From Baselines to Transport Geodesics: Axiomatic Attribution via Optimal Generative Flows

Cenwei Zhang*
Shanghai Jiao Tong University
Shanghai, China
cwzhang2001@gmail.com

Lin Zhu*
Aalto University
Espoo, Finland
lin.1.zhu@aalto.fi

Manxi Lin
Alibaba
Hangzhou, China
linmanxi.lmx@alibaba-inc.com

Lei You†
Technical University of Denmark
Copenhagen, Denmark
leiyo@dtu.dk

Abstract

Feature attributions often hide a critical modeling choice: they explain a prediction along a counterfactual path from a reference state to an input. Different baselines, interpolations, and generative trajectories define different paths and can therefor produce different explanations. We study this path ambiguity as a modeling problem. Our central question is whether the path can be chosen by the data-generating transport process, rather than by a hand-designed interpolation or by the sensitivity geometry of the model being explained. We separate attribution into fixed-path credit allocation and path selection. For a fixed path, we prove that the Aumann-Shapley line integral is the unique attribution rule under standard fixed-path axioms and explicit coordinate-trace regularity. For path selection, we minimize kinetic action over flows that transport a reference distribution to the data distribution, yielding a transport-geodesic attribution principle. We approximate this ideal with Rectified Flow and Reflow and derive stability bounds linking vector-field error to attribution error. Experiments show that lower-action, transport-consistent paths produce more stable and structured explanations, preserving competitive deletion faithfulness, without claiming data-manifold membership. Our code is available at https://github.com/cenweizhang/OTFlowSHAP.

1 Introduction

Feature attribution asks a simple question: which input coordinates made a model score go up or down for a particular example? This question is simple only at the surface. To explain an image classifier, one usually compares the observed image with a reference state where some information is absent. A black image, a blurred image, a mean image, a masked image, and an inpainted image all define different meanings of absence. They also define different transitions from absence to presence. Since a neural network is nonlinear, the contribution assigned to a pixel or region can change when this transition changes. We therefore view the main ambiguity in attribution as a *path ambiguity*. The problem is not only which attribution formula one should use, but also which counterfactual path one should trust.

Classical Shapley values give a principled answer for finite cooperative games [1]. They allocate the total payoff according to efficiency, symmetry, dummy, and additivity. Model explanation methods

*Equal contribution.
†Corresponding author.

Preprint.

such as SHAP and KernelSHAP adapt this idea to predictors by treating features as players [2, 3]. This adaptation is powerful, but it also exposes the absence problem. To evaluate a coalition, one must define the model value when other features are missing. Different choices give different Shapley games and different explanations [4]. In images, many such counterfactuals look unnatural or low density, so the model may be queried in regions that it never learned to handle.

Path-integral methods reduce the combinatorial burden. Integrated Gradients connects a baseline to the input by a straight line and integrates the gradient along that path [5]. Expected Gradients averages this idea over references [6]. Guided or blurred paths modify the interpolation to reduce visual noise [7]. Recent geometric variants go further by replacing the straight line with Riemannian geodesics on a learned data geometry or on a model-induced geometry [8, 9]. These works make clear that path choice matters. We do not try to show that a transport path is universally stronger than these geometric paths. Explanation methods are difficult to rank by a single number because deletion, insertion, stability, localization, and visual structure measure different properties. Our aim is different, namely, we ask:

> *Whether the attribution path can be defined by the data-generating process rather than by the sensitivity geometry of the model being explained?*

In other words, we study a different path-selection object: not a hand-designed instance-level curve and not a model-induced Riemannian geodesic, but a least-action generative transport process from a reference distribution to the data distribution.

This viewpoint changes the role of the generative model. A diffusion or flow model can certainly provide a structured trajectory, but a structured trajectory is not automatically a principled explanation path. We use optimal transport to specify the target path-selection principle. Among all flows that move a reference distribution p_0 to the data distribution p_1, we choose the one with minimum kinetic action. By the dynamic formulation of optimal transport, this is the Wasserstein-2 geodesic in distribution space [10, 11]. A sample explanation then follows a characteristic curve of this least-action flow and integrates the predictor gradient along that curve.

