University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Bartłomiej Sadlej
Student no. 429589

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

Bachelor's thesis
in MACHINE LEARNING

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

Warsaw, September 2025

Abstract

Understanding the decision-making process of deep neural networks is an active area of research in machine learning. Current state-of-the-art methods focus on finding human-interpretable concepts or features that influence predictions in known data samples. However, 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. The main contribution of this work 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. This work first introduces a new framework for generative XAI and then proposes an efficient diffusion-based method for exploring it. Evaluation of this method on popular models such as ResNet-50 or Sparse Autoencoders (SAE) highlights significant limitations of current XAI methods.

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 explainable AI methods focus on the narrow slice of reality represented in training datasets, leaving vast territories of the input manifold completely unexplored and potentially harboring unexpected model behaviors. The robustness implications are staggering: if a model can be fooled into seeing a dog in a random arrangement of geometric shapes with the same confidence as it sees a dog in an actual photograph of a Golden Retriever, what does this say about its reliability in real-world deployment? The bias detection possibilities are equally concerning—systematic patterns within these datasets might reveal that the model has learned to associate certain irrelevant features (perhaps related to image compression artifacts, camera settings, or demographic markers in the background) with specific classes, perpetuating hidden biases that would never be discovered through traditional dataset analysis. For fairness evaluation, understanding these equivalence classes becomes critical: if the model makes identical predictions for inputs that vary along protected attributes while maintaining other spurious correlations, we need to map these relationships to ensure equitable treatment. Finally, the structure of these datasets provides unprecedented insight into how models generalize beyond their training distribution—revealing whether generalization relies on semantically meaningful features or on arbitrary statistical regularities that happen to correlate with class labels in the training data.

1.2. Proposed Approach: Generative Explainable AI

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

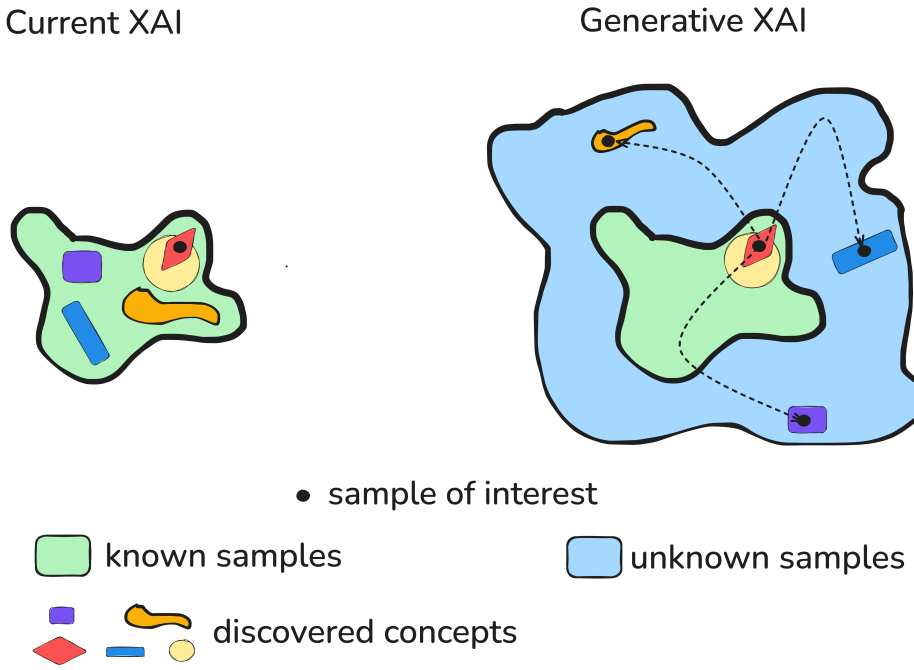


Figure 1.1: Conceptual comparison between traditional XAI methods and Generative XAI. Traditional methods analyze known data samples (left), while proposed approach synthesizes diverse examples from the invariant set that yield identical predictions (right).

1.3. Contributions

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

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

The second contribution introduces 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 explainable AI approach. Chapter 2 reviews the relevant literature across explainable AI 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 explores practical applications of this framework in real-world scenarios, while Chapter 6 discusses the broader implications of this work, current limitations, and directions for future research. Chapter 7 synthesizes contributions and their significance for the field of explainable AI.

Chapter 2

Related Work

This chapter reviews the relevant literature across several interconnected areas that form the foundation of 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 [**?**]. 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. The Realism Imperative in Explainable AI

For explainable AI applications, the requirement for realistic generation transcends technical considerations to address fundamental epistemological questions about the validity of synthetic explanations. If interpretation methods systematically produce unrealistic visualizations, they risk providing misleading insights about how neural networks process natural inputs.

Consider the implications of unrealistic explanations for different stakeholders: researchers developing new architectures may draw incorrect conclusions about feature learning if their analysis tools produce artificial patterns; practitioners deploying models in safety-critical applications may develop false confidence if explanations appear to show reasonable behavior on unrealistic inputs; and users of AI systems may lose trust if explanations appear divorced from recognizable visual concepts.

The realism constraint serves as a crucial validity check that ensures generated explanations remain grounded in the visual world that neural networks are designed to understand. By constraining interpretation methods to produce realistic images, this work ensures that insights reflect genuine computational strategies rather than artifacts of the interpretation process itself.

2.7.3. 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.

2.7.4. 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.

2.7.5. 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 compre-

hensive 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.

Invariance Set Analysis: Current approaches lack the theoretical and methodological foundations for systematically analyzing the complete manifold of inputs that produce identical model outputs. Without this capability, interpretability remains fundamentally incomplete, missing crucial insights about model robustness, generalization, and potential failure modes.

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 iterative 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 $\{\mathbf{x}_i\}_{i=1}^N$ such that:

1. ***Invariance constraint**: $\|f(\mathbf{x}_i) - f(\mathbf{x}^*)\|_2 < \epsilon$ for small $\epsilon > 0$*
2. ***Realism constraint**: $\mathbf{x}_i$ lies within the natural image manifold as defined by a pre-trained diffusion model*
3. ***Diversity constraint**: The generated samples exhibit semantic and visual diversity while maintaining the invariance constraint*

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.

