ATTRIBUTION VIA DISTRIBUTIONAL PATHS FOR INFORMATION REVELATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature attribution methods often either compare predictions under different subsets of observed feature information, as in SHAP-style methods, or integrate gradients along a path from a baseline to the input, as in Integrated Gradients. We introduce Reveal-IG, which combines these perspectives by defining a continuous path of distributions that progressively reveal information about an input. At each point along this path, the model is evaluated in expectation over partially revealed inputs, and attribution integrates the resulting change in expected prediction. This replaces IG's trajectory through individual input values with a trajectory through information states, while retaining a completeness property: feature attributions sum to the change in expected model output across the path. The framework naturally supports structured perturbations, including variable-resolution image probes and feature-wise uncertainty in tabular data. On synthetic diagnostics, Reveal-IG reduces path artifacts compared to pointwise input-space paths. Across ImageNet classification and tabular regression, Reveal-IG produces stable, signed attributions, leading on sign-sensitive image benchmarks while remaining competitive across standard attribution metrics.

1 INTRODUCTION

Explainable AI is an important component of modern machine learning systems, and feature attribution is one of its most widely used forms, assigning importance scores to input features in order to explain a model's prediction (Samek et al., 2021; Molnar, 2022). Two central approaches are Shapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) and Integrated Gradients (IG) (Sundararajan et al., 2017). Both satisfy completeness: attributions sum to the change in model output between a reference state and the input being explained. Both can be understood as path-based methods (Fig. 1). SHAP attributes to each feature its average contribution along discrete paths that iteratively reveal feature values, moving from a global reference distribution to the observed input. In contrast, IG decomposes a continuous path integral, typically along a trajectory in input space.

Despite this shared foundation, SHAP and IG expose distinct limitations. SHAP requires averaging over combinatorially many feature orderings, and its discrete reveal operation treats each feature as either unknown or fully observed. This discontinuous reveal operation is a severe intervention: a feature jumps to fully observed without allowing evidence to enter the explanation gradually. IG provides a continuous decomposition, but explains the model through a trajectory of individual input points, which may pass through regions of input space that are not semantically meaningful and reflect only pointwise evaluations of the model (Sturmfels et al., 2020). Several variants modify this trajectory, for example by replacing the straight-line path with a scale-space path (Xu et al., 2020) or with an adaptive path chosen to reduce noisy attributions (Kapishnikov et al., 2021). These methods improve the choice of path, but the model is still evaluated pointwise along a trajectory of inputs.

Crucially, completeness is a property of the path integral, not of the particular space in which the path is drawn. It only requires a trajectory connecting a reference state to the explained prediction, along which changes in the model response can be aggregated and then decomposed into feature contributions. This opens the possibility of taking the path integral in a lifted space of structured probes rather than in the original input space. In this work, we instantiate this idea by defining continuous paths through distributional probes of the model: transformations that aggregate the model's behavior over structured perturbations of the input. In practice, these probes are implemented as expectations over parameterized distributions, enabling efficient evaluation via Monte Carlo

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

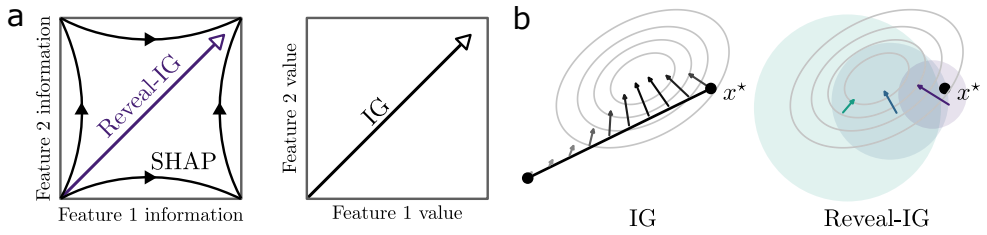

Figure 1: **(a)** SHAP reveals feature information in discrete steps, averaging contributions along all such paths. We introduce Reveal-IG, a method that gradually reveals feature information, traversing a continuous path analogous to the path of Integrated Gradients (IG) in the space of feature values. **(b)** IG evaluates along a single path from a baseline to the point being explained, $x^\star$. Reveal-IG integrates over a sequence of distributions that progressively concentrate around the input, capturing contributions across multiple locations and scales.

sampling and reparameterization. We call the resulting method Reveal-IG, reflecting the gradual revelation of information about the datapoint being explained within an IG-style attribution objective. The result is a continuous analogue of SHAP-style feature revelation, replacing discrete feature coalitions with a differentiable reveal path from broad to concentrated probes around the input.

This lifted view gives a common language for attribution across domains. In each case, attribution is defined by a path through structured probe distributions—from a broad, high-uncertainty reference to a probe concentrated on the observed example—that specifies not only which feature values are shown to the model but the resolution at which they are queried. The probe family is the only domain-specific choice: it adapts to the structure of images or tabular data while the path-integral construction stays fixed. We instantiate Reveal-IG for image classification and tabular regression, where it yields interpretable attribution trajectories and competitive-to-strong performance against gradient-, perturbation-, and Shapley-style baselines.

2 Method

Consider a model f and an input $x^\star$ whose prediction we wish to explain. We index the features of $x^\star$ by i, and seek feature-wise attributions that decompose the change in model response between a reference state and the explained input.

Path-based attribution methods formalize such decompositions by applying the fundamental theorem of calculus along a path. Let θ denote the coordinates being varied along the path, and let $G(\theta)$ be the scalar model quantity whose change we wish to decompose. For a path $\mathcal{S}$ from θ^{start} to θ^{end},

$$G(\theta^{\text{end}}) - G(\theta^{\text{start}}) = \int_{\mathcal{S}} \nabla_\theta G(\theta) \cdot d\theta = \sum_i \int_{\mathcal{S}} \frac{\partial G}{\partial \theta_i}\, d\theta_i = \sum_i a_i. \tag{1}$$

The terms a_i define the path-dependent, per-feature attributions, where θ_i denotes the block of path coordinates associated with feature i and the sum runs over features.

In standard input-space formulations, the path coordinates are the input values themselves: $\theta = x$ and $G(\theta) = f(x)$, so each feature's block θ_i is the single scalar x_i, and the attribution decomposes the change in model output as the input moves from a baseline x^{start} to the example $x^{\text{end}} = x^\star$. In our formulation, by contrast, θ parameterizes a family of probe distributions $q_\theta(z \mid x^\star)$, and

$$G(\theta; x^\star) = \mathbb{E}_{z \sim q_\theta(\cdot \mid x^\star)}[f(z)]. \tag{2}$$

The path-integral decomposition is applied to a path through the space of distributions over inputs, rather than through input space directly. Attribution is obtained by integrating gradients of this expected response along the path of probe distributions:

$$\Delta G = \int_{\mathcal{S}} \nabla_\theta \mathbb{E}_{z \sim q_\theta(z \mid x^\star)}[f(z)] \cdot d\theta. \tag{3}$$

For practical implementation, we assume a factorized form for the probe distribution,

$$q_\theta(z \mid x^\star) = \prod_i q_{\theta_i}(z_i \mid x_i^\star). \tag{4}$$

When a feature is represented by multiple probe parameters, its attribution is the sum of the corresponding line-integral terms. For Gaussian probes, each feature is represented by a parameter block $\theta_i = (\mu_i, \ell_i)$, where $\ell_i = \log \sigma_i^2$, and its attribution is the line-integral contribution of the block:

$$a_i = \int_0^1 \left[\frac{\partial G}{\partial \mu_i} \frac{d\mu_i}{dt} + \frac{\partial G}{\partial \ell_i} \frac{d\ell_i}{dt} \right] dt. \tag{5}$$

Thus the reported feature attribution is the sum of the mean-path and variance-path contributions; this sum is required by completeness and is not an additional weighting heuristic.

Under this parameterization, the key design choice is the trajectory of feature-wise probe distributions. We define a path that begins at a broad reference probe and ends at a probe concentrated around the example $x^\star$. The framework does not require it, but in both of our instantiations the reference is tied to the data distribution: it is the factorized marginal of p_{data}, realized as a moment-matched Gaussian for images and as the empirical marginal for tabular features. This makes the start of the path a maximum-uncertainty state under the factorized approximation, and grounds the reference in the population. The endpoint is a probe concentrated on $x^\star$—a product of finite-variance distributions —corresponding to near-full specification of the explained input.

Image data. For image data, we define a local probe distribution around the explained image $x^\star$ by modeling each color channel of each pixel as an independent Gaussian variable. Let z denote an image sampled from the probe. We parameterize

$$q_\theta(z \mid x^\star) = \prod_i \mathcal{N}(z_i; \mu_i, \sigma_i^2), \tag{6}$$

where i indexes pixel-channel components. The path begins at a reference probe with $\mu_i = 0$ and $\sigma_i = 1$, corresponding to the normalized image-space prior, and ends at a probe centered on the observed image, with $\mu_i = x_i^\star$ and $\sigma_i = \sigma_{\mathrm{final}}$, with parameters linearly interpolated along the path.

