University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Bartłomiej Sadlej
Student no. 429589

Title in English

Bachelor's thesis
in MACHINE LEARNING

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

Warsaw, May 2017

Abstract

Understanding the decision-making process of deep neural networks is an active area of research in machine learning. Current state of the art methods focus on finding human-interpretable concepts or features that influence predictions in known data samples. However, we argue that this approach is limited in its ability to provide a comprehensive understanding of model behavior due to vast unexplored regions of the data manifold, not present in investigated datasets, which can potentially lead to the same predictions. This work is a continuation of pioneering work on Generative Explainable AI (XAI) **?** published at International Conference on Learning Representations 2025 and introduces a novel framework of Invariants - sets of data points that yield identical predictions under a given model, thereby establishing an equivalence relation. Our method EquiDiff allows for sampling meaningful and diverse high quality realistic examples from these invariant sets and as a result opens new possibilities for evaluating existing explainability methods on unknown data - simulating the real world, and fully preventing training data leaks by generating synthetic samples.

Keywords

blabaliza różnicowa, fetory σ-ρ, fooizm, blarbarucja, blaba, fetoryka, baleronik

Thesis domain (Socrates-Erasmus subject area codes)

11.4 Sztuczna inteligencja

Subject classification

D. Software
D.127. Blabalgorithms
D.127.6. Numerical blabalysis

Tytuł pracy w języku polskim

Tytuł po polsku

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 **?**, concept activation vectors **?**, and gradient-based attribution methods **?**. 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.

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?

This question is not merely academic. Understanding the full scope of inputs that lead to identical model predictions is crucial for several reasons:

1. **Robustness Assessment**: Identifying the complete set of equivalent inputs reveals potential vulnerabilities and edge cases that might not be present in training data.

2. **Bias Detection**: Systematic patterns within invariant sets can reveal spurious correlations and biases that the model has learned.

3. **Fairness Evaluation**: Understanding what variations preserve predictions helps assess whether models make decisions based on relevant features rather than protected attributes.

4. **Generalization Understanding**: The structure of invariant sets provides insight into how models generalize beyond their training distribution.

1.2. Our Approach: Generative Explainable AI

This thesis introduces a paradigm shift from traditional interpolative XAI methods to a generative approach. Instead of analyzing existing data points, we propose synthesizing new, meaningful exam-

ples that preserve model predictions, thereby exploring the *Invariant Set* – the complete collection of inputs that yield identical outputs under a given objective function.

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

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

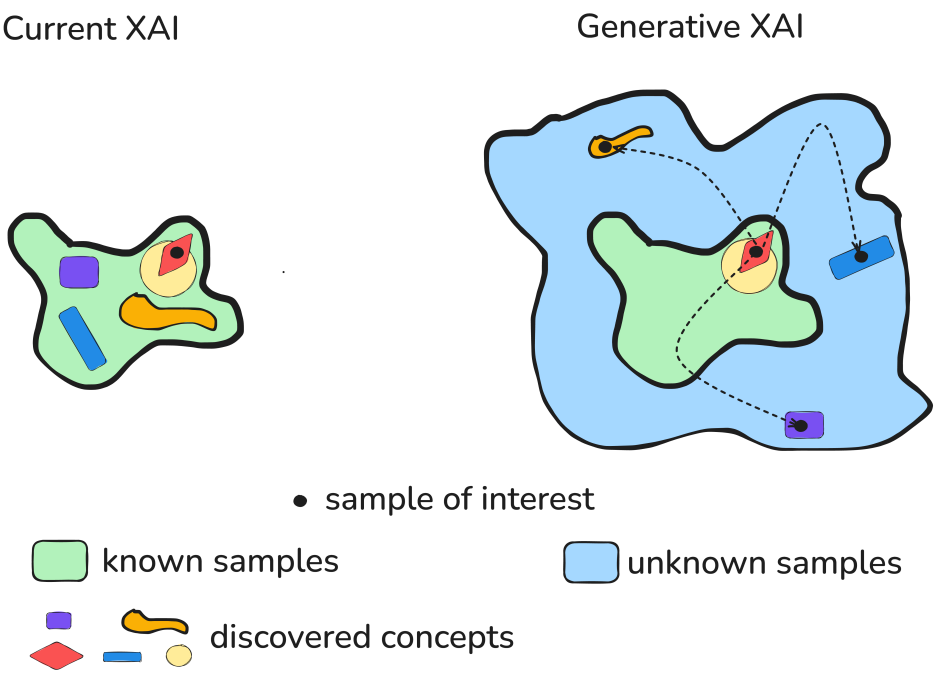

Figure 1.1: Conceptual comparison between traditional XAI methods and Generative XAI. Traditional methods analyze known data samples (left), while our approach synthesizes diverse examples from the invariant set that yield identical predictions (right).