We separate the construction into two parts. The first part is allocation along a fixed path. Once a smooth path γ from a reference state to the input is given, we ask which rule can split the score change $f(\gamma(1)) - f(\gamma(0))$ among coordinates. We show that natural fixed-path axioms force the Aumann-Shapley line integral. This result says that the allocation rule is not the place where one should add more heuristics. The second part is path selection. We choose the path by the least-action transport principle above. Thus, the paper's main object is not a new saliency heuristic, but a decomposition of the attribution problem into fixed-path credit allocation and distribution-level transport path selection.

This story differs from an "on-manifold Shapley" story. We do not claim that every intermediate point of every sample trajectory lies on a true low-dimensional image manifold. Such a claim would require a formal data manifold, a certificate of membership, and a proof that the learned sampler respects it. Our construction does not need this claim. We only require a reference distribution, a data distribution, and a transport flow between them. The intermediate marginals are generated by this flow. For this reason, we use terms such as *transport-consistent*, *generative*, and *transport-geodesic* rather than strict on-manifold.

The practical method follows the same principle. The ideal Wasserstein geodesic is not available in high-dimensional image spaces. We approximate it with Rectified Flow and Reflow, which learn vector fields that move samples from a simple prior toward the data distribution with increasingly straight transport trajectories [12, 13]. Given an input $\boldsymbol{x}$, we trace a learned flow trajectory between its reference endpoint and $\boldsymbol{x}$, evaluate the target logit gradient along the trajectory, and accumulate coordinatewise products between gradients and path increments. This gives a computable approximation of the transport-geodesic Aumann-Shapley attribution.

We make four contributions. First, we make explicit a separation that is implicit in path-based attribution methods: fixed-path credit allocation is a different problem from path selection. Second, we prove a fixed-path uniqueness theorem showing that the Aumann-Shapley line integral is the unique attribution rule under explicit path axioms and coordinate-trace regularity. Third, we explore a distribution-level path-selection principle based on kinetic-action minimization, which turns baseline choice into an optimal transport problem. Fourth, we instantiate the ideal principle with Rectified

Flow and Reflow, and we empirically study whether lower-action, transport-consistent paths improve stability and visual structure while retaining competitive faithfulness.

Due to the page limit, we present a detailed related work section in the appendix.

2 Problem setup

Let $f_c : \mathbb{R}^d \to \mathbb{R}$ be the scalar score of a fixed predictor for class or target c. We write inputs as bold vectors $\boldsymbol{x} \in \mathbb{R}^d$. We write $\partial_i f(\boldsymbol{x})$ for the partial derivative of a scalar function with respect to coordinate x_i, and we write $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ for the full input gradient. When the input variable is clear, we also write $\nabla f(\boldsymbol{x})$. Let $p_1 \in \mathcal{P}_2(\mathbb{R}^d)$ denote the data distribution and let $p_0 \in \mathcal{P}_2(\mathbb{R}^d)$ denote a reference distribution such as a standard Gaussian or another simple prior. Here $\mathcal{P}_2(\mathbb{R}^d)$ is the set of probability measures with finite second moment. For a specific observed input $\boldsymbol{x}_1$, an explanation should account for the score change between a reference endpoint $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$. The endpoint $\boldsymbol{x}_0$ can be produced by a backward generative flow or by a coupling between p_0 and p_1.

We describe a counterfactual transition by a continuously differentiable path

$$\gamma : [0,1] \to \mathbb{R}^d, \qquad \gamma(0) = \boldsymbol{x}_0, \quad \gamma(1) = \boldsymbol{x}_1.$$

The parameter $t \in [0,1]$ is the path time. For a C^1 path, we write $\dot{\gamma}(t) = \frac{\mathrm{d}\gamma(t)}{\mathrm{d}t}$ for its time derivative, and $\dot{\gamma}_i(t) = \mathrm{d}\gamma_i(t)/\mathrm{d}t$ for the derivative of its i-th coordinate. A C^1 path is called regular when $\dot{\gamma}(t) \neq \mathbf{0}$ for all $t \in [0,1]$. This dot notation is only shorthand; in the main attribution formula we write the derivative explicitly. The path may be a straight line, a diffusion trajectory, a flow trajectory, or an optimal transport characteristic.

