Attribution Methods for Explainability of Predictive and Deep Generative Diffusion Models

By

Debela Desalegn Yadeta

Submitted to the School of Information Technology and Engineering

In partial fulfillment of the requirements for the degree of

Master of Science in Artificial Intelligence

Supervised by: Dr. Beakal Gizachew

College of Technology and Built Environment

Addis Ababa University

Addis Ababa, Ethiopia

June, 2025

APPROVAL

This is to certify that this thesis titled **"Attribution Methods for Explainability of Predictive and Deep Generative Diffusion Models"** is prepared by Debela Desalegn Yadeta and submitted in partial fulfillment of the thesis-option requirements for the Degree of Master of Science in Artificial Intelligence at the School of Information Technology & Engineering, College of Technology and Built Environment.

Name	**Signature**	**Date**
_____	_____	_____
(Advisor)		
_____	_____	_____
(External Examiner)		
_____	_____	_____
(Internal Examiner)		

ABSTRACT

As machine learning models grow in complexity and their deployment in high-stakes domains becomes more common, the demand for transparent and faithful explainability methods has become increasingly urgent. However, most existing attribution techniques remain fragmented, targeting either predictive or generative models, and lack a hybrid approach that offers coherent interpretability across both domains. While predictive modeling faces challenges such as faithfulness, sparsity, stability, and reliability, generative diffusion models introduce additional complexity due to their temporal dynamics, token-to-region interactions, and diverse architectural designs. This work presents a hybrid attribution method designed to improve explainability for both predictive black-box models and generative diffusion models. We propose two novel methods: **FIFA** (Firefly-Inspired Feature Attribution), an optimization-based approach for sparse and faithful attribution in tabular models; and **DiffuSAGE** (Diffusion Shapley Attribution with Gradient Explanations), a temporally and spatially grounded method that attributes generated image content to individual prompt tokens using Aumann-Shapley values, Integrated Gradients, and cross-attention maps. FIFA applied to the Random Forest, XGBoost, CatBoost, and TabNet models in three benchmark datasets: Adult Income, Breast Cancer, and Diabetes, outperforming SHAP and LIME in key metrics: +6.24% sparsity, +9.15% Insertion AUC, -8.65% Deletion AUC, and +75% stability. DiffuSAGE evaluated on Stable Diffusion v1.5 trained on the LAION-5B dataset, yielding a 12.4% improvement in Insertion AUC and a 9.1% reduction in Deletion AUC compared to DF-RISE and DF-CAM. A qualitative user study further validated DiffuSAGE's alignment with human perception. Overall, these contributions establish the first hybrid attribution methods for both predictive and generative models, addressing fundamental limitations in current XAI approaches and enabling more interpretable, robust, and human-aligned AI systems.

Keywords: Diffusion models, Explainable AI, Feature attribution, Integrated Gradients, Shapley values.

Acknowledgements

First and foremost, I thank God for His unfailing love, mercy, and grace. Through every challenge, moment of doubt, and period of exhaustion, His strength and presence sustained me. This work would not have been possible without His divine guidance, provision, and the peace He placed in my heart. To God be the glory for every step of this journey.

I express my heartfelt gratitude to my MSc advisor, Dr. Beakal Gizachew, for his exceptional mentorship and insightful guidance throughout this academic journey. His unwavering support and constructive feedback were vital to both my academic progress and personal growth. I am also deeply thankful to Dr. Adane Letta, Dr. Fantahun Bogale, and Dr. Natnael Argaw for their guidance, encouragement, and the lasting impact they have had on my development as a researcher.

My deepest appreciation goes to my brother Megersa, who has meant everything to me, a pillar of strength, wisdom, and unwavering belief. His sacrifices and care laid the foundation for my progress. I also warmly thank my beloved family: Hunde, Lense, Gadisa, Hawi, Bayise, and Eskedar. Your constant encouragement and presence were a true source of comfort and inspiration.

I am grateful to my friends Mr. Azmeraw Bekele, Mr. Robbel Habtamu, and Mr. Challa Bekabil for their valuable insights and encouragement, which helped refine this work. I also acknowledge Mr. Wesagn Dawit for his collaborative spirit and impactful idea-sharing, which enriched my academic experience. Lastly, I thank my friends and fellow students: Adisu, Bontu, Elbethel, Tigist, Mintesnot, Ermias, Migbar, and Hussein for the camaraderie, shared challenges, and motivation that made this journey more meaningful.

Dedication

This thesis is dedicated to the memory of my beloved mother, my first teacher, my greatest strength, and the deepest love of my life. You were not only my mother but also stood in the place of a father. You were my provider, my protector, my guide, my everything. It has been 2,075 days since I lost you, yet not a single day has passed without your presence in my heart. Your love gave me purpose. Your sacrifices gave me hope. Your prayers gave me wings. This achievement is not mine alone. It is yours. Every step I took was guided by the light you left behind. And if I had to walk this journey again, every sleepless night and every challenge, I would do it all again for you.

I made it, Mom. I did it for you.

Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

AI systems have achieved remarkable success across diverse domains such as healthcare [7], finance [8], justice [9], military [10], transportation [11], and education [12], driving innovation and transforming traditional workflows [13]. With performance levels rivaling or surpassing those of human experts in many tasks, AI continues to be increasingly integrated into organizational operations and decision-making processes [14].

However, as AI systems gain autonomy and influence in making high-stakes decisions, a pressing challenge emerges: Can we trust and understand how these systems arrive at their decisions? The opacity of modern machine learning models, especially deep learning architectures, has raised significant concerns about trust, safety, fairness, and accountability [15, 16]. This growing demand for transparency has led to the emergence of the field of XAI, which aims to make AI systems and their decisions more interpretable to human users. XAI has become a central concern in modern AI research due to the increasing reliance on complex machine learning models whose internal decision-making processes are not readily understandable. Although the term "Explainable Artificial Intelligence" was formally introduced by Van Lent et al. [17] in the context of simulation-based training systems, the concept of explainability predates the term itself. Researchers in the 1970s focused on building interpretable models such as rule-based expert systems and Bayesian networks [18, 19], valuing them for their transparency and logical structure.

Throughout the 1990s, researchers extended these efforts to neural networks, exploring ways to interpret their increasingly powerful but opaque behaviors [20]. The 2000s saw a surge of interest in explainability for recommendation systems, which were rapidly becoming integral to digital platforms [21, 22]. However, as the field shifted toward achieving superior predictive accuracy in the late 2000s, the emergence of deep learning and ensemble models led to the dominance of black-box systems. These models achieved

unprecedented performance but lacked transparency, introducing significant concerns in domains where decision-making affects human lives, such as healthcare, criminal justice, and finance [23]. As illustrated in Figure 1.1, there exists a fundamental trade-off between model performance and explainability. While rule-based and linear models are highly interpretable, they often fall short in predictive power. In contrast, deep learning and ensemble methods offer superior performance but are typically opaque, reinforcing the need for robust post-hoc explainability techniques.

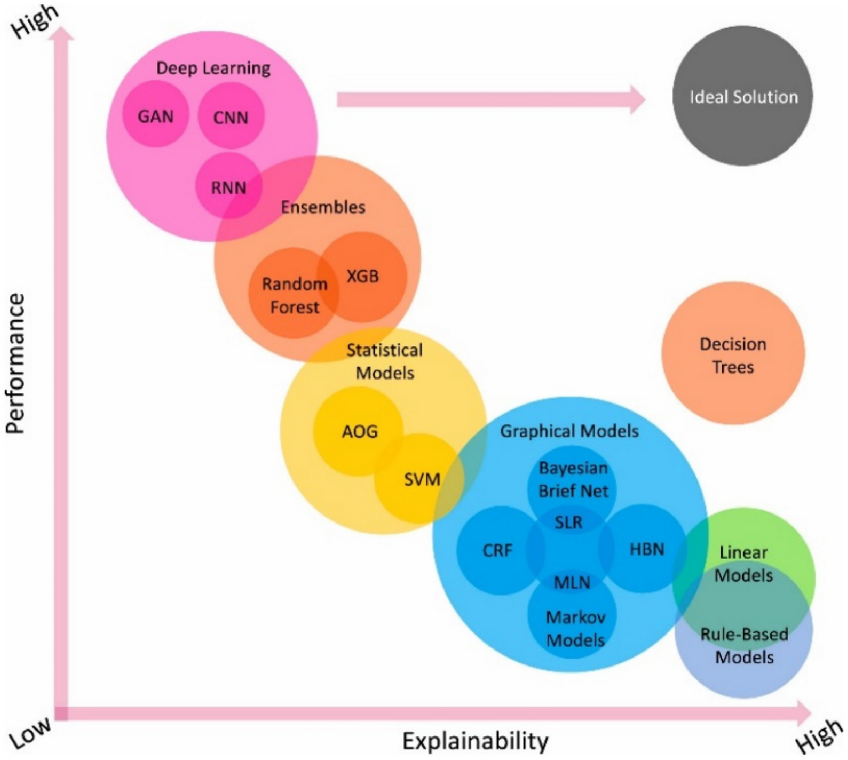


Figure 1.1: Trade-off between model performance and explainability across AI model types[1].

This opacity has given rise to growing concerns among the public, industry, and regulators. In particular, the European Union's General Data Protection Regulation (GDPR) [24] has established the right to explanation for individuals subject to automated decision-making, placing legal pressure on developers to design interpretable AI systems. Similar initiatives, such as DARPA's XAI program in the United States, have emphasized the importance of maintaining interpretability in high-performance systems, especially in mission-critical applications.

Despite the growing interest in explainability, the field of XAI remains multidisci-

plinary and lacks a universally accepted definition. As Anjomshoae et al. [25] note, interpretations of explainability vary significantly across disciplines such as machine learning, cognitive science, human-computer interaction, and software engineering. A widely cited definition by Guidotti et al. [26] describes explainability as the ability to make both the outputs and the underlying processes of an AI model understandable to humans. Explanations function as the interface between users and models, helping to bridge the gap between algorithmic logic and human reasoning. As AI continues to evolve and impact sensitive decision-making environments, the need for reliable, faithful, and human-centered explanations is more critical than ever.

1.1 Motivation

Despite the growing adoption of AI in critical sectors, the opacity of both predictive and generative models raises serious concerns about fairness, safety, and accountability. This study is motivated by the need for robust attribution methods that can explain decisions across model types, particularly in deep generative diffusion models.

A. Bias and Fairness Risks

Predictive models trained on historical or observational data often replicate and amplify structural inequalities [27, 28]. A well-documented example is the COMPAS tool, which disproportionately flagged Black defendants as high-risk [29]. Similarly, Obermeyer et al. [30] showed that healthcare models underestimated the needs of Black patients by relying on cost-based proxies. Facial recognition systems have also demonstrated higher error rates for individuals with darker skin tones [31]. These cases reveal how a lack of model transparency can conceal bias and undermine equitable outcomes [32, 33].

B. Trust, Safety, and Industrial Accountability

Beyond fairness, explainability is essential for diagnosing system failures in real-world deployments. For example, DeGrave et al. [34] found that several AI models trained for COVID-19 diagnosis relied on confounding image artifacts rather than clinical features.

In other domains, a 2021 Tesla Autopilot crash raised questions about how decisions are logged and interpreted by autonomous systems [35]. Similar challenges have surfaced in law enforcement and hiring, where proprietary AI systems have led to wrongful arrests [36] and alleged discrimination [37]. In industrial settings, opaque model behavior has even resulted in fatalities, as in the case of a misclassified warehouse worker [38]. These examples underscore the operational and ethical importance of transparent AI behavior.

C. Regulatory and Ethical Imperatives

Legal frameworks around the world are evolving to mandate explainability as part of responsible AI use. The GDPR enshrines a "right to explanation"[39], and newer proposals such as the EU AI Act[40] and U.S. AI Bill of Rights [41] further emphasize transparency, accountability, and nondiscrimination. However, fulfilling these obligations requires technical mechanisms that can expose model reasoning in a manner both faithful to the model and understandable to humans. Bridging this policy–technology gap is essential to meet legal standards and build public trust.

D. The Mystification of Generative Model Decisions

Generative diffusion models bring unique interpretability challenges. Unlike classifiers, which produce discrete outputs, diffusion models generate high-dimensional content, such as images or text, through multi-step denoising processes [42, 43]. This makes it difficult to trace how input prompts influence specific features or when structural decisions are made. Traditional XAI methods are ill-suited to capture the temporal and compositional dynamics of generation.

The risks of this opacity are increasingly evident. Luccioni et al. [44] found that diffusion models exhibit stereotypical associations in image generation without transparent mechanisms to detect or mitigate them. Carlini et al. [45] revealed that these models can memorize and regenerate copyrighted content, while Birhane et al. [46] linked toxic generations to biased training data. Without the ability to attribute visual elements to prompt tokens or diffusion steps, developers cannot reliably audit or control generative behavior. This highlights the urgent need for tailored XAI methods that make diffusion

models more transparent, controllable, and aligned with human intent.

1.2 Statement of the Problem

Explainable AI is increasingly critical for ensuring transparency and accountability in AI systems, particularly in high-stakes domains such as healthcare, finance, and law [26, 47]. Feature attribution methods play a central role in XAI by identifying which input features influence model predictions [48, 49, 50, 51]. However, these methods face distinct limitations depending on the underlying model architecture and nature of the task. For predictive models, main challenges include faithfulness [52, 53], sparsity [54, 55], stability [56, 57, 58], and reliability [59].

Among these, a key challenge for feature attribution methods is faithfulness, how accurately an explanation reflects a model's true reasoning. LIME [60] approximates local decision boundaries with linear surrogates but often fails in non-linear or complex regions [53]. SHAP [61], especially TreeSHAP, assumes feature independence and tree structure, which may not hold in practice [62]. KernelSHAP generalizes SHAP for any model but suffers from sensitivity to background data and sampling variance [58]. Refinements like OptiLIME [52] and Sig-LIME [53] improve performance in specific domains, yet a general, faithful attribution method across architectures continues to pose a significant challenge.

Another persistent challenge in explainable AI is achieving sparsity, the ability to highlight a minimal yet informative subset of features. Sparse explanations are more interpretable, especially in high-dimensional settings [63], but common methods like LIME and SHAP often produce dense outputs [60, 61]. LIME includes many features in local surrogates, while SHAP distributes scores across all coalitions. Heuristic approaches such as L1 regularization or thresholding can enforce sparsity but risk omitting meaningful features [64]. More principled methods include SEV [55], which optimizes for sparsity under monotonic assumptions, and DiCE [65], which generates sparse counterfactuals but may produce unrealistic examples. These methods highlight ongoing trade-offs between sparsity, fidelity, and semantic plausibility, underscoring the need for concise yet faithful explanations.

In addition, attribution methods often struggle with stability and reliability. LIME is sensitive to input perturbations and yields inconsistent outputs due to its reliance on random sampling [57]. KernelSHAP faces similar issues from stochastic coalition formation and dependence on background data [58]. Several variants have been proposed to address these issues. DLIME [59] introduces deterministic sampling strategies to enhance stability, although this comes at the cost of reduced local accuracy. ALIME [56] utilizes latent space representations to achieve more robust sampling, though this adds complexity and dependence on representation quality. GLIME [57] improves robustness by incorporating locality-aware sampling distributions but demands careful parameter tuning. ST-SHAP [58] seeks to enhance reliability by enforcing deterministic coalition selection, thereby reducing variance in explanation outputs. Among these, TreeSHAP [61] remains the most stable and reliable due to its model-specific, deterministic path tracing. Still, a general solution that ensures reliable attribution across diverse model types remains elusive.

In the context of text-to-image diffusion models, explainability presents unique challenges due to the stochastic and iterative nature of the generative process [43]. In generative diffusion models, core challenges include temporal attribution, token-to-region mapping, and architecture-specific reliability of attribution methods, as many existing techniques are limited to particular model classes [66, 67, 68]. Unlike predictive models that yield outputs in a single step, diffusion models progressively denoise latent variables to produce coherent images, embedding semantic content at varying stages of generation [69, 6]. Attribution techniques such as DF-RISE and DF-CAM [66] adapt saliency methods by visualizing spatial relevance either through perturbation-based approaches or activation maps from U-Net architectures. While these methods offer some temporal insight into when and where visual features are formed, they lack the ability to directly associate specific visual regions with corresponding prompt tokens.

To bridge the prompt-to-region attribution gap in diffusion models, recent methods, such as DAAM[67] and ConceptAttention [68], utilize attention-based mechanisms to align spatial heatmaps with individual tokens through cross-attention or transformer embeddings. However, they reduce temporal dynamics to static views and are limited to

transformer-based architectures, which hinders their applicability to U-Net-based models. Exponential sampling [66] adds temporal insight by identifying key denoising steps, but lacks token-level resolution and relies on output-level comparisons. These limitations highlight the need for a token-level, temporally-aware attribution method applicable across diffusion architectures.

Overall, current attribution methods struggle to provide explanations that are simultaneously faithful, stable, sparse, and reliable. These challenges are especially pronounced in generative models like diffusion architectures, where token-level influence and temporal dynamics remain underexplored. Existing approaches often lack consistency, fine-grained resolution, or semantic alignment, limiting their effectiveness across diverse AI systems. This work addresses these gaps by introducing hybrid attribution methods that deliver robust, interpretable, and temporally-aware explanations. Designed to operate across both predictive and generative settings, the proposed method enhances the reliability, generalizability, and practical value of explainable AI in real-world applications.

1.3 Research Question

RQ1: How can faithful, sparse, stable, and reliable feature attributions be generated for black-box predictive models?

RQ2: How can the contributions of individual prompt tokens to specific visual regions in generated images be faithfully identified and quantified?

RQ3: How do prompt tokens influence image synthesis throughout the diffusion timesteps?

1.4 Objective

1.4.1 General objective

To develop hybrid attribution methods that generate faithful and interpretable explanations for both predictive black-box models and text-to-image diffusion models.

1.4.2 Specific objective

- To develop a feature attribution method for black-box predictive models that produces faithful, sparse, stable, and reliable explanations.
- To quantify the contribution of individual prompt tokens to the image generated by text-to-image diffusion models.
- To identify the correspondence between individual prompt tokens and specific visual regions in text-to-image generation.
- To quantify how the influence of prompt tokens evolves throughout diffusion timesteps.
- To determine which timesteps in the diffusion process are the most influential in shaping the final output image.

1.5 Significance

Ensuring faithful, interpretable, and regulation-compliant explainability in predictive and generative AI models is essential, particularly in domains where trust, transparency, and accountability are critical. This study introduces two attribution-based explainability methods that address key methodological gaps and deliver practical value to a wide range of stakeholders:

For Decision-Makers in High-Stakes Domains

Clinicians, legal experts, and financial analysts need explanations they can trust. The predictive attribution method ensures faithfulness by reflecting the model's actual reasoning and supports sparsity by highlighting only the most influential features. This facilitates rapid decision verification, reduces cognitive overhead, and supports compliance with legal frameworks such as the GDPR [26, 70].

For Practitioners in Real-Time and Operational Settings

In dynamic environments such as autonomous systems, fraud detection, or patient monitoring, stability and reliability are essential. The method delivers stable attributions that

are resistant to small input changes and reliable outputs that remain consistent across runs and datasets, supporting robust decision-making under fluctuating conditions [71].

For AI Developers and Researchers

Effective model development and analysis depend on deep interpretability. The method provides time-step aware explanations, revealing when key generative decisions occur, and token-level attribution, which connects individual prompt tokens to specific outcomes. These tools assist in debugging, prompt engineering, and understanding generative dynamics [67, 66].

For Creative Professionals and Designers

Artists, content creators, and designers using generative AI benefit from greater transparency and control. By visualizing token-to-output mappings, the method enables iterative design, creative exploration, and fine-tuning of prompt inputs, making generative systems more interactive and intuitive [68].

For End-Users

Individuals impacted by AI outputs, such as patients, customers, or students, require explanations that they can understand. This method emphasizes human-centered interpretability, translating complex technical attributions into clear, actionable insights, aligned with GDPR requirements for accessible justification of automated decisions [60, 72].

For Regulators, Auditors, and Fairness Advocates

Policymakers, oversight bodies, and ethicists rely on tools that support transparency and fairness. The method facilitates bias and fairness auditing by identifying reliance on sensitive or proxy features, ensuring models align with non-discrimination principles and legal frameworks such as the GDPR and the EU AI Act[73, 74].

This work addresses the urgent need for explainability techniques that apply to both predictive and generative AI systems. By offering a human-centered approach, it advances

the practical deployment of AI in high-stakes, creative, and regulated domains, supporting trustworthy and transparent decision-making for a diverse range of stakeholders.

1.6 Scope

The scope of the research encompasses the development and evaluation of a unified attribution-based explainability method for predictive black-box models and deep generative diffusion models. The study focuses on post hoc interpretability techniques aimed at generating faithful, sparse, stable, and reliable attributions. For predictive tasks, the scope includes structured tabular data using models such as Random Forests, XGBoost, CatBoost, and Tabnet. For generative tasks, the study is limited to pretrained text-to-image models, with Stable Diffusion serving as the primary generative model for experimentation.

While the study includes statistical testing and human feedback for evaluation, it does not involve modifying or retraining the underlying models, also excludes interface design, deployment, or domain-specific customization. Subjective aspects such as user preferences or application-specific trust calibration are acknowledged but remain outside the scope. The focus is strictly on the computational and algorithmic aspects of attribution, evaluated through controlled experiments using established interpretability metrics and comparison with baseline methods.

1.7 Contribution

This work presents a hybrid attribution methods that advances the explainability of both predictive and generative diffusion models. Unlike most existing XAI approaches that focus exclusively on a single model type, this work addresses critical gaps across both domains. The proposed methods, FIFA and DiffuSAGE, deliver sparse, faithful, temporally aligned, and semantically grounded explanations. The main contributions are summarized as follows:

- **Firefly-Inspired Feature Attribution (FIFA):** We propose FIFA, a global optimization-based attribution method for predictive models. Inspired by the Fire-

fly Algorithm, FIFA identifies compact, high-impact feature subsets and achieves significant improvements over LIME and SHAP in terms of sparsity, faithfulness (Insertion and Deletion AUC), and stability, without compromising reliability.

- **Multi-Model Evaluation of FIFA:** FIFA is implemented across four model types: Random Forest, XGBoost, CatBoost, and TabNet, and evaluated on three benchmark datasets: Adult Income, Breast Cancer, and Diabetes. Results show +6.24% improvement in sparsity, +9.15% in Insertion AUC, –8.65% in Deletion AUC, and +75% in stability compared to baselines.

- **Diffusion Shapley Attribution with Gradient Explanations (DiffuSAGE):** We propose DiffuSAGE, a novel attribution technique for generative models that unifies Aumann-Shapley values, Integrated Gradients, and cross-attention maps to enable both spatial (token-to-region) and temporal (stepwise) attribution in Stable Diffusion. Empirical results on Stable Diffusion v1.5 (trained on LAION-5B) show a +12.4% improvement in Insertion AUC and a –9.1% reduction in Deletion AUC compared to DF-RISE and DF-CAM. A qualitative user study further validates the human-aligned interpretability of token-to-region mappings.

1.8 Thesis Structure

The subsequent sections of this document present the development of the proposed explainability method. Chapter 2 provides an extensive review of foundational concepts, including background knowledge, types of models, core explainability techniques, and prior work in the field. Chapter 3 outlines the research methodology, covering data acquisition, preprocessing, and the design and development of the proposed methods, FIFA and DiffuSAGE. It also presents the evaluation metrics used to assess attribution performance. Chapter 4 discusses the experimental setup and results, while Chapter 5 concludes the study by summarizing the key findings and offering recommendations for future research.

Chapter 2

Literature review

2.1 Background

Machine learning (ML) has undergone a remarkable evolution, transitioning from early rule-based systems to complex neural architectures capable of performing creative and perceptual tasks. Classical algorithms, widely adopted in the early decades, enabled automated decision-making in areas such as credit scoring, fraud detection, and diagnostics. Yet, their reliance on manually engineered features limited their flexibility in dealing with high-dimensional data such as images or text [75, 76].

The shift toward more sophisticated learning paradigms, including unsupervised and reinforcement learning, paved the way for models that could uncover structure in data or learn from interaction and feedback. The most transformative leap came with the rise of deep learning, where neural networks learned data representations automatically through stacked layers of abstraction [75]. These models, including convolutional and recurrent architectures, enabled breakthroughs in domains like vision and language, but also introduced new opacity. Unlike traditional statistical models, deep networks often operate as "black boxes", offering little insight into how decisions are made. This lack of interpretability has raised concerns around trust, accountability, and fairness, especially when such models are deployed in high-stakes settings [26, 77].

Adding to this complexity, the emergence of generative models marked a new paradigm: rather than predicting labels or outcomes, these models learn to simulate the data distribution itself. Popular approaches such as Generative Adversarial Networks (GANs) [78, 79], Variational Autoencoders (VAEs) [2], and Diffusion Models [6] now underpin advances in image synthesis, text generation, and even molecular design [43, 80, 81, 82]. However, as generative models become embedded in creative industries, research, and sci-

entific discovery, the challenge of understanding how these models generate outputs has grown more urgent. Explainability in this context is not only a matter of transparency, but also a key to controlling, steering, and improving generative behavior [83, 84].

Recent work in explainable AI has introduced tools aimed at interpreting generative systems. Techniques such as feature attribution, attention visualization, and latent space probing help trace which inputs affect outputs, or which internal components drive certain behaviors [66, 67, 68]. While still in their early stages compared to Predictive model explainability, these tools are essential for building trustworthy and interactive generative systems. As industries continue to adopt generative AI for design, media, healthcare, and more, the demand for interpretable, controllable models will only intensify. By making these systems more transparent, we can enable more responsible deployment, encourage creative exploration, and ultimately build generative technologies that are both powerful and human-aligned [26, 77].

2.2 From Predictive to Generative models

2.2.1 Predictive Algorithms

Predictive models are essential tools in data science and machine learning, enabling systems to forecast future outcomes based on historical data [85, 86]. These models identify patterns and relationships within datasets to make informed predictions. Their applications span diverse fields, including healthcare, finance, marketing, and engineering [86]. The central aim is to extract actionable insights from data that can guide decision-making [87]. Predictive models can broadly be classified into two major categories: regression and classification [86]. These categories differ in terms of the nature of the output variable they predict. While regression models predict continuous numerical values, classification models predict categorical outcomes [85]. Understanding the distinction is crucial for selecting the appropriate modeling technique based on the problem context.

Regression models aim to predict a continuous response variable by establishing a mathematical relationship between one or more independent variables and a dependent variable [87]. These models are widely used in economic forecasting, risk management,

and demand estimation. The accuracy of regression models is often evaluated using metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared [87]. Linear regression is the simplest and most interpretable regression technique [87]. It assumes a linear relationship between features and the target variable. Variants include multiple linear regression, ridge regression (with L2 regularization), and lasso regression (with L1 regularization) [88]. The elastic net combines both regularization methods to balance bias and variance.

When relationships between variables are nonlinear, models like polynomial regression or Support Vector Regression (SVR) are preferred [89]. Polynomial regression fits a higher-order curve to capture nonlinearities. SVR, a variant of Support Vector Machines, is robust to outliers and can model complex patterns using kernel functions [89]. Decision trees and ensemble-based methods are powerful nonlinear regression tools [90]. Regression trees split data recursively to reduce prediction error. Ensemble methods such as Random Forest Regression and Gradient Boosting Regression aggregate predictions from multiple trees, increasing robustness and accuracy. XGBoost and LightGBM are state-of-the-art gradient boosting algorithms commonly used in tabular data tasks [91].

Classification models are used when the output is categorical [85]. These models assign inputs to discrete classes and are applied in spam detection, fraud detection, medical diagnosis, and image recognition. Their performance is commonly evaluated using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) [92]. Logistic regression is a widely used classification algorithm for binary outcomes [86]. It models the probability that a sample belongs to a class using the logistic function. Naive Bayes is a probabilistic classifier based on Bayes' theorem and assumes feature independence. It remains effective in applications like text categorization and spam filtering [93].

Support Vector Machines (SVMs) are effective in both linear and nonlinear classification tasks [94]. They work by finding the optimal hyperplane that maximally separates different classes. Using kernel tricks, SVMs can handle complex and high-dimensional data structures [85].

Similar to regression, classification tasks benefit from decision trees and ensemble classifiers. Methods like Random Forest Classifier, Gradient Boosting Classifier, and tools like XGBoost and CatBoost improve predictive accuracy and robustness [95]. These models are particularly adept at handling categorical variables, missing data, and complex interactions among features.

In recent years, deep learning models such as Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) have demonstrated impressive performance in both regression and classification tasks [85]. These models automatically learn feature representations from raw data, making them suitable for image and speech recognition. Notably, TabNet has emerged as an interpretable and powerful deep learning model specifically designed for structured tabular data [96]. TabNet uses sequential attention mechanisms to focus on relevant features at each decision step. Its variants aim to improve computational efficiency and extend its application to semi-supervised and multi-task learning contexts.

2.2.2 Generative Models

Imagine a painter who has studied countless landscapes, portraits, and abstract pieces. Over time, they master the art of combining colors, shapes, and textures to create entirely new works that feel as if they belong in an art gallery alongside the originals. This is how generative models work in the world of artificial intelligence: they learn from vast collections of data, such as images, text, or audio, and then craft new outputs that mimic the patterns and nuances of what they've seen. These models, like creative minds, can generate photorealistic images, compose music, write stories, and even assist in scientific breakthroughs. Notable variants include Generative Adversarial Networks (GANs)[78, 79, 5, 97, 98, 99], Variational Autoencoders (VAEs)[2, 82, 100], and Diffusion Models[43, 69, 6, 101], each with unique approaches and capabilities. Different families of generative models employ different mathematical strategies and architectures to achieve this goal, with varying trade-offs in sample quality, training stability, and interpretability.

Variational Autoencoders (VAEs) are one of the earliest and most foundational ap-

proaches to probabilistic generative modeling [2]. As shown in Figure 2.1, VAE consists of an encoder that maps input data into a latent space and a decoder that reconstructs the data from this latent representation. Unlike traditional autoencoders, VAEs model the latent space probabilistically, typically assuming a Gaussian prior. During training, VAEs optimize a loss function that includes a reconstruction loss and a Kullback-Leibler divergence term, ensuring the latent variables approximate the prior distribution. This probabilistic formulation allows VAEs to generate new data by sampling from the latent space and decoding the samples.

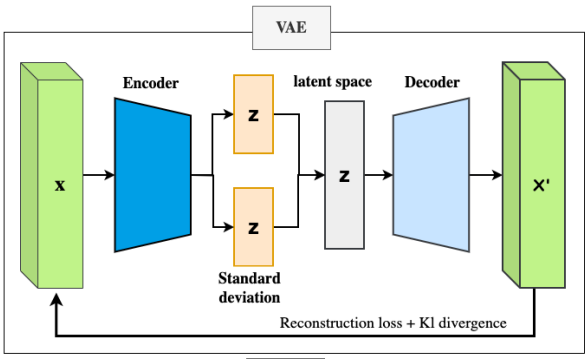


Figure 2.1: Variational Auto Encoder (VAE)[2]

One of the key strengths of VAEs is their structured latent space, which makes them interpretable and useful for tasks like interpolation and anomaly detection [100]. VAEs have been applied in domains such as image generation, drug discovery, and semi-supervised learning [82, 102, 103]. However, a notable challenge with VAEs is the generation of blurry images, especially when compared to other generative models like GANs [104]. This is due in part to the pixel-wise reconstruction loss and the regularization imposed by the KL divergence.

Flow-Based Models offer an alternative approach by constructing an invertible transformation between the input space and a latent space [3]. These models use a sequence of bijective functions to map data into a latent representation and back, ensuring exact likelihood computation, as shown in Figure 2.2. Notable examples include RealNVP, NICE, and Glow. The advantage of flow-based models lies in their tractable likelihoods and reversible mappings, which make them useful for both generation and inference.

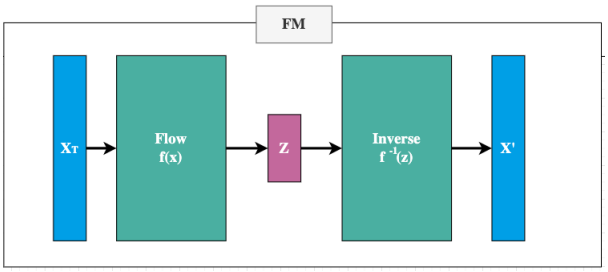


Figure 2.2: Flow Based Model [3]

These models shine in applications where likelihood estimation and exact inference are important. However, designing invertible and expressive transformations is non-trivial. In practice, flow-based models often require large model sizes to match the performance of GANs in sample quality [105]. Additionally, their architectural constraints can make them less flexible when compared to models like VAEs and GANs.

To address these limitations, autoregressive models offer an alternative generative strategy by decomposing the joint distribution into a sequence of conditional probabilities [4]. For example, in language modeling, an autoregressive model like GPT predicts the next word given the previous ones [106, 107]. Similarly, in image modeling, PixelRNN and PixelCNN generate pixels one at a time, conditioned on previously generated pixels [108]. These models are simple in concept and often produce high-fidelity outputs.

