University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Bartłomiej Sadlej
Student no. 429589

EquiDiff: Generative Exploration of Neural Network Invariant Sets through Diffusion-Based Sampling

Bachelor's thesis
in MACHINE LEARNING

Supervisor:
prof. dr hab. inż. Przemysław Biecek
Wydział Matematyki Informatyki i Mechaniki

Warsaw, September 2025

Abstract

Understanding the decision-making process of deep neural networks is an active area of research in machine learning. Current state of the art methods focus on finding human-interpretable concepts or features that influence predictions in known data samples. However, we argue that this approach is limited in its ability to provide a comprehensive understanding of model behavior due to vast unexplored regions of the data manifold, not present in investigated datasets, which can potentially lead to the same predictions. The main contribution of this work is paradigm shift from traditional explainable AI (XAI) methods where we find human-interpretable features in known data to generative XAI methods that synthesize new samples. This work first introduces new framework *Invariants* - a theoretical backbone for generative XAI methods and then proposes new method EquiDiffhich an efficient implementation of this framework. Evaluation of this methon on popular models such as ResNet-50 or Sprase AutoEncoders (SAE) points out significant limitations of current XAI methods.

This work is a continuation of pioneering work on Generative Explainable AI (XAI) [Sobieski et al., 2024] published at International Conference on Learning Representations 2025 and introduces a novel framework of Invariants - sets of data points that yield identical predictions under a given model, thereby establishing an equivalence relation. Our method EquiDiff allows for sampling meaningful and diverse high quality realistic examples from these invariant sets and as a result opens new possibilities for evaluating existing explainability methods on unknown data - simulating the real world, and fully preventing training data leaks by generating synthetic samples.

Keywords

explainable AI, generative models, diffusion models, invariant sets, level sets, neural network interpretability, mechanistic interpretability, sparse autoencoders, visual explanations, counterfactual generation, score-based generative models

Thesis domain (Socrates-Erasmus subject area codes)

11.4 Sztuczna inteligencja
11.3 Informatyka
11.1 Matematyka

Subject classification

I. Computing Methodologies
I.2 Artificial Intelligence
I.2.6 Learning
I.2.10 Vision and Scene Understanding

I.4 Image Processing and Computer Vision
I.4.8 Scene Analysis

Tytuł pracy w języku polskim

EquiDiff: Generatywna Eksploracja Zbiorów Niezmienniczych Sieci Neuronowych przez Próbkowanie Oparte na Dyfuzji

Contents

Chapter 1

Introduction

The remarkable success of deep neural networks in computer vision has been accompanied by an equally pressing need to understand their decision-making processes. As these models are deployed in critical applications ranging from medical diagnosis to autonomous driving, the ability to explain and interpret their behavior becomes paramount for building trust, ensuring fairness, and identifying potential failure modes.

Current explainable AI (XAI) methods have made significant strides in providing insights into model behavior through various approaches including saliency maps [Simonyan et al., 2014], concept activation vectors [Kim et al., 2018], and gradient-based attribution methods [Sundararajan et al., 2017]. However, these approaches share a fundamental limitation: they primarily operate within the confines of known training data or slight perturbations thereof, leaving vast regions of the input manifold unexplored.

More recent advancements that expand the scope of interpretability try to address those limitations. Approaches such as Rate-Distortion Explanation (RDE) frameworks systematically perturb input signals across diverse data modalities to identify truly relevant features, thereby moving beyond local sensitivity [Kolek et al., 2022]. These frameworks also explicitly aim for in-distribution interpretability by leveraging generative models like in-painting GANs, thereby guarding against explanations corrupted by evaluations in undeveloped or unrealistic regions of the model's function. Similarly, new techniques for interpreting deep generative models (GANs) enable the identification of human-understandable concepts within latent spaces, allowing for interactive image generation and editing. This actively explores the input manifold by creating new data, offering insights into how realistic images are composed from deep representations. [Zhou, 2022, Karimi et al., 2022]

Furthermore, XAI is seeing a shift towards building transparency into models from the outset, often referred to as "interpretable-by-design" methods [Karimi et al., 2022, Holzinger et al., 2022]. Research into interpretable reinforcement learning via programmatic policies aims to train policies in the form of human-readable programs (e.g., decision trees, state machines), which are inherently more interpretable, verifiable, and robust than traditional deep neural network policies [Marcos et al., 2022, Inala et al., 2020, Verma et al., 2019]. Likewise, Explainable Neural-Symbolic Learning (X-NeSyL) represents another design-based approach, fusing deep learning representations with expert knowledge graphs to encourage neural networks to learn structures akin to human expert reasoning, ensuring interpretability is embedded throughout the training process [Díaz-Rodríguez et al., 2022, Karimi et al., 2020].

These advancements align with what is sometimes referred to as "RED XAI" – a model-centric culture focused on questioning models, extracting knowledge, spotting, and fixing bugs, and ultimately improving the reliability and safety of AI systems [Biecek and Samek, 2024]. This perspective is critical for using explanations not just to justify decisions, but to drive model development and

verification [Tsai and Carroll, 2022]. This includes attributing importance to feature interactions and groups, which can then be used to directly improve model generalization or to distill complex models into simpler forms, often validated through "reality checks" [Singh et al., 2022]

1.1. Motivation and Problem Statement

Consider a trained image classifier that correctly identifies both a standard photograph of a dog and a highly stylized artistic rendering of the same animal. Traditional XAI methods would analyze these two specific instances, potentially identifying common features like shape or texture patterns. However, they would miss the broader question: what other visual representations would this model also classify as a dog with the same confidence?

When we say "a dog with the same confidence," we mean something far more profound and potentially disturbing than might initially appear. We are not merely talking about different breeds of dogs, or dogs photographed from different angles, or even dogs rendered in different artistic styles. We are talking about the complete universe of visual patterns—no matter how bizarre, abstract, or seemingly unrelated to dog anatomy—that trigger identical neural responses in the classifier's decision-making apparatus. This could include a Jackson Pollock painting with just the right splatter of paint, a close-up photograph of tree bark with particular texture patterns, a geometric arrangement of colored pixels that bears no resemblance whatsoever to any living creature, or even a photograph of a kitchen appliance that happens to contain the precise combination of edges, curves, and color distributions that the model has learned to associate with "dog-ness." The classifier would assign these wildly disparate inputs exactly the same probability score—perhaps 0.8347 for "dog"—despite their complete lack of semantic relationship to actual dogs.

This question is not merely academic but reveals a fundamental blindness in our understanding of machine learning models. Traditional explainable AI methods focus on the narrow slice of reality represented in training datasets, leaving vast territories of the input manifold completely unexplored and potentially harboring unexpected model behaviors. The robustness implications are staggering: if a model can be fooled into seeing a dog in a random arrangement of geometric shapes with the same confidence as it sees a dog in an actual photograph of a Golden Retriever, what does this say about its reliability in real-world deployment? The bias detection possibilities are equally concerning—systematic patterns within these datasets might reveal that the model has learned to associate certain irrelevant features (perhaps related to image compression artifacts, camera settings, or demographic markers in the background) with specific classes, perpetuating hidden biases that would never be discovered through traditional dataset analysis. For fairness evaluation, understanding these equivalence classes becomes critical: if the model makes identical predictions for inputs that vary along protected attributes while maintaining other spurious correlations, we need to map these relationships to ensure equitable treatment. Finally, the structure of these datasets provides unprecedented insight into how models generalize beyond their training distribution—revealing whether generalization relies on semantically meaningful features or on arbitrary statistical regularities that happen to correlate with class labels in the training data.

1.2. Our Approach: Generative Explainable AI

This thesis introduces a paradigm shift from traditional interpolative XAI methods to a generative approach. Instead of analyzing existing data points, we propose synthesizing new, meaningful examples that preserve model predictions, thereby exploring the *Invariant Set* – the complete collection of inputs that yield identical outputs under a given objective function.

Our method, EquiDiff (Equivariant Diffusion Sampling), combines score-based generative models with classifier guidance to sample high-quality, diverse images from these invariant sets. By leveraging the powerful generative capabilities of diffusion models, we can explore regions of the input space that may never have been encountered during training, providing a more comprehensive understanding of model behavior.

Figure 1.1 illustrates the conceptual distinction between our approach and current XAI methods. While traditional methods focus on explaining decisions within known data boundaries, generative XAI does not have this limitation and can explore the broader space of possible inputs that lead to the same predictions.