A path-based attribution rule returns a vector $\boldsymbol{\phi}(f_c, \gamma) \in \mathbb{R}^d$ whose i-th coordinate represents the contribution of input coordinate i to the score change along γ. Efficiency asks that these contributions sum to the finite difference:

$$\sum_{i=1}^{d} \phi_i(f_c, \gamma) = f_c(\gamma(1)) - f_c(\gamma(0)). \tag{1}$$

Equation (1) is the continuous analogue of Shapley efficiency. It is necessary but not sufficient. It says that the accounting balances, but it does not say how the accounting should be done or which path should be used.

The central design principle of our method is not to design a new saliency heuristic, but to separate two decisions that are often entangled. For an input $\boldsymbol{x}$, write $\gamma_{\boldsymbol{x}}$ for a counterfactual path ending at $\boldsymbol{x}$, write $\Phi(f, \gamma)$ for the Aumann-Shapley attribution vector along a fixed path, and write $\mathrm{Char}_{\boldsymbol{x}}(\boldsymbol{v})$ for the characteristic curve of a velocity field $\boldsymbol{v}$ that ends at $\boldsymbol{x}$. Then the organizing equation is

$$\mathrm{Attr}(f, \boldsymbol{x}) = \underbrace{\Phi(f, \gamma_{\boldsymbol{x}})}_{\substack{\text{Aumann-Shapley} \\ \text{credit allocation}}} \quad \text{with} \quad \underbrace{\begin{array}{c}(\rho^\star, \boldsymbol{v}^\star) \in \arg \min_{(\rho,\boldsymbol{v}):p_0 \to p_1} \mathcal{A}(\rho, \boldsymbol{v}), \\ \gamma_{\boldsymbol{x}} = \mathrm{Char}_{\boldsymbol{x}}(\boldsymbol{v}^\star)\end{array}}_{\substack{\text{transport-geodesic} \\ \text{path selection}}}.$$

Here $\mathcal{A}$ denotes kinetic action, which we define formally in Eq. (5). The blue term asks how to allocate score change once a path is fixed. This is the part addressed by Aumann-Shapley path integrals. The orange term asks which counterfactual path should be used. This is the part that is often fixed by a baseline or a hand-designed interpolation rule; we instead pose it as an optimal generative transport problem.

A time-dependent velocity field $\boldsymbol{v}_t : \mathbb{R}^d \to \mathbb{R}^d$ induces trajectories through the ordinary differential equation

$$\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{v}_t(\boldsymbol{x}_t), \qquad t \in [0,1]. \tag{2}$$

For any time-dependent state $\boldsymbol{x}_t$, the shorthand $\dot{\boldsymbol{x}}_t$ means $\mathrm{d}\boldsymbol{x}_t/\mathrm{d}t$. If $\boldsymbol{x}_0 \sim p_0$ and the solution of Eq. (2) has marginal law ρ_t, then mass conservation is described by the continuity equation

$$\partial_t \rho_t + \nabla \cdot (\rho_t \boldsymbol{v}_t) = 0, \qquad \rho_0 = p_0, \qquad \rho_1 = p_1. \tag{3}$$

Equation (3) is a distribution-level constraint. It does not say that $\rho_t = p_1$ for intermediate times. The intermediate laws ρ_t form a generative bridge between the reference and data distributions.

We will use two types of objects. A fixed path γ is the object used by the attribution integral. A flow pair $(\rho, \boldsymbol{v})$, where $\rho = (\rho_t)_{t\in[0,1]}$ and $\boldsymbol{v} = (\boldsymbol{v}_t)_{t\in[0,1]}$, is the object used to select paths. In experiments, we parameterize the learned velocity field by a neural network $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$. When we compare theory and implementation, we write $\hat{\boldsymbol{v}}_t(\boldsymbol{x}) = \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ for the learned field. Its characteristic curves supply the numerical paths for attribution.

3 Axiomatic attribution along a fixed path

We first solve the allocation problem while treating the path as fixed. This part of the theory is deliberately independent of optimal transport. It says that, after the user or a generative model has chosen a smooth path, the natural attribution rule is forced by axioms.