Tabular data. For tabular data, we do not assume a dense input space from which meaningful perturbations can be sampled freely. Instead, we define probe distributions over empirical feature values observed in the training data. For each feature, we construct a distribution that transitions from the empirical marginal distribution to a concentrated distribution around the observed value $x_i^\star$.

We parameterize each feature-wise probe using a temperature-controlled similarity kernel,

$$q_{\tau_i}(z_i \mid x_i^\star) \propto p_{\mathrm{data}}(z_i) \exp(s_i(z_i, x_i^\star)/\tau_i)\,, \tag{7}$$

where z_i denotes a candidate value for feature i, $p_{\mathrm{data}}(z_i)$ is the empirical marginal distribution, and $s_i(\cdot, \cdot)$ is a feature-specific similarity function. In our experiments, all features are treated as numeric, and we use negative squared distance as the similarity function.

The path is defined by decreasing the temperatures τ_i, which progressively concentrates each feature-wise probe around the observed value $x_i^\star$. At high temperature, each probe approaches the empirical marginal distribution; at low temperature, it approaches a near-deterministic distribution at the observed feature value. Expectations under the resulting product probe are estimated using empirical feature values from the training data under a factorized, interventional approach. That is, feature values are recombined across datapoints, as in interventional SHAP (Lundberg & Lee, 2017; Covert et al., 2020), with feature-wise importance weights determined by Eqn. 7. We further allow feature-wise temperature schedules parameterized by fractional entropy reduction, so that features with different marginal entropies reveal information at matched rates and reach near-deterministic endpoint distributions simultaneously. This prevents low-entropy features from being effectively revealed earlier than high-entropy features simply because their marginal distributions are already concentrated.

Practical evaluation of gradients. The gradient in Eqn. 3 is taken with respect to the probe parameters θ, not directly with respect to the input values. When the probe distribution admits a differentiable reparameterization, we write samples as $z = g_\theta(\epsilon; x^\star)$ for fixed noise ϵ, and estimate gradients by differentiating $f(g_\theta(\epsilon; x^\star))$ with respect to θ using automatic differentiation (Kingma & Welling, 2014). For Gaussian image probes, this corresponds to sampling $z_i = \mu_i + \sigma_i \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 1)$. For tabular data, where probes are defined over empirical feature values, we estimate expectations using training-set feature values with weights induced by the feature-wise

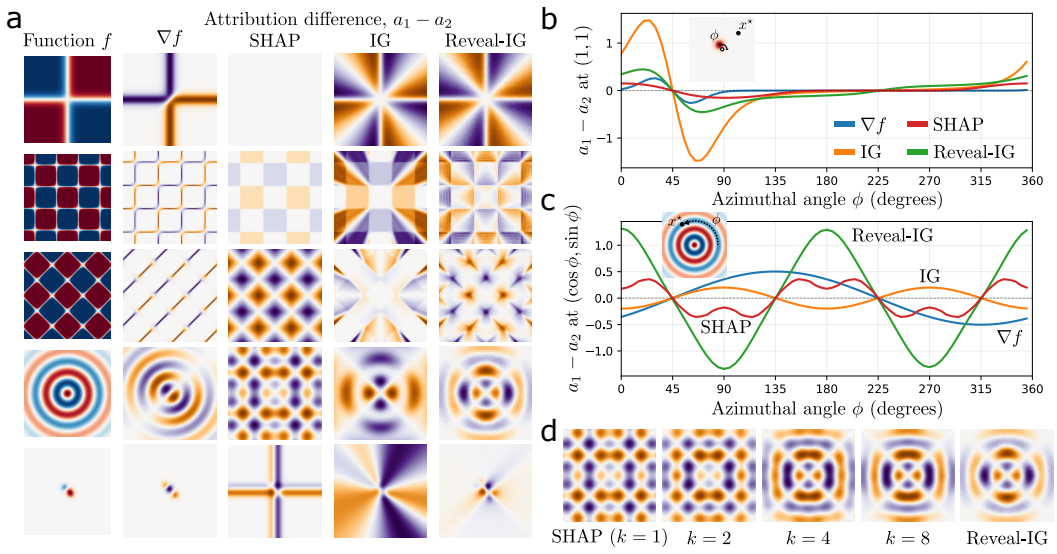

Figure 2: Attribution fields in two dimensions. **(a)** For a selection of functions (left column), we evaluate attributions according to gradients, SHAP, IG, and Reveal-IG. The difference of attribution components, $a_1 - a_2$ is displayed as a heatmap. **(b)** By systematically varying the function, we obtain response curves for the attribution methods. A bump in the function's output is rotated around the origin and the attribution at $(1, 1)$ is recorded. **(c)** For the radial rings in panel **a**, we extract the attribution difference around a circle of radius one. **(d)** By averaging over paths with k steps of feature information reveal, the radial rings' attribution interpolates from SHAP toward Reveal-IG.

probe distributions. The empirical feature values are fixed and only the weights depend on the probe parameters; differentiating the weighted expectation therefore estimates the gradient of the expected model response with respect to the probe parameters. This estimator requires only model evaluations, making the method applicable to tree ensembles and other non-differentiable predictors.

3 Experiments

3.1 Synthetic diagnostics: attribution fields in two dimensions

We begin with scalar functions on a two-dimensional input space (Fig. 2a), where both the input-output landscape and the resulting attribution field can be visualized directly. These examples are not benchmarks; they are controlled diagnostics for how attribution methods respond to structure away from the point being explained. For complete two-feature methods, the attribution vector (a_1, a_2) satisfies $a_1 + a_2 = f(x^{\text{end}}) - f(x^{\text{start}})$. We visualize the remaining degree of freedom, $a_1 - a_2$, which shows how each method allocates the same total change across the two coordinates. For gradients, which have two free components, we likewise plot their difference.

The first rows of Fig. 2a show agreement in simple cases that diminishes as the landscape becomes more structured. For the XOR function, IG and Reveal-IG produce closely related attribution fields. For the checkerboard functions, both methods highlight sharp changes near corners and edges, though with different emphasis induced by their respective paths. SHAP produces more axis-aligned structure, particularly visible in the last two rows, reflecting that its explanations are built from coordinate-wise feature reveal operations rather than continuous motion through the input or distribution space.

In the final row, IG produces "shadows": attribution at a point can be strongly influenced by features encountered along the straight-line path from the baseline, even when those features are far from the point being explained. Here, localized structure in the function's response near the origin casts long imprints through the attribution field. Reveal-IG shows a much less pronounced effect because the early, large-length scale probes average model responses over a broad region of the landscape and dilute the influence of distant structure before the distributions concentrate near $x^\star$.

We isolate this shadow effect using a controlled function sweep in Fig. 2b. A localized bump is moved around a circle centered at the origin, while the attribution at a fixed point $x^\star$ is measured. Although

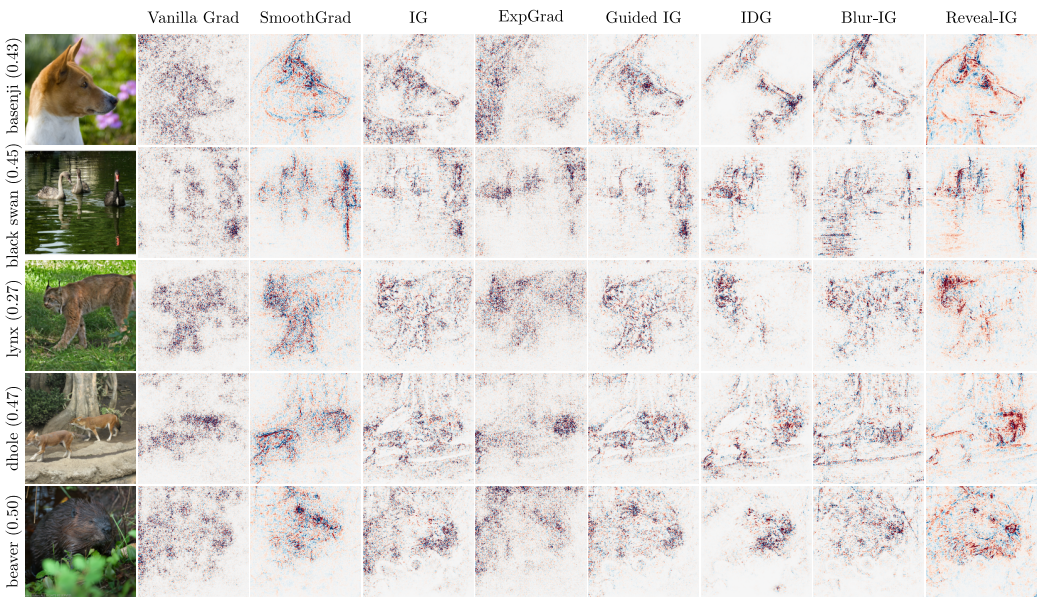

Figure 3: For a random selection of ImageNet validation images, we show signed saliency maps for the methods evaluated in Tab. 1. The left column shows the input image together with the ResNet-50 top-1 prediction and confidence. Red is positive attribution and blue is negative.

the bump remains far from $x^\star$, IG changes substantially when the bump crosses the straight-line path from the baseline to $x^\star$. Reveal-IG shows a weaker, though nonzero version of the same effect.

