Visualizing the Impact of Data Perturbation on Image Models using Explainable AI

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

[1] University of California, Davis – Visualization & Interface Design Innovation (VIDi) Laboratory

*Corresponding Author: csemartin@ucdavis.edu

December 5, 2025

UCDAVIS COMPUTER SCIENCE

VISUALIZATION & INTERFACE DESIGN INNOVATION

Motivation

Generative Image Models Are Trained on Scraped Art at Planetary Scale

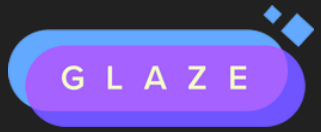
- Text-to-image systems (Stable Diffusion, DALL•E, Midjourney) rely on massive, scraped image–text compilations
- Training data overwhelmingly collected without artist consent or compensation
- Models can learn an artist's style from as few as a handful of examples

This Creates a New Technical Vulnerability:

- Artists are now exposed to:
  - Style Mimicry
  - Signature Reproduction
  - Dataset Contamination

- Training pipelines assume benign data
- That assumption is now false

Two Fundamentally Different Production Philosophies



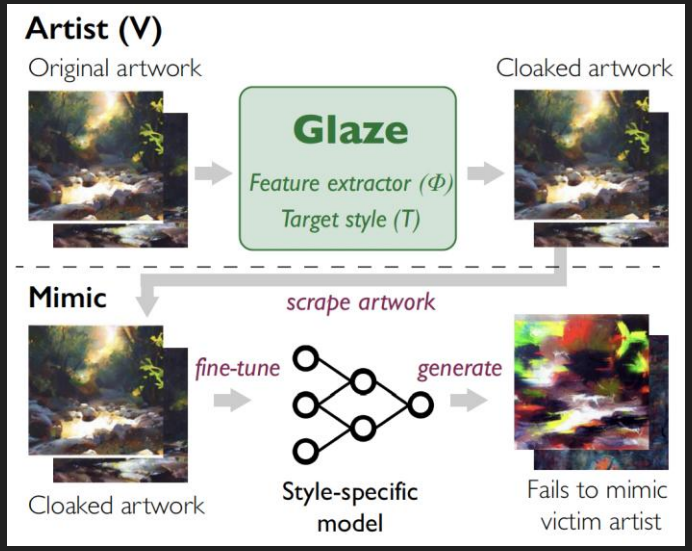
Defensive Style Cloaking Shan et al. (2023)

- Perturbs artwork toward a **surrogate style**
- Preserves content but **disrupts style learning**
- Acts in feature space, not pixel space
- **Goal**: Prevent style mimicry

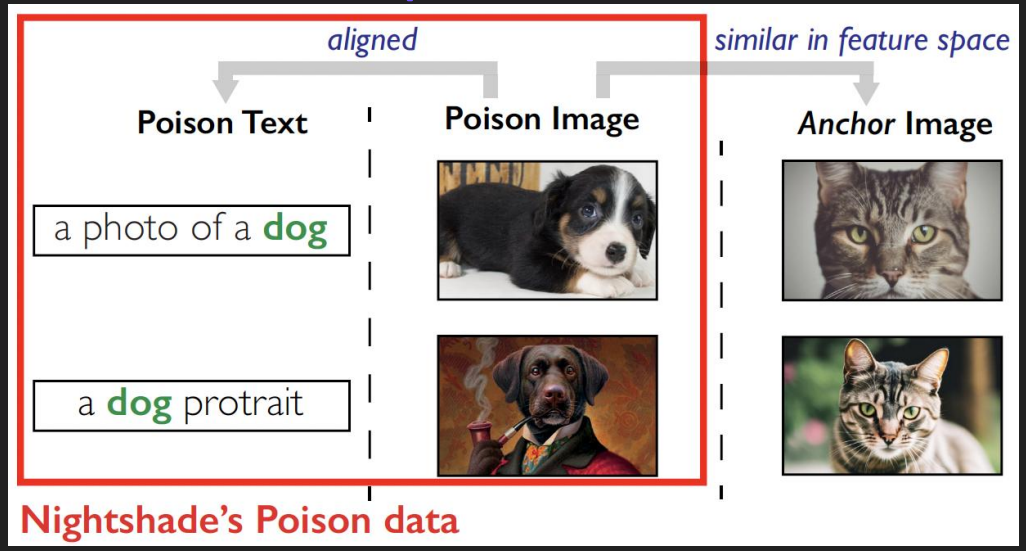
Offensive Concept Corruption Shan et al. (2024)

- Prompt-specific poisoning
- Misaligns visual features and semantic labels
- Corrupts the model's internal concept representation
- Goal: Destroy class-level understanding

Mechanistic Formation of Feature-Space Poisons



Shan et al. (2023)

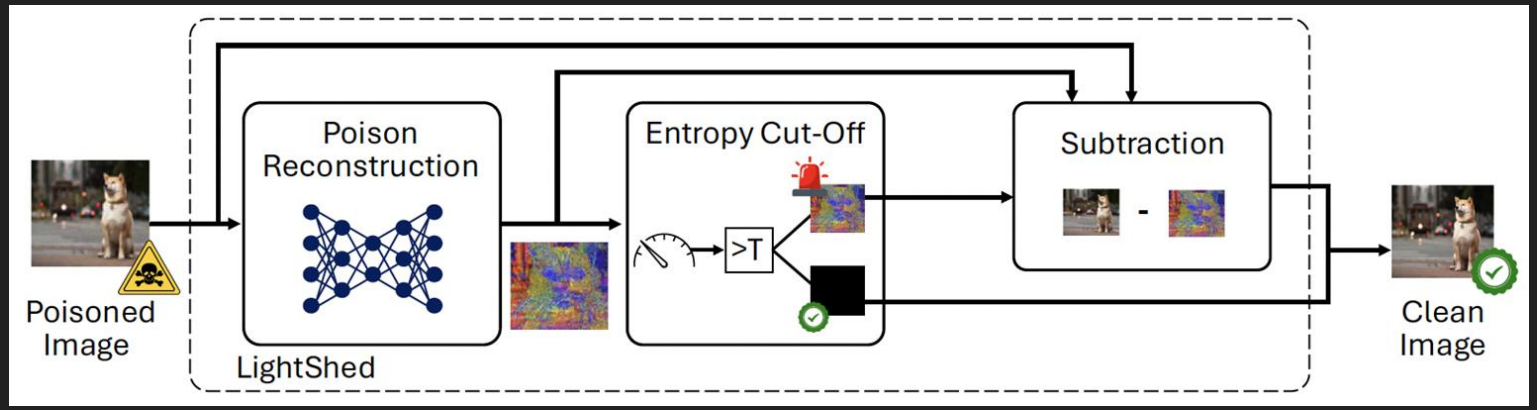


Shan et al. (2024)

- Both methods operate in feature space, not pixel space
- Glaze maximizes feature-space displacement toward surrogate styles
- Nightshade enforces cross-modal misalignment during training
- These mechanisms inject structured, low-entropy perturbations
- Distinct objectives imply distinct internal & spectral signatures