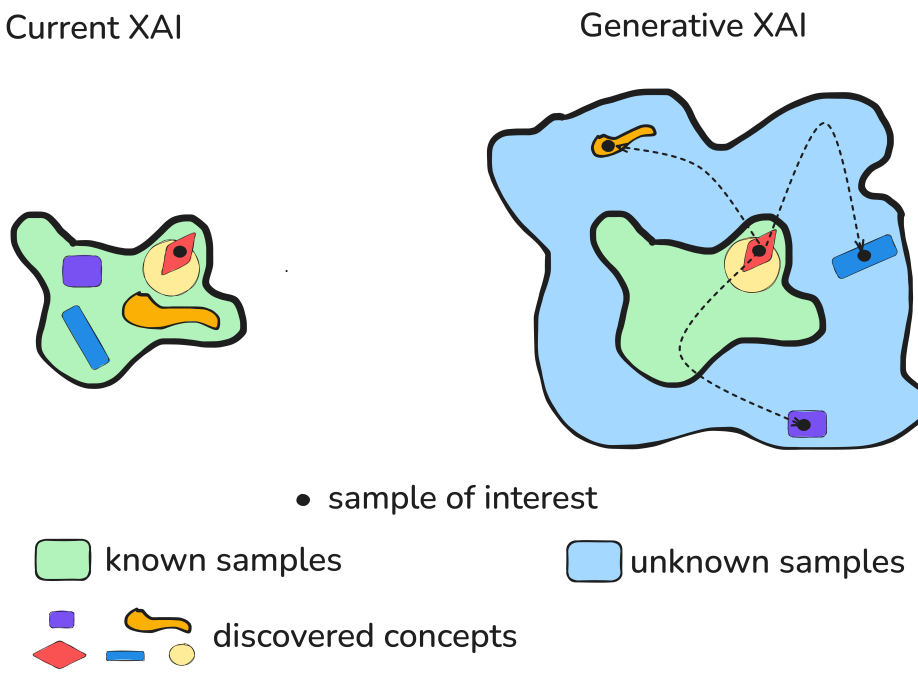


Figure 1.1: Conceptual comparison between traditional XAI methods and Generative XAI. Traditional methods analyze known data samples (left), while our approach synthesizes diverse examples from the invariant set that yield identical predictions (right).

1.3. Contributions

This thesis makes three key contributions to the field of explainable AI, as outlined in the abstract and detailed in Chapter 3:

The first contribution represents a **paradigm shift from traditional explainable AI methods** that analyze human-interpretable features in known data samples to generative XAI methods that synthesize new samples. This fundamental change in perspective allows for exploration of vast regions of the input manifold that remain unexplored by current approaches, providing a more comprehensive understanding of model behavior beyond the confines of training datasets.

The second contribution introduces the **Invariants framework** – a novel theoretical backbone for generative XAI methods. This framework provides formal mathematical definitions of invariant sets and establishes their properties as equivalence relations, offering a rigorous foundation for understanding and generating diverse examples that yield identical model predictions.

The third contribution presents **EquiDiff** – an efficient algorithmic implementation of the Invariants framework. This method combines score-based diffusion models with guided sampling to generate high-quality, diverse examples from invariant sets, enabling practical application of the theoretical framework to real-world neural network analysis and interpretation.

These contributions are comprehensively detailed and evaluated in Chapter 3, where we present both the theoretical foundations and empirical validation of our approach.

1.4. Thesis Organization

The remainder of this thesis is structured to provide a comprehensive exploration of our generative explainable AI approach. Chapter 2 reviews the relevant literature across explainable AI methods, generative modeling, and diffusion models, establishing the theoretical foundation for our work. Chapter 3 presents our core theoretical framework and details the EquiDiff algorithm, providing the mathematical foundation for invariant set generation and the practical implementation of our approach.

Chapter 4 presents a comprehensive experimental evaluation demonstrating the effectiveness of our method across multiple neural network analysis paradigms, from individual neuron activation to complete classifier output preservation. Chapter 5 explores practical applications of our framework in real-world scenarios, while Chapter 6 discusses the broader implications of our work, current limitations, and directions for future research. Chapter 7 synthesizes our contributions and their significance for the field of explainable AI.

Chapter 2

Related Work

This chapter reviews the relevant literature across several interconnected areas that form the foundation of our work. We begin with an overview of explainable AI methods, followed by background on score-based generative models, conditional generation techniques, and related work on activation maximization and concept discovery.

2.1. Explainable Artificial Intelligence

The field of explainable AI has evolved rapidly in response to the growing complexity and opacity of modern deep learning models. Current approaches can be broadly categorized into several paradigms:

2.1.1. Attribution Methods

Attribution methods aim to identify which input features are most important for a model's prediction in a form of a heatmap. Gradient-based methods like Integrated Gradients [Sundararajan et al., 2017] and GradCAM [Selvaraju et al., 2017] compute the gradient of the output with respect to input features to determine importance scores. While computationally efficient, these methods are limited to local explanations around specific data points and can be sensitive to model architecture and input preprocessing.

Perturbation-based methods such as LIME [Ribeiro et al., 2016] and SHAP [Lundberg and Lee, 2017] evaluate feature importance by measuring how predictions change when features are masked or altered. These methods provide more model-agnostic explanations but are computationally expensive and may not capture complex feature interactions.

2.1.2. Concept-Based Methods

Concept-based explainability methods attempt to understand models in terms of human-interpretable concepts. Concept Activation Vectors (CAVs) [Kim et al., 2018] learn linear directions in activation space that correspond to human-defined concepts. Network Dissection [Bau et al., 2017] automatically discovers concepts by correlating individual neurons with semantic segmentation labels.

More recent work has focused on discovering concepts automatically without human supervision. ACE (Automatic Concept Extraction) [Ghorbani et al., 2019] uses unsupervised segmentation to identify important concepts, while TCAV (Testing with CAVs) [Kim et al., 2018] provides statistical significance testing for concept importance.

2.1.3. Counterfactual Explanations

Counterfactual explanations answer the question "What would need to change for the model to make a different prediction?" This paradigm has gained popularity due to its intuitive nature and practical utility. Although earlier work has explored generative models for visual counterfactual explanations, [Sobieski et al., 2024] advanced this direction in a way that directly inspired our approach. Our method is, to our knowledge, the first to combine high-quality results with almost real-time performance.

In counterfactual methods, the objective is generally defined as finding the smallest possible changes to an input that alter the model's decision—for example, modifying a few pixels in an image so that the predicted class changes. In contrast, our goal is to generate diverse examples that preserve the original prediction.

2.2. Score-Based Generative Models

Score-based generative models (SGMs) have emerged as a powerful framework for high-quality image generation. Following the seminal work of [Song et al., 2021], these models can be understood through the lens of stochastic differential equations (SDEs).

2.2.1. Mathematical Foundation

The core idea behind SGMs is to transform samples from a complex data distribution p_0 (e.g. natural images) to a simple noise distribution p_1 (typically Gaussian) through a forward diffusion process and then learn to reverse this transformation. The forward SDE is given by:

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t \tag{2.1}$$

where $\mathbf{x}_t$ represents the noisy version of a clean image at time $t \in [0, 1]$, $\mathbf{f}(\mathbf{x}_t, t)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and $\mathbf{w}_t$ is a Wiener process.

The corresponding reverse SDE, which enables generation, is:

$$\mathrm{d}\mathbf{x}_t = [\mathbf{f}(\mathbf{x}_t, t) - g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}_t \tag{2.2}$$

The key term $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}t)$ is the score function, which must be learned by a neural network $\mathbf{s}\boldsymbol{\theta}(\mathbf{x}_t, t)$, since it cannot be computed analytically without access to the final, fully denoised image.

2.2.2. Training and Sampling

Score networks are typically trained using denoising score matching [Vincent, 2011, Song and Ermon, 2020]:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}} \left[\lambda(t)\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \boldsymbol{\epsilon}\|_2^2\right] \tag{2.3}$$

where $\mathbf{x}_t = \alpha(t)\mathbf{x}_0 + \sigma(t)\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, and $\lambda(t)$ is a weighting function.

During sampling, we start from pure noise $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I})$ and integrate the reverse SDE using numerical solvers, with the learned score function $\mathbf{s}_{\boldsymbol{\theta}}$ approximating the true score.

2.3. Conditional Generation and Classifier Guidance

Conditional generation extends SGMs to produce samples conditioned on additional information $\mathbf{y}$, such as class labels or other attributes. The conditional score function can be decomposed using Bayes' theorem:

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t \mid \mathbf{y}, t) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{y} \mid \mathbf{x}_t, t) \tag{2.4}$$

2.3.1. Classifier Guidance

Classifier guidance [Dhariwal and Nichol, 2021] implements conditional generation by training an auxiliary time-dependent classifier $p_{\boldsymbol{\phi}}(\mathbf{y} \mid \mathbf{x}_t, t)$ on noisy images and incorporating its gradients into the sampling process:

$$\tilde{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{y}) = \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) + s \cdot \nabla_{\mathbf{x}_t} \log p_{\boldsymbol{\phi}}(\mathbf{y} \mid \mathbf{x}_t, t) \tag{2.5}$$

where s is the guidance scale that controls the trade-off between sample quality and diversity.