Finally, Fig. 2c examines angular structure in attributions when the underlying function is radially symmetric. Gradients show first-harmonic dependence while the complete methods are dominated by the second harmonic. SHAP contains additional higher-order angular components. We test whether the severity of SHAP's revelation operation—a single all-or-nothing step—is responsible for the axis-aligned structure by subdividing each feature's reveal into k equal-uncertainty steps and averaging over the possible orderings. $k = 1$ is equivalent to SHAP, and $k \to \infty$ yields a path-ensembling attribution that approaches the continuous reveal path underlying Reveal-IG. As k increases, higher-order angular components diminish toward the nonzero value retained by Reveal-IG itself (Fig. 2d). The severity of the reveal step thus has a measurable effect on attributions, independent of whether that effect is desirable; across functions, intermediate k produces attribution maps that interpolate between the SHAP and Reveal-IG extremes (Appx. C).

3.2 Image Classification

Data and model. We evaluate image attributions on 1,000 images sampled from the ImageNet-1K validation set (Deng et al., 2009; Russakovsky et al., 2015), using a pre-trained ResNet-50 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2021). Inputs are resized to 256 px, center-cropped to 224×224, and normalized with ImageNet statistics. Unless otherwise stated, the attribution target for image experiments is the ground-truth class if its predicted probability exceeds 0.05, and the model's top-1 predicted class otherwise.

Attribution methods. We compare Reveal-IG against gradient- and path-based baselines: Vanilla Gradient (Simonyan et al., 2013), SmoothGrad (Smilkov et al., 2017), IG (Sundararajan et al., 2017), Guided IG (Kapishnikov et al., 2021), Integrated Decision Gradients (IDG) (Walker et al., 2024), Blur-IG (Xu et al., 2020), and Expected Gradients (Erion et al., 2021). Reveal-IG is evaluated with a terminal scale $\sigma = 0.25$, 50 integration steps and 10 Monte Carlo samples per step; we explore other endpoints and an adaptive-σ variant in Appx. B. All methods return signed channel-level attributions, which are collapsed by summing over channels. Full hyperparameters are given in Appx. A.

Table 1: Image attribution across architectures (95% CIs). Insertion/deletion and Sensitivity-n are evaluated on $n = 1{,}000$ ImageNet validation images, sampled as one image per class. Insertion/deletion uses a Gaussian-blur baseline following the RISE protocol; *Ins-Del|abs*: pixels ranked by $|\mathrm{attr}|$; *Ins-Del|relu*: pixels ranked by $\mathrm{relu}(\mathrm{attr})$ (positive evidence only). RMA is evaluated on $n = 879$ ImageNet-S validation images with ground-truth segmentation masks for ResNet-50 ($n = 914$ for ViT due to higher accuracy). Shading indicates best, underline second best.

	Method	Ins-Del|abs ↑	Ins-Del|relu ↑	RMA ↑	Sens.-n ↑
ResNet-50	Vanilla Grad	0.0206 ± 0.0056	0.0042 ± 0.0037	0.4884 ± 0.0184	-0.0036 ± 0.0039
	SmoothGrad	0.1652 ± 0.0076	0.0195 ± 0.0048	0.5672 ± 0.0163	-0.0014 ± 0.0040
	IG	0.0798 ± 0.0063	0.0760 ± 0.0048	0.5177 ± 0.0179	0.0203 ± 0.0039
	ExpGrad	0.0849 ± 0.0063	0.0858 ± 0.0047	0.5047 ± 0.0170	0.0137 ± 0.0039
	Guided IG	0.1329 ± 0.0069	0.1043 ± 0.0053	0.5787 ± 0.0164	0.0358 ± 0.0038
	IDG	0.1772 ± 0.0078	0.1003 ± 0.0055	0.6139 ± 0.0161	0.0210 ± 0.0041
	Blur-IG	0.1987 ± 0.0078	0.0847 ± 0.0044	0.5663 ± 0.0168	0.0032 ± 0.0039
	Reveal-IG	0.1960 ± 0.0081	0.2343 ± 0.0091	0.5956 ± 0.0153	0.0802 ± 0.0043
ViT-B/16	Vanilla Grad	0.0725 ± 0.0119	0.0213 ± 0.0073	0.4205 ± 0.0173	-0.0025 ± 0.0040
	SmoothGrad	0.3263 ± 0.0140	0.0526 ± 0.0092	0.5283 ± 0.0157	-0.0059 ± 0.0040
	IG	0.1641 ± 0.0124	0.2090 ± 0.0105	0.4709 ± 0.0163	0.0226 ± 0.0040
	ExpGrad	0.1493 ± 0.0113	0.1980 ± 0.0092	0.4442 ± 0.0164	0.0138 ± 0.0039
	Guided IG	0.2249 ± 0.0127	0.2426 ± 0.0110	0.5139 ± 0.0154	0.0335 ± 0.0039
	IDG	0.2800 ± 0.0133	0.2604 ± 0.0117	0.5359 ± 0.0157	0.0246 ± 0.0041
	Blur-IG	0.3658 ± 0.0130	0.2689 ± 0.0103	0.5238 ± 0.0154	0.0113 ± 0.0040
	Reveal-IG	0.3774 ± 0.0140	0.4951 ± 0.0146	0.5545 ± 0.0147	0.0808 ± 0.0043

Evaluation protocol. We use three complementary families of metrics. First, insertion and deletion measure whether highly ranked pixels have large effects on the target logit when pixels are progressively restored from, or removed to, the same blurred baseline (Petsiuk et al., 2018) (blur $\sigma = 10$). We report both magnitude-based ranking, $|\mathrm{attr}|$, which treats positive and negative evidence as saliency, and positive-evidence ranking, $\mathrm{ReLU}(\mathrm{attr}, 0)$, which asks specifically whether the method identifies features supporting the target class. Second, Sensitivity-n evaluates whether the signed attribution assigned to a random subset of pixels predicts the signed change in target logit when that subset is removed (Ancona et al., 2018). Third, object-concentration metrics measure how much attribution mass falls within externally defined object regions. We report Relevance Mass Accuracy (RMA) (Arras et al., 2022), the fraction of positive attribution mass inside the object mask, using ImageNet-S ground-truth segmentation masks (Gao et al., 2022). RMA is evaluated only on examples for which the foreground mask covers more than 2% of the image.

Image attribution results. The quantitative results are summarized in Tab. 1, with representative saliency maps shown in Fig. 3. Across both architectures, Reveal-IG is best or second-best on every reported metric. Its largest gains appear on evaluations that use attribution sign: Reveal-IG leads both positive-evidence insertion/deletion and Sensitivity-n by a wide margin.

The closest competitors emphasize different strengths. Blur-IG is strongest for magnitude-ranked insertion/deletion, showing that a coarse-to-fine image path can be highly effective for unsigned saliency. IDG is strongest for ResNet-50 localization, indicating that its path can produce spatially concentrated maps. Both Blur-IG and Reveal-IG replace direct point-to-point interpolation with paths organized around progressive information revelation, but they smooth different objects: Blur-IG smooths the input space by progressively restoring spatial detail, whereas Reveal-IG smooths the output response by integrating predictions over distributions that progressively concentrate around the observed image.

