Research Direction: Mechanistic Interpretability of Structured Signals in Generative Vision Systems

Project Title: Interpreting Structured Perturbations in Image Protection Methods for Diffusion Models

Doctoral Researcher, Computer Science: **Michael R. Martin**

Advisor: Dr. Kwan Liu Ma

VIDi Lab (Visualization & Interface Design Innovation), University of California, Davis

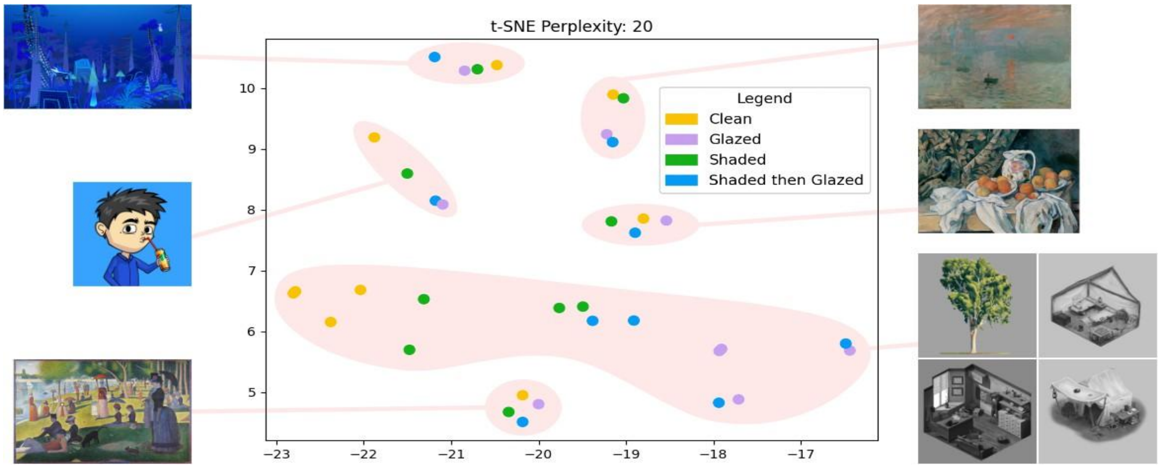

Figure 1: *Latent-space t-SNE visualization of clean, Glaze-protected, Nightshade-protected, and sequentially protected images. Semantic content remains the dominant organizing factor, while protection methods introduce consistent, structured subclusters. This demonstrates that modern protection mechanisms deform localized representation structure without causing global drift.*

1. Research Summary

Recent image-protection mechanisms, including Glaze and Nightshade, introduce visually imperceptible perturbations intended to inhibit the misuse of images by diffusion-based generative models. Although these mechanisms demonstrate practical success, their internal behavior has not been systematically characterized. This work provides a mechanistic interpretability analysis of structured image-protection perturbations in diffusion models. Through latent-space clustering, feature-channel activation statistics, occlusion-based spatial sensitivity mapping, and frequency-domain analysis, the study demonstrates that protection perturbations behave as structured, low-entropy, content-aligned deformations rather than random noise. These deformations preserve semantic organization while introducing predictable representational substructure, offering a principled foundation for understanding and evaluating protection strategies in generative vision systems.

2. Motivation

Modern diffusion-based generative models rely on multi-scale feature hierarchies and iterative denoising, making them highly sensitive to localized, low-entropy perturbations. Protection methods exploit this sensitivity but lack a mechanistic account of:

- How perturbations alter latent and intermediate feature representations
- Which diffusion stages and architectural components encode perturbation structure
- How spatial, spectral, and representational domains interact to govern detectability
- Why visually minimal perturbations persist through denoising
- How protection modifies generative behavior without disrupting semantics