Definition 3.1 (Admissible score class and path attribution rule)**.** Let $\mathcal{F}$ be an admissible class of scalar scores. In this paper, admissible means that every $f \in \mathcal{F}$ is C^1 on an open set containing the image of the path under consideration, and that $\mathcal{F}$ is closed under finite linear combinations. For a C^1 path $\gamma : [0, 1] \to \mathbb{R}^d$, a path attribution rule assigns a number $A_i(f, \gamma) \in \mathbb{R}$ to each score $f \in \mathcal{F}$ and each coordinate $i \in \{1, \dots, d\}$. The vector $\boldsymbol{A}(f, \gamma) = (A_1(f, \gamma), \dots, A_d(f, \gamma))$ explains the score difference $f(\gamma(1)) - f(\gamma(0))$ along γ.

For a score $f \in \mathcal{F}$, a path γ, and a coordinate i, define the coordinate trace

$$h_i^f(t) = \frac{\partial f(\gamma(t))}{\partial x_i} \frac{\mathrm{d}\gamma_i(t)}{\mathrm{d}t}, \qquad t \in [0, 1].$$

This trace is the infinitesimal score change attributed to coordinate i at path time t. The following assumption collects the fixed-path axioms. We state the axioms as an assumption because the representation theorem needs to know exactly what information the rule is allowed to use.

Assumption 3.2 (Fixed-path attribution axioms)**.** For every fixed regular C^1 path γ, the rule $A_i(f, \gamma)$ satisfies the following conditions for all admissible scores. It satisfies efficiency, namely Eq. (1). It is linear in the model score: for real numbers a, b and scores $f, g \in \mathcal{F}$,

$$A_i(af + bg, \gamma) = aA_i(f, \gamma) + bA_i(g, \gamma).$$

It satisfies the dummy property: if $h_i^f(t) = 0$ for all $t \in [0, 1]$, then $A_i(f, \gamma) = 0$. It is invariant to smooth increasing reparameterizations of the path: for any C^1 bijection $\sigma : [0, 1] \to [0, 1]$ with $\sigma(0) = 0$, $\sigma(1) = 1$, and $\mathrm{d}\sigma(t)/\mathrm{d}t > 0$, we have $A_i(f, \gamma \circ \sigma) = A_i(f, \gamma)$. Finally, it is coordinate-trace determined and continuous: for each i, the value $A_i(f, \gamma)$ depends on f along γ only through the scalar trace h_i^f, and this dependence is continuous under uniform convergence of that trace.

The coordinate-trace condition is not a technical trick. It encodes what a coordinate attribution along a path means. If the attribution assigned to coordinate i changed when all coordinate-i infinitesimal contributions along the path stayed the same, then the rule would be using information outside the coordinate trace to assign coordinate-i credit. That behavior would no longer be a coordinatewise path attribution.

Assumption 3.3 (Coordinate-trace richness)**.** For the fixed path γ, the admissible score class $\mathcal{F}$ is rich enough to separate coordinate traces. Concretely, for any coordinate i and any scalar trace that can be written as $h(t) = a(t)\,\mathrm{d}\gamma_i(t)/\mathrm{d}t$ with continuous a, there is an admissible score $g \in \mathcal{F}$ such that

$$\frac{\partial g(\gamma(t))}{\partial x_i} \frac{\mathrm{d}\gamma_i(t)}{\mathrm{d}t} = h(t) \quad \text{and} \quad \frac{\partial g(\gamma(t))}{\partial x_j} \frac{\mathrm{d}\gamma_j(t)}{\mathrm{d}t} = 0 \quad \text{for all } j \neq i.$$

Assumption 3.3 makes explicit a standard richness requirement behind representation theorems. It rules out degenerate score classes where a coordinate trace cannot be varied without also changing all other coordinate traces. For smooth embedded paths, this condition can be obtained by locally extending prescribed first-order information along the curve.

Definition 3.4 (Aumann-Shapley path attribution)**.** Let $f \in \mathcal{F}$ be differentiable in a neighborhood of the image of γ. The Aumann-Shapley attribution of coordinate i along γ is

$$\Phi_i(f, \gamma) = \int_0^1 \frac{\partial f(\gamma(t))}{\partial x_i} \frac{\mathrm{d}\gamma_i(t)}{\mathrm{d}t}\, \mathrm{d}t. \tag{4}$$

The definition is the coordinate decomposition of the line integral of ∇f along γ. Summing Eq. (4) over coordinates gives

$$\sum_{i=1}^{d} \Phi_i(f,\gamma) = \int_0^1 \nabla f(\gamma(t))^\top \frac{\mathrm{d}\gamma(t)}{\mathrm{d}t}\,\mathrm{d}t = f(\gamma(1)) - f(\gamma(0)),$$

so efficiency follows from the chain rule. Reparameterization invariance follows because a change of the time variable changes the path derivative and the integration variable in compensating ways.

