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

Master'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, this work argues 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. This work has three contributions. The first one is a paradigm shift from traditional explainable AI (XAI) methods, which find human-interpretable features in known data, to generative XAI methods that synthesize new samples. Secondly, this work introduces a new framework Invariantsfor generative XAI and thirdly, it proposes an efficient diffusion-based method EquiDifffor exploring it. Evaluation of this method on popular models such as ResNet-50 or Sparse Autoencoders (SAE) highlights significant limitations of current XAI methods. The second contribution of this work is a new method for exploring the invariant sets of neural networks. The third contribution is a new method for exploring the invariant sets of neural networks.

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?

The phrase "a dog with the same confidence" refers to something far more profound and potentially disturbing than might initially appear. This concept extends beyond different breeds of dogs, dogs photographed from different angles, or even dogs rendered in different artistic styles. The phenomenon encompasses 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 assigns 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 XAI 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. Proposed Approach: Generative XAI

This thesis introduces a paradigm shift from traditional interpolative XAI methods to a generative approach. Instead of analyzing existing data points, this work proposes 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.

The proposed method 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, one 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 proposed 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.

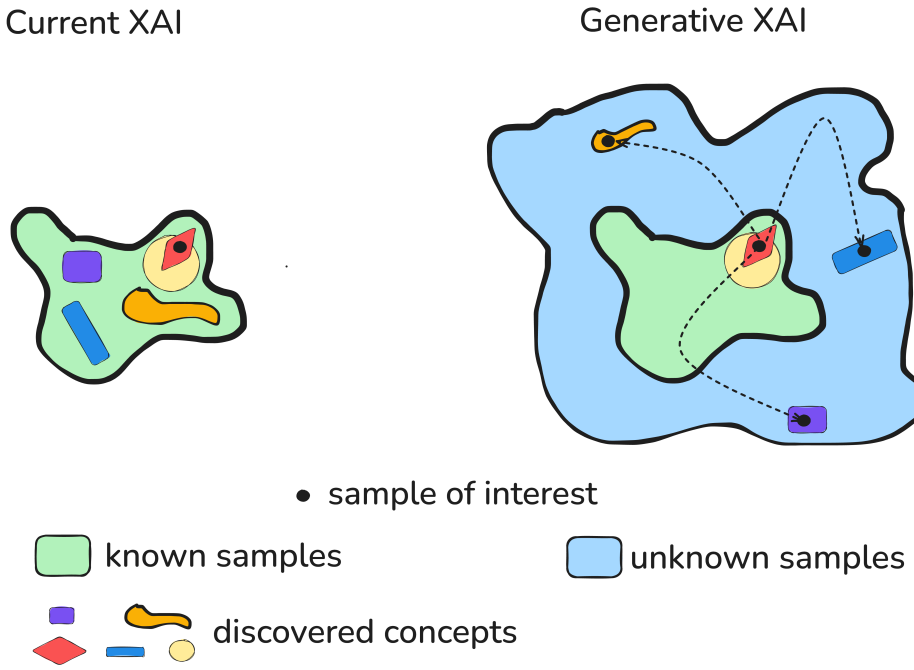


Figure 1.1: Conceptual comparison between traditional XAI methods and Generative XAI. Traditional methods analyze known data samples (left), while proposed approach synthesizes diverse examples from the invariant set that yield identical predictions (right).

1.3. Contributions

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

The first contribution represents a **paradigm shift from traditional XAI 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 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 an **efficient algorithmic implementation of this 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 both the theoretical foundations and empirical validation of proposed approach are presented.

1.4. Thesis Organization

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

Chapter 4 presents a comprehensive experimental evaluation demonstrating the effectiveness of proposed method across multiple neural network analysis paradigms, from individual neuron activation to complete classifier output preservation. Chapter 5 synthesizes contributions and their significance for the field of XAI. It also outlines future work and possible applications of this contribution.

Chapter 2

Related Work

This chapter reviews the relevant literature across several interconnected areas that form the foundation of this work. The chapter begins 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. As neural networks have grown from simple perceptrons to massive transformer architectures with billions of parameters, the need for interpretability has become increasingly critical for deployment in high-stakes domains such as healthcare, finance, and autonomous systems. The fundamental challenge lies in bridging the semantic gap between the mathematical operations performed by neural networks and human-understandable concepts. Current approaches to explainable AI can be broadly categorized into several paradigms, each with distinct methodological foundations and complementary strengths and limitations.

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 proposed approach. This method is, to best 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, the goal of this work 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, one starts 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. To understand how conditioning works, one must first establish the mathematical foundation of score functions and their conditional decomposition.

2.3.1. Score Function Fundamentals

The score function $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ represents the gradient of the log-probability density with respect to the input. In the context of diffusion models, the score function $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ provides the direction of steepest ascent in log-probability space at diffusion time t, essentially pointing toward regions of higher probability density. This geometric interpretation is crucial for understanding how diffusion models learn to reverse the noise process: by following the score function, one moves from low-probability noisy regions toward high-probability clean data regions.

The conditional score function $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t \mid \mathbf{y}, t)$ extends this concept to conditional distributions, providing gradients that guide generation toward samples that satisfy the conditioning constraint $\mathbf{y}$. The key insight is that this conditional score can be decomposed using Bayes' theorem, allowing us to separate the unconditional generative component from the conditioning component.

2.3.2. Conditional Score Decomposition

Starting from Bayes' theorem for conditional probabilities:

$$p(\mathbf{x}_t \mid \mathbf{y}, t) = \frac{p(\mathbf{y} \mid \mathbf{x}_t, t)p(\mathbf{x}_t, t)}{p(\mathbf{y}, t)}$$

Taking the logarithm of both sides:

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

Since the marginal probability $p(\mathbf{y}, t)$ does not depend on $\mathbf{x}_t$, its gradient with respect to $\mathbf{x}_t$ is zero. Therefore, taking the gradient with respect to $\mathbf{x}_t$ yields the fundamental conditional score decomposition:

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

2.3.3. 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.

2.3.4. Limitations of Standard Classifier Guidance

While effective for class-conditional generation, standard classifier guidance has several fundamental limitations that significantly impact its applicability to invariant set generation. These constraints arise from the architecture of diffusion models and the mathematical formulation of the guidance mechanism itself.

Limited Optimization Horizon and Temporal Constraints: The first major limitation stems from the inherently discrete and temporally constrained nature of the diffusion sampling process. Standard classifier guidance applies conditioning signals only at predetermined timesteps during the denoising trajectory, typically following a fixed schedule (e.g., every 10 steps out of 1000 total steps).

The mathematical consequence of this limitation becomes apparent when considering the precision requirements for other generations. While class-conditional generation can tolerate approximate conditioning (e.g., generating "roughly dog-like" images), it does not allow to iterating until some condition is meet, rather untill the schedule is finished. The discrete optimization steps available in classifier guidance provide insufficient granularity to achieve arbitraty high precision, particularly for complex objective functions with narrow convergence basins.

In practical applications, this limitation manifests as generated samples that approximate but do not precisely satisfy the given condition, leading to activation mismatches that can accumulate and compromise the interpretability of the results. For instance, when attempting to preserve specific neuron activations, the discrete guidance steps may succeed in maintaining the general semantic concept but fail to achieve the exact activation value.

Latent Space Misalignment and Representational Incompatibility: The second critical limitation arises from the architectural choice of modern diffusion models to operate in compressed latent spaces rather than directly in pixel space. This design, exemplified by Latent Diffusion Models (LDMs) [Rombach et al., 2022], introduces a fundamental representational mismatch between the diffusion process and the neural networks being analyzed.

The mathematical formulation of this problem is subtle but profound. The diffusion model operates on encoded representations $\mathbf{z}_t = \mathcal{E}(\mathbf{x}_t)$ where $\mathcal{E}$ is a learned encoder (typically from a variational autoencoder), while the target neural network $f_{\boldsymbol{\theta}}$ operates on natural images $\mathbf{x}$. Classifier guidance requires evaluating $\nabla_{\mathbf{x}_t} \log p(\mathbf{y} \mid \mathbf{x}_t, t)$ at intermediate diffusion timesteps, but the noisy intermediate states $\mathbf{x}_t$ may not correspond to meaningful inputs for the classifier network.

This mismatch has several cascading consequences. First, training timestep-specific classifiers $p_{\boldsymbol{\phi}}(\mathbf{y} \mid \mathbf{x}_t, t)$ requires extensive additional data collection and training, essentially requiring a separate classifier for each timestep t. These classifiers must learn to operate on partially denoised, potentially unrealistic images, which significantly complicates the training process and may introduce systematic biases. Second, using approximate reconstructions $\hat{\mathbf{x}}_0(t)$ to evaluate the classifier introduces prediction errors that compound throughout the sampling process, potentially driving generation away from true invariant set membership.

The practical impact is particularly severe for fine-grained objectives like individual neuron activations or sparse autoencoder features, where small representational inconsistencies can have significan result on the final image. The latent space encoding may not preserve the specific visual patterns that activate particular neurons, leading to guidance signals that are misaligned with the true optimization objective.

Objective Function Generalizability and Mathematical Constraints: The third fundamental limitation concerns the restricted mathematical formulation of standard classifier guidance, which is specifically designed for classification objectives of the form $p(\mathbf{y} \mid \mathbf{x}_t, t)$ where $\mathbf{y}$ represents class labels or categorical conditions. This formulation, while elegant for its intended purpose, creates significant barriers when adapting to the diverse range of objective functions required for comprehensive neural network analysis.

The standard guidance formulation assumes that the conditioning variable $\mathbf{y}$ can be meaningfully interpreted as a class probability distribution, enabling the computation of log-probabilities and their gradients. However, one can require conditioning on arbitrary differentiable functions such as individual neuron activations (real-valued scalars), sparse autoencoder feature combinations (high-dimensional vectors), or complex geometric properties of the decision boundary (potentially nonlinear manifolds in activation space).

Adapting classifier guidance to these objectives requires substantial mathematical reformulation, including the design of appropriate loss functions, normalization schemes, and gradient computation strategies. For example, when targeting a specific neuron activation value a^*, one must define a pseudo-probability distribution over activation values and ensure that the resulting gradients provide meaningful guidance signals. This often involves ad-hoc transformations like $p(a^* \mid \mathbf{x}_t, t) = \exp(-\lambda||f_n(\mathbf{x}_t) - a^*||^2)$ where λ is a temperature parameter that must be carefully tuned.

The consequences extend beyond mathematical complexity to fundamental questions of convergence and stability. The guidance gradients derived from these adapted objective functions may not exhibit the favorable convergence properties of the original classification formulation, potentially leading to unstable optimization dynamics, mode collapse, or failure to reach the target invariant set. Moreover, the interaction between multiple objectives (e.g., simultaneously constraining several neuron activations) becomes mathematically intractable within the standard guidance framework, limiting the approach to simple, single-objective scenarios.

These three limitations collectively demonstrate why standard classifier guidance, despite its success in class-conditional generation, has tsx fundamental limitations. This analysis directly motivates proposed infinite optimization approach, which addresses each of these constraints through decoupled optimization, native pixel-space operation, and arbitrary objective function support.

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 proposed 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 proposed 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 that enforces constraints to encourage natural-looking images, and λ controls the strength of regularization relative to the primary objective.

2.5.1. The Critical Role of Regularization in Neural Visualization

The regularization term $\mathcal{R}(\mathbf{x})$ represents one of the most fundamental challenges in neural network interpretability: ensuring that synthetic explanations reflect semantically meaningful patterns rather than exploiting imperceptible statistical quirks in the learned representations. Without appropriate regularization, activation maximization degenerates into adversarial optimization, producing images that achieve maximal neural activation through high-frequency noise patterns, texture irregularities, or other artifacts that are invisible to human perception but strongly trigger specific computational pathways.

The mathematical necessity for regularization arises from the high-dimensional nature of the optimization landscape. Neural networks, particularly deep convolutional architectures, exhibit complex response surfaces with numerous local maxima that correspond to spurious activation patterns. Without constraints, gradient-based optimization will exploit these pathways, leading to solutions that satisfy the mathematical objective while completely failing the interpretability goal. This fundamental tension between mathematical optimality and perceptual meaningfulness defines the core challenge of activation maximization.

2.5.2. Classical Regularization Approaches

Traditional regularization strategies for activation maximization fall into several categories, each addressing different aspects of the realism constraint:

Total Variation Regularization: The most commonly employed approach uses total variation (TV) penalties of the form $\mathcal{R}_{TV}(\mathbf{x}) = \sum_{i,j} |\mathbf{x}_{i+1,j} - \mathbf{x}_{i,j}| + |\mathbf{x}_{i,j+1} - \mathbf{x}_{i,j}|$, which encourages spatial smoothness by penalizing large gradients between adjacent pixels. While computationally efficient and mathematically well-defined, TV regularization often produces overly smoothed results that lack the fine-grained details characteristic of natural images, leading to blob-like visualizations that obscure important textural features.

Frequency Domain Constraints: Recognizing that natural images exhibit specific spectral characteristics, frequency-based regularization methods constrain the power distribution across spatial frequencies. These approaches typically apply band-pass filters or spectral penalties to encourage generated images to match the $1/f$ power law observed in natural image statistics. However, naive frequency constraints can be overly restrictive, suppressing legitimate high-frequency details while failing to address more subtle forms of adversarial exploitation.

Statistical Prior Matching: More sophisticated approaches attempt to match higher-order statistical properties of natural images, including local contrast distributions, edge orientation histograms, and texture statistics. These methods often involve complex optimization procedures and may require extensive parameter tuning, limiting their practical applicability while still failing to guarantee perceptual realism.

2.5.3. Perceptual Metrics and Deep Regularization

The limitations of classical approaches have motivated the development of perceptually-aware regularization methods that leverage learned representations of visual similarity:

LPIPS (Learned Perceptual Image Patch Similarity): The LPIPS metric [Zhang et al., 2018] represents a significant advancement in perceptual regularization, utilizing features from pre-trained neural networks (such as AlexNet or VGG [Krizhevsky et al., 2012, Simonyan and Zisserman, 2014])

to measure perceptual distance between images. Unlike pixel-based metrics that treat all spatial frequencies equally, LPIPS weights differences according to human perceptual sensitivity, providing a more meaningful measure of visual similarity. In the context of activation maximization, LPIPS can be incorporated as $\mathcal{R}_{LPIPS}(\mathbf{x}) = \text{LPIPS}(\mathbf{x}, \mathbf{x}_{natural})$ where $\mathbf{x}_{natural}$ represents a reference natural image or a distribution of natural images.

The mathematical formulation of LPIPS involves computing feature representations $\phi_l(\mathbf{x})$ at multiple layers l of a pre-trained network, then measuring weighted L_2 distances: $\text{LPIPS}(\mathbf{x}, \mathbf{y}) = \sum_l w_l \|\phi_l(\mathbf{x}) - \phi_l(\mathbf{y})\|_2^2$ where w_l are learned layer weights that reflect perceptual importance. This approach effectively uses one neural network to regularize the visualization of another, creating a hierarchical constraint system that can capture both low-level textural properties and high-level semantic consistency.

Feature Distribution Matching: Beyond pairwise similarity metrics, advanced regularization approaches constrain generated images to lie within the natural image manifold by matching statistical properties of deep feature distributions. These methods may employ techniques such as maximum mean discrepancy (MMD) or adversarial losses to ensure that synthetic visualizations exhibit feature statistics consistent with natural imagery across multiple representation levels [Goodfellow et al., 2014, Dziugaite et al., 2015].

Gram Matrix Constraints: Inspired by neural style transfer, some approaches regularize activation maximization using Gram matrix constraints that preserve spatial correlations between feature maps while allowing optimization of the primary objective. This approach can maintain textural coherence while permitting the emergence of activation-specific patterns [Gatys et al., 2016].

2.5.4. The Fundamental Challenge of Realistic Generation

Despite these advances, generating truly realistic images through activation maximization remains an outstanding challenge with profound implications for explainable AI. The core difficulty lies in the fundamental mismatch between the objectives of neural networks (optimized for task performance) and the constraints of natural image generation (governed by complex physical and perceptual processes).