The Adversarial Arms Race: Purification Emerges

- Protection methods rapidly triggered **automated purification**
- Early countermeasures: **noisy upscaling, filtering, reconstruction**
- **LightShed**: autoencoder for detection, reconstruction, and removal
- Detection based on **entropy of reconstructed perturbation**
- Establishes purification as the current state of the art

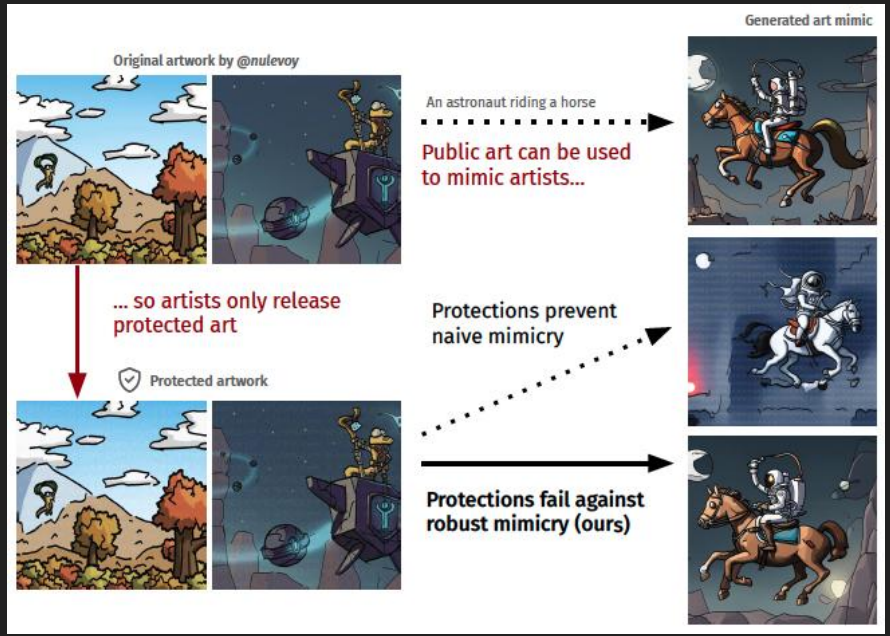


Foerster et al. (2025)

Limits of Current Protection and Purification

- Public art enables **robust style mimicry**
- Protections prevent **naïve imitation**
- Protections fail under **robust training pipelines**

- Shows limits of **both poisoning and purification**
- Motivates need for **interpretability, not just detection**



Hönig et al. (2024)

⇒ Few interpretability/XAI studies on protection tools/workarounds

Research Gaps

Prior work evaluates **whether poisoning succeeds or purification removes it.**

Missing understanding of:

- How poisoned images are represented **inside** detection models
- What structured signal the perturbations **inject into** the image itself

No prior work unifies:

- White-box neuron and latent-space analysis
- Black-box spatial sensitivity
- Black-box frequency-domain characterization

Empirical behavior rather than grounded in interpretable explainability lens.

Research Questions

- 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

Overall Analysis Pipeline

Clean images → poisoned with Glaze, Nightshade, Nightshade→Glaze

White-box branch (LightShed):

- Latent bottleneck embeddings (t-SNE)
- Layer-wise encoder activations
- Reconstruction entropy for detectability

Black-box branch (model-agnostic):

- Spatial sensitivity via occlusion maps
- Frequency structure via FFT and radial spectra

Integrates **inside-model** and **inside-signal** perspectives

Dataset and Poisoning Protocol

9 total images spanning distinct artistic styles

4 variants per image:

- Clean
- Glaze
- Nightshade
- Nightshade → Glaze

Default protection settings for realism and minimal visual distortion

All variants processed through a uniform analysis pipeline.

Latent Representation of Clean Vs. Poisoned Images

Encoder **bottleneck embeddings** extracted from LightShed

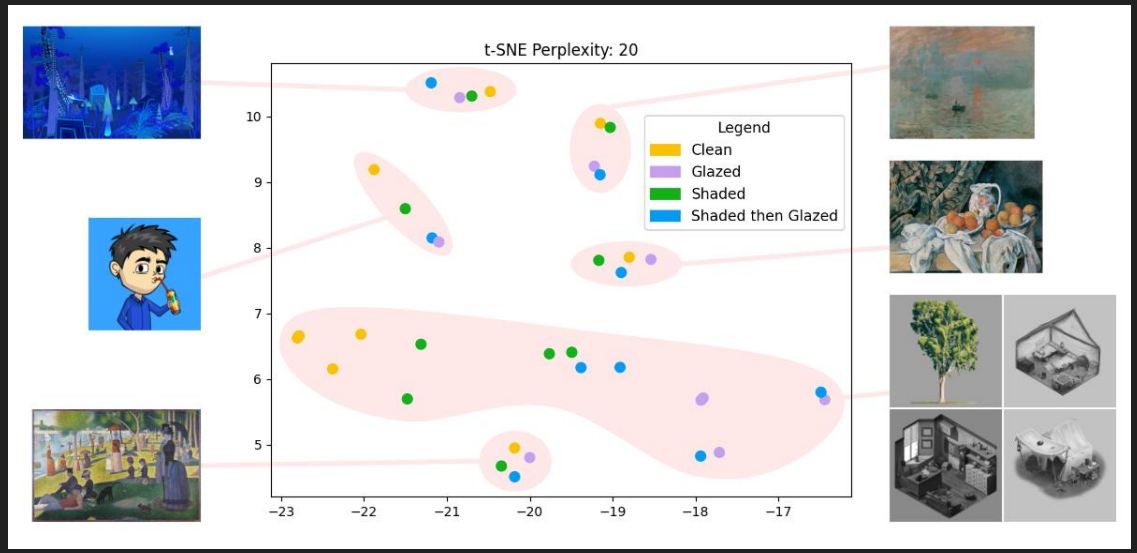
t-SNE used for 2D projection of latent space

Compares Clean, Glaze, Nightshade, Nightshade → Glaze

Tests whether poisoning creates **separable latent structure**

Latent Clustering in Bottleneck Space

- Clean images form tight clusters by artistic style
- Nightshade variants separate clearly from clean images
- Glaze variants form distinct subclusters within each style
- Nightshade → Glaze overlaps most closely with Glaze-only
- No consistent global shift vector across poisoning methods

