Diffusion Integrated Gradients: Controllable Path Generation for Flexible Feature Attribution

Soyeon Kim[1,3], Kyowoon Lee[2]†, and Jaesik Choi[1,3]†

[1]KAIST, Korea [2]Ajou University, Korea [3]INEEJI Corp., Korea
kyowoon.lee1924@gmail.com, {soyeon.k,jaesik.choi}@kaist.ac.kr
†Co-corresponding authors.

Abstract. Path-based attribution methods such as Integrated Gradients (IG) are widely adopted for their strong axiomatic properties and effectiveness in attributing model predictions to input features by integrating gradients along a path from a baseline to the input. However, the choice of the attribution path largely affects the quality of explanations, and existing approaches rely on fixed or hand-crafted paths that often produce noisy or distorted attributions. To address this limitation, we propose *Diffusion Integrated Gradients* (DiffIG), a novel method that reformulates path generation as a conditional generative modeling problem. DiffIG first trains a diffusion model to learn a distribution over paths generated from a Stick-Breaking Process, then employs guided sampling to embed user guidance during the sampling procedure. We demonstrate that DiffIG quantitatively matches or outperforms existing path-based methods, achieving perceptually aligned explanations. This work introduces a new generative perspective for flexible, inference-time controllable Explainable Artificial Intelligence (XAI) methods.

Keywords: Feature Attribution · Path-Based Attribution Method · Explainable Artificial Intelligence

1 Introduction

Deep Neural Networks (DNNs) have demonstrated exceptional performance in a wide range of domains; however, their opaque, black-box nature limits their use in high-stakes settings such as medical diagnosis, autonomous driving, or weather forecasting [24], where trust, debugging, safety, and fairness are essential [7, 30]. Explainable Artificial Intelligence (XAI), particularly input attribution, addresses this by assigning each input feature an importance score reflecting its contribution to the model output. Consequently, a wide array of methods, such as Saliency Maps [48], SHAP [38], and LRP [39] have been proposed. However, the trustworthiness of early attribution methods remains uncertain, since they often rely on heuristics and have been shown to fail basic sanity checks [1]. This raises concerns that such methods may generate explanations that are misleading or not faithful to the underlying model behavior [27].

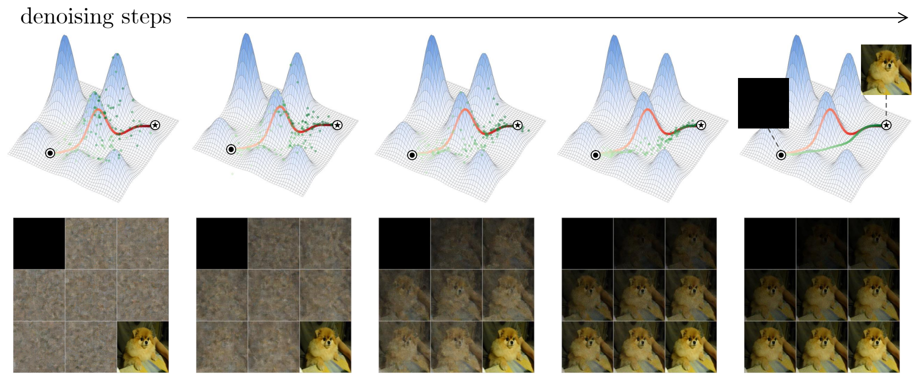


Fig. 1: Illustration of the path generation process of DiffIG. (Top) The diffusion model generates an attribution path from random noise through an iterative denoising process. The blue meshed field refers to the logit surface of the model. The red path represents the IG straight-line path, which is susceptible to noisy gradient accumulation and often traverses irregular regions of the model surface. In contrast, the green path denotes the DiffIG path, which leverages guided sampling to avoid regions that induce noisy gradients and produces more faithful attributions through controllable guidance (see Sec. 3). **(Bottom)** Visualization in the image space, where DiffIG transforms a noisy initial trajectory into a non-linear path connecting the baseline point ◉ to the input point ⍟.

To provide more principled explanations, Integrated Gradients (IG) [55] emerged as a prominent solution. IG is theoretically rooted in the AumannShapley value framework and path-based attribution formulations [3, 12], which provide strong theoretical guarantees. In particular, IG uniquely satisfies critical axioms such as *Efficiency* or *Completeness*, ensuring that the resulting feature attributions add up to the difference between the models prediction for the input and its prediction for the baseline. IG assigns importance scores by integrating gradients along a straight-line path from a neutral baseline, such as a black image, to the target input. Owing to this strong axiomatic foundation, IG has become widely used as a standard feature attribution method for complex models.

Despite these strengths, IG critically depends on the choice of integration path. The standard straight-line path is theoretically justified under the AumannShapley framework [3], where it is uniquely consistent with the axioms and probabilistically optimal for functions whose gradient fields are *conservative*, i.e., the path integral is path-independent and any baseline-to-input path yields the same attribution. This assumption breaks down for modern rectified DNNs: their piecewise-linear nature and discontinuous gradients at linear-region boundaries make the gradient field *highly discontinuous*, so the attribution integral becomes *path-dependent* and the straight-line path loses its optimality. Worse, this path can traverse complex regions of the function space, aggregating spurious feature contributions and misleading attributions [20, 53, 58, 60].

We propose **Diff**usion **I**ntegrated **G**radients (**DiffIG**), a path-based attribution method that learns integration paths via generative modeling. We train a diffusion model on the distribution of non-linear paths from a Stick-Breaking Process, capturing diverse trajectories of varying form and complexity. DiffIG further leverages guided sampling to embed user-defined priors at inference, steering path generation toward properties such as faithfulness and complexity. This transforms path-based attribution from a fixed procedure into a flexible, controllable generative process.

Our main contributions are as follows: **(1)** We reformulate attribution path generation as a conditional generative modeling problem. **(2)** We propose **Diff**usion **I**ntegrated **G**radients (DiffIG), a novel method that enables flexible and inference-time controllable feature attribution. **(3)** We demonstrate through quantitative and qualitative experiments that DiffIG matches or outperforms existing path-based methods, yielding more faithful and perceptually aligned explanations.