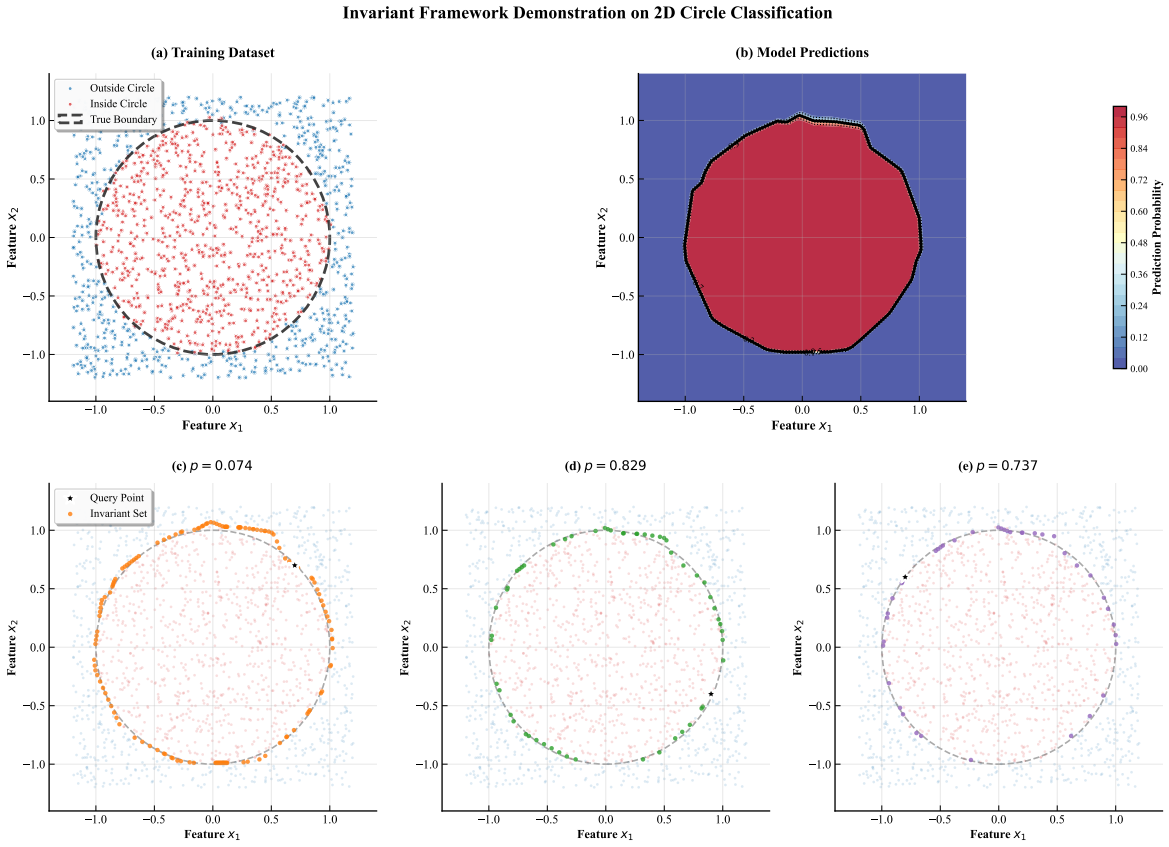


Figure 3.1: Demonstration of the Invariant Framework on a 2D Concentric Circles Dataset. (a) Training dataset with 1,500 samples classified by their position relative to a unit circle (dashed line). Blue points represent the outer class, pink points the inner class. (b) Learned decision boundary and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

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 Iterative 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 detailed algorithm specification is provided in Appendix .1.

3.4. Quality and Realism Assurance

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

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. Specifically, this work introduces a low-pass filter $\mathcal{F}$ before the objective function $\mathcal{L}$ and measures deviation from the original measurement across different cutoff frequencies f_c.

This frequency-aware approach ensures that:

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

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

Chapter 4

Experiments

This chapter presents a comprehensive experimental evaluation of 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 addresses the following core research questions:

1. Can EquiDiff generate visually diverse samples that maintain identical activation patterns for interpretable neurons?

2. Do generated samples reveal semantic patterns beyond those present in typical training data?

3. How effectively does the method preserve complex feature representations learned by sparse autoencoders?

4. Can the framework maintain classifier predictions while generating semantically meaningful variations?

4.1.1. Infrastructure and Implementation

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

4.1.2. Evaluation Framework

This work employs a multi-faceted evaluation approach combining quantitative precision metrics with qualitative semantic analysis:

Quantitative Metrics:

- **Activation Fidelity**: L_1 and L_2 norm deviations from target values

- **Probability Preservation**: Kullback-Leibler (KL) divergence for probability distributions, defined as

$$D_{KL}(P \parallel Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \tag{4.1}$$

  where $P(x)$ represents the target probability distribution and $Q(x)$ the generated distribution

- **Spectral Coherence**: Frequency domain analysis using ideal low-pass filters (see **??**)

- **Image Quality**: Fréchet Inception Distance (FID) relative to natural image statistics

Qualitative Assessment:

- Semantic diversity within invariant sets

- Visual coherence and absence of adversarial artifacts

- Alignment between generated patterns and expected neuron/feature selectivity

4.2. Individual Neuron Activation Analysis

Building upon mechanistic interpretability advances, 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:

- **Neuron #1656 (Zebra Striping)**: Alignment score $r = 0.945$, responds to black-white striped patterns

- **Neuron #1052 (Honeycomb Structure)**: Alignment score $r = 0.880$, activates on hexagonal cellular structures

- **Neuron #421 (Gyromitra Morphology)**: Alignment score $r = 0.952$, responds to convoluted, brain-like surface textures

4.2.2. Experimental Protocol

For each target neuron n, the following protocol was followed:

1. Select a query image $\mathbf{x}^*$ that strongly activates the neuron

2. Define the invariant set constraint: $n(\mathbf{x}) = n(\mathbf{x}^*)$

3. Apply EquiDiff to generate 32 samples maintaining this constraint

4. Evaluate activation fidelity, visual quality, and semantic diversity

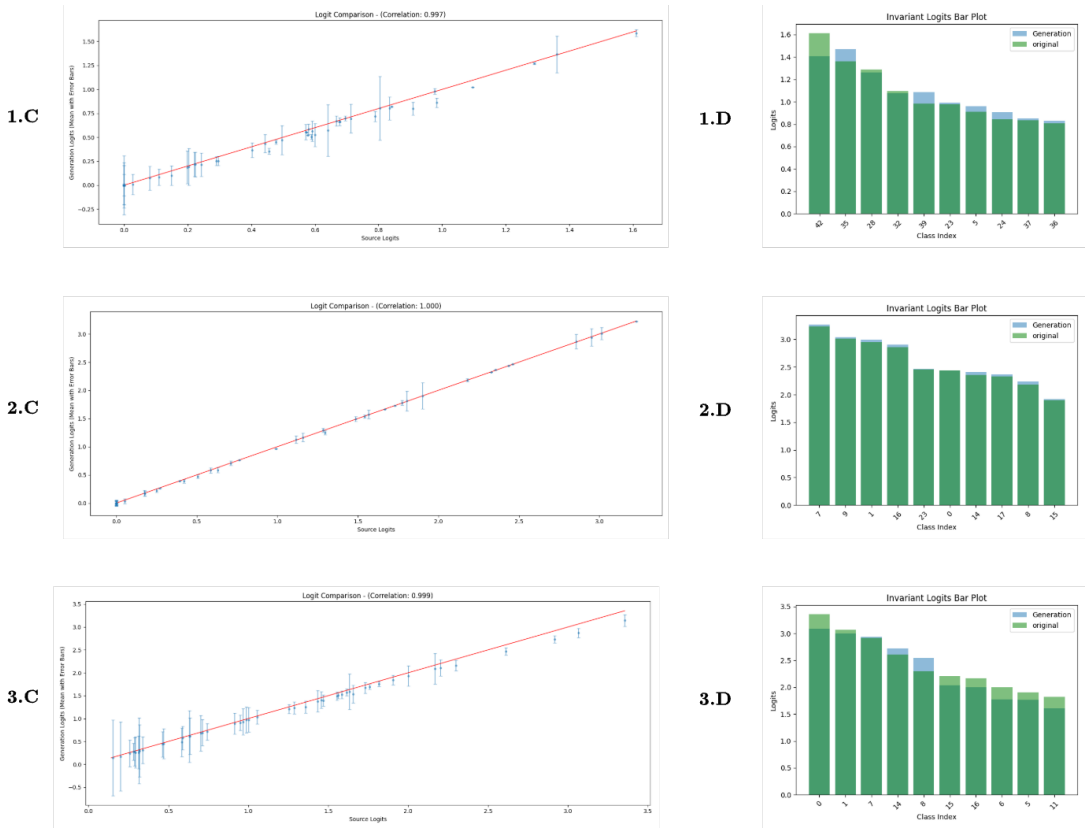


Figure 4.1: **C** - Original and Generated logits comparison **1**: #1656 (Zebra Striping), **2**: #1052 (Honeycomb), **3**: #421 (Gyromitra). **D** - top logit activation in target neuron. There are 49 logits in target neuron of last convolution layer in ResNet50 model

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

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

4.2.3. 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.

4.2.4. Qualitative Analysis

Figure 4.7 demonstrates the semantic diversity achieved within the invariant set for targeted neurons. Generated samples exhibit various patterns beyond typical imagery, including architectural elements,

textile patterns, and abstract geometric designs, all maintaining identical activation levels.

4.2.5. Cross-Neuron Comparison

The consistent performance across neurons with different semantic specializations (geometric patterns, biological textures, structural elements) demonstrates the generality of proposed approach. Notably, the inverse relationship between semantic alignment scores and generation difficulty suggests that more specialized neurons provide clearer optimization targets.

4.3. Sparse Autoencoder Feature Analysis

Sparse autoencoders (SAEs) have emerged as powerful tools for decomposing neural network representations into interpretable features. 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 SAE features from ViT models that exhibit clear semantic interpretability were targeted:

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

4.3.2. Expected Results

Based on the neuron experiments, this work anticipates:

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

4.3.3. Qualitative Results

Figure 4.8 shows representative results for SAE feature #6547, demonstrating 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: (1) the feature exhibits broader semantic scope than suggested by typical training examples, (2) invariant set membership can be maintained across significant stylistic and compositional variations, and (3) 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. This work investigates:

4.4.1. Experimental Design

This work investigates invariant set generation for:

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

4.4.2. Frequency Domain Analysis

Figure 4.10 illustrates proposed frequency domain evaluation methodology, examining how invariant set membership changes across different spectral bands. This analysis ensures that generated samples achieve invariance through semantically meaningful rather than imperceptible high-frequency variations.

4.4.3. Preliminary Observations

Initial experiments demonstrate:

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

Comprehensive results forthcoming upon experimental completion.

4.5. Discussion

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. This work concludes:

4.5.1. Key Findings

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

4.5.2. Limitations and Future Work

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

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

4.6. Experimental Design

The experimental evaluation addresses the following core research questions:

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

4.6.1. 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 **??**). Each experimental condition generates 32-256 samples due to computational constraints, representing a balance between statistical validity and resource efficiency.

4.6.2. Evaluation Framework

This work employs a multi-faceted evaluation approach combining quantitative precision metrics with qualitative semantic analysis:

Quantitative Metrics:

- **Activation Fidelity**: L_1 and L_2 norm deviations from target values
- **Probability Preservation**: Kullback-Leibler (KL) divergence for probability distributions, defined as

$$D_{KL}(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \tag{4.2}$$

  where $P(x)$ represents the target probability distribution and $Q(x)$ the generated distribution

- **Spectral Coherence**: Frequency domain analysis using ideal low-pass filters (see **??**)

- **Image Quality**: Fréchet Inception Distance (FID) relative to natural image statistics

Qualitative Assessment:

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

4.7. 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.

4.7.1. Target Neuron Selection

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

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

4.7.2. Experimental Protocol

For each target neuron n, the following protocol was followed:

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

4.7.3. 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.

4.7.4. Qualitative Analysis

Figure 4.7 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.

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.

4.7.5. Cross-Neuron Comparison

The consistent performance across neurons with different semantic specializations (geometric patterns, biological textures, structural elements) demonstrates the generality of proposed approach. Notably, the inverse relationship between semantic alignment scores and generation difficulty suggests that more specialized neurons provide clearer optimization targets.

4.8. 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].

4.8.1. Experimental Setup

The SAE features from ViT models that exhibit clear semantic interpretability were targeted:

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

4.8.2. Expected Results

Based on the neuron experiments, this work anticipates:

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

4.8.3. Qualitative Results

Figure 4.8 shows representative results for SAE feature #6547, demonstrating 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: (1) the feature exhibits broader semantic scope than suggested by typical training examples, (2) invariant set membership can be maintained across significant stylistic and compositional variations, and (3) 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.

4.9. Classifier Output Preservation

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

4.9.1. Experimental Design

This work investigates invariant set generation for:

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

4.9.2. Frequency Domain Analysis

Figure 4.10 illustrates proposed frequency domain evaluation methodology, examining how invariant set membership changes across different spectral bands. This analysis ensures that generated samples achieve invariance through semantically meaningful rather than imperceptible high-frequency variations.

4.9.3. Preliminary Observations

Initial experiments demonstrate:

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

Comprehensive results forthcoming upon experimental completion.