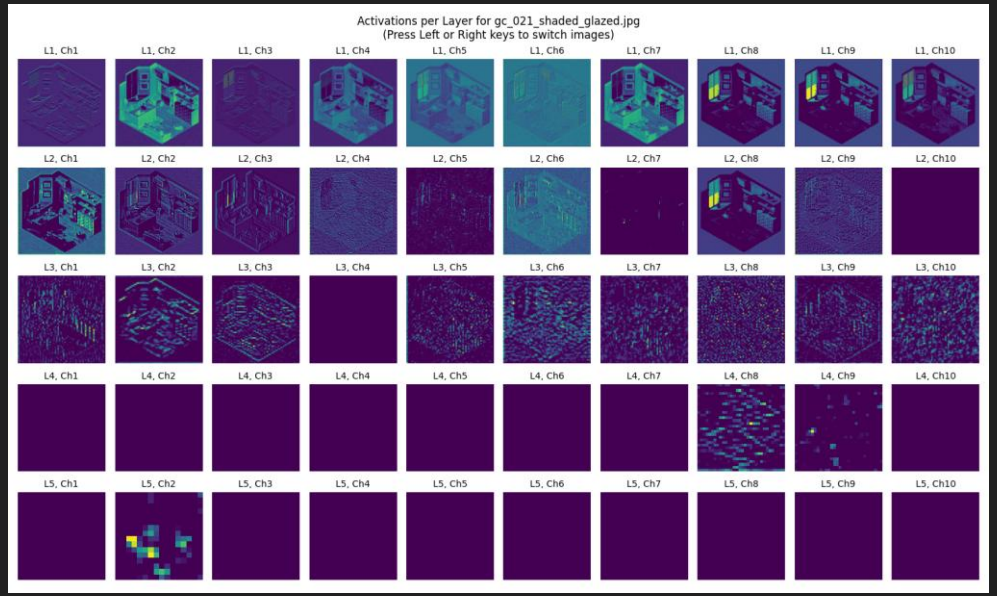


Feature Activation

- Analyze LightShed encoder convolutional layers
- Visualize channel-wise activation maps
- Track where perturbation responses first emerge
- Localize layers and neurons driving reconstruction
- Links internal activation patterns to detectability

Feature Activation

Identify layers/neurons responsible for poison reconstruction



Layer-Wise Activation Emergence

Layer 1 - minor differences

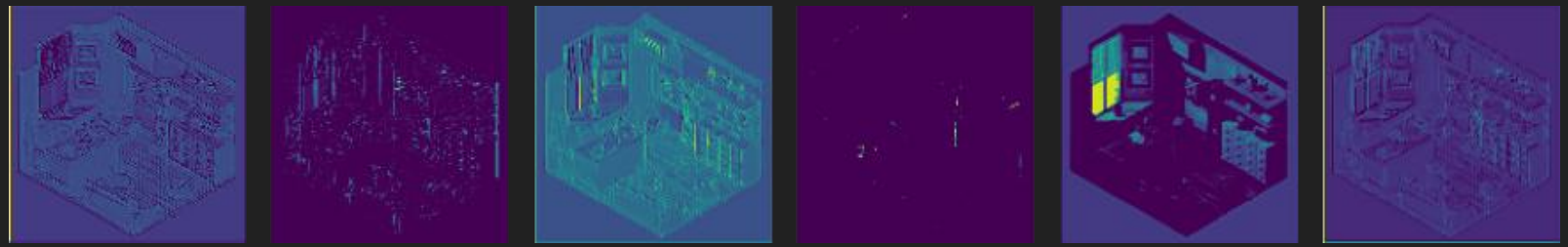
Clean

Shaded + Glazed

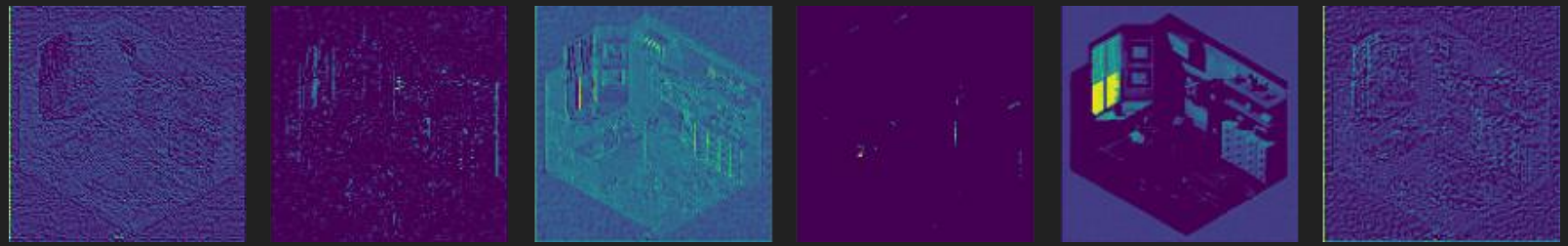
Layer-Wise Activation Emergence

Layer 2 - noticeable differences

Clean



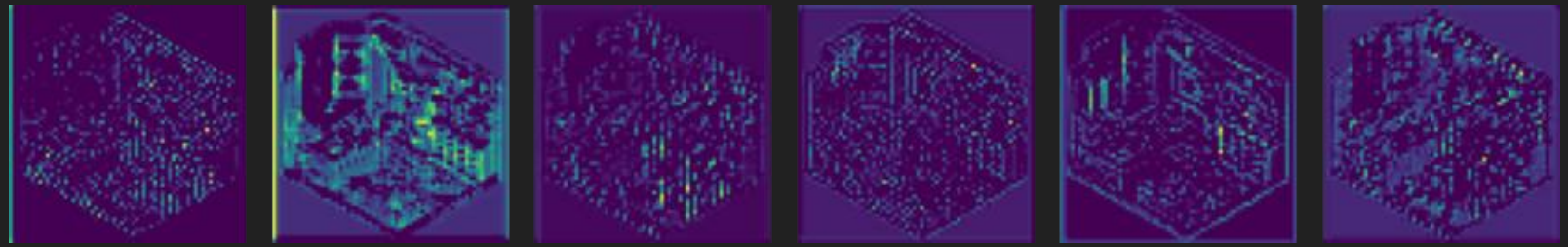
Shaded + Glazed



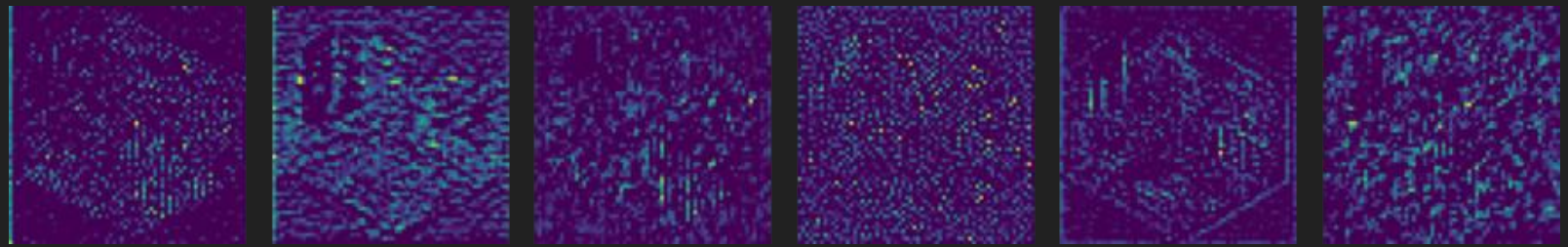
Layer-Wise Activation Emergence

Layer 3 - clear differences

Clean



Shaded + Glazed



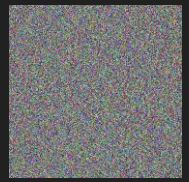
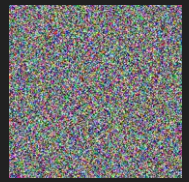
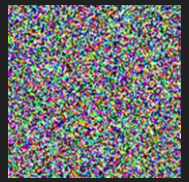
Detectability, Entropy & Evasion

