlers are running amok. *NPR*, 2024. Online article. Accessed: 2025-10-29. 1, 8

[2] Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. Impress: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai. In *Advances in Neural Information Processing Systems*, 2023. 1

[3] N. S. Chong. Beyond robots.txt: Modern anti-crawler mechanisms. United Nations University Centre for Policy Research (C3) Blog, 2025. Online article. Accessed: 2025-10-29. 1, 8

[4] Benj Edwards. Flooded with ai-generated images, some art communities ban them completely. Ars Technica, 2022. 1, 8

[5] Hannes Foerster, Sadegh Behrouzi, Philipp Rieger, Murtuza Jadliwala, and Ahmad-Reza Sadeghi. Light-shed: Defeating perturbation-based image copyright protections. In *34th USENIX Security Symposium (USENIX Security 25)*. USENIX Association, 2025. 1, 2, 3, 6

[6] 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, 2022. 1

[7] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems 32*, 2019. 2

[8] J. C. A. Guevara. Artist shows evidence that ai is possibly stealing others' art. LevelUp, 2022. Accessed: 2025-10-31. 1, 8

[9] Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative ai. In *International Conference on Learning Representations*, 2025. 1, 2, 6, 8

[10] Victoria Hood. How to opt-out of meta ai. *TechRadar*, 2025. Online article. Accessed: 2025-10-29. 8

[11] Harry H. Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. Ai art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*, pages 363–374, New York, NY, USA, 2023. Association for Computing Machinery. 1, 8

[12] Rebecca Leppert. What we know about energy use at u.s. data centers amid the ai boom. *Pew Research Center, Short Read*, 2025. Published October 24, 2025. Accessed: 2025-10-29. 8

[13] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples, 2023. 1

[14] 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. 1

[15] Michael R. Martin, Garrick Chan, and Kwan-Liu Ma. Adversarial-perturbation. *GitHub repository*, 2026. 8

[16] Midjourney, Inc. Midjourney. Commercial text-to-image generation system, 2025. Official website. 1, 8

[17] Liz Mineo. Is art generated by artificial intelligence real art? *The Harvard Gazette*, 2023. Published August 15, 2023. Accessed: 2025-10-29. 1, 8

[18] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. Online article. 2

[19] OpenAI. Dall·e. Commercial text-to-image generation system, 2023. 1, 8

[20] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 1, 2

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2

[22] Rob Salkowitz. Midjourney founder david holz on the impact of ai on art, imagination and the creative economy. Forbes interview, 2022. Accessed: 2025-10-29. 8

[23] Christoph Schuhmann, Romain Beaumont, Robin Vencu, Cade W. Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R. Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022. 1

[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 2, 4

[25] 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)*, Anaheim, CA, 2023. USENIX Association. 1, 3

[26] Shawn Shan, Jennifer Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Frequently asked questions (faq) — glaze: Protecting artists from generative ai. Official FAQ for Glaze, 2023. 1, 3

[27] Shawn Shan, Stanley Wu, Haitao Zheng, and Ben Y. Zhao. A response to glaze purification via impress, 2023. 1, 6

[28] Shawn Shan, Wenbo Ding, John Passananti, Shiyu Wu, Haitao Zheng, and Ben Y. Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 807–825, 2024. 1, 3

[29] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th In-*

ternational Conference on Machine Learning (ICML), pages 3319–3328, 2017. 2, 4

[30] University of Chicago. Nightshade user guide. Official documentation for Nightshade, 2024. 1, 3

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 3

[32] Chloe Xiang. Artists are revolting against ai art on artstation. VICE (Motherboard), 2022. 1, 8

[33] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems*, pages 1319–1329, 2019. 2

[34] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer. 2