2 Background

2.1 Path-based Attribution Methods

For an input $x \in \mathbb{R}^n$ and a differentiable model $f : \mathbb{R}^n \to \mathbb{R}$, path-based attribution methods accumulate gradients along a path $\gamma(t)$ for $t \in [0,1]$. This path connects a baseline x' to the input x (i.e., $\gamma(0) = x'$ and $\gamma(1) = x$). In vision models, the baseline x' is typically chosen as either a black or a white image. The attribution $\mathcal{A}_i$ for the i-th feature is defined as [55]:

$$\mathcal{A}_i(\gamma) = \int_0^1 \frac{\partial f(\gamma(t))}{\partial \gamma_i(t)} \frac{\partial \gamma_i(t)}{\partial t} \,\mathrm{d}t, \tag{1}$$

where the integral consists of two key terms: $\frac{\partial f(\gamma(t))}{\partial \gamma_i(t)}$, the model gradient at $\gamma(t)$, and $\frac{\partial \gamma_i(t)}{\partial t}$, the path direction.

Integrated Gradients (IG). When the path is the straight line $\gamma(t) = x' + t(x - x')$, Equation (1) recovers the canonical *Integrated Gradients* (IG) [55]. For this path, the direction term $\frac{\partial \gamma_i(t)}{\partial t}$ simplifies to $(x_i - x'_i)$, yielding:

$$\mathcal{A}_i(\gamma) = (x_i - x'_i) \int_0^1 \frac{\partial f(x' + t(x - x'))}{\partial x_i} \,\mathrm{d}t. \tag{2}$$

Path-Dependency Problem in Rectified DNNs. The theoretical elegance of IG relies on the assumption that the gradient field of f is *conservative*. In this case, the path integral is path-independent, and any γ connecting x' and x yields a unique attribution. However, this assumption does not hold for modern rectified DNNs. Due to their piecewise-linear structure [8], the input space is partitioned into numerous distinct decision regions. Consequently, the gradient field of such models is highly discontinuous [4], making the resulting feature-wise attribution integral highly sensitive to the chosen path and susceptible to numerical errors [23].

In practice, the standard straight-line path remains a convenient yet naive heuristic that ignores the complex geometry of the underlying decision surface of the model. As a result, it often traverses discontinuities at region boundaries, accumulating noisy or spurious gradients. This can lead to misleading feature attribution allocation, resulting in attributions that are unfaithful to the true decision-making process of the model. Consequently, the key challenge for improving path-based attribution lies in discovering an adaptive and model-aware integration path.

2.2 Diffusion Probabilistic Models

Diffusion probabilistic models [17,51] are generative models that learn to reverse a fixed noising process. The learned reverse process $p_\theta(\gamma^{\tau-1} \mid \gamma^{\tau})$ inverts a predefined forward process $q(\gamma^{\tau} \mid \gamma^{\tau-1})$ that gradually adds noise to data until it becomes a standard Gaussian prior. The model learns a data distribution represented as a Markov chain with Gaussian transitions:

$$p_\theta(\gamma^0) = \int p(\gamma^M) \prod_{\tau=1}^{M} p_\theta(\gamma^{\tau-1} \mid \gamma^{\tau}) \, \mathrm{d}\gamma^{1:M}. \tag{3}$$

Here, γ^0 represents the clean data, and $p(\gamma^M)$ is the standard Gaussian prior after M noising steps. The training objective is to find parameters θ that minimize a variational bound on the negative log-likelihood of the reverse process: $\theta^* = \arg\min_\theta -\mathbb{E}_{\gamma^0}[\log p_\theta(\gamma^0)]$. In practice, the reverse transition $p_\theta(\gamma^{\tau-1} \mid \gamma^{\tau})$ is often modeled as a Gaussian distribution with learnable means:

$$p_\theta(\gamma^{\tau-1} \mid \gamma^{\tau}) = \mathcal{N}(\gamma^{\tau-1} \mid \mu_\theta(\gamma^{\tau}), \Sigma^{\tau}), \tag{4}$$

while the forward process $q(\gamma^{\tau} \mid \gamma^{\tau-1})$ is fixed and is typically specified by a predefined variance schedule.

3 Diffusion Integrated Gradients

We first formulate attribution path generation as conditional generative modeling, then introduce **Diffusion Integrated Gradients (DiffIG)**, a path-based attribution method that learns integration paths through generative modeling and guided sampling. As highlighted in Figure 2, DiffIG offers distinct advantages over existing path-based methods in adaptive path generation, global optimization, and controllability, and Figure 1 overviews its path generation process.

3.1 Path Generation with Diffusion

As discussed in Sec. 2.1, the quality of the resulting attributions heavily depends on the chosen path. In particular, methods such as Integrated Gradients (IG) employ a model-agnostic linear path, making their attributions highly sensitive to the underlying decision surface of the model. To mitigate the accumulation of noise and unfaithful attributions caused by such paths, path-based attribution methods aim to identify a path γ that connects a baseline x' to an input x. Formally, this can be formulated as finding an optimal path γ^* that maximizes an objective function $\mathcal{J}$:

$$\gamma^* = \underset{\gamma \in \Gamma}{\arg\max}\ \mathcal{J}(\gamma), \tag{5}$$

where Γ denotes the set of all possible paths, γ represents a path, and $\mathcal{J}$ indicates the corresponding objective value of that path.

For example, Guided IG (GIG) [23] seeks a path that maximizes the objective $\mathcal{J}(\gamma)$, defined as the negative of a cost combining path instability ℓ_{noise} and deviation from the straight-line path ℓ_{distance}:

$$\begin{aligned}
\mathcal{J}(\gamma) &= -(\ell_{\text{noise}} + \lambda \ell_{\text{distance}}) \\
\ell_{\text{noise}} &= \sum_{i=1}^{n} \int_{t=0}^{1} \left| \frac{\partial f(\gamma(t))}{\partial \gamma_i(t)} \frac{\partial \gamma_i(t)}{\partial t} \right| \mathrm{d}t \\
\ell_{\text{distance}} &= \int_{t=0}^{1} ||\gamma(t) - \gamma_{\text{IG}}(t)|| \,\mathrm{d}t,
\end{aligned}$$

where $\gamma_{\text{IG}}(t)$ is the straight-line path, ℓ_{noise} quantifies the accumulated gradient magnitude on irrelevant features, and ℓ_{distance} measures the deviation from the linear path. However, optimizing such a complex objective over a continuous function space, particularly under the complex topology of the underlying decision surface of the model, is generally intractable and necessitates heuristic or greedy approximations. Consequently, GIG relies on a *greedy*, iterative approximation that performs heuristic local updates, preventing it from identifying a globally consistent integration path.