1.3. Contributions

This thesis makes the following key contributions to the field of explainable AI:

1. **Theoretical Framework**: We provide a formal mathematical definition of Invariants and establish their properties as equivalence relations.

2. **Algorithmic Innovation**: We develop EquiDiff, a practical algorithm that combines score-based diffusion models with guided sampling to generate high-quality examples from invariant sets.

3. **Comprehensive Evaluation**: We conduct extensive experiments demonstrating the effectiveness of our approach across multiple computer vision tasks and architectures.

4. **Novel Insights**: We show how invariant set analysis can reveal model biases, spurious correlations, and learned invariances that are not detectable through traditional XAI methods.

5. **Quality Assurance**: We develop rigorous methods for ensuring the generated images maintain both semantic quality and mathematical invariance properties.

1.4. Thesis Organization

The remainder of this thesis is organized as follows. Chapter **??** reviews relevant literature in explainable AI, generative modeling, and diffusion models. Chapter **??** presents our theoretical framework and the EquiDiff algorithm. Chapter **??** details our experimental setup and results. Chapter **??** explores practical applications of our framework. Chapter **??** discusses implications, limitations, and future directions. Finally, Chapter **??** summarizes our contributions and concludes the thesis.

Chapter 2

Related Work

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

2.1. Explainable Artificial Intelligence

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

2.1.1. Attribution Methods

Attribution methods aim to identify which input features are most important for a model's prediction in a form of a heatmap. Gradient-based methods like Integrated Gradients **?** and GradCAM **?** 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 **?** and SHAP **?** 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) **?** learn linear directions in activation space that correspond to human-defined concepts. Network Dissection **?** 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) **?** uses unsupervised segmentation to identify important concepts, while TCAV (Testing with CAVs) **?** 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, **?** advanced this direction in a way that directly inspired our approach. Our method is, to our knowledge, the first to combine high-quality results with almost real-time performance.

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

2.2. Score-Based Generative Models

Score-based generative models (SGMs) have emerged as a powerful framework for high-quality image generation. Following the seminal work of **?**, 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 **??**:

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

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

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

2.3. Conditional Generation and Classifier Guidance

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

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

2.3.1. Classifier Guidance

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

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

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

2.3.2. Limitations of Standard Classifier Guidance

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

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

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

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 **??**. 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.

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

2.4.1. Diverse Posterior Sampling

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

2.5. Activation Maximization and Feature Visualization

Activation maximization techniques attempt to synthesize inputs that maximally activate specific neurons or model outputs **??**. The basic approach optimizes an input image $\mathbf{x}$ to maximize an objective function $\mathcal{L}(\mathbf{x})$:

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

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

However, activation maximization methods often produce unrealistic images accompanied by high-frequency artifacts that are imperceptible to humans yet strongly activate neurons **???**. To address this, various regularization techniques have been proposed, including total variation penalties, which encourage smoother and more coherent outputs **?**, and frequency domain constraints that reduce high-frequency noise and improve interpretability **?**.

2.5.1. Limitations and Relationship to Our Work

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

1. **Single Solution vs. Diverse Sets**: Activation maximization typically finds one optimal input, while we aim to generate a set of new data points.
2. **Maximum Activation vs. Preserved Predictions**: Activation maximization seeks to maximize responses while we preserve specific prediction values.
3. **Quality Issues**: Traditional activation maximization often produces unrealistic images, while our diffusion-based approach leverages strong visual priors.

2.6. Concept Discovery and Spurious Feature Detection

Understanding what concepts neural networks learn has been an active area of research. **?** developed SpRAy, an automatic pipeline for exploring shortcuts and biases learned by models, often referred to as "Clever Hans" effects **?**. **?** investigates methods for automatically finding spurious features in training data.

Recent work by **?** addresses the question of what concepts were learned by models and where in the training data they were present. However, **?** argues that automatically discovered concepts may lack atomicity and completeness.

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

2.7. Detection of Synthetic Images

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

Recent advances in synthetic image detection **???** have achieved near-perfect accuracy on images generated by GANs and diffusion models by analyzing frequency domain artifacts.

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

2.8. Gap in Current Literature

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

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

Chapter 3

Method

This chapter presents our theoretical framework and algorithmic approach for generating Invariants. We begin with formal definitions that relate our concept to classical level sets from differential topology **???**, establish our theoretical foundation, detail our algorithm, and conclude with implementation specifics and quality assurance measures.