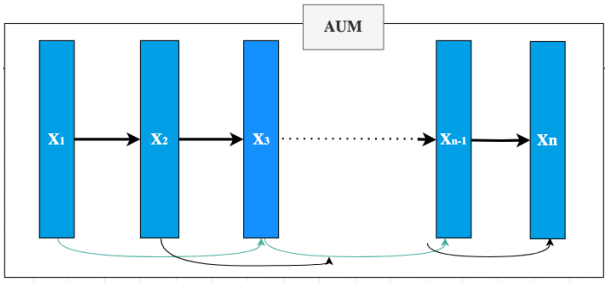


Figure 2.3: Auto regressive Model (AUM) [4]

One major advantage of autoregressive models is their likelihood-based training, which avoids some of the stability issues seen in adversarial approaches [108]. They are widely used in natural language processing, audio generation (e.g., WaveNet), and image generation [106, 109]. However, they suffer from slow sampling times, as they generate outputs sequentially [110]. Moreover, they lack an explicit latent representation, which limits their utility in representation learning and data compression [110].

Among the most influential advancements in generative modeling, Generative Adversarial Networks (GANs) represent a breakthrough in generative modeling by introducing an adversarial training paradigm [5]. As shown in Figure 2.4, GAN consists of a generator that creates fake data and a discriminator that tries to distinguish real from fake. The generator learns to produce increasingly realistic data to fool the discriminator, while the discriminator improves its classification ability. This dynamic, game-theoretic training process enables GANs to produce highly realistic samples.

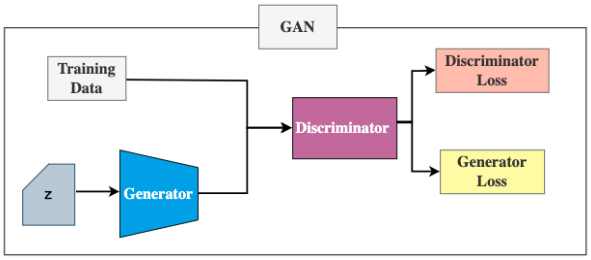


Figure 2.4: Generative Adversarial Network (GAN) [5]

As shown in the figure 2.4, the GAN optimization process is governed by the following objective function: Here, represents samples from the real data distribution, and denotes random noise sampled from a prior distribution, such as a Gaussian or uniform distribution. The generator maps to the data space, creating synthetic samples. The discriminator outputs the probability that is real to maximize the probability for real samples and minimize it for synthetic ones. During training, and iteratively optimize this loss function, striving to accurately classify inputs and striving to deceive.

Generative Adversarial Networks (GANs) have evolved significantly, resulting in various specialized variants that address specific challenges and applications. Conditional GANs (cGANs) enable controlled data generation by incorporating auxiliary information like labels or features [97]. Wasserstein GANs (WGANs) improve training stability by replacing the Jensen-Shannon divergence with the Wasserstein distance [99], while Progressive Growing GANs (PGGANs) enhance high-resolution image synthesis through gradual training [79]. StyleGAN introduces a disentangled latent space that allows fine-grained control over image attributes [78]. Moreover, CycleGAN targets unpaired image-to-image translation tasks by learning domain mappings [111].

Despite their success, GANs face persistent challenges such as mode collapse, where the generator outputs lack diversity [99], and training instability due to issues like vanishing gradients and hyperparameter sensitivity [98]. Evaluating GAN performance is also difficult, as existing metrics like Fréchet Inception Distance (FID) and Inception Score (IS) do not fully capture perceptual quality. Nevertheless, GANs have broad applications across domains: they are used in image synthesis, super-resolution [112], inpainting [113], deepfakes and style transfer, medical imaging [114], data augmentation [115], and scientific simulations, such as astronomical image generation and molecular structure synthesis for drug discovery [80].

Diffusion Models are a more recent and increasingly dominant class of generative models that learn to generate complex data distributions by iteratively denoising a sample initialized with random noise [116]. Inspired by non-equilibrium thermodynamics, diffusion models map data to latent representations through a forward diffusion process that incrementally adds noise to the data, and then reconstruct the original data through a reverse generative process. Originally introduced by Sohl-Dickstein et al. [117], these models have gained attention for their ability to generate high-quality images and other structured data. The recent advancements in diffusion models, such as Denoising Diffusion Probabilistic Models (DDPMs) by Ho et al. [6] and improved variants like DDIM [69], have demonstrated state-of-the-art results in image synthesis and beyond.

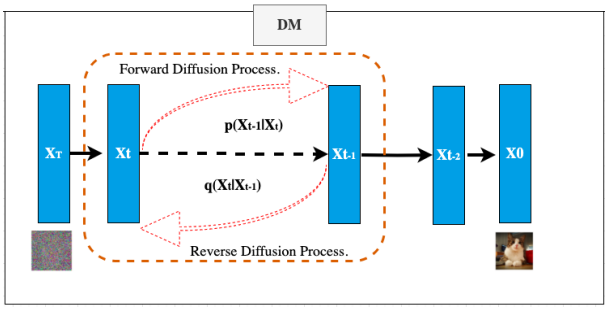


Figure 2.5: Diffusion Model [6]

Diffusion models generate data by learning to reverse a noising process that progressively corrupts an input, typically using Gaussian noise. In the forward process, a clean data point x_0, such as an image, is gradually noised over T steps, producing intermediate

states $x_1, x_2, \ldots, x_T$. This is achieved using a variance schedule β_t, with the transition defined as $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$. The cumulative noise across steps is captured by $\bar{\alpha}_t$, allowing for direct sampling of any intermediate noisy version x_t from x_0, improving computational efficiency.

The reverse process is modeled as $p_\theta(x_{t-1}|x_t)$, a Gaussian distribution whose parameters, mean μ_θ and variance Σ_θ, are predicted by a neural network trained to denoise the noisy sample at each step. The objective is to minimize the discrepancy between the actual noise and the noise predicted by the model, formalized as:

$$L = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right] \tag{2.1}$$

This learning process equips the model to reverse the diffusion and reconstruct high-fidelity data from random noise through iterative refinement.

While diffusion models excel at generating high-quality and diverse outputs, they face challenges such as computationally intensive training and slow sampling due to the large number of sequential steps required [118]. Their performance is also sensitive to the choice of noise schedules. Despite these limitations, diffusion models have seen rapid adoption across domains. They power advanced image generation tools like Stable Diffusion [119], contribute to speech synthesis [101], and are used in scientific applications like molecular generation and video synthesis [43]. Additionally, they are effective for inpainting and denoising tasks, demonstrating versatility in both creative and analytical contexts [43].

2.3 Explainability Techniques

Explainability techniques in AI help uncover how models make decisions, promoting transparency, trust, and accountability. These methods differ by data modality, model applicability (model-specific vs. model-agnostic), and their role in the ML lifecycle, before, during, or after training. Mechanisms include perturbation, gradients, rules, and surrogate models, with explanations ranging from global (model-level) to local (instance-level). Outputs range from feature importance and counterfactuals to visual or textual

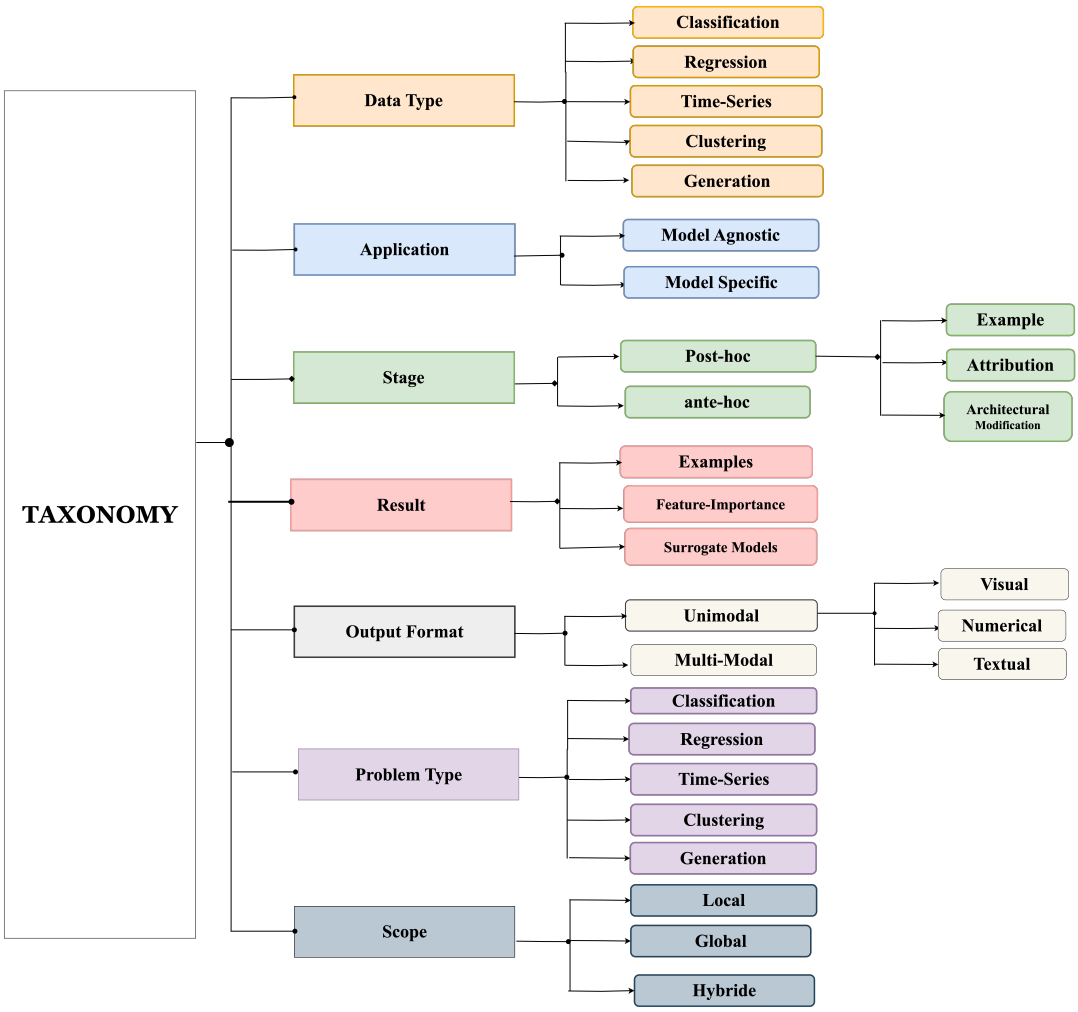


Figure 2.6: A Taxonomy of XAI Techniques

justifications. The following taxonomy organizes XAI methods for better method selection based on the use case.

2.3.1 Data Modalities in XAI

The type of data a model processes plays a central role in shaping which explainability techniques are appropriate. In tabular data, commonly found in domains such as finance, healthcare, and logistics, models such as random forests and neural networks offer high predictive accuracy but limited interpretability. To address this, model-agnostic techniques, including SHAP[61], LIME[60], and counterfactual explanations, are widely used to quantify feature influence. However, these methods often face challenges related to computational cost, high dimensionality, and fairness, particularly when trained on

biased datasets[120, 121].

In the vision domain, explainability is especially crucial in high-stakes areas such as medical imaging, where understanding model decisions is essential for clinical trust. Gradient-based techniques such as CAM[122] and Integrated Gradients[51], as well as perturbation-based methods including SHAP and LIME, are used to highlight relevant image regions. For domain-specific imagery, such as Optical Coherence Tomography (OCT), tailored interpretability approaches are often required [123, 124]. Hybrid strategies that combine multiple techniques can enhance robustness and produce more reliable explanations [125].

For textual data, explainability methods aim to improve transparency in models used for sentiment analysis, legal classification, and medical document processing. While transformer-based models such as BERT and GPT achieve strong performance, their decision processes are often opaque. Post-hoc tools such as LIME and SHAP provide word-level attribution, while attention mechanisms offer model-internal interpretability by highlighting which parts of the input were most relevant[126, 127]. Nonetheless, the context-dependent nature of language, idiomatic expressions, and potential embedded biases make consistent and reliable explanations difficult to achieve [70, 23].

Graph-structured data introduces unique interpretability challenges due to its relational and often dynamic nature. Graph Neural Networks (GNNs) are widely applied in areas such as social network analysis, molecular interaction modeling, and recommendation systems. Techniques such as GNNExplainer[128], GraphLIME[129], and attention-based approaches help identify influential nodes, edges, or subgraphs. Still, scalability remains a key concern, particularly for large or evolving graphs, and explanation fairness is an ongoing issue when graph structure reflects real-world biases[130, 131].

Time series data, prevalent in domains such as health monitoring, energy forecasting, and finance, requires XAI techniques that respect temporal dependencies. Deep learning models like LSTMs, GRUs, and transformers capture complex sequential patterns but often lack interpretability. Feature attribution methods, attention scores, and perturbation-based techniques such as SHAP and LIME have been adapted to high-

light influential time steps[132, 133]. However, challenges persist in explaining long-term dependencies, managing multivariate signals, and developing scalable and reliable evaluation metrics[134].

Finally, multimodal poses a complex challenge for explainability due to the non-linear and interdependent nature of cross-modal interactions. Modern models such as multimodal transformers and deep fusion architectures integrate diverse information streams, making it difficult to isolate the contribution of each modality. Attention-based visualization techniques and extensions of SHAP and LIME have been used to disentangle these contributions[135, 136]. However, scalability, interpretability across modalities, and the absence of standardized evaluation frameworks continue to hinder progress in this area [137].

2.3.2 Target Problems for Explainability

XAI techniques can be organized by the specific tasks they aim to support, as each type of machine learning problem presents distinct interpretability challenges. Classification, regression, time series forecasting, clustering, and generative modeling all require tailored explanation strategies to ensure model transparency and usability. This categorization helps researchers and practitioners match XAI methods to the nature of the task and the demands of the application.

In classification tasks, where the goal is to assign discrete labels to inputs, XAI is essential for demystifying black-box models such as deep neural networks. Post-hoc methods like SHAP[61], LIME[60], and Grad-CAM[138] are commonly used to quantify feature importance and visualize decision-making pathways. Prototype-based models and counterfactual explanations have also gained traction for making models more interpretable by design[139, 140]. However, balancing interpretability and performance remains difficult. Post-hoc methods can sometimes offer misleading explanations, and their quality is often subjective and hard to evaluate[141, 142]. Furthermore, ethical concerns such as exposing embedded biases emphasize the importance of responsible implementation[143].

For regression tasks, which involve predicting continuous values, interpretability chal-

lenges differ. While traditional models like linear regression are inherently explainable, modern approaches such as support vector regressors, decision trees, and neural networks trade transparency for performance. SHAP, LIME, partial dependence plots (PDPs)[90], and ICE plots[144] have been adapted for regression, enabling insight into how features influence predictions. Counterfactuals and monotonic neural networks offer ways to integrate transparency into model design[145, 146]. However, interpreting continuous outputs across a wide range of values is inherently harder than explaining categorical predictions. Evaluating explanation quality, computational overhead, and fairness in output predictions remain persistent challenges [147, 148].

In time series modeling, where inputs are sequential and temporally dependent, the opacity of deep learning models such as LSTMs and transformers presents serious risks in critical applications like healthcare and autonomous systems. XAI aims to provide visibility into how these models use temporal patterns to make predictions. Tools like SHAP, DeepLIFT, LRP, and attention maps have been extended to this domain, helping identify which time steps or features influence forecasts [149, 150]. In finance, for instance, XAI improves trust in high-risk decisions by revealing model reasoning[151]. However, many methods originally designed for image or text tasks do not directly transfer to time series. Unique evaluation frameworks that account for temporal dynamics are needed to validate the reliability and faithfulness of explanations.

Clustering tasks, as unsupervised learning problems, bring distinct explainability challenges. Since labels are not provided, interpretability must focus on understanding why specific data points are grouped together. Approaches such as ClAMP, which uses rule-based prototypes, and Shapley value-based scoring have been introduced to make clustering outcomes more interpretable and actionable [152, 153]. These methods are particularly valuable in domains like healthcare or manufacturing, where stakeholder trust depends on transparent, meaningful groupings. Despite progress, challenges remain in representing complex cluster structures, evaluating explanation quality, and ensuring explanations generalize across domains [154, 155].

Generative tasks, such as image synthesis, text generation, and data augmentation,

present some of the most difficult challenges for explainability. Deep generative models like GANs, VAEs, and diffusion models operate in high-dimensional latent spaces, making their behavior inherently opaque. XAI techniques in this space aim to trace how inputs map to outputs and how latent variables influence the generated content. Tools like LRP and attention maps offer partial transparency by highlighting which parts of the input drive generation[156]. Recent work also explores the interpretability of latent dynamics in diffusion models. However, explaining generative processes remains difficult due to their stochastic nature, sensitivity to small changes, and computational demands. Additionally, explanations must be adapted to their context—clinical image generation, for instance, requires alignment with domain knowledge, whereas creative applications like music or art demand more subjective interpretability criteria[157, 15]. Addressing these challenges calls for integrated architectures, better evaluation metrics, and domain-specific explanation frameworks.

2.3.3 Model Applicability of XAI Methods

Explainability methods can be broadly categorized based on their applicability to machine learning models: model-specific methods, which are tailored to particular architectures, and model-agnostic methods, which can be applied across a range of models regardless of their internal structure. This classification helps guide the selection of appropriate XAI techniques depending on the model type, performance requirements, and interpretability goals.

A. Model Specific

Model-specific XAI methods are designed with detailed knowledge of a model's internal mechanisms, allowing them to produce highly precise and computationally efficient explanations. These methods are often faster at inference time, as they leverage the structure of the underlying model rather than approximating it, as in model-agnostic techniques[158]. For example, in time series classification, DEMUX generates class-specific saliency maps that highlight the distinctive patterns contributing to a model's prediction [159]. In NLP, model-specific techniques can dissect individual components of deep architectures, such

as word embeddings, attention layers, or recurrent units, providing granular explanations that are well-aligned with how the model operates [160]. Despite their advantages in speed and fidelity, model-specific methods are inherently limited in scope. They must be redesigned or retuned for each model class, which can restrict their generalizability and scalability across applications. Nevertheless, their performance and tailored insights make them valuable in settings where computational efficiency and detailed model understanding are priorities [161].

B. Model Agnostic

Model-agnostic XAI methods aim to explain predictions without requiring access to the internal workings of the model. These approaches treat the model as a black box and infer feature importance or input-output relationships through systematic perturbations or surrogate modeling. Techniques such as LIME[162] and SHAP[61] exemplify this paradigm, offering flexibility and broad applicability across various models. Their utility extends to critical domains such as healthcare, where model-agnostic tools enhance the interpretability of risk scores for clinicians[163], and environmental modeling, where feature shuffling and occlusion analysis are used to identify influential variables[161]. However, model-agnostic methods come with limitations. Because they approximate the model's behavior from the outside, they are often more computationally intensive and may produce less faithful explanations. Moreover, they can be prone to misinterpretation, especially when users conflate correlation with causation or overlook feature dependencies[164, 165]. As a result, careful application and contextual understanding are essential to avoid misleading conclusions.

Figure 2.7 summarizes the distinction between model-specific and model-agnostic XAI techniques within the context of generative models. Model-specific approaches are aligned with particular generative architectures, such as variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models, whereas model-agnostic techniques offer general-purpose insights using perturbation, attention, or feature attribution strategies applicable across architectures.

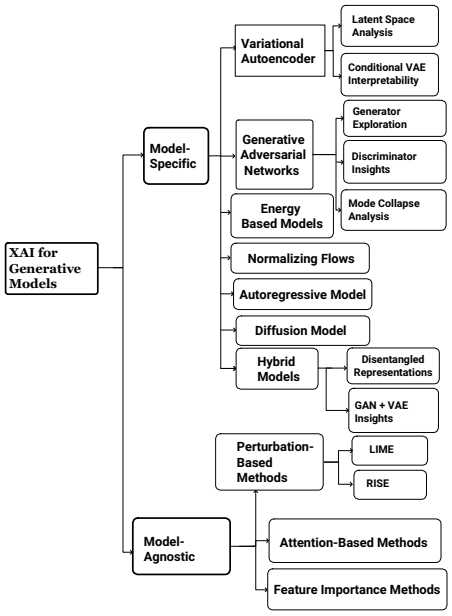


Figure 2.7: A Classification of XAI Methods for Generative models

2.3.4 Integration Stage of XAI Techniques

The integration stage of XAI techniques refers to the point in the model lifecycle at which explainability is introduced, either during model development or after the model has been trained. Depending on the timing and method of integration, XAI techniques are broadly categorized into ante-hoc and post-hoc approaches. This distinction is crucial for selecting appropriate methods based on application-specific needs, such as performance, interpretability, or regulatory compliance.

A. Ante-Hoc Explainability

Ante-hoc methods incorporate interpretability directly into the model architecture from the outset. These techniques aim to construct inherently transparent models, either through explicit reasoning mechanisms or by using simple, interpretable structures. Ante-hoc interpretability can be described in three levels: simulatability, decomposability, and algorithmic transparency [15]. Simulatable models, like small decision trees, allow a human to fully understand the model's behavior. Decomposable models offer interpretability at the component level, such as interpretable inputs or parameters, while algorithmically transparent models are based on well-understood learning rules but lack

human-readability in complex scenarios. Although ante-hoc methods ensure faithful explanations, they are usually restricted to simpler models, such as decision trees, linear models, and rule-based systems, potentially limiting predictive performance in complex tasks[166].

B. Post-Hoc Explainability

Post-hoc methods, in contrast, are applied after the model has been trained. These techniques do not require modifying the model's architecture and are often used to interpret complex black-box models. Post-hoc approaches include feature importance scores, counterfactual explanations, saliency maps, and surrogate models that approximate the original model's behavior. Their flexibility makes them suitable for a wide range of models and applications. In addition, they are often more practical to implement since they do not require changes to existing high-performing systems. However, post-hoc explanations may not always be faithful to the actual reasoning of the model, which can lead to misleading interpretations if not used carefully.

Table 2.1: Comparison of ante-hoc and post-hoc explainability approaches based on desired properties

Stages	**Performance**	**Faithfulness**	**Additional Storage**	**Additional Time**	**Model Training**	**Model Modification**	**Flexibility**	**Bias Detection**	**Development Cost**
Ante-hoc	Low	High	X	X	✓	✓	X	X	Expensive
Post-hoc	High	Low	✓	✓	X	X	✓	✓	Cheap

As shown in Table 2.1, the choice between ante-hoc and post-hoc methods involves a trade-off. Ante-hoc approaches typically offer higher faithfulness but come with limitations in model flexibility and performance. Post-hoc methods, while easier to implement and scalable to complex models, may compromise the fidelity of explanations. Therefore, the selection of an XAI approach should be informed by the application's goals, whether

prioritizing interpretability and trust or maximizing model performance and generalizability.

2.3.5 Mechanisms of XAI Methods

Understanding the mechanisms that underpin different XAI techniques is fundamental to improving model transparency, trust, and usability. Broadly, explainability methods operate through one or more of the following mechanisms: architecture modification, example-based reasoning, and attribution. Each offers distinct strategies for uncovering how models arrive at their decisions.

A. Architecture Modification

Architecture modification involves designing models with interpretability embedded directly into their structure. These methods adjust the underlying architecture either through simplification or principled incorporation of domain knowledge to enhance both transparency and performance. For instance, ToyArchitecture introduces simplified, hierarchical structures combining unsupervised learning with symbolic and sub-symbolic integration to maintain interpretability across tasks [167]. Interpretable Neural Networks (INNs) incorporate rule-based reasoning to inject domain-specific knowledge into deep architectures, improving both decision reliability and human understanding[168]. Another approach, Multi-Objective Neural Architecture Search (NAS), uses optimization techniques like NSGA-II to balance predictive performance with interpretability during model design[169]. Similarly, heuristic-based construction methods apply rule-guided model selection to support the development of transparent systems[170]. These techniques aim to preserve model introspectability without compromising effectiveness.

B. Examples

Example-based methods explain model behavior through representative data instances. Instead of summarizing feature contributions, these approaches provide tangible examples that illustrate how and why the model made a particular decision. This mechanism aligns closely with human reasoning, using concrete cases to understand abstract behav-

ior. It is especially effective in domains like bioinformatics or legal AI, where specific case studies can enhance stakeholder comprehension and trust[171]. Common example-based techniques include counterfactuals[72], adversarial examples[172], prototypes and criticisms[173], and influential instance detection. These methods are generally model-agnostic and particularly well-suited to tasks involving visual or textual data. However, for high-dimensional tabular data, summarizing or selecting meaningful instances remains a challenge unless dimensionality reduction or instance aggregation is used.

C. Attribution

Attribution techniques aim to quantify how different parts of the input influence the model's output. These mechanisms can be classified into three categories: perturbation-based, meta-explanation, and propagation-based methods.

i) Perturbation-Based Techniques: These methods evaluate how changes to input features affect the model's prediction. Examples include occlusion analysis[174], prediction difference analysis (PDA)[175] and meaningful perturbations[176]. While they are flexible and model-agnostic, they are also computationally expensive due to repeated evaluations. Gradient-based alternatives like Sensitivity Analysis (SA)[177] offer efficiency but often suffer from issues such as gradient shattering, which undermines explanation stability.

ii) Meta-Explanations: Meta-level techniques analyze explanation methods themselves, offering insights into their behavior, fidelity, and consistency across contexts. Techniques like Spectral Relevance Analysis (SpRAy)[178] cluster heatmaps to uncover model-wide behavior, while network dissection[179, 180] interprets learned representations by identifying semantic concepts encoded in hidden units. These approaches help users understand what the model has learned, not just how it behaves.

iii) Propagation-Based Methods: These techniques trace how information flows through a model from input to output. Layer-wise Relevance Propagation (LRP) is a prominent example, redistributing predictions back through the layers of a network to highlight important features[181, 182]. LRP has been extended to recurrent models[183] and adapted to clustering and anomaly detection tasks through neuralization[184]. Unlike

basic gradient-based methods, LRP offers stable, efficient explanations with just one forward and backward pass and avoids pitfalls like explanation discontinuity[185].

Other propagation approaches include Deconvolution[174] and Guided Backpropagation[186], which focus more on visualizing activation patterns than tracing decision causality. Recent heuristics and optimization-based propagation variants, such as Grad-CAM[138] and PatternAttribution[49], further extend the propagation paradigm. The iNNvestigate toolbox[187] provides implementations for many of these methods, making them accessible for practical analysis of neural networks.

2.3.6 Granularity of Explanations

The granularity of explanations in Explainable AI (XAI) refers to the level of detail at which a model's decision-making process is interpreted. This distinction is critical because different users and use cases require varying depths of insight. In general, XAI techniques operate at either the local, global or hybrid level of granularity. Each serves a distinct purpose and contributes to a more comprehensive understanding of how models behave.

A. Local

Local explanations focus on individual predictions, offering insights into why a model produced a specific output for a particular input. These methods are especially valuable in high-stakes applications such as medical diagnosis or credit approval, where understanding the rationale behind a single decision can directly impact outcomes. Techniques like LIME and SHAP are widely used for local explanation, attributing feature importance to a model's prediction at the instance level [50, 188]. Because they are intuitive and case-specific, local explanations are particularly appealing to non-expert users and are often preferred in domains where transparency for individual decisions is required [188].

B. Global

Global explanations, by contrast, aim to uncover how a model behaves across the entire input space. These methods are useful for developers and domain experts who seek to

understand the model's learned patterns, decision boundaries, and feature interactions at a holistic level. Techniques such as PCAIME (Principal Component Analysis-Enhanced Approximate Inverse Model Explanations) exemplify global explanation approaches by summarizing global feature importance and highlighting overall model behavior [189]. Global methods are essential for validating whether a model adheres to ethical principles, fairness constraints, or domain expectations.

C. Hybrid

Hybrid approaches aim to integrate the strengths of both local and global explanations, providing a more nuanced understanding of model behavior. For example, Concept Relevance Propagation (CRP) bridges local relevance attribution with global concept analysis, helping users interpret both "where" and "why" a model made a prediction [50]. PCAIME also fits into this category by enabling visual interpretation of both individual and aggregated feature contributions. Such techniques help address the limitations of relying solely on local or global views.

However, one of the persistent challenges in XAI is the inconsistency or disagreement among explanation methods. Different techniques may yield divergent results for the same model and input, making it difficult to determine which explanation is most reliable. To address this, methods like Functional Decomposition (FD) offer region-based explanations that aim to minimize conflicting feature attributions by reducing local feature interactions [190]. Furthermore, there is growing recognition of the need for more human-centered explanation strategies that consider user diversity, domain context, and interpretability goals [191].

2.3.7 Explanation Result

The result-based approach to explainability focuses on using the output of an explanation method as the primary component for interpreting model behavior. These methods are typically grouped into three main categories: Feature Importance, Surrogate Models, and Examples.

A. Feature Importance-Based Methods

Feature importance-based methods quantify and visualize the influence of input features on a model's predictions. These methods are essential for identifying which variables most significantly affect outputs, thereby enhancing transparency and trust.

Various visualization techniques assist in interpreting feature importance. Partial Dependence Plots (PDPs) show the average marginal effect of a feature, while Individual Conditional Expectation (ICE) plots reveal individual instance-level responses [144]. Sensitivity Analysis (SA) measures changes in outputs caused by perturbing inputs [192]. In image classification, saliency maps highlight pixel-level importance using gradients or occlusion [193].

Relevance propagation techniques, such as Layer-wise Relevance Propagation (LRP), provide pixel-wise or feature-wise contributions by tracing model outputs backward through the network [181]. LRP ensures conservation of relevance across layers and is often visualized via heatmaps.

SHAP [61] is another widely used method for local feature attribution. It uses Shapley values from cooperative game theory to assign fair importance scores across all possible feature combinations. Variants of SHAP improve scalability and expand utility: L-Shapley and C-Shapley reduce computation costs [194]; NeuronSHAP [195] evaluates neural unit relevance; and DataSHAP [196] ranks training examples by their impact on the model.

B. Surrogate Models

Surrogate models are interpretable models that approximate the behavior of complex black-box systems. These are useful when direct interpretation is impractical.

LIME [60] is a widely used method that fits a local linear model around a prediction to approximate the black-box decision boundary. Its variants include K-LIME [197], which uses k-means to partition input space; LIME-SUP [198], which incorporates supervised tree-based partitioning; and NormLIME [199], which aggregates normalized surrogates to yield global class-level insights.

Symbolic surrogates provide an alternative by constructing global symbolic representations. Shih et al. [200] introduced symbolic decision diagrams (ODDs) to explain Bayesian classifiers. Albini et al. [201] used influence graphs to capture relationships between inputs and outputs. Ignatiev et al. [202] proposed abductive reasoning with formal guarantees to ensure logically sound explanations. These methods contribute structured and verifiable insights into model behavior.

C. Example-Based Methods

Example-based explanations clarify predictions by referencing representative or similar instances from the dataset. These approaches help users understand decisions through analogies, which align with human reasoning.

Doshi-Velez et al. [203] proposed topic models using concept-word examples for topic interpretation. Bien and Tibshirani [204] formulated prototype selection as a set-cover problem to extract minimal representative subsets. Kim et al. [173] introduced MMD-critic, an algorithm that selects both prototypes and criticisms to enhance interpretability by showing what the model considers representative and non-representative.

These instance-based methods are particularly useful when inputs can be naturally understood by users, such as images, text, or structured tabular records, and they play a growing role in human-centered explainable AI.

2.3.8 Explanation Output Modalities

Explanation outputs can be categorized into two primary types: uni-modal and multi-modal, each tailored to deliver model insights most effectively based on user requirements and the characteristics of the data. These modalities are critical for facilitating a deeper understanding of model behavior, ensuring that users receive explanations that align with their specific needs and the context in which the model is applied. By selecting the appropriate output format, explanations can be optimized for clarity, accessibility, and utility across different application domains.

A. Uni-modal

Uni-modal explanations deliver model insights through a single output format—visual, numerical, or textual—based on specific user needs and data types.

Visual explanations offer intuitive and accessible insights into complex model behavior through graphical representations. These explanations are particularly effective at highlighting key input features that drive a model's predictions, such as salient image regions via heatmaps or relevant words in text. Visualizations can also illustrate internal model mechanisms, such as neural network layers and their connections. The strength of visual formats lies in their ability to simplify complex dynamics, making them understandable not only to technical users but also to non-specialists, thereby broadening accessibility and trust.

Numerical explanations provide precise, quantitative insights into how input features influence model predictions. These are typically represented as values, vectors, or matrices. They allow users to assess the relative importance of various inputs, supporting detailed evaluation of the model's decision-making process. Techniques like Concept Activation Vectors (CAVs) and other model-agnostic tools often employ numerical formats to convey information concisely. However, numerical outputs may be less intuitive than visual or textual formats and are best suited for users with a strong technical background. To improve interpretability, numerical explanations are often integrated with complementary formats in comprehensive explanation pipelines.