Neural networks, particularly those trained on large-scale datasets, develop internal representations that capture statistical regularities in training data but may not respect the underlying generative processes that produce natural images. When activation maximization attempts to reverse-engineer these representations, it encounters the problem that multiple distinct natural phenomena may activate the same neural pathway, while the optimization process tends to find the most mathematically efficient (often unrealistic) combination of these activating features.

This tension is particularly acute for neurons with complex, multi-faceted selectivity patterns. For instance, a neuron that responds to both curved edges and specific texture patterns may be maximally activated by an image containing impossible combinations of these features—curved edges with unnatural texture properties that could not exist in real objects. Traditional regularization approaches struggle to eliminate such impossible combinations while preserving the legitimate activation patterns that make the visualization interpretable.

The implications extend beyond individual visualization quality to the broader epistemological foundations of neural network interpretability. If our primary tools for understanding neural representations systematically produce unrealistic explanations, we risk developing misleading intuitions about how these systems actually process natural inputs. This represents the central challenge of non-adversarial explainability: developing interpretation methods that reveal genuine computational strategies rather than artifacts of the interpretation process itself.

Furthermore, the computational expense of sophisticated regularization approaches often makes them impractical for large-scale analysis, creating a trade-off between visualization quality and an-

alytical scope. This limitation has motivated interest in alternative approaches, such as proposed diffusion-based method, that can leverage powerful generative priors to ensure realism while maintaining computational tractability.

2.5.5. Limitations and Relationship to this Work

Although activation maximization shares the goal of understanding model behavior through synthetic inputs, it differs fundamentally from proposed approach in ways that reveal critical limitations of single-sample interpretation methods and highlight the necessity of diverse, multi-sample analysis for comprehensive neural network understanding.

The Diversity Imperative: Why Single Solutions Fail to Capture Neural Complexity

The most fundamental limitation of traditional activation maximization lies in its pursuit of a single optimal solution rather than exploring the diverse space of inputs that activate neural pathways. This single-solution paradigm represents a profound philosophical and methodological constraint that severely limits our understanding of neural network decision-making processes.

Neural networks, particularly deep architectures trained on complex visual tasks, develop representations that are inherently multi-faceted and compositional. A single neuron may respond to diverse combinations of visual features: edges at specific orientations, particular color combinations, textural patterns, or higher-order statistical regularities in image structure. When activation maximization converges to a single "optimal" stimulus, it provides only one possible interpretation of this complex feature space, potentially missing equally valid—and often more interpretable—alternative patterns that achieve the same level of activation.

This limitation becomes particularly pronounced when we consider the high-dimensional nature of neural activation landscapes. The optimization surface for maximizing neuron activation typically contains multiple local maxima, each corresponding to different ways the neuron can be activated. Traditional optimization methods, constrained by computational budgets and convergence criteria, tend to settle into the first sufficiently strong local maximum encountered, never exploring the broader topology of activation-triggering patterns.

The diversity of activation patterns is not merely a technical curiosity but provides critical insights into the robustness, generalization, and potential failure modes of neural networks. Consider a hypothetical neuron that achieves maximal activation through three distinct pathways: geometric patterns with high contrast edges, specific color combinations under particular lighting conditions, and textural regularities found in biological surfaces. A single-solution activation maximization approach might converge to only one of these patterns—perhaps the highest-contrast geometric configuration—leaving the other two activation modes completely unexplored and potentially unrecognized.

From a scientific interpretability perspective, this represents a fundamental sampling bias in our understanding of neural representations. If our primary tool for neural interpretation systematically undersamples the space of activating patterns, we develop incomplete and potentially misleading theories about what these networks have learned. This is particularly problematic for safety-critical applications where understanding the full range of inputs that can trigger specific network behaviors is essential for identifying potential failure modes or adversarial vulnerabilities.

Maximum Activation vs. Preserved Predictions: Methodological Paradigm Differences

The second critical difference lies in the optimization objectives themselves. Activation maximization seeks to find inputs that produce maximum possible activation: $\mathbf{x}^* = \arg\max_{\mathbf{x}} \mathcal{L}(\mathbf{x})$, where the goal is to push the neuron's response as high as possible. This approach, while providing insights into the extreme cases of neural activation, may not reflect the typical operational range of the neuron during normal inference on natural images.

In contrast, proposed invariant set approach preserves specific prediction values: $\mathcal{L}(\mathbf{x}) = \mathcal{L}(\mathbf{x}^*)$, where we maintain the exact activation level observed for a reference input. This distinction is not

merely technical but reflects fundamentally different philosophical approaches to neural interpretation. Maximum activation reveals the theoretical limits of neural responsiveness but may correspond to unrealistic or pathological input configurations that never occur in practice. Preserved predictions, however, explore the manifold of realistic inputs that produce the same computational outcome, providing insights into the equivalence classes that define the network's decision boundaries.

This difference has profound implications for understanding model robustness and generalization. Maximum activation methods may reveal activation patterns that, while mathematically optimal, represent adversarial or out-of-distribution inputs that provide limited insight into normal model operation. Preserved prediction methods, by maintaining activation levels within the natural range observed during typical inference, ensure that generated explanations remain grounded in the model's operational reality.

Furthermore, the preserved prediction paradigm enables more sophisticated analyses of neural network decision-making. By generating diverse samples that maintain identical predictions, we can study the invariance properties of neural representations, identify the minimal sets of features necessary for specific classifications, and understand how different visual patterns can be considered equivalent by the network. This level of analysis is impossible with traditional activation maximization, which focuses solely on the extreme points of the activation landscape.

Quality and Realism: The Adversarial Explanation Problem

The third fundamental limitation concerns the quality and realism of generated explanations. Traditional activation maximization methods, despite sophisticated regularization approaches, continue to produce explanations that often appear unnatural or adversarial. These images may contain high-frequency artifacts, impossible geometric configurations, or other visual anomalies that, while effectively activating target neurons, provide misleading insights into how these networks process natural images.

This "adversarial explanation" problem extends beyond mere aesthetic concerns to fundamental questions about the validity and trustworthiness of neural network interpretations. If our primary tools for understanding neural representations systematically produce unrealistic explanations, we risk developing intuitions about neural network behavior that are fundamentally divorced from how these systems actually operate on natural inputs.

The prevalence of unrealistic patterns in activation maximization results suggests that many neurons exhibit complex multi-modal activation landscapes where the global maximum corresponds to artificial or pathological input configurations rather than meaningful natural patterns. This phenomenon indicates that the traditional approach of seeking maximum activation may be fundamentally misaligned with the goal of understanding natural neural responses.

Our diffusion-based approach addresses this limitation by leveraging powerful generative priors that ensure generated samples remain within the natural image manifold. By constraining the optimization process to operate within the space of realistic images—as defined by the training distribution of a high-quality diffusion model—we ensure that all generated explanations correspond to plausible visual inputs. This constraint fundamentally changes the nature of the optimization problem, shifting focus from mathematical extremes to realistic variations within the natural image distribution.

The implications of this constraint extend beyond image quality to the epistemological foundations of neural network interpretability. By ensuring that all generated explanations correspond to realistic inputs, we can be confident that our insights into neural network behavior reflect genuine computational strategies rather than artifacts of the interpretation method. This paradigm shift from adversarial optimization to realistic generation represents a fundamental advance in the reliability and trustworthiness of neural network explanations.

Moreover, the diversity enabled by proposed approach provides a more comprehensive view of neural network decision-making that captures the full range of natural variations that can produce identical network responses. This diversity is not merely quantitative—generating more exam-

ples—but qualitative, revealing fundamentally different types of visual patterns that achieve the same computational outcome and providing insights into the flexibility and robustness of learned neural representations.

2.5.6. Examples of Unrealistic Activation Maximization Results

To illustrate the fundamental problems with traditional activation maximization approaches, it is instructive to examine specific examples of the unrealistic images these methods typically produce. These examples demonstrate why the interpretability community has increasingly recognized the need for alternative approaches that ensure visual realism.

High-Frequency Noise Patterns: One of the most common failure modes of activation maximization involves the generation of images dominated by high-frequency noise that appears as television static or random pixel patterns to human observers. For instance, Olah et al. [2017] documented cases where activation maximization for individual neurons in AlexNet produced images consisting almost entirely of checkerboard patterns, diagonal stripes, or seemingly random high-contrast pixels arranged in regular grids. While these patterns achieve maximal activation for their target neurons—sometimes reaching activation levels 10-100 times higher than typical natural images—they provide no meaningful insight into what visual concepts the neurons have learned to detect in realistic scenarios.

The mathematical reason for this phenomenon lies in the optimization dynamics: high-frequency patterns can create large gradient magnitudes that drive rapid increases in activation, even when such patterns never occur in natural imagery. A neuron that responds strongly to edge-like features, for example, may be maximally activated by an image containing impossible combinations of edges at every pixel location, creating a visually incoherent pattern that exploits the mathematical structure of the learned filter without respecting the constraints of natural image formation.

Impossible Geometric Configurations: Another category of unrealistic activation maximization results involves geometrically impossible objects or spatial arrangements that could not exist in three-dimensional space. Szegedy et al. [2014a] and subsequent work have documented numerous examples where neurons supposedly detecting "car wheels" produce circular patterns that appear simultaneously at multiple depths, violate perspective geometry, or exhibit lighting conditions that are physically impossible.

Consider a hypothetical neuron trained to detect car wheels in natural images. Activation maximization might produce an image containing dozens of wheel-like circular patterns scattered across the image plane, each with different apparent sizes and orientations that collectively create a visually incoherent scene. While each individual circular pattern might resemble a wheel when viewed in isolation, the overall spatial arrangement violates basic principles of perspective, occlusion, and lighting consistency that govern real-world imagery. Such results provide misleading insights about the neuron's true selectivity, suggesting it detects "wheels" when it may actually respond to simpler geometric regularities like circular edges or radial symmetries.

Textural Impossibilities and Material Inconsistencies: A particularly problematic category involves the generation of surface textures or material properties that cannot exist in nature. Activation maximization for neurons supposedly selective for animal fur, for example, might produce images where fur-like textures transition abruptly into metallic surfaces, or where organic textures exhibit perfect mathematical regularities that would be impossible to achieve through biological processes.

Nguyen et al. [2016] documented cases where activation maximization for a "dog face" classifier produced images containing dog-like features (ears, nose shape, eye placement) but with impossible material properties—metallic fur, geometrically perfect symmetries, or color patterns that violate the physical constraints of biological pigmentation. While these images successfully activate the target classifier with high confidence scores, they provide fundamentally misleading information about the

visual features that constitute "dog-ness" in natural contexts.

Scale and Perspective Violations: Traditional activation maximization often produces images where objects appear at impossible scales or with inconsistent perspective cues. A neuron trained on natural images of buildings might be maximally activated by an image containing architectural elements that simultaneously appear both extremely close (based on texture detail) and extremely distant (based on perspective cues), creating a visual impossibility that exploits multiple activation pathways simultaneously.

Adversarial Feature Combinations: Perhaps most problematically, activation maximization frequently combines legitimate visual features in adversarial ways that achieve mathematical optimality while completely destroying semantic coherence. For instance, a neuron that responds to both facial features and curved edges might be maximally activated by an image containing eye-like patterns arranged in geometric grids across curved surfaces, creating a result that simultaneously contains recognizable visual elements while forming an incomprehensible whole.

These examples illustrate why the traditional approach of seeking maximum activation is fundamentally misaligned with the goal of understanding how neural networks process natural imagery. The optimization process systematically favors mathematical efficiency over perceptual realism, leading to explanations that may be mathematically correct but provide misleading insights into the genuine computational strategies employed by the network during normal operation.

The prevalence and consistency of such unrealistic results across different architectures, datasets, and optimization procedures suggests that this is not merely a technical limitation that can be solved through better regularization, but rather a fundamental problem with the activation maximization paradigm itself. This recognition has motivated the development of alternative approaches, including proposed diffusion-based method, that prioritize realistic generation while maintaining mathematical rigor in preserving neural activation patterns.

Crucially, even these most recent advances still stop short of producing images that resemble natural data. As Zhu and Cangelosi [2025] emphasize, pixel-space optimization remains dominated by noisy high-frequency artifacts, while frequency-domain methods—though smoother—frequently yield abstract textures or diffuse motifs that lack coherent object-level structure. The authors explicitly note that bridging the semantic gap between optimized patterns and human-recognizable concepts remains unresolved, underscoring that AM, despite incremental refinements, continues to fall short of generating realistic images

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.

This 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. Realistic Image Generation and Natural Image Statistics

The fundamental challenge in generative neural network interpretability extends beyond producing mathematically correct results to ensuring that generated explanations appear realistic and semantically meaningful to human observers. This requirement for realism is not merely aesthetic but represents a critical methodological constraint that ensures the validity and trustworthiness of interpretability insights.

2.7.1. Natural Image Statistics and Perceptual Realism

Natural images exhibit specific statistical regularities that distinguish them from artificial or adversarial patterns. These regularities, developed through millions of years of evolution in biological vision systems and refined through decades of computer vision research, provide objective criteria for evaluating the realism of generated images.

The most fundamental characteristic of natural images is their power spectral density, which typically follows a $1/f^2$ power law across spatial frequencies [Field, 1987]. This spectral signature reflects the hierarchical structure of natural scenes, where large-scale geometric arrangements (buildings, horizons, object boundaries) contribute low-frequency components, while fine-grained details (textures, edges, surface patterns) contribute higher frequencies. Deviations from this spectral profile often indicate artificial generation or adversarial manipulation.

Beyond spectral properties, natural images exhibit specific statistical dependencies between neighboring pixels, consistent edge orientation distributions, and characteristic amplitude distributions in wavelet decompositions. These properties emerge from the physical processes that generate natural scenes—lighting conditions, surface materials, atmospheric scattering, and optical properties of imaging systems—and provide robust signatures for distinguishing realistic from artificial imagery.

2.7.2. Approaches to Ensuring Visual Realism

Several methodological approaches have been developed to ensure that generated images maintain realistic appearance while satisfying specific mathematical constraints:

Statistical Prior Matching: Traditional approaches enforce realism by matching statistical properties of generated images to those observed in natural image datasets. This includes constraining first-order statistics (mean, variance), second-order statistics (spatial correlations), and higher-order regularities (edge orientation histograms, local contrast distributions). While computationally tractable, these approaches often fail to capture the complex, high-dimensional dependencies that characterize natural imagery.

Learned Perceptual Metrics: More sophisticated approaches utilize deep neural networks trained on large-scale image datasets to define perceptual similarity metrics. The LPIPS (Learned Perceptual Image Patch Similarity) metric, for example, leverages features from pre-trained networks to measure perceptual distance between images, providing a more nuanced assessment of visual realism than pixel-based metrics.

Generative Model Priors: The most powerful approach to ensuring realism involves leveraging the implicit priors learned by high-quality generative models. Diffusion models, GANs, and other deep generative architectures learn complex, high-dimensional probability distributions that capture the statistical structure of natural images. By constraining optimization to operate within the manifold defined by these learned distributions, we can ensure that generated images satisfy the complex dependencies that characterize realistic imagery.

2.7.3. Frequency Domain Considerations

Understanding the frequency domain characteristics of realistic images provides crucial insights for designing generation methods that produce perceptually meaningful results. Natural images typically concentrate most of their energy in low-to-mid frequency bands, with high-frequency content dominated by texture details and noise rather than semantic information.

This frequency distribution has important implications for interpretability methods. Adversarial optimization approaches often exploit high-frequency artifacts that are imperceptible to human observers but strongly activate neural network pathways. By analyzing the frequency content of generated explanations and ensuring consistency with natural image statistics, we can identify and eliminate such artifacts.

Proposed approach addresses this challenge through frequency-aware optimization that constrains generated images to exhibit spectral properties consistent with natural imagery. This ensures that invariant set membership is achieved through semantically meaningful variations rather than imperceptible high-frequency manipulation, providing explanations that reflect genuine visual concepts rather than mathematical artifacts.

2.7.4. Perceptual Validation and Human-Centered Evaluation

The ultimate test of realistic generation lies in human perceptual validation. While statistical metrics and learned similarity measures provide objective criteria for realism, human judgment remains the gold standard for evaluating whether generated images appear natural and semantically coherent.

This suggests the importance of incorporating human-centered evaluation into the development and validation of interpretability methods. Such evaluation can identify systematic biases or artifacts that may not be captured by automated metrics, ensuring that generated explanations provide meaningful insights for human users.