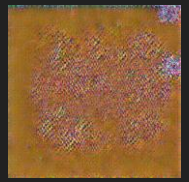
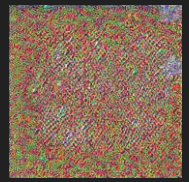
- Study when **perturbations trigger detection**
- LightShed detects based on **reconstruction entropy**
- Test synthetic noise and **procedural perturbations**
- Vary mask lightness and **spatial structure**
- Quantify **detectability vs. entropy relationship**

Experimental Design: Base, Noise & Masks

Base Image

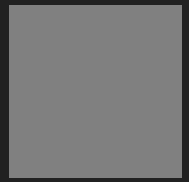
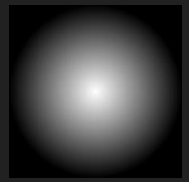
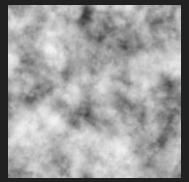
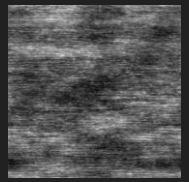
Noises

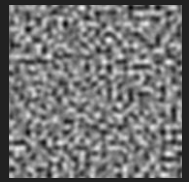




Gauss & Upscaled Gauss *Glaze & Nightshade Noise*

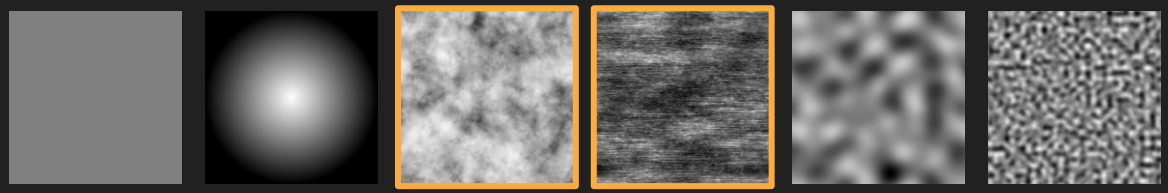
Procedurals



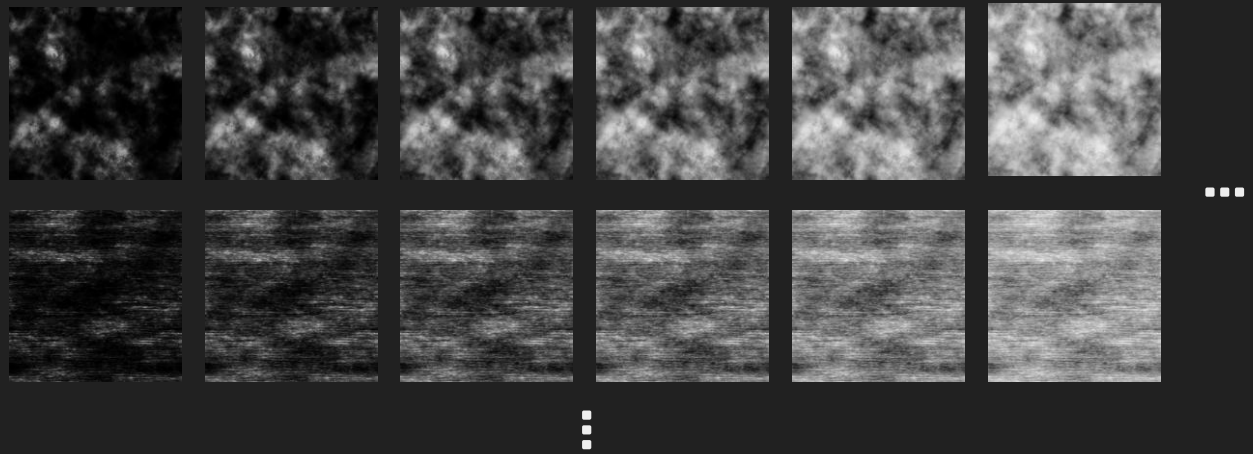
Reducing Detectability

Procedurals



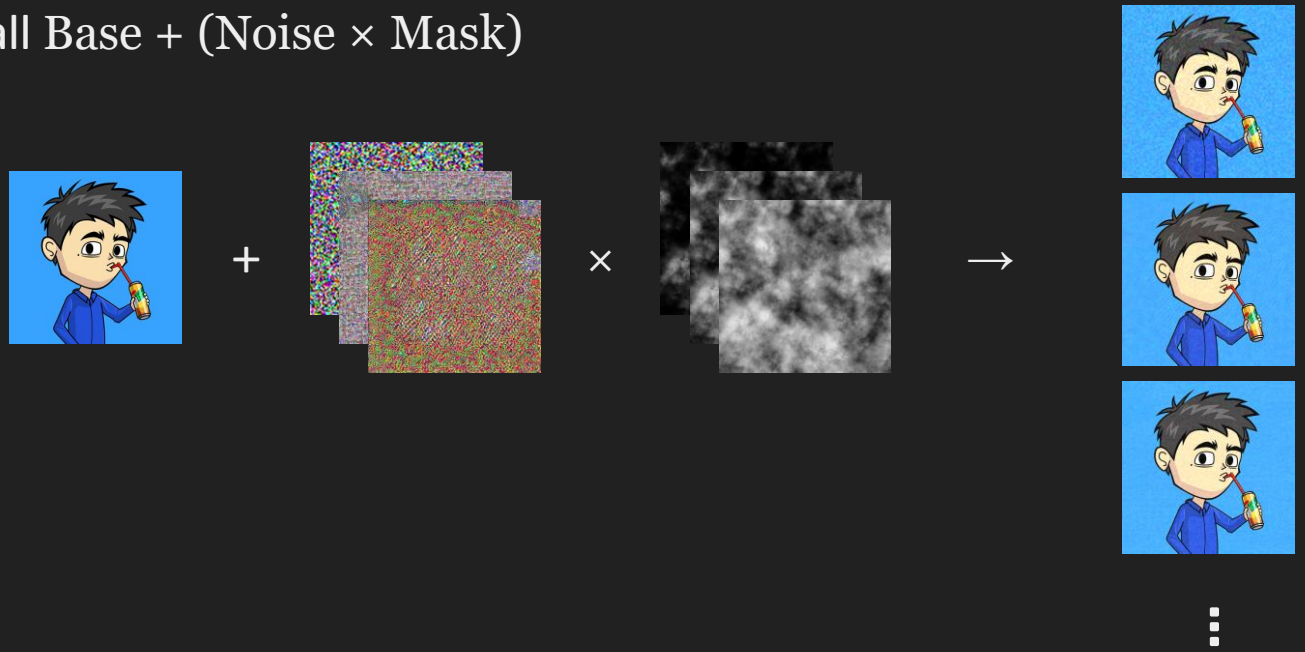
Lightness: *0.1* *0.2* *0.3* *0.4* *0.5* *0.6*