Theorem 3.5 (Fixed-path uniqueness)**.** *Fix a regular C^1 path γ. Any path attribution rule satisfying Assumptions 3.2 and 3.3 coincides with the Aumann-Shapley path attribution in Eq.* (4)*. That is, for every $f \in \mathcal{F}$ and every coordinate i,*

$$A_i(f,\gamma) = \Phi_i(f,\gamma).$$

Theorem 3.5 removes one degree of freedom from the explanation problem. Once a path is fixed, the credit-allocation rule is not an additional design choice. The remaining question is therefore not how to modify the attribution formula, but which path should be used.

This result also clarifies the relation to Integrated Gradients. Integrated Gradients is Eq. (4) for the straight path $\gamma(t) = \boldsymbol{x}_0 + t(\boldsymbol{x}_1 - \boldsymbol{x}_0)$. Our method keeps the same fixed-path allocation principle, but replaces the straight line by a path selected through optimal generative transport.

4 Transport-geodesic path selection by optimal generative flows

We now address the path-selection problem. If the path determines the explanation, then the path should not be an arbitrary drawing in pixel space. We explore an optimal transport principle for this choice. Among all flows that move p_0 to p_1, the ideal object is the flow with minimum kinetic action. This is a path-selection principle, not a claim that it is the best attribution method under every downstream metric.

Assumption 4.1 (Transport regularity)**.** The reference and data distributions $p_0, p_1 \in \mathcal{P}_2(\mathbb{R}^d)$ are absolutely continuous with respect to Lebesgue measure on their supports, and the quadratic-cost optimal transport problem between them admits a unique optimal dynamic plan. The corresponding velocity field $\boldsymbol{v}_t^\star$ is locally Lipschitz on the compact region that contains the trajectories used for attribution.

Assumption 4.1 is a standard regularity condition for the ideal theory. It is stronger than what we can certify for high-dimensional learned image models. We use it to define the target object. The learned Rectified Flow is an approximation to this target, not a proof that exact optimal transport has been recovered.

We write $W_2(p_0, p_1)$ for the quadratic Wasserstein distance between p_0 and p_1. For any admissible flow pair $(\rho, \boldsymbol{v})$, with $\rho = (\rho_t)_{t\in[0,1]}$ and $\boldsymbol{v} = (\boldsymbol{v}_t)_{t\in[0,1]}$, satisfying Eq. (3), define its kinetic action by

$$\mathcal{A}(\rho,\boldsymbol{v}) = \int_0^1 \int_{\mathbb{R}^d} \|\boldsymbol{v}_t(\boldsymbol{x})\|_2^2 \,\mathrm{d}\rho_t(\boldsymbol{x})\,\mathrm{d}t. \tag{5}$$

When ρ_t admits a density, this measure integral is the same as $\int \|\boldsymbol{v}_t(\boldsymbol{x})\|_2^2 \rho_t(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. The Benamou-Brenier formula states that

$$W_2^2(p_0,p_1) = \inf_{(\rho,\boldsymbol{v}) \text{ satisfying Eq. (3)}} \mathcal{A}(\rho,\boldsymbol{v}). \tag{6}$$

The minimizer $(\rho_t^\star, \boldsymbol{v}_t^\star)$ is the Wasserstein-2 geodesic in distribution space [10, 11]. This is the sense in which we use the word geodesic. The path of distributions is shortest in the Wasserstein geometry. A single sample trajectory is a characteristic curve of this distributional flow, not a Riemannian geodesic in a prescribed input-space metric.