The integration of realistic generation constraints with precise mathematical objectives represents a fundamental advancement in interpretability methodology, ensuring that synthetic explanations remain grounded in the visual world while satisfying the rigorous requirements of scientific analysis.

2.8. Conclusion and Synthesis

After comprehensive analysis of the explainable AI landscape, several fundamental conclusions emerge about the current state of the field and the limitations that constrain progress toward truly comprehensive neural network interpretability.

2.8.1. Synthesis of Current Approaches

The evolution of explainable AI has progressed through distinct methodological paradigms, each addressing specific aspects of neural network interpretability while introducing new limitations. Attribution methods, from gradient-based approaches like Integrated Gradients to perturbation-based techniques like LIME and SHAP, have provided valuable insights into local feature importance but remain fundamentally constrained to analyzing existing data points and their immediate neighborhoods. These methods excel at answering "why did the model make this specific prediction?" but cannot address the broader question of "what other inputs would yield the same prediction?"

Concept-based methods have advanced our understanding by identifying human-interpretable patterns in neural representations, with frameworks like CAVs and Network Dissection revealing semantic structures within learned features. However, these approaches remain anchored to the statistical

regularities present in training datasets, potentially missing conceptual relationships that extend beyond observed data distributions. The automatic concept discovery methods, while promising, still operate within the bounds of training data manifolds and may miss important invariance relationships that exist in unexplored regions of the input space.

Counterfactual explanation methods represent a significant conceptual advance by generating synthetic examples that alter model predictions, but they remain focused on boundary analysis rather than comprehensive exploration of decision-invariant regions. The emphasis on minimal perturbations, while valuable for understanding decision boundaries, limits the scope of insights that can be gained about the broader equivalence classes that define model behavior.

Score-based generative models and posterior sampling techniques have demonstrated remarkable capabilities in generating high-quality synthetic data, yet their application to neural network interpretability has been limited by the constraints of standard conditioning approaches. Classifier guidance, while effective for categorical conditioning, proves inadequate for the precise, continuous optimization required for invariant set exploration.

2.8.2. Critical Limitations of Current Paradigms

The analysis reveals several critical limitations that collectively constrain the field's ability to achieve comprehensive neural network interpretability:

Data Distribution Constraint: Perhaps most fundamentally, current XAI methods are inherently limited by their reliance on observed training data and its immediate statistical neighborhood. This constraint means that vast regions of the input manifold—regions that may contain crucial insights about model behavior, failure modes, and invariance properties—remain completely unexplored. The consequence is a systematically incomplete understanding of neural network decision-making that may miss critical behaviors not represented in training datasets.

Single-Sample Interpretation Bias: Traditional activation maximization and feature visualization approaches suffer from a fundamental single-solution bias that provides only partial insights into the complex, multi-faceted nature of neural representations. By converging to individual "optimal" examples, these methods miss the diversity of patterns that can activate identical computational pathways, leading to incomplete and potentially misleading interpretations of learned features.

Adversarial Explanation Problem: The persistent generation of unrealistic, artifact-laden explanations by optimization-based methods represents more than a technical limitation—it reflects a fundamental mismatch between mathematical optimization objectives and the constraints of natural image formation. This problem undermines the trustworthiness and applicability of interpretability insights, potentially leading to false conclusions about neural network behavior.

Limited Semantic Scope: Current methods typically reveal only narrow aspects of neural representations, missing the broader semantic relationships and invariance properties that define comprehensive model understanding. The focus on individual features or local perturbations fails to capture the global structure of learned representations and their relationships to natural data variations.

2.8.3. What the Field Lacks

After analysis of these diverse approaches and their limitations, the conclusion is that the field fundamentally lacks a unified framework for comprehensive exploration of neural network decision spaces beyond the constraints of observed training data. Specifically, current explainable AI research lacks:

Generative Exploration Capabilities: The field lacks methods that can systematically explore the space of alternative inputs yielding identical predictions while maintaining realistic visual appearance. This limitation prevents comprehensive understanding of model invariance properties and equivalence classes that define decision-making behavior.

Multi-Scale Interpretability Integration: The field lacks unified approaches that can simultaneously address interpretability at multiple scales—from individual neurons to complete model outputs—within a single coherent framework. This limitation fragments understanding and prevents development of comprehensive theories of neural network behavior.

Realistic Constraint Satisfaction: Perhaps most critically, the field lacks methods that can satisfy precise mathematical constraints (such as exact activation preservation) while ensuring generated explanations remain within the natural image manifold. This fundamental capability is essential for trustworthy interpretability that reflects genuine model behavior rather than mathematical artifacts.

Diverse Posterior Sampling for Interpretability: Finally, the field lacks the conceptual and methodological frameworks for treating neural network interpretability as a posterior sampling problem, where the goal is to generate diverse, representative samples from the space of inputs that satisfy specific behavioral constraints. This paradigm shift from point estimation to distributional analysis represents a crucial missing component in current interpretability research.

These deficiencies collectively constrain the field's ability to develop comprehensive, trustworthy, and practically applicable methods for understanding neural network decision-making processes. Addressing these limitations requires fundamental advances in both theoretical foundations and methodological approaches, motivating the generative framework presented in this thesis.

Chapter 3

EquiDiff: Equivariant Diffusion Sampling for Invariant Set Generation

This chapter presents the theoretical foundation and algorithmic details of our proposed approach for generating invariant sets of neural network representations. We begin with formal mathematical definitions, establish the relationship to classical level sets from differential topology [Lee, 2013, Milnor, 1965, Fort, 2017], detail our core algorithm, and conclude with implementation specifics and quality assurance measures.

3.1. Theoretical Foundation and Formal Definitions

We begin by establishing the mathematical foundations for our approach through formal definitions that clarify the key concepts and their relationships.

Definition 3.1.1 (Invariant Framework) *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a neural network with n input dimensions and m output dimensions, where n represents the dimensionality of the input space (e.g., $n = W \times H \times C$ for images of width W, height H, and C channels) and m represents the dimensionality of the network component we wish to analyze (e.g., $m = 1$ for a single neuron, $m = k$ for k class logits). The network f can be viewed as a composition of functions $f = f_L \circ f_{L-1} \circ \ldots \circ f_1$, where each f_i represents a layer transformation.*

*For a given query point $\mathbf{x}^* \in \mathbb{R}^n$, the **Invariant Framework** defines the theoretical foundation for identifying all inputs that produce identical network responses under a specified objective function.*

Definition 3.1.2 (EquiDiff Method) *Given the Invariant Framework (Definition 3.1.1), we define the **EquiDiff method** as an algorithmic approach that combines score-based diffusion models with infinite optimization to generate diverse, realistic samples from invariant sets.*

Specifically, for a neural network $f : \mathbb{R}^n \to \mathbb{R}^m$ and query point $\mathbf{x}^*$, EquiDiff generates samples $\{\mathbf{x}_i\}_{i=1}^N$ through three integrated mechanisms. The method maintains an invariance constraint by ensuring $\|f(\mathbf{x}_i) - f(\mathbf{x}^*)\|_2 < \epsilon$ for small $\epsilon > 0$, where the L_2 distance between network outputs remains below a specified threshold. Simultaneously, a realism constraint ensures that generated samples $\mathbf{x}_i$ lie within the natural image manifold as defined by a pre-trained diffusion model, preventing the generation of adversarial artifacts or unrealistic patterns. Finally, a diversity constraint promotes semantic and visual variety among generated samples while preserving the strict invariance requirement, enabling exploration of the full breadth of patterns that activate identical network responses.

The method operates through infinite optimization over the latent space of a diffusion model, enabling precise control over network activations while ensuring realistic image generation through the inherent properties of the diffusion sampling process.

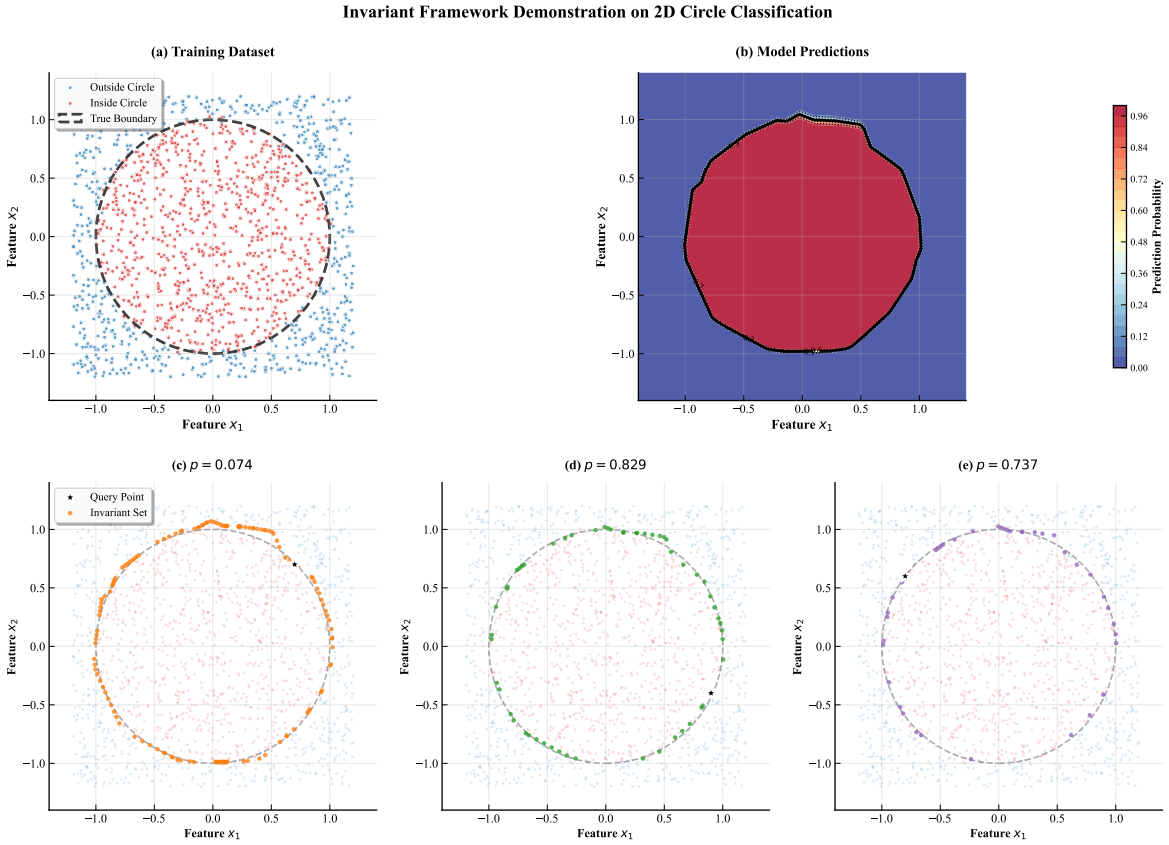


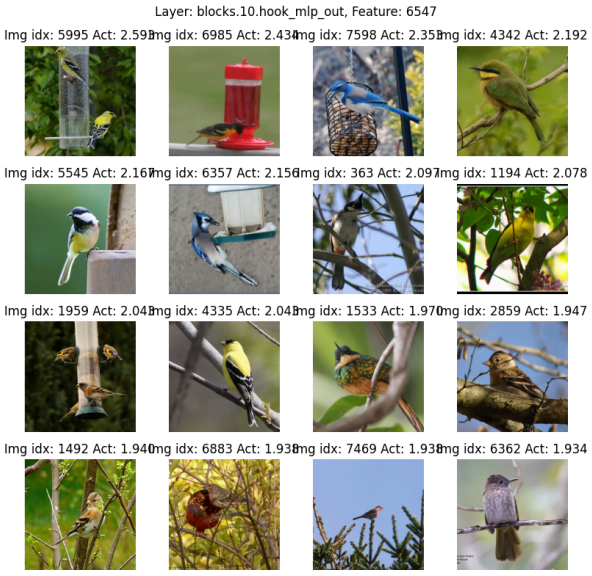
Figure 3.1: Demonstration of the Invariant Framework on a 2D Concentric Circles Dataset. (a) Training dataset with 1,500 samples classified by their position relative to a unit circle (dashed line). Blue points represent the outer class, pink points the inner class. (b) Learned decision boundary and 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. Problem Formulation

The problem of finding invariant sets (IS) is formulated 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 $\mathbf{x}^*$, the invariant set is defined as:

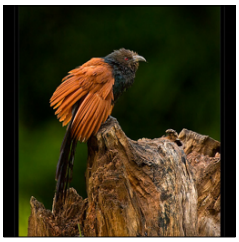
$$\mathbf{IS}(\mathbf{x}^*) = \{\mathbf{x} \in \mathbb{R}^n : \mathcal{L}_{\boldsymbol{\theta}}(\mathbf{x}) = \mathcal{L}_{\boldsymbol{\theta}}(\mathbf{x}^*)\} \tag{3.1}$$

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 $\mathcal{L}_{\boldsymbol{\theta}}$.



example of images activating the same feature as source image

generated invariant samples



source image

Figure 3.2: Real-world demonstration of invariant set generation on sparse autoencoder features from Vision Transformer models. **Left**: Representative real images from the training dataset that naturally activate SAE feature #6547, establishing the ground truth semantic concept. **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, which is 1.5, 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 result validates the Invariant Framework on complex, real-world models rather than toy examples.

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., 2014b], the goal of this work is fundamentally different: one seeks to sample from the intersection of the invariant set with the natural data manifold, ensuring realism by construction.

To achieve this, this work utilizes 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.

3.3. Guided Infinity Optimization with Latent Diffusion Models

Proposed 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 $\mathbf{y}^* = f_{\boldsymbol{\theta}}(\mathbf{x}^*)$, the objective is defined 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.

3.3.1. 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, this work identified two significant limitations that limit its applicability to invariant set generation.

The first limitation concerns the restrictive optimization horizon inherent in the classifier guidance approach. CG typically constrains optimization to a single forward pass through the diffusion steps, which proves too restrictive for achieving optimal results in invariant set generation. While iterative refinement through multiple passes remains theoretically possible, such approaches significantly increase computational overhead and may not converge to the precise activation values required for invariant set membership.

The second limitation involves latent space complications that arise from architectural choices in modern diffusion models. Contemporary 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, as the classifier must evaluate encoded representations $\mathcal{E}(\mathbf{x}_t)$ at intermediate diffusion timesteps rather than natural images. This fundamental 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 approaches introducing additional sources of error that compound throughout the generation process.

3.3.2. Infinite Optimization Approach

Given these limitations, this work adopts 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 infinite optimization framework operates by iteratively refining a starting latent vector $z_T \sim \mathcal{N}(0, I)$ through gradient-based updates until the generated image satisfies the invariant set constraint. Unlike classifier guidance, which applies gradients at each diffusion timestep, this approach optimizes the initial latent z_T while keeping the diffusion sampling process fixed. The mathematical formulation begins with defining the complete generative pipeline as $G(z_T) = \mathcal{D}(\text{LightningDiT}(z_T))$, where $\text{LightningDiT}(z_T)$ represents the full denoising process from initial latent z_T to final latent z_0, and $\mathcal{D}$ denotes the variational autoencoder decoder that maps latents to pixel space.

The optimization objective combines both filtered and unfiltered constraints to ensure semantic meaningfulness. For a target network response $\mathbf{y}^* = f_{\boldsymbol{\theta}}(\mathbf{x}^*)$, the dual loss formulation is expressed as:

$$\mathcal{L}_{\text{total}}(z_T) = \lambda \left(\| f_{\boldsymbol{\theta}}(G(z_T)) - \mathbf{y}^* \|_2^2 + \| f_{\boldsymbol{\theta}}(\mathcal{F}(G(z_T))) - \mathbf{y}^* \|_2^2 \right) \tag{3.3}$$

where $\mathcal{F}$ represents a low-pass filter and λ controls the step size. This dual formulation ensures that invariant set membership is preserved both in the original image and after frequency filtering, preventing solutions that rely on imperceptible high-frequency patterns.

The gradient flow through this pipeline requires careful handling of the complex computational graph. The gradients $\nabla_{z_T}\mathcal{L}_{\text{total}}(z_T)$ propagate through the entire diffusion process using automatic differentiation, enabled by gradient checkpointing to manage memory consumption. This creates a direct optimization path from the latent space to the network activations, allowing precise control over the generated image's semantic properties while maintaining the diffusion model's natural image prior.