4.10. 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. This work concludes:

4.10.1. Key Findings

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

4.10.2. Limitations and Future Work

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

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

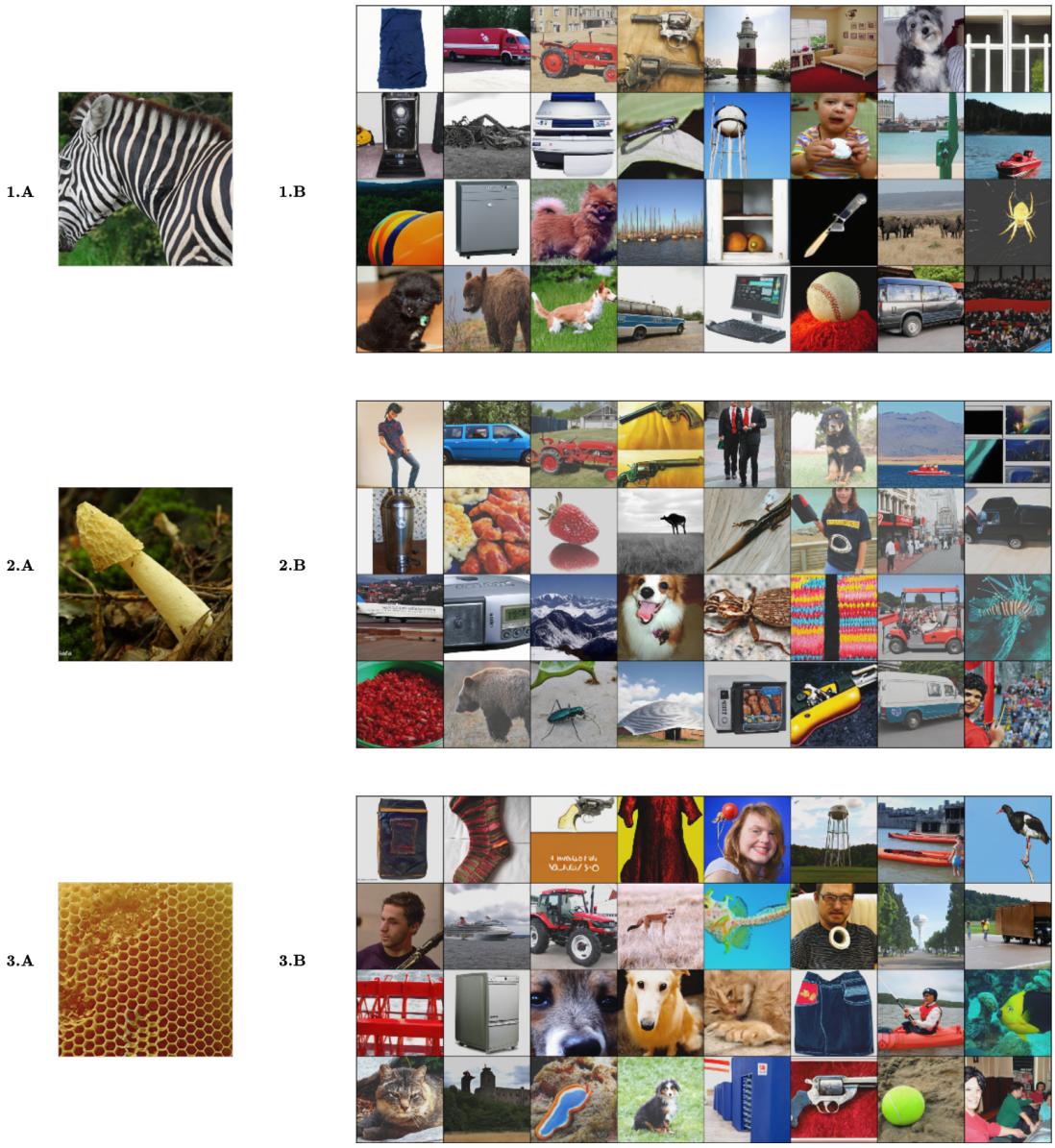


Figure 4.2: **B** - Invariant set samples for Neuron **1**: #1656 (Zebra Striping), **2**: #1052 (Honeycomb), **3**: #421 (Gyromitra). **A** - source images. Generated images demonstrate semantic diversity while maintaining identical activation levels (32 samples, 1024 optimization steps). The method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

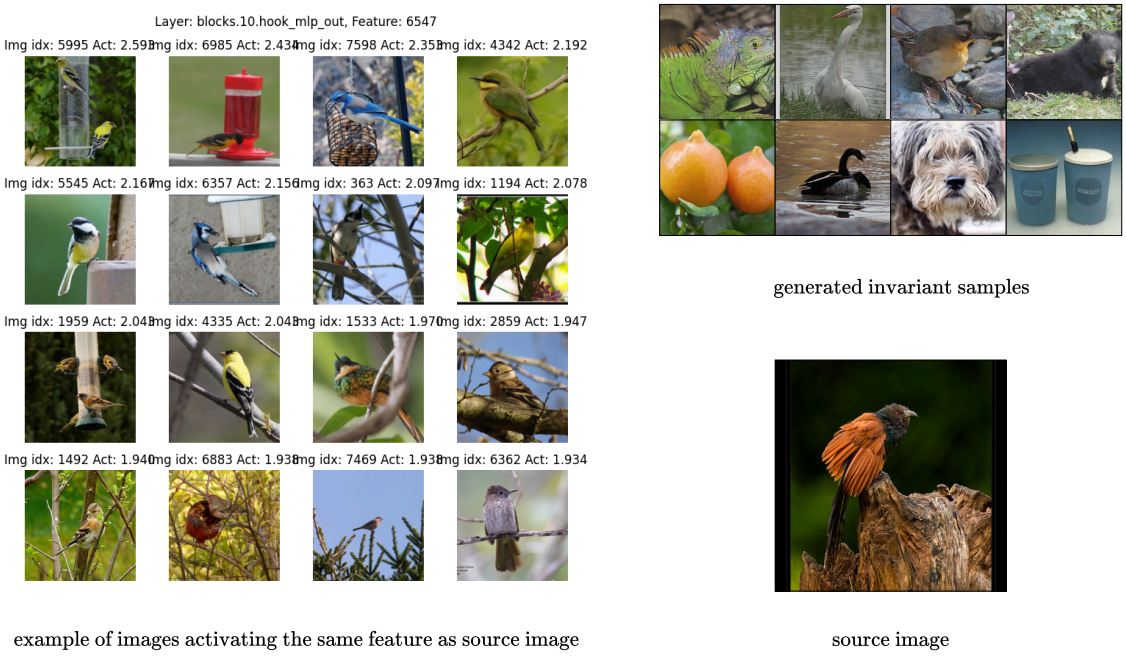


Figure 4.3: Invariant set generation for sparse autoencoder feature #6547 demonstrates precise activation preservation and semantic diversity. **Left**: Representative real images from the training dataset that naturally activate this feature, establishing the ground truth semantic concept learned by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