2.3.2. Limitations of Standard Classifier Guidance

While effective for class-conditional generation, standard classifier guidance has several limitations for our application:

1. **Limited Optimization Steps**: Guidance is applied only at discrete points along the denoising trajectory, due to the fixed number of diffusion steps, which can constrain the precision of conditioning.
2. **Latent Space Considerations**: Many state-of-the-art diffusion models operate in latent space, requiring careful alignment of the conditioning signal with the model's latent representation.
3. **Objective Function Alignment**: In its standard form, classifier guidance is tailored for classification objectives (predicting $p(y \mid x)$). While the conditioning variable y can, in principle, represent a wide range of targets beyond class labels, adapting it to arbitrary objective functions may require additional formulation effort.

These limitations motivate our approach, which we detail in Chapter 3.

2.4. Inverse Problems and Posterior Sampling

Recent work has explored the use of diffusion models as priors for solving inverse problems in image restoration [Song et al., 2023, Chung et al., 2024]. The general inverse problem can be formulated as:

$$\mathbf{y} = \mathcal{A}(\mathbf{x}) + \boldsymbol{\epsilon} \tag{2.6}$$

where $\mathcal{A}$ is a (possibly nonlinear) forward operator, $\mathbf{x}$ is the unknown signal, $\mathbf{y}$ is the observed measurement, and $\boldsymbol{\epsilon}$ is an additive noise term, which may follow different distributions (e.g., Gaussian, Poisson) and can exhibit nontrivial covariance structures.

[Chung et al., 2024] showed that diffusion models can address nonlinear inverse problems for arbitrary differentiable forward systems by incorporating the measurement likelihood into the reverse SDE. Their framework accommodates various noise models, including Gaussian and Poisson. This is particularly relevant to our approach, as neural network predictions can be interpreted as nonlinear measurements of the input image.

2.4.1. Diverse Posterior Sampling

More recently, [Cohen et al., 2024] extended inverse problem solvers to generate diverse solutions rather than a single best estimate. This paradigm shift from point estimation to posterior sampling aligns closely with our goal of generating new data samples.

2.5. Activation Maximization and Feature Visualization

Activation maximization techniques attempt to synthesize inputs that maximally activate specific neurons or model outputs [Erhan et al., 2009, Mordvintsev et al., 2015]. The basic approach optimizes an input image $\mathbf{x}$ to maximize an objective function $\mathcal{L}(\mathbf{x})$:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \mathcal{L}(\mathbf{x}) - \lambda \mathcal{R}(\mathbf{x}) \tag{2.7}$$

where $\mathcal{R}(\mathbf{x})$ is a regularization term to encourage natural-looking images.

However, activation maximization methods often produce unrealistic images accompanied by high-frequency artifacts that are imperceptible to humans yet strongly activate neurons [Shinkle and Lescroart, 2025, Nanfack et al., 2023, Zhu and Cangelosi, 2025]. To address this, various regularization techniques have been proposed, including total variation penalties, which encourage smoother and more coherent outputs [Mahendran and Vedaldi, 2014], and frequency domain constraints that reduce high-frequency noise and improve interpretability [Olah et al., 2017].

2.5.1. Limitations and Relationship to Our Work

Although activation maximization shares the goal of understanding model behavior through synthetic inputs, it differs fundamentally from our approach.

1. **Single Solution vs. Diverse Sets**: Activation maximization typically finds one optimal input, while we aim to generate a set of new data points.

2. **Maximum Activation vs. Preserved Predictions**: Activation maximization seeks to maximize responses while we preserve specific prediction values.

3. **Quality Issues**: Traditional activation maximization often produces unrealistic images, while our diffusion-based approach leverages strong visual priors.

2.6. Concept Discovery and Spurious Feature Detection

Understanding what concepts neural networks learn has been an active area of research. [Lapuschkin et al., 2019] developed SpRAy, an automatic pipeline for exploring shortcuts and biases learned by models, often referred to as "Clever Hans" effects [Pfungst, 1911]. [Neuhaus et al., 2023] investigates methods for automatically finding spurious features in training data.

Recent work by [Dreyer et al., 2025] addresses the question of what concepts were learned by models and where in the training data they were present. However, [Leask et al., 2025] argues that automatically discovered concepts may lack atomicity and completeness.

Our work complements this line of research by exploring the space of inputs that preserve predictions, potentially revealing spurious correlations and biases that may not be apparent from training data analysis alone.

2.7. Detection of Synthetic Images

As generative models become increasingly sophisticated, detecting synthetic images has become an important research area. Modern architectures using resampling operations (upsampling, downsampling, interpolation) introduce specific periodic correlations between pixels that are rarely present in natural images [Popescu and Farid, 2005].

Recent advances in synthetic image detection [Zhang et al., 2019, Wang et al., 2023, Zhang and Xu, 2023] have achieved near-perfect accuracy on images generated by GANs and diffusion models by analyzing frequency domain artifacts.

While our goal is not to evade detection methods, we acknowledge that our images are synthetic. Drawing on insights from this literature, we aim for the Fourier spectrum of our generated images to match that of real images, ensuring that model predictions are not driven by imperceptible frequency artifacts but by signal patterns consistent with natural image statistics.

2.8. Gap in Current Literature

Despite significant advances in XAI, a fundamental gap remains: current methods primarily analyze known training data and model behavior on observed inputs. This leaves vast regions of the input manifold unexplored, potentially missing important insights about model behavior.

Our work addresses this gap by introducing a principled framework for exploring the space of alternative inputs that yield identical predictions. By leveraging powerful generative models, we sample from regions of the input space that were observed during training but represent continuous interpolations and variations of the training data, providing a more comprehensive understanding of model behavior within the learned data manifold.

Chapter 3

Method

This chapter presents our theoretical framework and algorithmic approach for generating Invariants. We begin with formal definitions that relate our concept to classical level sets from differential topology [Lee, 2013, Milnor, 1965, Fort, 2017], establish our theoretical foundation, detail our algorithm, and conclude with implementation specifics and quality assurance measures.

3.1. Problem Formulation

We formulate the problem of finding invariant sets (IS) as discovering members of an equivalence relation. Given a neural network with parameters $\boldsymbol{\theta}$ and objective function $\mathcal{L}_{\boldsymbol{\theta}} : \mathbb{R}^n \to \mathbb{R}^m$, and a query point $\mathbf{x}^*$, we define the invariant set as:

$$\mathbf{IS}(\mathbf{x}^*) = \{\mathbf{x} \in \mathbb{R}^n : \mathcal{L}_{\boldsymbol{\theta}}(\mathbf{x}) = \mathcal{L}_{\boldsymbol{\theta}}(\mathbf{x}^*)\} \tag{3.1}$$

We use the notation $\mathbf{x}^* \sim_{\mathcal{L}_{\boldsymbol{\theta}}} \mathbf{x}$ to denote that two elements $\mathbf{x}^*$ and $\mathbf{x}$ belong to the same invariant set under the equivalence relation defined by $\mathcal{L}_{\boldsymbol{\theta}}$.

The objective function $\mathcal{L}_{\boldsymbol{\theta}}$ can represent various neural network components: a single neuron's activation, class logits for one or multiple classes, or any differentiable function for which gradients can be computed. While adversarial examples can be viewed as specific perturbations that may belong to invariant sets under certain conditions [Szegedy et al., 2014], our goal is fundamentally different: we seek to sample from the intersection of the invariant set with the natural data manifold, ensuring realism by construction.

To achieve this, we utilize a trained diffusion model, specifically LightningDIT [Yao et al., 2025] [Yao et al., 2024], which excels at generating high-quality images while maintaining the mathematical constraints of invariant set membership. The diversity of examples emerges naturally from exploring different regions of this manifold intersection.

3.2. Guided Iterative Optimization with Latent Diffusion Models

Our algorithm integrates signals from the neural network function $f_{\boldsymbol{\theta}} : \mathbb{R}^{W \times H} \to \mathbb{R}^m$ through a scalar loss function $\ell : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ to conditionally synthesize images from invariant sets. Given a target output $\mathbf{y}^* = f_{\boldsymbol{\theta}}(\mathbf{x}^*)$, we define our objective as:

$$\mathcal{L}(\mathbf{x}) = \ell(f_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{y}^*) \tag{3.2}$$

where ℓ is typically the ℓ_2 norm or another appropriate distance metric. This formulation enables gradient computation for optimization while maintaining the invariant set constraint $\mathcal{L}(\mathbf{x}) = 0$. There are two primary approaches for conditioning generation using this objective.