Masks



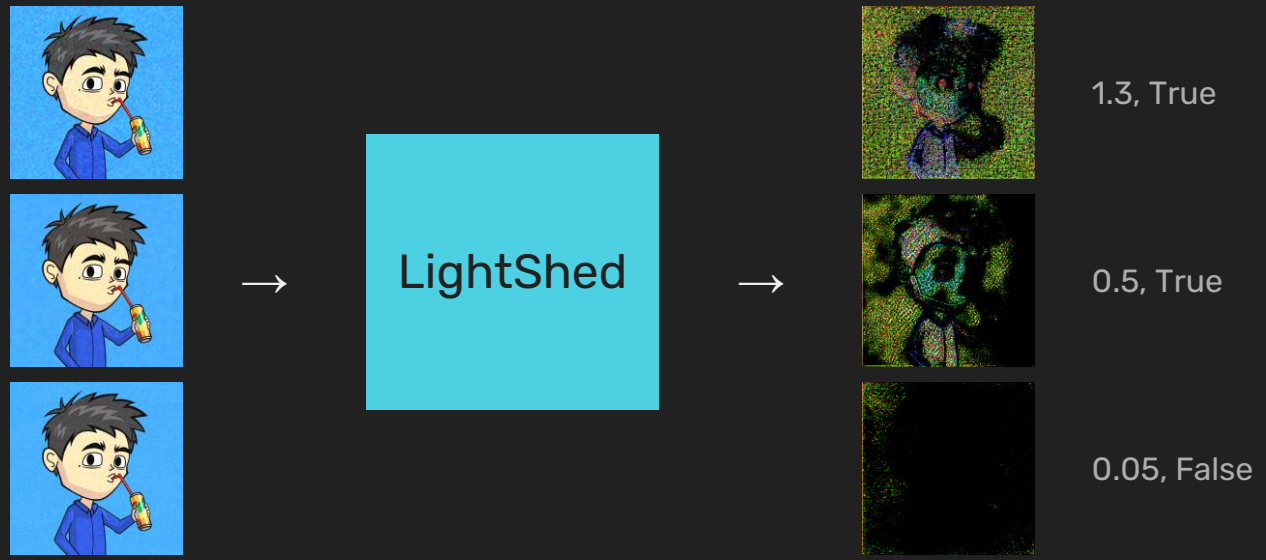
Reducing Detectability

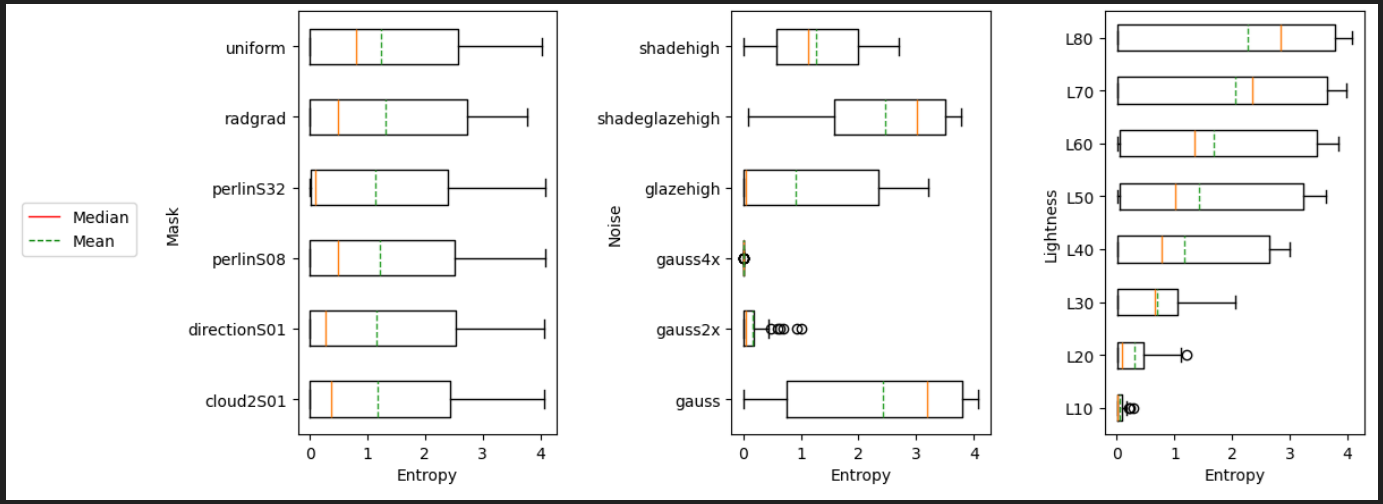
Permute all Base + (Noise × Mask)



Reducing Detectability

Process with LightShed
Return Entropy & Detected (> 0.07) for each image





Mask	Entropy	Detect	Noise	Entropy	Detect	$\mathcal{L}$	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%

Detectability vs. Entropy

- Upscaled Gaussian noise shows lowest detection rate
- Glaze and Nightshade noise detected most reliably
- Detection peaks at mid-range mask lightness (~0.5–0.6)
- Very low and very high densities reduce detection reliability
- Confirms entropy alone is not sufficient for detectability

Mask	Entropy	Detect	Noise	Entropy	Detect	$\mathcal{L}$	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%

Black-Box Analysis #1: Spatial & Frequency Structure

Analyze perturbations using input–output only methods

- No access to model internals

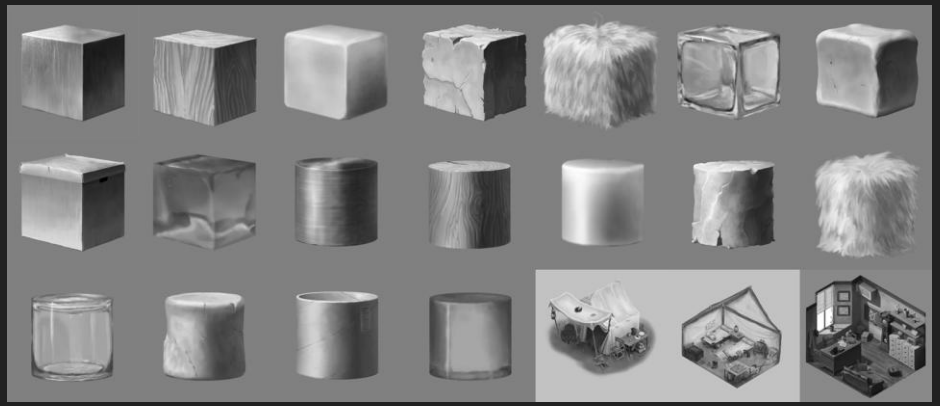
Two complementary views:

- Spatial sensitivity (**occlusion** analysis)
- Frequency structure (**FFT** and **radial profiles**)

Probes how the signal behaves, not how the model is built.