Textual explanations use natural language to articulate the reasoning behind a model's decisions. These are typically descriptive statements highlighting which features influenced the outcome and how. Textual formats are highly effective in bridging the gap between complex AI systems and users, especially those without technical expertise. The clarity, accuracy, and relevance of the language used determine the effectiveness of textual explanations. When paired with visual or numerical outputs, textual explanations enhance understanding by translating technical insights into accessible narratives.

B. Multi-modal

Multi-modal explanations combine multiple output formats—such as numerical, visual, and textual—to create a more holistic and accessible understanding of model behavior. This integration leverages the strengths of each modality to accommodate a broader range of user needs and preferences.

For instance, methods like Functional ANOVA decomposition generate both numerical metrics and visual plots to illustrate variable interactions and their influence on predictions. Justification Narratives combine bar charts with explanatory text, offering detailed yet accessible breakdowns of classification results. These hybrid formats are particularly beneficial in collaborative or interdisciplinary settings, where users may vary in technical expertise. By aligning detailed technical information with user-friendly narratives, multi-modal explanations improve transparency, support decision-making, and build user trust in AI systems.

2.4 Related work

In the field of machine learning, interpretability techniques are crucial for understanding and trusting complex models. These methods help provide insights into how a model arrives at a specific decision, especially when the model is a black box, like deep learning models or ensemble methods. A variety of techniques have been developed to address this need, each with its unique approach and use cases. Among these, LIME, SHAP (including KernelSHAP, TreeSHAP, and DeepSHAP), Integrated Gradients, Saliency Maps, LRP, and CAM are some of the most widely used methods, offering different ways to interpret the behavior of machine learning models [162, 205].

LIME (Local Interpretable Model-agnostic Explanations) is one of the pioneering methods in model interpretability [162]. It operates by approximating a complex model's decision boundary locally using simpler, interpretable models, such as linear regression or decision trees. This technique works by perturbing the input data and observing the corresponding changes in the model's predictions. LIME then fits a surrogate model to this perturbed data, explaining the model's behavior in a local context. This makes

LIME highly versatile, as it can be applied to any model, regardless of its complexity or architecture.

Building on the Shapley values from game theory, KernelSHAP offers an alternative model-agnostic approach for explaining predictions [205]. Shapley values assign a contribution score to each feature based on its influence on the prediction, considering all possible combinations of features. KernelSHAP computes these values using a kernel-based approximation, allowing for a more efficient calculation of Shapley values in the case of complex models. This method is particularly useful for obtaining a global understanding of feature importance and understanding how individual features contribute to the model's predictions, especially when the model is too complex for traditional analytical methods.

TreeSHAP, on the other hand, is specifically optimized for tree-based models like decision trees, random forests, and gradient boosting machines [205]. While KernelSHAP applies to any model, TreeSHAP takes advantage of the structure of tree-based algorithms to compute exact Shapley values more efficiently. By evaluating the marginal contribution of each feature at each decision node, TreeSHAP can quickly and accurately determine the importance of features in decision-making. This specialization makes TreeSHAP particularly well-suited for understanding tree-based models, which are commonly used in practice due to their interpretability and performance.

For deep learning models, more sophisticated methods are required due to their complexity. DeepSHAP is an extension of SHAP designed specifically for deep learning models [206]. By combining the principles of Shapley values with techniques tailored to neural networks, such as layer-wise relevance propagation (LRP), DeepSHAP can explain the output of complex neural networks by determining the contribution of each input feature across different layers of the model. This allows for a more granular understanding of how inputs are processed through the layers of a deep neural network, making it possible to interpret the results even in models with millions of parameters.

In addition to these Shapley-based methods, several other techniques focus on visualizing which parts of an input are most influential in the model's decision-making process.

Integrated Gradients, Saliency Maps, and Layer-wise Relevance Propagation (LRP) are commonly used for image-based tasks [207, 208, 181]. Integrated Gradients measure the importance of each feature by computing the gradient of the model's output with respect to the input and integrating it along a path from a baseline to the input. Saliency Maps use the gradients to highlight areas in the input that have the most influence on the model's prediction. LRP, on the other hand, propagates the relevance of the model's output backward through its layers, attributing importance to each neuron in the network. For convolutional neural networks (CNNs), Class Activation Mapping (CAM) identifies regions in an image that are most responsible for a particular prediction by weighting the feature maps of the final convolutional layer [209]. These methods offer intuitive visualizations that are particularly helpful in understanding how deep learning models interpret images and text.

To evaluate the practical effectiveness of XAI methods, it is important to consider a range of desirable properties: *stability*, *faithfulness*, *sparsity*, *reliability*, and *computational complexity* [26]. These properties help characterize the quality and usability of explanations, which in turn affect user trust and decision-making. As shown in Table 2.2, this section explores how representative XAI techniques perform across these dimensions and why. We begin with stability, which refers to the consistency of an explanation when small perturbations are applied to the input or model. As shown in Table 2.2, TreeSHAP excels in this property because it leverages a model-specific, deterministic algorithm that exactly computes Shapley values for tree-based models [210]. Similarly, Integrated Gradients (IG) and Layer-wise Relevance Propagation (LRP) provide stable explanations, owing to their reliance on structured gradient backpropagation and conservation principles, respectively [51, 181]. In contrast, LIME and KernelSHAP generate local surrogate models by randomly sampling perturbed inputs, introducing stochasticity that leads to explanation variability [60, 61]. Saliency maps and DeepSHAP also lack stability due to their dependence on noisy gradients and model approximations [211].

Table 2.2: Comparison of prominent XAI techniques based on key interpretability criteria.

XAI	PRE	GEN	STB	FTH	SPR	REL	CCX	TA	TL	TRM
LIME [162]	✓	×	×	×	✓	×	✓			
KernelSHAP [205]	✓	×	×	✓	✓	✓	×	-	-	-
TreeSHAP [205]	✓	×	✓	✓	×	✓	✓	-	-	-
DeepSHAP [206]	✓	×	×	×	×	×	×	-	-	-
IG [207]	✓	×	✓	✓	×	✓	✓	-	-	-
Saliency Maps [208]	✓	×	×	×	×	×	✓	-	-	-
LRP [212]	✓	×	✓	✓	✓	×	✓	-	-	-
CAM [209]	✓	×	×	✓	✓	×	✓	-	-	-
Diffusion XAI [116]	×	✓	✓	✓	×	✓	✓	✓	×	✓
DAAM[67]	×	✓	✓	✓	×	✓	×	✓	✓	✓
Concept Attention [213]	×	✓	✓	×	✓	✓	×	×	✓	✓

Key:

XAI: *Explainable AI Technique*, PRE: *Predictive*, GEN: *Generative*, STB: *Stability*, FTH: *Faithfulness*, SPR: *Sparsity*, REL: *Reliability*, CCX: *Computational Complexity*, TA: *Timestep Aware*, TL: *Token Level*, TRM: *Token-to-region mapping*

While stability ensures consistency, faithfulness—the extent to which an explanation accurately reflects the model's decision-making process—is arguably the most critical property. TreeSHAP and KernelSHAP perform well here because they are grounded in cooperative game theory, ensuring that the contribution of each feature aligns with the model's output [61]. IG and LRP are also highly faithful as they trace back model decisions via gradients or layer-wise relevance scores, maintaining fidelity to the learned function [51, 181]. On the other hand, LIME often struggles with faithfulness, as its linear surrogate models oversimplify complex non-linear decision boundaries [60]. Saliency maps and DeepSHAP may produce visually appealing outputs but often suffer from gradient saturation and class insensitivity [214].

The third property, sparsity, is key to generating explanations that are easily understandable to humans, especially in high-dimensional settings. LIME and CAM intentionally select or highlight a small number of important features or regions, thereby offering sparse, concise outputs [60, 122]. KernelSHAP also exhibits sparsity by assigning zero weights to non-influential features in many cases [61]. However, methods such as IG, LRP, and TreeSHAP typically yield dense attributions because they aim to provide comprehensive explanations [51, 181], which enhances faithfulness but can hinder interpretability.

Closely tied to the practical utility of explanations is reliability, which reflects the consistency and semantic validity of explanations across varied conditions. TreeSHAP is highly reliable due to its deterministic formulation and theoretical soundness [210]. IG and LRP also show strong reliability because they consistently identify relevant features in a principled manner that adheres to model internals [51, 181]. By contrast, LIME, saliency maps, and DeepSHAP can generate unreliable outputs depending on the choice of perturbations, baselines, or model approximations [211, 214].

The other key consideration is computational complexity, particularly important in real-time applications or resource-constrained environments. LIME, CAM, and saliency maps are relatively lightweight, making them appealing for fast deployment [60, 122]. TreeSHAP is efficient for tree-based models due to its optimized structure [210], while IG benefits from gradient-based computations that can be parallelized [51]. However, KernelSHAP incurs substantial computational overhead due to its sampling and kernel-weighted estimation process [61]. DeepSHAP, combining aspects of SHAP and DeepLIFT, also suffers from this inefficiency when applied to deep networks [211].

In recent advancements of explainable AI applied to generative diffusion models, several methods have emerged to interpret model behavior during the iterative denoising process. Techniques such as DF-RISE, DF-CAM, DAAM (Diffusion Attentive Attribution Maps), and Concept Attention aim to bridge the interpretability gap in generative tasks [116, 67, 213].

DF-RISE and DF-CAM leverage perturbation-based and gradient-based mechanisms

respectively to generate attribution maps across multiple denoising steps. These techniques provide partial faithfulness and moderate stability, with some ability to localize token influences over time. However, they involve substantial computational overhead due to repeated sampling and inference steps, and they lack fine-grained token-level resolution [116, 73].

DAAM introduces a novel approach by aggregating internal cross-attention maps of diffusion models to derive word-to-region attributions without requiring model perturbation [67]. By tracing how individual tokens influence spatial attention across denoising timesteps, DAAM produces stable, semantically meaningful token-to-region maps. Its architecture allows it to operate efficiently compared to sampling-based methods, and its explanations are token-level, interpretable, and consistent. However, since DAAM reflects attention patterns rather than causal attributions, its faithfulness and reliability remain partial. Still, it supports timestep-aware reasoning and facilitates prompt dissection, making it a practical tool for real-world diffusion explainability workflows.

The Concept Attention framework by Chen et al. [213] instead modifies the generative process itself by injecting semantic guidance, improving segmentation quality and coherence. While it contributes to interpretability via structured generation, it does not provide post-hoc explanations and does not support predictive or timestep-aware analysis.

Together, these diffusion-based methods highlight the evolving nature of XAI for generative models. DAAM offers an efficient middle ground between faithfulness and scalability, while DF-RISE and DF-CAM prioritize attribution fidelity at a higher computational cost. Concept Attention advances semantic control but lacks diagnostic transparency. These trade-offs underscore the need to align explanation tools with application-specific interpretability goals.

Synthesizing these evaluations, it becomes clear that each XAI method involves trade-offs. TreeSHAP stands out for its faithfulness, stability, and reliability, but lacks sparsity, making its explanations dense. LIME offers sparse and computationally efficient explanations, but at the cost of faithfulness and stability. IG and LRP provide high fidelity and reliability but generate dense outputs that can overwhelm end-users. KernelSHAP

is both faithful and sparse but computationally demanding. These trade-offs underscore the necessity of aligning method selection with specific application requirements and user expectations.

Furthermore, each technique's performance is deeply intertwined with model architecture. TreeSHAP is tailored for decision tree-based models like XGBoost and random forests, while IG, DeepSHAP, and LRP are more suited for deep neural networks. CAM is effective for CNN-based visual tasks due to its spatially aligned heatmaps, but is not adaptable to arbitrary model types [122]. These dependencies highlight the need to match explanation methods not just to desirable properties but also to architectural compatibility.

Chapter 3

Methodology

This study introduces **FIFA–DiffuSAGE**, a pair of hybrid attribution methods developed to enhance the explainability of both predictive models and deep generative diffusion models. The first method, **Firefly-Inspired Feature Attribution (FIFA)**, leverages the Firefly Algorithm [215] to generate sparse, stable, and faithful feature attributions for black-box predictive models on tabular data. The second, **DiffuSAGE** (Diffusion Shapley Attribution with Gradient Explanations), combines Aumann-Shapley values [216] with Integrated Gradients [51] and cross-attention alignment to provide step-aware and token-level attributions for text-to-image diffusion models. The following sections present the theoretical formulation, design rationale, and experimental evaluation of both methods.

3.1 Research Methodology

This study adopts the Design Science Research Process (DSRP) as the guiding methodology to develop novel attribution methods for explainability in both predictive and generative deep learning models. DSRP is a systematic, iterative framework widely used in computer science and information systems to design, build, and rigorously evaluate innovative solutions [217]. Within this methodology, we introduce **FIFA** for predictive models and **DiffuSAGE** for deep generative diffusion models. These methods are designed to address trade-offs between attribution faithfulness and interpretability.

DSRP structures our research through progressive stages: problem identification, objective formulation, design and development, evaluation, and communication. Figure 3.1 illustrates these phases, aligning them with the formulation and validation of our attribution methods [217].

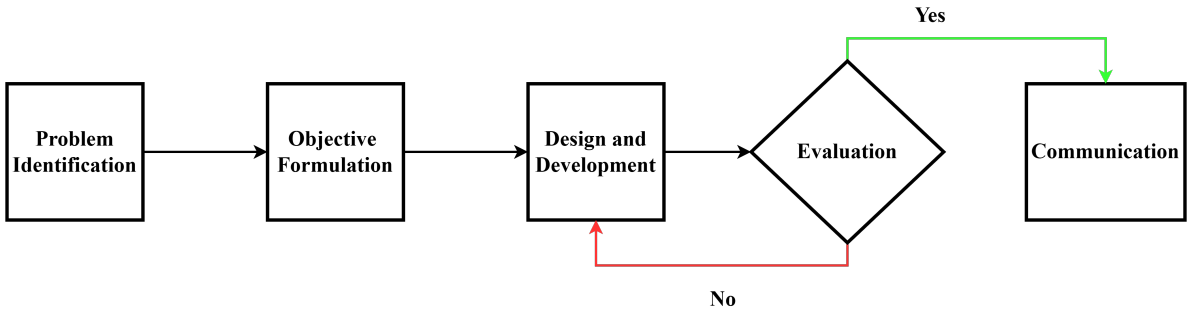


Figure 3.1: Research Design

Problem Identification

To identify the research problem, we conducted a focused exploration of explainability techniques, particularly examining the challenges of balancing faithfulness and interpretability. Our review of feature attribution methods critically analyzed their trade-offs and limitations, which informed the development of more principled and effective solutions.

We examined foundational works in explainable AI, including SHAP, LIME, Integrated Gradients, and Aumann-Shapley values. This literature review highlighted persistent issues, including baseline sensitivity, instability, and a lack of fairness guarantees. These findings motivated a deeper evaluation of the limitations' root causes and practical implications, guiding the formulation of our research direction.

Through this assessment, we identified two core research problems: (1) the need for a biologically inspired, optimization-based attribution method to improve stability, faithfulness, and interpretability in predictive models; and (2) the need for a hybrid temporal attribution method combining Aumann-Shapley values and Integrated Gradients to enhance the interpretability of deep generative diffusion models. These problems form the basis of our investigation, driving the development of FIFA and DiffuSAGE as key contributions to the field.

Objective Formulation

Following problem identification, we formulated the central research question, as detailed in Section 1.3. To address this, we established clear objectives, starting with a general

goal of advancing explainability for both predictive and generative deep learning models.

Building on this foundation, we articulated specific objectives 1.4 to guide the development of FIFA for predictive models and DiffuSAGE for generative diffusion models. This structured roadmap ensured that each research stage remained aligned with the core goals of improving faithfulness and interpretability.

Design and Development

Based on the insights gathered, we propose two complementary approaches to address the trade-offs between faithfulness, interpretability, and performance in explainability. For predictive models, we introduce FIFA, an optimization-based method designed to enhance stability, sparsity, and faithfulness in feature attributions. For deep generative models, we develop DiffuSAGE, a hybrid framework that combines the fairness and robustness of Aumann-Shapley values with the interpretability of Integrated Gradients, enabling balanced and interpretable attributions across the multi-step generative process.

We explored strategies like diversity-driven search in FIFA and baseline selection, path refinement, and temporal attribution aggregation in DiffuSAGE. Additionally, regularization techniques and structural adaptations ensure generalizability across different model architectures and datasets. Through iterative design and empirical evaluation, we developed practical, scalable solutions for achieving faithful and interpretable explainability in both predictive and generative AI systems.

Evaluation

In the evaluation phase, we assess the performance and viability of our methods using a set of well-defined metrics and evaluation criteria. As detailed in Chapter 4, we evaluate FIFA and DiffuSAGE by analyzing attribution faithfulness, stability, and robustness across different scenarios. We benchmark both methods on models of varying complexity, including advanced architectures.

We also examine their practical applicability by simulating real-world use cases, ensuring that FIFA and DiffuSAGE are not only theoretically sound but also effective in applied explainability settings. This comprehensive evaluation validates the adaptability

and performance of our hybrid attribution methods across both predictive and generative AI contexts.

Communication

In the communication phase, we focus on effectively presenting the results and insights from our proposed methods: FIFA and DiffuSAGE. This involves documenting the methodologies, design rationales, and evaluation outcomes through structured presentations and visualizations. We emphasize how each approach addresses the trade-offs between faithfulness and interpretability in model explainability.

Our findings are shared through peer-reviewed research papers, technical presentations, and targeted communication channels, ensuring our contributions advance academic understanding and support practical adoption in real-world AI systems.

3.2 Data Acquisition

Acquiring diverse and representative datasets is crucial for benchmarking the performance and reliability of explainability methods. To support the development of our predictive attribution method, FIFA, we selected three widely recognized datasets: Breast Cancer (WDBC), Diabetes, and Adult Income (Census). These datasets differ in domain, size, and complexity, offering a well-rounded exploration of FIFA's capability to handle various predictive modeling scenarios.

The Breast Cancer (WDBC) dataset [218] consists of 30 features derived from digitized fine needle aspirate (FNA) images, with 569 samples, and is used for binary classification of tumor malignancy. The Diabetes dataset [219], sourced from Kaggle, contains 8 medical attributes collected from 768 individuals and is used for regression tasks related to disease progression. The Adult Income (Census)dataset [220] includes 14 demographic and economic features across 48,842 samples and is commonly used for binary income classification tasks. Table 3.1 provides an overview of these datasets.

These datasets were selected due to their widespread use in explainable AI research, as well as their ability to highlight the challenges and opportunities in generating faithful

and sparse feature attributions [221, 222, 223]. The variation in feature dimensionality and dataset size enables a robust comparison across different learning tasks and data characteristics.

Table 3.1: Datasets used for feature attribution benchmarking.

Dataset	**# Features**	**# Samples**
Breast Cancer (WDBC) [218]	30	569
Diabetes [219]	8	768
Adult Income (Census) [220]	14	48,842

3.3 Data preprocessing

To ensure consistent and fair comparison across predictive tasks, we applied standardized preprocessing procedures to all datasets evaluated with FIFA. These steps supported robust model training while maintaining the interpretability of features for attribution analysis.

For numerical features, min-max normalization was applied to scale values into the range $[0, 1]$, ensuring uniform contribution across features. This normalization was applied uniformly to the Diabetes and Breast Cancer (WDBC) datasets, which consist entirely of continuous attributes.

The Adult Income (Census) dataset, which includes both numerical and categorical features, was preprocessed using one-hot encoding for categorical variables. Missing values (e.g., in `workclass` and `native-country`) were treated as a separate category to avoid information loss.

All datasets were randomly split into training and test sets using an 80/20 ratio. For classification tasks (Breast Cancer and Adult Income), stratified sampling was used to maintain class balance. Preprocessing was implemented using `scikit-learn` and applied uniformly to ensure consistency and compatibility with FIFA.

3.4 Design and Development

The proposed method produces robust, model-specific explanations for both predictive and generative diffusion models. It adopts a unified, modular architecture that accommodates heterogeneous input types and delivers faithful, stable, and semantically coherent attributions.

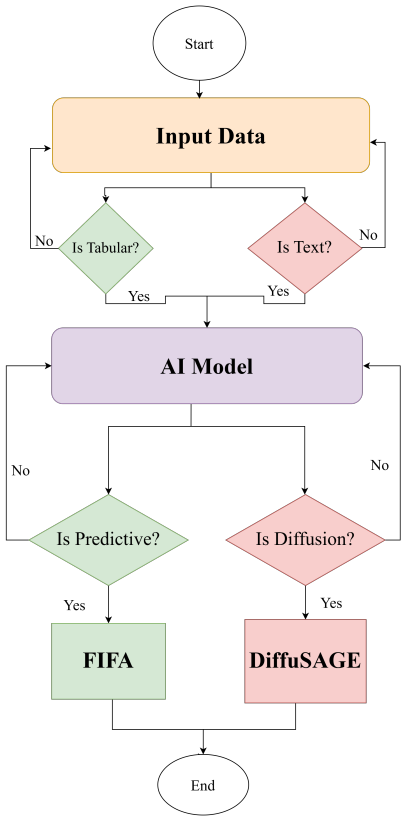


Figure 3.2: System flow of the attribution-based explainability framework for predictive and diffusion models.

As shown in Figure 3.2, the system begins by ingesting structured tabular features or natural language prompts. It then determines the nature of the target model, predictive (classifiers or regressors) or generative (text-to-image diffusion models), based on input modality and available model metadata.

Based on this classification, the framework routes the input-output pair to one of two dedicated explanation modules. For predictive models, it invokes **FIFA**, which computes sparse, stable, and faithful feature importance scores. For generative models, it activates **DiffuSAGE**, a token-level attribution mechanism that captures the spatial and temporal

influence of prompt tokens throughout the diffusion process.

This routing mechanism enables task-specific attribution without requiring access to internal model parameters or gradients, making the system architecture agnostic while preserving interpretability and reliability.

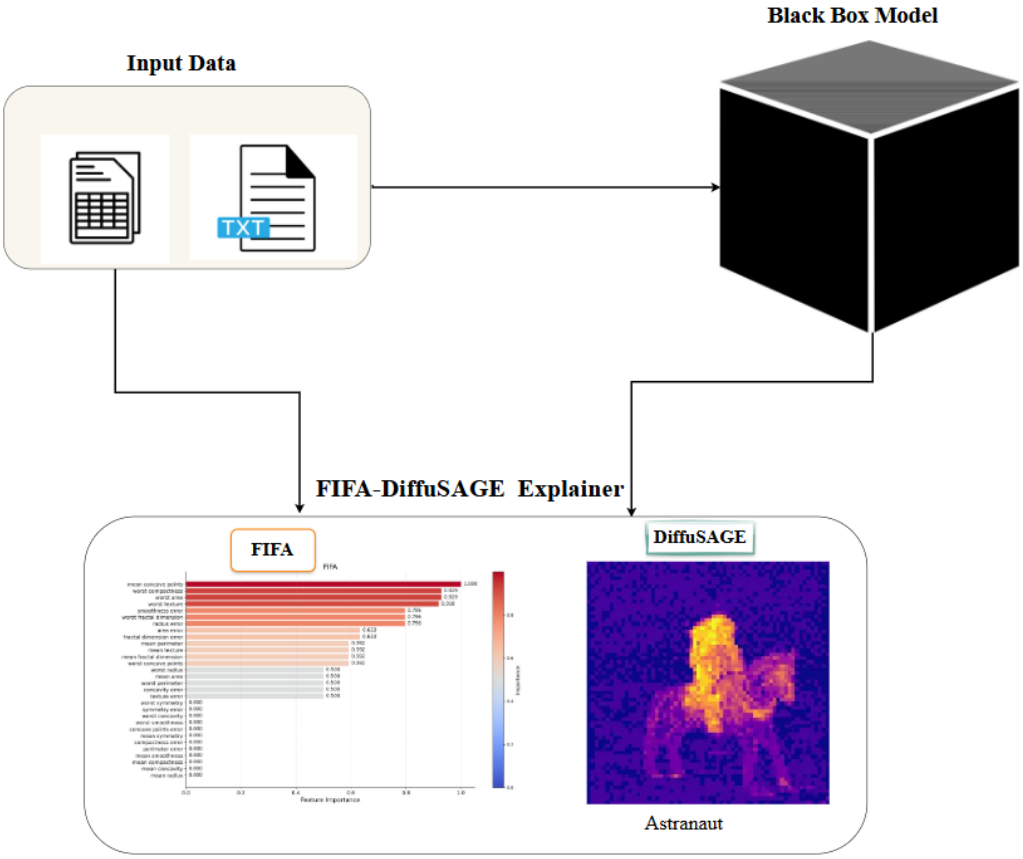


Figure 3.3: Architectural overview of the FIFA–DiffuSAGE Explainer.

Figure 3.3 shows the internal architecture of the explanation engine, which comprises two parallel paths tailored to each model type, followed by a unified evaluation and visualization module.

In the predictive path, FIFA analyzes structured inputs and produces ranked feature importance scores, guided by four core principles: sparsity, to identify concise yet informative feature subsets; stability, to ensure robustness under perturbation; faithfulness, to reflect the model's true reasoning; and reliability, to maintain consistency across runs and input variations. Results are visualized using bar plots for interpretability.

In the generative path, DiffuSAGE aggregates Integrated Gradients across multiple

diffusion steps and aligns them with cross-attention maps extracted from intermediate layers. This results in temporally grounded, token-to-region attributions. Outputs include heatmaps, relevance curves, and token saliency profiles that visualize how prompt tokens influence the image generation process over time.

Both modules feed into a shared evaluation and visualization interface that standardizes output formats and supports quantitative comparison across tasks. Evaluation metrics such as faithfulness, stability, sparsity, and insertion and deletion AUC are employed to assess explanation quality.

By integrating task-specific attribution strategies within a coherent, extensible design, the FIFA–DiffuSAGE method enables interpretable, trustworthy explanations across a wide range of machine learning applications.

3.4.1 Firefly-Inspired Feature Attribution (FIFA)

FIFA is a model-agnostic, optimization-based feature attribution method inspired by the collective behavior of fireflies [215]. Frame explanation as a search for the smallest subset of input features that preserves the output of a model. Each candidate subset is encoded as a binary vector and is evaluated based on a fitness score that reflects the consistency of the prediction.

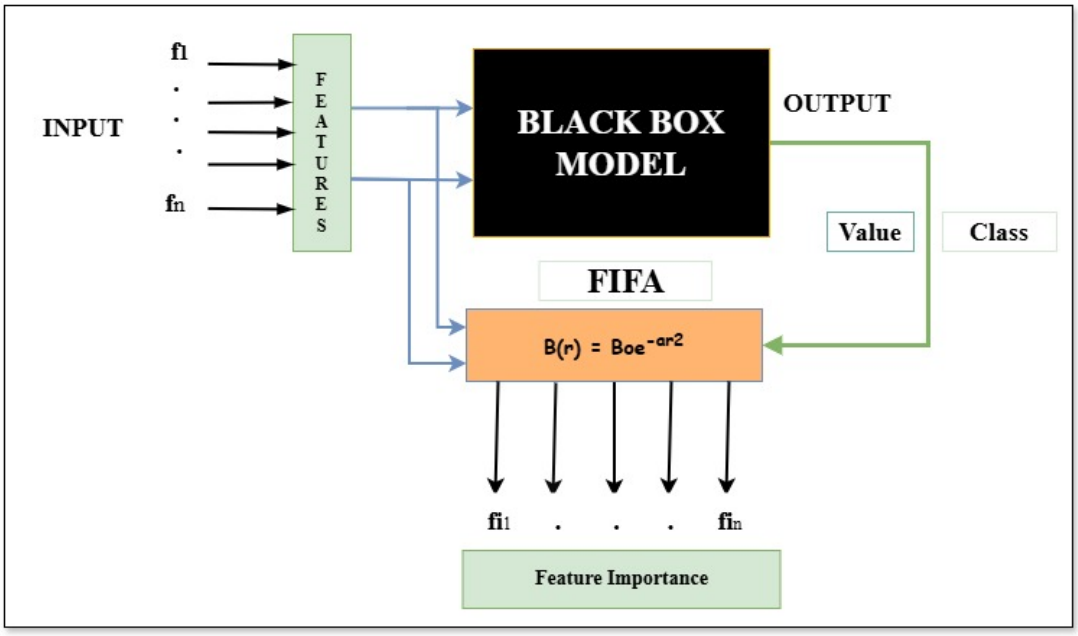


Figure 3.4: Proposed architecture of FIFA.

As shown in Figure 3.4, FIFA iteratively samples feature subsets and evaluates their fidelity using a black-box model. The optimization is guided by a brightness function that quantifies similarity between subsets:

$$B(r) = \beta_0 e^{-\gamma r^2} \tag{3.1}$$

where r is the Hamming distance between two subsets, and $B(r)$ determines attractiveness in the swarm update step.

Problem Formulation

Formally, given an input $x = \{f_1, \ldots, f_n\}$ and a trained model f, FIFA seeks a feature subset $S^* \subseteq \{1, \ldots, n\}$ that maximizes predictive fidelity:

$$S^* = \arg \max_{S \subseteq \{1,\ldots,n\}} \text{Fitness}(S) \tag{3.2}$$

For classification tasks:

$$I(S) = P_{\text{model}}(y = c \mid x_S) \tag{3.3}$$

where c is the original class prediction and x_S is the subset of features in S.

For regression tasks, metrics such as R^2, MSE, or MAE are used to define Fitness(S). A sparsity constraint can be incorporated to discourage large subsets and enhance interpretability.

Feature importance is quantified as the marginal drop in confidence when a feature is removed from the best-performing subset:

$$\Delta_k = P_{\text{model}}(c \mid x_{S^*}) - P_{\text{model}}(c \mid x_{S^* \setminus \{k\}}) \tag{3.4}$$

The output is a ranked list of features that reflect the model's reasoning sparsely and faithfully.

Working Principle

FIFA operates on three fundamental principles: Initialization, Movement and Update, and Convergence.

Initialization: Feature subsets are encoded as binary vectors, and a firefly population is randomly initialized to ensure a diverse coverage of the search space. Each firefly represents a potential subset of features. Prior knowledge (e.g., from simpler models) can optionally guide this process.

Movement and Update: Fitness for each firefly is evaluated by querying the model with its feature subset. Less fit fireflies move toward brighter (higher-fitness) ones based on:

$$x_i \leftarrow x_i + \beta_0 e^{-\gamma r_{ij}^2}(x_j - x_i) + \alpha\epsilon_i \tag{3.5}$$

where r_{ij} is the Hamming distance between x_i and x_j, and α adds stochasticity for exploration. This process promotes convergence toward optimal feature subsets while preserving diversity.

Convergence: The population evolves over multiple iterations, with the algorithm halting after reaching a maximum iteration count $T_{\max}$ or when improvement plateaus. The top performing subset S^* is retained, and the importance of the features is calculated as Equation 3.4.

Hyperparameters: The performance of FIFA is governed by a set of key hyperparameters that balance exploration and exploitation, guide convergence, and control computational efficiency. Each parameter plays a distinct role in steering the optimization process toward sparse, stable, and faithful feature attributions.

- α **(Randomness factor):** Determines the level of stochastic perturbation in the movement of each firefly. Higher α encourages broad exploration and helps avoid premature convergence, while lower α promotes focused exploitation of high-fitness

regions. Careful tuning of α is essential for maintaining a balance between diversity and stability.

- β_0 **(Maximum attractiveness):** Sets the attractiveness of a firefly at zero distance, controlling how strongly fireflies are drawn toward better-performing peers. Higher values accelerate convergence but risk local optima; lower values allow slower, more deliberate progress across the search space.

- γ **(Attractiveness decay rate):** Controls how quickly attractiveness decreases with distance. A larger γ restricts movement to nearby fireflies, supporting local refinement; a smaller γ facilitates global exploration by permitting long-range influence.

- $T_{\mathbf{max}}$ **(Iteration limit):** Specifies the maximum number of iterations allowed. Higher values enable more thorough optimization at the cost of increased runtime, while lower values provide faster termination but may limit convergence. It acts as a practical time-bound for the search process.

- N **(Population size):** Determines the number of fireflies in the swarm. Larger populations improve coverage and search diversity, increasing the likelihood of identifying optimal subsets. Smaller populations reduce computational load but may risk premature convergence or limited exploration.

Together, these hyperparameters allow FIFA to efficiently navigate the combinatorial space of feature subsets and generate robust, interpretable explanations. Its model-agnostic nature ensures broad applicability across domains and predictive modeling tasks.

The pseudocode in Algorithm 1 outlines the core working principle of the FIFA method. It begins by initializing a population of binary fireflies, where each firefly represents a candidate subset of features encoded as a binary vector. The fitness of each firefly is evaluated using the prediction confidence of a pre-trained model when restricted to the selected features. Over multiple iterations, each firefly compares itself with brighter (higher-fitness) peers and updates its position using a binary adaptation of the Firefly

Algorithm's movement rule, guided by a combination of attraction and random perturbation. The algorithm progressively refines the population toward subsets that better explain the model's prediction. Once convergence is reached, the best-performing subset is identified, and feature importance scores are computed by measuring the marginal drop in prediction confidence when each feature is individually removed.