Instead of explicitly minimizing an objective as in Eq. (5), we draw on offline reinforcement learning, where sequential decision-making is cast as generating desired trajectories conditioned on their outcomes [2, 19]. Following this view, we reformulate finding an optimal path as conditional generative modeling:

$$\theta^* = \underset{\theta}{\arg\max}\ \mathbb{E}_{\gamma \sim \mathcal{D}} \big[\log p_\theta(\gamma | y = \mathcal{J}(\gamma)) \big], \tag{6}$$

where $\mathcal{D}$ represents the dataset of pre-collected paths, and the conditional label y denotes the score $\mathcal{J}(\gamma)$ associated with a path γ from this dataset. By learning this conditional data distribution p_θ, we can later generate new paths γ that possess high $\mathcal{J}(\gamma)$ values.

Unlike unconstrained image synthesis, our objective is path generation between fixed endpoints. To ensure that the generated path $\gamma(t)$ indeed connects the baseline and the input, we explicitly enforce the boundary conditions

$\gamma(0) = x'$ and $\gamma(1) = x$ at every denoising step during sampling. This guarantees that the diffusion model generates only the valid intermediate attribution trajectories that start from the baseline and terminate at the input.

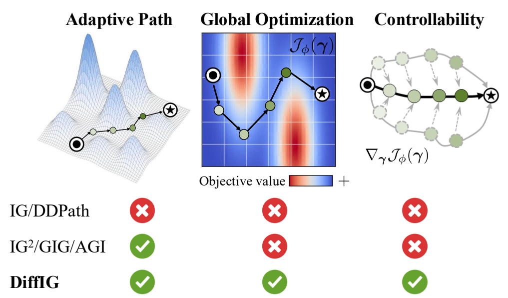


Fig. 2: DiffIG capabilities. Adaptive Path indicates whether a method can adapt its integration path to the complex decision surface of the model. **Global Optimization** refers to the ability to find a globally optimal path, rather than relying on greedy or local search heuristics. **Controllability** denotes the ability to flexibly guide path generation at inference time toward desired properties.

Training. We use diffusion probabilistic models [17, 51] to learn the conditional distribution $p_\theta(\gamma|y = \mathcal{J}(\gamma))$. However, instead of training this conditional distribution directly, we adopt an unconditional diffusion model with classifier guidance [9]. This approach involves training two separate models. First, we train an unconditional diffusion model to learn the unconditional distribution of paths $p_\theta(\gamma)$ from the dataset $\mathcal{D}$. As part of this process, we train an $\boldsymbol{\epsilon}$-model, $\boldsymbol{\epsilon}_\theta$, to estimate the source noise rather than predicting the mean $\boldsymbol{\mu}_\theta$. This enables a simplified objective [17]:

$$\mathcal{L}(\theta) := \mathbb{E}_{\tau,\boldsymbol{\epsilon},\gamma^0}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\gamma^\tau)\|_2^2], \tag{7}$$

where $\tau \in \{1, ..., M\}$ is the diffusion timestep, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the target noise, and γ^τ is the path corrupted by the noise $\boldsymbol{\epsilon}$ from the noiseless path γ^0. Second, to steer the diffusion model towards generating paths with high objective scores, we use energy-guided sampling. This samples from a perturbed distribution $\tilde{p}_\theta(\gamma^0)$ that prioritizes paths with higher objective scores $\mathcal{J}(\gamma^0)$, defined as:

$$\tilde{p}_\theta(\gamma^0) \propto p_\theta(\gamma^0) \exp\left(\mathcal{J}(\gamma^0)\right). \tag{8}$$

To enable this guidance, we train a separate regression network, $\mathcal{J}_\phi$. This network is trained to predict the objective score $\mathcal{J}(\gamma^0)$ from a noisily perturbed version γ^τ at any timestep τ. This is accomplished through the following mean-square-error (MSE) objective:

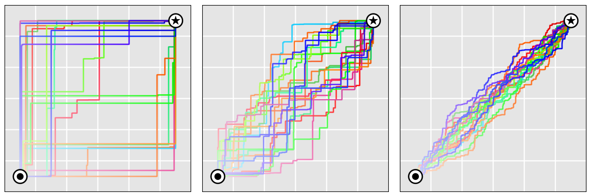
$$\min_\phi \mathbb{E}_{\tau,\boldsymbol{\epsilon},\gamma^0}\left[\|\mathcal{J}_\phi(\gamma^\tau) - \mathcal{J}(\gamma^0)\|_2^2\right]. \tag{9}$$

During the sampling stage, classifier guidance [9] is employed, incorporating the gradients of $\mathcal{J}_\phi$ into the reverse diffusion process. Specifically, the mean of the reverse transition is updated as follows:

$$\tilde{p}_\theta(\gamma^{\tau-1}|\gamma^\tau) = \mathcal{N}(\gamma^{\tau-1}|\boldsymbol{\mu}_\theta(\gamma^\tau) + \omega\boldsymbol{\Sigma}^\tau g, \boldsymbol{\Sigma}^\tau), \tag{10}$$

where $g = \nabla_\gamma \mathcal{J}_\phi(\gamma)|_{\gamma=\boldsymbol{\mu}_\theta(\gamma^\tau)}$ and ω is the guidance scale that controls the strength of the guidance.

3.2 Synthetic Path Data Generation



(a) $\alpha = 1$ **(b)** $\alpha = 10$ **(c)** $\alpha = 100$

Fig. 3: 2D visualization of stochastic paths generated by the Stick-Breaking Process (SBP). Each curve represents a sampled path from the baseline ⦿ to the input ✪. The concentration parameter α controls the distribution of paths. A small α encourages high variance, producing diverse non-linear paths, while a large α constrains the paths to be smoother and closer to the linear path.

Training the unconditional diffusion model and the guidance network requires a large and diverse dataset $\mathcal{D}$ of candidate paths. To this end, we adopt the Stick-Breaking Process (SBP) [21, 47] to generate a rich set of non-linear paths. Conceptually, SBP sequentially breaks a unit-length stick into random segments, where the resulting proportions define a random measure over the interval $[0, 1]$.