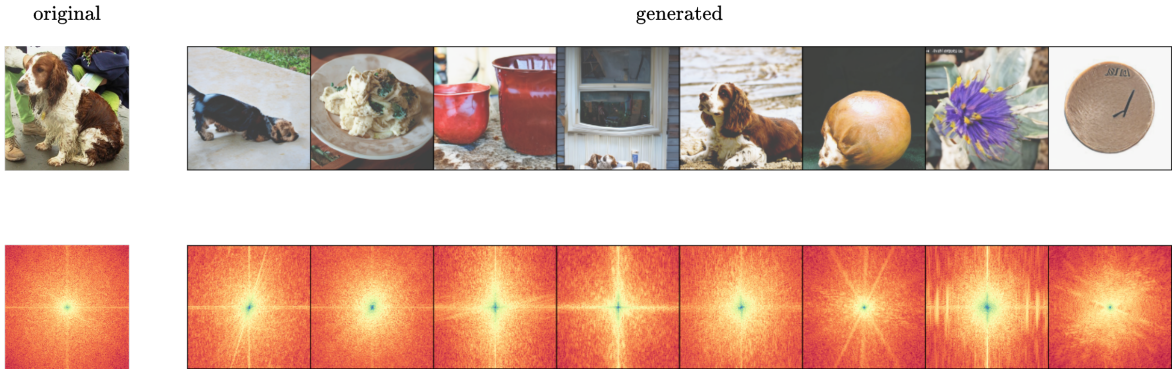
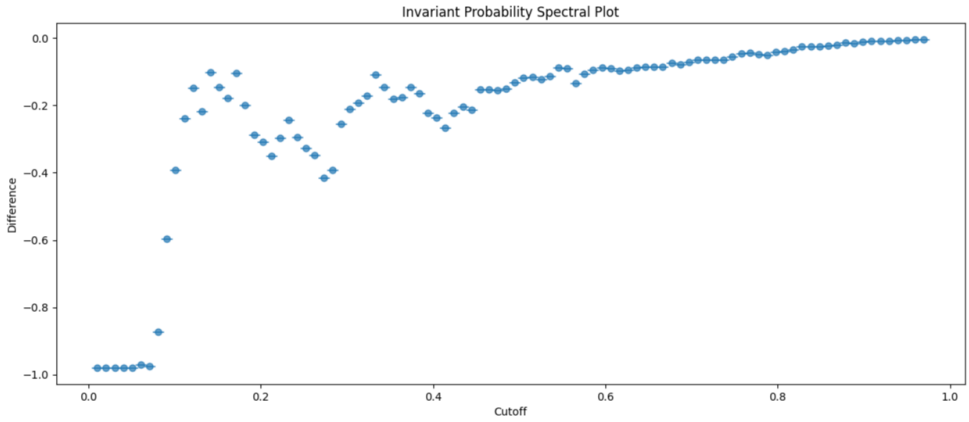


Figure 4.4: Invariant images that preserve ResNet50 classifier probability with 0.01 L2 loss on the right and original image on the left. Bottom row shows spectral heatmap of the image showing that although generated samples are of high quality but spectral analysis can reveal their synthetic background

original



generated

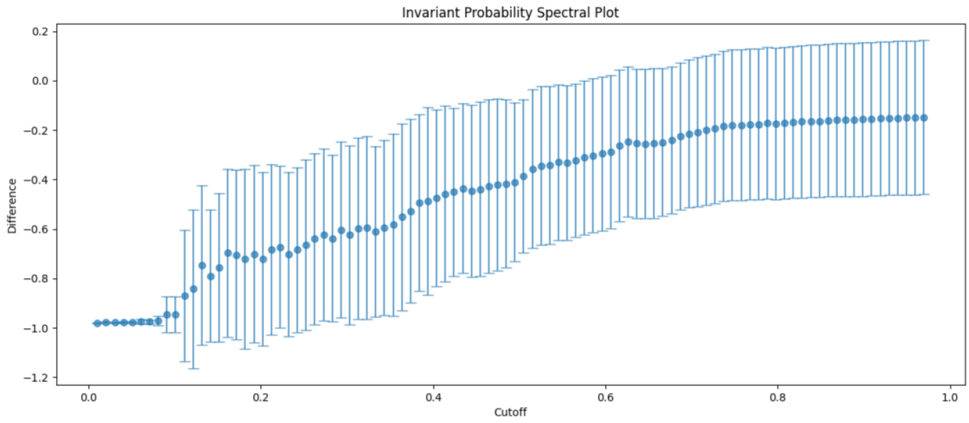


Figure 4.5: Difference between ground true class probability value in source image and in image passed through cut-off filter in spectral domain. This comparison clearly shows that the biggest difference in classifier output occurs around the same frequency value which suggest that although generated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

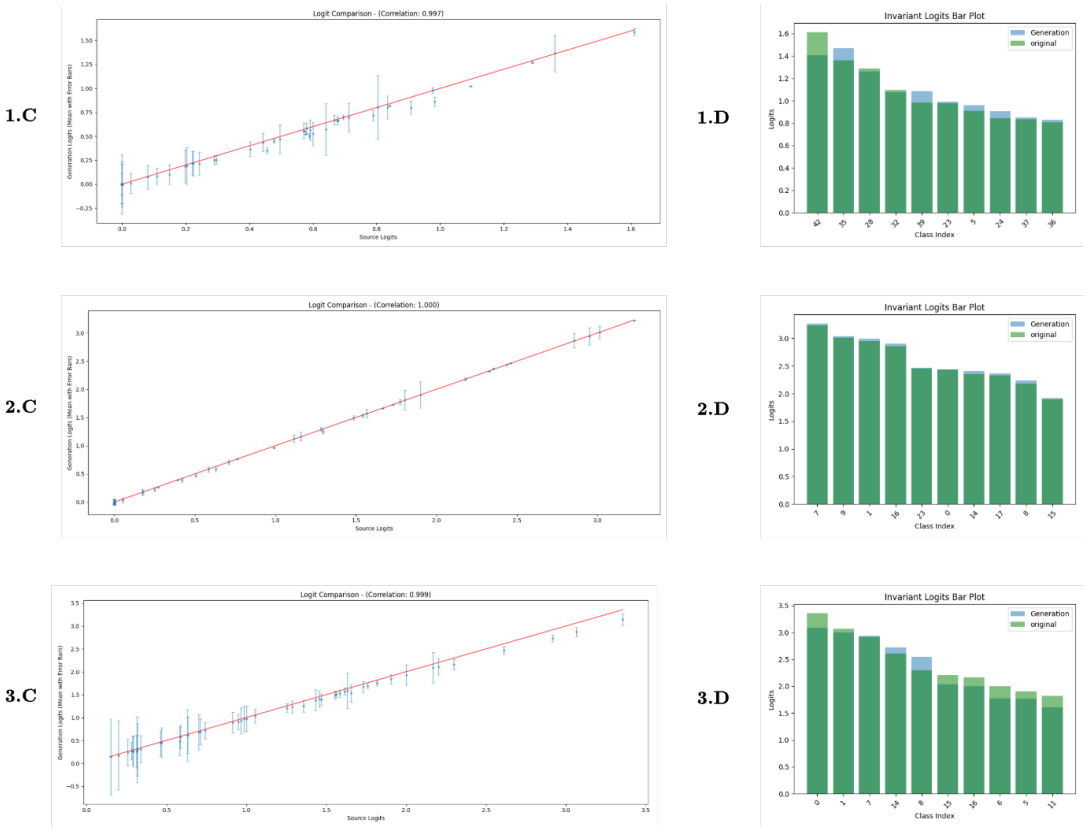


Figure 4.6: **C** - Original and Generated logits comparison **1**: #1656 (Zebra Striping), **2**: #1052 (Honeycomb), **3**: #421 (Gyromitra). **D** - top logit activation in target neuron. There are 49 logits in target neuron of last convolution layer in ResNet50 model