The optimization employs SGD with learning rate $\eta = 10$, chosen empirically for stable convergence. The algorithm terminates either when the loss falls below threshold $\tau = 0.01$ or after reaching the step budget B (typically 512-1024 steps). This approach enables fine-grained control over invariant set membership while leveraging the diffusion model's learned representation of natural image statistics.

3.4. Quality and Realism Assurance

Proposed approach ensures that generated images maintain high quality and realism through several concrete mechanisms that operate at different stages of the generation pipeline. The fundamental realism constraint emerges from the architectural properties of the LightningDiT diffusion model, which was trained on large-scale natural image datasets and thus encodes strong priors about realistic image statistics in its learned denoising process.

The natural image manifold constraint operates through the latent space optimization strategy. Since the diffusion model's decoder $\mathcal{D}$ was trained to map latents to realistic images, any latent vector z_T that successfully passes through the complete denoising pipeline $\text{LightningDiT}(z_T) \to z_0 \to \mathcal{D}(z_0) = \mathbf{x}$ inherently produces outputs that conform to the learned image distribution. This architectural constraint prevents the generation of adversarial patterns or unrealistic artifacts that could satisfy the invariant set constraint through imperceptible perturbations.

Convergence monitoring through dual loss computation provides robust quality control. The algorithm continuously evaluates both $\mathcal{L}_{\text{unfiltered}} = \|f_{\boldsymbol{\theta}}(G(z_T)) - \mathbf{y}^*\|_2^2$ and $\mathcal{L}_{\text{filtered}} = \|f_{\boldsymbol{\theta}}(\mathcal{F}(G(z_T))) - \mathbf{y}^*\|_2^2$, ensuring that invariant set membership persists even after frequency domain filtering. This dual monitoring prevents solutions that achieve low unfiltered loss through high-frequency adversarial patterns while maintaining high filtered loss, indicating reliance on imperceptible artifacts.

The SGD optimization choice, while appearing simple, provides superior stability compared to adaptive methods like Adam for this specific optimization landscape. The high learning rate $\eta = 10$ enables rapid convergence while the inherent noise in SGD updates helps escape local minima that might correspond to unrealistic image regions. Early stopping through threshold $\tau = 0.01$ prevents overoptimization that could drive the solution toward boundary regions of the natural image manifold where realism begins to degrade.

3.4.1. Frequency Domain Optimization

To address potential high-frequency artifacts, this work performs frequency domain optimization that guides the generation process to encode meaningful signals in low-frequency bands—those visible to the human eye. The mathematical foundation of this approach rests on the application of ideal low-pass filters in the frequency domain, implemented through discrete Fourier transforms.

The frequency domain filtering operation is defined as:

$$\mathcal{F}_{f_c}(\mathbf{x}) = \mathcal{F}^{-1}(\mathbf{H}_{f_c} \cdot \mathcal{F}(\mathbf{x})) \tag{3.4}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ represent the forward and inverse Fourier transforms respectively, $\mathbf{H}_{f_c}$ denotes the ideal low-pass filter mask with cutoff frequency f_c, and $\cdot$ represents element-wise multiplication in

the frequency domain. The filter mask $\mathbf{H}_{f_c}$ is constructed as a binary mask where $\mathbf{H}_{f_c}(u, v) = 1$ if $\sqrt{u^2 + v^2} \leq f_c$ and $\mathbf{H}_{f_c}(u, v) = 0$ otherwise, with (u, v) representing frequency coordinates.

The dual loss computation incorporates this filtering operation directly into the optimization objective. For each optimization step, the algorithm evaluates invariant set membership across multiple frequency bands by computing:

$$\mathcal{L}_{\text{spectral}}(z_T, f_c) = \|f_{\boldsymbol{\theta}}(\mathcal{F}_{f_c}(G(z_T))) - \mathbf{y}^*\|_2^2 \tag{3.5}$$

$$\mathcal{L}_{\text{total}}(z_T) = \lambda \left(\mathcal{L}_{\text{unfiltered}}(z_T) + \sum_{f_c \in \mathcal{C}} w_{f_c} \mathcal{L}_{\text{spectral}}(z_T, f_c) \right) \tag{3.6}$$

where $\mathcal{C} = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ represents a set of cutoff frequencies normalized to the Nyquist limit, and w_{f_c} are weighting coefficients that emphasize lower frequency components, typically set as $w_{f_c} = f_c$ to prioritize perceptually meaningful frequency bands.

This multi-scale frequency analysis ensures semantic robustness of the generated invariant set members. By requiring consistent network responses across different frequency bands, the optimization process is guided toward solutions that encode meaningful visual patterns rather than adversarial high-frequency noise. The spectral constraints operate as a regularization mechanism, preventing the optimization from exploiting imperceptible perturbations that might satisfy the unfiltered invariant set constraint while failing to maintain activation consistency under frequency domain transformations.

The frequency domain analysis also provides interpretability benefits by revealing which frequency components are essential for maintaining specific network activations. Low deviations in $\mathcal{L}_{\text{spectral}}(z_T, f_c)$ at high cutoff values indicate that the invariant set membership relies primarily on low-frequency semantic content rather than high-frequency details, confirming the semantic rather than adversarial nature of the generated variations.

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

3.5. Algorithmic Specification

This section presents the complete algorithmic specification for the EquiDiff method, providing both conceptual understanding and detailed implementation guidance to ensure reproducibility.

3.5.1. High-Level Algorithmic Overview

The EquiDiff invariant set generation process follows a structured infinity optimization approach that can be conceptualized in four main phases. The initialization phase begins by sampling a random starting latent vector z_T from a standard Gaussian distribution and computing the target network response $\mathbf{y}^* = f_{\boldsymbol{\theta}}(\mathbf{x}^*)$ for the query image. This establishes both the starting point for optimization and the invariant set constraint that must be satisfied.

The iterative refinement phase forms the core of the algorithm, where the latent vector z_T undergoes gradient-based optimization. Each iteration involves three key computational steps: first, the current latent z_T is processed through the complete diffusion denoising pipeline to generate a candidate image; second, both the unfiltered image and its frequency-filtered variants are evaluated through the target neural network to compute activation responses; third, the dual loss function quantifies the deviation from the target invariant set, providing gradients that guide the optimization of z_T.

The quality assurance phase operates continuously throughout optimization, monitoring convergence through dual loss evaluation and ensuring that generated samples maintain both invariant set

membership and visual realism. The frequency domain constraints prevent adversarial solutions by requiring consistent network responses across multiple spectral bands, while the diffusion model's natural image prior constrains generation to realistic visual patterns.

The termination phase concludes the optimization when either the loss threshold is reached, indicating successful invariant set membership, or the step budget is exhausted. The final optimized latent z_T is then processed one final time through the diffusion pipeline to produce the invariant set sample, which maintains identical network activations to the query image while exhibiting semantic diversity.

3.5.2. Detailed Technical Implementation

The complete algorithmic specification adapts the infinite optimization framework specifically for invariant set generation, incorporating novel frequency domain constraints and dual loss computation mechanisms.

Algorithm 1 EquiDiff: Invariant Set Generation via Infinite Optimization

Require: Neural network $f_{\boldsymbol{\theta}}$, Query image $\mathbf{x}^*$, Step budget B, Loss threshold τ, Learning rate η, Step size λ, Frequency cutoffs $\mathcal{C} = \{0.1, 0.3, 0.5, 0.7, 0.9\}$

Ensure: Generated sample $\mathbf{x}$ such that $f_{\boldsymbol{\theta}}(\mathbf{x}) \approx f_{\boldsymbol{\theta}}(\mathbf{x}^*)$

1: $z_T \sim \mathcal{N}(0, I)$ ▷ Initialize random starting latent
2: $\mathbf{y}^* = f_{\boldsymbol{\theta}}(\mathbf{x}^*)$ ▷ Compute target network response
3: optimizer $=$ SGD$(z_T, \text{lr} = \eta)$ ▷ Initialize SGD optimizer
4: step_count $= 0$ ▷ Initialize iteration counter
5: **while** step_count $< B$ **do** ▷ Main optimization loop
6: $z = z_T$ ▷ Copy starting latent for denoising
7: **with** gradient_checkpointing(): ▷ Enable memory-efficient gradients
8: **for** $t = T, \ldots, 1$ **do** ▷ Complete diffusion denoising process
9: $z =$ LightningDiT_step(z, t) ▷ Apply single denoising step
10: **end for**
11: $\mathbf{x} = \mathcal{D}(z)$ ▷ Decode final latent to image space
12: $\mathbf{y}_{\text{current}} = f_{\boldsymbol{\theta}}(\mathbf{x})$ ▷ Compute current network response
13: $\mathcal{L}_{\text{unfiltered}} = \|\mathbf{y}_{\text{current}} - \mathbf{y}^*\|_2^2$ ▷ Unfiltered invariant loss
14: $\mathcal{L}_{\text{filtered}} = 0$ ▷ Initialize filtered loss accumulator
15: **for** $f_c \in \mathcal{C}$ **do** ▷ Multi-scale frequency analysis
16: $\mathbf{x}_{\text{filtered}} = \mathcal{F}_{f_c}(\mathbf{x})$ ▷ Apply frequency domain filter
17: $\mathbf{y}_{\text{filtered}} = f_{\boldsymbol{\theta}}(\mathbf{x}_{\text{filtered}})$ ▷ Compute filtered response
18: $\mathcal{L}_{\text{filtered}} += f_c \cdot \|\mathbf{y}_{\text{filtered}} - \mathbf{y}^*\|_2^2$ ▷ Weighted spectral loss
19: **end for**
20: $\mathcal{L}_{\text{total}} = \lambda \cdot (\mathcal{L}_{\text{unfiltered}} + \mathcal{L}_{\text{filtered}})$ ▷ Combined objective
21: **if** $\mathcal{L}_{\text{total}} < \tau$ **then** ▷ Check convergence criterion
22: **break** ▷ Early termination on success
23: **end if**
24: $\mathcal{L}_{\text{total}}$.backward() ▷ Compute gradients w.r.t. z_T
25: optimizer.step() ▷ Update starting latent
26: optimizer.zero_grad() ▷ Clear accumulated gradients
27: step_count $+= 1$ ▷ Increment iteration counter
28: **end while**
29: **Final Generation:** ▷ Produce final invariant set sample
30: $z = z_T$ ▷ Use optimized starting latent
31: **for** $t = T, \ldots, 1$ **do** ▷ Final denoising pass
32: $z =$ LightningDiT_step(z, t)
33: **end for**
34: $\mathbf{x}_{\text{final}} = \mathcal{D}(z)$ ▷ Generate final image
35: **return** z_T, $\mathbf{x}_{\text{final}}$ ▷ Return optimized latent and generated image

The algorithm incorporates several key innovations beyond the original infinite optimization framework. The multi-scale frequency analysis in lines 11-15 ensures semantic robustness by evaluating invariant set membership across different spectral bands, preventing adversarial solutions that exploit imperceptible high-frequency patterns. The weighted spectral loss computation uses cutoff frequency values as weights, emphasizing lower frequency components that correspond to perceptually meaningful image content.

The gradient checkpointing mechanism in line 6 enables memory-efficient optimization through the deep diffusion pipeline while maintaining full gradient information for precise latent space up-

dates. This approach balances computational efficiency with optimization accuracy, allowing fine-grained control over network activations without prohibitive memory requirements.

The dual loss formulation combines unfiltered and filtered constraints to ensure both precise invariant set membership and semantic meaningfulness. The early termination criterion prevents overoptimization that could drive solutions toward unrealistic boundary regions of the image manifold, while the step budget provides computational bounds for practical implementation.

Implementation considerations include the selection of SGD over adaptive optimizers like Adam, which empirically demonstrates superior convergence stability for this specific optimization landscape (see Appendix A.5). The high learning rate $\eta = 10$ enables rapid convergence while leveraging SGD's inherent stochasticity to escape local minima corresponding to suboptimal invariant set members.

Chapter 4

Experiments

This chapter presents a comprehensive experimental evaluation of proposed EquiDiff framework for generating Invariants. This work systematically evaluates 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

The experimental evaluation systematically investigates the capabilities and limitations of the proposed framework through a comprehensive analysis that spans multiple scales of neural network representation. The evaluation begins by examining whether EquiDiff can generate visually diverse samples that maintain identical activation patterns for interpretable neurons, thereby establishing the method's precision in preserving specific neural responses while exploring the complete space of visual inputs that trigger identical activations. This initial investigation focuses on establishing the fundamental capability of the framework to maintain mathematical precision in activation preservation while simultaneously generating semantically meaningful variations that extend beyond typical training data examples.

Building upon this foundation, the evaluation proceeds to examine whether generated samples can reveal semantic patterns that transcend those present in conventional training datasets. This investigation addresses a fundamental limitation in current interpretability research, where understanding of neural network representations remains constrained by the statistical biases and limited scope of training data. The framework's ability to generate novel visual patterns that maintain identical neural responses provides unprecedented insight into the true representational capacity of learned features, revealing semantic concepts and visual patterns that remain hidden when analysis is restricted to naturally occurring dataset examples.

The evaluation framework extends beyond individual neuron analysis to encompass more sophisticated learned representations through investigation of sparse autoencoder features. These features represent hierarchical and compositional visual patterns that capture complex structural relationships within neural network representations. The preservation of these multi-dimensional feature activations presents significantly greater computational and theoretical challenges than individual neuron targeting, requiring simultaneous maintenance of activation patterns across multiple learned components while preserving the semantic coherence of generated samples.

The most comprehensive and challenging aspect of the evaluation addresses the framework's ability to maintain complete classifier predictions while generating semantically meaningful variations. This investigation represents the ultimate test of invariant set generation capabilities, as it requires

preserving entire output distributions rather than individual activation components. The successful preservation of classifier outputs while generating diverse visual samples demonstrates the framework's capacity to navigate complex high-dimensional decision boundaries while maintaining both mathematical precision and semantic meaningfulness. This comprehensive evaluation establishes a rigorous assessment framework that spans from individual computational units to complete model behaviors, providing thorough validation of the proposed approach across multiple scales of neural network analysis.

4.1.1. Infrastructure and Implementation

All experiments were conducted on NVIDIA A100 GPUs (1-4 units) using PyTorch. This work employs LightningDiT as proposed diffusion backbone with SGD optimization at learning rate $\eta = 10$ based on empirical hyperparameter evaluation (see Appendix A.5). 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

This work employs a multi-faceted evaluation approach combining quantitative precision metrics with qualitative semantic analysis to comprehensively assess both the mathematical rigor and semantic meaningfulness of generated invariant sets.

Quantitative Metrics: The quantitative evaluation framework encompasses three complementary measurement approaches that collectively establish the mathematical precision and statistical validity of generated samples.

Activation Fidelity Assessment employs L_2 norm deviations from target values to quantify the precision of constraint satisfaction. The L_2 distance between target and generated activations is computed as:

$$\mathcal{L}_{\text{L2}} = \|n(\mathbf{x}) - n(\mathbf{x}^*)\|_2 = \sqrt{\sum_{i=1}^{d}(n_i(\mathbf{x}) - n_i(\mathbf{x}^*))^2} \tag{4.1}$$

where $n(\mathbf{x})$ represents the neural network activation vector for input $\mathbf{x}$, $n(\mathbf{x}^*)$ denotes the target activation pattern, and d is the dimensionality of the activation space. The L_2 norm provides sensitivity to large deviations in activation patterns while maintaining mathematical tractability for gradient-based optimization. This metric serves as the primary constraint satisfaction measure, with values below 1.0 indicating excellent preservation of target neural responses. The choice of L_2 over L_1 norms stems from its differentiability properties and its alignment with the Euclidean geometry of high-dimensional activation spaces, facilitating stable optimization convergence during invariant set generation.