For each feature dimension i, we sample a measure G_i via SBP:

$$G_i(t) = \sum_{k=1}^{\infty} \pi_k \delta_{t_k}(t), \tag{11}$$

$$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_j), \tag{12}$$

$$\beta_k \sim \mathrm{Beta}(1, \alpha), \quad t_k \sim H, \tag{13}$$

where $\delta_{t_k}(t)$ is a Dirac delta function centered at t_k, and H is a base distribution. Following [21], we set the base distribution H to the uniform distribution $U(0, 1)$. The cumulative distribution function (CDF) $F_{G_i}(t)$ of G_i is then used to define a non-linear interpolation between the baseline x'_i and the input x_i:

$$\gamma_i(t) = x'_i + F_{G_i}(t)(x_i - x'_i). \tag{14}$$

This construction ensures that $\gamma_i(0) = x'_i$ and $\gamma_i(1) = x_i$. Since F_{G_i} is a non-decreasing CDF with $F_{G_i}(0) = 0$ and $F_{G_i}(1) = 1$, $\gamma_i(t)$ transitions monotonically from x'_i to x_i. The concentration parameter α controls the diversity of generated paths. A large α produces paths close to the straight-line path used in standard IG, while a small α yields more diverse and non-linear trajectories. Figure 3 illustrates how different α values lead to distinct path characteristics. To build $\mathcal{D}$, we generate paths across a wide range of α, ensuring a rich diversity of trajectories ranging from nearly linear to highly complex.

3.3 Guidance Objectives

To guide the diffusion model toward generating high-quality attribution paths, we train the guidance network $\mathcal{J}_\phi$ to predict objective values that reflect the quality of each path. Specifically, we define two complementary objectives that capture distinct aspects of attribution quality: *faithfulness* and *complexity*.

Faithfulness. A faithful attribution path should accurately reflect how the model prediction changes when important or unimportant features are perturbed. We quantify this by measuring the difference in model prediction confidence under two complementary perturbations, *insertion* and *deletion*. The *faithfulness score* of a path γ is defined as

$$\text{Faithfulness}(\gamma) = \mathbb{E}_{r \in \mathcal{R}} \Big[\text{Conf}^{\text{ins}}(r) - \text{Conf}^{\text{del}}(r) \Big], \tag{15}$$

where $\mathcal{R}$ is the set of perturbation ratios (e.g., $r \in \{0.1, 0.2, ..., 0.9\}$). Here, $\text{Conf}^{\text{ins}}(r)$ and $\text{Conf}^{\text{del}}(r)$ denote the softmax confidence of the model for the correct class when the most important features are inserted and deleted, respectively. A higher faithfulness score indicates that the attribution path better captures features that truly drive the decision of the model, leading to a larger drop in confidence when important features are removed and stable confidence when unimportant ones are altered.

Complexity. While high faithfulness is desirable, overly complex attributions that rely on many dispersed features may hinder interpretability. To regularize this, we introduce a *complexity* measure inspired by attribution entropy [5]. Given an attribution map $\mathcal{A} \in \mathbb{R}^n$ for a path γ, we define normalized weights $p_i = \frac{|\mathcal{A}_i|}{\sum_j |\mathcal{A}_j|}$ and the entropy-based complexity score:

$$\text{Comp}(\gamma) = -\sum_{i=1}^{n} p_i \log(p_i + \varepsilon), \tag{16}$$

where ε is a small constant added for numerical stability. Lower complexity corresponds to sparser, more human-interpretable explanations, while higher complexity indicates diffuse and harder-to-interpret attributions.

Combined Guidance. To guide the diffusion model toward generating paths that are both faithful and interpretable, we employ two complementary guidance networks: $\mathcal{J}_{\phi_1}$ for *faithfulness* and $\mathcal{J}_{\phi_2}$ for *complexity*. Each network is independently trained to regress its respective objective from noisy path samples following the regression formulation in Eq. (9).

During the reverse diffusion process, the dual guidance signals are combined into a unified gradient field that jointly influences the path generation:

$$g = \Big(\lambda_{\text{faith}} \nabla_\gamma \mathcal{J}_{\phi_1}(\gamma) + \lambda_{\text{comp}} \nabla_\gamma \mathcal{J}_{\phi_2}(\gamma) \Big) \Big|_{\gamma = \boldsymbol{\mu}_\theta(\gamma^\tau)}, \tag{17}$$

where λ_{faith} and λ_{comp} are coefficients balancing the two guidance strengths. This combined gradient g is injected into the reverse transition of the diffusion model in Eq. (10), steering the denoising trajectory toward attribution paths that are both faithful to the model reasoning and sufficiently simple to interpret. In all our main experiments, we fix the global guidance scale ω=1 in Eq. (10) and control the overall guidance strength directly through λ_{faith} and λ_{comp}; the effect of varying ω is examined separately in Appendix K of the supplementary material.

3.4 Practical Algorithm

Path Generation over Latent Space. Generating diffusion paths directly in the pixel space is computationally prohibitive due to the high dimensionality of images. Instead, we employ a pre-trained β-VAE [16, 34] encoder $\mathcal{E}_{\text{enc}}$ and decoder $\mathcal{D}_{\text{dec}}$ to construct paths in a lower-dimensional latent space. Given an input x and its baseline x', we first obtain their latent representations $z = \mathcal{E}_{\text{enc}}(x)$ and $z' = \mathcal{E}_{\text{enc}}(x')$. The diffusion model is trained in the latent space to learn the distribution of latent trajectories $\gamma_z(t)$ that, when decoded, yield non-linear paths $\gamma(t)$ in the image space. At inference time, it generates latent trajectories $\gamma_z(t)$, which are decoded back to the image domain via $\mathcal{D}_{\text{dec}}$:

$$\gamma(t) = \mathcal{D}_{\text{dec}}(\gamma_z(t)), \quad \text{where } \gamma_z(0) = z',\ \gamma_z(1) = z. \tag{18}$$

Stochastic Multi-path Sampling. Diffusion sampling naturally enables stochastic generation of multiple attribution paths from the guidance-conditioned distribution $\tilde{p}_\theta(\gamma) \propto p_\theta(\gamma) \exp(\mathcal{J}(\gamma))$. To leverage this property, we generate N distinct path candidates $\{\gamma^{(k)}(t)\}_{k=1}^N$ between the baseline x' and the input x, where the superscript (k) indexes each sampled path. We consider two strategies to utilize these multiple paths:

(i) Best-of-N Search: Among the sampled candidates, we select the path that best balances the two guidance objectives of faithfulness and complexity:

$$\gamma^* = \arg\max_{\gamma^{(k)}} \Big[\lambda_{\text{faith}}\, \mathcal{J}_{\phi_1}(\gamma^{(k)}) + \lambda_{\text{comp}}\, \mathcal{J}_{\phi_2}(\gamma^{(k)})\Big]. \tag{19}$$

The selected path γ^* is then used to compute the final attribution via Integrated Gradients (IG).