Algorithm 1 Firefly-Inspired Feature Attribution (FIFA)

Input: Trained model f, input sample x, number of fireflies N, number of iterations T
Output: Feature importance scores $\{\Delta_k\}$
Initialize N binary fireflies $\{x_1, x_2, \ldots, x_N\} \in \{0,1\}^d$
for each firefly x_i **do**
    Compute fitness $I_i \leftarrow f(x_i)$
end for
for $t = 1$ to T **do**
    for each firefly x_i **do**
        for each firefly x_j where $I_j > I_i$ **do**
            Compute Hamming distance r_{ij}
            Compute attractiveness $\beta \leftarrow \beta_0 e^{-\gamma r_{ij}^2}$
            Update position:
                $x_i \leftarrow x_i + \beta(x_j - x_i) + \alpha \cdot \epsilon$
                (Apply rounding and clipping for binary domain)
        end for
        Update fitness $I_i \leftarrow f(x_i)$
    end for
end for
Select best firefly x^* with maximum fitness
for each feature k in x^* **do**
    Compute importance:
        $\Delta_k \leftarrow f(x^*) - f(x^* \setminus \{k\})$
end for
Assign zero importance to features not in x^*
return Feature importance scores $\{\Delta_k\}$

3.5 Diffusion Shapley Attribution with Gradient Explanations(DiffuSAGE)

The architectural overview of DiffuSAGE, shown in Figure 3.5, outlines a framework designed to generate faithful and fine-grained explanations for text-to-image diffusion models. Its core objective is to quantify when, where, and how each input token influences the generated image. The process begins with a textual input prompt, for example,

"An astronaut riding a horse on Mars", which is tokenized and cleaned of stop words. The resulting tokens are encoded into semantic embeddings via a text encoder and subsequently fed into a diffusion model that performs iterative denoising over timesteps $t = 1, 2, \ldots, T$, progressively refining the image.

At each timestep, the model produces intermediate outputs: $\mathbf{T}$, the timestep index; $\mathbf{A}_t$, cross-attention maps linking image regions and prompt tokens; and $\mathbf{I}_t$, latent representations of the image. These components are passed to the DiffuSAGE module, which integrates Integrated Gradients (IG)[51] and Aumann-Shapley Values (ASV)[216] to compute token-level attributions. The resulting outputs include (1) **Token Importance**, which quantifies each prompt token's overall influence on the generated image; (2) **Timestep Importance**, which identifies the diffusion steps that had the most significant impact; and (3) **Token-to-Region Mapping**, which localizes the spatial influence of each token in the image.

By combining IG's gradient-based sensitivity, ASV's fairness properties, and the spatial grounding provided by cross-attention, DiffuSAGE delivers temporally grounded and spatially precise explanations that reflect the model's internal decision-making process.

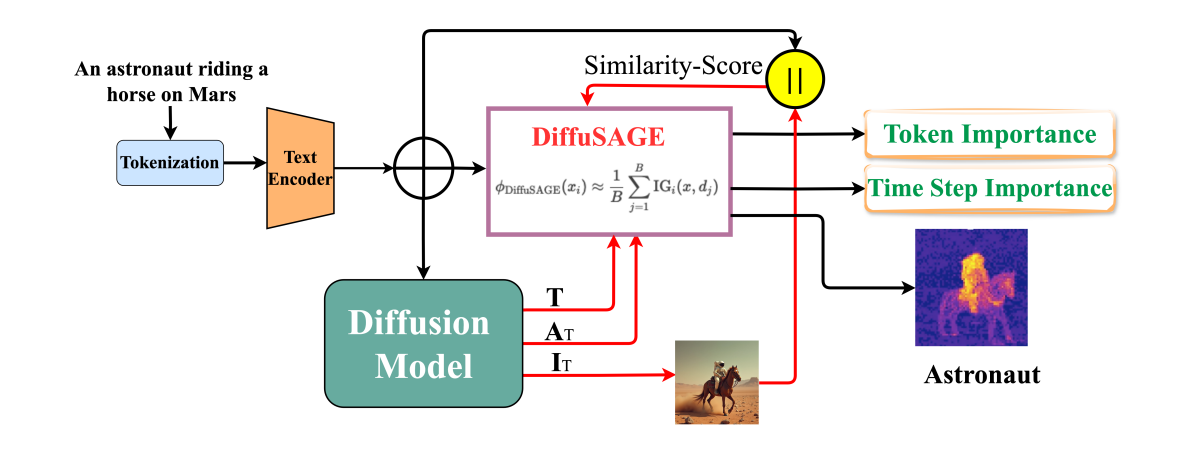


Figure 3.5: Proposed Architecture for DiffuSAGE

3.5.1 Problem Formulation

In generative diffusion models, the image is synthesized through a sequence of latent transformations from pure noise to a coherent output[119]. Unlike classification tasks,

where attribution maps directly to class probabilities, text-to-image diffusion models involve temporal, semantic, and spatial factors. Therefore, attributing image content to specific input tokens demands a temporal decomposition of influence across the denoising trajectory.

Let the input prompt be tokenized into a sequence of tokens $\{x_1, x_2, \ldots, x_n\}$, where each x_i is transformed into an embedding E_i using a frozen text encoder. These embeddings condition the UNet denoising model at each timestep $t \in \{1, 2, \ldots, T\}$.

The generation trajectory can be represented as a series of latent states $\{\mathbf{z}_t\}$, where the model predicts noise $\epsilon_\theta(\mathbf{z}_t, t, \{E_i\})$ and updates the latent $\mathbf{z}_{t-1}$.

We define a scalar scoring function F (similarity between prompt and final image) to quantify how well the generated image aligns with the input semantics.

Our goal is to compute: 1) Token Attribution $\phi_i \in \mathbb{R}$: how much each token x_i contributes to the final score F, 2) Timestep Attribution $\psi_t \in \mathbb{R}$: how important each timestep is in shaping the final image, 3) Token-to-Region Mapping $R_i \in \mathbb{R}^{H \times W}$: a heatmap localizing the impact of token x_i across the image space.

The challenge lies in capturing nonlinear, non-monotonic, and temporally-distributed influence patterns, which we address using a combination of Integrated Gradients and Aumann-Shapley Value approximations.

3.5.2 Integrated Gradients (IG)

Integrated Gradients (IG) [51] is a path-based attribution method that assigns importance scores to input features by integrating gradients along a straight-line path from a baseline input to the actual input. IG satisfies key axioms such as *sensitivity* and *implementation invariance*, making it a principled approach for attributing model predictions to inputs.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and $x \in \mathbb{R}^n$ the input of interest. Given a baseline input x', the IG attribution for the i-th feature is defined as:

$$\mathrm{IG}_i(x) = (x_i - x'_i) \cdot \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} \, d\alpha \tag{3.6}$$

where $\alpha \in [0, 1]$ is the interpolation parameter, and x' represents a reference input such

as a zero vector.

In text-to-image diffusion models, the input is a sequence of token embeddings $\{E_1, E_2, \ldots, E_n\}$, with each $E_i \in \mathbb{R}^d$ obtained from a frozen text encoder. These embeddings condition the generative denoising process across T timesteps. Let the final image representation be $\mathbf{z}_T$, and define a scoring function $F(\{E_i\}, \mathbf{z}_T)$ that quantifies prompt-image alignment, such as CLIP-based cosine similarity.

The attribution to token embedding E_i is computed by applying IG over the input embedding path:

$$\phi_i = (E_i - E_i')^\top \cdot \int_{\alpha=0}^{1} \frac{\partial F(\{E_j^{(\alpha)}\}, \mathbf{z}_T)}{\partial E_i} \, d\alpha \tag{3.7}$$

where $E_j^{(\alpha)} = E_j' + \alpha(E_j - E_j')$, and E_i' is a baseline embedding token or null vector). In practice, the integral is approximated using a Riemann sum over m steps:

$$\phi_i \approx (E_i - E_i')^\top \cdot \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F\left(\left\{E_j' + \frac{k}{m}(E_j - E_j')\right\}, \mathbf{z}_T\right)}{\partial E_i} \tag{3.8}$$

This formulation yields a faithful and differentiable token-level attribution that captures how each prompt token affects the semantic properties of the final image, directly applying the IG framework without modifying the underlying model architecture.

3.5.3 Aumann-Shapley Values (ASV)

The Aumann-Shapley Value [216] is a continuous generalization of the discrete Shapley value for settings with infinitely divisible inputs. Unlike the original Shapley formulation, which averages marginal contributions over all possible permutations of discrete features, ASV defines a path-integral over marginal contributions, aligning naturally with gradient-based models.

Given a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ and an input vector $x \in \mathbb{R}^n$, the Aumann-Shapley value for the i-th input is:

$$\phi_i^{\text{ASV}}(x) = \int_0^1 \frac{\partial f(\alpha x)}{\partial x_i} \cdot x_i \, d\alpha \tag{3.9}$$

which is equivalent to:

$$\phi_i^{\text{ASV}}(x) = x_i \cdot \int_0^1 \frac{\partial f(\alpha x)}{\partial x_i}\, d\alpha \tag{3.10}$$

This formulation has a direct analogy to Integrated Gradients (IG), with ASV arising as the unique value function satisfying efficiency, linearity, and symmetry under infinitesimal inputs.

In the diffusion setting, let the input be a sequence of token embeddings $\{E_1, \ldots, E_n\}$ from a frozen text encoder, and the generation process proceed over T timesteps, producing latent representations $\{\mathbf{z}_t\}$. Define a scoring function $F(\{E_i\}, \mathbf{z}_T)$ that evaluates how well the final image $\mathbf{z}_T$ aligns with the prompt semantics.

The Aumann-Shapley attribution to token embedding E_i is given by:

$$\phi_i^{\text{ASV}} = E_i^\top \cdot \int_0^1 \frac{\partial F(\{\alpha E_j\}, \mathbf{z}_T)}{\partial E_i}\, d\alpha \tag{3.11}$$

where $\{\alpha E_j\}$ denotes the scaled embedding path. As with IG, the integral is approximated via discrete summation over m steps:

$$\phi_i^{\text{ASV}} \approx E_i^\top \cdot \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F\left(\left\{\frac{k}{m} E_j\right\}, \mathbf{z}_T\right)}{\partial E_i} \tag{3.12}$$

This formulation distributes credit for output changes smoothly along a continuous scaling of input embeddings and supports attribution with theoretical fairness properties. In diffusion models, it allows us to interpret how much each token contributes to the final image by averaging influence across all scaling paths, making ASV a natural fit for time-distributed, gradient-compatible generative processes.

3.5.4 Token Importance

To compute the influence of each input token on the final generated image, DiffuSAGE utilizes both gradient information and Shapley-inspired temporal weighting. At each timestep $t \in \{1, \ldots, T\}$, the UNet model consumes the text embeddings and produces latent predictions $\hat{\epsilon}_t$. Given a scalar scoring function $F(\mathbf{z}_T, \{E_i\})$—such as CLIP similar-

ity—the token attribution ϕ_i is computed as:

$$\phi_i = \sum_{t=1}^{T} w_t \cdot \left(\frac{\partial F(\mathbf{z}_T)}{\partial E_i^{(t)}} \right) \tag{3.13}$$

where $E_i^{(t)}$ denotes the embedding of token x_i at timestep t, and w_t is the Shapley-inspired weight representing the marginal contribution of timestep t in the generation process.

To improve faithfulness, w_t is not assumed uniform; rather, it is derived from CLIP Drop or gradient magnitude to reflect the relative importance of each timestep in shaping the final image.

To capture the dynamics of influence propagation over the denoising trajectory, DiffuSAGE maintains a sequence of partial token attributions $\phi_i^{(t)}$ at each step:

$$\phi_i^{(t)} = \frac{\partial F(\mathbf{z}_t)}{\partial E_i^{(t)}} \tag{3.14}$$

These are then aggregated using ASV-derived weights:

$$\phi_i = \sum_{t=1}^{T} w_t \cdot \phi_i^{(t)} \tag{3.15}$$

This formulation explicitly models token importance propagation over time, allowing us to observe how each token's influence accumulates or fades during the image synthesis process.

DiffuSAGE ensures attribution is temporally distributed, avoiding the bias of attributing all credit to the final steps. The final attribution vector $\{\phi_1, \ldots, \phi_n\}$ is normalized to sum to 1, satisfying the Shapley efficiency axiom.

3.5.5 Timestep Importance

Timestep importance in DiffuSAGE measures the significance of each timestep $t \in \{1, 2, \ldots, T\}$ in shaping the final image output. This is computed via two complementary strategies:

1) Gradient-Based Aggregation: We aggregate token sensitivities across all tokens at each timestep:

$$\psi_t = \sum_{i=1}^{n} \left| \frac{\partial F(\mathbf{z}_t)}{\partial E_i} \right| \tag{3.16}$$

This captures how responsive the model is to token embeddings at each t, highlighting diffusion stages with high attribution dynamics.

2) Semantic Degradation: We halt the diffusion process prematurely at timestep t and decode the intermediate latent $\mathbf{z}_t$ into an image. The CLIP similarity between this early image and the input prompt is then compared to the full image:

$$\Delta_t = F(\mathbf{z}_T) - F(\mathbf{z}_t) \tag{3.17}$$

A large drop Δ_t implies that the timestep t contributes significantly to maintaining or establishing semantic alignment.

Together, these metrics offer a robust interpretation of which stages in the denoising trajectory are critical, enabling temporal analysis of semantic formation in diffusion-based generation.

3.5.6 Token-to-Visual Region Mapping

To understand not just which tokens are important and when, but also where they influence the image, DiffuSAGE computes token-to-region attributions by integrating attention maps with gradient-based sensitivity.

At each timestep t, the UNet produces a cross-attention map $\mathbf{A}_t \in \mathbb{R}^{H \times W \times n}$ aligning image regions to input tokens. In parallel, gradients $\nabla_t^i \in \mathbb{R}^{H \times W}$ are computed for each token x_i, capturing how sensitive the generated pixels are to changes in the token embedding.

The token-to-region attribution heatmap $R_i \in \mathbb{R}^{H \times W}$ is defined as:

$$R_i = \sum_{t=1}^{T} w_t \cdot \left(\text{Attn}_t^i \odot |\nabla_t^i| \right) \tag{3.18}$$

Where: w_t are timestep weights derived from Shapley-based averaging or other importance weighting schemes, $\text{Attn}_t^i \in \mathbb{R}^{H \times W}$ is the attention map for token i, $\nabla_t^i \in \mathbb{R}^{H \times W}$

is the gradient of the image score with respect to the latent features at t, $\odot$ denotes element-wise multiplication.

This formulation ensures that token-region attributions are: Localized via spatial attention, Faithful via gradients from the actual model output, and Temporally-aware through weighted integration over timesteps.

The resulting heatmap R_i can be overlaid on the final image to visualize the spatial influence of each token in the prompt. This completes the multi-dimensional attribution strategy of DiffuSAGE, offering interpretability across semantic, temporal, and spatial axes.

Algorithm 2 DiffuSAGE Attribution Algorithm

Require: Tokenized prompt $\{x_1, \ldots, x_n\}$, Text encoder, Diffusion model with T steps, Scoring function F

Ensure: Token importances ϕ_i, Timestep importances ψ_t, Region heatmaps R_i

1: Encode tokens: $E_i \leftarrow \text{text\_encoder}(x_i)$
2: Initialize latent noise: $\mathbf{z}_T \sim \mathcal{N}(0, I)$
3: Initialize attribution containers: $\phi_i^{(t)} \leftarrow 0$, $\psi_t \leftarrow 0$, $R_i^{(t)} \leftarrow \mathbf{0}$
4: **for** $t = T$ **down to** 1 **do**
5: Predict noise: $\hat{\epsilon}_t \leftarrow \text{UNet}(\mathbf{z}_t, t, \{E_i\})$
6: Update latent: $\mathbf{z}_{t-1} \leftarrow \text{scheduler\_step}(\hat{\epsilon}_t, \mathbf{z}_t, t)$
7: Compute gradients: $g_i^{(t)} \leftarrow \frac{\partial F(\mathbf{z}_t)}{\partial E_i}$
8: Token attribution: $\phi_i^{(t)} \leftarrow g_i^{(t)}$
9: Timestep attribution: $\psi_t \leftarrow \sum_i \left| g_i^{(t)} \right|$
10: Extract cross-attention maps: $\mathbf{A}_t^i \leftarrow \text{CrossAttn}(x_i, t)$
11: Compute region heatmap: $R_i^{(t)} \leftarrow \mathbf{A}_t^i \odot |\nabla_{\mathbf{z}_t} F|$
12: **end for**
13: Compute Shapley timestep weights w_t (normalized ψ_t or CLIP drop)
14: Aggregate token importance: $\phi_i = \sum_{t=1}^{T} w_t \cdot \phi_i^{(t)}$
15: Aggregate region map: $R_i = \sum_{t=1}^{T} w_t \cdot R_i^{(t)}$
16: **return** $\{\phi_i\}$, $\{\psi_t\}$, $\{R_i\}$

The pseudocode presented in Algorithm 2 encapsulates the operational flow of DiffuSAGE and its three-pronged attribution objectives. The process begins by encoding each token x_i in the input prompt into embeddings E_i using a frozen text encoder (Line 1). The diffusion process is initialized with pure Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(0, I)$ (Line 2), and containers for storing timestep-wise token attributions $\phi_i^{(t)}$, timestep importance ψ_t, and token-to-region maps $R_i^{(t)}$ are instantiated (Line 3). Within the reverse diffusion

loop (Lines 4–12), the model iteratively predicts noise $\hat{\epsilon}_t$ (Line 5) and updates the latent variable $\mathbf{z}_{t-1}$ using a scheduler (Line 6). At each step, gradients $g_i^{(t)}$ of a scoring function (CLIP similarity) are computed with respect to the input embeddings (Line 7). These gradients quantify token sensitivity and are stored as per-timestep attributions $\phi_i^{(t)}$ (Line 8), while their sum provides timestep importance ψ_t (Line 9). Simultaneously, the UNet's cross-attention maps $\mathbf{A}_t^i$ are extracted (Line 10), and token-to-region maps $R_i^{(t)}$ are computed by modulating these with pixel-level gradients (Line 11), thus anchoring token influence in image space.

Following the diffusion loop, timestep weights w_t are derived by normalizing ψ_t or by computing prompt-image alignment drops across truncated generation (Line 13). These weights serve to aggregate token attributions over time, ensuring that each step's influence is proportionally considered (Line 14). Likewise, token-to-region maps are accumulated using the same temporal weights to yield spatially faithful heatmaps R_i (Line 15). Finally, the algorithm returns the full attribution package: token importance scores ϕ_i, timestep relevance values ψ_t, and spatial region maps R_i (Line 16). This structured approach allows DiffuSAGE to decompose generative decisions across semantic, temporal, and spatial axes, ensuring that the resulting explanations are faithful to the model's actual internal behavior during image synthesis.

3.6 Evaluation Metrics

To assess the quality of attribution methods, we adopt a suite of evaluation metrics grounded in the theoretical principles of faithful and interpretable explanations. These metrics are applicable to both predictive and generative models, and they evaluate attribution quality across five key dimensions: *faithfulness*, *stability*, *reliability*, *sparsity*, and, for diffusion models, *temporal sensitivity* and *spatial alignment*.

Faithfulness Metrics

Faithfulness assesses whether features assigned high importance are indeed causally influential to the model's decision. It is typically evaluated using insertion and deletion

curves.

Insertion AUC

Let $x \in \mathbb{R}^d$ be an input instance and $\phi \in \mathbb{R}^d$ its corresponding attribution vector. The insertion curve measures the model's prediction as features are progressively added in descending order of importance.

Let $S_k = \text{Top-}k(\phi)$ denote the top-k features. Define the masked input $x^{(k)} \in \mathbb{R}^d$ as:

$$x_i^{(k)} = \begin{cases} x_i & \text{if } i \in S_k \\ 0 & \text{otherwise} \end{cases} \tag{3.19}$$

The insertion score at step k is:

$$I_k = f(x^{(k)}), \quad \text{for } k = 1, \ldots, d \tag{3.20}$$

The area under the insertion curve (AUC) is approximated using the trapezoidal rule:

$$\text{AUC}_{\text{ins}} \approx \sum_{k=1}^{d-1} \frac{I_k + I_{k+1}}{2} \tag{3.21}$$

Deletion AUC

Conversely, the deletion curve measures the model prediction as top features are removed:

$$x_i^{(-k)} = \begin{cases} x_i & \text{if } i \notin S_k \\ 0 & \text{otherwise} \end{cases} \tag{3.22}$$

$$D_k = f(x^{(-k)}), \quad \text{AUC}_{\text{del}} = \sum_{k=1}^{d-1} \frac{D_k + D_{k+1}}{2} \tag{3.23}$$

A high insertion AUC and low deletion AUC indicate that the attribution aligns well with the model's actual decision process.

Stability

Stability captures the robustness of an attribution method to small perturbations in the input. Let $x' = x + \delta$, where $\delta \sim \mathcal{N}(0, \epsilon^2)$, and let ϕ and ϕ' be the corresponding attributions. Then:

$$\text{Stability} = \frac{\langle \phi, \phi' \rangle}{\|\phi\|_2 \|\phi'\|_2} \tag{3.24}$$

Averaging over N perturbations yields:

$$\text{Stability} = \frac{1}{N} \sum_{i=1}^{N} \text{cosine}(\phi, \phi^{(i)}) \tag{3.25}$$

Higher values indicate greater consistency and robustness.

Reliability

Reliability quantifies the variance of attributions produced by stochastic explanation methods across multiple runs:

$$\text{Reliability} = \frac{1}{d} \sum_{j=1}^{d} \text{StdDev}(\phi_j^{(1)}, \phi_j^{(2)}, \ldots, \phi_j^{(n)}) \tag{3.26}$$

Lower values indicate more stable and reproducible attributions.

Sparsity

Sparsity reflects the conciseness of an explanation. Let $\|\phi\|_0$ be the number of non-zero entries in the attribution vector:

$$\text{Sparsity} = 1 - \frac{\|\phi\|_0}{d} \tag{3.27}$$

High sparsity is desirable for interpretability as it indicates that only a small number of features are deemed important.

Qualitative Inspection

We overlay R_i on the generated image and visually assess whether the highlighted regions correspond to the expected semantic elements (e.g., "dog", "umbrella", "astronaut"). This visual inspection is standard in evaluating generative attribution when no ground truth is available.

Together, these metrics provide a rigorous and multifaceted framework for evaluating attribution quality. In predictive models, faithfulness, stability, and sparsity are key. In generative models, temporal and spatial alignment become equally critical. This evaluation suite supports fair and transparent comparison across attribution techniques.

Chapter 4

Experimentation

4.1 Experimental Setup

We implemented both the FIFA and DiffuSAGE of our attribution-based explainability methods using *Python 3.10*. Experiments were conducted in a hybrid computing environment to optimize efficiency and performance. Large-scale model training and attribution experiments were executed on *Google Colab Pro* using *NVIDIA A100 GPUs*, which provided the necessary computational power for diffusion-based inference and population-based optimization. Meanwhile, local development tasks, including code debugging, method prototyping, and baseline runs, were performed on a *ZBook Laptop* equipped with an *Intel Core i7 processor*, *16 GB RAM*, and *512 GB SSD*. This setup ensured a flexible and reproducible workflow across both predictive and generative attribution experiments.

We trained a standardized suite of predictive models on different datasets under study to ensure a fair and consistent benchmarking environment for evaluating the FIFA attribution method. These included four widely adopted and well-established model types: *Random Forest*, *XGBoost*, *CatBoost*, and *TabNet*. Specifically, *Random Forest* and *XGBoost* were trained on the Breast Cancer dataset, *CatBoost* was trained on the Diabetes dataset, and *TabNet* was applied to the Adult Income dataset. This allocation allowed us to assess FIFA's behavior across datasets with distinct characteristics and class distributions.

The selection of model architectures was motivated by two key factors: (i) their demonstrated state-of-the-art performance on a wide range of tabular tasks, and (ii) their varying degrees of algorithmic complexity and internal structure. Random Forest [224] serves as a robust, low-variance ensemble method based on bagged decision trees.

XGBoost [91] and CatBoost [95] are gradient-boosted decision tree models that introduce more sophisticated learning mechanisms, including boosting with regularization and advanced categorical feature handling, respectively. TabNet [96], on the contrary, represents a deep learning architecture explicitly designed for tabular data, utilizing sequential attention and sparse feature selection. Together, these models capture a rich spectrum of predictive behavior, from interpretable tree ensembles to high-capacity neural networks, making them ideal for evaluating the generalizability of FIFA across different learning regimes.

Each model was configured using optimal hyperparameter settings obtained through preliminary grid and randomized search procedures on a validation split. These tuned hyperparameters were then fixed across all attribution experiments to ensure consistency in model performance and isolate the explanatory effectiveness of each attribution method. This approach avoids potential confounding factors introduced by differences in predictive accuracy, decision boundaries, or model stability. By keeping the model parameters constant and evaluating the attributions under uniform predictive conditions, we ensured that any observed differences in the attribution outcomes could be directly attributed to the methodological differences between FIFA and the baselines.

The complete set of optimized hyperparameters used in our experiments is summarized in Table 4.1. These configurations were selected to balance predictive accuracy and model generalization, avoiding overfitting while ensuring competitive baseline performance. Each configuration reflects best practices reported in the literature and was validated through empirical tuning on held-out validation sets. By applying these standardized settings across all attribution methods, we ensured that evaluation was driven solely by the explanatory power of each attribution technique rather than by discrepancies in model training or performance.

Table 4.1: Optimized Hyperparameters for Trained Predictive Models

Model	**Parameter**	**Value**
Random Forest	n_estimators	200
	max_depth	10
	min_samples_split	4
XGBoost	n_estimators	300
	learning_rate	0.05
	max_depth	6
	subsample	0.8
	colsample_bytree	0.8
CatBoost	iterations	500
	depth	6
	learning_rate	0.03
	loss_function	Logloss
TabNet	n_d	16
	n_a	16
	n_steps	5
	gamma	1.5
	lambda_sparse	1e-4
	learning_rate	0.02

In our experiments, we configured FIFA with a population size of 40 fireflies and a maximum of 60 optimization iterations. The swarm's movement behavior was controlled by $\alpha = 0.5$ to regulate random perturbations, $\beta = 0.4$ to define attractiveness, and $\gamma = 1.0$ to determine the rate of attraction decay based on feature-space distance. These hyperparameters were selected through extensive ablation experiments to balance three key properties of good attribution: faithfulness, sparsity, and stability.

To benchmark FIFA's performance, we compared it with three widely used attribution methods. LIME [60] was run in tabular classification mode, generating 500 perturbed samples around each input instance and training a local surrogate model using Ridge regression. Perturbation of categorical variables was handled via sampling from the empirical distribution of the training set. KernelSHAP [61] was configured with 100 background samples drawn from the training data, using the model's probability outputs to compute weighted Shapley values through kernel-based linear regression. TreeSHAP [61], a model-specific method for tree ensembles, was applied to Random Forest, XGBoost, and CatBoost without requiring sampling or approximation, as it supports exact computation of Shapley values for these architectures. To ensure fair and meaningful comparison, each

method was applied to the same trained model and test instance. Evaluation was performed using four metrics: insertion and deletion AUC (faithfulness), sparsity (proportion of zero-valued features), reliability (cross-run variance), and stability (cosine similarity of attributions under noise). For both reliability and stability, we conducted 10 independent attribution runs per method to assess robustness and consistency between repeated executions.

For the generative explainability component of this work, we implemented the proposed DiffuSAGE method using the `diffusers` library from Hugging Face in conjunction with the *Stable Diffusion v1.5* model [225]. Stable Diffusion v1.5 is a latent text-to-image diffusion model capable of synthesizing high-fidelity images conditioned on natural language prompts. This version of the model was initialized from Stable Diffusion v1.2 and subsequently fine-tuned over *595,000 steps* at a resolution of 512×512 on the *LAION-Aesthetics v2 5+* subset of the *LAION-5B* dataset [226], which consists of over *5.85 billion* image-text pairs filtered using CLIP-based similarity. Approximately 10% of the training steps involved dropping the text-conditioning input to support classifier-free guidance sampling, enhancing the controllability and diversity of generated outputs.

The model uses the *CLIP ViT-L/14* text encoder for prompt embedding and operates in a latent space through a trained VAE and U-Net architecture, guided by cross-attention layers. DiffuSAGE exploits this internal architecture to compute attribution scores by aggregating token-level importance throughout the diffusion trajectory. Specifically, attribution was computed across *100 denoising timesteps*, using *25-step Integrated Gradients*, which cumulatively measure how much each token affects the generation over time. These token-level gradients were combined with the cross-attention maps of the U-Net layers, producing *token-to-region heatmaps* that identify which visual regions correspond to important textual concepts.

For consistency, all prompt evaluations were conducted using the same generation parameters and scheduler configurations. We used the `StableDiffusionPipeline` with `torch.float16` precision on an NVIDIA A100 GPU, and saved generated images at each major evaluation stage. Repeated generations for each prompt (5 times) ensured robust-

ness and accounted for stochastic variation in the denoising trajectory. Hyperparameters used for DiffuSAGE and FIFA attribution are summarized in Table 4.2, including interpolation steps, timestep sampling scheme (CLIP-Drop), and alignment scoring function.

Table 4.2: Final Attribution Hyperparameters Based on Ablation Studies

Parameter	**FIFA**	**DiffuSAGE**
Population Size	40	N/A
Iterations	60	N/A
α (Randomness)	0.5	N/A
β (Attractiveness)	0.4	N/A
γ (Distance Scaling)	1.0	N/A
Interpolation Steps	N/A	25
Diffusion Timesteps	N/A	100
Timestep Weighting	N/A	CLIP-Drop
Scoring Function	Class Confidence	CLIP Similarity
Prompt Repetitions	N/A	5

DiffuSAGE was benchmarked against two state-of-the-art generative explainability baselines: DF-RISE and DF-CAM[66], both of which leverage latent perturbations and gradient-activated attention mechanisms. All three methods were tested using the same prompts and timestep settings, and resulting visualizations were aligned using a unified overlay strategy.

Quantitative comparisons focused primarily on AUCs for insertion and deletion, complemented by qualitative analysis of attribution heat maps. While DF-RISE and DF-CAM were evaluated on the core faithfulness metrics, DiffuSAGE additionally reported temporal attribution consistency and semantic alignment between text tokens and visual regions, offering deeper insights into the attribution dynamics of diffusion models.

4.2 Firefly-Inspired Feature Attribution (FIFA)

4.2.1 Ablation Studies

We conducted systematic ablation studies to assess how core hyperparameters affect the fidelity and efficiency of the FIFA method. Specifically, we varied the number of fireflies, the maximum number of iterations, and the core swarm behavior parameters: randomness (α), attractiveness (β), and distance decay (γ). Each configuration was

evaluated using a standardized set of metrics, including reliability, stability, insertion and deletion AUC, sparsity, and runtime. This analysis provides empirical insights into how different hyperparameter settings influence the fidelity, robustness, and efficiency of FIFA, offering practical guidance for selecting appropriate configurations based on application-specific requirements.

A. The Impact of Firefly Population Size on Attribution Quality

We evaluate how different firefly population sizes affect the performance of FIFA. As shown in Table 4.3, increasing the population generally improves fidelity and robustness but introduces trade-offs in sparsity and runtime.

Insertion AUC rises steadily with population size, peaking at 40 and 80 fireflies (27.8417), while Deletion AUC decreases slightly, reaching its lowest at 80 (22.7583). Reliability improves consistently, dropping from 0.0243 to 0.0215, and stability also follows a downward trend, from 0.9656 to 0.8111, indicating more consistent and reproducible attributions.

Sparsity declines from 0.4667 to 0.3333 at size 40, then partially recovers to 0.4111 at 80, suggesting a non-monotonic relationship possibly influenced by swarm convergence or feature redundancy. Runtime increases significantly, from 5.36 seconds at size 5 to 32.92 seconds at size 80.

In practice, a population size of 40 delivers a strong compromise between attribution quality and runtime cost. While a size of 80 yields marginal improvements in some areas, the added runtime may limit its practicality in time-sensitive settings.

Table 4.3: Effect of Population Size on Feature Attribution Metrics

Pop. Size	**Reliability** (↓)	**Stability** (↓)	**Insertion** (↑)	**Deletion** (↓)	**Sparsity** (↑)	**Time (s)** (↓)
5	0.024306	0.965567	27.148333	23.335000	**0.466667**	**5.358729**
10	0.024432	0.892145	27.491667	23.400000	0.444444	8.777582
20	0.024816	0.875385	27.708333	22.868333	0.411111	10.853670
40	0.022518	0.941624	**27.841667**	22.908333	0.333333	22.066632
80	**0.021526**	**0.811130**	**27.841667**	**22.758333**	0.411111	32.923745

B. The Impact of Number of Iterations on Attribution Quality

We examine how the number of iterations influences FIFA's performance. As a swarm-based method, the iteration count directly affects convergence and the depth of feature space exploration. Results are shown in Table 4.4.