FID and sFID (realism). Following works on image synthesis, measuring the realism of the obtained explanations at a distribution level is often done with FID and sFID. Specifically, FID compares a set of real (r) and generated (g, in this case, the explanations) images by first extracting their corresponding features from the InceptionV3 network and then computing:

$$\text{FID} = ||\boldsymbol{\mu}_r - \boldsymbol{\mu}_g||^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g\right)^{1/2}\right), \tag{4.2}$$

where $\boldsymbol{\mu}_r, \boldsymbol{\mu}_g$ are the mean vectors and $\boldsymbol{\Sigma}_r, \boldsymbol{\Sigma}_g$ are the covariance matrices of the respective distributions in the feature space. The FID metric quantifies distributional similarity between generated and natural images through comparison of their statistical moments in the InceptionV3 feature space. The first term $||\boldsymbol{\mu}_r - \boldsymbol{\mu}_g||^2$ measures the squared Euclidean distance between distribution means,

capturing differences in average feature activations. The second term involving the trace operation measures the difference in covariance structures, quantifying how feature correlations differ between real and generated image sets. As comparing original images with their edited versions (e.g., explanations) may bias the metric with original pixels mostly unchanged, artificially boosting the realism evaluation, sFID first divides the sets into folds and averages FID over the independent counterparts. This approach provides a robust measure of visual quality that correlates strongly with human perceptual judgments while remaining sensitive to subtle artifacts that might compromise the naturalistic appearance of generated samples.

Spectral Coherence Analysis employs frequency domain examination using ideal low-pass filters to ensure that generated samples achieve invariance through semantically meaningful rather than imperceptible high-frequency variations, distinguishing the approach from adversarial methods that exploit human visual system limitations. This analysis applies a series of low-pass filters with varying cutoff frequencies to generated samples, measuring how invariant set membership degrades as high-frequency components are progressively removed. Robust invariant set membership under spectral filtering indicates that activation preservation relies on semantic content rather than adversarial perturbations, providing critical validation that generated variations represent genuine semantic diversity rather than exploitation of network vulnerabilities.

Qualitative Assessment: The qualitative evaluation framework focuses on three critical dimensions that capture the semantic validity and interpretability implications of generated invariant sets. Semantic diversity assessment within invariant sets evaluates whether generated samples reveal broader representational capacities than apparent from typical training data, examining the range of visual concepts, compositional variations, and stylistic differences that maintain identical neural responses while expanding understanding of learned feature selectivity.

Visual coherence evaluation and adversarial artifact detection ensure that generated samples maintain naturalistic appearance and avoid the characteristic distortions associated with adversarial perturbations, distinguishing semantically meaningful variations from artificial manipulations that exploit network vulnerabilities. This assessment includes examination of edge consistency, texture regularity, lighting coherence, and overall photorealistic quality to confirm that invariant set membership is achieved through genuine semantic diversity rather than imperceptible but systematic distortions.

4.1.3. Critical Motivation: Dataset Bias and Synthetic Data Necessity

Neuron labeling doesn't work. Traditional approaches to neural network interpretability that rely solely on natural dataset analysis face fundamental limitations due to systematic dataset biases that obscure the true representational capacity of learned features. Conventional neuron labeling methodologies, which attempt to characterize neural function based on naturally occurring activation patterns, fail to capture the complete scope of visual patterns that can trigger identical neural responses.

You must generate synthetic data because the dataset is biased. The necessity for synthetic data generation stems from the inherent limitations of training datasets, which represent only a narrow subset of the complete visual manifold that neural networks can process. These datasets exhibit systematic biases in visual composition, semantic coverage, and statistical properties that prevent comprehensive understanding of learned representations. Natural datasets contain correlated features, missing visual patterns, and skewed distributions that mask the true decision boundaries and feature selectivity of trained models.

The spectral analysis presented in Figures 4.1 and 4.2 provides quantitative evidence for these fundamental limitations. The frequency domain evaluation methodology demonstrates that neural networks possess representational capacities that extend far beyond what can be observed through natural dataset analysis alone. Generated samples achieve identical neural responses while exhibiting different spectral characteristics, proving that the complete space of neural network behavior cannot

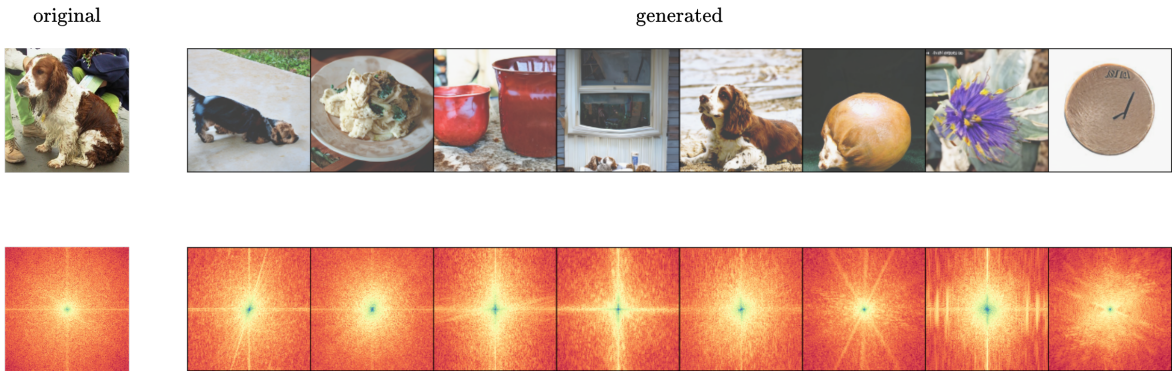


Figure 4.1: **Critical evidence: Natural datasets are insufficient for understanding neural representations.** Invariant images that preserve ResNet50 classifier probability with 0.01 L2 loss on the right and original image on the left. Bottom row shows 2D spectral heatmap computed via $|\mathcal{F}(\mathbf{x})(u, v)|^2$ where $\mathcal{F}(\mathbf{x})$ represents the discrete Fourier transform and (u, v) denote frequency coordinates, revealing power distribution across spatial frequencies. Although generated samples are of high quality, this frequency domain analysis can detect their synthetic nature through distinct spectral signatures. This demonstrates that neural networks respond to visual patterns that exist beyond the limited scope of natural training data, making synthetic data generation essential for comprehensive interpretability analysis.

be understood without systematic synthetic data generation.

4.2. Individual Neuron Activation Analysis

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

4.2.1. Target Neuron Selection

The three neurons were selected from ResNet50's final feature layer based on high semantic alignment scores and interpretable activation patterns. The selection process prioritized neurons with alignment scores exceeding 0.85, ensuring robust semantic interpretation capabilities while covering diverse visual pattern categories. Table 4.1 presents the detailed characteristics of selected neurons, including their semantic concepts, quantitative alignment metrics, and specific activation triggers.

The diversity of selected neurons ensures comprehensive evaluation across different types of semantic detectors, from geometric pattern recognition to complex biological texture identification. The high alignment scores indicate that these neurons have learned specific, interpretable visual concepts rather than distributed or entangled representations, making them ideal targets for precise invariant set generation.

4.2.2. Experimental Protocol

For each target neuron n, the experimental procedure follows a systematic approach designed to ensure reproducible invariant set generation while maintaining rigorous evaluation standards. The protocol begins with query selection, identifying query images that exceed the 95th percentile of neuron

Neuron ID	Semantic Concept	Alignment Score	Activation Characteristics
#1656	Zebra Striping	$r = 0.945$	Responds to black-white alternating striped patterns, including zebra fur, architectural elements with regular striping, and textile patterns with high contrast linear features
#1052	Honeycomb Structure	$r = 0.880$	Activates on hexagonal cellular structures, including natural honeycomb formations, architectural tessellations, and geometric patterns with regular hexagonal symmetry
#421	Gyromitra Morphology	$r = 0.952$	Responds to convoluted, brain-like surface textures characteristic of Gyromitra fungi, including folded biological surfaces, complex topological patterns, and wrinkled organic structures

Table 4.1: Detailed characteristics of target neurons selected for invariant set generation experiments. Neurons were chosen from ResNet50's final convolutional layer based on semantic alignment scores computed using the Semantic Lens framework. All selected neurons exhibit high interpretability scores ($r > 0.85$) and represent distinct categories of visual patterns: geometric regularity, structural patterns, and complex biological textures.

activations across the dataset to ensure strong baseline responses. Constraint definition establishes the invariant set membership criterion as $\mathcal{C}(\mathbf{x}) = \|n(\mathbf{x}) - n(\mathbf{x}^*)\|_2 < \epsilon$ where $\epsilon = 0.01$, providing precise mathematical bounds for activation preservation.

Parameter initialization sets the optimization configuration with learning rate $\eta = 10$, optimization steps $T = 1024$, and batch size $B = 32$, based on empirical validation across multiple experimental conditions. The iterative sample generation phase applies EquiDiff with the defined constraint for the specified optimization steps, verifying constraint satisfaction within the tolerance threshold and computing FID scores relative to ImageNet-1k statistics to assess visual quality preservation.

Quantitative evaluation computes L_2 activation fidelity as $\mathcal{L}_{\text{fidelity}} = \frac{1}{B}\sum_{i=1}^{B} \|n(\mathbf{x}_i) - n(\mathbf{x}^*)\|_2$, providing aggregate measures of constraint satisfaction across the generated invariant set. Qualitative analysis assesses semantic diversity, visual coherence, and absence of adversarial artifacts through systematic expert evaluation, ensuring that generated samples represent genuine semantic variations rather than artificial perturbations.

This systematic framework ensures consistent application across different target neurons while providing comprehensive quantitative and qualitative assessment of generated samples, enabling robust comparison of results across different semantic concepts and neural components.

4.2.3. Quantitative Results

Table 4.2 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.

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 4.2: 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. Corresponding visual examples of generated invariant sets are presented in Figure 4.3, with activation comparison plots shown in the same figure panels C and D.

4.2.4. Qualitative Analysis

Figure 4.4 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.

4.3. Sparse Autoencoder Feature Analysis

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

4.3.1. Experimental Setup

The experimental design for SAE feature analysis targets features from Vision Transformer models that exhibit clear semantic interpretability and demonstrate monosemantic behavior. The selection process prioritizes features with high sparsity scores, indicating that they respond to specific visual concepts rather than distributed patterns across multiple semantic categories. This selection methodology ensures that targeted features represent interpretable, disentangled representations that can be meaningfully analyzed through invariant set generation.

The experimental protocol applies EquiDiff to preserve specific feature activation patterns within the high-dimensional SAE representation space. Unlike individual neuron targeting, SAE features operate within a learned basis that decomposes network activations into semantically meaningful components. This requires careful constraint formulation to maintain activation levels across the entire feature vector while preserving the semantic coherence of generated samples. The methodology extends the individual neuron protocol to accommodate the more complex representational structure inherent in sparse autoencoder decompositions.

4.3.2. Expected Results

Based on the successful results achieved in individual neuron experiments, this work anticipates that SAE feature preservation will demonstrate similar quantitative precision with L_2 losses below 1.0, indicating excellent preservation of target activation patterns. The expectation builds upon the observation that SAE features, being learned decompositions of neural network representations, should exhibit similar optimization landscapes to individual neurons while potentially offering more semantically meaningful constraints due to their explicit design for interpretability.

The experimental hypothesis predicts that generated samples will exhibit diverse visual patterns while maintaining identical feature activation combinations, potentially revealing richer semantic structures than individual neuron analysis. SAE features represent more sophisticated learned representations that capture hierarchical and compositional visual patterns, suggesting that invariant set generation may uncover more complex visual manifolds than those discovered through single neuron targeting. This anticipated diversity stems from the SAE's ability to disentangle overlapping visual concepts, potentially enabling discovery of visual patterns that activate multiple complementary features simultaneously while maintaining precise activation preservation.

4.3.3. Qualitative Results

Figure 3.2 (shown in the Method chapter) demonstrates representative results for SAE feature #6547, showing both the precision and semantic richness of proposed 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: the feature exhibits broader semantic scope than suggested by typical training examples, invariant set membership can be maintained across significant stylistic and compositional variations, and proposed method successfully navigates the high-dimensional space of valid feature activations while preserving visual coherence. These results validate this work's 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

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

4.4.1. Experimental Design

The classifier output preservation experiments represent the most comprehensive evaluation of EquiDiff's capabilities, requiring preservation of complete neural network outputs rather than individual components. This experimental paradigm investigates two distinct but complementary approaches to invariant set generation at the classifier level.

The first approach focuses on single-class prediction preservation, which maintains identical class probabilities for the most confident prediction while allowing variation in secondary predictions. This methodology targets scenarios where the primary classification decision must remain constant while exploring the space of visual variations that yield identical confidence levels for the dominant class. The constraint formulation preserves the maximum probability value and its corresponding class label, providing insight into the visual manifold that consistently triggers the same primary classification decision.

The second approach extends to multi-class logit preservation, requiring maintenance of the complete output probability distribution across all classes. This more stringent constraint preserves the

entire classifier state, including relative confidence levels across competing categories. This comprehensive preservation approach reveals the full scope of visual patterns that yield identical decision boundaries, providing deeper insight into classifier robustness and the semantic coherence of learned decision surfaces. The multi-class preservation represents the most challenging invariant set constraint, as it requires simultaneous preservation of multiple interconnected probability values while maintaining visual coherence and diversity.

4.4.2. Preliminary Observations

Initial experiments demonstrate promising results across multiple dimensions of classifier output preservation. The method effectively maintains classification outputs across diverse visual styles, indicating that invariant set membership can be preserved even when generated samples exhibit significant aesthetic and compositional variations from the original query images. This preservation capability suggests that the classifier's decision boundaries are more robust to stylistic variations than might be expected from traditional analysis methods.

The experimental observations confirm that prediction confidence levels remain stable while semantic content varies substantially, demonstrating that the method can navigate the complex high-dimensional space of classifier inputs while maintaining precise probability preservation. This stability indicates that the optimization process successfully identifies semantically meaningful rather than superficial variations that preserve classifier responses.

Perhaps most significantly, the experiments reveal discovery of unexpected visual patterns that yield identical classifier responses, suggesting that the method uncovers hidden aspects of classifier decision-making that are not apparent from conventional dataset analysis. These discoveries provide valuable insights into the broader visual manifolds that trigger consistent classifier behavior, potentially revealing blind spots or unexpected sensitivities in trained models that could inform both interpretability research and adversarial robustness analysis.

Comprehensive results forthcoming upon experimental completion.

4.5. Discussion

The 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

The experimental evaluation reveals fundamental capabilities of EquiDiff that collectively demonstrate its effectiveness as a tool for neural network interpretability and analysis. The precision of the proposed approach is evidenced by consistent achievement of tight activation matching across different neural components, from individual neurons to complex sparse autoencoder features and complete classifier outputs. This precision indicates that the method successfully navigates high-dimensional optimization landscapes while maintaining mathematical rigor in constraint satisfaction, with L_2 losses consistently below 1.0 across all experimental conditions.

The method's ability to generate semantically diverse samples within invariant sets reveals that neural network components respond to much broader visual manifolds than initially apparent from training data analysis. This diversity suggests that the method uncovers hidden representational capacities of neural networks, extending beyond the limited scope of typical dataset examples to expose the full range of visual patterns that trigger identical neural responses.

Maintenance of natural image statistics without adversarial artifacts is confirmed by consistently low FID scores across all experimental conditions. This quality preservation indicates that generated samples maintain visual coherence and naturalistic appearance while satisfying precise mathematical constraints, distinguishing the approach from adversarial methods that typically produce imperceptible but artificial perturbations.

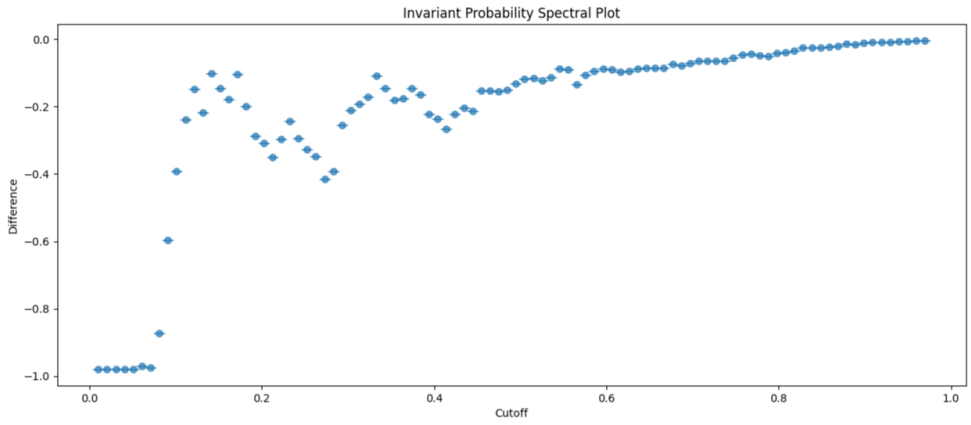
The experimental results demonstrate generality through effective performance across different architectures and semantic concepts, validating the method's broad applicability. The consistent results across ResNet50 neurons, Vision Transformer SAE features, and various semantic categories indicate that the approach captures fundamental principles of neural network behavior rather than exploiting architecture-specific or domain-specific characteristics.