(ii) Multi-path Aggregation: Alternatively, we compute attributions for all sampled paths and aggregate them to obtain a more consistent explanation. Let $\mathcal{A}(x; \gamma^{(k)})$ denote the attribution map computed along a specific path $\gamma^{(k)}$. We then aggregate the path-dependent attributions to form the final attribution map for x:

$$\bar{\mathcal{A}}(x) = \mathcal{M}\big(\{\mathcal{A}(x; \gamma^{(k)})\}_{k=1}^N\big), \tag{20}$$

where $\mathcal{M}$ denotes an aggregation operator such as mean, median, or variance-weighted mean. This aggregation reduces the stochastic variance among sampled paths and yields more reliable and consistent attributions. Further details on each aggregation strategy are provided in Appendix D.

Axiomatic Properties. Since DiffIG computes attributions via the path integral in Eq. (1), it is a path-based attribution method and naturally inherits the axiomatic properties established for this class of methods [55]. In particular, we formally prove in Appendix B that DiffIG satisfies *Completeness* (attributions sum to $f(x) - f(x')$), *Sensitivity* (features with zero partial derivatives receive zero attribution), and *Implementation Invariance* (attributions depend solely on the model's input-output behavior). These guarantees hold for any individual path generated by the diffusion model, and completeness is additionally preserved under mean aggregation of multiple paths.

4 Experiments

In this section, we empirically evaluate the effectiveness of our proposed DiffIG method. Specifically, we examine **(1)** whether DiffIG produces explanations that are more perceptually aligned in both quantitative and qualitative terms, **(2)** whether its performance can be further enhanced through multi-path sampling, and **(3)** whether inference-time guidance can effectively steer path generation toward desired objectives. A complete description of our experimental setup and implementation details is provided in Appendices C and Q of the supplementary material.

4.1 Experimental Setup

Datasets and Models We evaluate the effectiveness of DiffIG on two real-world image datasets: (1) the Oxford-IIIT Pet dataset [42], which contains images of 37 pet categories, and (2) the Mini-ImageNet dataset [54], a subset of the larger ImageNet corpus comprising 100 categories. For classification, we use publicly available pre-trained models including VGG16 [49], InceptionV1 [56], and ResNet18 [14], which are further finetuned on each dataset.

Baselines To demonstrate the effectiveness of DiffIG, we compare it against seven representative path-based attribution methods. Our comparison includes IG [55]; optimization-based methods (GIG [23], IG^2 [63], and AGI [41]); stochastic process-based methods (SPI [21]); and manifold-informed methods (EIG [22], MIG [60]).

Evaluation Metrics To quantitatively evaluate the quality of attribution maps, we adopt the widely used pixel perturbation framework [43], which evaluates how the model output varies as important pixels are progressively inserted or deleted. Specifically, we use the Insertion and Deletion metrics [43], and their difference, the DiffID score [59]. The Insertion metric computes the area under the confidence curve (AUC) as pixels are added in order of importance, while Deletion does so as pixels are removed. The DiffID score captures the gap between the two AUCs, offering a unified measure of both sensitivity and selectivity, where higher values indicate more faithful and discriminative attributions.

4.2 Qualitative Results

We qualitatively compare the DiffIG attribution maps with those from other path-based methods. Attribution results are computed for random validation images from the Oxford-IIIT Pet and Mini-ImageNet datasets. As illustrated in Figure 4, DiffIG produces attribution maps that more precisely highlight important pixels and more comprehensively localize salient objects. For instance, in the *Meerkat* example, while other methods emphasize only the two central figures, DiffIG additionally identifies the third meerkat on the far right(Appendix N). Furthermore, local optimization approaches such as IG^2 and AGI tend to follow

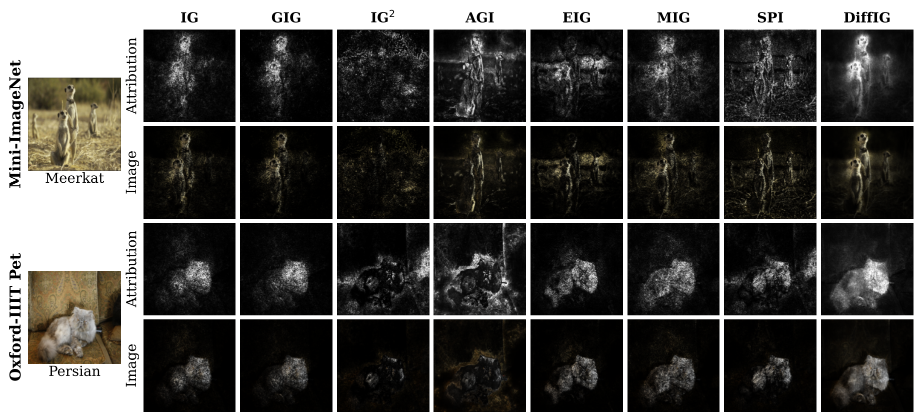


Fig. 4: Qualitative comparison of path-based attribution methods on Oxford-IIIT Pet and Mini-ImageNet using VGG16. The *Attribution* rows show the attribution maps, and the *Image* rows show them overlaid on the input images.

spurious path features, leading to misleading attribution aggregation. This issue is evident in the *Persian* example, where these methods are distracted by the high-frequency texture of the sofa, incorrectly assign importance to the background, and create spurious *ghost cat* attributions on the upper right side of the image. In comparison, DiffIG accurately attributes importance to both the cats face and body, reflecting a more complete and faithful interpretation of the model reasoning process. Additional qualitative results appear in Appendix R.

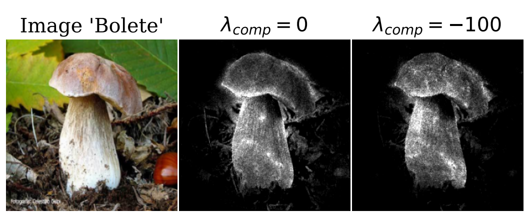


Fig. 5: Effect of Complexity Guidance. The attributions for $\lambda_{\text{comp}} = 0$ (no guidance) and -100 (strong), showing sparser and more focused attributions as the guidance strength increases.