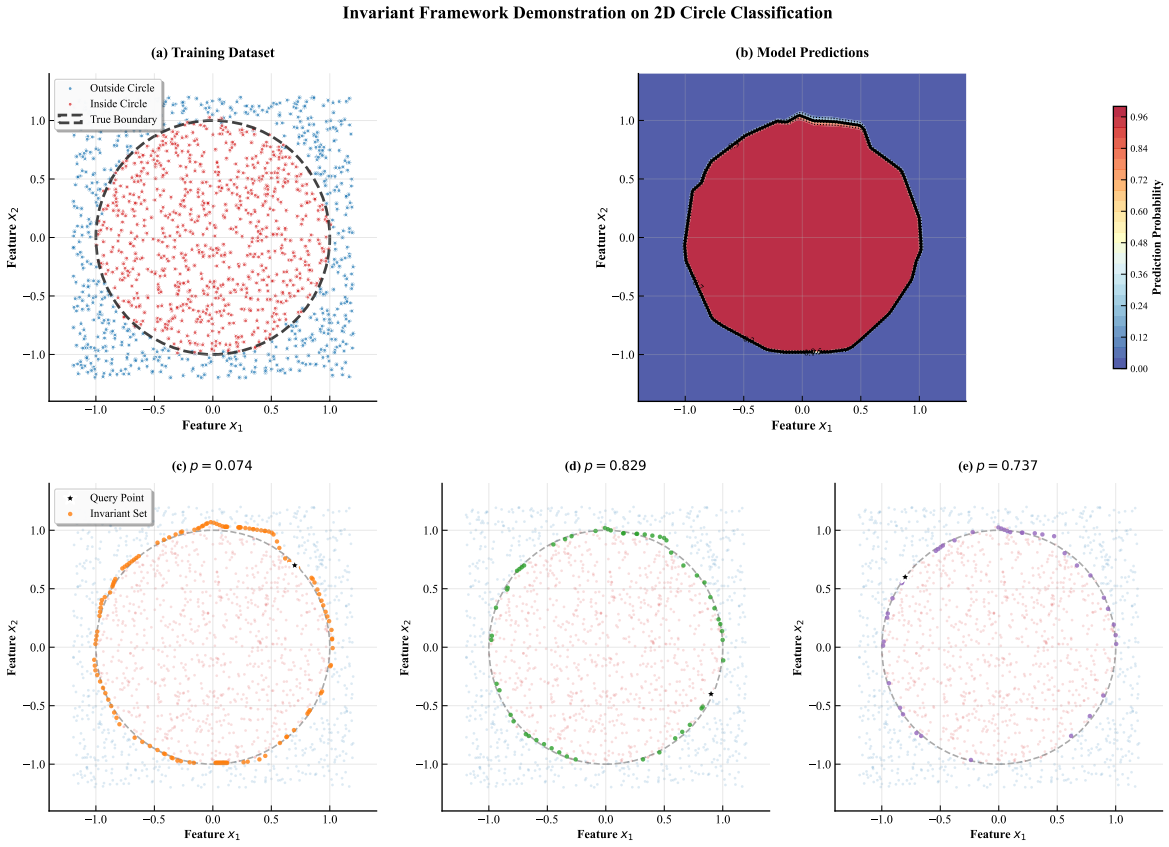


Figure 3.1: Demonstration of the Invariant Framework on a 2D Concentric Circles Dataset. (a) Training dataset with 1,500 samples classified by their position relative to a unit circle (dashed line). Blue points represent the outer class, pink points the inner class. (b) Learned decision boundary and prediction probability heatmap from a 3-layer MLP (test accuracy: 0.983). The black contour shows the 0.5 decision boundary. (c-e) Invariant sets for three query points (black stars) with prediction values p. Orange points represent all input locations that yield identical predictions under the trained model, demonstrating the equivalence relation established by the model's output. The invariant sets approximate level curves of the learned decision function, revealing the geometric structure of the model's decision space.

3.2.1. Classifier Guidance Limitations

Classifier Guidance (CG) [Dhariwal and Nichol, 2021] offers a simple, computationally efficient method for trading diversity for fidelity using gradients from the objective function at each denoising step. However, we identified two significant limitations:

- **Restrictive optimization horizon**: CG typically constrains optimization to a single forward pass through the diffusion steps, which can be too restrictive for achieving optimal results in invariant set generation. While iterative refinement through multiple passes is possible, it significantly increases computational overhead.

- **Latent space complications**: Modern diffusion models often employ the Latent Diffusion Model (LDM) approach [Rombach et al., 2022], which operates in a compressed latent space rather than directly on pixel values. This architectural choice introduces additional complexity when conditioning on neural network outputs: the classifier must evaluate encoded represen-

tations $\mathcal{E}(\mathbf{x}_t)$ at intermediate diffusion timesteps rather than natural images. This mismatch between the diffusion model's latent space and the classifier's expected input domain requires either training timestep-specific classifiers or using approximate reconstructions $\hat{\mathbf{x}}_0(t)$, both of which introduce additional sources of error.

3.2.2. Infinite Optimization Approach

Given these limitations, we adopt an *Infinite Optimization* strategy, specifically adapting Algorithm 1 from [Augustin et al., 2024]. This approach decouples the optimization process from the diffusion sampling steps, allowing for more flexible and thorough exploration of the invariant set while maintaining image quality and realism. The detailed algorithm specification is provided in Appendix .1.

3.3. Quality and Realism Assurance

Our approach ensures that generated images maintain high quality and realism through several mechanisms. We build upon state-of-the-art frameworks for synthetic image detection and leverage the inherent properties of diffusion models, which naturally generate samples from the learned data distribution. Unlike optimization-based adversarial methods that may introduce imperceptible high-frequency artifacts, our diffusion-based approach constrains generation to the natural image manifold, ensuring that invariant set samples remain visually coherent and realistic.

3.3.1. Frequency Domain Optimization

To address potential high-frequency artifacts, we perform frequency domain optimization that guides the generation process to encode meaningful signals in low-frequency bands—those visible to the human eye. Specifically, we introduce a low-pass filter $\mathcal{F}$ before the objective function $\mathcal{L}$ and measure deviation from the original measurement across different cutoff frequencies f_c.

This frequency-aware approach ensures that:

- Generated images appear natural to human observers
- Invariant set membership is achieved through semantically meaningful variations rather than imperceptible noise
- The generated samples maintain the visual characteristics expected from the underlying data distribution

The combination of infinite optimization with frequency domain constraints allows our method to generate diverse, high-quality samples from invariant sets while preserving both mathematical rigor and visual realism.

Chapter 4

Experiments

This chapter presents a comprehensive experimental evaluation of our EquiDiff framework for generating Invariants. We systematically evaluate the method's ability to generate diverse, high-quality samples while maintaining invariant set membership across three complementary experimental paradigms: individual neuron activation analysis, sparse autoencoder (SAE) feature investigation, and classifier output preservation.

4.1. Experimental Design

Our experimental evaluation addresses the following core research questions:

1. Can EquiDiff generate visually diverse samples that maintain identical activation patterns for interpretable neurons?
2. Do generated samples reveal semantic patterns beyond those present in typical training data?
3. How effectively does the method preserve complex feature representations learned by sparse autoencoders?
4. Can the framework maintain classifier predictions while generating semantically meaningful variations?

4.1.1. Infrastructure and Implementation

All experiments were conducted on NVIDIA A100 GPUs (1-4 units) using PyTorch. We employ LightningDiT as our diffusion backbone with SGD optimization at learning rate $\eta = 10$ based on empirical hyperparameter evaluation (see **??**). Each experimental condition generates 32-256 samples due to computational constraints, representing a balance between statistical validity and resource efficiency.

4.1.2. Evaluation Framework

We employ a multi-faceted evaluation approach combining quantitative precision metrics with qualitative semantic analysis:

Quantitative Metrics:

- **Activation Fidelity**: L_1 and L_2 norm deviations from target values
- **Probability Preservation**: KL divergence for probability distributions

- **Spectral Coherence**: Frequency domain analysis using ideal low-pass filters (see **??**)
- **Image Quality**: Fréchet Inception Distance (FID) relative to natural image statistics

Qualitative Assessment:

- Semantic diversity within invariant sets
- Visual coherence and absence of adversarial artifacts
- Alignment between generated patterns and expected neuron/feature selectivity

4.2. Individual Neuron Activation Analysis

Building upon mechanistic interpretability advances, we target neurons with well-characterized semantic properties identified through the Semantic Lens framework [Dreyer et al., 2025]. Our analysis investigates whether EquiDiff can generate diverse visual patterns that consistently activate specific semantic detectors.

4.2.1. Target Neuron Selection

We selected three neurons from ResNet50's final feature layer based on high semantic alignment scores and interpretable activation patterns:

- **Neuron #1656 (Zebra Striping)**: Alignment score $r = 0.945$, responds to black-white striped patterns
- **Neuron #1052 (Honeycomb Structure)**: Alignment score $r = 0.880$, activates on hexagonal cellular structures
- **Neuron #421 (Gyromitra Morphology)**: Alignment score $r = 0.952$, responds to convoluted, brain-like surface textures

4.2.2. Experimental Protocol

For each target neuron n, we:

1. Select a query image $\mathbf{x}^*$ that strongly activates the neuron
2. Define the invariant set constraint: $n(\mathbf{x}) = n(\mathbf{x}^*)$
3. Apply EquiDiff to generate 32 samples maintaining this constraint
4. Evaluate activation fidelity, visual quality, and semantic diversity

4.2.3. Quantitative Results

Table 4.1 presents quantitative evaluation metrics across target neurons. The consistently low L_2 losses (< 1.0 on unbounded logits) demonstrate precise activation preservation, while FID scores indicate maintenance of natural image statistics.