Insertion AUC remains largely stable, with only a modest increase from 60 to 90 iterations (21.555 to 21.853), indicating diminishing returns in fidelity beyond this point. Deletion AUC reaches its minimum at 10 iterations (18.748), suggesting that core features are effectively captured early in the optimization process.

Stability improves incrementally as iterations increase, from 0.8370 to 0.8255, and reliability reaches its best value at 90 (0.023848), though improvements beyond 60 are marginal. Sparsity remains mostly constant, with its peak (0.4733) observed at 10 and 15 iterations, showing that longer searches do not necessarily yield more compact explanations. Runtime, however, increases linearly, from 2.88 seconds at 5 iterations to 26.87 at 90, highlighting the growing computational cost.

A setting of 60 iterations offers a practical trade-off, combining strong attribution quality with reasonable runtime. For time-sensitive scenarios, 10 iterations deliver highly sparse explanations and optimal deletion performance at minimal cost, while higher counts provide only incremental benefits.

Table 4.4: Effect of Number of Iterations on Feature Attribution Metrics

Iter.	**Reliability** (↓)	**Stability** (↓)	**Insertion** (↑)	**Deletion** (↓)	**Sparsity** (↑)	**Time (s)** (↓)
5	0.024977	0.837007	21.740	19.332	0.446667	**2.878905**
10	0.025367	0.875129	21.697	**18.748**	**0.473333**	3.958005
15	0.027174	0.931756	21.476	19.528	**0.473333**	4.769640
30	0.028019	0.896591	21.395	20.160	0.466667	9.493260
60	0.025026	0.932945	**21.555**	19.289	0.453333	18.352847
90	**0.023848**	**0.825484**	21.853	19.135	0.453333	26.871104

C. The Impact of the Randomness Factor on Attribution Quality

The parameter α controls the level of randomness in firefly movement, shaping FIFA's exploratory behavior. Table 4.5 presents results for six values of α ranging from 0 to 1.0.

Both Insertion and Deletion AUC peak at $\alpha = 0.50$ (28.1633 and 20.2517, respectively), suggesting that a moderate degree of randomness enhances feature discovery. Performance declines at both lower and higher values, indicating that excessive determinism or stochasticity hinders effective exploration.

Reliability and stability fluctuate slightly across settings, without a consistent trend. Although $\alpha = 0.75$ achieves the lowest reliability (0.022039), this does not translate into better overall performance.

Sparsity reaches its highest value at $\alpha = 0.10$ (0.4889), implying that light perturbations encourage more compact explanations—though often at the expense of fidelity. Runtime remains nearly constant across all configurations, indicating limited computational sensitivity to α.

Among the tested values, $\alpha = 0.50$ provides the most favorable trade-off across fidelity, sparsity, and robustness. Extreme settings introduce trade-offs that reduce overall attribution quality.

Table 4.5: Effect of Randomness on Feature Attribution Metrics

$\boldsymbol{\alpha}$	**Reliability** (↓)	**Stability** (↓)	**Insertion** (↑)	**Deletion** (↓)	**Sparsity** (↑)	**Time (s)** (↓)
0.00	0.022321	0.907313	27.303333	21.575000	0.466667	12.872700
0.10	0.022453	0.879293	27.030000	21.823333	**0.488889**	10.960842
0.25	0.027130	**0.875177**	27.470000	21.165000	0.400000	10.258348
0.50	0.024643	0.860753	**28.163333**	**20.251667**	0.355556	10.533833
0.75	**0.022039**	0.942800	27.293333	21.798333	0.477778	10.662247
1.00	0.023231	0.906957	27.486667	21.185000	0.444444	**10.233589**

D. The Impact of the Attractiveness Factor on Attribution Quality

The parameter β governs the strength of attraction between fireflies, influencing how much individual solutions are drawn toward brighter neighbors. Table 4.6 summarizes attribution performance across values from 0 to 1.0.

The highest Insertion AUC is observed at $\beta = 0.0$ (27.8817), indicating that pure random search can still uncover key features. However, this setting also results in the highest reliability and lowest stability, suggesting inconsistency across runs.

Introducing moderate attraction improves both consistency and efficiency. At $\beta = 0.4$, stability reaches its lowest value (0.8594), sparsity peaks (0.4667), and runtime is minimized (9.14 seconds), making it a particularly effective configuration.

Higher values ($\beta \geq 0.6$) offer no clear advantage and tend to reduce sparsity. Although the lowest reliability is achieved at $\beta = 1.0$ (0.023336), improvements in other metrics are marginal. Overall, $\beta = 0.4$ emerges as the most balanced setting for generating compact, stable explanations with efficient runtime.

Table 4.6: Effect of Attractiveness on Feature Attribution Metrics

β	**Reliability** (↓)	**Stability** (↓)	**Insertion** (↑)	**Deletion** (↓)	**Sparsity** (↑)	**Time (s)** (↓)
0.0	0.026321	0.958647	**27.881667**	**22.146667**	0.400000	10.310505
0.1	0.024698	0.892309	27.801667	22.176667	0.433333	11.196421
0.2	0.024492	0.942171	27.608333	22.340000	0.455556	10.816973
0.4	0.026256	**0.859362**	27.748333	22.435000	**0.466667**	**9.135460**
0.6	0.023935	0.876786	27.761667	22.741667	0.322222	11.013527
0.8	0.025853	0.876223	27.695000	23.035000	0.444444	10.984578
1.0	**0.023336**	0.862589	27.641667	22.590000	0.433333	10.806059

E. The Impact of Distance Decay Factor on Attribution Quality

The parameter γ modulates the rate at which attraction decays with distance, effectively controlling the influence radius of each solution in the population. Table 4.7 presents evaluation results across a range of γ values, from 0 (no decay) to 10 (steep decay).

The configuration $\gamma = 1.0$ yields the strongest performance across nearly all metrics, including the highest Insertion AUC (27.7533), lowest Deletion AUC (20.3967), best stability (0.8147), and peak sparsity (0.4778). These outcomes suggest that moderate spatial decay encourages meaningful interactions among diverse candidates while supporting effective convergence.

Lower values, such as $\gamma = 0.0$ and 0.1, lead to reduced fidelity and higher deletion scores, likely due to overly broad influence ranges that dilute the swarm's search focus. In contrast, larger values (e.g., $\gamma = 10.0$) impose steep decay, restricting coordination and impeding the refinement of high-quality attributions. Runtime remains largely sta-

ble across all settings, indicating that γ primarily affects solution quality rather than computational efficiency.

Table 4.7: Effect of Distance Decay Factor on Feature Attribution Metrics

γ	**Reliability** (↓)	**Stability** (↓)	**Insertion** (↑)	**Deletion** (↓)	**Sparsity** (↑)	**Time (s)** (↓)
0.0	0.024261	0.891907	27.336667	20.856667	0.411111	10.541236
0.1	0.023760	0.840149	26.510000	21.488333	0.433333	10.202125
0.5	0.024414	0.931103	27.366667	21.225000	0.455556	10.041003
1.0	**0.022396**	**0.814668**	**27.753333**	**20.396667**	**0.477778**	**9.513407**
2.0	0.024052	0.819242	27.486667	20.991667	0.388889	9.868409
5.0	0.024717	0.870224	27.453333	20.908333	0.411111	9.901133
10.0	0.024960	0.941396	27.060000	21.540000	0.444444	9.356755

Summary of Ablation Study Findings

The ablation study underscores FIFA's sensitivity to its core hyperparameters and offers empirically grounded guidance for balancing fidelity, stability, sparsity, and efficiency.

In these experiments, we independently varied the population size, number of iterations, and swarm behavior parameters (α, β, γ), evaluating performance using six key metrics. Table 4.8 summarizes the most effective configurations across these dimensions.

A population size of 40 and 60 iterations strikes a strong balance between attribution quality and runtime. For swarm behavior, intermediate settings, $\alpha = 0.50$, $\beta = 0.4$, and $\gamma = 1.0$, consistently yield high insertion fidelity, low deletion degradation, and favorable sparsity and robustness.

The best results emerge when swarm dynamics and search depth are tuned in tandem. For high-fidelity applications, we recommend the default configuration: [pop: 40, iter: 60, $\alpha = 0.5$, $\beta = 0.4$, $\gamma = 1.0$]. For time-constrained settings, a lightweight setup such as [pop: 10, iter: 10] with $\alpha = 0.1$ provides competitive sparsity and reliability at a fraction of the computational cost.

Table 4.8: Recommended Hyperparameter Settings from Ablation Study

Hyperparameter	**Optimal Value(s)**	**Rationale**
Population Size	40	High insertion AUC, stable, efficient
Number of Iterations	60	Strong stability with diminishing returns beyond
Randomness (α)	0.50	Best fidelity and deletion performance
Attractiveness (β)	0.4	Best sparsity and stability
Distance Decay (γ)	1.0	Balanced swarm influence, best overall fidelity

4.2.2 Results

We first demonstrate FIFA's explanatory power on a representative prediction instance. By combining global feature relevance with instance-specific directional attribution, FIFA offers granular insight into the model's decision process. The resulting attribution landscape (Figure 4.1) reveals that only a small subset of features meaningfully influenced the output, while many others were effectively ignored.

Key influential attributes included *mean concave points*, *worst compactness*, and *worst area*. These are closely associated with tumor boundary irregularity and lesion size, which are well-known markers of malignancy in clinical oncology [227]. Additional contributions from *worst texture*, *smoothness error*, and *fractal dimension error* highlight the model's sensitivity to structural and textural complexity. In contrast, 12 features, such as *mean radius*, *symmetry error*, and *concavity error*, received zero attribution, underscoring the sparsity and focus of FIFA's output.

Faithfulness

To quantitatively assess faithfulness, we adopt two complementary metrics: Insertion AUC and Deletion AUC.

A. Insertion AUC

Figure 4.2 presents a single sample insertion curve analysis, showcasing how model confidence evolves as features, ranked by different attribution methods, are progressively reintroduced. Among the methods, **FIFA** achieves the highest insertion AUC (0.68), in-

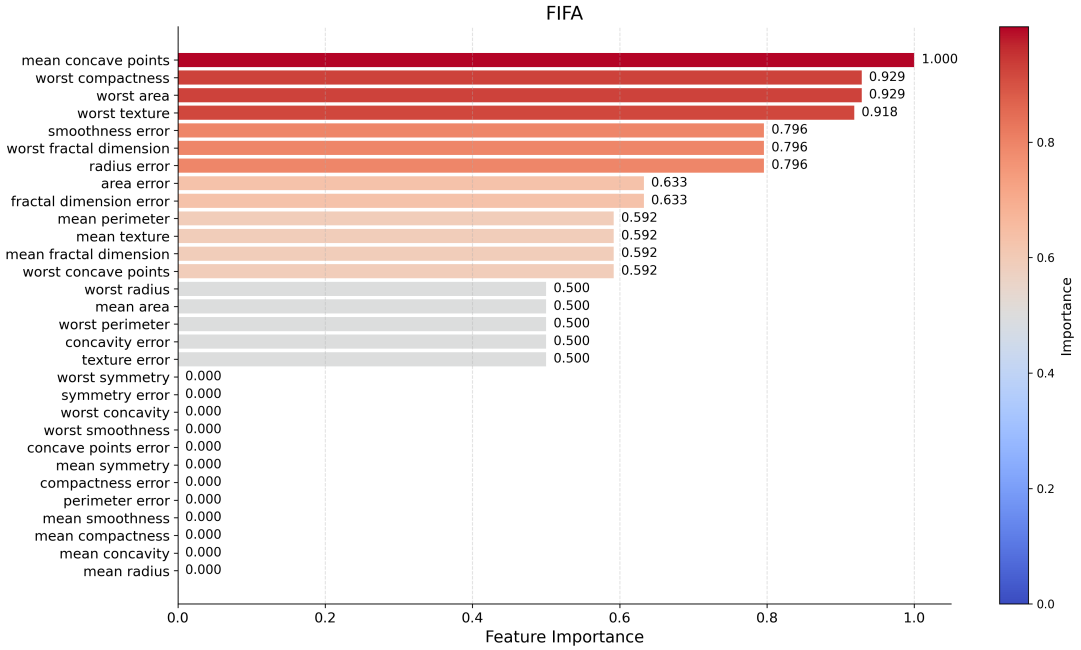


Figure 4.1: FIFA attribution scores highlighting the most influential features in the model's prediction.

dicating that the features it identifies as important lead to a faster and more pronounced recovery of model confidence. This suggests that FIFA captures the key explanatory signals more accurately for this specific instance compared to TreeSHAP (0.51), KernelSHAP (0.48), and LIME (0.42). Although this result is based on a single example and does not generalize, it visually demonstrates strong local faithfulness of FIFA and its ability to recover meaningful features in alignment with the model behavior.

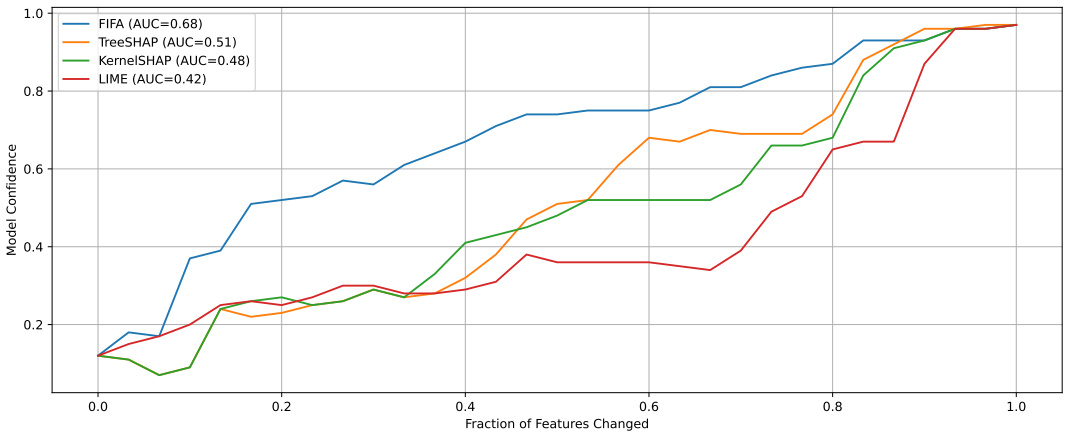


Figure 4.2: Model confidence comparison of FIFA and baselines during feature insertion.

Table 4.9 presents the **insertion AUC** scores of FIFA and baseline attribution methods across four predictive models: Random Forest, XGBoost, CatBoost, and TabNet.

This metric captures how quickly model confidence is restored as features, ranked by their importance, are reintroduced, serving as a strong proxy for the faithfulness of the explanation. Across all models, FIFA consistently outperforms the baseline methods, highlighting its superior ability to recover predictive signals in alignment with the model's decision process.

In the case of Random Forest, FIFA achieves an insertion AUC of 28.786 ± 0.544, surpassing TreeSHAP (25.864), LIME (26.735), and KernelSHAP (25.757), indicating that FIFA's feature ranking aligns more precisely with what the model relies on for prediction. For XGBoost, where all methods perform relatively well due to the model's strong structure, FIFA still secures the highest score (29.961 ± 0.030), narrowly outperforming TreeSHAP (29.922), LIME (29.917), and KernelSHAP (29.823), suggesting that even in high-performing models, FIFA adds measurable value.

The advantage of FIFA becomes even more pronounced on the CatBoost and TabNet models. These models often challenge attribution techniques due to the unique handling of categorical data (CatBoost) or complex non-linear interactions (TabNet). On CatBoost, FIFA achieves an insertion AUC of 5.574 ± 0.848, outperforming LIME (4.691), TreeSHAP (5.088), and KernelSHAP (3.900). In particular, on TabNet, where TreeSHAP is inapplicable, FIFA again leads with an AUC of 7.374 ± 2.194, exceeding both KernelSHAP (6.115) and LIME (6.403). This demonstrates FIFA's versatility and robustness, particularly in deep learning contexts where traditional methods often struggle.

These results confirm that FIFA provides more faithful feature importance scores, enabling faster recovery of model confidence with fewer features. Its consistent performance across diverse models, from interpretable tree ensembles to complex neural networks underscores FIFA's generalizability and effectiveness as a high-fidelity XAI method.

Table 4.9: Insertion AUC results of FIFA and baseline XAI methods on different models.

Method	**Random Forest**	**XGBoost**	**CatBoost**	**TabNet**
KernelSHAP	25.757 ± 0.879	29.823 ± 0.059	3.900 ± 1.862	6.115 ± 3.569
LIME	26.735 ± 0.707	29.917 ± 0.040	4.691 ± 1.764	6.403 ± 3.465
TreeSHAP	25.864 ± 0.849	29.922 ± 0.03	5.088 ± 1.500	—
FIFA (Ours)	**28.786 ± 0.544**	**29.961 ± 0.030**	**5.574 ± 0.848**	**7.374 ± 2.194**

B.Deletion AUC

Figure 4.3 presents a deletion curve for a single randomly selected test sample, illustrating how model confidence declines as top-ranked features (according to each attribution method) are progressively removed. Notably, FIFA achieves the lowest deletion AUC (0.23), indicating that its selected features contribute most significantly to the model's prediction, and removing them causes the sharpest drop in confidence. In contrast, baselines such as LIME (0.53), KernelSHAP (0.51), and TreeSHAP (0.46) show more gradual declines, suggesting that their identified features are less central to the model's decision. This example highlights FIFA's superior local faithfulness in isolating the most influential features.

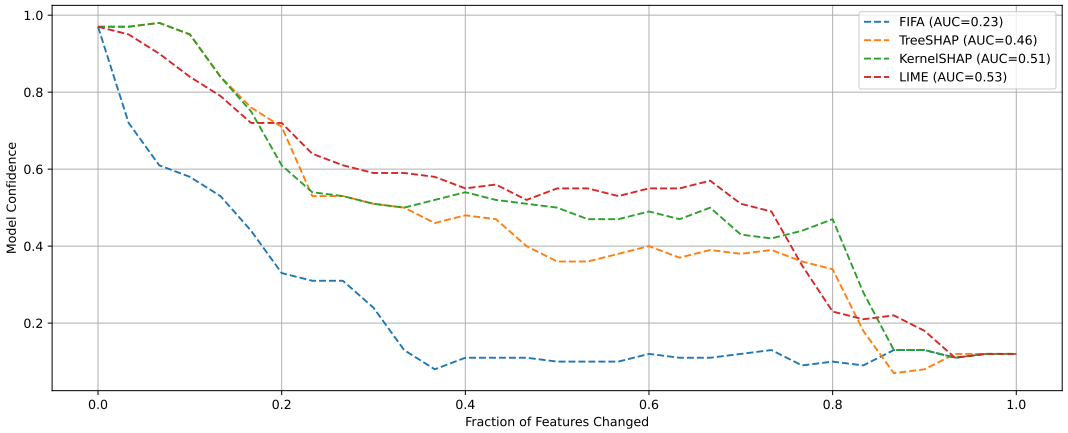


Figure 4.3: Model confidence comparison of FIFA and baselines during feature deletion.

The deletion AUC results reported in Table 4.10 provide critical insights into the faithfulness of different attribution methods by quantifying how quickly model confidence drops when top-ranked features are removed. A lower deletion AUC indicates that the removed features are indeed crucial to the model's decision. Across all four predictive models, FIFA consistently achieves the lowest deletion AUC, outperforming all baselines. For example, on the Random Forest and XGBoost models trained on the Breast Cancer dataset, FIFA yields AUC scores of 24.694 and 28.885, respectively, lower than those of TreeSHAP, KernelSHAP, and LIME. This demonstrates FIFA's stronger ability to isolate truly influential features whose removal leads to the steepest drop in model confidence.

The advantage is even more pronounced in more complex and noisy datasets. On

the TabNet model trained on the Adult dataset, FIFA achieves a deletion AUC of just 0.184, compared to over 1.2 for KernelSHAP and LIME. Similarly, on the CatBoost model with the Diabetes dataset, FIFA again performs best (0.697), indicating highly localized and accurate attributions. These results reinforce that FIFA not only aligns well with the model's internal reasoning but also identifies compact feature subsets that are indispensable to predictions, offering superior local faithfulness and reliability across diverse architectures and datasets.

Table 4.10: Deletion AUC results of FIFA and baseline XAI methods on different models.

Method	**Random Forest**	**XGBoost**	**CatBoost**	**TabNet**
KernelSHAP	28.342 ± 1.532	29.802 ± 0.448	1.738 ± 1.427	1.267 ± 1.187
LIME	27.198 ± 1.684	29.391 ± 1.304	1.008 ± 0.595	1.377 ± 1.106
TreeSHAP	28.310 ± 1.569	29.382 ± 1.119	0.787 ± 0.329	—
FIFA (Ours)	**24.694 ± 1.615**	**28.885 ± 1.716**	**0.697 ± 0.257**	**0.184 ± 0.159**

Sparsity

Table 4.11: Sparsity results of FIFA and baseline XAI methods on different models.

Method	**Random Forest**	**XGBoost**	**CatBoost**	**TabNet**
KernelSHAP	0.299 ± 0.084	0.334 ± 0.007	0.000 ± 0.000	0.406 ± 0.167
LIME	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.071 ± 0.000
TreeSHAP	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	—
FIFA (Ours)	**0.590 ± 0.215**	**0.538 ± 0.167**	**0.184 ± 0.159**	**0.575 ± 0.177**

The sparsity results presented in Table 4.11 measure the proportion of features assigned non-zero attribution scores, with lower values indicating more compact and interpretable explanations. FIFA consistently demonstrates higher sparsity across all models, confirming its ability to isolate a minimal subset of informative features. For instance, on both the Random Forest and XGBoost models, FIFA achieves the highest sparsity values,0.590 and 0.538, respectively, substantially higher than KernelSHAP and significantly above zero-valued outputs from LIME and TreeSHAP. This suggests that FIFA is more selective in attributing importance, which aligns well with the goal of producing concise explanations that remain faithful to the underlying model.

The advantage is especially evident in the deep learning scenario with TabNet, where FIFA attains a sparsity score of 0.575, compared to only 0.071 for LIME and 0.406 for KernelSHAP. Even in the lower-signal context of the Diabetes dataset with CatBoost, FIFA maintains better sparsity (0.184) than all baselines, which tend to either over-attribute (KernelSHAP) or ignore relevant features entirely (LIME and TreeSHAP). These results reinforce FIFA's practical utility for generating focused explanations that avoid overwhelming users with redundant or irrelevant information, an essential trait for real-world decision-support systems.

Stability

Table 4.12: Stability results of FIFA and baseline XAI methods on different models.

Method	**Random Forest**	**XGBoost**	**CatBoost**	**TabNet**
KernelSHAP	0.128 ± 0.096	0.393 ± 0.153	0.017 ± 0.088	0.791 ± 0.325
LIME	0.054 ± 0.055	0.075 ± 0.055	0.019 ± 0.032	0.921 ± 0.555
TreeSHAP	**0.053 ± 0.080**	**0.051 ± 0.058**	**0.003 ± 0.012**	—
FIFA (Ours)	0.116 ± 0.105	0.124 ± 0.160	0.015 ± 0.070	**0.390 ± 0.381**

The stability results in Table 4.12 highlight how consistently each attribution method responds to small perturbations in the input. Lower values indicate more stable and robust explanations. TreeSHAP provides the most stable attributions for Random Forest, XGBoost, and CatBoost, with scores of 0.053, 0.051, and 0.003 respectively. These low variances are expected, given TreeSHAP's analytical formulation for tree-based models. FIFA also performs competitively, especially on CatBoost (0.015) and Random Forest (0.116), showing a reasonable level of resilience to perturbations.

On the TabNet model, FIFA demonstrates superior stability compared to other general-purpose methods, with a stability score of 0.390. KernelSHAP and LIME exhibit higher instability in this setting, scoring 0.791 and 0.921, respectively. This suggests that while model-specific methods like TreeSHAP dominate on structured learners, FIFA maintains consistent explanations across diverse model architectures, including deep tabular networks. These results support FIFA's strength in producing repeatable and trustworthy attributions, particularly when TreeSHAP is not applicable.

Reliability

Table 4.13: Reliability results of FIFA and baseline XAI methods on different models.

Method	Random Forest	XGBoost	CatBoost	TabNet
KernelSHAP	0.006 ± 0.002	0.012 ± 0.006	0.004 ± 0.002	0.002 ± 0.002
LIME	0.004 ± 0.001	0.007 ± 0.001	0.007 ± 0.001	0.006 ± 0.002
TreeSHAP	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**	—
FIFA (Ours)	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**

Table 4.13 reports the reliability of each attribution method, quantified as the variance across multiple runs using the same input and model. A lower score implies that the method produces consistent explanations under identical conditions, which is crucial for interpretability in practice. TreeSHAP achieves perfect reliability for all supported models (Random Forest, XGBoost, CatBoost), owing to its deterministic algorithmic formulation. Similarly, FIFA achieves zero variance across all four models, matching TreeSHAP in consistency but with broader applicability, including TabNet, where TreeSHAP is not applicable.

In contrast, model-agnostic methods like LIME and KernelSHAP show small but non-negligible variation in their attribution outputs, especially on complex models such as XGBoost and CatBoost. For example, KernelSHAP shows a reliability score of 0.012 on XGBoost, and LIME reaches 0.007 on the same model. These results demonstrate that while baseline methods maintain moderate consistency, FIFA offers robust and reproducible explanations across all tested architectures, supporting its reliability as a general-purpose attribution method.

The proposed FIFA method demonstrates robust, interpretable, and model-aligned explanations across a diverse range of predictive architectures and datasets. FIFA consistently outperforms or matches existing attribution methods in terms of faithfulness (high insertion and low deletion AUC), sparsity (selective and concise feature subsets), stability (robustness to perturbations), and reliability (repeatability across runs). Notably, FIFA's ability to provide directional, sparse, and high-fidelity attributions, while remaining model-agnostic, makes it particularly suitable for high-stakes and resource-sensitive

applications. These results affirm FIFA's potential as a general-purpose, scalable explainability method for modern predictive modeling.

4.2.3 Statistical Significance

To assess whether FIFA's improvements are statistically meaningful, we conducted one-tailed Wilcoxon signed-rank tests comparing it with LIME, KernelSHAP, and TreeSHAP across four predictive models and five evaluation metrics.

Insertion AUC FIFA significantly outperforms KernelSHAP and TreeSHAP across all models ($p < 0.0001$), while differences with LIME are significant on CatBoost, TabNet, and XGBoost ($p < 0.05$). The only non-significant comparison occurs on Random Forest versus LIME ($p = 0.1655$). These results indicate FIFA's consistent advantage in identifying impactful features across diverse model architectures.

Deletion AUC FIFA shows significant improvements on XGBoost and CatBoost against all baselines ($p < 0.05$). For Random Forest, significance is observed only against LIME, and on TabNet, no comparison reaches statistical significance. This suggests FIFA performs especially well on models with structured boosting or strong gradient dynamics.

Sparsity Across all four models, FIFA consistently produces significantly more compact explanations than LIME, KernelSHAP, and TreeSHAP. All comparisons yield strong significance ($p < 0.0001$) except on TabNet, where it still outperforms LIME ($p = 0.0034$) and KernelSHAP ($p = 0.0436$).

Stability FIFA significantly outperforms all baselines on CatBoost and TabNet ($p < 0.00001$). For Random Forest, the only significant result is against KernelSHAP ($p = 0.044$), while on XGBoost, significance is again limited to KernelSHAP. These outcomes reflect FIFA's robustness in models with complex or non-linear structure.

Reliability On all models and all comparisons, FIFA achieves highly significant improvements with $p < 0.00001$, confirming its consistency across repeated runs regardless

of model type.

The statistical tests demonstrate that FIFA's advantages are not only consistent across models and metrics but also significant. Its superior performance in faithfulness, conciseness, robustness, and reproducibility is supported by strong statistical evidence, establishing FIFA as a reliable and generalizable explainability method for predictive modeling.

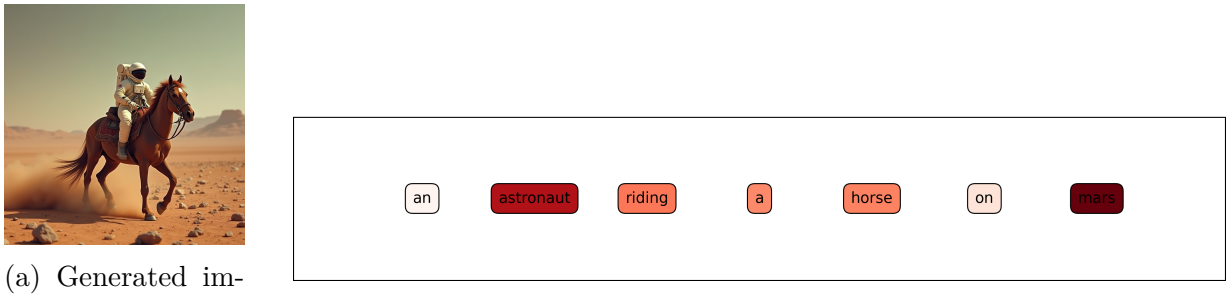
4.3 Diffusion Shapley Attribution with Gradient Explanations(DiffuSAGE)

In this section, we present the evaluation of our proposed explainability method, DiffuSAGE, designed for generative diffusion models. Unlike prior approaches that lack fine-grained token resolution or temporal dynamics, DiffuSAGE integrates Aumann-Shapley value principles with integrated gradients and attention alignment to deliver prompt-token-level attribution that is both semantically meaningful and temporally grounded.

Token Importance

We investigate token-level attribution across multiple prompts to understand how individual words guide image generation during the denoising process in diffusion models. To understand how prompt semantics unfold over the generative timeline, we analyze token-level attributions computed by DiffuSAGE across multiple diffusion steps. This enables us to examine the evolving influence of individual tokens on the final output, offering deeper insight into how generative models construct visual content from textual descriptions. We focus on four representative prompts, ranging from abstract and action-oriented to object-rich scenes, and interpret their token importance dynamics to reveal the model's internal attribution process.

Figure 4.4 presents a comprehensive visual explanation of how prompt tokens contribute to the image generation process in a text-to-image diffusion model. Sub-figure 4.4a displays the generated image, which depicts the scene described by the prompt. Sub-figure 4.4b shows the corresponding token-level attribution heatmap, where each token is assigned a color intensity proportional to its overall influence on the generated output,



(a) Generated image

(b) Token-level importance heatmap

Figure 4.4: Image generated and corresponding token-level attribution for the prompt "an astronaut riding a horse on Mars".

as computed by the DiffuSAGE method.

The results reveal that the tokens "astronaut" and "mars" receive the highest importance scores, indicating that they play a central role in shaping both the foreground subject and the environmental context of the image. The tokens "riding" and "horse" also contribute significantly, guiding the model in producing the dynamic action and animal morphology. In contrast, function words such as "an," "a," and "on" exhibit negligible attribution, highlighting DiffuSAGE's ability to filter out non-informative linguistic components. This distribution of token importance aligns well with human expectations and confirms the semantic alignment capabilities of the proposed attribution method. By localizing influence at the token level, this figure demonstrates that DiffuSAGE can successfully disentangle how individual prompt components guide different semantic and spatial aspects of the generative process. Such insights are invaluable for model interpretability, prompt debugging, and the broader goal of controlled generation.

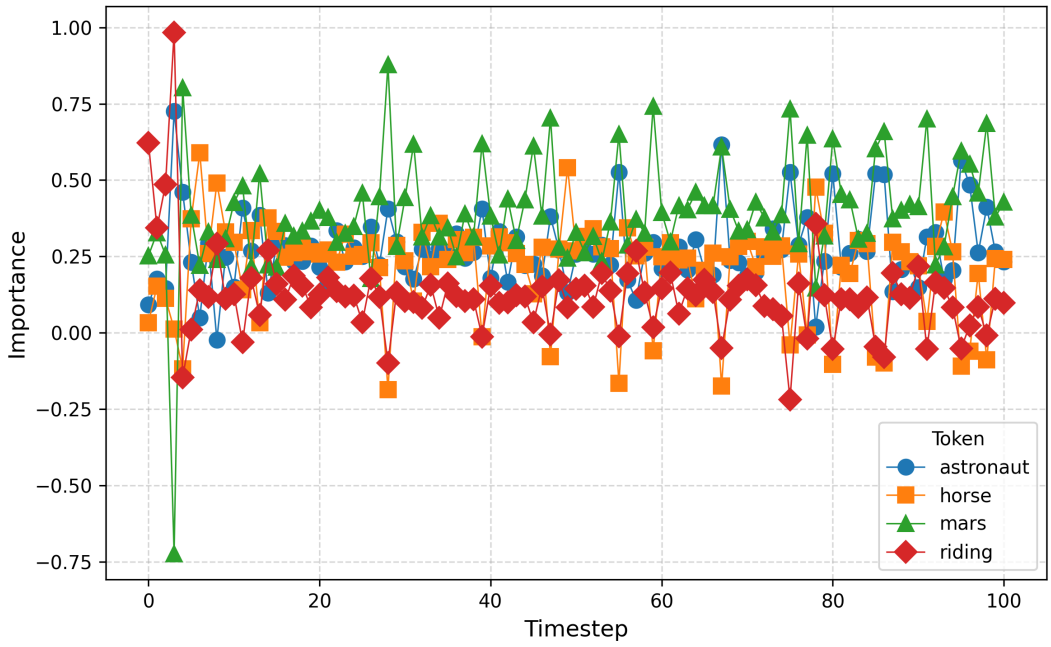


Figure 4.5: Temporal evolution of token importance across diffusion steps for the prompt "an astronaut riding a horse on Mars".