Definition 4.2 (Transport-geodesic characteristic path)**.** Under Assumption 4.1, let $(\rho^\star, \boldsymbol{v}^\star)$ minimize Eq. (5). For an observed endpoint $\boldsymbol{x}_1$ in the support of p_1, the transport-geodesic characteristic path $\gamma^\star_{\boldsymbol{x}_1}$ is the solution of

$$\frac{\mathrm{d}\gamma^\star_{\boldsymbol{x}_1}(t)}{\mathrm{d}t} = \boldsymbol{v}_t^\star(\gamma^\star_{\boldsymbol{x}_1}(t)), \qquad \gamma^\star_{\boldsymbol{x}_1}(1) = \boldsymbol{x}_1, \tag{7}$$

traced backward to a reference endpoint $\gamma^\star_{\boldsymbol{x}_1}(0) = \boldsymbol{x}_0$.

If $\boldsymbol{x}_0 \sim p_0$ and $\boldsymbol{x}_t$ follows Eq. (7) forward in time, then $\boldsymbol{x}_t \sim \rho_t^\star$ for each t. Notice the precise statement: $\rho_t^\star$ interpolates between p_0 and p_1. We do not need or claim $\rho_t^\star = p_1$ for all t.

Definition 4.3 (Transport-geodesic Aumann-Shapley attribution)**.** For a differentiable target score f_c and an input $\boldsymbol{x}_1$, the transport-geodesic Aumann-Shapley attribution is

$$\Psi_i(f_c, \boldsymbol{x}_1) = \int_0^1 \frac{\partial f_c(\gamma^\star_{\boldsymbol{x}_1}(t))}{\partial x_i} \frac{\mathrm{d}\gamma^\star_{\boldsymbol{x}_1,i}(t)}{\mathrm{d}t}\,\mathrm{d}t. \tag{8}$$

Definition 4.3 combines the two parts of the paper. The Aumann-Shapley integral allocates credit along a path, and the optimal transport problem selects the path. The explanation is canonical only relative to the ideal transport problem and its regularity assumptions. It is best understood as a principled transport alternative to baseline paths and Riemannian path choices, not as a universal dominance claim. This definition separates two roles that are often entangled. The generative transport flow determines where the explanation path goes, while the predictor f_c determines how the score changes along that path. Thus, the path is not chosen by the same sensitivity geometry that is being explained.

Theorem 4.4 (Axiomatic and transport-geodesic characterization)**.** *Suppose Assumptions 3.2, 3.3, and 4.1 hold. Among all attribution rules that first choose a kinetic-action-minimizing flow from p_0 to p_1 and then apply a fixed-path attribution rule to its characteristic path, Eq. (8) is the unique rule satisfying the fixed-path axioms.*

Theorem 4.4 says that there are two sources of uniqueness. The fixed-path axioms force the line integral, while the kinetic-action principle chooses the distributional path. This is the precise version of the paper's main claim. It is not a statement of strict manifold membership, and it does not impose a total ordering over all attribution methods.

The construction reduces to familiar cases. If γ is the straight line from a fixed baseline to $\boldsymbol{x}_1$, Eq. (4) recovers Integrated Gradients. If the model is additive, namely $f(\boldsymbol{x}) = b + \sum_i f_i(x_i)$, and the transport path separates across coordinates, Eq. (8) returns the coordinatewise finite differences $f_i(x_{1,i}) - f_i(x_{0,i})$, which match the classical Shapley allocation for the induced additive game.

We also need a stability guarantee because the ideal flow is not observed. Let $\hat{\boldsymbol{v}}_t(\boldsymbol{x}) = \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ denote the learned vector field. For an observed endpoint $\boldsymbol{x}_1$, let $\hat{\gamma}_{\boldsymbol{x}_1}$ be the learned characteristic defined by

$$\frac{\mathrm{d}\hat{\gamma}_{\boldsymbol{x}_1}(t)}{\mathrm{d}t} = \hat{\boldsymbol{v}}_t(\hat{\gamma}_{\boldsymbol{x}_1}(t)), \qquad \hat{\gamma}_{\boldsymbol{x}_1}(1) = \boldsymbol{x}_1.$$

Define $\hat{\Psi}_i$ by replacing $\gamma^\star_{\boldsymbol{x}_1}$ with $\hat{\gamma}_{\boldsymbol{x}_1}$ in Eq. (8).