Attribution along the path. To examine how different path methods distribute attribution across the trajectory, Fig. 4 plots the fraction of total absolute attribution accumulation as a function of path progress, averaged over 100 validation images. The normalization forces each curve to sum to one,

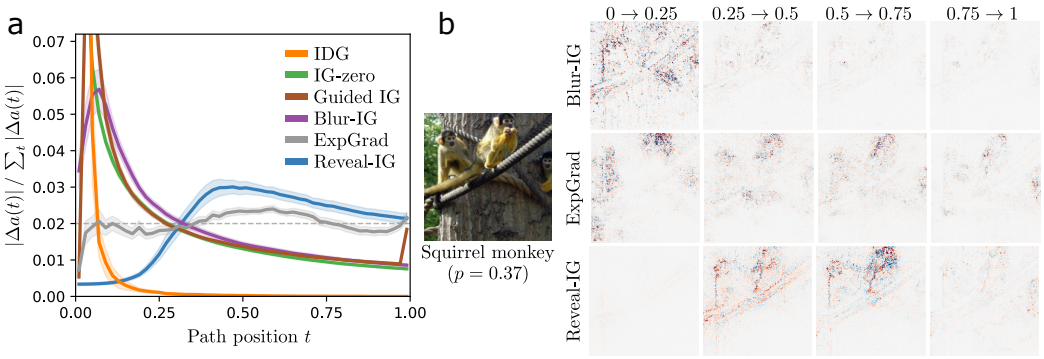

Figure 4: Attribution over the path. **(a)** We show the fractional rate of change of the attributions, $|\Delta a(t)| / \sum_t |\Delta a(t)|$, averaged over 100 images. **(b)** Contribution to final attribution maps across segments of the trajectory for a randomly selected image.

so the plot compares where along the path attribution is assigned rather than the total magnitude or completeness of the attribution. The methods form three natural groups. IDG, IG, Guided IG, and Blur-IG assign the majority of their attribution early in the trajectory, while the path is still far from the input. Expected Gradients allocates attribution nearly uniformly along the trajectory, while Reveal-IG assigns most attribution late in the path. This late accumulation is more consistent with the reveal-based interpretation: attribution is assigned as the probe becomes informative about the explained input, rather than being dominated by regions near the baseline. It also echoes the synthetic diagnostics, where early path dependence produced shadow artifacts (Fig. 2); delaying attribution until the probe concentrates near the explained image reduces the opportunity for distant or baseline-adjacent structure to dominate the explanation.

3.3 Tabular regression

Data and model. We evaluate tabular attributions on two regression datasets: the UCI Bike Sharing dataset (Fanaee-T & Gama, 2014) and Wine Quality (the combined red and white UCI Wine Quality dataset) (Cortez et al., 2009). For each dataset, we train a two-hidden-layer MLP regressor with 64 hidden units per layer and ReLU activations, and compute attributions with respect to the scalar model prediction. All features and regression targets are standardized using training-set statistics. Unless otherwise stated, tabular results are evaluated on 1,000 held-out test examples.

Attribution methods. We compare Reveal-IG against gradient-, path-, perturbation-, and local-surrogate baselines: Vanilla Gradient, SmoothGrad, IG, Expected Gradients, KernelSHAP (Lundberg & Lee, 2017), LIME (Ribeiro et al., 2016), and feature ablation. For Reveal-IG, feature-wise probe distributions are defined over empirical marginal feature values and annealed toward the observed feature values along a normalized entropy path. The path begins near the empirical marginal distribution, at 0.99 of each feature's marginal entropy, and ends at a concentrated probe with 0.05 of that entropy. For IG, the reference point is the training-set mean after standardization, i.e., zero. Feature ablation replaces each feature with samples from its empirical marginal distribution and averages the resulting prediction change. Full hyperparameters are given in Appx. A.

Evaluation protocol. Sufficiency and comprehensiveness measure whether the top-ranked features are respectively adequate to preserve the prediction and necessary for the original prediction, following DeYoung et al. (2020). We rank features by attribution magnitude and use the top 20% of features. For sufficiency, we retain only these features and replace the rest using the training-set mean; lower values indicate that the selected features preserve the original prediction. For comprehensiveness, we replace the top-ranked features and measure the absolute change in prediction; higher values indicate that the selected features capture information the model relies on. Both metrics are computed as absolute differences in the scalar model output. Sensitivity-max measures local stability of the attribution map under small input perturbations (Yeh et al., 2019). We estimate Sensitivity-max by sampling 50 random perturbation directions on the ℓ_2 sphere of radius $r = 0.1$ in standardized feature space, and report the largest normalized attribution change, $\max_m \frac{\|\Phi(x+\delta_m) - \Phi(x)\|_2}{r}$.

Table 2: Tabular attribution evaluation (95% CIs). Shading indicates best, underline second best.

Data	Method	Ins-Dir ↑	Suff. ↓	Comp. ↑	Sens.-max ↓
Bike	Vanilla Grad	0.098 ± 0.049	0.500 ± 0.034	0.675 ± 0.041	30.119 ± 2.638
	SmoothGrad	0.131 ± 0.049	0.496 ± 0.033	0.643 ± 0.040	13.063 ± 0.747
	IG	0.819 ± 0.044	0.514 ± 0.036	0.641 ± 0.040	6.441 ± 0.369
	ExpGrad	0.636 ± 0.050	0.486 ± 0.032	0.706 ± 0.041	38.358 ± 0.860
	KernelSHAP	1.010 ± 0.044	0.378 ± 0.024	0.848 ± 0.044	4.122 ± 0.093
	LIME	0.082 ± 0.039	0.570 ± 0.040	0.543 ± 0.030	2.012 ± 0.021
	Feat. Ablation	0.881 ± 0.043	0.401 ± 0.025	0.801 ± 0.042	5.125 ± 0.185
	Reveal-IG	0.973 ± 0.046	0.367 ± 0.024	0.817 ± 0.043	2.474 ± 0.137
Wine	Vanilla Grad	-0.587 ± 0.065	0.748 ± 0.045	0.761 ± 0.050	21.175 ± 0.687
	SmoothGrad	-0.660 ± 0.069	0.699 ± 0.044	0.794 ± 0.050	7.555 ± 0.161
	IG	1.371 ± 0.065	0.704 ± 0.043	0.894 ± 0.057	5.493 ± 0.233
	Expected Grad	1.116 ± 0.064	0.726 ± 0.048	0.852 ± 0.058	18.808 ± 0.310
	KernelSHAP	1.505 ± 0.068	0.638 ± 0.042	1.008 ± 0.064	3.932 ± 0.059
	LIME	-0.657 ± 0.056	0.773 ± 0.050	0.672 ± 0.046	2.968 ± 0.032
	Feat. Ablation	1.220 ± 0.062	0.681 ± 0.045	1.137 ± 0.064	6.705 ± 0.125
	Reveal-IG	1.438 ± 0.065	0.665 ± 0.044	0.928 ± 0.061	2.188 ± 0.102

The directional insertion diagnostic adapts insertion/deletion to signed regression attributions. We reveal features in descending attribution order and in ascending attribution order, compute the AUC of each prediction curve as a function of the fraction of features revealed, and report their difference, $\mathrm{AUC}_{\mathrm{desc}} - \mathrm{AUC}_{\mathrm{asc}}$.

Tabular attribution results. The tabular results in Table 2 are more mixed than the image results, but show that the distributional-path idea remains competitive outside image geometry. Reveal-IG is consistently among the strongest methods on directional insertion and sufficiency, and achieves the best or second-best sensitivity-max score on both datasets. This stability advantage is especially notable relative to the other gradient and path-based methods: vanilla gradients and SmoothGrad often fail under directional insertion, while Expected Gradients can be highly sensitive despite reasonable sufficiency or comprehensiveness scores. Compared to input-space IG, the distributional reveal path is more natural for tabular data, which present no obvious spatial or geometric interpolation between a baseline and an example. Additionally, whereas IG requires differentiability along a continuous input-space path, tabular Reveal-IG moves probability mass over empirical feature values.

KernelSHAP is strong with regard to perturbation metrics, as its explanations are themselves defined through feature replacement. This makes sufficiency and comprehensiveness a favorable evaluation setting for perturbation-native methods. Reveal-IG instead occupies a different tradeoff: it approaches the perturbation baselines on removal-style metrics while producing substantially more stable attributions.

4 Related work

Path attribution and completeness. Integrated Gradients (IG) attributes a prediction by integrating gradients along a path from a reference input to the example being explained, yielding a complete decomposition of the change in model output (Sundararajan et al., 2017). This path-integral view is closely related to Aumann–Shapley cost sharing (Aumann & Shapley, 1974), and has motivated a family of methods that preserve completeness while modifying the choice of reference point or path. Several works study the sensitivity of IG to its baseline and propose alternative baseline choices (Sturmfels et al., 2020). Others modify the path itself, for example by replacing the straight-line path with a scale-space path (Xu et al., 2020; Kim et al., 2026), an adaptive path chosen to avoid high-gradient regions (Kapishnikov et al., 2021), or a path designed to emphasize decision-relevant regions (Walker et al., 2024). These methods demonstrate that the path is not merely an implementation detail: it determines which regions of the model's input-output landscape are incorporated into the explanation.