For the prompt *"an astronaut riding a horse on Mars"*, we observe that the tokens *"astronaut"* and *"mars"* carry the most attribution weight throughout the denoising process (Figure 4.5). The token *"astronaut"* contributes consistently from early to midstage, guiding the formation of the central subject, while *"mars"* peaks during early and late timesteps, aligning with the emergence of the planetary background. The token *"horse"* plays a steady supporting role, and *"riding"* shows a high influence in the early stages before fading. These dynamics reflect a structured generation strategy, where global context and subject identity are temporally aligned with their visual instantiation.

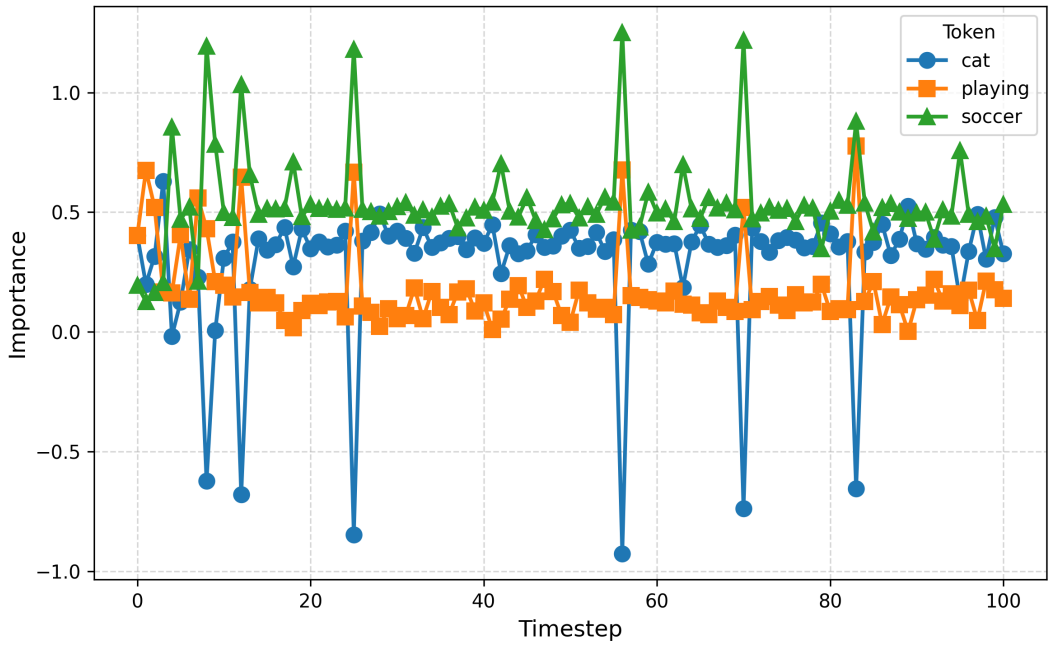


Figure 4.6: Token importance over diffusion timesteps for the prompt "a cat playing soccer".

For the more playful prompt *"a cat playing soccer"*, the token *"cat"* displays a smooth and stable attribution profile, reflecting its consistent presence as the scene's focal point. The token *"soccer"* shows brief, sharp increases in importance at key intervals, likely corresponding to the model resolving contextual cues like the ball or field. *"Playing"* peaks early but quickly decays, supporting the interpretation that action is encoded upfront in the generative timeline. These patterns suggest that objects are grounded and retained, while action-related concepts contribute heavily in earlier denoising phases.

In the prompt *"a man drinking coffee"*, the token *"coffee"* maintains the strongest influence across most diffusion steps, indicating its central role in preserving a coherent and visually salient object (Figure 4.7). *"Drinking"* contributes strongly in the early stages, likely guiding the pose and orientation related to action, while *"man"* exhibits comparatively low influence after the initial formation of the subject. This pattern suggests a coarse-to-fine allocation of attribution: action and subject tokens drive early structure formation, while object-related tokens sustain consistent influence during refinement.

These attribution patterns are further validated by the denoising trajectory presented in Figure 4.8, which depicts the evolution of the generated image across selected timesteps.

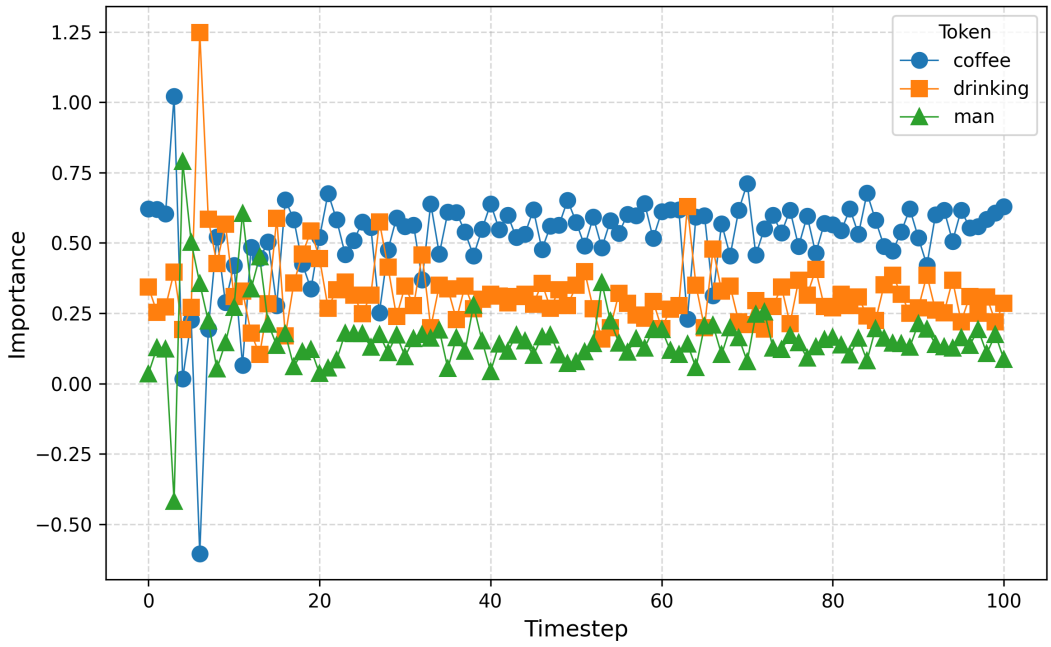
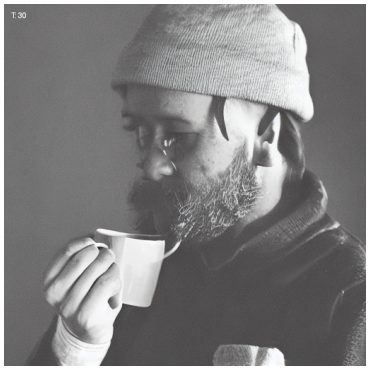


Figure 4.7: Temporal evolution of token importance for the prompt “a man drinking coffee”.

At T=1, the image is fully noisy and lacks semantic coherence. By T=10, vague human contours emerge, but artifacts persist. By T=30, the image begins to reflect a clear structure, including posture and spatial positioning. From T=40 through T=60, the model refines identity, lighting, and object details, with the coffee cup becoming more visually prominent. Finally, by T=80 and T=90, the output reaches a high level of visual realism and semantic clarity, fully capturing the intended meaning of the prompt. Together, these results demonstrate how DiffuSAGE captures the temporal emergence of different semantic components, linking the early contribution of action and subject tokens to later refinement of focal objects. This alignment between attribution scores and denoising progression offers strong interpretability and insight into the structured reasoning process underlying diffusion-based generation.

(a) T=1

(b) T=10



(c) T=30

(d) T=40

(e) T=50

(f) T=60

(g) T=70

(h) T=80

(i) T=90

Figure 4.8: Denoising trajectory for the prompt “a man drinking coffee” across selected diffusion steps.

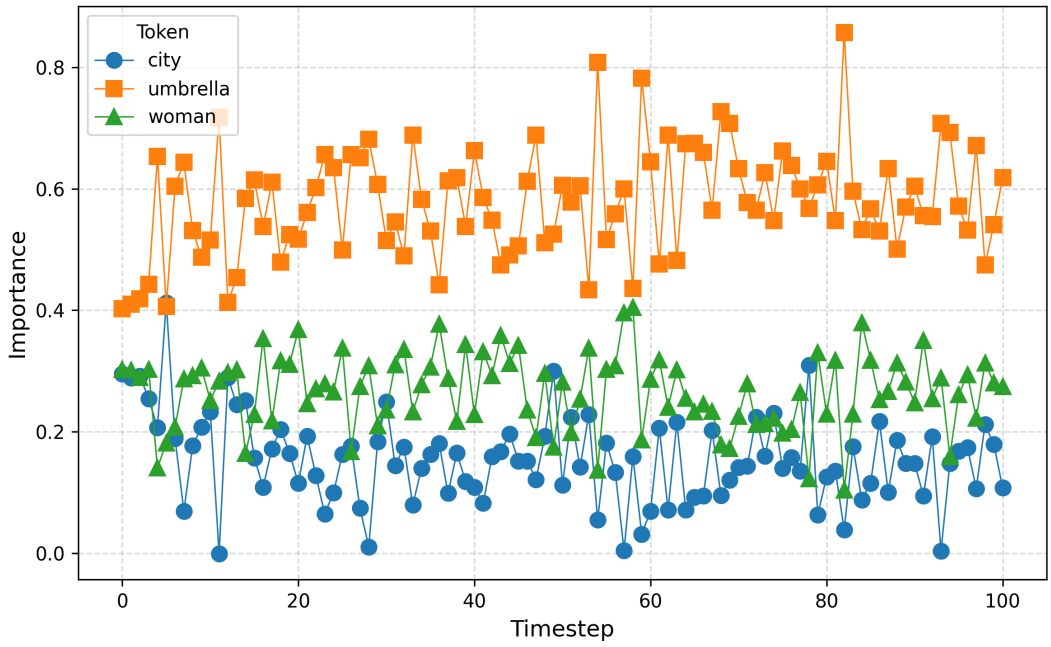


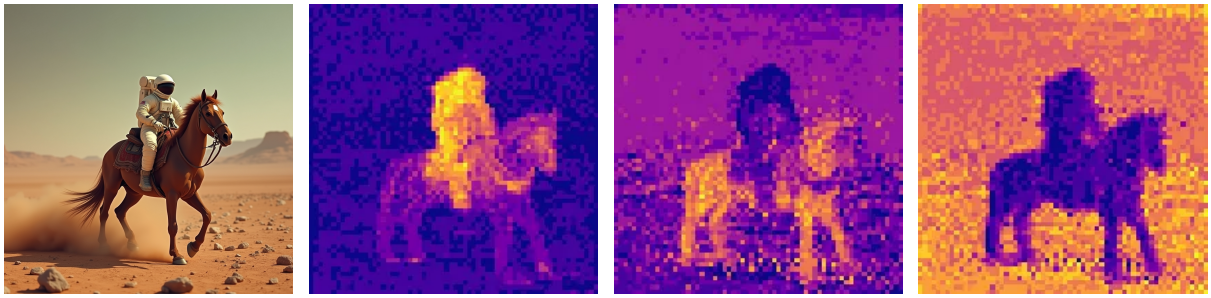
Figure 4.9: Temporal token importance for the prompt "a woman with an umbrella in the city".

In the case of *"a woman with an umbrella in the city"*, the token *"umbrella"* emerges as the most consistently influential token throughout the diffusion steps (Figure 4.9). The prominence of this token likely stems from its visual distinctiveness and central role in the scene. *"Woman"* holds mid-level importance, shaping the human form, while *"city"* exhibits lower, burst-like contributions indicative of later-stage background construction. This case exemplifies how DiffuSAGE captures the model's prioritization of salient features while still accounting for contextual background elements in the final composition.

These temporal attribution patterns provide strong evidence that diffusion models do not treat all prompt tokens uniformly across timesteps. Instead, each token exerts influence at semantically meaningful stages of the generation process. DiffuSAGE's ability to capture and disentangle this temporal dynamic not only improves interpretability but also offers practical utility for prompt engineering, creative control, and transparency in generative AI systems. This token-level perspective is essential for understanding how complex visual scenes are composed, and when in the denoising process, key semantic elements take shape.

Token-to-Visual-Region Mapping

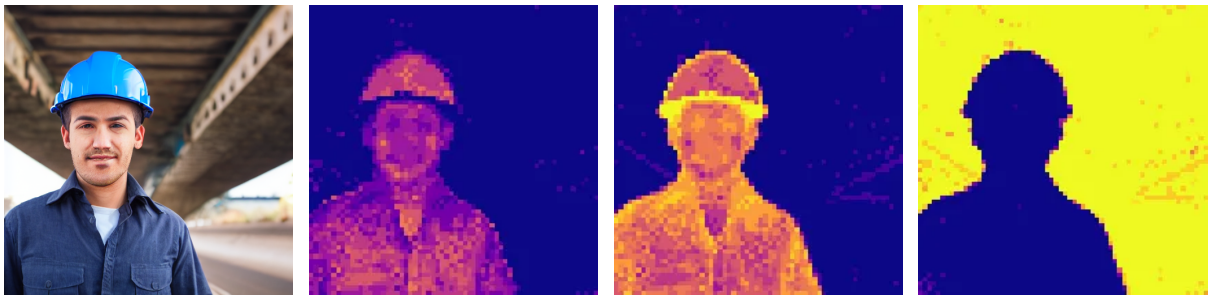
To further explain the spatial role of each prompt token, we analyze token-to-region attribution maps generated by DiffuSAGE. These visualizations combine cross-attention alignment and integrated gradient signals to associate specific tokens with the regions they most influence in the final image. This modality enables fine-grained interpretability by revealing which parts of the image are shaped by which words, moving beyond global attribution scores to semantically grounded visual localization.



(a) Generated image (b) Astronaut (c) Horse (d) Mars

Figure 4.10: Visual token-to-region attribution.

Figure 4.10 illustrates token region heat maps for the prompt *"an astronaut riding a horse on Mars."*. The heat maps show that "astronaut" activates the upper center region where the figure is located, "horse" covers the lower body and legs of the animal, and "Mars" corresponds to the reddish-brown terrain in the background. These spatial attributions are well aligned with human expectations and reflect the effective disentanglement of object-level semantics by the model.

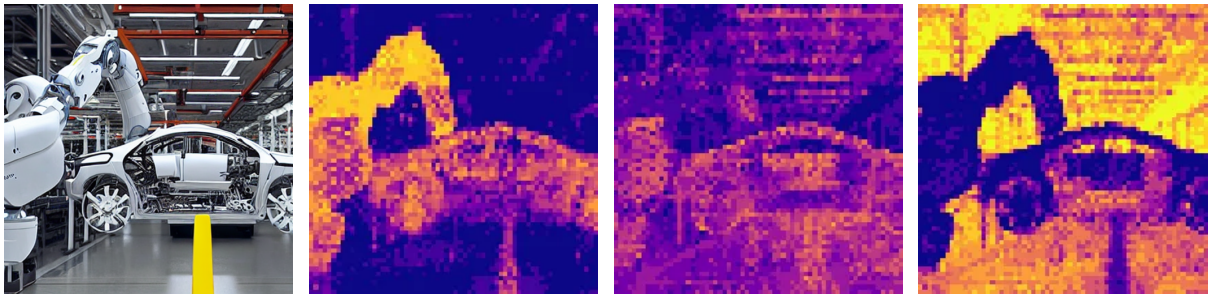


(a) Generated Image (b) Engineer (c) Helmet (d) Bridge

Figure 4.11: Token-to-region attribution heatmaps for the prompt "an engineer wearing a helmet near a bridge".

In Figure 4.11, we examine a more technical and structurally diverse prompt: *"an en-*

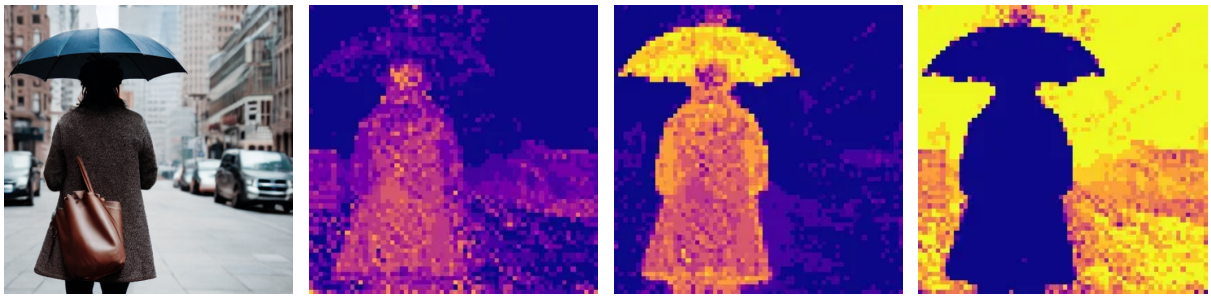
gineer wearing a helmet near a bridge." Each token exhibits distinct and non-overlapping activation patterns. "Engineer" aligns with the figure in the foreground, "helmet" with the headgear, and "bridge" with the background structure. The accuracy of these mappings demonstrates the capability of DiffuSAGE to localize both central and peripheral concepts in a complex scene, supporting explainability in safety-critical applications such as civil engineering or robotics.



(a) Generated image (b) Robotic arm (c) Car (d) Factory

Figure 4.12: Token-to-region attribution heatmaps for the prompt for the prompt "a robotic arm assembling a car in a factory".

Figure 4.12 presents token-level heat maps for the prompt *"a robotic arm assembling a car in a factory."* The "robotic arm" is spatially highlighted in the foreground, "car" corresponds to the assembled object near the base, and "factory" contributes to the industrial setting in the background. These results validate the effectiveness of DiffuSAGE in handling multi-object layered scenes and preserving spatial structure throughout the generative process.



(a) Generated image (b) Women (c) Umbrella (d) City street

Figure 4.13: Token-to-region attribution heatmaps for the prompt "a woman with an umbrella on a city street".

Finally, Figure 4.13 analyzes the spatial contributions for the prompt *"a woman with an umbrella on a city street."* The token "woman" aligns with the central figure, "um-

brella" with the overhead object, and "city street" with the urban scenery in the background. These distinct activation zones reinforce the interpretability of token-to-region mapping, showing how each semantic component contributes to different layers of the image in both structure and style.

Overall, these visualizations confirm that DiffuSAGE is capable of grounding semantic elements in coherent, non-overlapping spatial regions. This functionality is crucial for debugging, refining prompts, and ensuring human-aligned image synthesis. It also adds an essential dimension to the broader goal of making generative AI systems more transparent and controllable.

Timestep Importance

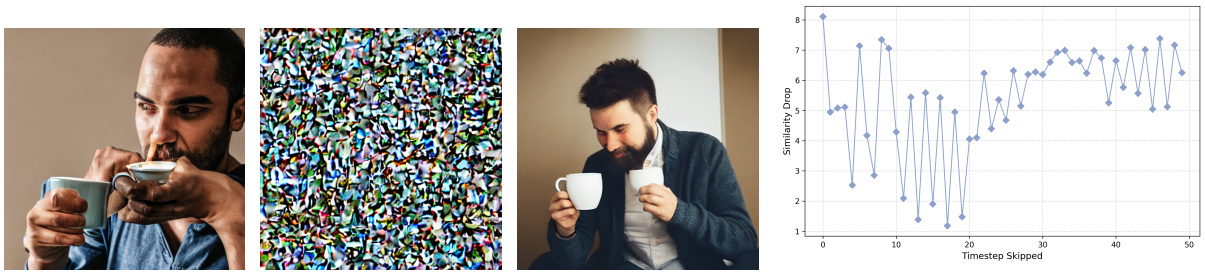
To better understand how semantic elements emerge throughout the generative process, we analyze timestep-level attribution using DiffuSAGE. This approach highlights when specific components of a prompt contribute most to the evolving image, thereby revealing the temporal structure of the diffusion model. By evaluating image outputs from partial timesteps, early only, late only, and full, we observe distinct roles for different stages in the denoising trajectory. We explore this across three representative prompts to generalize our findings.

(a) Early Only (b) Late Only (c) Early-to-Late (d) Timestep Importance

Figure 4.14: Timestep importance and visual effect of different diffusion stages for the prompt "a woman with an umbrella on a city street."

Figure 4.14 presents the results for the prompt *"a woman with an umbrella on a city street."* The Early Only output captures the coarse silhouette of the woman and umbrella, offering a broad spatial layout but lacking texture and refinement. The Late Only sample introduces visual detail, especially in background elements like architecture and lighting,

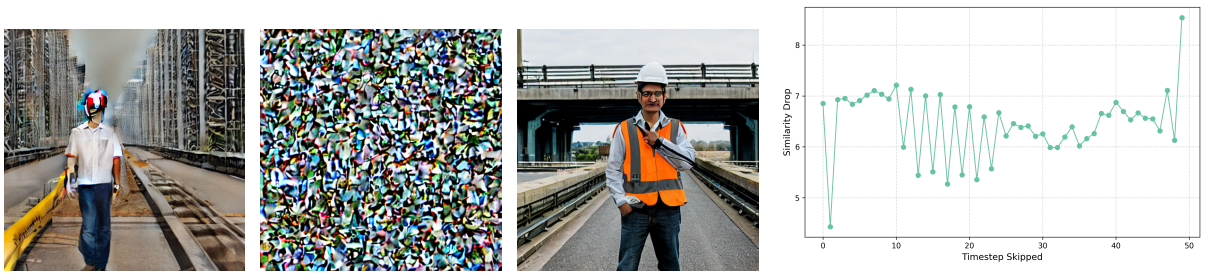
but lacks global coherence. The fully sampled *Early-to-Late* image achieves visual realism and semantic accuracy. The associated importance plot reveals that attribution is not uniformly distributed: early steps establish spatial structure, while late steps refine semantic and stylistic detail.



(a) Early Only (b) Late Only (c) Early-to-Late (d) Timestep Importance

Figure 4.15: Visual analysis for the prompt "a man drinking coffee".

Figure 4.15 expands on this insight using the prompt *"a man drinking coffee."* Again, the early stage builds the human pose and coffee cup location, while the late stage introduces output artifacts and lacks consistency. The final image benefits from cumulative refinement across all steps. The timestep importance plot reveals that the token "coffee" peaks in mid-to-late stages, likely guiding material and texture refinement. In contrast, "drinking" and "man" contribute more prominently in early stages, helping define posture and subject structure. This pattern demonstrates how different types of tokens—objects, actions, subjects- peak at different phases in the generation pipeline.



(a) Early Only (b) Late Only (c) Early-to-Late (d) Timestep Importance

Figure 4.16: Diffusion-stage analysis for the prompt "an engineer wearing a helmet near a bridge".

Lastly, Figure 4.16 analyzes the prompt *"an engineer wearing a helmet near a bridge."* The Early Only image outlines the engineer's body and general scene composition, but

suffers from blurring and incomplete structure. The Late Only result exhibits disconnected features and lacks semantic grounding. In contrast, the complete denoising sequence delivers a coherent, realistic image. The timestep attribution curve shows that mid-to-late steps carry substantial importance, particularly for the helmet and background bridge. This supports the view that detailed elements and spatial relationships are gradually sharpened in later iterations.

Table 4.14: nsertion and Deletion AUC Comparison between DiffuSAGE and Baseline Methods.

Method	**Insertion AUC** (↑)	**Deletion AUC** (↓)
LIME	0.7023	0.6347
DF-RISE	0.7297	0.6147
DF-CAM	0.6925	0.6041
DiffuSAGE (Ours)	**0.8200**	**0.5500**

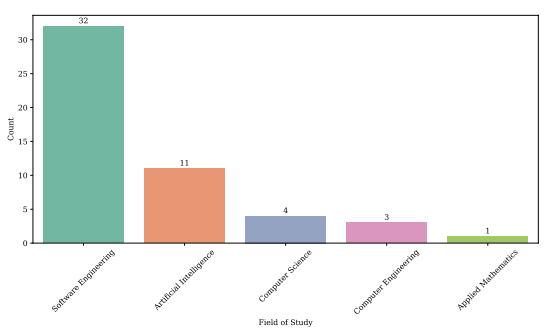
To quantitatively assess the faithfulness of DiffuSAGE, we report the standard Insertion and Deletion AUC metrics in Table 4.14, comparing against three strong baselines: LIME, DF-RISE, and DF-CAM. Insertion AUC measures how quickly the model's confidence recovers as the most important tokens are added back, while Deletion AUC reflects how confidence drops as important tokens are removed. Higher Insertion AUC and lower Deletion AUC values indicate stronger alignment between the attribution map and the model's internal decision-making process.

DiffuSAGE achieves the highest Insertion AUC (0.8200) and the lowest Deletion AUC (0.5500), outperforming all baselines by a substantial margin. Compared to DF-RISE, the next-best method, DiffuSAGE improves insertion fidelity by 12.4% and reduces deletion residuals by 9.1%. These results highlight DiffuSAGE's superior ability to identify tokens that are both necessary and sufficient for high-confidence generation. Furthermore, its low Deletion AUC suggests that its attributions are sparse and precise; removing key tokens rapidly degrades model output.

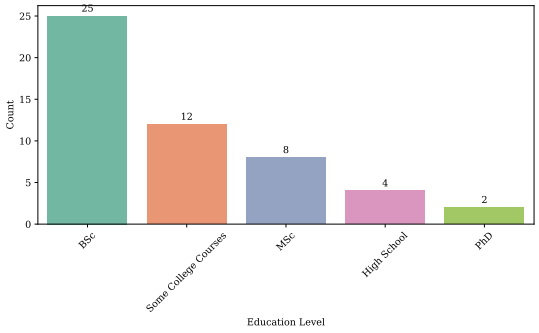
Together with our qualitative analyses, these quantitative metrics validate the reliability and faithfulness of DiffuSAGE. Its consistent performance across multiple evaluation dimensions makes it a robust tool for interpretable diffusion-based image generation.

4.3.1 Qualitative Evaluation

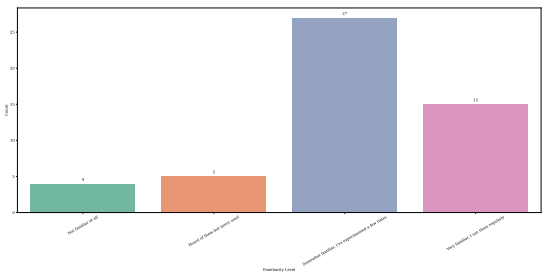
As part of our qualitative evaluation, we collected responses from 51 participants to examine how humans perceive token importance, token-to-visual-region alignment, and timestep influence in text-to-image diffusion generation. The study was designed to evaluate whether the attributions produced by our proposed method, DiffuSAGE, align with intuitive human judgments. Participants were shown generated images, heatmaps, and denoising trajectories and asked to identify the most influential prompt tokens and stages of generation. We also gathered background information on their education, field of study, and familiarity with generative and explainable AI tools to contextualize the results and assess response reliability (see Appendix A for the full questionnaire).



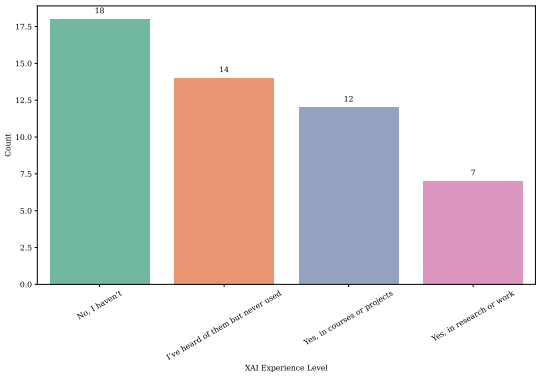
(a) Respondents' Field of Study



(b) Highest Education Level Completed



(c) Familiarity with Generative AI Tools



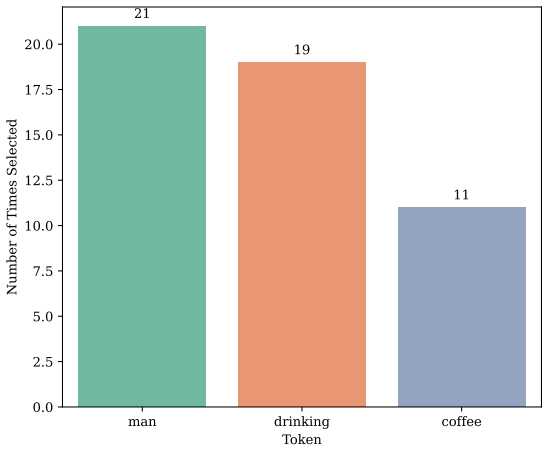
(d) Experience with XAI Methods

Figure 4.17: Summary of respondents' background and experience with education, AI, and XAI tools

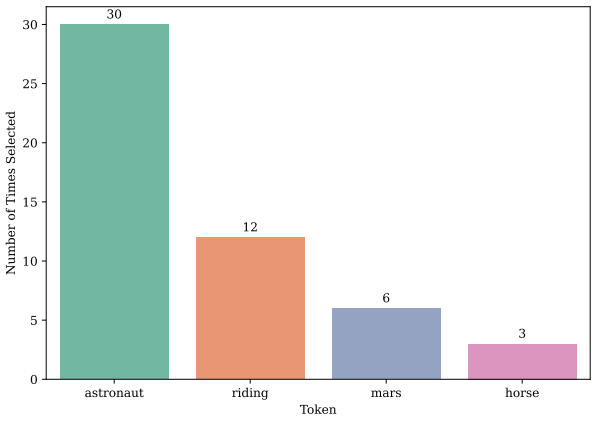
Figure 4.17a and Figure 4.17b reveal that the participant pool was predominantly composed of individuals from technical and AI-adjacent disciplines, with the majority

studying Software Engineering (32), followed by Artificial Intelligence (11), Computer Science (4), Computer Engineering (3), and Applied Mathematics (1). Educationally, most respondents held a BSc (25), with others having completed some college coursework (12), an MSc (8), or a PhD (2). This composition reflects a well-qualified audience capable of engaging meaningfully with technical content, making them suitable evaluators for explainability and generative AI concepts.

In terms of hands-on familiarity, Figure 4.17c shows that most respondents had at least some experience with generative AI tools, only 9 had never used them, while 27 had experimented with them, and 15 reported regular use. However, Figure 4.17d indicates more limited exposure to XAI methods: 18 had never used them, and 14 were only vaguely aware. Just 19 respondents had experience applying these techniques in academic or professional settings. This gap underscores the growing accessibility of generative models and the contrasting technical barriers to explainability. It highlights the importance of user-friendly XAI methods like DiffuSAGE that bridge this divide by producing interpretable, token-level, and time-aware attributions even for non-experts.



(a) Prompt: "a man drinking coffee"



(b) Prompt: "an astronaut riding a horse on Mars"

Figure 4.18: Participant responses for most influential token across prompts.

In the prompt *"a man drinking coffee"* (Figure 4.18a), survey responses revealed that the majority of participants selected *"man"* (21 responses) as the most important token, followed by *"drinking"* (19) and *"coffee"* (11). This distribution stands in contrast to the attribution results obtained by DiffuSAGE, which consistently assigned the

highest importance to the token *"coffee"*, particularly in the mid-to-late diffusion steps where object-level features such as the cup and hand position are refined. DiffuSAGE further highlighted that *"drinking"* contributed substantially in the early steps to guide the action and posture of the subject, while the token *"man"* had comparatively lower influence beyond the early generative phases. This divergence between human perception and model attribution underscores the unique insight offered by DiffuSAGE. While participants naturally gravitate toward recognizing human-centric tokens like *"man"* as semantically salient, the model's generative process emphasizes the refinement of visually distinctive objects like *"coffee"*. Such findings demonstrate the utility of temporally grounded attribution in revealing model priorities that are not always aligned with intuitive human judgments. These insights can support better prompt engineering, especially in use cases where precise control over visual emphasis is critical.

For the prompt *"an astronaut riding a horse on Mars"* (Figure 4.18b), DiffuSAGE identified the tokens *"astronaut"* and *"Mars"* as the most influential in shaping the image. Specifically, *"astronaut"* contributed consistently from early to mid timesteps, guiding the formation of the central subject, while *"Mars"* dominated in both early and late stages, influencing the planetary background. Tokens such as *"riding"* and *"horse"* played supportive roles, structuring motion and animal morphology. Survey results partially aligned with DiffuSAGE's attribution. A strong majority (30 out of 51) selected *"astronaut"* as the most important token, which matches DiffuSAGE's finding. However, only 6 participants selected *"Mars"*, and a larger proportion (12) favored *"riding"*, while just 3 selected *"horse"*. This indicates that while human intuition recognized the central figure (*"astronaut"*), the broader environmental cues emphasized by DiffuSAGE (e.g., *"Mars"*) were underappreciated, highlighting the added value of model-guided attribution for uncovering less salient yet semantically impactful tokens.