Black-Box Dataset & Occlusion Sensitivity Method

- 42 total images for black-box analysis
- 21 stylized 3D models with stylized shading
- 21 digital illustrations

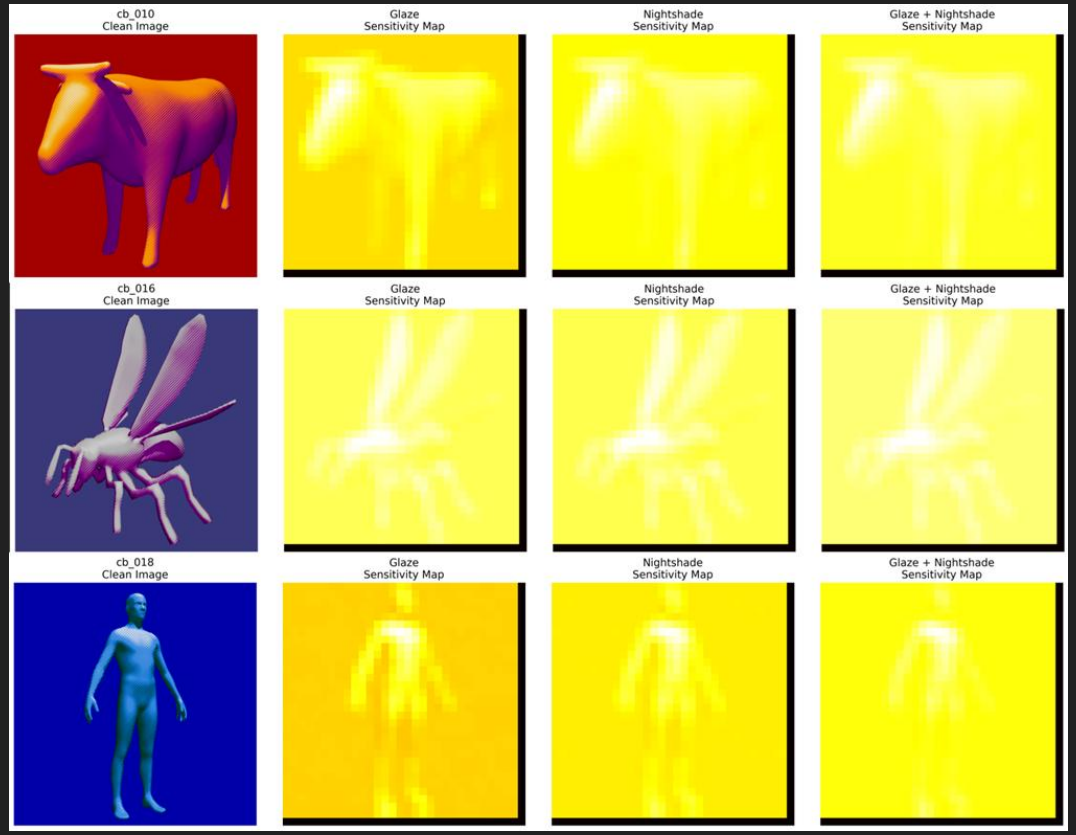


Using this Dataset:

- Spatial sensitivity probed via sliding-window occlusion
- Measures output change under localized perturbation

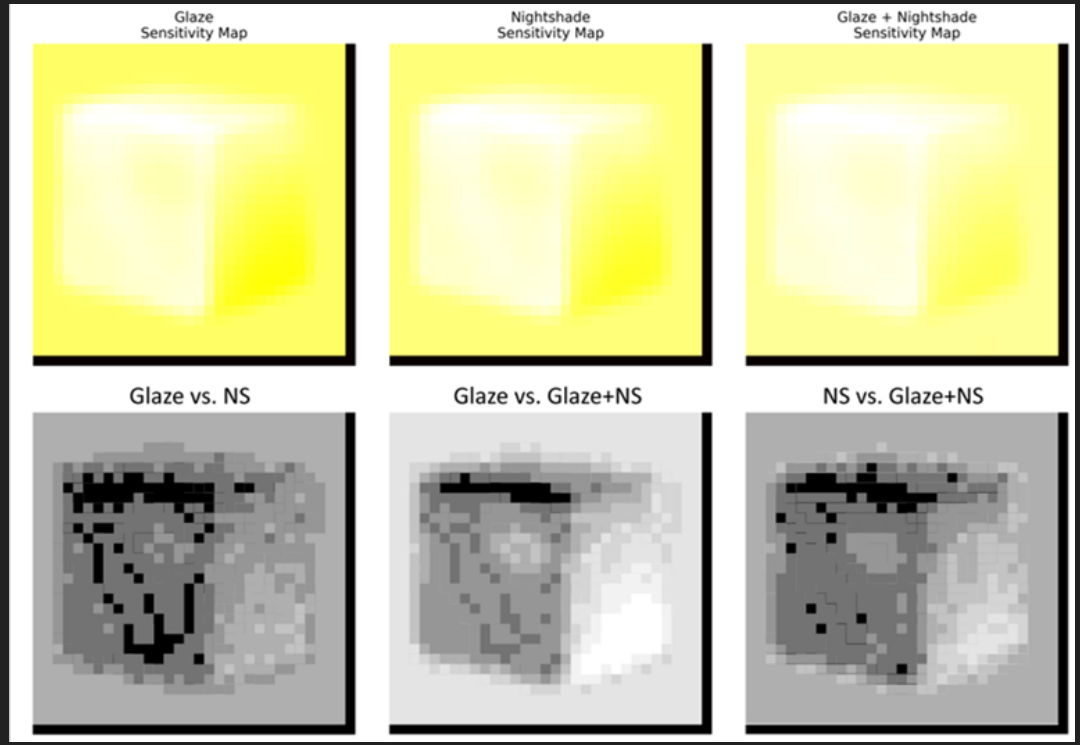
Black-Box Spatial Analysis: Raw Occlusion Limits

- Sensitivity patterns are structurally similar across all perturbations
- Differences appear primarily as tonal, not geometric
- Glaze: more object-localized contrast
- Nightshade: smoother, more globally distributed response



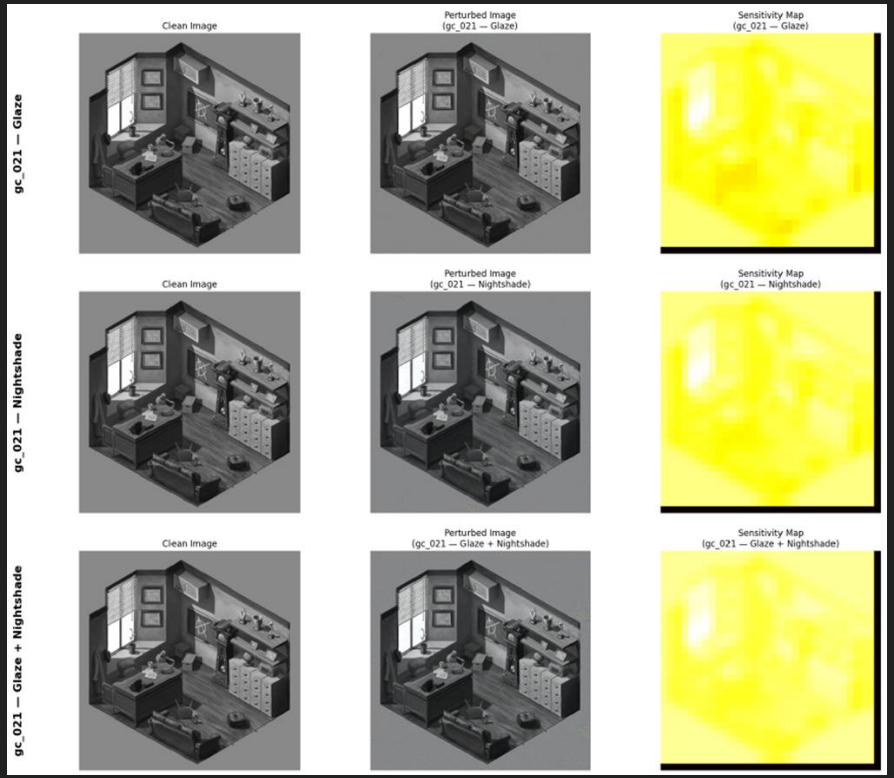
Black-Box Results: Difference Maps Reveal Hidden Structure