Stochastic and distributional path variants. A related line of work incorporates stochasticity into attribution by averaging over baselines, paths, or perturbations. Expected Gradients averages IG over reference samples drawn from the data distribution, replacing a single baseline with a distribution of baselines (Erion et al., 2021). Other methods introduce stochasticity through sampled paths, baseline mixtures, diffusion-based trajectories, or iterative attribution procedures (Lundstrom et al., 2022; Barkan et al., 2023b;a; Lei et al., 2024; Elisha et al., 2024). These approaches introduce distributions into the attribution procedure, but the model is still evaluated pointwise along sampled input-space trajectories, and the final attribution is obtained by averaging over such trajectories. Reveal-IG differs by making the object of integration itself an expected model response under a parameterized probe distribution. The path therefore lives in a lifted space of distributions rather than only in the original input space.

Shapley values and information revelation. Shapley-value methods average each feature's marginal contribution over coalitions of observed features (Lundberg & Lee, 2017). From a path perspective, this corresponds to averaging over discrete feature-reveal paths, where features move one at a time from unobserved to observed. Reveal-IG keeps this information-revelation interpretation, but replaces discrete feature inclusion with a continuous path of probe distributions. Related work on additive feature importance clarifies the connection between Shapley values, feature removal, and marginal contributions (Covert et al., 2020).

Information-restriction attribution. Information bottlenecks for attribution explain a fixed model by injecting noise into intermediate features and identifying the information that must be preserved to maintain the prediction (Schulz et al., 2020; Jiang et al., 2020; Hong et al., 2025). Related information-constrained approaches jointly optimize feature information and prediction, exposing which variables support task-relevant structure in the data (Murphy & Bassett, 2023; 2024). Reveal-IG uses a related principle in a complete path-attribution setting: the model is fixed, and attributions integrate changes in expected model response as feature information is revealed continuously from a low-information reference to a high-information endpoint.

5 Discussion

Completeness is both the strength and the burden of path-based attribution. A complete path method must explain the total change between a reference state and the input, but it does so by aggregating all changes encountered along the chosen trajectory. The central design problem is therefore not only choosing a neutral baseline, but choosing intermediate probes that are relevant to the explanation. Reveal-IG addresses this by moving path design from input space to distribution space: attribution proceeds by controlled information revelation around the target input.

This distributional view reframes path attribution as a smoothing problem. Blur-IG smooths the input directly, moving from coarse to fine spatial structure; Reveal-IG instead smooths the model response by evaluating it under probe distributions whose uncertainty decreases along the path. The attribution therefore depends not only on where the path goes, but on the scale at which the model is queried. This makes the path a choice of explanatory resolution: early probes ask how the model responds to coarse or uncertain information, while later probes assign credit as the input becomes more specified.

The two-dimensional diagnostics suggest a broader geometric view of attribution. After factoring out the completeness constraint, attribution fields contain degrees of freedom that reveal method-specific artifacts, such as IG path shadows or SHAP's axis-aligned reveal structure. These controlled diagnostics complement high-dimensional benchmarks, where such artifacts are harder to isolate.

Limitations. Reveal-IG shifts the main design choice from a baseline input to a family of probe distributions and a path through them. Reveal-IG is also more expensive than single-gradient methods because it estimates gradients of an expected response over multiple path steps and Monte Carlo samples. For ResNet-50, IG with 50 integration steps takes ~ 165 milliseconds, whereas Reveal-IG's extra 10 Monte Carlo evaluations per integration step brings the time per attribution to ~ 1.1 seconds, an increase of about $7\times$. In tabular settings, the penalty is far greater: on Bikeshare, it's closer to $1,000\times$ (IG, KernelSHAP, and LIME are 1-3 milliseconds per attribution, and Reveal-IG takes about 1.5 seconds). Finally, completeness holds for the lifted function being integrated: finite-variance endpoints decompose the change in expected response under the endpoint distribution, rather than exactly decomposing the deterministic model output.

AI USE STATEMENT

We used LLM-based programming assistance to help implement and edit experimental code. The LLM was not used as a component of the proposed method, was not used to generate experimental results autonomously, and did not determine the experimental design, analysis, or conclusions. LLMs were also used for writing support such as copyediting, wording suggestions, LaTeX formatting, and brainstorming. All code, figures, numerical results, references, and scientific claims were reviewed and verified by the authors.

REFERENCES

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Sy21R9JAW.

Leila Arras, Ahmed Osman, and Wojciech Samek. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022.

R.J. Aumann and L.S. Shapley. *Values of Non-atomic Games*. Princeton Legacy Library. Princeton University Press, 1974. ISBN 9780691081038. URL https://books.google.com/books?id=SIvUAQAACAAJ.

Oren Barkan, Yehonatan Elisha, Yuval Asher, Amit Eshel, and Noam Koenigstein. Visual explanations via iterated integrated attributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023a.

Oren Barkan, Yehonatan Elisha, Jonathan Weill, Yuval Asher, Amit Eshel, and Noam Koenigstein. Stochastic integrated explanations for vision models. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 938–943. IEEE, 2023b.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4): 547–553, 2009.

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in neural information processing systems*, 33:17212–17223, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4443–4458, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Yehonatan Elisha, Oren Barkan, and Noam Koenigstein. Probabilistic path integration with mixture of baseline distributions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 570–580, 2024.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631, 2021.

Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2):113–127, 2014.

Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jung-Ho Hong, Ho-Joong Kim, Kyu-Sung Jeon, and Seong-Whan Lee. Comprehensive information bottleneck for unveiling universal attribution to interpret vision transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25166–25175, 2025. doi: 10.1109/CVPR52734.2025.02343.

Zhiying Jiang, Raphael Tang, Ji Xin, and Jimmy Lin. Inserting Information Bottlenecks for Attribution in Transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3850–3857, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.findings-emnlp.343.

Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: an adaptive path method for removing noise. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5048–5056, 2021. doi: 10.1109/CVPR46437.2021.00501.

Soyeon Kim, Seongwoo Lim, Kyowoon Lee, and Jaesik Choi. 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. URL http://arxiv.org/abs/2605.19607.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020.

Yiming Lei, Zilong Li, Junping Zhang, and Hongming Shan. Denoising diffusion path: Attribution noise reduction with an auxiliary diffusion model. *Advances in Neural Information Processing Systems*, 37:54003–54025, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Daniel D Lundstrom, Tianjian Huang, and Meisam Razaviyayn. A rigorous study of integrated gradients method and extensions to internal neuron attributions. In *International Conference on Machine Learning*, pp. 14485–14508. PMLR, 2022.

Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2022.

Kieran A Murphy and Dani S. Bassett. Interpretability with full complexity by constraining feature information. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=R_OL5mLhsv.

Kieran A Murphy and Dani S Bassett. Information decomposition in complex systems via machine learning. *Proceedings of the National Academy of Sciences*, 121(13):e2312988121, 2024.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos 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.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1xWh1rYwB.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034*, 2013.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020. doi: 10.23915/distill.00022. https://distill.pub/2020/attribution-baselines.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Chase Walker, Sumit Jha, Kenny Chen, and Rickard Ewetz. Integrated decision gradients: Compute your attributions where the model makes its decision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5289–5297, 2024.

Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

A Implementation Details

Code is attached to this submission in the supplemental material. Hyperparameters are listed in Tables 3&4.

A.1 Image classification.

Models. We used `IMAGENET1K_V2` weights for ResNet-50 (He et al., 2016) and `IMAGENET1K_SWAG_LINEAR_V1` weights for ViT-B/16 (Dosovitskiy et al., 2021).

Baselines. All baselines use publicly available implementations with their default configurations except where noted. Vanilla Grad, SmoothGrad, IG, and Expected Gradients are implemented using Captum (Kokhlikyan et al., 2020). IDG uses the authors' public implementation,[1] and Blur-IG and Guided-IG use the PAIR saliency implementation.[2][3] The one exception to default settings is Blur-IG: we use maximum blur scale $\sigma = 10$ rather than the PAIR saliency default $\sigma = 50$, which performed worse under all reported metrics in our setting.

ImageNet-S Filtering for RMA. We use ground-truth segmentation masks from ImageNet-S (Gao et al., 2022). Evaluation is restricted to images (1) for which the model assigns greater than p=0.05 to the ground truth class, and (2) have a foreground mask covering more than 2% of the image area.

[1] https://github.com/chasewalker26/Integrated-Decision-Gradients/tree/main/IDG

[2] https://github.com/PAIR-code/saliency/blob/master/saliency/core/blur_ig.py

[3] https://github.com/PAIR-code/saliency/blob/master/saliency/core/guided_ig.py

Adaptive σ selection. For Reveal-IG (adaptive), the endpoint noise scale σ_{final} is determined per image by binary search over $[1/256, 1.0]$ to find the largest σ satisfying $\mathbb{E}_\varepsilon[f(x+\sigma\varepsilon)] \geq 0.95 \cdot f(x)$, where $\varepsilon \sim \mathcal{N}(0, I)$ and the expectation is approximated with 10 samples. This ensures the path endpoint preserves the model's confidence in the target class while allowing image-specific noise tolerance.

A.2 Tabular regression.