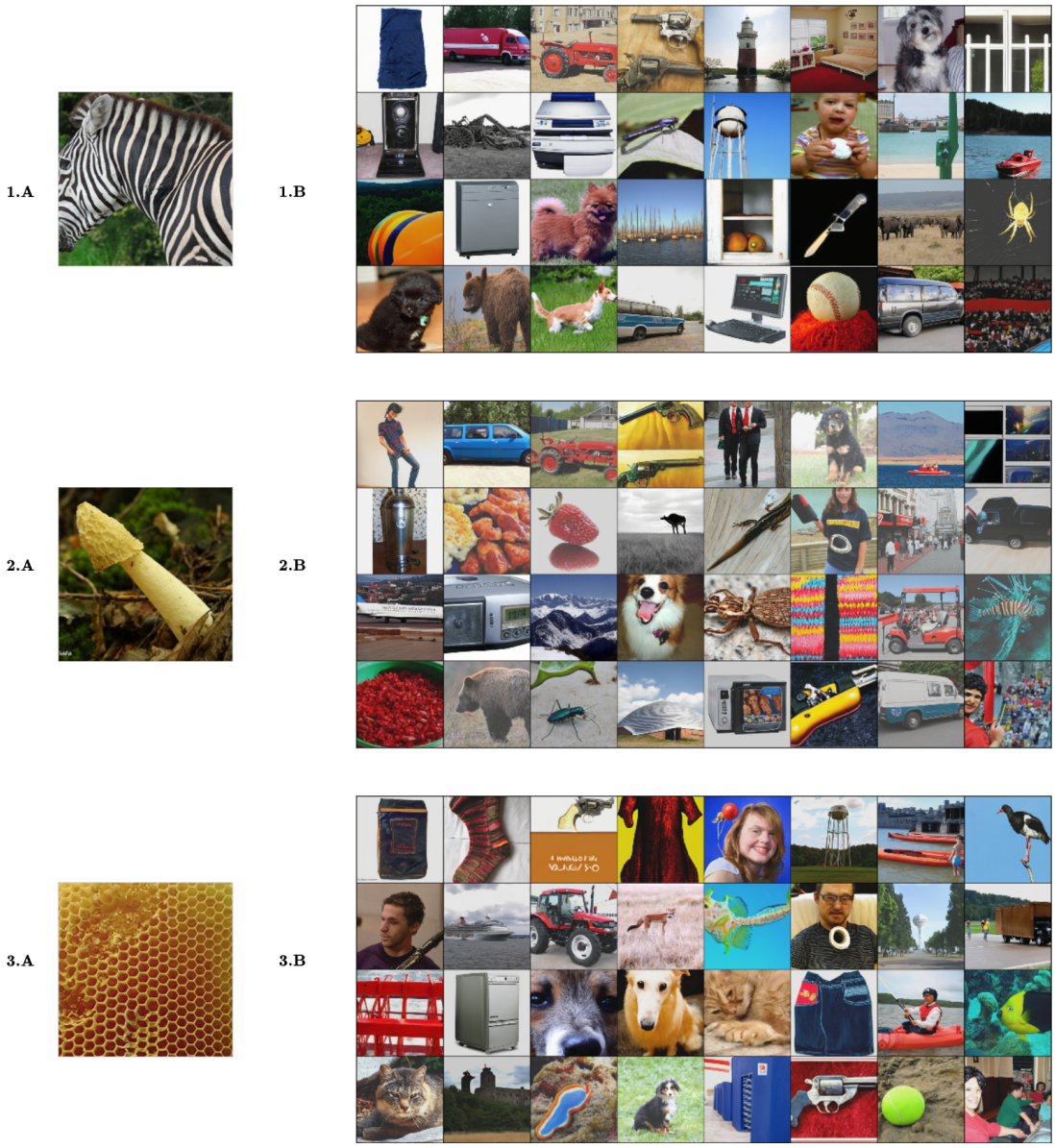


Figure 4.7: **B** - Invariant set samples for Neuron **1**: #1656 (Zebra Striping), **2**: #1052 (Honeycomb), **3**: #421 (Gyromitra). **A** - source images. Generated images demonstrate semantic diversity while maintaining identical activation levels (32 samples, 1024 optimization steps). The method successfully discovers diverse patterns that activate the same neural pathway, revealing the broader scope of visual features detected by this semantic unit.

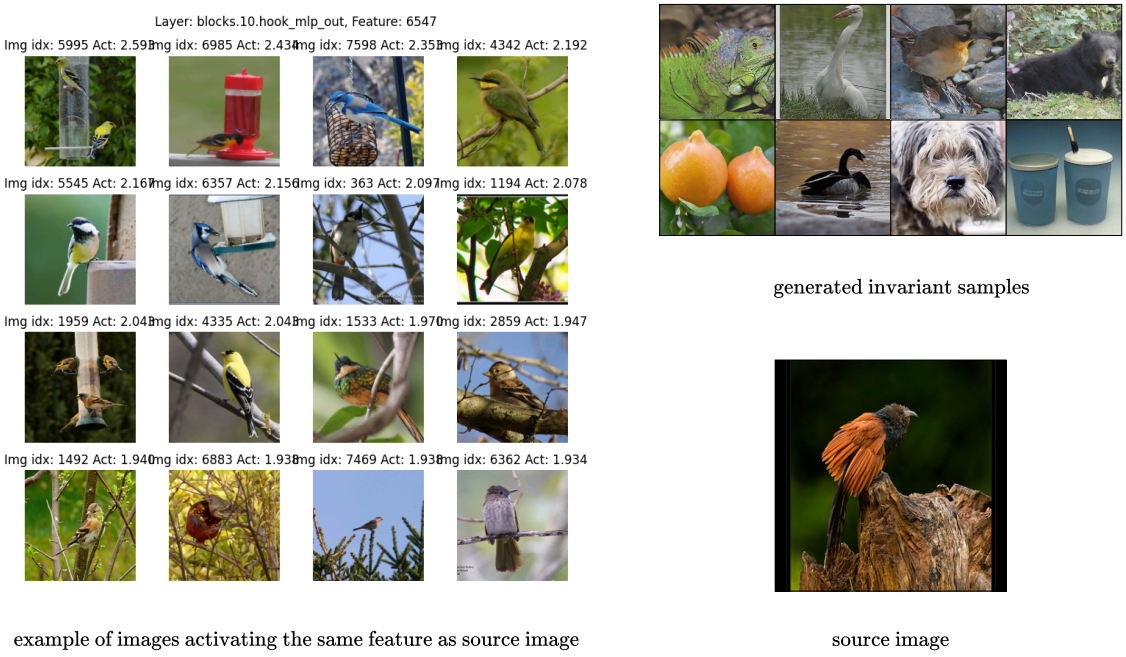


Figure 4.8: Invariant set generation for sparse autoencoder feature #6547 demonstrates precise activation preservation and semantic diversity. **Left**: Representative real images from the training dataset that naturally activate this feature, establishing the ground truth semantic concept learned by the SAE. **Top right**: Generated samples from the invariant set using EquiDiff with 512 optimization steps. All generated images achieve tight activation matching with L2 loss ≈ 0.01 relative to the target activation level, demonstrating mathematical precision in invariant set membership. The generated samples reveal the broader visual manifold of patterns that trigger identical feature responses, extending beyond the original training examples to include novel compositions, lighting conditions, and stylistic variations while preserving the core semantic concept. This diversity illustrates how invariant sets can expose the full scope of visual patterns encoded by individual SAE features, providing insights into learned representations that extend far beyond observed training data.