3. Key Findings

- **Structured perturbations remain tightly coupled to image content:** Frequency- and spatial-domain analyses show that protection perturbations redistribute magnitude along image-aligned frequency axes rather than producing diffuse noise. These perturbations maintain coherent structure aligned with underlying content.
- **Semantic organization is preserved in latent space:** t-SNE visualizations (Figure 1) show that clean and protected versions of each image cluster within the same global semantic manifold, demonstrating that protection preserves high-level content organization. Protection introduces consistent, method-specific subclusters without inducing global drift.
- **Feature-channel activations encode perturbation pathways:** Analysis of channel activations identifies localized pathways through which protection perturbations are represented. Perturbation-activated channels correlate with structured distortions in reconstruction, revealing how protection signals propagate internally.
- **Detectability is shaped by entropy, spatial deployment, and frequency alignment:** Spatial sensitivity mapping and spectral decomposition indicate that detectability depends on interactions between perturbation entropy, spatial patterning, and frequency-axis alignment. Sequentially applied protections amplify structured energy, increasing detectability even when visual differences remain minimal.
- **Protection acts through structured deformation rather than semantic disruption:** Across all analyses, protection perturbs feature-level organization without altering semantic layout. The perturbations are coherent, low-entropy signals that modify internal pathways while preserving content.

4. Methodological Approach

This study introduces a unified interpretability framework composed of complementary analytical tools:

- **Comparative evaluation:** consistent analysis across Glaze, Nightshade, and sequential protection variants to identify shared and distinct mechanistic signatures.
- **Latent-space analysis:** T-SNE clustering to examine manifold structure and method-specific subclusters.
- **Feature-channel activation statistics:** identification of channel activation components responsible for encoding perturbations.
- **Spatial sensitivity mapping:** occlusion-based analysis to determine where perturbations influence generative behavior.
- **Frequency-domain characterization:** Fourier-transform magnitude and directionality analysis to quantify spectral alignment and energy redistribution.

5. Significance

This work advances understanding of generative model robustness and interpretability in several ways:

- Establishes a mechanistic explanation of how structured protection perturbations interact with diffusion architectures.
- Explains why visually imperceptible perturbations remain detectable at the representational, spatial, and spectral levels.
- Provides analytical tools to evaluate and compare protection mechanisms based on latent geometry, activation pathways, and frequency behavior.
- Demonstrates that protection mechanisms function through structured deformation rather than semantic alteration.
- Creates a framework for developing protection and detection methods informed by model behavior.
- Positions structured perturbation analysis as a general tool for probing robustness, controllability, and failure modes in generative vision systems beyond image protection.

6. Future Extensions

- Building on these findings, several immediate extensions include.
- Extending structured perturbation analysis to video and multimodal diffusion architectures.
- Leveraging mechanistic signatures to design controllable or steerable generative behaviors under adversarial or constrained settings.
- Developing quantitative measures of representational stability under structured perturbations.
- Designing detection mechanisms informed by spectral and activation-level signatures.
- Formalizing perturbation manifold geometry to characterize protection-induced subspace deformation.

7. Conclusion

This work delivers a comprehensive mechanistic interpretation of structured image-protection perturbations in diffusion models. By integrating latent, spatial, spectral, and activation-based analyses, it reveals how protection signals propagate through generative architecture and their detectability despite their minimal visual footprint. The results establish a rigorous scientific foundation for evaluating current protection techniques and guiding the development of interpretable and robust generative vision systems.

8. Role and Leadership

Michael Martin originated and led the project, formulating research questions, designing the methodological framework, and structuring the overall investigation. He developed and implemented the complete perturbation-interpretability pipeline, including model instrumentation, procedural perturbation synthesis, black-box spectral analysis, and visualization systems for neuron activations, latent-space organization, and attribution behavior. He conducted all quantitative evaluations, performed result interpretation, and produced reproducible visual analyses. Martin served as the lead author of the resulting preprint. Dr. Kwan-Liu Ma provided research guidance.

Reference

Martin, M. R., Chan, G., Ma, K.-L. (2025). *Interpreting Structured Perturbations in Image Protection Methods for Diffusion Models,* http://arxiv.org/abs/2512.08329.