Tabular model training details. For all tabular regression experiments, we use the same MLP architecture:

$$\text{Linear}(D, 64) \to \text{ReLU} \to \text{Linear}(64, 64) \to \text{ReLU} \to \text{Linear}(64, 1).$$

Models are trained with Adam using learning rate 10^{-3} and batch size 512. We use a 90/10 train/test split with random shuffling. All features are treated as continuous variables and are standardized using the training-set mean and standard deviation; regression targets are standardized using training-set target statistics. The Bike Sharing Demand model trained for 500 epochs; the Wine Quality model trained for 800 epochs. Training uses mean-squared error loss for all regression datasets.

Attribution. For each evaluation point $x^\star$, we draw a fixed reference pool of 2047 other dataset points, which is used throughout the Reveal-IG attribution calculation. For each feature, the initial entropy is computed from the empirical counts of unique feature values in this pool. The path is parameterized by feature-wise temperatures τ_i, which define categorical distributions over the pool values through their similarity to $x_i^\star$. At each path step, each τ_i is chosen by binary search so that the feature's marginal entropy linearly interpolates between its initial entropy and its endpoint entropy. Holding these temperatures fixed within a path step, each Monte Carlo sample constructs a context point by independently sampling one value per feature from these categorical distributions. The model is then queried on the sampled context point, and the contribution to the path gradient is obtained using the closed-form derivative of the categorical probabilities with respect to τ_i; no derivative of the model with respect to its input features is required.

Hyperparameters are listed in Tab. 4. All baselines use publicly available implementations with their default configurations except where noted.

Table 3: Hyperparameters for image attribution methods.

Method	**Hyperparameters**
Vanilla Gradient	Single forward-backward pass
SmoothGrad	50 noise samples, $\sigma_{sg} = 0.15\times$ input range
IG	50 Gauss-Legendre steps, zero baseline
Expected Gradients	50 samples drawn uniformly from a pool of 100 randomly selected ImageNet training images
Guided IG	200 steps, feature fraction 0.25, max-distance 0.02 (public defaults)
IDG	50 steps, slope-weighted alpha redistribution, zero baseline
Blur-IG	50 steps, max blur σ=10.0
Reveal-IG	50 integration steps, 10 MC samples per step, σ_{final}=0.25

B Ablation experiments

Figure 5 presents ablation studies on the computational requirements to close the completeness gap for Reveal-IG. Note that the completeness gap, defined as $|\sum a_i - (\mathbb{E}_{q_{\text{end}}}[f(x)] - \mathbb{E}_{q_{\text{start}}}[f(x)])|$, compares the sum of attributions across features, a_i, to the Reveal-IG path's start and end points. These will differ in general from the start and end points for other methods: for images, the end point is the expected model response taken over a normal distribution centered at the image and not the image itself as would be the case for IG. For tabular data, the difference is more apparent in the starting point: while SHAP uses the expectation of the model response taken over the product of

Table 4: Hyperparameters for tabular attribution methods.

Method	**Hyperparameters**
Vanilla Gradient	Single forward-backward pass
SmoothGrad	50 noise samples, $\sigma_{sg} = 0.15\times$ per-feature training std
IG	64 Riemann midpoint steps, baseline is training set mean
Expected Gradients	64 random training points via Captum `GradientShap` with `stdevs` $= 0.0$
KernelSHAP	`n_samples` $= 512$, `n_background` $= 50$
LIME	`n_samples` $= 1000$
Feat. Ablation	Captum `FeatureAblation`, baseline is sample from marginal
Reveal-IG	Feature fractional entropy range: $0.99 \rightarrow 0.05$, 40 integration steps, 2048 pool size, 40 MC samples per step

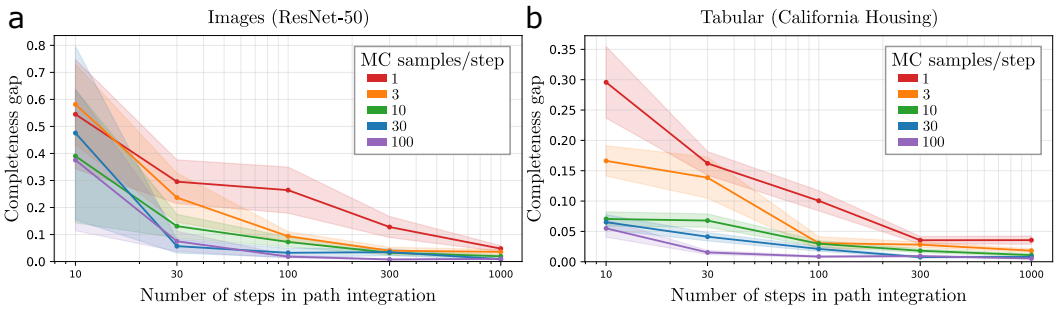

Figure 5: **Reveal-IG completeness convergence.** The completeness gap, measured as $|\sum a_i - (\mathbb{E}_{q_{\text{end}}}[f(x)] - \mathbb{E}_{q_{\text{start}}}[f(x)])|$, with a_i the attribution for component i, for **(a)** ResNet-50 on eight ImageNet samples and **(b)** an MLP trained on the California Housing dataset, evaluated for 20 samples. Shaded regions display standard error.

feature marginal distributions as the baseline for all datapoints, Reveal-IG uses a broad distribution whose entropy per feature is $0.99\times$ the feature's marginal entropy, resulting in a slightly different baseline value per datapoint.

B.1 Sigma variation

We compare different σ endpoints for Reveal-IG in Tab. 5 and Fig. 6. The motivation behind the adaptive endpoint is to identify the noise scale at which the model is sensitive to the particular class; σ is calculated via binary search over $[1/256, 1.0]$ to satisfy $\mathbb{E}[f(x+\sigma\varepsilon)] \geq 0.95 \cdot f(x)$. We find that the performance across metrics is best with a fixed endpoint $\sigma = 0.25$. Qualitatively, the attribution maps lose focus when the σ endpoint is smaller than around σ=1/8.

C Extended synthetic attribution results

In Fig. 7, we evaluate SHAP, Reveal-IG, and the SHAP analogue where information is revealed in k steps. Specifically, the functions are defined over the unit square and the data distribution is taken to be uniform over this area. Whereas SHAP would reveal each feature value entirely in its reveal operation (Fig. 7a), the SHAP analogue with k reveal steps gradually shrinks the extent of the uncertainty for a feature during each step. Here we use a linear interpolation from the full extent to zero over the steps to match Reveal-IG. We note that once subdivisions are included, choice over revelation path, and multiplicity of valid attributions with it, emerges that is desirably absent from SHAP.

The attribution fields in Fig. 7b roughly interpolate between those of SHAP and Reveal-IG, controlled by the number of subdivisions k. Whereas the simple `xor` function acquires structure as soon as more

Table 5: Reveal-IG σ endpoint ablation (ResNet-50, 1k images for Ins/Del/Sens-n, ImageNet-S for RMA ($n = 879$)).

Endpoint σ	**Ins-Del|abs ↑**	**Ins-Del|relu ↑**	**RMA ↑**	**Sens.-n PCC ↑**
σ=1/256	0.1040 ± 0.0066	0.1236 ± 0.0060	0.5365 ± 0.0172	0.0350 ± 0.0042
σ=0.0625	0.1696 ± 0.0077	0.1944 ± 0.0082	0.5756 ± 0.0159	0.0701 ± 0.0043
σ=0.125	0.1854 ± 0.0080	0.2171 ± 0.0086	0.5850 ± 0.0156	0.0749 ± 0.0044
σ=0.25	0.1960 ± 0.0081	0.2343 ± 0.0091	0.5956 ± 0.0153	0.0802 ± 0.0043
σ=0.5	0.1956 ± 0.0082	0.2270 ± 0.0089	0.6060 ± 0.0153	0.0720 ± 0.0042
σ=1	0.1697 ± 0.0082	0.1773 ± 0.0083	0.5970 ± 0.0164	0.0536 ± 0.0041
Adaptive	0.1912 ± 0.0082	0.2193 ± 0.0086	0.6094 ± 0.0154	0.0668 ± 0.0043