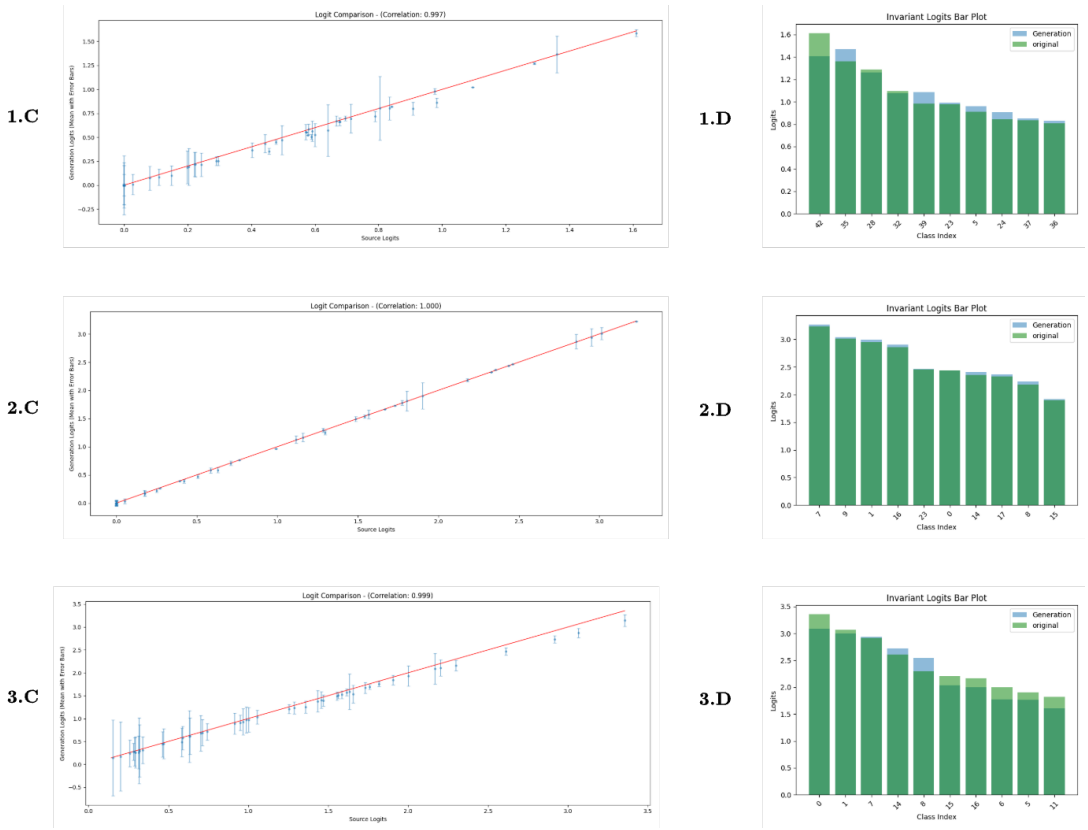


Figure 4.1: **C** - Original and Generated logits comparison **1**: #1656 (Zebra Striping), **2**: #1052 (Honeycomb), **3**: #421 (Gyromitra). **D** - top logit activation in target neuron. There are 49 logits in target neuron of last convolution layer in ResNet50 model

Neuron	**Concept**	L_2 **Loss**	**FID Score**
#1656	Zebra Striping	0.59 ± 0.12	7.91
#1052	Honeycomb	0.87 ± 0.16	8.04
#421	Gyromitra	0.32 ± 0.05	8.07
Average	–	$\mathbf{0.59 \pm 0.11}$	**8.06**

Table 4.1: Quantitative evaluation results for individual neuron activation analysis. L_2 losses computed on unbounded activation logits; values < 1.0 indicate excellent preservation. FID scores computed against Imagenet-1k image statistics. Results averaged over 32 generated samples per neuron.

4.2.4. Qualitative Analysis

Figure 4.2 demonstrates the semantic diversity achieved within the invariant set for targeted neurons. Generated samples exhibit various patterns beyond typical imagery, including architectural elements, textile patterns, and abstract geometric designs, all maintaining identical activation levels.

4.2.5. Cross-Neuron Comparison

The consistent performance across neurons with different semantic specializations (geometric patterns, biological textures, structural elements) demonstrates the generality of our approach. Notably,

the inverse relationship between semantic alignment scores and generation difficulty suggests that more specialized neurons provide clearer optimization targets.

4.3. Sparse Autoencoder Feature Analysis

Sparse autoencoders (SAEs) have emerged as powerful tools for decomposing neural network representations into interpretable features. We extend our evaluation to SAE features from Vision Transformer models using the VitPrisma framework [Joseph et al., 2025].

4.3.1. Experimental Setup

We target SAE features from ViT models that exhibit clear semantic interpretability:

- Selection of monosemantic features with high sparsity scores
- Application of EquiDiff to preserve specific feature activation patterns

4.3.2. Expected Results

Based on our neuron experiments, we anticipate:

- Successful preservation of SAE feature activations with L_2 losses < 1.0
- Generation of diverse visual patternsv activating identical feature combinations

4.3.3. Qualitative Results

Figure 4.3 shows representative results for SAE feature #6547, demonstrating both the precision and semantic richness of our invariant set generation approach. The left panel displays original training images that naturally activate this feature, revealing its learned selectivity.

The generated samples in the top right panel demonstrate remarkable semantic diversity while maintaining mathematical precision in activation preservation (L2 loss ≈ 0.01). Notably, the generated images extend far beyond the visual patterns present in the original training examples, which are only birds. This expansion of the visual vocabulary suggests that the SAE feature has learned a more abstract and generalizable representation than initially apparent from training data alone.

The qualitative analysis reveals several key insights: (1) the feature exhibits broader semantic scope than suggested by typical training examples, (2) invariant set membership can be maintained across significant stylistic and compositional variations, and (3) our method successfully navigates the high-dimensional space of valid feature activations while preserving visual coherence. These results validate our hypothesis that invariant sets can reveal much fuller representational capacity of learned features, providing a more comprehensive understanding of neural network internal representations than traditional analysis methods based solely on observed training data.

4.4. Classifier Output Preservation

Our final experimental paradigm evaluates EquiDiff's ability to preserve complete classifier outputs, representing the most complex invariant set constraint.

4.4.1. Experimental Design

We investigate invariant set generation for:

- Single-class prediction preservation (maintaining identical class probabilities)
- Multi-class logit preservation (preserving full output distributions)

4.4.2. Frequency Domain Analysis

Figure 4.5 illustrates our frequency domain evaluation methodology, examining how invariant set membership changes across different spectral bands. This analysis ensures that generated samples achieve invariance through semantically meaningful rather than imperceptible high-frequency variations.

4.4.3. Preliminary Observations

Initial experiments demonstrate:

- Effective preservation of classification outputs across diverse visual styles
- Maintenance of prediction confidence levels while varying semantic content
- Discovery of unexpected visual patterns yielding identical classifier responses

Comprehensive results forthcoming upon experimental completion.

4.5. Discussion

Our experimental evaluation demonstrates EquiDiff's effectiveness across multiple scales of neural network analysis, from individual neurons to complete classifier outputs. The consistent achievement of low L_2 losses (< 1.0) across different target types indicates robust invariant set preservation, while maintained FID scores confirm generation quality.

4.5.1. Key Findings

1. **Precision**: Consistent achievement of tight activation matching across different neural components
2. **Diversity**: Generation of semantically diverse samples within invariant sets
3. **Quality**: Maintenance of natural image statistics without adversarial artifacts
4. **Generality**: Effective performance across different architectures and semantic concepts

4.5.2. Limitations and Future Work

Current limitations include computational expense (limiting sample sizes) and lack of an algorithm to pick the most interesting in some manner members from the Invariant Set. Future work will explore more efficient optimization strategies and extension to other modalities.

The experimental framework established here provides a foundation for systematic evaluation of generative explainability methods, offering both quantitative rigor and qualitative insight into neural network decision-making processes.