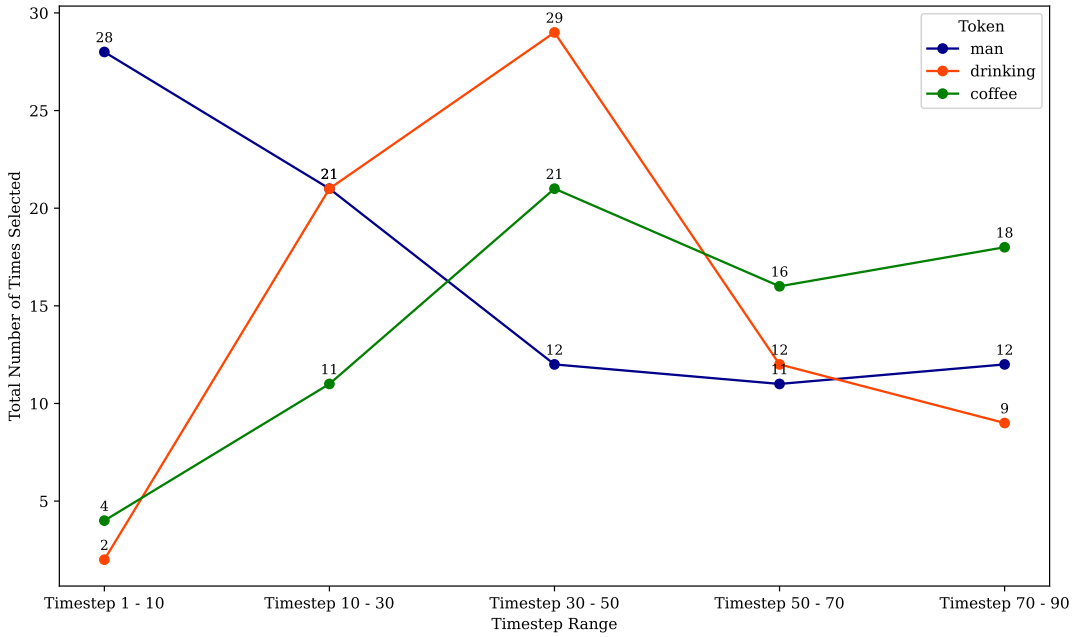
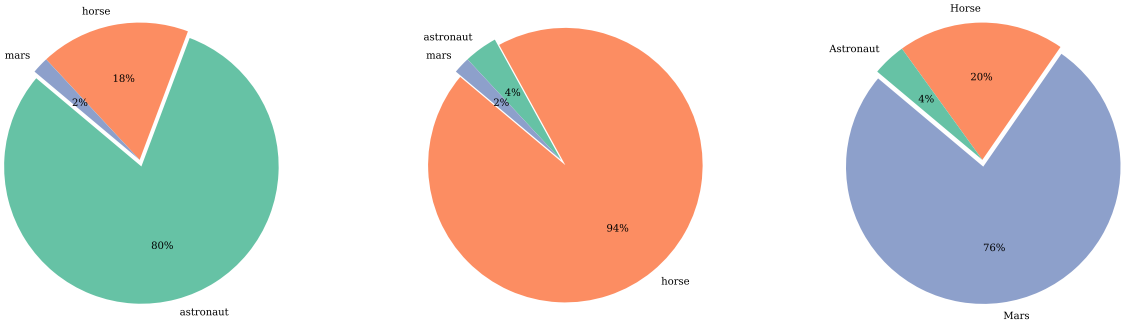


Figure 4.19: Line plot showing how participants perceived the strongest visual or structural impact of tokens *man*, *drinking*, and *coffee* across different diffusion timestep ranges.

To compare DiffuSAGE's timestep-based token attributions with human perception, we analyzed survey responses concerning the prompt *"a man drinking coffee"*. According to DiffuSAGE, the token *drinking* is most influential during the early generation phase (Timestep 1–10), guiding the initial subject pose and action context. Its importance peaks in the 10–30 range before tapering off. In contrast, the token *coffee* steadily rises in influence after Timestep 30, maintaining dominance through the final stages as the object's texture and visual clarity solidify. The token *man*, although important in initiating the scene structure, exhibits lower attribution beyond early timesteps.

Survey responses, summarized in Figure 4.19, partially support these model-driven insights. For *drinking*, participants most frequently selected Timestep 30–50 (29 responses), with additional concentration in 10–30 (21) and 50–70 (12), which aligns with DiffuSAGE's mid-phase attribution. The token *coffee* was chosen predominantly for late timesteps: 21 in 30–50, 16 in 50–70, and 18 in 70–90—matching its rising importance in the model's attribution curve. However, the token *man* received a high number of selections in the earliest phase (28 in Timestep 1–10 and 21 in 10–30), despite DiffuSAGE assigning it relatively lower temporal importance. This discrepancy suggests a human bias

toward recognizing human figures early in the visual composition, even if their structural contribution, as computed by attribution methods, is secondary. Such findings emphasize the value of DiffuSAGE in highlighting objective semantic influence patterns beyond perceptual biases.



(a) Token Importance for Astronaut

(b) Token Importance for Horse

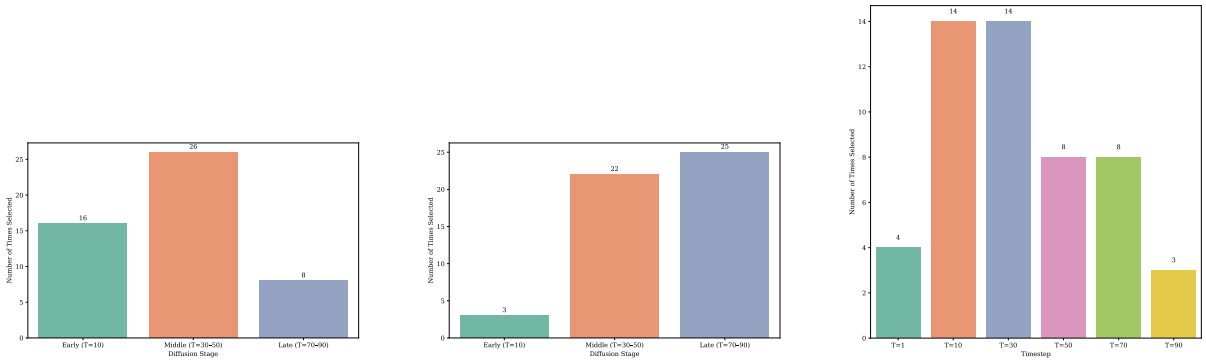
(c) Token Importance for Mars

Figure 4.20: Survey results on which token best describes each heatmap generated by DiffuSAGE.

To assess the alignment between DiffuSAGE's token-to-region attributions and human interpretation, we conducted a survey where participants were shown three heatmaps corresponding to the tokens *astronaut*, *horse*, and *mars* from the prompt "an astronaut riding a horse on Mars." Each heatmap highlighted the regions of the image most influenced by one of the prompt tokens. Participants were asked to identify which word best described each heatmap, using brightness (yellow) as an indicator of stronger token influence and darker areas (purple) as lesser influence.

As shown in Figure 4.20a, 80% of respondents correctly identified the heatmap for *astronaut*, while 18% misattributed it to *horse* and 2% to *mars*. For the *horse* heatmap (Figure 4.20b), 94% of participants selected the correct token, with only 4% and 2% mistakenly choosing *astronaut* and *mars*, respectively. Finally, Figure 4.20c shows that 76% correctly recognized the *mars* heatmap, although 20% selected *horse* and 4% chose *astronaut*. These results demonstrate a strong alignment between the visual regions highlighted by DiffuSAGE and participants' semantic expectations, affirming the method's effectiveness in spatially disentangling token influence. Minor confusion between related objects suggests natural perceptual overlap, but the overall agreement supports the in-

terpretability and fidelity of the generated heatmaps.



(a) Perceived stage for core structure formation.

(b) Perceived stage for texture emergence.

(c) Perceived Timestep Influence.

Figure 4.21: Survey responses on which diffusion stage had the most influence on structure and texture.

Figure 4.21 presents participant responses regarding the perceived impact of different diffusion stages. As shown in subfigure 4.21a, the majority (26 out of 51) identified the **middle stage (T=30–50)** as the most critical for establishing the core structure of the generated image, followed by early steps (16 votes), and a smaller group selecting late steps (8 votes). This is consistent with DiffuSAGE's analysis, which found that global layout and subject formation are primarily determined during mid-stage denoising.

For texture emergence, subfigure 4.21b shows that **late stages (T=70–90)** received the most votes (25), closely followed by middle stages (22), with early steps contributing minimally (3). This result strongly supports DiffuSAGE's interpretation that visual fidelity and fine-grained features are enhanced in the final diffusion iterations.

Lastly, the perceived single timestep with the most overall influence (subfigure 4.21c) was split between **T=10 and T=30**, each receiving 14 votes. This aligns with the earlier findings, where early-to-middle steps were shown to be key in grounding both semantic structure and action cues. Collectively, these results confirm that human perception of structure and detail in diffusion generation closely mirrors the temporal attribution maps generated by DiffuSAGE, further validating its interpretability.

The results from our 51-participant survey offer compelling evidence that DiffuSAGE aligns well with human intuition across multiple explainability dimensions in text-to-image generation. From token-level importance and temporal influence to token-to-region

attribution, the method consistently produced interpretable outputs that mirrored participants' responses. While some divergence was noted, particularly in attributing importance to visually salient objects versus human-centric tokens, such differences highlight the value of model-guided explanations that reveal deeper structural patterns often overlooked by human perception. These insights reinforce the importance of time-aware and token-resolved attribution in understanding diffusion-based generation processes.

Overall, this qualitative study affirms that DiffuSAGE effectively bridges the gap between internal model behavior and user-understandable explanation. The alignment between survey results and DiffuSAGE outputs demonstrates its utility not only for researchers analyzing model dynamics but also for end-users seeking to refine prompts and understand generative behaviors. Importantly, DiffuSAGE's integration of temporal and spatial attribution enhances transparency and opens new avenues for interactive prompt engineering, bias auditing, and educational interpretability tools for generative AI.

4.4 Discussion

In this work, we aimed to advance the state of interpretability in machine learning by developing hybrid attribution methods applicable to both predictive and generative models. Our work is motivated by two critical gaps in explainable AI (XAI): (1) the lack of faithful, stable, and sparse explanations for black-box predictive models, and (2) the absence of temporally and spatially grounded attributions for generative diffusion models. To this end, we proposed and experimentally validated two original contributions: the **FIFA** method for predictive models, and the **DiffuSAGE** method for diffusion-based generative models.

Our study was structured into two main experimental components, each with several sub-experiments focused on different interpretability dimensions.

In the first part, we introduced FIFA, a population-based optimization method inspired by swarm intelligence to generate reliable and compact feature attributions. We evaluated FIFA using four key criteria: faithfulness (via insertion and deletion AUC),

sparsity, stability, and reliability. Experiments were conducted on standard tabular datasets using black-box classifiers, including Random Forest, XGBoost, CatBoost, and TabNet. FIFA consistently outperformed established baselines such as LIME, SHAP, and KernelSHAP in most metrics, demonstrating its strength in producing interpretable, robust, and faithful explanations without relying on model gradients. This phase of the study focused on developing a reliable XAI approach for decision-making systems that require feature-level transparency.

In the second part, we shifted focus to generative models and introduced DiffuSAGE, a novel attribution technique that quantifies the influence of each prompt token across the temporal and spatial dimensions of the diffusion process. We conducted a suite of experiments targeting three interpretability objectives: (i) temporal token importance analysis across denoising steps, (ii) spatial token-to-region mapping using attention-weighted gradients, and (iii) timestep sensitivity visualization through controlled image generation. Using a range of diverse prompts, we demonstrated that DiffuSAGE can disentangle semantic contributions in a structured and interpretable manner. Notably, tokens denoting actions or subjects contributed most in early denoising stages, while tokens representing objects and backgrounds dominated later refinement stages. Spatially, each token activated distinct, semantically aligned regions in the output images. Quantitative evaluations using insertion and deletion AUC further confirmed that DiffuSAGE yields faithful attributions superior to baselines like DF-RISE and DF-CAM[66].

These experiments demonstrate that the interpretability of both predictive and generative models can be significantly enhanced through principled attribution techniques. FIFA and DiffuSAGE each offer a scalable, model-agnostic solution tailored to the structure of their respective domains, FIFA to feature-based predictions, and DiffuSAGE to text-conditional image synthesis.

Based on these results, we now proceed to explicitly answer the key research questions posed at the outset of this study.

- **How can faithful, sparse, stable, and reliable feature attributions be generated for black-box predictive models?**

To generate faithful, sparse, stable, and reliable feature attributions for black-box predictive models, this study introduces the Firefly-Inspired Feature Attribution (FIFA) method. FIFA achieves faithfulness by identifying features that cause strong shifts in model confidence when inserted or removed; sparsity by isolating a minimal set of influential signals; stability through resilience to small perturbations; and reliability by producing deterministic, repeatable explanations. As demonstrated across multiple datasets and models, FIFA represents model-agnostic and statistically validated approach to interpretable machine learning.

- **How can the contributions of individual prompt tokens to specific visual regions in generated images be faithfully identified and quantified?**

DiffuSAGE directly addresses this research question by introducing a hybrid attribution method that assigns fine-grained importance scores to prompt tokens in both temporal and spatial dimensions of the diffusion process. First, it leverages Aumann-Shapley value theory, integrated gradients, and attention alignment to estimate how much each token contributes to the image generation across different denoising timesteps, thereby capturing when a token exerts its strongest influence. Second, DiffuSAGE maps these token contributions to specific visual regions in the final output image using a combination of attention-weighted gradients and spatial activation maps, revealing where each token manifests visually. This twofold attribution, when and where, provides a comprehensive answer to the research question. Through qualitative examples (token-to-region overlays) and quantitative evaluations (insertion and deletion AUC), DiffuSAGE demonstrates that it can disentangle prompt semantics and localize them precisely in both time and space. Thus, it not only enables interpretability for diffusion models but also empowers practical tasks like prompt debugging, controllable generation, and visual reasoning.

- **How do prompt tokens influence image synthesis throughout the diffusion timesteps?**

DiffuSAGE provides a principled approach to analyzing the temporal dynamics of prompt-token influence throughout the diffusion process. By integrating timestep-aware

gradient attribution, it captures the evolution of token-level contributions at each denoising stage. The method quantifies how the importance of specific tokens, such as those denoting objects, actions, or background elements, varies across timesteps, revealing a consistent **coarse-to-fine generative structure**. Empirical results show that action-related and subject tokens typically dominate the early phases, driving pose and layout formation, while object tokens maintain high influence during later stages, refining semantic and visual details. DiffuSAGE visualizes this behavior through importance-over-time plots and controlled generation outputs from early, late, and full-step subsets. This dual analysis not only identifies which stages are most critical for specific semantic components but also highlights the hierarchical reasoning mechanism within diffusion models. Overall, DiffuSAGE offers clear insights into how and when tokens shape the generative trajectory, enhancing interpretability and opening avenues for more temporally controllable synthesis.

4.4.1 Key Findings

- **FIFA introduces a faithful and stable feature attribution method for black-box models.** FIFA uses swarm optimization to find a compact set of influential features and consistently outperforms LIME [60], SHAP [61], and KernelSHAP [61] in faithfulness, sparsity, reliability and stability. This makes it well-suited for real-world interpretability in predictive models.

- **DiffuSAGE offers token-level interpretability in generative diffusion models.** DiffuSAGE combines Aumann-Shapley values, attention alignment, and gradients to assign temporally and spatially grounded importance to prompt tokens. It reveals how different tokens shape early structure versus later refinement, offering deeper insight into prompt-to-image generation.

- **Token importance in diffusion is temporally structured.** Analysis across prompts shows that action and subject tokens (e.g., *"drinking," "man"*) shape early pose and layout, while object tokens (e.g., *"coffee," "umbrella"*) refine texture and detail in later steps. This coarse-to-fine pattern reflects the hierarchical reasoning

of diffusion models.

- **Prompt tokens map to coherent, semantically meaningful visual regions.** Token-to-region heatmaps show that DiffuSAGE disentangles prompt semantics into distinct spatial zones (e.g., *"astronaut"* mapped to the figure, *"mars"* to the background). This spatial grounding enhances user trust and supports controllable generation.

- **Timestep-level attribution reveals stage-specific semantic emergence.** By visualizing early-only, late-only, and full-step generations, DiffuSAGE highlights how early steps establish composition while later ones refine detail. This supports temporal control and deeper insight in generative design.

- **Quantitative metrics confirm attribution faithfulness.** DiffuSAGE achieves the highest Insertion AUC (0.8200) and lowest Deletion AUC (0.5500) compared to DF-RISE, DF-CAM, and LIME, confirming its ability to reliably identify the most influential tokens for model confidence and output fidelity.

4.4.2 Limitations

- During the DiffuSAGE experiments, token importance for abstract or metaphorical prompts (e.g., prompts involving symbolic language or personification) failed to produce coherent or interpretable attribution maps. The model often assigned high importance to less relevant tokens, indicating a lack of semantic grounding in cases where the textual prompt was not directly visualizable.

- Several attempts to apply DiffuSAGE on low-resolution or corrupted images produced unstable attributions, where overlapping or diluted heatmaps hindered token-to-region mapping, highlighting the method's sensitivity to image quality.

- The computational cost of both DiffuSAGE and FIFA limited scalability for real-time or large-scale use. DiffuSAGE was particularly inefficient with long prompts or models exceeding 100 denoising steps, and some high-resolution generations were excluded due to GPU memory constraints.

Chapter 5

Conclusion and recommendation

5.1 Conclusion

Explainability in machine learning has become an essential component of trustworthy and transparent AI systems, particularly as models increase in complexity and are deployed in high-stakes domains. Despite notable progress in feature attribution for predictive models and interpretability methods for generative models, critical gaps remain, especially in ensuring temporal, spatial, and semantic alignment in attribution for generative diffusion models, and in producing stable, sparse, and faithful attributions for black-box predictive models.

In this work, we introduced two novel methods that address these challenges in a systematic and principled way. First, we proposed the Firefly-Inspired Feature Attribution (FIFA) algorithm, a black-box feature importance technique based on population-based optimization. FIFA achieves high levels of faithfulness, sparsity, and stability, and outperforms baseline methods such as LIME, SHAP, and KernelSHAP across multiple metrics and model types. Second, we introduced DiffuSAGE, a gradient-based attribution method for generative diffusion models. DiffuSAGE uniquely captures both when (temporal) and where (spatial) a prompt token influences the output, allowing for fine-grained interpretability of text-to-image generation.

Through extensive experiments, we demonstrated the effectiveness of FIFA in producing reliable feature attributions on benchmark tabular datasets, while DiffuSAGE successfully revealed the evolving role of prompt tokens across diffusion steps and localized their impact to distinct visual regions. DiffuSAGE also achieved higher insertion AUC and lower deletion AUC compared to baselines like DF-RISE and DF-CAM, validating its faithfulness and precision.

These contributions collectively represent a step forward in developing robust, interpretable AI systems. They show that it is possible to generate transparent and meaningful explanations even in highly complex generative and predictive settings by leveraging optimization-based attribution and integrated gradient strategies.

Nevertheless, our work is not without limitations. DiffuSAGE exhibits sensitivity to prompt ambiguity and image quality, and FIFA can face challenges in highly correlated feature spaces. Moreover, both methods incur non-trivial computational costs, which can limit scalability for real-time or high-throughput scenarios.

Future work can address these limitations in several directions. For FIFA, improving convergence strategies and dimensionality reduction techniques may increase scalability and applicability to high-dimensional datasets. For DiffuSAGE, incorporating multimodal inputs, supporting more abstract prompts, and optimizing timestep attribution could further enhance interpretability. Additionally, integrating both methods into real-world human-in-the-loop systems may offer practical insights into usability and impact in applied domains such as healthcare, design, and policy decision-making.

5.2 Recommendation

- Attribution robustness: Based on observed limitations in DiffuSAGE's performance on abstract prompts and low-quality images, future research should focus on improving the robustness of token attribution. This can be addressed by augmenting training data with more diverse and semantically rich prompts, including non-literal language and figurative expressions. Additionally, incorporating auxiliary modules that validate attribution coherence, such as vision-language alignment models or caption-based grounding, could help refine the semantic precision of token-to-region mappings. These improvements would enhance the generalizability of DiffuSAGE across a wider range of prompts and image qualities.

- Timestep attribution scalability: Given the computational cost associated with full denoising trajectories, optimizing DiffuSAGE's timestep attribution strategy is recommended. Future work could explore adaptive sampling across timesteps, where

attribution is concentrated on key decision points identified through variance or entropy measures. This would reduce resource usage while preserving interpretability. Alternatively, approximating full trajectories using learned temporal priors or lightweight distillation models could enable real-time explanations for time-sensitive applications.

- Optimization efficiency for FIFA: To improve the convergence speed and scalability of FIFA, researchers are encouraged to experiment with more advanced metaheuristic optimization algorithms. Approaches such as adaptive swarm intelligence, hybrid genetic-firefly algorithms, or differentiable surrogate models could reduce computation time while maintaining attribution quality. These strategies would be particularly beneficial when applying FIFA to high-dimensional data or in production settings with strict runtime constraints.

- Multi-modal generalization: Both FIFA and DiffuSAGE are currently optimized for specific input modalities, tabular data and text-to-image diffusion, respectively. A valuable direction for future work is to generalize these methods to multi-modal settings. For example, extending DiffuSAGE to incorporate audio or motion-conditioned generation, or adapting FIFA to handle visual tabular representations, could significantly broaden the practical utility of these attribution techniques. Such extensions would require novel attention mechanisms or joint embedding strategies to align features across modalities.

- Integration with user feedback: To enhance trust and usability, future implementations of FIFA and DiffuSAGE could incorporate human-in-the-loop systems for feedback-driven attribution refinement. Allowing users to validate, adjust, or interactively explore attribution outputs can help uncover hidden model biases and improve model alignment with domain-specific expectations. This feedback loop can be particularly impactful in high-stakes environments such as healthcare, finance, or scientific modeling.

REFERENCES

[1] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Information Fusion*, vol. 77, pp. 29–52, 2022.

[2] D. Liang, R. G. Krishnan, M. D. Hoffman, and T. Jebara, "Variational autoencoders for collaborative filtering," in *Proceedings of the 2018 World Wide Web Conference*, pp. 689–698, 2018.

[3] J. Dumas, A. Wehenkel, D. Lanaspeze, B. Cornélusse, and A. Sutera, "A deep generative model for probabilistic energy forecasting in power systems: normalizing flows," *Applied Energy*, vol. 305, p. 117871, 2022.

[4] T. Teräsvirta, "Specification, estimation, and evaluation of smooth transition autoregressive models," *Journal of the American Statistical Association*, vol. 89, pp. 208–218, 1994.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, pp. 139–144, 2020.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[7] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T.-S. Chua, "Disease inference from health-related questions via sparse deep learning," *IEEE Transactions on knowledge and Data Engineering*, vol. 27, no. 8, pp. 2107–2119, 2015.

[8] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," *Expert Systems with Applications*, vol. 83, pp. 187–205, 2017.

[9] G. Goswami, R. Bhardwaj, R. Singh, and M. Vatsa, "Mdlface: Memorability augmented deep learning for video face recognition," in *IEEE international joint conference on biometrics*, pp. 1–7, IEEE, 2014.

[10] J. Lundén and V. Koivunen, "Deep learning for hrrp-based target recognition in multistatic radar systems," in *2016 IEEE Radar Conference (RadarConf)*, pp. 1–6, IEEE, 2016.

[11] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena, "Car that knows before you do: Anticipating maneuvers via learning temporal driving models," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3182–3190, 2015.

[12] X. Zhai, X. Chu, C. S. Chai, M. S. Y. Jong, A. Istenic, M. Spector, J.-B. Liu, J. Yuan, and Y. Li, "A review of artificial intelligence (ai) in education from 2010 to 2020," *Complexity*, vol. 2021, pp. 1–18, 2021.

[13] P. Georgiev, S. Bhattacharya, N. D. Lane, and C. Mascolo, "Low-resource multi-task audio sensing for mobile and embedded devices via shared deep neural network representations," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–19, 2017.

[14] I. M. Enholm, E. Papagiannidis, P. Mikalef, and J. Krogstie, "Artificial intelligence and business value: A literature review," *Information Systems Frontiers*, vol. 24, no. 5, pp. 1709–1734, 2022.

[15] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[16] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[17] M. Van Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in *Proceedings of the national conference on artificial intelligence*, pp. 900–907, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

[18] W. R. Swartout, "Xplain: A system for creating and explaining expert consulting programs," *Artificial intelligence*, vol. 21, no. 3, pp. 285–325, 1983.

[19] J. D. Moore and W. R. Swartout, *Explanation in expert systems: A survey.* University of Southern California, Information Sciences Institute Marina del . . . , 1988.

[20] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge-based systems*, vol. 8, no. 6, pp. 373–389, 1995.

[21] N. Tintarev and J. Masthoff, "A survey of explanations in recommender systems," in *2007 IEEE 23rd international conference on data engineering workshop*, pp. 801–810, IEEE, 2007.

[22] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pp. 241–250, 2000.

[23] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[24] G. Sartor, F. Lagioia, *et al.*, "The impact of the general data protection regulation (gdpr) on artificial intelligence." European Parliament, 2020.

[25] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pp. 1078–1088, International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[26] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[27] K. Crawford, "Artificial intelligence's white guy problem," *The New York Times*, vol. 25, no. 06, p. 5, 2016.

[28] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[29] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," in *Ethics of data and analytics*, pp. 254–264, Auerbach Publications, 2022.

[30] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[31] P. Grother, M. Ngan, and K. Hanaoka, *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019.

[32] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, 2017.

[33] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730, 2015.

[34] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "Ai for radiographic covid-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, vol. 3(7), pp. 610–619, 2021.

[35] S. Ballingall, M. Sarvi, and P. Sweatman, "Safety assurance for automated systems in transport: A collective case study of real-world fatal crashes," *Journal of Safety Research*, vol. 92, pp. 27–39, 2025.

[36] K. Hill, "Wrongfully accused by an algorithm," in *Ethics of Data and Analytics*, pp. 138–142, Auerbach Publications, 2022.

[37] M. H. LeRoy, "Algorithmic bias in hiring: Amending title vii to prohibit ai discrimination," *J. Legis.*, vol. 51, p. 261, 2025.

[38] J. Starr and C. Quick, *Robotic Safety Systems: An Applied Approach*. CRC Press, 2024.

[39] C. J. Hoofnagle, B. Van Der Sloot, and F. Z. Borgesius, "The european union general data protection regulation: what it is and what it means," *Information & Communications Technology Law*, vol. 28, no. 1, pp. 65–98, 2019.

[40] European Commission, "Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," Apr. 2021. COM(2021) 206 final.

[41] White House Office of Science and Technology Policy, "Blueprint for an ai bill of rights: Making automated systems work for the american people," October 2022.

[42] A. Oussidi and A. Elhassouny, "Deep generative models: Survey," in *2018 International conference on intelligent systems and computer vision (ISCV)*, pp. 1–8, 2018.

[43] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–39, 2023.

[44] A. S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite, "Stable bias: Analyzing societal representations in diffusion models," *arXiv preprint arXiv:2303.11408*, 2023.

[45] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, and E. Wallace, "Extracting training data from diffusion models," in *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

[46] A. Birhane, V. U. Prabhu, and E. Kahembwe, "Multimodal datasets: misogyny, pornography, and malignant stereotypes," *arXiv preprint arXiv:2110.01963*, 2021.

[47] Z. Sadeghi, R. Alizadehsani, M. A. Cifci, S. Kausar, R. Rehman, P. Mahanta, P. K. Bora, A. Almasri, R. S. Alkhawaldeh, S. Hussain, *et al.*, "A review of explainable artificial intelligence in healthcare," *Computers and Electrical Engineering*, vol. 118, p. 109370, 2024.

[48] Y. Wang, T. Zhang, X. Guo, and Z. Shen, "Gradient based feature attribution in explainable ai: A technical review," *arXiv preprint arXiv:2403.10415*, 2024.

[49] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne, "Learning how to explain neural networks: Patternnet and patternattribution," *arXiv preprint arXiv:1705.05598*, 2017.

[50] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, and S. Lapuschkin, "From attribution maps to human-understandable explanations through concept relevance propagation," *Nature Machine Intelligence*, vol. 5, pp. 1006–1019, 2023.

[51] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*, pp. 3319–3328, PMLR, 2017.

[52] G. Visani, E. Bagli, and F. Chesani, "Optilime: Optimized lime explanations for diagnostic computer algorithms," *arXiv preprint arXiv:2006.05714*, 2020.

[53] T. A. A. Abdullah, M. S. M. Zahid, A. F. Turki, W. Ali, A. A. Jiman, M. J. Abdulaal, N. M. Sobahi, and E. T. Attar, "Sig-lime: a signal-based enhancement of lime explanation technique," *IEEE access*, vol. 12, pp. 52641–52658, 2024.

[54] J. Chen, L. Song, M. Wainwright, and M. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," in *International conference on machine learning*, pp. 883–892, PMLR, 2018.

[55] Y. Sun, Z. Chen, V. Orlandi, T. Wang, and C. Rudin, "Sparse and faithful explanations without sparse models," *arXiv preprint arXiv:2402.09702*, 2024.

[56] S. M. Shankaranarayana and D. Runje, "Alime: Autoencoder based approach for local interpretability," in *Intelligent Data Engineering and Automated Learning–IDEAL 2019: 20th International Conference, Manchester, UK, November 14–16, 2019, Proceedings, Part I 20*, pp. 454–463, Springer, 2019.

[57] Z. Tan, Y. Tian, and J. Li, "Glime: General, stable and local lime explanation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 36250–36277, 2023.

[58] G. Kelodjou, L. Rozé, V. Masson, L. Galárraga, R. Gaudel, M. Tchuente, and A. Termier, "Shaping up shap: Enhancing stability through layer-wise neighbor selection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38(12), pp. 13094–13103, 2024.

[59] M. R. Zafar and N. M. Khan, "Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," *arXiv preprint arXiv:1906.10263*, 2019.

[60] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[61] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, vol. 30, 2017.

[62] C. Molnar, *Interpreting Machine Learning Models with SAP: A Guide with Python Examples and Theory on Shapley Values*. Chistoph Molnar c/o MUCBOOK, Heidi Seibold, 2023.

[63] I. D. Mienye, G. Obaido, N. Jere, E. Mienye, K. Aruleba, I. D. Emmanuel, and B. Ogbuokiri, "A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges," *Informatics in Medicine Unlocked*, p. 101587, 2024.

[64] P. Li, S. S. Rangapuram, and M. Slawski, "Methods for sparse and low-rank recovery under simplex constraints," *Statistica Sinica*, vol. 30, no. 2, pp. 557–577, 2020.

[65] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 607–617, 2020.

[66] J.-H. Park, Y.-J. Ju, and S.-W. Lee, "Explaining generative diffusion models via visual analysis for interpretable decision-making process," *Expert Systems with Applications*, vol. 248, p. 123231, 2024.

[67] H. Chefer, N. Sclar, and L. Wolf, "What the daam: Interpreting stable diffusion using cross attention," *arXiv preprint arXiv:2304.08641*, 2023.

[68] Z. Chen, Y. Sun, T. Wang, and C. Rudin, "Concept attention makes diffusion models better zero-shot segmenters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[69] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[70] R. Tiwari, "Explainable ai (xai) and its applications in building trust and understanding in ai decision making," *International J. Sci. Res. Eng. Manag*, vol. 7, pp. 1–13, 2023.

[71] E. Sepulveda, F. Vandervorst, B. Baesens, and T. Verdonck, "Enhancing explainability in real-world scenarios: Towards a robust stability measure for local interpretability," *Expert Systems with Applications*, vol. 274, p. 126922, 2025.

[72] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harvard Journal of Law & Technology*, vol. 31, p. 841, 2017.

[73] J. Adebayo, J. Gilmer, I. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[74] M. Veale and F. Zuiderveen Borgesius, "Demystifying the draft eu artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach," *Computer Law Review International*, vol. 22, no. 4, pp. 97–112, 2021.

[75] S. Cohen, "The evolution of machine learning: Past, present, and future," in *Artificial Intelligence in Pathology*, pp. 3–14, Elsevier, 2025.

[76] N. Rane, S. Choudhary, and J. Rane, "Machine learning and deep learning: A comprehensive review on methods, techniques, applications, challenges, and future directions," *Techniques, Applications, Challenges, and Future Directions (May 31, 2024)*, 2024.

[77] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021.

[78] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

[79] T. Karras, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[80] M. Durgadevi *et al.*, "Generative adversarial network (gan): A general review on different variants of gan and applications," in *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, pp. 1–8, 2021.