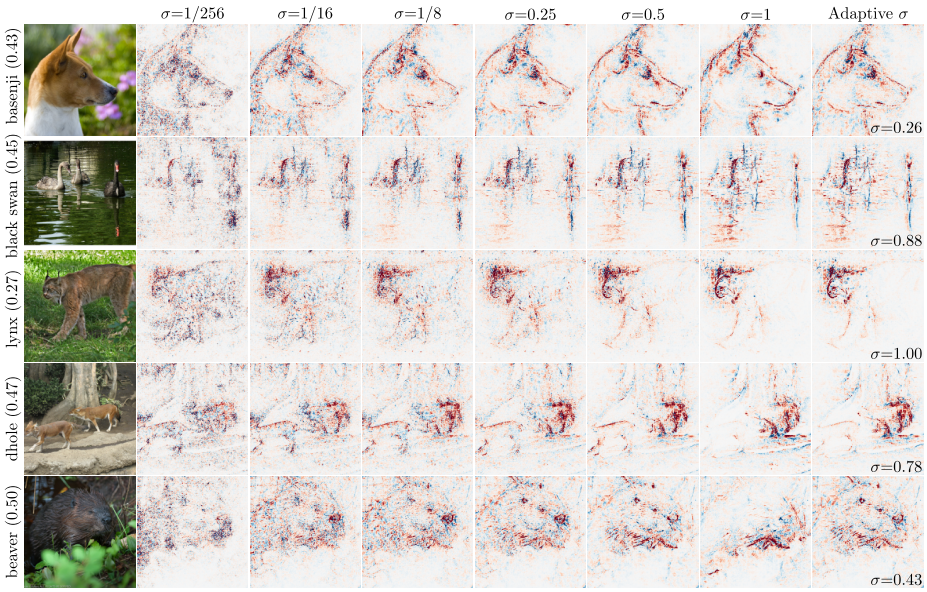

Figure 6: Comparison of Reveal-IG with different σ endpoints, including the adaptive (per-image) endpoint, whose values are displayed inside the corresponding attribution maps in the right column.

than one subdivision is used, the other functions' attributions evolve with k. The angular structure in the radial rings diminishes with k (Fig. 7c), as do the line-shadows for the hill-valley near the origin in the final row.

D Qualitative Results

Here we include additional attribution maps for ResNet-50 and ViT.

E Declarations

Broader impacts. This work develops a feature-attribution method based on distributional paths for information revelation. Potential positive impacts include improved tools for interpreting model predictions, diagnosing model behavior, and comparing how different attribution methods assign credit across features. Such tools may support model debugging, scientific analysis, and more transparent use of machine-learning systems.

The potential negative impacts arise primarily from misuse or overinterpretation of explanations. Attribution methods can produce visually or numerically plausible explanations even when the underlying model is unreliable, biased, or sensitive to the evaluation protocol. They may therefore be misused to justify model decisions in high-stakes settings such as healthcare, hiring, lending, or

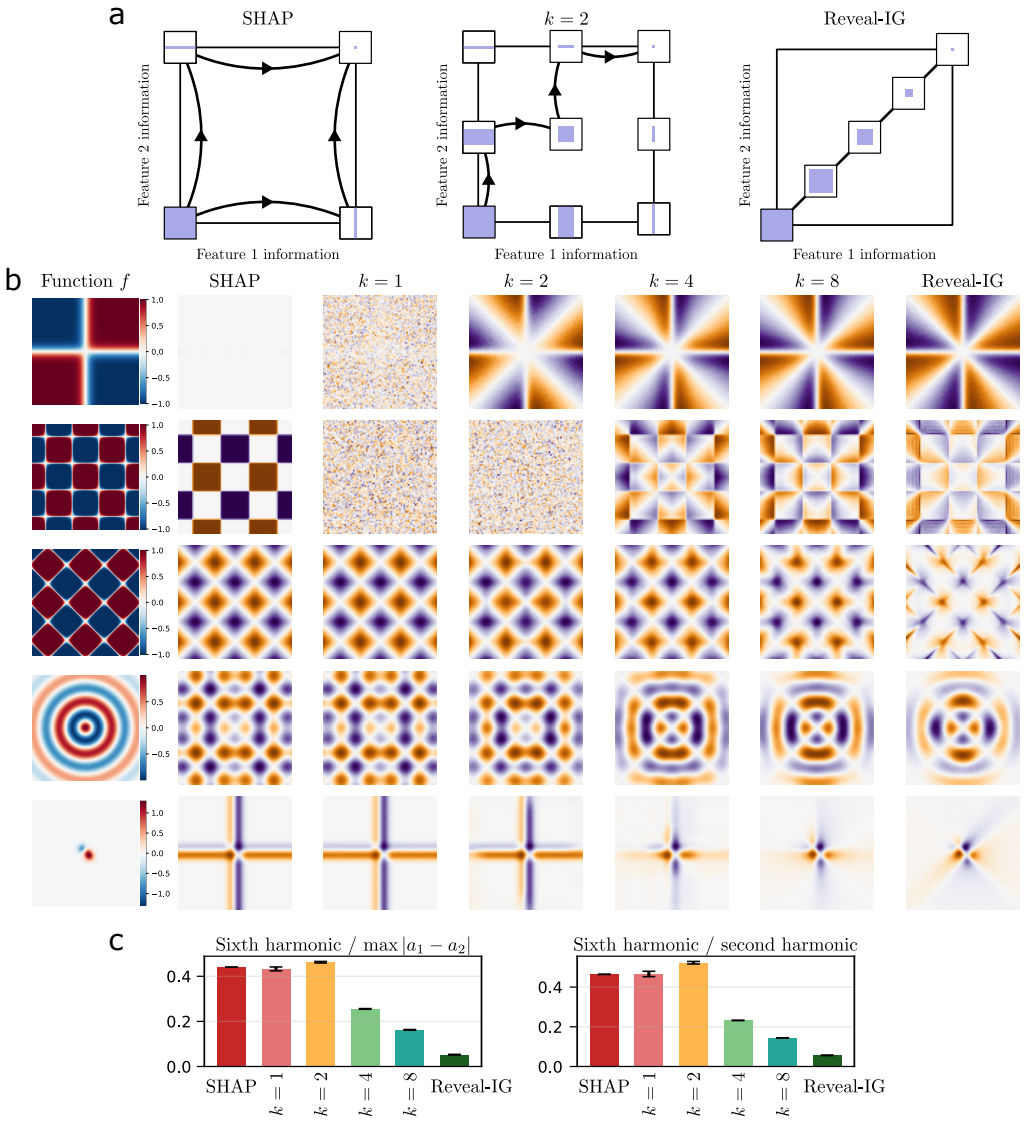

Figure 7: **(a)** The manner of feature information revelation compared schematically for SHAP, the subdivided variant discussed in Sec. 3.1, and Reveal-IG. For SHAP, the feature contributions are averaged over all paths (here, two possibilities). For $k = 2$, one path is shown out of the six possibilities. The insets' shaded regions visualize the distribution over feature values used for the expectation in that step's calculation, and corresponds to a linear interpolation in the distribution's extent. Reveal-IG reveals the feature information along one continuous path. **(b)** For the functions in Fig. 2, we evaluate SHAP, Reveal-IG, and the SHAP analogue. Note the attribution heatmaps' color range is $[-b, b]$ with $b = \max(0.10, |a_1 - a_2|)$. **(c)** Repeating the azimuthal sweep at radius one as done in Sec. 3.1, the amplitude of the second and sixth harmonics are compared; as k increases, the sixth harmonic decays toward that of Reveal-IG's attribution.

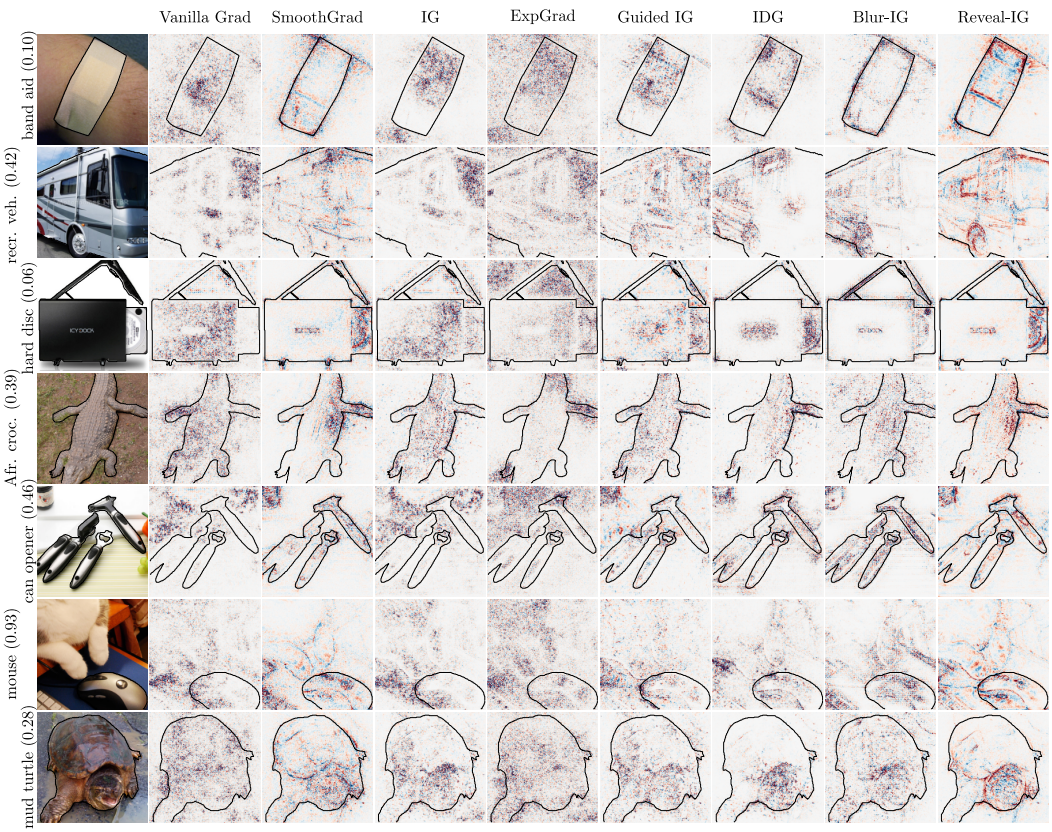

Figure 8: For a random selection of images from the localization set with accompanying ImageNet-S masks and assigned probability to the ground truth class greater than 0.05, we show the outlines of masks in black and the signed attribution maps (red is positive attribution, blue is negative).