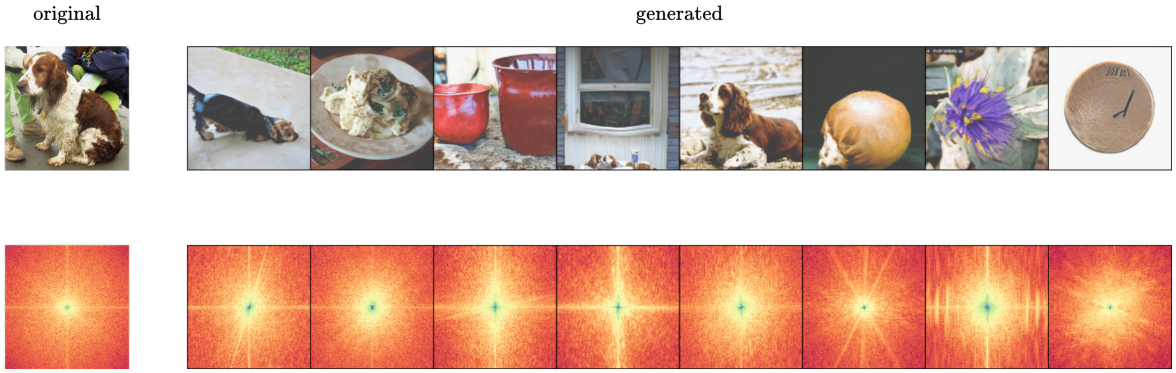
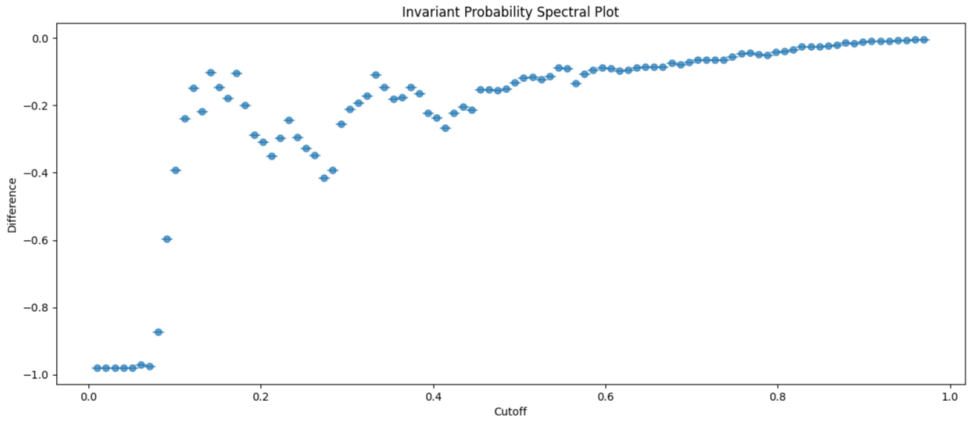


Figure 4.9: Invariant images that preserve ResNet50 classifier probability with 0.01 L2 loss on the right and original image on the left. Bottom row shows spectral heatmap of the image showing that although generated samples are of high quality but spectral analysis can reveal their synthetic background

original



generated

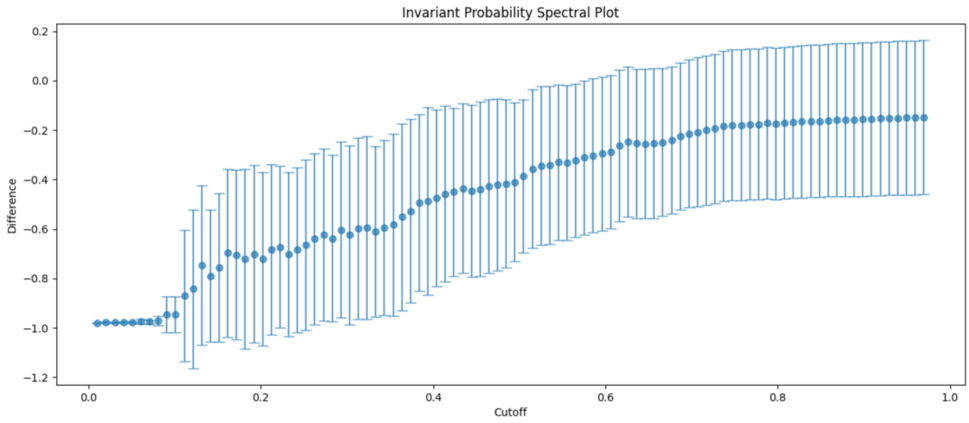


Figure 4.10: Difference between ground true class probability value in source image and in image passed through cut-off filter in spectral domain. This comparison clearly shows that the biggest difference in classifier output occurs around the same frequency value which suggest that although generated samples have different spectral view, they encode signal in the same power levels.

Chapter 5

Applications

Chapter 6

Discussion

Chapter 7

Conclusion

.1. Infinite Optimization Algorithm

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

Algorithm 1 Invariant Set Generation via Infinite Optimization

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

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

.1.1. Key Differences from Original Algorithm

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

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

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

.1.2. Computational Considerations

The infinite optimization approach requires careful management of computational resources:

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

.2. Level Set Theory Foundation

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

.2.1. Basic Definition

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

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

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

.2.2. Neural Network Case

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

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

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

.2.3. Why This Works

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

.3. Implementation Details

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

.3.1. Optimization Configuration

Based on empirical evaluation, the following parameters were selected:

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

.3.2. Hardware Configuration

All experiments were conducted on:

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

.4. Frequency Domain Analysis

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

.4.1. Filter Implementation

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

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

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

.4.2. Analysis Protocol

For each generated sample, the following protocol was followed:

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

.4.3. Quality Interpretation

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

.5. Neuron Selection Methodology

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

.5.1. Selection Criteria

Neurons were chosen based on:

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

.5.2. Selected Neurons

The three target neurons were selected:

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

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

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	–	**0.76 ± 0.24**	**7.91**	**0.22**

Table 1: 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.