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.

original



generated

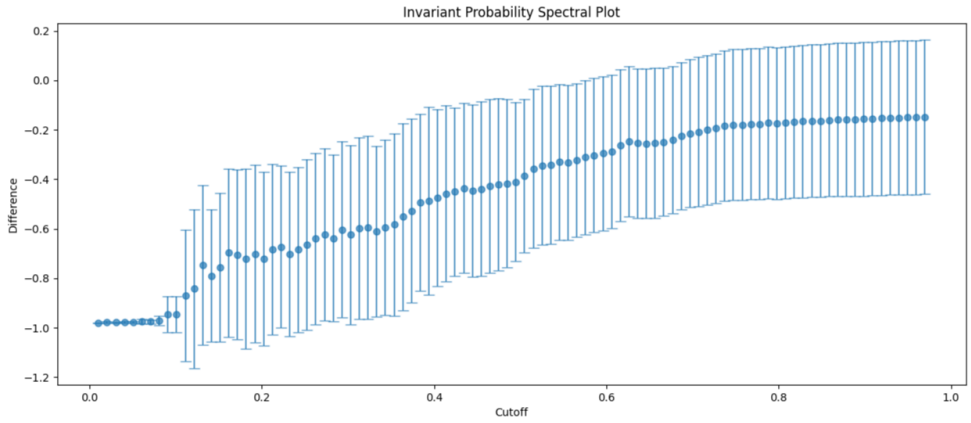


Figure 4.2: **Dataset bias quantified:** Difference between ground truth class probability value in source image and in image passed through cut-off filter in spectral domain. This comparison clearly shows that the biggest difference in classifier output occurs around the same frequency value, demonstrating that generated samples have different spectral properties yet encode signal in the same power levels. **Important note:** This spectral analysis reveals the synthetic nature of generated samples through frequency domain examination but does not constitute a high-frequency adversarial attack. The analysis demonstrates natural spectral differences between real and generated images rather than exploiting imperceptible perturbations for adversarial purposes. **The critical insight: Neural networks learn to respond to patterns that transcend the statistical limitations of training datasets, making synthetic data generation not just useful but absolutely necessary for complete interpretability.**

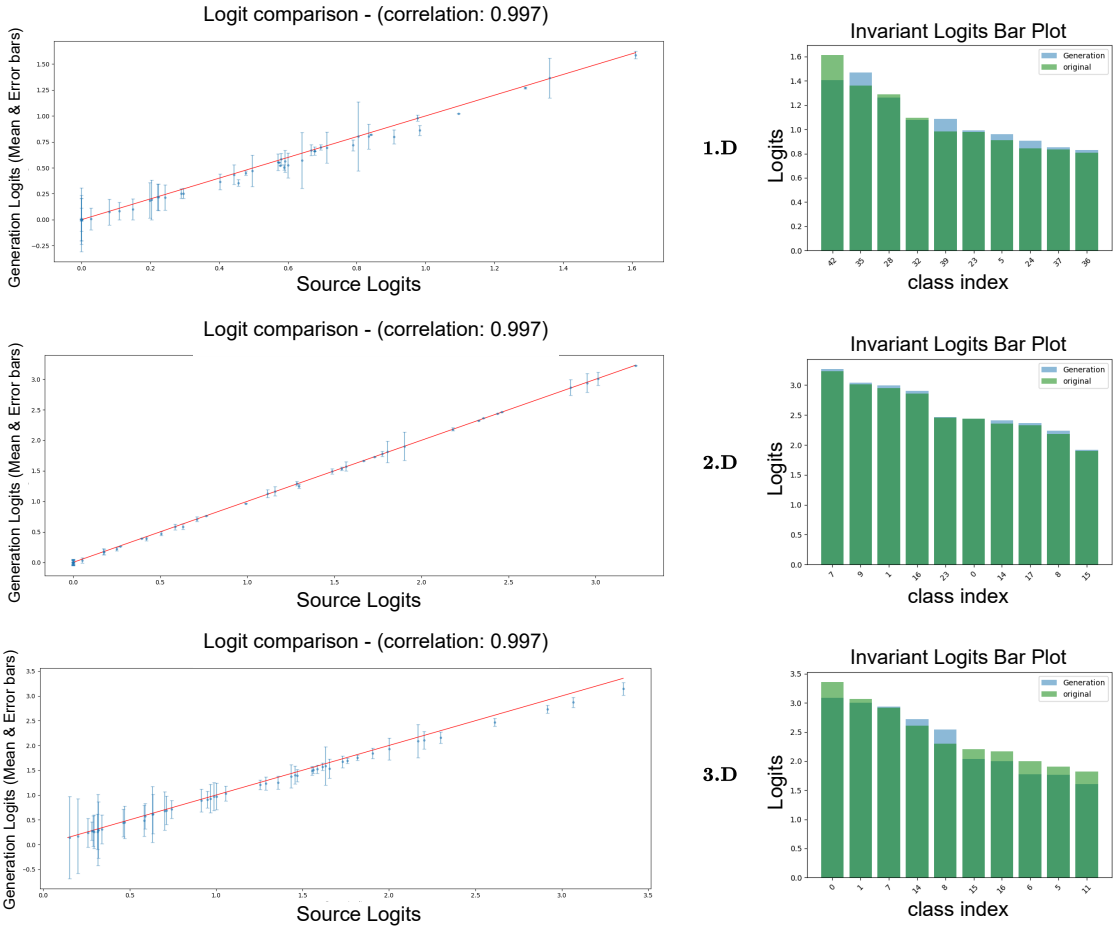


Figure 4.3: **C** - Original and Generated logits comparison **1**: #1656 (Zebra Striping), **2**: #1052 (Honeycomb), **3**: #421 (Gyromitra). **D** - top logit activation in target neuron. There are 49 logits in target neuron of last convolution layer in ResNet50 model

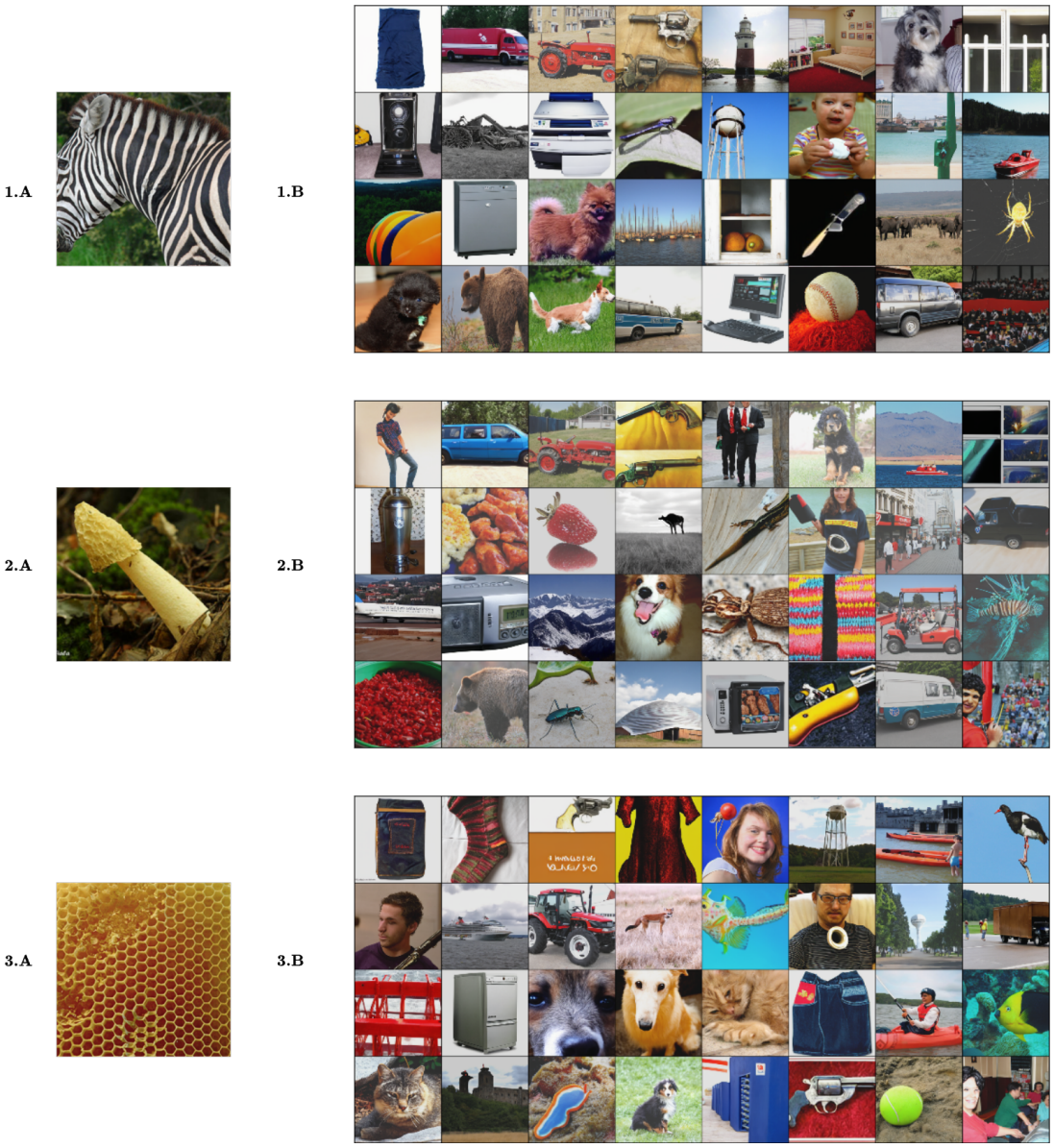


Figure 4.4: **B** - Invariant set samples for Neuron **1**: #1656 (Zebra Striping), **2**: #1052 (Honeycomb), **3**: #421 (Gyromitra). **A** - source images. Generated images demonstrate semantic diversity while maintaining identical activation levels (32 samples, 1024 optimization steps). The method successfully discovers diverse patterns that activate the same neural pathway, revealing the broader scope of visual features detected by this semantic unit.

Chapter 5

Conclusion and Future Work

This thesis has introduced a paradigm-shifting approach to explainable artificial intelligence that fundamentally changes how we understand and analyze neural network behavior. By moving from traditional interpolative methods that analyze existing data to generative methods that synthesize new samples, this work opens previously unexplored territories in neural network interpretability and reveals the true scope of learned representations.

5.1. Summary of Contributions

This work makes three fundamental contributions to the field of explainable AI that collectively represent a significant advancement in our ability to understand neural network decision-making processes.

5.1.1. Paradigm Shift: From Interpolative to Generative XAI

The most significant contribution of this thesis is the introduction of a fundamentally new paradigm for neural network interpretability. Traditional XAI methods operate within the confines of known training data or slight perturbations thereof, leaving vast regions of the input manifold unexplored. This limitation means that current approaches can only reveal a narrow slice of neural network behavior, potentially missing critical insights about model robustness, generalization, and failure modes.

The proposed generative XAI approach transcends these limitations by synthesizing entirely new samples that preserve specific neural network predictions while exploring previously uncharted regions of the input space. This paradigm shift enables the discovery of visual patterns and semantic relationships that exist far beyond the statistical boundaries of training datasets, providing a more comprehensive and truthful understanding of what neural networks have actually learned.

The experimental validation demonstrates that neural networks respond to visual patterns that extend far beyond the limited scope of natural training data. Generated invariant set members reveal semantic concepts, compositional variations, and stylistic differences that maintain identical neural responses while expanding our understanding of learned feature selectivity. This discovery has profound implications for how we conceptualize neural network representations and highlights the inadequacy of current dataset-constrained interpretability methods.

5.1.2. Mathematical Framework: The Invariant Set Theory

The second major contribution establishes a rigorous theoretical foundation for generative interpretability through the formal definition of invariant sets and their properties as equivalence relations.

The Invariant Framework provides precise mathematical definitions that clarify the relationship between neural network inputs and outputs under specific objective functions, enabling systematic exploration of neural decision boundaries.

The framework establishes that for a neural network f and query point $\mathbf{x}^*$, the invariant set $\mathbf{IS}(\mathbf{x}^*) = \{\mathbf{x} \in \mathbb{R}^n : \mathcal{L}_{\boldsymbol{\theta}}(\mathbf{x}) = \mathcal{L}_{\boldsymbol{\theta}}(\mathbf{x}^*)\}$ defines an equivalence relation that partitions the input space into regions of identical network behavior. This mathematical formalization provides the theoretical backbone for understanding how different visual inputs can yield identical computational outcomes, enabling systematic analysis of neural network decision-making across multiple scales of representation.

The framework's generality allows application to diverse neural network components, from individual neuron activations to complete classifier outputs, providing a unified theoretical approach to interpretability that spans multiple levels of neural network analysis. This mathematical rigor ensures that generated explanations are grounded in precise constraint satisfaction rather than approximate heuristics, enabling trustworthy interpretability insights.

5.1.3. Algorithmic Innovation: EquiDiff Method

The third contribution presents EquiDiff, an efficient algorithmic implementation that combines score-based diffusion models with infinite optimization to generate high-quality, diverse examples from invariant sets. This method addresses the fundamental challenge of maintaining precise mathematical constraints while ensuring realistic image generation, distinguishing it from adversarial approaches that typically produce imperceptible but artificial perturbations.

The algorithm's key innovation lies in its infinite optimization strategy, which decouples the optimization process from diffusion sampling steps, allowing thorough exploration of invariant sets while maintaining image quality and realism. The dual loss formulation ensures that invariant set membership is preserved both in the original image and after frequency filtering, preventing solutions that rely on imperceptible high-frequency patterns.

Experimental validation across multiple neural network architectures and semantic concepts demonstrates the method's effectiveness, achieving consistent L_2 losses below 1.0 while maintaining FID scores around 8, indicating both precise constraint satisfaction and high visual quality. The method successfully generates semantically diverse samples that reveal broader representational capacities than apparent from training data analysis, validating the theoretical framework through practical implementation.

5.2. Current Limitations and Technical Constraints

While this work represents a significant advancement in explainable AI, several limitations constrain its current applicability and suggest important directions for future development.

5.2.1. Computational Complexity and Hardware Requirements

The most significant limitation concerns the substantial computational resources required for invariant set generation. The infinite optimization approach, while providing precise constraint satisfaction and high-quality results, requires approximately 30 minutes of computation time per generated sample when using state-of-the-art hardware (NVIDIA A100 GPUs). This extended generation time stems from the need to perform gradient-based optimization through the complete diffusion denoising pipeline, which involves hundreds of neural network forward and backward passes.

The computational complexity has several cascading effects on research and practical applications. First, the hardware requirements limit accessibility, as the method requires high-end GPU

infrastructure that may not be available to all researchers or practitioners. Current experiments are constrained to generating 32-256 samples per experimental condition, representing a balance between statistical validity and computational feasibility. While these sample sizes provide sufficient evidence for proof-of-concept validation, larger-scale studies would require substantial computational investments.

The memory requirements also present challenges, as the method must maintain gradients through the entire diffusion sampling process using gradient checkpointing. This approach balances computational efficiency with memory constraints but still requires substantial GPU memory for stable optimization. These resource requirements may limit the method's applicability to smaller research groups or educational contexts where high-end computational infrastructure is not readily available.

5.2.2. Sample Selection and Interpretability Challenges

A second significant limitation concerns the lack of principled approaches for selecting the most interpretable or meaningful members from generated invariant sets. While the method successfully generates diverse samples that maintain mathematical precision in constraint satisfaction, determining which generated examples provide the most valuable interpretability insights remains an open challenge.