- Difference overlays reveal hidden divergences
- Glaze vs. Nightshade diverge on top and right cube faces
- NS→Glaze remains largely Glaze-dominated
- Nightshade vs. NS→Glaze shows broad similarity with localized Glaze structure

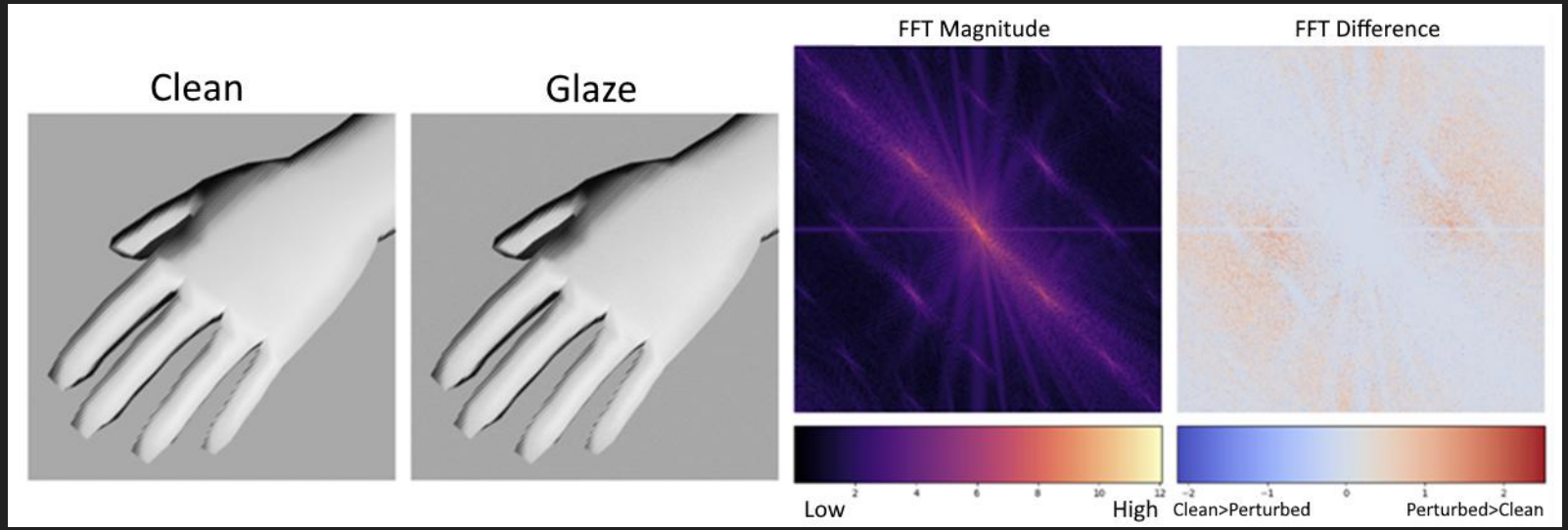


Black-Box Results: Occlusion Sensitivity

- Perturbation influence is **spatially localized, not uniform**
- Nightshade shows **diffuse global sensitivity**
- Glaze shows more object-aligned sensitivity
- Nightshade → Glaze follows Nightshade's spatial pattern
- **Finding:** All perturbations/poison signals remain spatially anchored



Black-Box Analysis #2: Frequency-Domain Analysis



- Analyze perturbations in the **spatial frequency domain**
- Apply **2D Fast Fourier Transform (FFT)** to perturbed images
- Study energy distribution across frequency bands
- Compare Glaze, Nightshade, and NS→Glaze spectra
- Probes structure beyond spatial visibility

Black-Box FFT Processing & Normalization Pipeline

- Compute **2D FFT magnitude** of perturbed images
- Apply **log scaling** for dynamic range compression
- **Center spectra** via FFT-shift
- Radially average to obtain **1D frequency profiles**
- Enables **direct spectral comparison across methods**

Algorithm 3A Log-Magnitude FFT Computation

Given image $g \in [0,1]$, size $H \times W$

$F \leftarrow \text{FFT2}(g)$ ▷ 2D Fast Fourier Transform

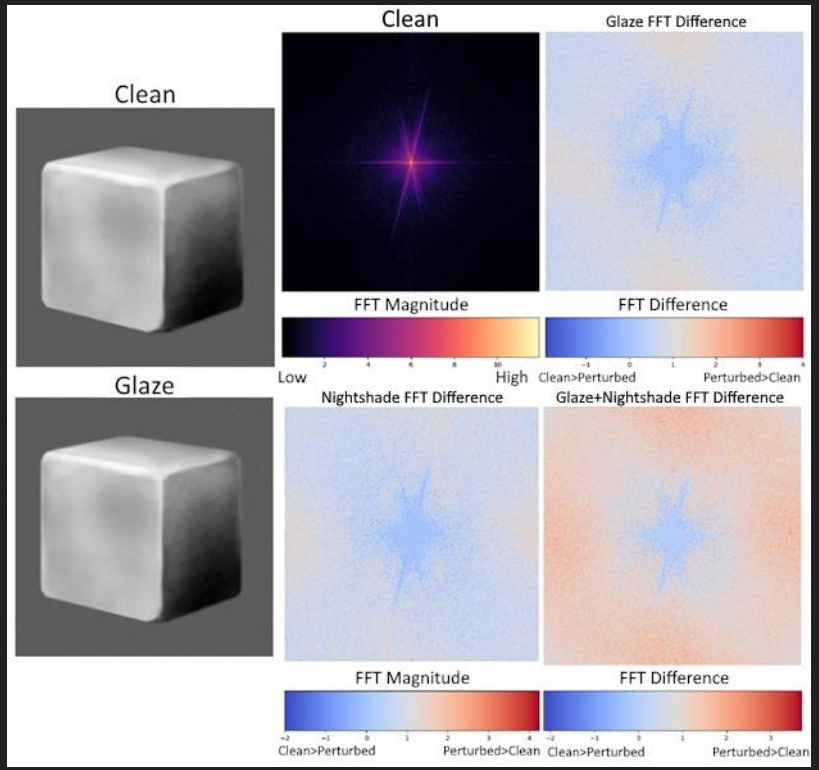
$F \leftarrow \text{FFTShift}(F)$ ▷ Shift DC component to the center