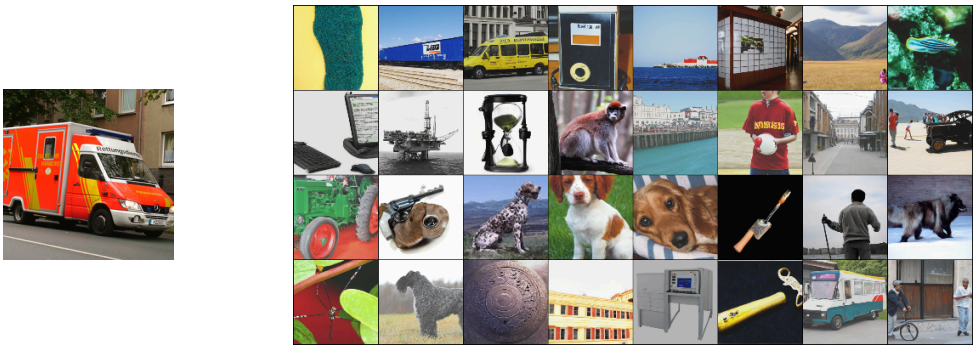


Figure 1: Appendix results - Class 407 (ambulance) - Neuron #1807: flashing emergency lights. Generated images demonstrate semantic diversity while maintaining identical activation levels (32 samples, 1024 optimization steps).

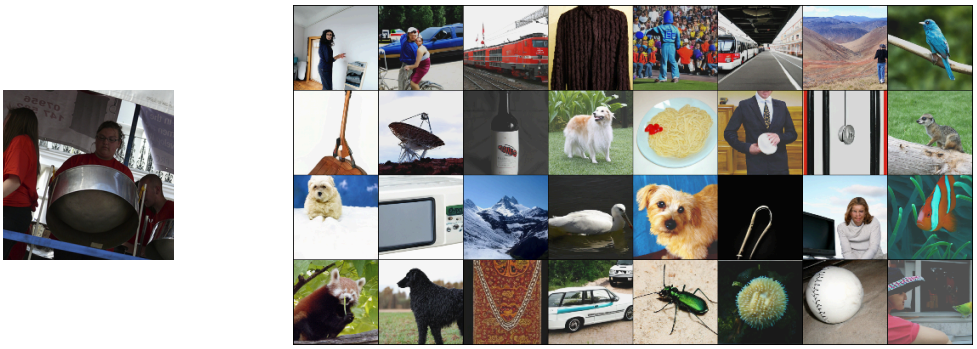


Figure 2: Appendix results - Class 822 (steel drum) - Neuron #1935: reflective metal finish. Generated images demonstrate semantic diversity while maintaining identical activation levels (32 samples, 1024 optimization steps).

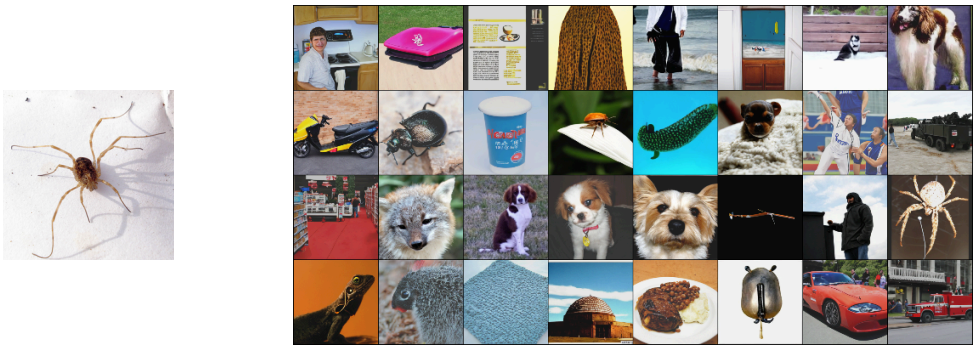


Figure 3: Appendix results - Class 70 (harvestman) - Neuron #1581: thin, wiry legs. Generated images demonstrate semantic diversity while maintaining identical activation levels (32 samples, 1024 optimization steps).

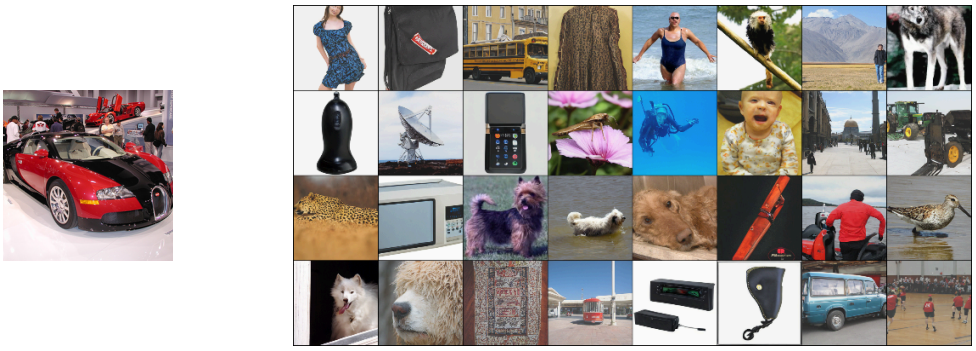


Figure 4: Appendix results - Class 817 (sports car) - Neuron #1507: wide tires. Generated images demonstrate semantic diversity while maintaining identical activation levels (32 samples, 1024 optimization steps).

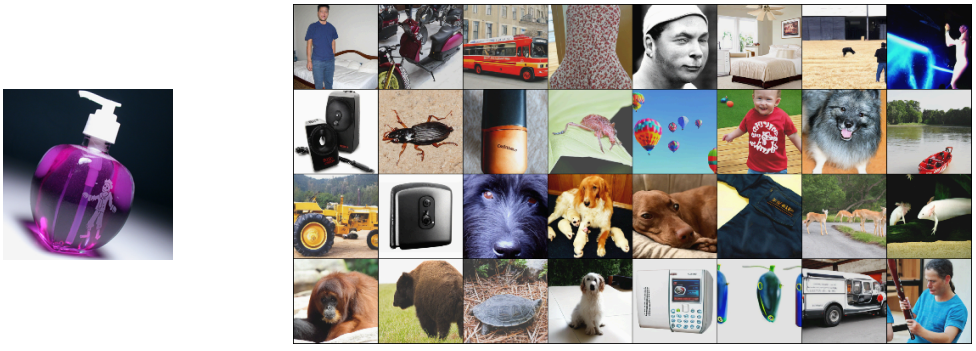


Figure 5: Appendix results - Class 804 (soap dispenser) - Neuron #1066: liquid soap inside. Generated images demonstrate semantic diversity while maintaining identical activation levels (32 samples, 1024 optimization steps).

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.

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

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

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

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

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.

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.