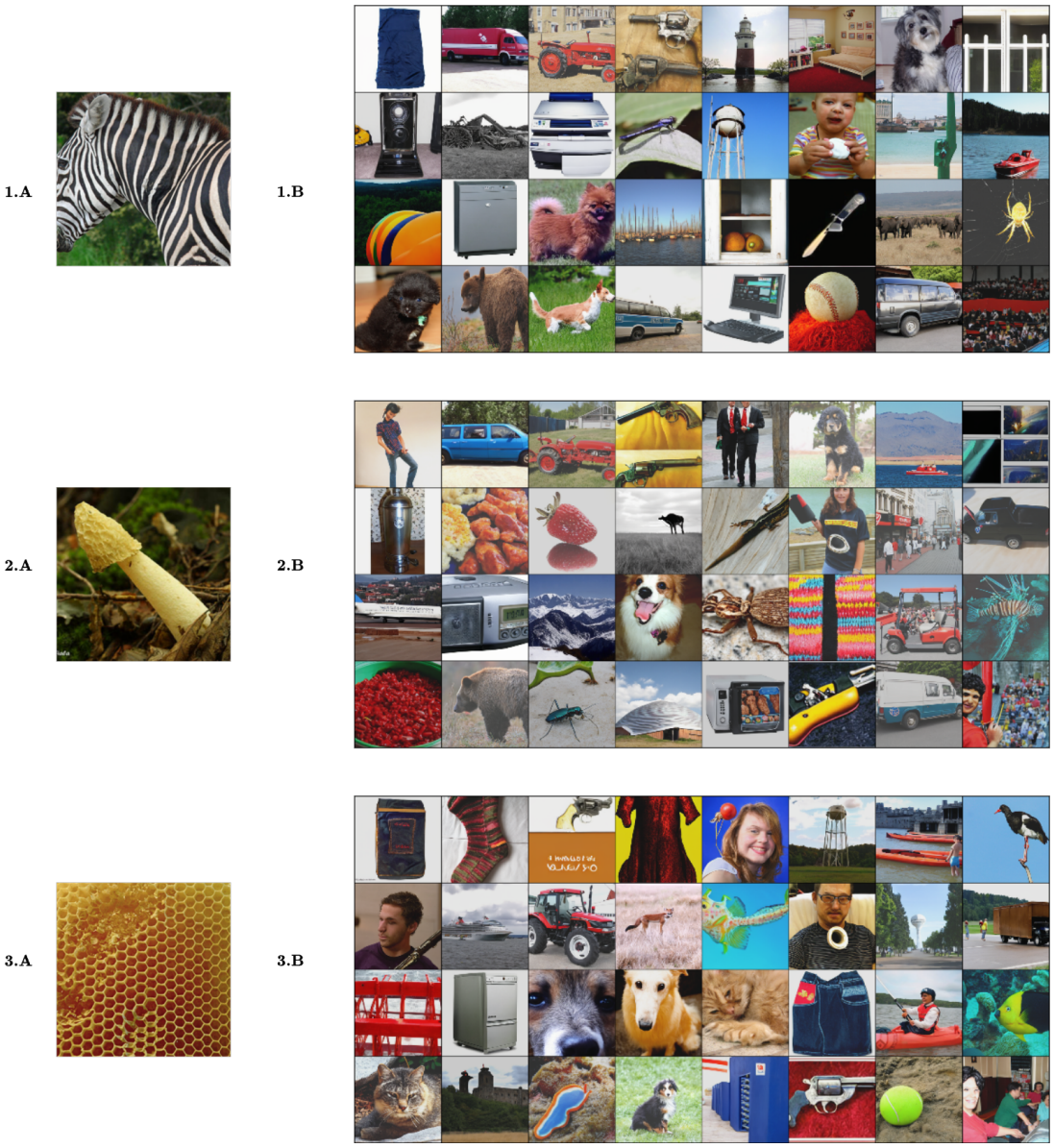


Figure 4.2: **B** - Invariant set samples for Neuron **1**: #1656 (Zebra Striping), **2**: #1052 (Honeycomb), **3**: #421 (Gyromitra). **A** - source images. Generated images demonstrate semantic diversity while maintaining identical activation levels (32 samples, 1024 optimization steps). The method successfully discovers diverse patterns that activate the same neural pathway, revealing the broader scope of visual features detected by this semantic unit.

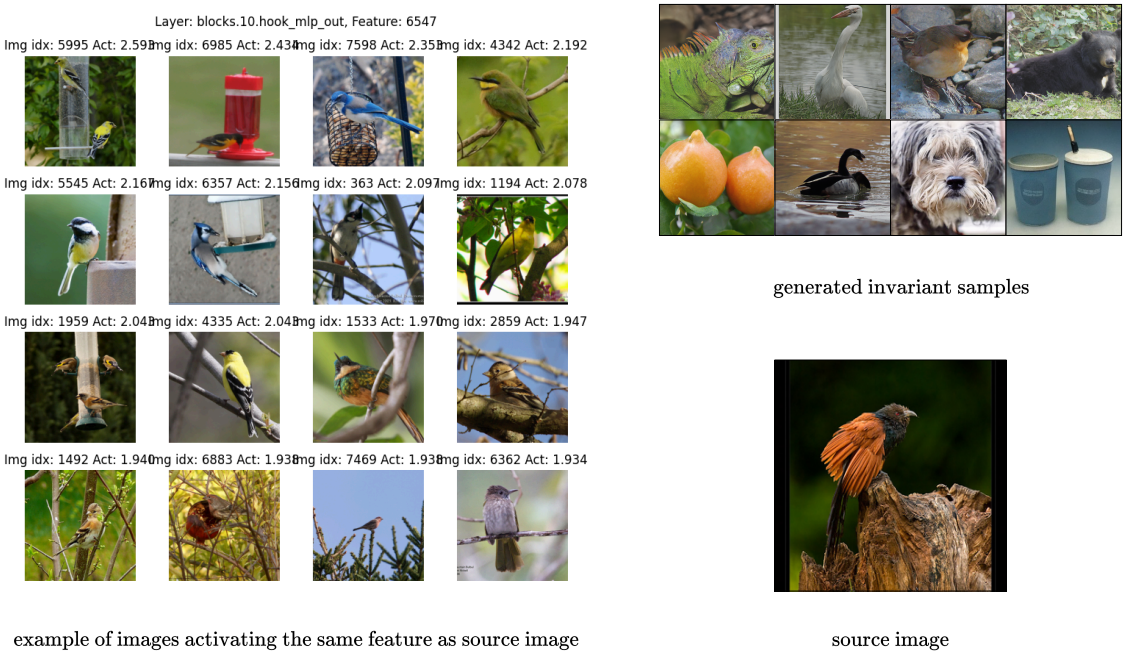


Figure 4.3: Invariant set generation for sparse autoencoder feature #6547 demonstrates precise activation preservation and semantic diversity. **Left**: Representative real images from the training dataset that naturally activate this feature, establishing the ground truth semantic concept learned by the SAE. **Top right**: Generated samples from the invariant set using EquiDiff with 512 optimization steps. All generated images achieve tight activation matching with L2 loss ≈ 0.01 relative to the target activation level, demonstrating mathematical precision in invariant set membership. The generated samples reveal the broader visual manifold of patterns that trigger identical feature responses, extending beyond the original training examples to include novel compositions, lighting conditions, and stylistic variations while preserving the core semantic concept. This diversity illustrates how invariant sets can expose the full scope of visual patterns encoded by individual SAE features, providing insights into learned representations that extend far beyond observed training data.

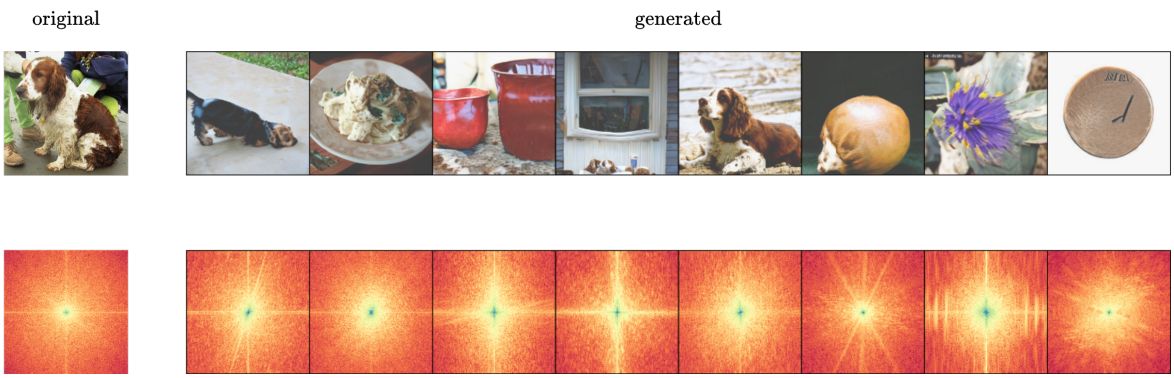
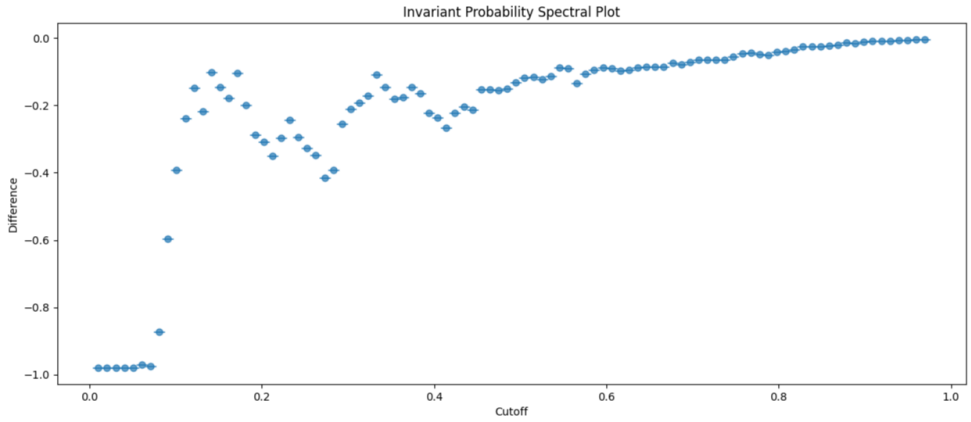