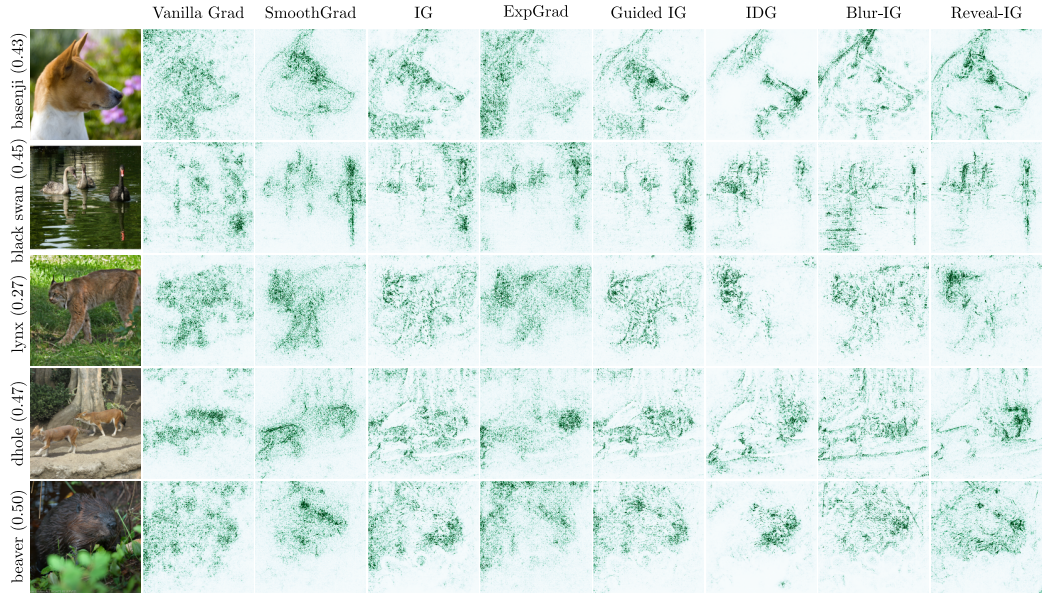

Figure 9: For the same random selection of images from the ImageNet validation set shown in Fig. 3, (ResNet top-1 class and assigned probability on left), we show absolute value saliency maps for the methods compared in Tab. 1.

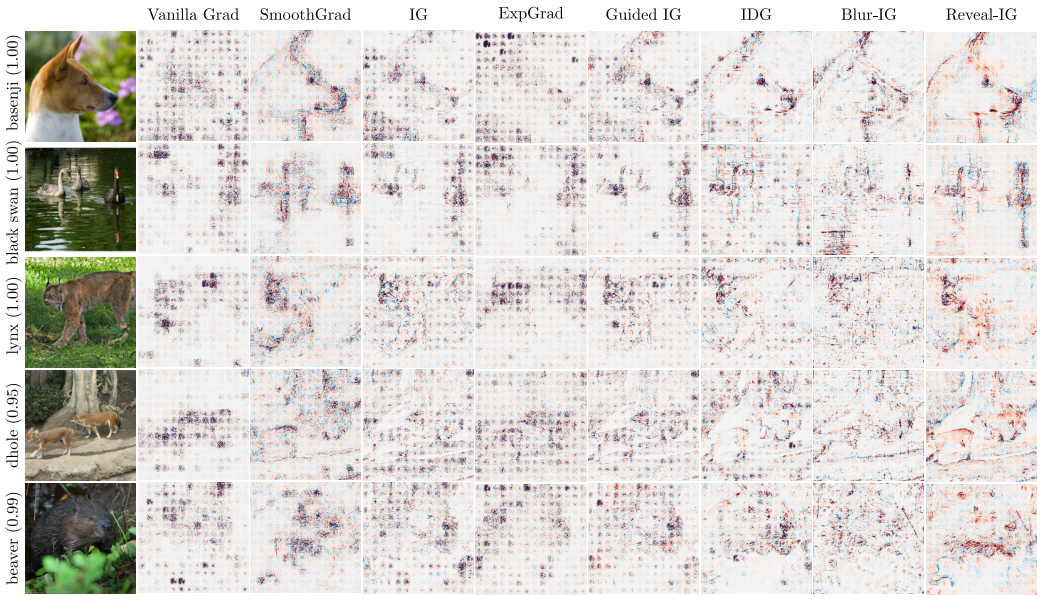

Figure 10: For the same random selection of images from the ImageNet validation set shown in Fig. 3, (ViT-B/16 top-1 class and assigned probability on left), we show signed saliency maps for the methods compared in Tab. 1. Red is positive attribution and blue is negative.

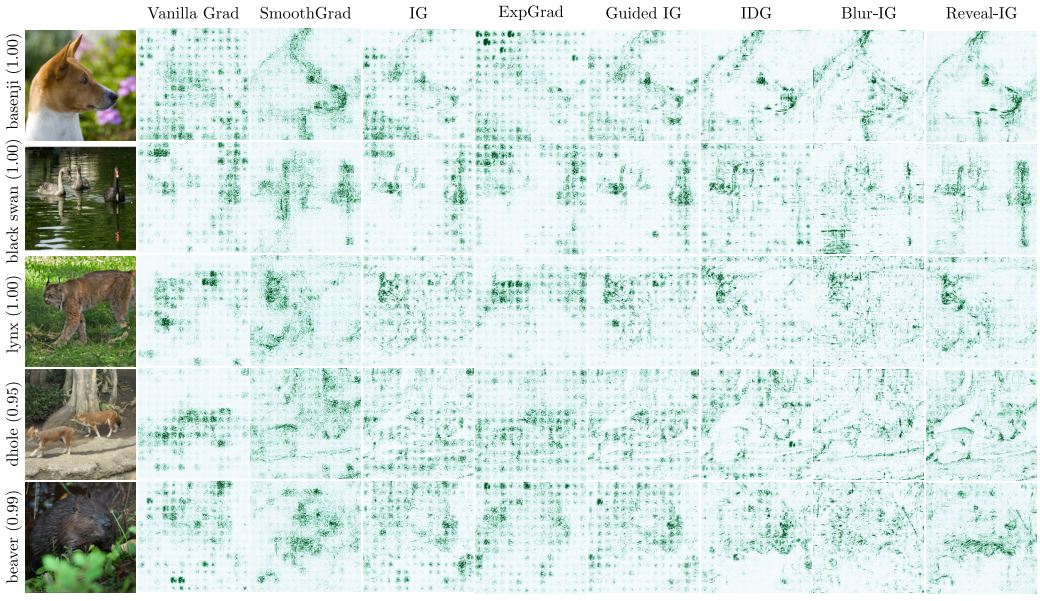

Figure 11: For the same random selection of images from the ImageNet validation set shown in Fig. 3, (ViT-B/16 top-1 class and assigned probability on left), we show absolute value saliency maps for the methods compared in Tab. 1.

surveillance without sufficient validation. Reveal-IG also involves choices of baseline distribution, path, sampling procedure, and perturbation protocol, and these choices can affect the resulting explanation. We mitigate these risks by presenting the method as an analysis tool rather than a certification of model reliability, comparing against established attribution baselines, reporting quantitative evaluation results, and explicitly discussing the role of path and baseline choices.

Compute resources. Attribution experiments were run on a single workstation with an NVIDIA RTX 4500 Ada GPU. The proposed attribution procedure does not require training additional large models. For image experiments, attributions were computed using fixed pretrained vision models, with compute dominated by repeated forward and backward passes for each input. For tabular experiments, we trained only lightweight predictive models on the evaluated tabular datasets; these models were inexpensive relative to the attribution sweeps.