3.1. Problem Formulation

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

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

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

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

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

3.2. Guided Iterative Optimization with Latent Diffusion Models

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

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

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

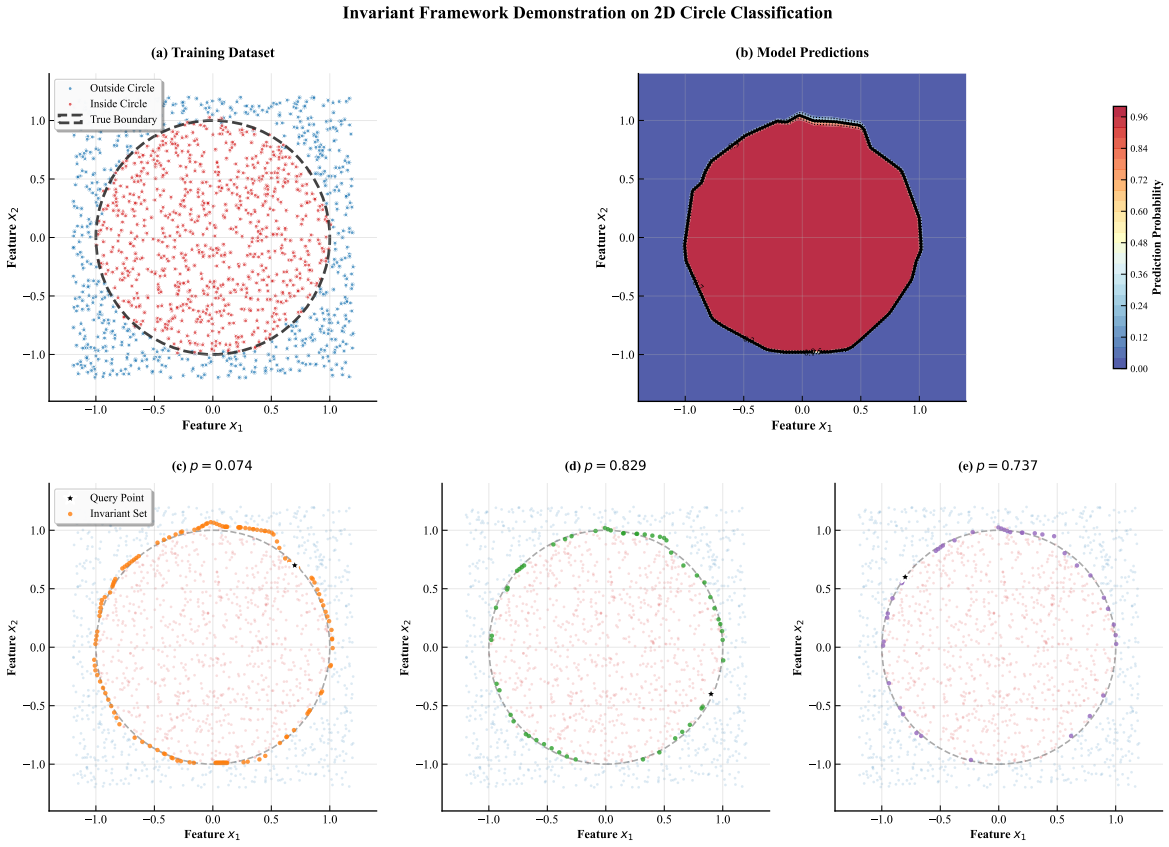

Figure 3.1: Demonstration of the Invariant Framework on a 2D Concentric Circles Dataset. (a) Training dataset with 1,500 samples classified by their position relative to a unit circle (dashed line). Blue points represent the outer class, pink points the inner class. (b) Learned decision boundary and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

3.2.1. Classifier Guidance Limitations

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

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

- **Latent space complications**: Modern diffusion models often employ the Latent Diffusion Model (LDM) approach **?**, which operates in a compressed latent space rather than directly on pixel values. This architectural choice introduces additional complexity when conditioning on neural network outputs: the classifier must evaluate encoded representations $\mathcal{E}(\mathbf{x}_t)$ at

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

3.2.2. Infinite Optimization Approach

Given these limitations, we adopt an *Infinite Optimization* strategy, specifically adapting Algorithm 1 from **?**. 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 **??**.

3.3. Quality and Realism Assurance

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

3.3.1. Frequency Domain Optimization

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

This frequency-aware approach ensures that:

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

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

Chapter 4

Experiments

Following recent work on monosemanticity and mechanistic understanding, we base our qualitative evaluation on

Chapter 5

Applications

Chapter 6

Discussion

Chapter 7

Conclusion

.1. Appendix

.2. Infinite Optimization Algorithm

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

Algorithm 1 Invariant Set Generation via Infinite Optimization

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

$z_T \sim \mathcal{N}(0, I)$ ▷ Draw starting latent
$target\_value = \mathcal{L}(\mathbf{x}^*)$ ▷ Store target invariant value
for $t = 1, \ldots, T$ **do** ▷ Initialize time step-dependent variables
    $C_t = \emptyset$ ▷ No conditioning (unconditional generation)
end for
$optim = \text{SGD}(z_T, \text{lr} = \eta)$ or $\text{Shampoo}(z_T, \text{lr} = \eta)$ ▷ 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

.2.1. Key Differences from Original Algorithm

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

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

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

.2.2. 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 and Shampoo optimizers **?** demonstrate 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.