Flexible Control of Sparseness. Figure 5 illustrates that DiffIG enables flexible control over the sparsity of attributions through the complexity guidance. By increasing the guidance strength, DiffIG produces sparser and more focused attributions while effectively suppressing irrelevant background features.

4.3 Quantitative Results

Table 1 presents the quantitative comparison of DiffIG with other path-based methods. Our approach consistently matches or surpasses the performance of existing baselines across various datasets and model architectures. We attribute these improvements to the ability of DiffIG to perform a guided, non-greedy search over generated path candidates, rather than relying on greedy or local search heuristics as in methods such as GIG, IG^2, and AGI. Figure 6 further

Table 1: Quantitative results across different datasets and model architectures. We compare DiffIG with various path-based attribution methods on the full Oxford-IIIT Pet validation set (370 images) and 500 randomly sampled images from the Mini-ImageNet validation set. **Bolded** values indicate the best performance, and underlined values indicate the second best.

Data	Method	ResNet18			VGG16			Inception V1		
		DiffID (↑)	Ins (↑)	Del (↓)	DiffID (↑)	Ins (↑)	Del (↓)	DiffID (↑)	Ins (↑)	Del (↓)
Oxford-IIIT Pet	IG [55]	0.3213	0.4892	0.1679	0.4654	0.5523	0.0869	0.3051	0.4333	0.1282
	GIG [23]	0.3486	0.5065	0.1579	0.4859	0.5659	0.0799	0.3085	0.4376	0.1290
	IG^2 [63]	0.2767	0.4385	0.1619	0.3016	0.4259	0.1243	0.2228	0.3650	0.1422
	AGI [41]	0.3242	0.4879	0.1637	0.4930	0.5861	0.0930	0.4092	0.5156	**0.1064**
	EIG [22]	0.2680	0.4532	0.1852	0.3950	0.4898	0.0947	0.2698	0.4108	0.1410
	MIG [60]	0.2552	0.4316	0.1764	0.3681	0.4723	0.1041	0.2602	0.3963	0.1361
	SPI [21]	0.2738	0.4414	0.1676	0.4202	0.5046	0.0844	0.2292	0.3751	0.1459
	DiffIG (Ours)	**0.5067**	**0.6300**	**0.1233**	**0.6368**	**0.7128**	**0.0760**	**0.4817**	**0.6112**	0.1296
Mini-ImageNet	IG [55]	0.2147	0.4522	0.2375	0.3158	0.4786	0.1628	0.2975	0.5168	0.2194
	GIG [23]	0.1896	0.4369	0.2474	0.3222	0.4803	0.1581	0.2861	0.5152	0.2291
	IG^2 [63]	0.1505	0.3858	0.2353	0.2251	0.4091	0.1840	0.2446	0.4423	**0.1977**
	AGI [41]	0.2949	0.5018	**0.2068**	0.4228	0.5695	**0.1466**	**0.3579**	0.5714	0.2135
	EIG [22]	0.1480	0.4166	0.2686	0.2385	0.4374	0.1988	0.2312	0.4795	0.2483
	MIG [60]	0.2227	0.4398	0.2171	0.3106	0.4660	0.1554	0.3057	0.5162	0.2105
	SPI [21]	0.2325	0.4491	0.2165	0.2763	0.4463	0.1700	0.2831	0.5112	0.2281
	DiffIG (Ours)	**0.3126**	**0.5747**	0.2621	**0.4298**	**0.6037**	0.1739	0.3578	**0.6285**	0.2707

demonstrates the effectiveness of the multi-path sampling in DiffIG. This highlights that generating and utilizing multiple diffusion-sampled paths leads to more stable and consistent attributions.

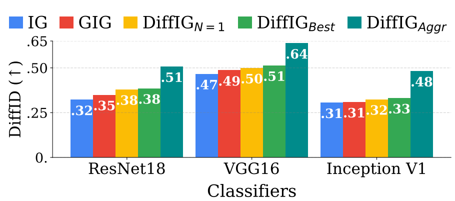


Fig. 6: Effect of Multi-path Sampling of DiffIG. The DiffIG variants correspond to *one-path generation (N=1)*, *Best-of-N Search (Best)*, and *multi-path mean aggregation (Aggr)*.

Validity of Guided Generation. DiffIG does not directly optimize the final evaluation pipeline, since the guidance predictor is a proxy regressor trained on noisy latent paths, while the final attribution faithfulness is evaluated later in pixel space. We also jointly optimize faithfulness and complexity, which discourages degenerate solutions that exploit a single score. This is further supported by the Pareto frontier analysis in Appendix Q, where the trade-off between Insertion and Deletion confirms that faithfulness improves through balancing multiple objectives. As further evidence that DiffIG does not merely fit the evaluation metric, it ranks first on the independent Faithfulness Correlation metric (computed on absolute attributions) from Quantus [15], where higher is better (DiffIG −0.0025 vs. AGI −0.0042 and GIG −0.0093 on Oxford-IIIT Pet with ResNet18; see Appendix M).

Additional Experiments. We provide further analyses in the supplementary material. Appendix E compares DiffIG against DDPath [33], a concurrent diffusion-based attribution method, demonstrating consistent improvements in DiffID and Insertion. Appendix F evaluates DiffIG on ViT-B/16 [10], confirming that performance gains generalize beyond CNNs to transformer architectures. Appendix G presents a runtime analysis showing that DiffIG achieves a favorable trade-off between computational cost and attribution quality. Appendix H reports sensitivity analysis across different VAE backbones (MAR [34], Stable Diffusion 2.1 [46], Kandinsky 2.1 [45]), confirming robustness to the choice of generative prior. Appendix I visualizes path dependence and shows robustness to the SBP concentration α, and Appendix J shows robustness to the baseline input, where a black–white mean mitigates the dark-region limitation. Appendix K examines the latent–pixel domain gap on a controlled toy example; Appendix L compares against IG smoothing variants [50]; and Appendices O and P give step-wise visualizations and a failure-case analysis.

5 Related Work

Path-based Attribution Methods. As discussed in Section 2.1, the path-dependent nature of attributions in highly discontinuous gradient fields poses a key challenge for reliable explanations. Therefore, the central problem lies in identifying a desirable or faithful integration path.