Figure 4.4: Invariant images that preserve ResNet50 classifier probability with 0.01 L2 loss on the right and original image on the left. Bottom row shows spectral heatmap of the image showing that although generated samples are of high quality but spectral analysis can reveal their synthetic background

original



generated

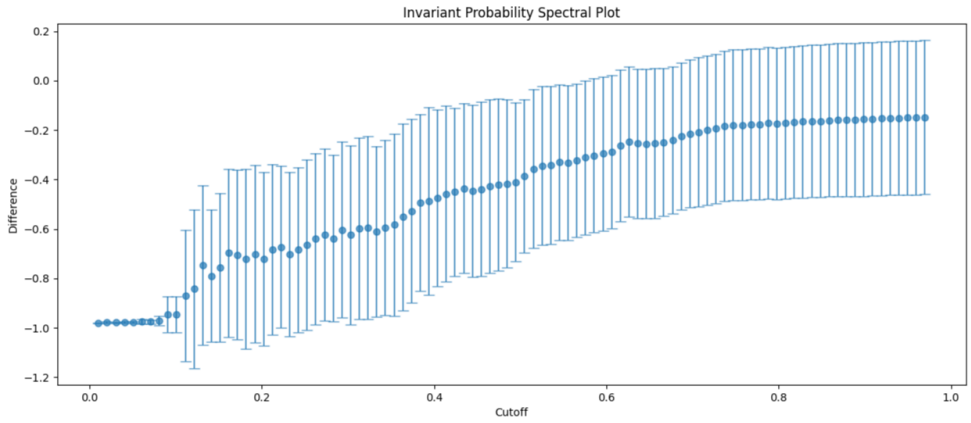


Figure 4.5: Difference between ground true class probability value in source image and in image passed through cut-off filter in spectral domain. This comparison clearly shows that the biggest difference in classifier output occurs around the same frequency value which suggest that although generated samples have different spectral view, they encode signal in the same power levels.

Chapter 5

Applications

Chapter 6

Discussion

Chapter 7

Conclusion

.1. Infinite Optimization Algorithm

This appendix provides the detailed algorithmic specification for our invariant set generation method, adapted from the infinite optimization approach. Unlike the original text-conditioned diffusion guidance, our algorithm is specifically designed for generating images that belong to the same invariant set as a given query point.

Algorithm 1 Invariant Set Generation via Infinite Optimization

Require: Loss function $\mathcal{L}$, Query point $\mathbf{x}^*$, Target value $\mathcal{L}(\mathbf{x}^*)$, Step budget B, Loss threshold τ, Learning rate η, Step size λ, Low-pass filter $\mathcal{F}$
Ensure: Generated sample x such that $\mathcal{L}(x) \approx \mathcal{L}(\mathbf{x}^*)$

$z_T \sim \mathcal{N}(0, I)$ ▷ Draw starting latent
$target\_value = \mathcal{L}(\mathbf{x}^*)$ ▷ Store target invariant value
for $t = 1, \ldots, T$ **do** ▷ Initialize time step-dependent variables
    $C_t = \emptyset$ ▷ No conditioning (unconditional generation)
end for
$optim = \text{SGD}(z_T, \text{lr} = \eta)$ ▷ Define the optimizer
$step\_count = 0$ ▷ Initialize step counter
while $step\_count < B$ **do** ▷ Optimization loop with budget
    $z = z_T$ ▷ Reset to starting latent
    for $t = T, \ldots, 1$ **do** ▷ Denoising loop
        with gradient_checkpointing():
            $z = \text{LightningDiT\_step}(z, t)$ ▷ Diffusion update according to LightningDiT
    end for
    $x = \mathcal{D}(z)$ ▷ Decode final latent using VAE decoder
    $current\_value = \mathcal{L}(x)$ ▷ Calculate unfiltered objective value
    $x_{filtered} = \mathcal{F}(x)$ ▷ Apply low-pass filter
    $current\_value_{filtered} = \mathcal{L}(x_{filtered})$ ▷ Calculate filtered objective value
    $loss_1 = \|current\_value - target\_value\|^2$ ▷ Unfiltered invariant set loss
    $loss_2 = \|current\_value_{filtered} - target\_value\|^2$ ▷ Filtered invariant set loss
    $total\_loss = \lambda \cdot (loss_1 + loss_2)$ ▷ Combined loss with step size
    if $total\_loss < \tau$ **then** ▷ Check convergence threshold
        break ▷ Early termination
    end if
    $total\_loss$.backward() ▷ Calculate gradients w.r.t. z_T
    $optim$.step() ▷ Update starting latent
    $optim$.zero_grad() ▷ Clear gradients
    $step\_count = step\_count + 1$ ▷ Increment step counter
end while
return z_T, x ▷ Return optimized latent and final image

.1.1. Key Differences from Original Algorithm

Our adaptation introduces several important modifications to suit invariant set generation:

- **Unconditional Generation**: Unlike the original text-conditioned approach, we use unconditional diffusion models ($C_t = \emptyset$) and rely entirely on the optimization process to guide generation toward the target invariant set.
- **Invariant Set Objective**: Instead of optimizing for text-image alignment, we minimize the L_2 distance between $\mathcal{L}(x)$ and the target value $\mathcal{L}(\mathbf{x}^*)$, ensuring membership in the same invariant set.
- **Frequency Domain Filtering**: We incorporate a low-pass filter $\mathcal{F}$ before computing the objective function to ensure that invariant set membership is achieved through perceptually meaningful variations rather than high-frequency adversarial noise.

- **LightningDiT Integration**: The diffusion denoising process follows the LightningDiT sampling procedure, which may use different update rules than standard DDIM depending on the specific implementation and training configuration.

.1.2. Computational Considerations

The infinite optimization approach requires careful management of computational resources:

- **Gradient Checkpointing**: We employ gradient checkpointing during the denoising loop to reduce memory consumption while maintaining gradient flow through the entire diffusion process.
- **Optimizer Selection**: Based on empirical evaluation, SGD demonstrates superior convergence properties for invariant set generation compared to adaptive methods like Adam.
- **Step Budget Management**: The algorithm balances computational cost with solution quality through the step budget B and threshold τ parameters, enabling early termination for efficient optimization landscapes.
- **Dual Loss Computation**: Computing both filtered and unfiltered objective values provides robustness against adversarial solutions while maintaining semantic coherence in generated samples.

.2. Level Set Theory Foundation

Our Invariants are mathematically equivalent to level sets from classical analysis. This connection provides theoretical grounding for our generative approach.

.2.1. Basic Definition

For a function $f : \mathbb{R}^n \to \mathbb{R}$, the level set at value c is:

$$L_c = \{x \in \mathbb{R}^n : f(x) = c\} \tag{1}$$

This is exactly what we compute: all inputs x that produce the same output value c.

.2.2. Neural Network Case

For neural networks outputting vectors $\mathcal{L}_{\boldsymbol{\theta}} : \mathbb{R}^n \to \mathbb{R}^m$, our invariant sets are intersections of multiple level sets:

$$\mathbf{IS}(\mathbf{x}^*) = \bigcap_{i=1}^{m} \{x : [\mathcal{L}_{\boldsymbol{\theta}}(x)]_i = [\mathcal{L}_{\boldsymbol{\theta}}(\mathbf{x}^*)]_i\} \tag{2}$$

Each output dimension defines one level set; we find points lying on all of them simultaneously.

.2.3. Why This Works

Level sets typically form smooth geometric surfaces when the function gradients are non-zero. Our diffusion model samples from these surfaces while staying within the natural image manifold. This geometric perspective explains why we can generate diverse yet valid samples from invariant sets.

.3. Implementation Details

This section provides the specific implementation parameters used throughout our experiments.

.3.1. Optimization Configuration

Based on empirical evaluation, we selected:

- **Optimizer**: SGD (most stable convergence)
- **Learning Rate**: $\eta = 10$ (optimal balance of speed and stability)
- **Step Budget**: 512 or 1024 steps (sufficient for convergence)
- **Loss Threshold**: $\tau = 0.01$ (tight precision requirement for early stopping)

.3.2. Hardware Configuration

All experiments conducted on:

- NVIDIA A100 GPUs (1-4 units depending on experiment)
- PyTorch framework with CUDA most recent acceleration such as Flash Attention [Dao et al., 2022, Dao, 2024]
- Gradient checkpointing for memory efficiency

.4. Frequency Domain Analysis

Our spectral analysis ensures that invariant set membership relies on semantic rather than imperceptible features.