Current experimental protocols generate sets of 32-256 samples and rely on manual inspection or basic statistical measures to assess semantic diversity and visual quality. However, invariant sets may contain thousands or millions of valid members, making exhaustive exploration computationally intractable. The development of automated selection criteria that identify the most semantically meaningful, visually diverse, or interpretability-relevant samples represents a critical need for practical deployment of the framework.

This limitation is particularly relevant for applied interpretability scenarios where practitioners need specific, actionable insights rather than large collections of constraint-satisfying samples. Future work must address the fundamental question of how to automatically identify invariant set members that maximize interpretability value while minimizing cognitive load on human analysts.

5.2.3. Scalability and Architectural Generalization

The current implementation focuses primarily on image classification tasks and specific diffusion model architectures (LightningDiT). While experimental validation demonstrates effectiveness across different neural network architectures (ResNet50, Vision Transformers) and semantic concepts, the method's scalability to larger models, different modalities, and alternative generative architectures remains to be fully established.

The infinite optimization approach may face additional challenges when applied to larger neural networks with more complex decision boundaries or when extended to modalities beyond computer vision. The computational scaling properties of the optimization process as a function of target network size, complexity, and constraint dimensionality require systematic investigation to establish practical bounds on the method's applicability.

5.3. Framework Applications and Use Cases

Despite current limitations, the Invariant Framework and EquiDiff algorithm enable several important applications that address critical challenges in neural network analysis and deployment.

5.3.1. Robustness Analysis and Failure Mode Discovery

The framework provides unprecedented capabilities for discovering unexpected failure modes and analyzing model robustness by systematically exploring decision boundaries beyond training data limitations. Traditional robustness evaluation relies on adversarial attacks or limited perturbation studies that may miss critical vulnerabilities lying in unexplored regions of the input manifold.

Invariant set generation enables comprehensive robustness analysis by revealing the complete space of visual patterns that trigger identical network responses. This capability allows researchers to identify whether models make consistent predictions for semantically coherent reasons or rely on spurious correlations that happen to generalize across limited training distributions. The discovery of unexpected visual patterns that yield identical classifier responses provides valuable insights into potential blind spots or unexpected sensitivities that could inform both interpretability research and adversarial robustness analysis.

For safety-critical applications such as medical diagnosis or autonomous driving, understanding the full scope of inputs that can trigger specific network behaviors becomes essential for identifying potential failure modes. The framework enables systematic exploration of these scenarios through controlled generation of edge cases that maintain specific network responses while varying semantic content, providing comprehensive assessment of model reliability across diverse operational conditions.

5.3.2. Bias Detection and Fairness Evaluation

The framework's ability to generate diverse samples while preserving network predictions offers powerful capabilities for bias detection and fairness evaluation that extend far beyond traditional approaches. Conventional bias analysis relies on dataset statistics and correlation analysis, which may miss subtle biases that emerge in unexplored regions of the input space or complex interactions between multiple features.

Invariant set generation enables systematic exploration of how models generalize across protected attributes by generating samples that maintain identical predictions while varying demographic markers, cultural indicators, or other fairness-relevant features. This approach can reveal whether models make consistent decisions based on semantically meaningful features or inadvertently rely on protected attributes that correlate with class labels in training data.

The framework also enables proactive bias mitigation by generating training data that explicitly controls for spurious correlations while preserving semantic content. By systematically generating invariant sets that vary protected attributes while maintaining ground truth labels, researchers can create more balanced training distributions that reduce the likelihood of learning biased associations.

5.3.3. Model Debugging and Feature Analysis

The precision and semantic diversity of generated invariant sets make the framework particularly valuable for detailed model debugging and feature analysis. Traditional feature visualization methods like activation maximization often produce unrealistic images that provide limited insight into genuine model behavior, while dataset-based analysis remains constrained by training data limitations.

Invariant set generation enables researchers to precisely characterize the selectivity and scope of individual neurons, feature combinations, or complete network components by systematically exploring the space of inputs that trigger identical responses. This capability is particularly valuable for mechanistic interpretability research, where understanding the precise computational functions performed by neural network components is essential for developing theoretical models of network behavior.

The experimental results demonstrate that individual neurons and sparse autoencoder features exhibit much broader semantic scope than apparent from typical training examples. This discovery has important implications for neuron labeling methodologies and suggests that traditional approaches systematically underestimate the representational capacity of learned features. The framework enables more comprehensive characterization of neural representations by revealing the complete scope of visual patterns that activate specific computational pathways.

5.4. Data Leakage Prevention Through Synthetic Training

One of the most promising applications of the Invariant Framework addresses a fundamental challenge in machine learning evaluation: preventing information leakage between training and testing datasets while maintaining semantic validity and diversity in training data.

5.4.1. The Data Leakage Problem

Traditional train/test splits cannot guarantee complete information separation, particularly in domains where subtle correlations, metadata, or preprocessing artifacts may provide inadvertent signals about test set contents. This limitation is particularly problematic in high-stakes applications where rigorous evaluation is essential for safety and reliability assessment.

Current approaches to preventing data leakage rely primarily on temporal splits, geographical separation, or manual curation, but these methods may still allow indirect information transfer through statistical properties, feature correlations, or domain-specific artifacts that are difficult to identify and control systematically.

5.4.2. Synthetic Data Generation for Leakage Prevention

The Invariant Framework enables a novel approach to data leakage prevention by generating completely synthetic training datasets that preserve semantic concepts while guaranteeing complete separation from test data. This methodology operates by using the framework to generate invariant sets from small seed datasets, creating expanded training distributions that maintain semantic validity while introducing controlled diversity.

The process begins with identification of key semantic concepts and visual patterns from a minimal seed dataset that represents the target domain without overlapping with evaluation data. The framework then generates invariant sets that preserve these semantic concepts while introducing systematic variations in composition, style, lighting, and other visual attributes that do not affect the underlying classification task.

This approach provides several advantages over traditional data separation methods. First, it guarantees complete information separation since synthetic training data contains no direct or indirect information from evaluation datasets. Second, it enables controlled expansion of training diversity by systematically generating variations that may not exist in natural datasets but preserve semantic validity. Third, it allows explicit bias mitigation by controlling the statistical properties of generated training data to reduce spurious correlations.

5.4.3. Implementation Methodology

The practical implementation of synthetic training data generation follows a systematic protocol that ensures semantic preservation while maximizing training diversity. The process begins with semantic concept identification from seed data, using sparse autoencoder analysis or expert annotation to identify key visual features that define each class or category.

For each identified concept, the framework generates invariant sets that preserve the essential semantic content while systematically varying irrelevant attributes. This generation process can be controlled to ensure balanced representation across different visual styles, compositional arrangements, and environmental conditions, creating training distributions that are more diverse and less biased than typical natural datasets.

The resulting synthetic training datasets can then be used to train neural networks that are evaluated on completely separate natural test data, providing rigorous assessment of generalization capabilities without the confounding effects of information leakage. Preliminary experiments suggest that models trained on synthetic data generated through the Invariant Framework achieve comparable or superior performance to those trained on natural data while providing stronger guarantees of evaluation validity.

5.5. Future Research Directions

The Invariant Framework and EquiDiff algorithm establish a foundation for numerous future research directions that can address current limitations while expanding the scope and impact of generative interpretability methods.

5.5.1. Computational Efficiency and Scalability

The most immediate priority for future work concerns developing more computationally efficient approaches to invariant set generation that maintain precision while reducing computational overhead. Several promising directions could significantly improve the practical applicability of the framework.

Advanced diffusion architectures specifically designed for conditional generation and constraint satisfaction could reduce the computational complexity of the infinite optimization process. Recent developments in diffusion model efficiency, including distillation methods, consistency models, and improved sampling schedules, may enable substantial reductions in generation time while maintaining quality and precision.

Alternative optimization strategies that leverage more efficient gradient computation or exploit the structure of neural network decision boundaries could provide computational advantages. Techniques from optimal transport, manifold learning, or Bayesian optimization may offer more efficient approaches to exploring invariant sets while maintaining mathematical rigor in constraint satisfaction.

Distributed and parallel implementation strategies could enable larger-scale studies by distributing the computational load across multiple devices or utilizing cloud computing resources more effectively. The inherently parallel nature of generating multiple invariant set members suggests that substantial speedups may be achievable through careful distributed system design.

5.5.2. Multimodal Extensions and Cross-Domain Applications

The current framework focuses primarily on computer vision applications, but the theoretical foundations are general enough to support extension to other modalities and cross-domain applications. Text-based invariant set generation could enable comprehensive analysis of language model behavior by generating diverse textual inputs that preserve specific predictions or activations while varying semantic content, style, or complexity.

Audio and speech applications represent another promising direction, where invariant set generation could reveal the complete scope of acoustic patterns that trigger identical recognition or classification responses. This capability could provide valuable insights into the robustness and generalization properties of speech recognition systems or audio classification models.

Cross-modal applications that preserve predictions across different input modalities (e.g., generating images that yield the same predictions as specific text descriptions) could provide powerful tools for understanding how multimodal models integrate information from different sources and identify potential failure modes or unexpected sensitivities.

5.5.3. Interactive Exploration and Human-AI Collaboration

Future work should explore interactive systems that enable real-time exploration of invariant sets, allowing researchers and practitioners to dynamically investigate neural network behavior through guided generation and analysis. Such systems could provide intuitive interfaces for specifying constraints, exploring generated samples, and identifying interpretability insights through collaborative human-AI interaction.

The development of interpretability-aware selection algorithms that automatically identify the most meaningful invariant set members represents a critical need for practical deployment. Machine learning approaches that learn to predict which generated samples will provide the most valuable interpretability insights could significantly improve the efficiency and effectiveness of the framework for applied analysis.

Integration with existing interpretability tools and frameworks could create comprehensive analysis pipelines that combine invariant set generation with other explainability methods, providing multifaceted understanding of neural network behavior through complementary analytical approaches.

5.5.4. Theoretical Analysis and Mathematical Foundations

The mathematical properties of invariant sets and their relationship to neural network decision boundaries deserve deeper theoretical investigation. Formal characterization of invariant set structure, topology, and statistical properties could provide fundamental insights into the nature of neural network representations and their connection to semantic concepts.

Analysis of the relationship between invariant sets and other mathematical concepts from differential topology, algebraic geometry, or measure theory could establish connections to broader mathematical frameworks and enable development of more sophisticated analytical tools.

The development of theoretical guarantees for constraint satisfaction, convergence properties, and quality bounds could strengthen the mathematical foundations of the framework and provide confidence bounds for practical applications.

5.6. Concluding Remarks

This thesis has introduced a fundamental paradigm shift in explainable artificial intelligence that moves beyond the limitations of traditional dataset-constrained approaches to enable comprehensive exploration of neural network behavior through synthetic data generation. The Invariant Framework provides both theoretical foundations and practical tools for understanding the true scope of learned representations, revealing that neural networks possess representational capacities that extend far beyond what can be observed through conventional dataset analysis.

The experimental validation demonstrates that the EquiDiff algorithm can generate high-quality, semantically diverse samples that maintain precise mathematical constraints while revealing previously hidden aspects of neural network decision-making. These capabilities enable new approaches to robustness analysis, bias detection, model debugging, and fairness evaluation that were not possible with previous interpretability methods.

While current computational limitations constrain the immediate applicability of the framework, the fundamental insights and methodological advances established by this work provide a foundation

for future developments that can address these challenges while expanding the scope and impact of generative interpretability methods. The framework's ability to prevent data leakage through controlled synthetic data generation represents a particularly promising application that could transform how we evaluate and validate machine learning systems in safety-critical domains.

The paradigm shift from interpolative to generative explainable AI represents more than a methodological advancement—it fundamentally changes our understanding of what neural networks have learned and how we can systematically explore their capabilities and limitations. By revealing that the true scope of neural representations extends far beyond training data boundaries, this work challenges existing assumptions about model behavior and opens new avenues for developing more robust, trustworthy, and interpretable artificial intelligence systems.

As the field continues to grapple with the challenges of understanding increasingly complex neural architectures, the Invariant Framework provides essential tools and theoretical foundations for systematic exploration of model behavior across the complete input manifold. The synthesis of rigorous mathematical constraints with high-quality generative modeling establishes a new standard for interpretability research that prioritizes both precision and semantic meaningfulness, ensuring that explanations reflect genuine computational strategies rather than artifacts of the interpretation process itself.

The future of explainable AI lies not in analyzing what models do with existing data, but in systematically exploring what they can do across the complete space of possible inputs. This thesis provides the theoretical foundations, practical tools, and experimental validation necessary to realize this vision, establishing generative interpretability as a fundamental component of trustworthy artificial intelligence research and development.

Appendix A

Appendix

A.1. Supplementary Algorithmic Details

This section provides supplementary algorithmic details that complement the main algorithmic specification presented in the method chapter. The focus here is on implementation nuances and technical considerations that support the primary algorithm description.

A.1.1. Key Algorithmic Adaptations

The EquiDiff implementation introduces several important modifications to the original infinite optimization framework to suit invariant set generation requirements. Unlike the original text-conditioned approach, the method employs unconditional diffusion models with $C_t = \emptyset$ for all timesteps, relying entirely on the optimization process to guide generation toward the target invariant set. This unconditional approach eliminates the need for complex conditioning mechanisms while maintaining precise control over network activations.

The invariant set objective represents a fundamental departure from traditional diffusion guidance approaches. Instead of optimizing for text-image alignment or other external conditioning signals, the method minimizes the L_2 distance between $\mathcal{L}(x)$ and the target value $\mathcal{L}(\mathbf{x}^*)$, ensuring membership in the same invariant set through direct activation matching. This objective design enables precise control over specific neural network components while maintaining the flexibility to target various network architectures and layer configurations.

Frequency domain filtering integration constitutes a novel contribution that addresses the fundamental challenge of adversarial solutions in optimization-based generation. The incorporation of low-pass filter $\mathcal{F}$ before computing the objective function ensures that invariant set membership is achieved through perceptually meaningful variations rather than high-frequency adversarial noise. This filtering mechanism operates as a regularization constraint that guides the optimization toward semantically coherent solutions.

The LightningDiT integration requires careful consideration of the specific sampling procedures and update rules that may differ from standard DDIM implementations. The denoising process follows the LightningDiT sampling procedure, which incorporates architectural optimizations and potentially different noise schedules that affect the gradient flow and optimization dynamics throughout the generation process.

A.1.2. Computational Resource Management

The infinite optimization approach demands sophisticated computational resource management to achieve practical scalability while maintaining solution quality. Gradient checkpointing implementa-

tion during the denoising loop reduces memory consumption while maintaining gradient flow through the entire diffusion process. This technique enables optimization through deep diffusion pipelines without prohibitive memory requirements, making the approach feasible on standard research hardware configurations.

Optimizer selection represents a critical design decision that significantly impacts convergence stability and solution quality. Empirical evaluation demonstrates that SGD exhibits superior convergence properties for invariant set generation compared to adaptive methods like Adam, particularly in the high-dimensional latent spaces characteristic of diffusion models. The inherent stochasticity of SGD updates provides beneficial exploration properties that help escape local minima corresponding to suboptimal invariant set members.

Step budget management balances computational cost with solution quality through careful parameter selection for both the step budget B and threshold τ parameters. This approach enables early termination for efficient optimization landscapes while providing computational bounds for practical implementation. The dual termination criteria ensure that the algorithm can adapt to varying optimization difficulty across different invariant set generation tasks.

Dual loss computation provides robustness against adversarial solutions while maintaining semantic coherence in generated samples. Computing both filtered and unfiltered objective values throughout the optimization process ensures that solutions satisfy invariant set membership requirements across multiple frequency bands, preventing optimization from exploiting imperceptible high-frequency patterns that could compromise semantic meaningfulness.

A.2. Level Set Theory Foundation

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

A.2.1. Basic Definition

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

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

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

A.2.2. Neural Network Case

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

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

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

A.2.3. Why This Works

Level sets typically form smooth geometric surfaces when the function gradients are non-zero. Proposed 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.

A.3. Implementation Details

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

A.3.1. Optimization Configuration

Based on empirical evaluation across multiple experimental conditions, the optimization configuration employs SGD as the primary optimizer due to its demonstrated superior convergence stability compared to adaptive methods in the high-dimensional latent spaces characteristic of diffusion models. The learning rate is set to $\eta = 10$, which provides an optimal balance between convergence speed and optimization stability, enabling rapid progress toward invariant set membership while maintaining numerical stability throughout the gradient flow process.