$M \leftarrow \log(1 + |F|)$ ▷ Log-magnitude for dynamic range compression

Return M ▷ Spectral energy representation

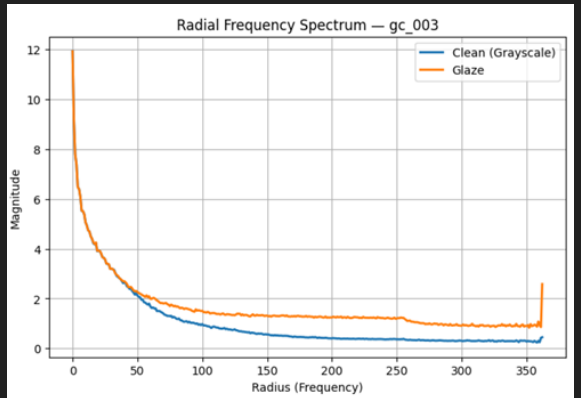
Black-Box Results: FFT – Distinct Spectral Poison Signatures

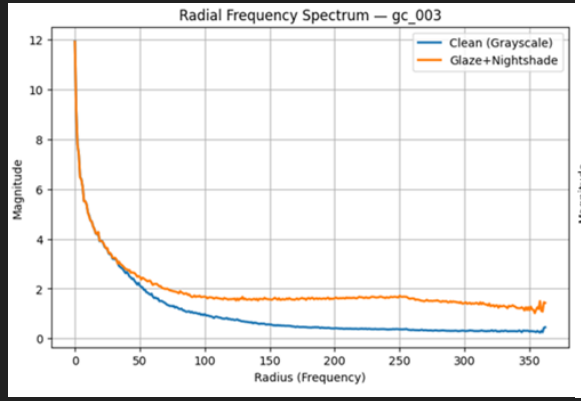
- Glaze concentrates energy in higher spatial frequencies
- Nightshade shows broader low-to-mid frequency distribution
- NS→Glaze exhibits a hybrid but Nightshade-dominant spectrum
- Spectral energy profiles remain stable across image styles
- Confirms poisons differ spectrally, not just spatially

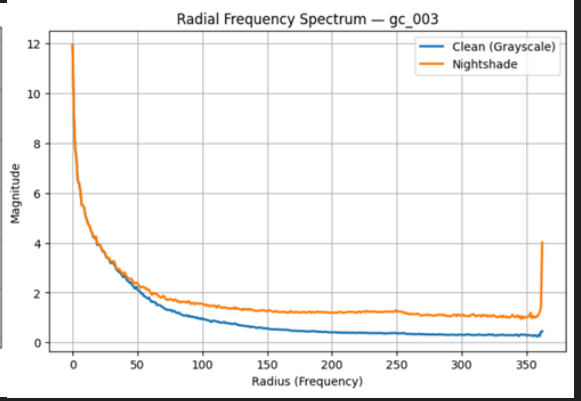


Black-Box Results: Radial Frequency Results (Quantitative Spectral Redistribution)

- All perturbations elevate spectral energy above the clean baseline
- Glaze: smooth, globally coherent upward spectral shift
- Nightshade: stronger low-frequency boost with a sharper high-frequency tail
- NS→Glaze: amplified low–mid frequencies with a regularized high-frequency spike
- **Confirms:** Structured frequency redistribution, not random noise injection







Synthesis: White-Box + Black-Box Picture

VISUALIZATION & INTERFACE DESIGN INNOVATION

WHITE-BOX (Model-Internal Interpretation)

- **Latent Behavior (RQ1, RQ3):**
  Clean and poisoned images form style-driven clusters with poison-specific substructure and clear ordering effects, revealing how perturbations reshape internal representations and influence detectability.
- **LightShed Feature Specialization (RQ2, RQ3):**
  Mid-level encoder layers and sparse channels specialize in reconstructing structured perturbation features, explaining when and why purification succeeds or fails.

BLACK-BOX (Model-Agnostic Signal Analysis)

- **Spatial Structure (RQ1, RQ3):**
  Perturbations are spatially anchored, globally coherent, and aligned with object geometry and background context, directly characterizing signal-level structure and evasion behavior.
- **Frequency Structure (RQ1, RQ2, RQ3):**
  Glaze, Nightshade, and NS→Glaze systematically redistribute existing spectral energy rather than injecting random noise, revealing both detectable band structure and low-entropy signatures exploited by purification.

Unified View: Across white-box and black-box analyses, protection methods act as **structured, low-entropy signal reshaping**, directly explaining both purification effectiveness and adversarial bypass conditions.

Conclusion

- Protection methods inject **structured, low-entropy perturbations**
- These structures are:
  - **Learned and localized internally** (white-box)
  - **Spatially anchored and spectrally organized** (black-box)
- Explains why purification succeeds—and when it fails
- Sequential poisons form hybrid but predictable signal signatures
- Points toward interpretability-driven design of future defenses

Future Work

- Expand to larger datasets and broader artistic style coverage
- Perform cross-model comparisons
  (CLIP-L, OpenCLIP, SigLIP)
- Extend frequency-domain analysis to wavelet-based representations
- Study temporal poisoning dynamics during fine-tuning
- Explore gradient-free surrogate detection models

Authorship and Contributions

Michael Martin

- Conceptualized the research idea and designed detailed methodology
- Developed the full experimental and analytical framework
- Developed and Imeplemented black-box spatial & frequency-domain analysis
- Led methodology development and study design
- Wrote the full research proposal and first-author manuscript

Garrick Chan

- Conducted white-box LightShed neuron and latent analysis
- Conducted t-SNE visualizations based on the designed methodology
- Produced original digital illustrations used in this dataset
- Assisted as co-author on the manuscript

Acknowledgements

We would like to sincerely thank Dr. Kwan-Liu Ma for his guidance, insight, and support throughout the development of this research. We are also grateful to the Visualization & Interface Design Innovation (VIDI) Laboratory at UC Davis for providing an outstanding research environment that made this work possible. Finally, we would like to thank Hanna Foerster at the University of Cambridge for providing access to the LightShed perturbation purification framework used in our experiments.

Thank you!

UC DAVIS
COMPUTER SCIENCE

VIDI
VISUALIZATION & INTERFACE
DESIGN INNOVATION

References

1. Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. 2023. Glaze: protecting artists from style mimicry by text-to-image models. In Proceedings of the 32nd USENIX Conference on Security Symposium (SEC '23). USENIX Association, USA, Article 123, 2187–2204.
2. S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng and B. Y. Zhao, "Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models," 2024 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2024, pp. 807-825, doi: 10.1109/SP54263.2024.00207.
3. Cao, Bochuan, 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." Advances in Neural Information Processing Systems 36 (2023): 10657-10677.
4. Hönig, Robert, Javier Rando, Nicholas Carlini, and Florian Tramèr. "Adversarial perturbations cannot reliably protect artists from generative ai." arXiv preprint arXiv:2406.12027 (2024).
5. Hanna Foerster, Sasha Behrouzi, Phillip Rieger, Murtuza Jadliwala, and Ahmad-Reza Sadeghi. 2025. LightShed: defeating perturbation-based image copyright protections. In Proceedings of the 34th USENIX Conference on Security Symposium (SEC '25). USENIX Association, USA, Article 373, 7271–7290.
6. Xiang, C. Artists are revolting against ai art on artstation. VICE (Dec 2022).