[81] S. Joo, M. S. Kim, J. Yang, and J. Park, "Generative model for proposing drug candidates satisfying anticancer properties using a conditional variational autoencoder," *ACS Omega*, vol. 5, pp. 18642–18650, 2020.

[82] M. Smith, S. Nichols, R. Henkelman, and M. Wood, "Application of autoregressive moving average parametric modeling in magnetic resonance image reconstruction," *IEEE Transactions on Medical Imaging*, vol. 5, pp. 132–139, 1986.

[83] J. Schneider, "Explainable generative ai (genxai): A survey, conceptualization, and research agenda," *Artificial Intelligence Review*, vol. 57, p. 289, 2024.

[84] L. Manduchi, K. Pandey, R. Bamler, R. Cotterell, S. Däubener, S. Fellenz, A. Fischer, T. Gärtner, M. Kirchler, M. Kloft, *et al.*, "On the challenges and opportunities in generative ai," *arXiv preprint arXiv:2403.00025*, 2024.

[85] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[86] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2nd ed., 2009.

[87] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*. Wiley, 5th ed., 2012.

[88] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[89] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Advances in Neural Information Processing Systems*, vol. 9, pp. 155–161, 1997.

[90] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[91] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, ACM, 2016.

[92] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[93] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.

[94] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[95] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31, NeurIPS, 2018.

[96] S. O. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35(8), pp. 6679–6687, AAAI, 2021.

[97] G. G. Chrysos, J. Kossaifi, and S. Zafeiriou, "Robust conditional generative adversarial networks," *arXiv preprint arXiv:1805.08657*, 2018.

[98] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[99] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, pp. 214–223, 2017.

[100] R. Wei and A. Mahmood, "Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey," *IEEE Access*, vol. 9, pp. 4939–4956, 2020.

[101] R. Huang, M. W. Y. Lam, J. Wang, D. Su, D. Yu, Y. Ren, and Z. Zhao, "Fastdiff: A fast conditional diffusion model for high-quality speech synthesis," *arXiv preprint arXiv:2204.09934*, 2022.

[102] A. Vahdat and J. Kautz, "Nvae: A deep hierarchical variational autoencoder," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19667–19679, 2020.

[103] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato, "Grammar variational autoencoder," in *International Conference on Machine Learning*, pp. 1945–1954, 2017.

[104] G. Bredell, K. Flouris, K. Chaitanya, E. Erdil, and E. Konukoglu, "Explicitly minimizing the blur error of variational autoencoders," *arXiv preprint arXiv:2304.05939*, 2023.

[105] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *Journal of Machine Learning Research*, vol. 22, pp. 1–64, 2021.

[106] X.-R. Gong, J.-X. Jin, and T. Zhang, "Sentiment analysis using autoregressive language modeling and broad learning system," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1130–1134, 2019.

[107] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 587–597, 2012.

[108] J. Xiong, G. Liu, L. Huang, C. Wu, T. Wu, Y. Mu, Y. Yao, H. Shen, Z. Wan, J. Huang, *et al.*, "Autoregressive models in vision: A survey," *arXiv preprint arXiv:2411.05902*, 2024.

[109] E. Coviello, Y. Vaizman, A. B. Chan, and G. R. Lanckriet, "Multivariate autoregressive mixture models for music auto-tagging," in *ISMIR*, pp. 547–552, 2012.

[110] A. Mikusheva, "Uniform inference in autoregressive models," *Econometrica*, vol. 75, pp. 1411–1452, 2007.

[111] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017.

[112] X. Wang, L. Sun, A. Chehri, and Y. Song, "A review of gan-based super-resolution reconstruction for optical remote sensing images," *Remote Sensing*, vol. 15, no. 20, p. 5062, 2023.

[113] H. Huang, P. S. Yu, and C. Wang, "An introduction to image synthesis with generative adversarial nets," *arXiv preprint arXiv:1803.04469*, 2018.

[114] I. Vaccari, V. Orani, A. Paglialonga, E. Cambiaso, and M. Mongelli, "A generative adversarial network (gan) technique for internet of medical things data," *Sensors*, vol. 21, no. 11, p. 3726, 2021.

[115] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[116] J. Park *et al.*, "X-diffusion: Explainable diffusion models via df-rise and df-cam," *arXiv preprint arXiv:2303.12345*, 2023.

[117] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*, pp. 2256–2265, 2015.

[118] A. B. Yenew, B. G. Assefa, and E. G. Belay, "Housegandi: A hybrid approach to strike a balance of sampling time and diversity in floorplan generation," *IEEE Access*, 2024.

[119] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

[120] A. Rawal, J. McCoy, D. B. Rawat, B. M. Sadler, and R. S. Amant, "Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 10, pp. 852–866, 2021.

[121] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. V. Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," *ACM Computing Surveys*, vol. 55, no. 10, pp. 1–42, 2023.

[122] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

[123] C. Vairetti, S. Maldonado, L. Cuitino, and C. A. Urzua, "Interpretable multimodal classification for age-related macular degeneration diagnosis," *PloS One*, vol. 19, p. e0311811, 2024.

[124] G. del Castillo Torres, M. F. Roig-Maimó, M. Mascaró-Oliver, E. Amengual-Alcover, and R. Mas-Sansó, "Understanding how cnns recognize facial expressions: a case study with lime and cem," *Sensors*, vol. 23, p. 131, 2022.

[125] L. K. Gupta, D. Koundal, and S. Mongia, "Explainable methods for image-based deep learning: a review," *Archives of Computational Methods in Engineering*, vol. 30, pp. 2651–2666, 2023.

[126] G. Attanasio, E. Pastor, C. D. Bonaventura, and D. Nozza, "ferret: a framework for benchmarking explainers on transformers," *arXiv preprint arXiv:2208.01575*, 2022.

[127] M. M. Hasan, "Understanding model predictions: A comparative analysis of shap and lime on various ml algorithms," *Journal of Scientific and Technological Research*, vol. 5, pp. 17–26, 2023.

[128] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[129] Q. Huang, M. Yamada, Y. Tian, D. Singh, and Y. Chang, "Graphlime: Local interpretable model explanations for graph neural networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, pp. 6968–6972, 2022.

[130] N. Liu, Q. Feng, and X. Hu, "Interpretability in graph neural networks," in *Graph Neural Networks: Foundations, Frontiers, and Applications* (L. Wu and C. C. Aggarwal, eds.), pp. 121–147, Cham: Springer, 2022.

[131] T. Zhao, D. Luo, X. Zhang, and S. Wang, "On consistency in graph neural network interpretation," *arXiv preprint arXiv:2205.13733*, vol. 9, 2022.

[132] J. I. Janjua, R. Ahmad, S. Abbas, A. S. Mohammed, M. S. Khan, A. Daud, T. Abbas, and M. A. Khan, "Enhancing smart grid electricity prediction with the fusion of intelligent modeling and xai integration," *International Journal of Advanced and Applied Sciences*, vol. 11, pp. 230–248, 2024.

[133] A. Theissler, F. Spinnato, U. Schlegel, and R. Guidotti, "Explainable ai for time series classification: a review, taxonomy and research directions," *IEEE Access*, vol. 10, pp. 100700–100724, 2022.

[134] G. Vilone and L. Longo, "Explainable artificial intelligence: a systematic review," *arXiv preprint arXiv:2006.00093*, 2020.

[135] M. H. Wang, K. K.-l. Chong, Z. Lin, X. Yu, and Y. Pan, "An explainable artificial intelligence-based robustness optimization approach for age-related macular degeneration detection based on medical iot systems," *Electronics*, vol. 12, p. 2697, 2023.

[136] B. M. de Vries, G. J. Zwezerijnen, G. L. Burchell, F. H. van Velden, C. W. Menke-van der Houven van Oordt, and R. Boellaard, "Explainable artificial intelligence (xai) in radiology and nuclear medicine: a literature review," *Frontiers in medicine*, vol. 10, p. 1180773, 2023.

[137] M. Belghachi, "A review on explainable artificial intelligence methods, applications, and challenges," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 11, no. 4, pp. 1007–1024, 2023.

[138] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localiza-

tion," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

[139] P. Angelov, D. Kangin, and Z. Zhang, "Towards interpretable-by-design deep learning algorithms," *arXiv preprint arXiv:2311.11396*, 2023.

[140] T. Vermeire, D. Brughmans, S. Goethals, R. M. B. De Oliveira, and D. Martens, "Explainable image classification with evidence counterfactual," *Pattern Analysis and Applications*, vol. 25, pp. 315–335, 2022.

[141] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, pp. 31–57, 2018.

[142] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[143] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, MA: MIT Press, 2023.

[144] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *Journal of Computational and Graphical Statistics*, vol. 24, pp. 44–65, 2015.

[145] T. Spooner, D. Dervovic, J. Long, J. Shepard, J. Chen, and D. Magazzeni, "Counterfactual explanations for arbitrary regression models," *arXiv preprint arXiv:2106.15212*, 2021.

[146] Y. Ji, Y. Sun, Y. Zhang, Z. Wang, Y. Zhuang, Z. Gong, D. Shen, C. Qin, H. Zhu, and H. Xiong, "A comprehensive survey on self-interpretable neural networks," *arXiv preprint arXiv:2501.15638*, 2025.

[147] S. Letzgus, P. Wagner, J. Lederer, W. Samek, K.-R. Müller, and G. Montavon, "Toward explainable artificial intelligence for regression models: A methodological perspective," *IEEE Signal Processing Magazine*, vol. 39, pp. 40–58, 2022.

[148] R. L. Bach, C. Kern, H. Mautner, and F. Kreuter, "The impact of modeling decisions in statistical profiling," *Data & Policy*, vol. 5, p. e32, 2023.

[149] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. Díaz-Rodríguez, "Explainable artificial intelligence (xai) on time-series data: A survey," *arXiv preprint arXiv:2104.00950*, 2021.

[150] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim, "Towards a rigorous evaluation of xai methods on time series," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 4197–4201, 2019.

[151] P.-D. Arsenault, S. Wang, and J.-M. Patenande, "A survey of explainable artificial intelligence (xai) in financial time series forecasting," *arXiv preprint arXiv:2407.15909*, 2024.

[152] S. Bobek, M. Kuk, M. Szelążek, and G. J. Nalepa, "Enhancing cluster analysis with explainable ai and multidimensional cluster prototypes," *IEEE Access*, vol. 10, pp. 101556–101574, 2022.

[153] J. Cohen, X. Huan, and J. Ni, "Shapley-based explainable ai for clustering applications in fault diagnosis and prognosis," *Journal of Intelligent Manufacturing*, pp. 1–16, 2024.

[154] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, *et al.*, "Explainable ai (xai): Core ideas, techniques, and solutions," *ACM Computing Surveys*, vol. 55, pp. 1–33, 2023.

[155] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, *et al.*, "Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions," *Information Fusion*, vol. 106, p. 102301, 2024.

[156] F. Mokhtar, R. Kansal, D. Diaz, J. Duarte, J. Pata, M. Pierini, and J.-R. Vlimant, "Explaining machine-learned particle-flow reconstruction," *arXiv preprint arXiv:2111.12840*, 2021.

[157] K. Werder, B. Ramesh, and R. Zhang, "Establishing data provenance for responsible artificial intelligence systems," *ACM Transactions on Management Information Systems (TMIS)*, vol. 13, pp. 1–23, 2022.

[158] D. Pan, N. Moniz, and N. Chawla, "Fast explainability via feasible concept sets generator," *arXiv preprint arXiv:2405.18664*, 2024.

[159] R. Doddaiah, P. Parvatharaju, E. Rundensteiner, and T. Hartvigsen, "Class-specific explainability for deep time series classifiers," in *2022 IEEE International conference on data mining (ICDM)*, pp. 101–110, 2022.

[160] J. E. Zini and M. Awad, "On the explainability of natural language processing deep models," *ACM Computing Surveys*, vol. 55, pp. 1–31, 2022.

[161] C. K. Wikle, A. Datta, B. V. Hari, E. L. Boone, I. Sahoo, I. Kavila, S. Castruccio, S. J. Simmons, W. S. Burr, and W. Chang, "An illustration of model agnostic explainability methods applied to environmental data," *Environmetrics*, vol. 34, p. e2772, 2023.

[162] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[163] W. K. Diprose, N. Buist, N. Hua, Q. Thurier, G. Shand, and R. Robinson, "Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator," *Journal of the American Medical Informatics Association*, vol. 27, pp. 592–600, 2020.

[164] C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, and B. Bischl, "General pitfalls of model-agnostic interpretation methods for machine learning models," in *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pp. 39–68, 2020.

[165] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications," *Information Fusion*, vol. 81, pp. 59–83, 2022.

[166] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[167] J. Vítků, P. Dluhoš, J. Davidson, M. Nikl, S. Andersson, P. Paška, J. Šinkora, P. Hlubuček, M. Stránský, M. Hyben, *et al.*, "Toyarchitecture: Unsupervised learning of interpretable models of the environment," *PloS One*, vol. 15, p. e0230432, 2020.

[168] X. Chen, Y. Zeng, S. Kang, and R. Jin, "Inn: An interpretable neural network for ai incubation in manufacturing," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, pp. 1–23, 2022.

[169] Z. Carmichael, T. Moon, and S. A. Jacobs, "Learning interpretable models through multi-objective neural architecture search," *arXiv preprint arXiv:2112.08645*, 2021.

[170] P. Pylov, A. Dyagileva, A. Protodyakonov, and R. Maitak, "Heuristics of constructing the architecture of an interpreted machine learning model," in *E3S Web of Conferences*, vol. 531, p. 03008, 2024.

[171] V. Lai, Y. Zhang, C. Chen, Q. V. Liao, and C. Tan, "Selective explanations: Leveraging human input to align explainable ai," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, pp. 1–35, 2023.

[172] C. Szegedy, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[173] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[174] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014*, pp. 818–833, Springer, 2014.

[175] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv preprint arXiv:1702.04595*, 2017.

[176] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3429–3437, 2017.

[177] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[178] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, p. 1096, 2019.

[179] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 2017.

[180] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Comparing the interpretability of deep networks via network dissection," in *Explainable AI: Interpreting, explaining and visualizing deep learning*, pp. 243–252, Springer, 2019.

[181] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[182] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," in *Explainable AI: Interpreting, explaining and visualizing deep learning*, pp. 193–209, Springer, 2019.

[183] L. Arras, J. Arjona-Medina, M. Widrich, G. Montavon, M. Gillhofer, K.-R. Müller, S. Hochreiter, and W. Samek, "Explaining and interpreting lstms," in *Explainable AI: Interpreting, explaining and visualizing deep learning*, pp. 211–238, Springer, 2019.

[184] J. Kauffmann, M. Esders, L. Ruff, G. Montavon, W. Samek, and K.-R. Müller, "From clustering to cluster explanations via neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, pp. 1926–1940, 2022.

[185] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[186] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[187] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans, "innvestigate neural networks!," *Journal of Machine Learning Research*, vol. 20, pp. 1–8, 2019.

[188] J. Aechtner, L. Cabrera, D. Katwal, P. Onghena, D. P. Valenzuela, and A. Wilbik, "Comparing user perception of explanations developed with xai methods," in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–7, 2022.

[189] T. Nakanishi, "Pcaime: Principal component analysis-enhanced approximate inverse model explanations through dimensional decomposition and expansion," *IEEE Access*, 2024.

[190] G. Laberge, Y. B. Pequignot, M. Marchand, and F. Khomh, "Tackling the xai disagreement problem with regional explanations," in *International Conference on Artificial Intelligence and Statistics*, pp. 2017–2025, 2024.

[191] C. T. Okolo, N. Dell, and A. Vashistha, "Making ai explainable in the global south: A systematic review," in *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, pp. 439–452, 2022.

[192] P. Cortez and M. J. Embrechts, "Using sensitivity analysis and visualization techniques to open black box data mining models," *Information Sciences*, vol. 225, pp. 1–17, 2013.

[193] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[194] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan, "L-shapley and c-shapley: Efficient model interpretation for structured data," *arXiv preprint arXiv:1808.02610*, 2018.

[195] A. Ghorbani and J. Zou, "Neuron shapley: Discovering the responsible neurons," in *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS'20)*, vol. 33, pp. 5922–5932, 2020.

[196] A. Ghorbani and J. Zou, "Data shapley: Equitable valuation of data for machine learning," in *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pp. 2242–2251, PMLR, 2019.

[197] P. Hall, N. Gill, M. Kurka, and W. Phan, "Machine learning interpretability with h2o driverless ai," *H2O. ai*, 2017.

[198] L. Hu, J. Chen, V. N. Nair, and A. Sudjianto, "Locally interpretable models and effects based on supervised partitioning (lime-sup)," *arXiv preprint arXiv:1806.00663*, 2018.

[199] I. Ahern, A. Noack, L. Guzman-Nateras, D. Dou, B. Li, and J. Huan, "Normlime: A new feature importance metric for explaining deep neural networks," *arXiv preprint arXiv:1909.04200*, 2019.

[200] A. Shih, A. Choi, and A. Darwiche, "A symbolic approach to explaining bayesian network classifiers," *arXiv preprint arXiv:1805.03364*, 2018.

[201] E. Albini, A. Rago, P. Baroni, F. Toni, *et al.*, "Relation-based counterfactual explanations for bayesian network classifiers.," in *IJCAI*, pp. 451–457, 2020.

[202] A. Ignatiev, N. Narodytska, and J. Marques-Silva, "Abduction-based explanations for machine learning models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33(1), pp. 1511–1519, 2019.

[203] F. Doshi-Velez, B. C. Wallace, and R. P. Adams, "Graph-sparse lda: A topic model with structured sparsity," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2575–2581, AAAI Press, 2015.

[204] J. Bien and R. Tibshirani, "Prototype selection for interpretable classification," *The Annals of Applied Statistics*, vol. 5(4), pp. 2403–2424, 2011.

[205] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[206] Z. T. Fernando, J. Singh, and A. Anand, "A study on the interpretability of neural retrieval models using deepshap," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*, (New York, NY, USA), pp. 1005–1008, Association for Computing Machinery, 2019.

[207] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 3319–3328, 2017.

[208] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2014.

[209] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, "Learning deep features for discriminative localization," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016.

[210] S. M. Lundberg, G. G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 252–259, 2020.

[211] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *International conference on machine learning*, pp. 3145–3153, 2017.

[212] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. Müller, W. Samek, and T. Schultz, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, p. e0130140, 2015.

[213] Z. Chen, Y. Sun, T. Wang, and C. Rudin, "Concept attention makes diffusion models better zero-shot segmenters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[214] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (un) reliability of saliency methods," *arXiv preprint arXiv:1711.00867*, 2017.

[215] X.-S. Yang, "Firefly algorithms for multimodal optimization," in *International symposium on stochastic algorithms*, pp. 169–178, Springer, 2009.

[216] R. J. Aumann and L. S. Shapley, *Values of non-atomic games*. Princeton University Press, 2015.

[217] K. Peffers, T. Tuunanen, C. E. Gengler, M. Rossi, W. Hui, V. Virtanen, and J. Bragge, "Design science research process: A model for producing and presenting information systems research," *arXiv preprint arXiv:2006.02763*, 2020.

[218] W. Wolberg, O. Mangasarian, N. Street, and W. Street, "Breast cancer wisconsin (diagnostic)." UCI Machine Learning Repository, 1993.

[219] Mathchi, "Diabetes data set," 2022.

[220] B. Becker and R. Kohavi, "Adult." UCI Machine Learning Repository, 1996.

[221] P. Chowdhury, M. Prabhushankar, and G. AlRegib, "Explaining explainers: Necessity and sufficiency in tabular data," in *NeurIPS 2023 Second Table Representation Learning Workshop*, 2023.

[222] A. Ghasemi, S. Hashtarkhani, D. L. Schwartz, and A. Shaban-Nejad, "Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review," *Cancer Innovation*, vol. 3, no. 5, p. e136, 2024.

[223] S. Ahmed, M. S. Kaiser, M. S. Hossain, and K. Andersson, "A comparative analysis of lime and shap interpreters with explainable ml-based diabetes predictions," *IEEE Access*, 2024.

[224] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[225] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

[226] C. Schuhmann, R. Beaumont, R. Vencu, and et al., "Laion-5b: An open large-scale dataset for training next generation multi-modal models," in *NeurIPS Datasets and Benchmarks Track*, 2022.

[227] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proceedings of the National Academy of Sciences*, vol. 87, no. 23, pp. 9193–9196, 1990.

APPENDICES

A Appendix A: Survey Questionnaire

This appendix contains the full questionnaire and the visual examples shown to participants during the qualitative evaluation of token-level and timestep-based attributions produced by the proposed DiffuSAGE method.

Survey Questionnaire

1. **Your Email**
2. **May we kindly have your name for record purposes?**
3. **What is your highest level of education completed?**
    - Some college or university
    - Bachelor's degree
    - Master's degree
    - PhD or equivalent
    - Other
4. **What is your field of study?**
5. **How familiar are you with Generative AI tools (e.g., DALL · E, Midjourney, Stable Diffusion)?**
    - Not familiar at all
    - Heard of them but never used

- Somewhat familiar, I've experimented a few times
- Very familiar, I use them regularly

6. **How would you rate your understanding of how AI models generate or interpret images?**
   - Not familiar at all
   - Somewhat familiar, I've learned or experimented a few times
   - Familiar, I've used them in courses or projects
   - Expert, I research or work in this field
7. **Have you used or studied any explainability (XAI) methods (e.g., SHAP, LIME, saliency maps)?**
   - Not familiar at all
   - Heard of them but never used
   - Familiar, I've used them in courses or projects
   - Expert, I research or work in this field
8. **Part I: Take a look at this image, generated with the prompt: *"An astronaut riding a horse on Mars."* Which word do you think played the biggest role in creating what you see?**

Figure A.1: Generated image from the prompt: “an astronaut riding a horse on Mars”

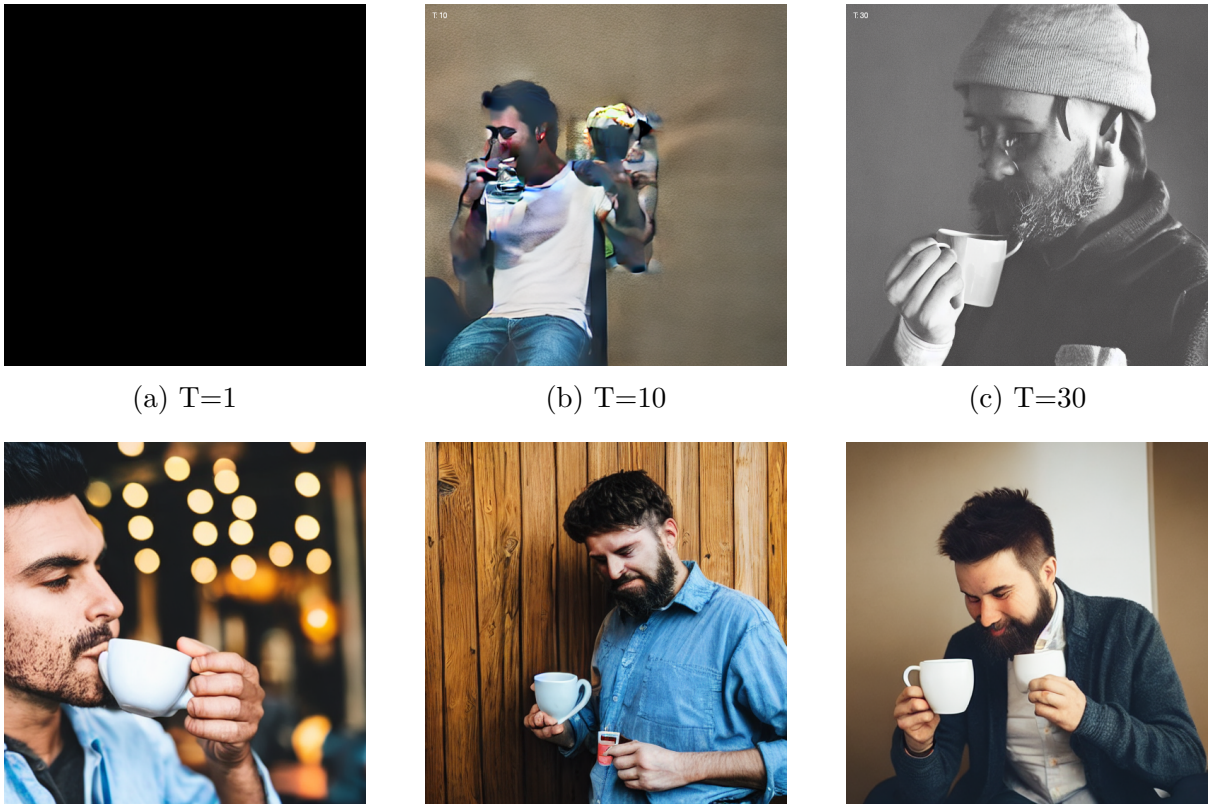
- An
- astronaut
- riding
- horse
- on
- mars
- a

9. **Below is an image generated using Stable Diffusion with the prompt: *“A man drinking coffee.”* Which of the following words do you think was most important in shaping the image content?**

Figure A.2: Generated image from the prompt: “A man drinking coffee.”

- A
- man
- drinking
- coffee

10. **Below are a few images showing the evolution of the generated image over time (Timestep 1, 10, 30, 50, 70, 90). Which stage do you think had the most influence on forming the final image?**



(a) T=1 (b) T=10 (c) T=30

(d) T=50 (e) T=70 (f) T=90

Figure A.3: Denoising trajectory for the prompt “a man drinking coffee” across selected diffusion steps.

11. **From Timesteps of question 10 above, at which stage of the diffusion process do you feel the core structure of the image was most strongly established?**

    - Early (Timestep=10)
    - Middle (Timestep 30–50)
    - Late (Timestep 70–90)

12. **From Timesteps of question 10 above, at which stage of the diffusion process do you feel the detailed textures (e.g., shading, surface features, fine visual elements) became most prominent in the image?**

    - Early (Timestep=10)
    - Middle (Timestep 30–50)

- Late (Timestep 70–90)

13. **Part II: Below are three heatmaps generated from the prompt: "An astronaut riding a horse on Mars." Each heatmap corresponds to one of the following tokens: astronaut, horse, or Mars. Match each token to the figure you believe represents it best. Tip for reading the heatmaps: Brighter (yellow) areas show a stronger influence of the token. Darker (purple) areas indicate lower influence.**

Figure A.4: Generated image from the prompt: "an astronaut riding a horse on Mars"

14. **Which token best describes Figure A.5?**

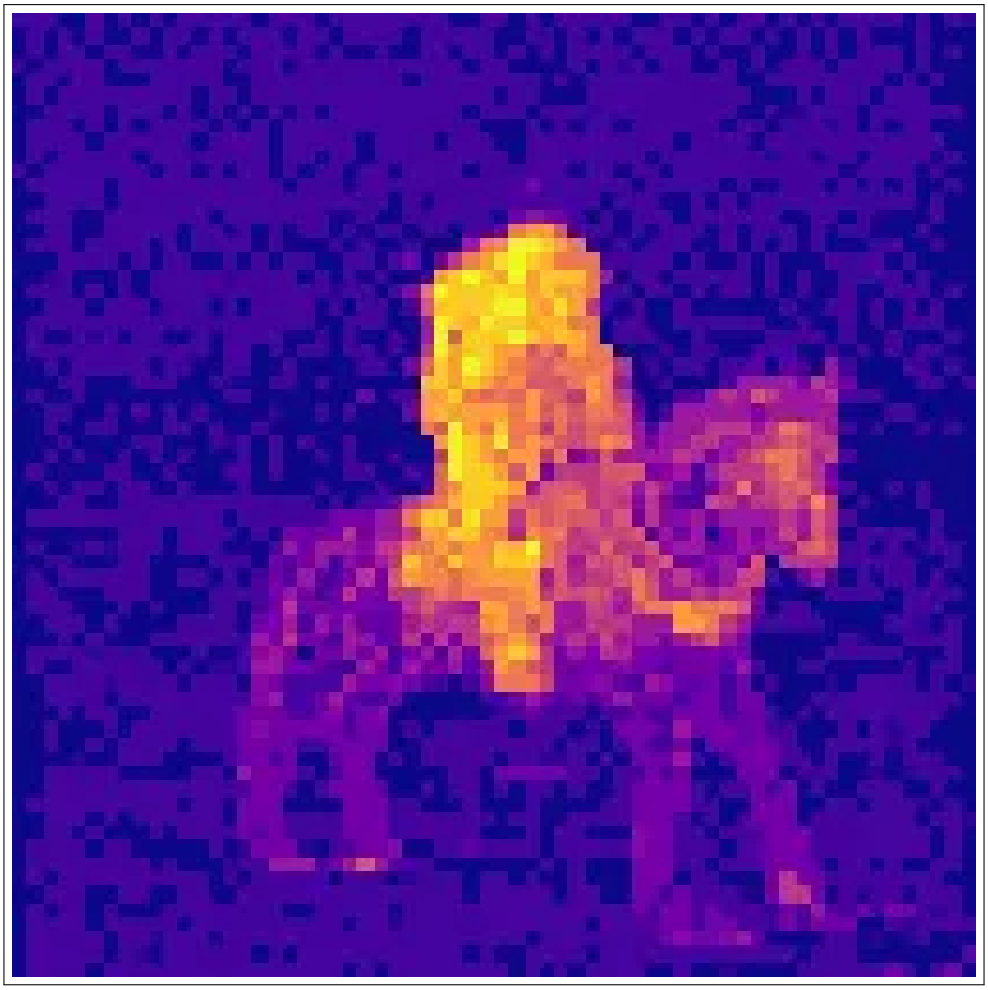


Figure A.5: Heatmap 1

- Astronaut
- Horse
- Mars

15. **Which token best describes Figure A.6?**

Figure A.6: Heatmap 2

- Astronaut
- Horse
- Mars

16. **Which token best describes Figure A.7**

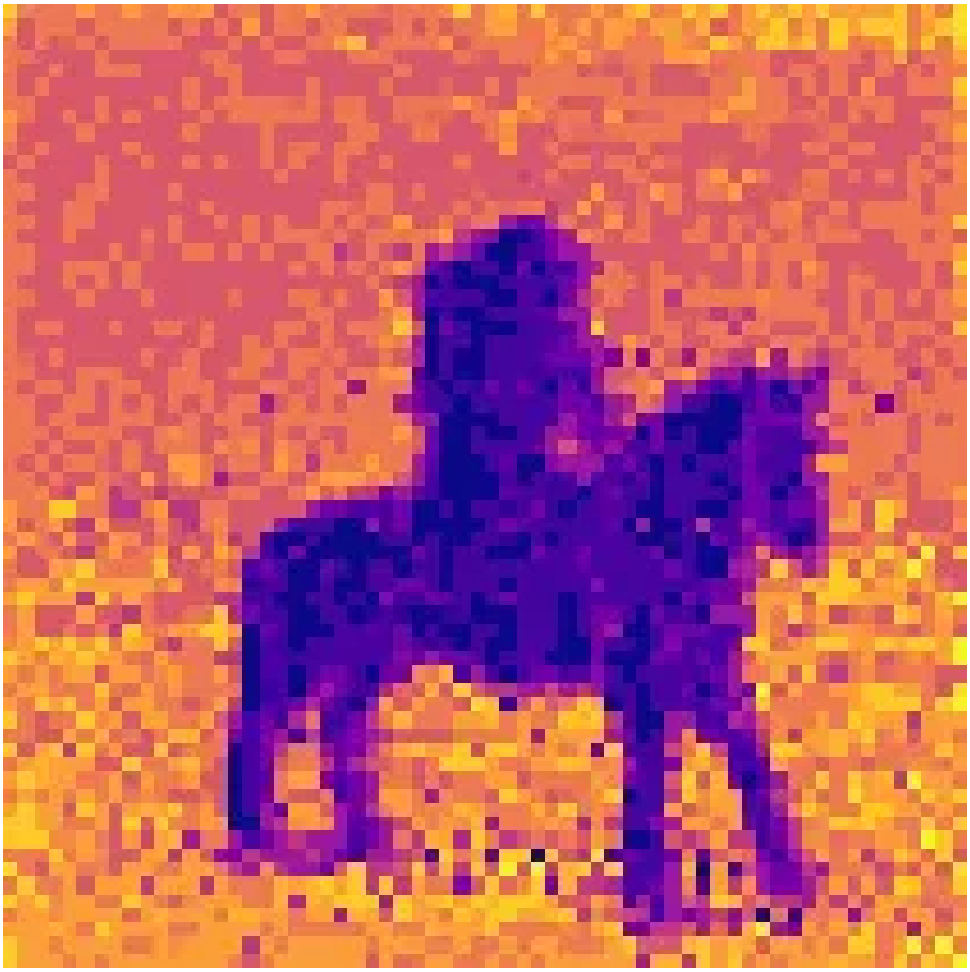


Figure A.7: Heatmap 3

- Astronaut
- Horse
- Mars