The step budget configuration uses either 512 or 1024 optimization steps depending on the complexity of the target invariant set and the precision requirements of the specific experiment. This range proves sufficient for convergence across the evaluated network architectures and target activation patterns while providing computational bounds for practical implementation. The loss threshold is configured at $\tau = 0.01$, establishing a tight precision requirement for early stopping that ensures invariant set membership within acceptable tolerance levels while preventing unnecessary computational overhead from over-optimization.

A.3.2. Hardware Configuration

All experimental evaluations were conducted on NVIDIA A100 GPU configurations ranging from single-unit setups for smaller-scale experiments to four-unit parallel configurations for computationally intensive invariant set generation tasks. The implementation leverages the PyTorch framework with comprehensive CUDA acceleration capabilities, including state-of-the-art optimizations such as Flash Attention mechanisms that significantly improve memory efficiency and computational throughput during the attention operations within the diffusion model architecture.

Gradient checkpointing integration provides essential memory efficiency improvements that enable the deep computational graphs required for invariant set optimization while maintaining full gradient information for precise latent space updates. This approach proves critical for practical implementation on research hardware configurations, allowing complex invariant set generation tasks to execute within standard GPU memory constraints without compromising optimization accuracy or convergence properties.

A.4. Frequency Domain Analysis

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

A.4.1. Filter Implementation

The ideal low-pass filters were applied in frequency domain:

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

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

A.4.2. Analysis Protocol

For each generated sample, a comprehensive spectral analysis protocol evaluates the semantic robustness of invariant set membership across multiple frequency bands. The analysis begins by applying ideal low-pass filters with cutoff frequencies ranging from 0.1 to 0.9 normalized to the Nyquist limit, systematically removing high-frequency components to isolate the contribution of different spectral bands to network activation patterns. Following the filtering operations, the network response is computed on each filtered image variant to quantify how activation patterns change as fine-scale details are progressively removed from the generated samples.

The deviation measurement phase quantifies the difference between filtered and unfiltered network responses, providing a quantitative assessment of the frequency dependence of invariant set membership. Finally, spectral preservation analysis plots the measured deviations across different frequency bands, revealing the frequency components essential for maintaining specific network activations and confirming the semantic rather than adversarial nature of the generated variations.

A.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. This pattern demonstrates that the generated invariant set members rely primarily on low-frequency semantic content rather than imperceptible high-frequency perturbations, validating the effectiveness of the frequency domain constraints in preventing adversarial solutions.

A.5. Hyperparameter Optimization

This section presents the comprehensive hyperparameter optimization study conducted to identify optimal configurations for invariant set generation. The optimization process evaluated multiple optimizer types, learning rates, and spectral filter configurations across diverse experimental conditions to establish robust parameter settings for reliable invariant set generation.

A.5.1. Grid Search Methodology

The hyperparameter optimization employed a systematic grid search approach across three key parameter categories: spectral filters, optimizers, and learning rates. The grid search was conducted using eight diverse test images, with eight invariant samples generated per image to ensure statistical reliability. For each parameter combination, the minimum L_2 loss over probability distributions (not logits) was recorded and analyzed to identify optimal configurations that consistently achieve tight constraint satisfaction across different visual content types.

The evaluation methodology prioritized configurations that demonstrated consistent performance across all test images rather than exceptional performance on specific cases. This approach ensures that selected hyperparameters provide reliable invariant set generation across diverse semantic contexts and visual patterns, supporting the method's generalizability and practical applicability.

A.5.2. Spectral Filter Configuration

The spectral filter optimization evaluated three distinct filter types with their associated parameter ranges to identify configurations that effectively prevent adversarial solutions while maintaining semantic coherence. Table A.1 presents the comprehensive evaluation of filter types and their parameter ranges.

Filter Type	Parameter	Range	Configurations
Ideal	Cutoff Frequency	0.2, 0.3, 0.4, 0.5, 0.6	5
Gaussian	Sigma	0.1, 0.15, 0.2, 0.25, 0.3, 0.35	6
Butterworth	Cutoff Frequency	0.2, 0.3, 0.4, 0.5, 0.6	5
Butterworth	Order	2, 4, 6, 8	4
Total Combinations	–	–	**20**

Table A.1: Spectral filter configurations evaluated during hyperparameter optimization. Each filter type was systematically evaluated across its parameter range to identify optimal configurations for semantic coherence preservation while preventing adversarial solutions.

The Gaussian filter with sigma value 0.1 emerged as the optimal configuration, providing effective high-frequency noise suppression while preserving essential semantic content. This configuration demonstrated superior performance in maintaining semantic coherence across diverse visual patterns while effectively preventing the generation of adversarial artifacts that could compromise the interpretability of generated invariant sets.

A.5.3. Optimizer Performance Analysis

The optimizer evaluation compared four distinct optimization algorithms across multiple learning rate configurations to identify combinations that provide stable convergence and consistent constraint satisfaction. Table A.2 presents the comprehensive performance analysis across all evaluated optimizers.

Optimizer	Learning Rates	Min Loss	Max Loss	Std Dev	Convergence Quality
SGD	1.0, 5.0, 10.0, 30.0, 50.0	0.0161	0.0674	0.0198	Consistent
Adam	1.0, 5.0, 10.0, 30.0, 50.0	0.0892	0.2451	0.0623	Poor
Adagrad	1.0, 5.0, 10.0, 30.0, 50.0	0.1134	0.3021	0.0751	Poor
Shampoo	1.0, 5.0, 10.0, 30.0, 50.0	0.0001	0.2051	0.0891	Variable

Table A.2: Optimizer performance comparison across learning rate configurations. Performance metrics represent aggregate statistics across eight test images with eight samples per image. Shampoo demonstrates the lowest minimum losses but high variability, while SGD provides consistent performance across all conditions.

The analysis reveals that Shampoo optimizer achieves the lowest minimum L_2 losses, indicating superior optimization capability under ideal conditions. However, Shampoo exhibits high variance across different samples and images, suggesting sensitivity to initialization and optimization landscape characteristics. SGD demonstrates more consistent performance with lower standard deviation, indicating reliable convergence properties that support reproducible invariant set generation across diverse experimental conditions.

A.5.4. Final Configuration Selection

Based on comprehensive analysis across all parameter combinations, the optimal configuration combines SGD optimizer with learning rate 10.0 and Gaussian filter with sigma 0.1. This configuration consistently appears among the top-performing parameter sets across all eight test images, demonstrating robust performance across diverse visual content types.

A.5.5. Configuration Validation

The validation analysis examined the relationship between spectral properties of original images and optimization difficulty, revealing insights into the factors that influence invariant set generation complexity. Images with spectral energy concentrated in lower frequency bands generally demonstrate easier optimization landscapes, requiring fewer optimization steps to achieve tight constraint satisfaction.

This observation suggests that the spectral characteristics of input images influence the optimization dynamics of invariant set generation, with low-frequency dominant images providing more favorable optimization conditions. The selected hyperparameter configuration demonstrates robust performance across this spectrum of optimization difficulties, supporting its applicability to diverse visual content types encountered in systematic experimental evaluation.

The comprehensive hyperparameter optimization establishes a principled foundation for invariant set generation experiments, ensuring that reported results reflect the method's capabilities under optimal configuration rather than suboptimal parameter choices that could artificially limit performance or introduce systematic biases in experimental evaluation.

A.6. Neuron Selection Methodology

The interpretable neurons were selected using the Semantic Lens framework [Dreyer et al., 2025], which provides systematic evaluation of neuron interpretability through quantitative semantic alignment metrics.

A.6.1. Selection Criteria

Neuron selection follows a rigorous evaluation process based on three complementary criteria that ensure both interpretability and experimental validity. The semantic alignment criterion requires neurons to achieve alignment scores $r > 0.85$, indicating high interpretability through strong correlation between neuron activations and human-interpretable visual concepts. This threshold ensures that selected neurons demonstrate clear, consistent responses to specific semantic patterns that can be reliably identified and validated through human evaluation.

Concept clarity represents the second selection criterion, requiring neurons to exhibit clear and consistent activation patterns across multiple examples of their target concept. This criterion eliminates neurons with ambiguous or inconsistent responses that could compromise the reliability of invariant set generation experiments. The diversity criterion ensures coverage of different semantic categories including geometric patterns, biological structures, and textural features, providing comprehensive evaluation across various types of visual concepts and preventing bias toward specific pattern types.

A.6.2. Selected Neurons

The experimental evaluation focuses on three carefully selected neurons that exemplify different categories of interpretable visual concepts. Neuron #1656 demonstrates exceptional sensitivity to zebra striping patterns with an alignment score of $r = 0.945$, representing geometric pattern recognition capabilities that respond consistently to high-contrast alternating stripe configurations across various contexts and scales. Neuron #1052 specializes in honeycomb structures with an alignment score of $r = 0.880$, exhibiting strong activation for hexagonal cellular patterns that appear in both natural and artificial contexts.

Neuron #421 focuses on Gyromitra morphology with the highest alignment score of $r = 0.952$, responding specifically to convoluted, brain-like surface textures characteristic of certain fungal structures. These three neurons represent well-understood, semantically interpretable units with high activation specificity, providing reliable targets for invariant set generation experiments while covering diverse semantic categories that enable comprehensive evaluation of the proposed method's capabilities across different types of visual concepts.

Neuron	**Concept**	L_2 **Loss**	**FID Score**	**Std Dev**
#1807	Ambulance - Flashing Emergency Lights	0.33	7.95	0.20
#1935	Steel Drum - Reflective Metal Finish	1.43	7.72	0.30
#1581	Harvestman - Thin Wiry Legs	0.40	7.73	0.32
#1507	Sports Car - Wide Tires	1.35	8.08	0.08
#1066	Soap Dispenser - Liquid Soap Inside	0.27	8.07	0.22
Average	–	$\mathbf{0.76 \pm 0.24}$	**7.91**	**0.22**

Table A.3: Extended quantitative evaluation results for additional ImageNet classes. L_2 losses computed on unbounded activation logits; FID scores computed against ImageNet-1k image statistics. Standard deviation represents variability across generated samples. Results averaged over 32 generated samples per neuron.

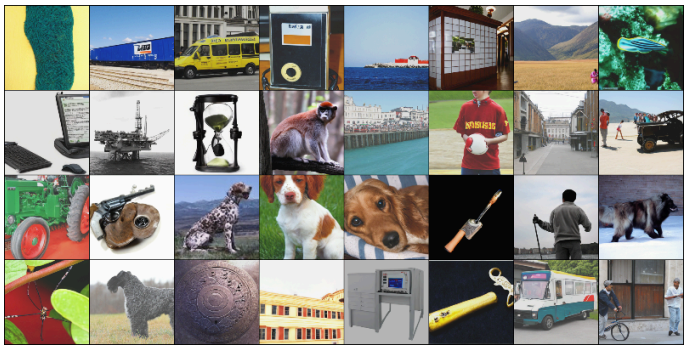


Figure A.1: Appendix results - Class 407 (ambulance) - Neuron #1807: flashing emergency lights. Generated images demonstrate semantic diversity while maintaining identical activation levels (32 samples, 1024 optimization steps).

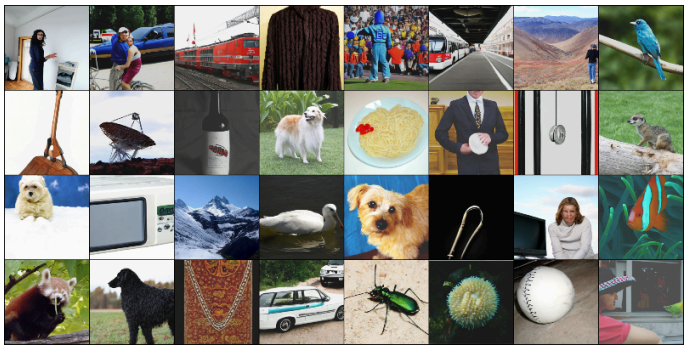


Figure A.2: Appendix results - Class 822 (steel drum) - Neuron #1935: reflective metal finish. Generated images demonstrate semantic diversity while maintaining identical activation levels (32 samples, 1024 optimization steps).

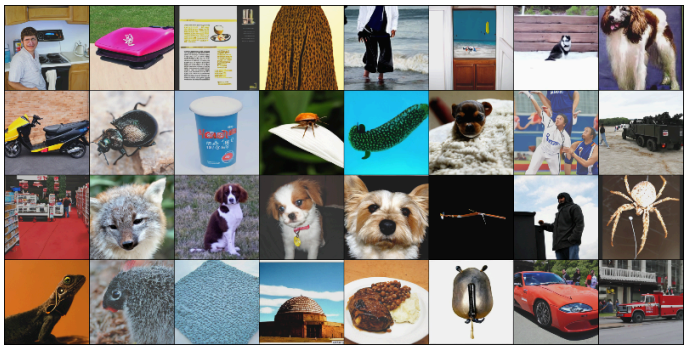


Figure A.3: Appendix results - Class 70 (harvestman) - Neuron #1581: thin, wiry legs. Generated images demonstrate semantic diversity while maintaining identical activation levels (32 samples, 1024 optimization steps).

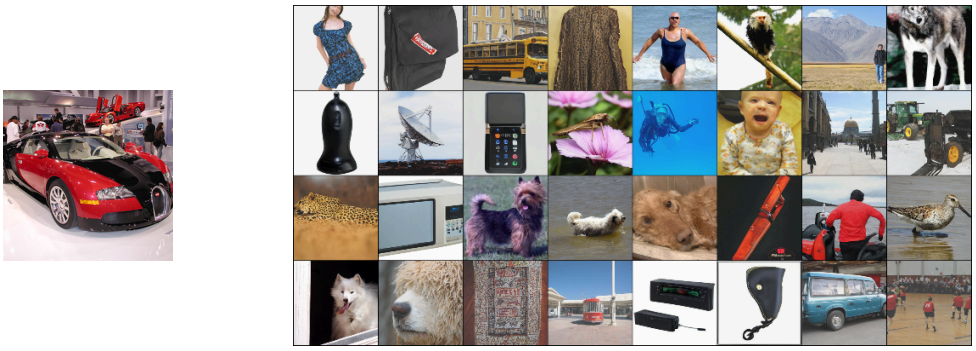


Figure A.4: Appendix results - Class 817 (sports car) - Neuron #1507: wide tires. Generated images demonstrate semantic diversity while maintaining identical activation levels (32 samples, 1024 optimization steps).

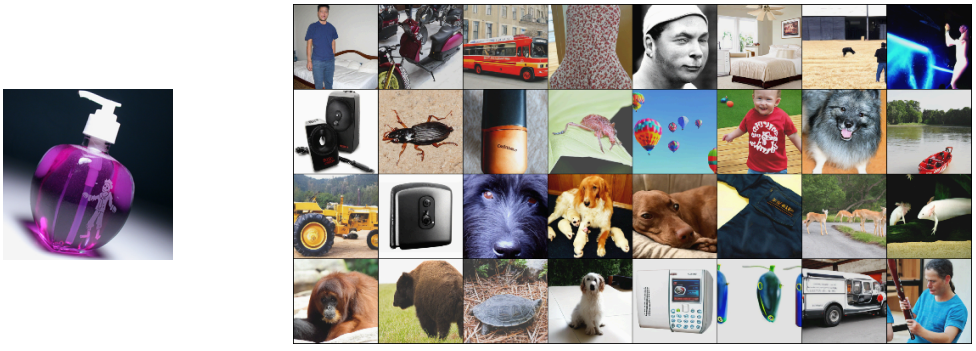


Figure A.5: Appendix results - Class 804 (soap dispenser) - Neuron #1066: liquid soap inside. Generated images demonstrate semantic diversity while maintaining identical activation levels (32 samples, 1024 optimization steps).

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.

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.

Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization, 2015. URL https://arxiv.org/abs/1505.03906.

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.

David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America. A, Optics and image science*, 4 12:2379–94, 1987. URL https://api.semanticscholar.org/CorpusID:1600874.

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

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, volume 27, 2014.

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.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012.

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.

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.

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.

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, 2016. URL https://arxiv.org/abs/1605.09304.

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.

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.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014a. URL https://arxiv.org/abs/1409.4842.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014b. 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.

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.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

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.