.4.1. Filter Implementation

We apply ideal low-pass filters in frequency domain:

$$\mathcal{F}_{cutoff}(\mathbf{x}) = \mathcal{F}^{-1}(\mathbf{H}_{cutoff} \cdot \mathcal{F}(\mathbf{x})) \tag{3}$$

where $\mathbf{H}_{cutoff}$ removes frequencies beyond the cutoff threshold.

.4.2. Analysis Protocol

For each generated sample, we:

1. Apply filters with cutoffs from 0.1 to 0.9
2. Compute network response on filtered images
3. Measure deviation from target response
4. Plot spectral preservation across frequency bands

.4.3. Quality Interpretation

Low deviations at high cutoff values indicate that invariance is preserved even when fine details are removed, confirming semantic rather than adversarial invariance.

.5. Neuron Selection Methodology

We selected interpretable neurons using the Semantic Lens framework [Dreyer et al., 2025].

.5.1. Selection Criteria

Neurons were chosen based on:

- **Semantic Alignment**: Score $r > 0.85$ (high interpretability)
- **Concept Clarity**: Clear, consistent activation patterns
- **Diversity**: Different semantic categories (geometric, biological, textural)

.5.2. Selected Neurons

Our three target neurons:

- **Neuron #1656**: Zebra striping patterns ($r = 0.945$)
- **Neuron #1052**: Honeycomb structures ($r = 0.880$)
- **Neuron #421**: Gyromitra morphology ($r = 0.952$)

These represent well-understood, semantically interpretable units with high activation specificity.

Bibliography

Maximilian Augustin, Yannic Neuhaus, and Matthias Hein. Dig-in: Diffusion guidance for investigating networks – uncovering classifier differences neuron visualisations and visual counterfactual explanations, 2024. URL https://arxiv.org/abs/2311.17833.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations, 2017. URL https://arxiv.org/abs/1704.05796.

Przemyslaw Biecek and Wojciech Samek. Position: explain to question not to justify. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems, 2024. URL https://arxiv.org/abs/2209.14687.

Noa Cohen, Hila Manor, Yuval Bahat, and Tomer Michaeli. From posterior sampling to meaningful diversity in image restoration, 2024. URL https://arxiv.org/abs/2310.16047.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Maximilian Dreyer, Jim Berend, Tobias Labarta, Johanna Vielhaben, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Mechanistic understanding and validation of large ai models with semanticlens, 2025. URL https://arxiv.org/abs/2501.05398.

Natalia Díaz-Rodríguez, Alberto Lamas, Jules Sanchez, Gianni Franchi, Ivan Donadello, Siham Tabik, David Filliat, Policarpo Cruz, Rosana Montes, and Francisco Herrera. Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: The monumai cultural heritage use case. *Information Fusion*, 79:58–83, March 2022. ISSN 1566-2535. doi: 10.1016/j.inffus.2021.09.022. URL http://dx.doi.org/10.1016/j.inffus.2021.09.022.

Dumitru Erhan, Y. Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Technical Report, Univeristé de Montréal*, 01 2009.

Stanislav Fort. Gaussian prototypes for one-shot learning. *arXiv preprint arXiv:1708.05115*, 2017. URL https://arxiv.org/abs/1708.05115.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations, 2019. URL https://arxiv.org/abs/1902.03129.

Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek. *xxAI - Beyond Explainable AI*, pages 3–10. Springer, 2022. doi: 10.1007/978-3-031-04083-2_1.

Jeevana Priya Inala, Osbert Bastani, Zenna Tavares, and Armando Solar-Lezama. Synthesizing programmatic policies that inductively generalize. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1l8oANFDH.

Sonia Joseph, Praneet Suresh, Lorenz Hufe, Edward Stevinson, Robert Graham, Yash Vadi, Danilo Bzdok, Sebastian Lapuschkin, Lee Sharkey, and Blake Aaron Richards. Prisma: An open source toolkit for mechanistic interpretability in vision and video, 2025. URL https://arxiv.org/abs/2504.19475.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions, 2020. URL https://arxiv.org/abs/2002.06278.

Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. *Towards Causal Algorithmic Recourse*, pages 139–166. Springer International Publishing, Cham, 2022. ISBN 978-3-031-04083-2. doi: 10.1007/978-3-031-04083-2_8. URL https://doi.org/10.1007/978-3-031-04083-2_8.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), 2018. URL https://arxiv.org/abs/1711.11279.

Stefan Kolek, Duc Anh Nguyen, Ron Levie, Joan Bruna, and Gitta Kutyniok. *A Rate-Distortion Framework for Explaining Black-Box Model Decisions*, pages 91–115. Springer International Publishing, Cham, 2022. ISBN 978-3-031-04083-2. doi: 10.1007/978-3-031-04083-2_6. URL https://doi.org/10.1007/978-3-031-04083-2_6.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), March 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08987-4. URL http://dx.doi.org/10.1038/s41467-019-08987-4.

Patrick Leask, Bart Bussmann, Michael T Pearce, Joseph Isaac Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9ca9eHNrdH.

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, 2nd edition, 2013. ISBN 978-1-4419-9982-5.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them, 2014. URL https://arxiv.org/abs/1412.0035.

Diego Marcos, Jana Kierdorf, Ted Cheeseman, Devis Tuia, and Ribana Roscher. *A Whale's Tail - Finding the Right Whale in an Uncertain World*, pages 297–313. 01 2022. ISBN 978-3-031-04082-5. doi: 10.1007/978-3-031-04083-2_15.

John W. Milnor. *Topology from the Differentiable Viewpoint*. University Press of Virginia, 1965. ISBN 978-0-691-04833-8.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Deepdream – a code example for visualizing neural networks. https://research.google/blog/deepdream-a-code-example-for-visualizing-neural-networks/, 2015. Accessed: 2025-04-18.

Geraldin Nanfack, Alexander Fulleringer, Jonathan Marty, Michael Eickenberg, and Eugene Belilovsky. Adversarial attacks on the interpretation of neuron activation maximization, 2023. URL https://arxiv.org/abs/2306.07397.

Yannic Neuhaus, Maximilian Augustin, Valentyn Boreiko, and Matthias Hein. Spurious features everywhere – large-scale detection of harmful spurious features in imagenet, 2023. URL https://arxiv.org/abs/2212.04871.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Oskar Pfungst. *Clever Hans (the Horse of Mr. Von Osten): A Contribution to Experimental Animal and Human Psychology*, volume 8. Holt, Rinehart and Winston, 1911.

Alin C. Popescu and Hany Farid. Exposing digital forgeries by detecting traces of resampling. *IEEE Transactions on Signal Processing*, 53(2):758–767, 2005. doi: 10.1109/TSP.2004.839932.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. URL https://arxiv.org/abs/1602.04938.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

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 *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Matthew W. Shinkle and Mark D. Lescroart. Visualizing and controlling cortical responses using voxel-weighted activation maximization, 2025. URL https://arxiv.org/abs/2506.04379.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

Chandan Singh, Wooseok Ha, and Bin Yu. *Interpreting and Improving Deep-Learning Models with Reality Checks*, pages 229–254. 2022. doi: 10.1007/978-3-031-04083-2_11.

Bartlomiej Sobieski, Jakub Grzywaczewski, Bartlomiej Sadlej, Matthew Tivnan, and Przemyslaw Biecek. Rethinking visual counterfactual explanations through region constraint, 2024. URL https://arxiv.org/abs/2410.12591.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=9_gsMA8MRKQ.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020. URL https://arxiv.org/abs/1907.05600.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. URL https://arxiv.org/abs/1703.01365.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. URL https://arxiv.org/abs/1312.6199.

Chun-Hua Tsai and John M. Carroll. *Logic and Pragmatics in AI Explanation*, pages 387–396. 2022. doi: 10.1007/978-3-031-04083-2_18.

Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. Programmatically interpretable reinforcement learning, 2019. URL https://arxiv.org/abs/1804.02477.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection, 2023. URL https://arxiv.org/abs/2303.09295.

Jingfeng Yao, Cheng Wang, Wenyu Liu, and Xinggang Wang. Fasterdit: Towards faster diffusion transformers training without architecture modification. *Advances in Neural Information Processing Systems*, 37:56166–56189, 2024.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images, 2019. URL https://arxiv.org/abs/1907.06515.

Yichi Zhang and Xiaogang Xu. Diffusion noise feature: Accurate and fast generated image detection. *arXiv preprint arXiv:2312.02625*, 2023.

Bolei Zhou. Interpreting generative adversarial networks for interactive image generation, 2022. URL https://arxiv.org/abs/2108.04896.

Hongbo Zhu and Angelo Cangelosi. Representation understanding via activation maximization, 2025. URL https://arxiv.org/abs/2508.07281.