One line of research formulates path identification as an optimization problem. For example, Guided IG (GIG) [23] argues that high-curvature areas of the output surface cause noisy gradients and should be avoided. Geometrically Guided IG (GGIG) [44] selects a path toward class-specific activation maximization. IG^2 [63] seeks a path with steeper gradients to avoid saturation areas. Adversarial Gradient Integration (AGI) [41] uses loss minimization paths from adversarial examples. Spectral IG [26] instead refines attributions in a coarse-to-fine manner using the singular value decomposition. However, these methods rely on heuristic, locally optimized search procedures rather than directly optimizing global objectives, which can result in sub-optimal paths that may not be holistically faithful.

Another line of work models the path as a distribution and applies stochastic processes. SAlient Manipulation Path (SAMP) [61] defines the path selection process as an additive stochastic process, akin to Brownian motion, and regards attribution as a cumulative process where contributions are allocated sequentially. Similarly, Stick-breaking Path Integration (SPI) [21] also defines a distribution over possible paths using a stick-breaking process. Despite capturing stochastic variability, these methods remain model-agnostic and fail to account for the underlying decision surface.

A recent line of work incorporates data manifold information to produce perceptually faithful attributions. For example, Enhanced Integrated Gradients (EIG) [22] and Manifold IG (MIG) [60] utilize the learned latent space of a variational autoencoder (VAE) to construct linear or geodesic paths that better respect the underlying data manifold. Manifold-Aligned Guided IG [25] similarly steers the path to remain aligned with the data manifold. Denoising Diffusion

Path (DDPath) [33] leverages a pre-trained diffusion model to construct the attribution path. The core idea is that the reverse diffusion process, which progressively transforms noise into an image, can be interpreted as a path that stays closer to the data manifold than a simple straight line. While these methods incorporate the data manifold into path generation, they remain constrained by fixed generative processes, lacking flexibility and controllable adaptation in path generation.

Planning with Diffusion Models. Diffusion probabilistic models capture complex data distributions by progressively denoising noise into structured samples [17,51]. This iterative process can be interpreted as learning the *score* of the data distribution [52], with connections to score matching [18] and energy-based models (EBMs) [11,13,40]. In offline reinforcement learning (RL), they have been applied to high-dimensional synthesis [29, 37, 57] and to planning [19, 28, 31, 32], where the goal is to synthesize action trajectories achieving high returns. For instance, Diffuser [19] steers an unconditional diffusion model toward high-return trajectories via a return predictor at inference [9], while Decision Diffuser [2] instead applies *classifier-free guidance* to condition directly on reward or constraint signals. Subsequent works extend this paradigm through adaptive fine-tuning [36], hierarchical control [6, 35], and multi-agent settings [62].

Inspired by these advances, DiffIG reformulates finding an optimal attribution path as a conditional generative modeling problem, highlighting diffusion models as a unified framework for generating adaptive paths for faithful and interpretable attribution.

6 Conclusion

We introduced *Diffusion Integrated Gradients* (DiffIG), a path-based attribution method that formulates integration path generation as a conditional generative modeling problem. By leveraging diffusion models trained on stochastic path distributions with guided sampling at inference, DiffIG enables adaptive, guided non-greedy search over generated path candidates, and inference-time controllable path generation, transforming attribution into a flexible generative process steerable toward user-defined objectives. Extensive experiments show that DiffIG matches or outperforms existing path-based methods, producing explanations that are both perceptually aligned and faithful, suggesting that generative modeling is a principled framework for controllable and interpretable attribution.

Limitations. DiffIG currently relies on a fixed baseline (e.g., a black image), which can yield less informative attributions for visually dark regions such as dark eyes. As a simple remedy, averaging the black- and white-baseline attributions mitigates this dark-region issue without retraining(Appendix J); adaptive or learned baselines and modalities such as text or tabular data are promising directions.

Acknowledgements

This work was supported by the Institute for Information & Communications Technology Planning & Evaluation (IITP), funded by the Ministry of Science and ICT (MSIT), Korea, under grants RS-2019-II190075 (Artificial Intelligence Graduate School Program (KAIST)), RS-2022-II220984 (Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation), RS-2024-00457882 (AI Research Hub Project), and RS-2026-26644375 (Leading Generative AI Human Resources Development), and by the MSIT InnoCORE program (N10250156).

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
2. Ajay, A., Du, Y., Gupta, A., Tenenbaum, J.B., Jaakkola, T.S., Agrawal, P.: Is conditional generative modeling all you need for decision-making? In: International Conference on Learning Representations (ICLR) (2023)
3. Aumann, R.J., Shapley, L.S.: Values of non-atomic games. Princeton University Press (1974)
4. Balduzzi, D., Frean, M., Leary, L., Lewis, J., Ma, K.W.D., McWilliams, B.: The shattered gradients problem: If resnets are the answer, then what is the question? In: International conference on machine learning. pp. 342–350. PMLR (2017)
5. Bhatt, U., Weller, A., Moura, J.M.: Evaluating and aggregating feature-based model explanations. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI). pp. 3016–3022 (2020)
6. Chen, C., Deng, F., Kawaguchi, K., Gulcehre, C., Ahn, S.: Simple hierarchical planning with diffusion. In: International Conference on Learning Representations (ICLR) (2024)
7. Chen, L., Lin, S., Lu, X., Cao, D., Wu, H., Guo, C., Liu, C., Wang, F.Y.: Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems **22**(6), 3234–3246 (2021)
8. Chu, L., Hu, X., Hu, J., Wang, L., Pei, J.: Exact and consistent interpretation for piecewise linear neural networks: A closed form solution. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1244–1253 (2018)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), `https://openreview.net/forum?id=YicbFdNTTy`
11. Du, Y., Mordatch, I.: Implicit generation and modeling with energy-based models. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
12. Friedman, E.J.: Paths and consistency in additive cost sharing. International Journal of Game Theory **32**(4), 501–518 (2004)

13. Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Zemel, R.: Learning the stein discrepancy for training and evaluating energy-based models without sampling. In: International Conference on Machine Learning (ICML) (2020)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Höhne, M.M.C.: Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. Journal of Machine Learning Research **24**(34), 1–11 (2023)
16. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: International conference on learning representations (2017)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
18. Hyvärinen, A.: Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research **6**, 695–709 (2005)
19. Janner, M., Du, Y., Tenenbaum, J.B., Levine, S.: Planning with diffusion for flexible behavior synthesis. In: International Conference on Machine Learning (ICML) (2022)
20. Jeon, G., Jeong, H., Choi, J.: Distilled gradient aggregation: Purify features for input attribution in the deep neural network. Advances in Neural Information Processing Systems **35**, 26478–26491 (2022)
21. Jeon, G., Jeong, H., Choi, J.: Beyond single path integrated gradients for reliable input attribution via randomized path sampling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2052–2061 (2023)
22. Jha, A., K. Aicher, J., R. Gazzara, M., Singh, D., Barash, Y.: Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study. Genome biology **21**(1), 149 (2020)
23. Kapishnikov, A., Venugopalan, S., Avci, B., Wedin, B., Terry, M., Bolukbasi, T.: Guided integrated gradients: An adaptive path method for removing noise. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5050–5058 (2021)
24. Kim, S., Choi, J., Choi, Y., Lee, S., Stitsyuk, A., Park, M., Jeong, S., Baek, Y.H., Choi, J.: Explainable ai-based interface system for weather forecasting model. In: International Conference on Human-Computer Interaction. pp. 101–119. Springer (2023)
25. Kim, S., Lim, S., Lee, K., Choi, J.: Manifold-aligned guided integrated gradients for reliable feature attribution (2026), `https://arxiv.org/abs/2605.02167`
26. Kim, S., Lim, S., Lee, K., Choi, J.: Spectral integrated gradients for coarse-to-fine feature attribution. In: Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2026), `http://arxiv.org/abs/2605.19607`
27. Kindermans, P.J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B.: The (un) reliability of saliency methods. Explainable AI: Interpreting, explaining and visualizing deep learning pp. 267–280 (2019)
28. Lee, K., Choi, J.: Local manifold approximation and projection for manifold-aware diffusion planning. In: Forty-second International Conference on Machine Learning (2025), `arXivpreprintarXiv:2506.00867`
29. Lee, K., Choi, J.: State-covering trajectory stitching for diffusion planners. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)

30. Lee, K., Kim, S., Choi, J.: Adaptive and explainable deployment of navigation skills via hierarchical deep reinforcement learning. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 11608–11614. IEEE (2023)
31. Lee, K., Kim, S., Choi, J.: Refining diffusion planner for reliable behavior synthesis by automatic detection of infeasible plans. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
32. Lee, K., Luo, Y., Tong, A., Choi, J.: Refining compositional diffusion for reliable long-horizon planning. arXiv preprint arXiv:2605.03075 (2026)
33. Lei, Y., Li, Z., Zhang, J., Shan, H.: Denoising diffusion path: Attribution noise reduction with an auxiliary diffusion model. Advances in Neural Information Processing Systems **37**, 54003–54025 (2024)
34. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. Advances in Neural Information Processing Systems **37**, 56424–56445 (2024)
35. Li, W., Wang, X., Jin, B., Zha, H.: Hierarchical diffusion for offline decision making. In: International Conference on Machine Learning (ICML) (2023)
36. Liang, Z., Mu, Y., Ding, M., Ni, F., Tomizuka, M., Luo, P.: Adaptdiffuser: Diffusion models as adaptive self-evolving planners. In: International Conference on Machine Learning (ICML) (2023)
37. Lu, C., Ball, P., Teh, Y.W., Parker-Holder, J.: Synthetic experience replay. Advances in Neural Information Processing Systems (2023)
38. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
39. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning pp. 193–209 (2019)
40. Nijkamp, E., Hill, M., Zhu, S.C., Wu, Y.N.: Learning non-convergent non-persistent short-run mcmc toward energy-based model. Advances in Neural Information Processing Systems **32** (2019)
41. Pan, D., Li, X., Zhu, D.: Explaining deep neural network models with adversarial gradient integration. In: Thirtieth International Joint Conference on Artificial Intelligence (IJCAI) (2021)
42. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
43. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: The British Machine Vision Conference (BMVC) (2018)
44. Rahman, M.M., Lewis, N., Plis, S.: Geometrically guided saliency maps. In: ICLR 2022 Workshop on PAIR^{2}Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data (2022)
45. Razzhigaev, A., Shakhmatov, A., Maltseva, A., Arkhipkin, V., Pavlov, I., Ryabov, I., Kuts, A., Panchenko, A., Kuznetsov, A., Dimitrov, D.: Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP) (2023)
46. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
47. Sethuraman, J.: A constructive definition of dirichlet priors. Statistica sinica pp. 639–650 (1994)

48. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2014)
49. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
50. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
51. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning (ICML) (2015)
52. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
53. Srinivas, S., Fleuret, F.: Full-gradient representation for neural network visualization. Advances in neural information processing systems **32** (2019)
54. Sun, Q., Liu, Y., Chua, T.S., Schiele, B.: Meta-transfer learning for few-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 403–412 (2019)
55. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
56. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
57. Wang, R., Frans, K., Abbeel, P., Levine, S., Efros, A.A.: Prioritized generative replay. In: International Conference on Learning Representations (ICLR) (2025)
58. Yang, P., Akhtar, N., Wen, Z., Mian, A.: Local path integration for attribution. Proceedings of the AAAI Conference on Artificial Intelligence **37**(3), 3173–3180 (Jun 2023). `https://doi.org/10.1609/aaai.v37i3.25422`, `https://ojs.aaai.org/index.php/AAAI/article/view/25422`
59. Yang, P., Akhtar, N., Wen, Z., Shah, M., Mian, A.S.: Re-calibrating feature attributions for model interpretation. In: The Eleventh International Conference on Learning Representations (ICLR) (2023)
60. Zaher, E., Trzaskowski, M., Nguyen, Q., Roosta, F.: Manifold integrated gradients: Riemannian geometry for feature attribution. In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 58090–58104. PMLR (21–27 Jul 2024), `https://proceedings.mlr.press/v235/zaher24a.html`
61. Zhang, B., Zheng, W., Zhou, J., Lu, J.: Path choice matters for clear attributions in path methods. In: The Twelfth International Conference on Learning Representations (2024), `https://openreview.net/forum?id=gzYgsZgwXa`
62. Zhu, Z., Liu, M., Mao, L., Kang, B., Xu, M., Yu, Y., Ermon, S., Zhang, W.: Madiff: Offline multi-agent learning with diffusion models. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
63. Zhuo, Y., Ge, Z.: IG2: Integrated Gradient on Iterative Gradient Path for Feature Attribution . IEEE Transactions on Pattern Analysis & Machine Intelligence **46**(11), 7173–7190 (2024). `https://doi.org/10.1109/TPAMI.2024.3388092`