Assumption 4.5 (Stability regularity)**.** The ideal and learned trajectories remain in a compact set $\mathcal{K} \subset \mathbb{R}^d$. On $\mathcal{K}$, the target score f_c has bounded gradient and Lipschitz gradient, and both vector fields are uniformly Lipschitz in $\boldsymbol{x}$. Moreover,

$$\sup_{t\in[0,1],\boldsymbol{x}\in\mathcal{K}} \|\hat{\boldsymbol{v}}_t(\boldsymbol{x}) - \boldsymbol{v}_t^\star(\boldsymbol{x})\|_2 \le \varepsilon.$$

Theorem 4.6 (Stability under flow approximation)**.** *Under Assumption 4.5, there is a constant C, depending only on the compact set, the time horizon, the Lipschitz constants of the vector fields, and the first two derivative bounds of f_c, such that for every coordinate i,*

$$|\Psi_i(f_c, \boldsymbol{x}_1) - \hat{\Psi}_i(f_c, \boldsymbol{x}_1)| \le C\varepsilon. \tag{9}$$

Theorem 4.6 gives the main engineering message. Improving the learned vector field improves the attribution in a controlled way. The theorem does not say that the learned path is exact optimal transport. It says that, when the learned vector field approaches the ideal field on the relevant region, the attribution approaches the ideal transport-geodesic attribution.

5 Implementation and experiments

We instantiate the ideal construction with Rectified Flow and Reflow. Rectified Flow learns a time-dependent vector field $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ that moves samples from a reference distribution to the data

Algorithm 1 Geodesic Aumann-Shapley attribution with a learned flow

Require: target score f_c, input $\boldsymbol{x}_1$, learned vector field $\boldsymbol{v}_{\boldsymbol{\theta}}$, number of steps K
integrate $\mathrm{d}\boldsymbol{x}_t/\mathrm{d}t = \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ backward from $\boldsymbol{x}_1$ to obtain $\hat{\boldsymbol{x}}_0$
integrate the same ODE forward from $\hat{\boldsymbol{x}}_0$ and store $\hat{\boldsymbol{x}}_0, \dots, \hat{\boldsymbol{x}}_K$
initialize $\hat{\boldsymbol{\Psi}} = \mathbf{0}$
for $k = 0, \dots, K-1$ **do**
    compute $\boldsymbol{g}_k = \nabla_{\boldsymbol{x}} f_c(\hat{\boldsymbol{x}}_k)$
    update $\hat{\boldsymbol{\Psi}} \leftarrow \hat{\boldsymbol{\Psi}} + \boldsymbol{g}_k \odot (\hat{\boldsymbol{x}}_{k+1} - \hat{\boldsymbol{x}}_k)$
end for
return $\hat{\boldsymbol{\Psi}}$

distribution. Given paired samples $(\boldsymbol{z}_0, \boldsymbol{z}_1) \sim \pi$, where π is a coupling between p_0 and p_1, the basic training objective is

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t,\boldsymbol{z}_0,\boldsymbol{z}_1} \left\| \boldsymbol{v}_{\boldsymbol{\theta}}((1-t)\boldsymbol{z}_0 + t\boldsymbol{z}_1, t) - (\boldsymbol{z}_1 - \boldsymbol{z}_0) \right\|_2^2 . \tag{10}$$

Reflow improves the coupling by first generating trajectories with a learned flow and then retraining on the induced endpoint pairs [12]. This procedure is useful for us because lower-curvature and lower-action paths are closer to the geodesic ideal in Eq. (6).

For a target input $\boldsymbol{x}_1$, we integrate the learned ODE backward to obtain a reference endpoint $\hat{\boldsymbol{x}}_0$, then integrate forward to obtain states $\hat{\boldsymbol{x}}_0, \dots, \hat{\boldsymbol{x}}_K$. Throughout the experiments, K denotes the number of integration steps, $t_k = k/K$, $\Delta t = 1/K$, and $\hat{\boldsymbol{x}}_k \approx \hat{\gamma}_{\boldsymbol{x}_1}(t_k)$. We compute gradients of the target score at these states and approximate Eq. (8) by a Riemann sum:

$$\hat{\Psi}_i(f_c, \boldsymbol{x}_1) = \sum_{k=0}^{K-1} \frac{\partial f_c(\hat{\boldsymbol{x}}_k)}{\partial x_i} (\hat{x}_{k+1,i} - \hat{x}_{k,i}). \tag{11}$$

We use the same number of integration steps for path-based baselines whenever possible. The cost of one explanation is linear in K: it requires ODE evaluations for the generative path and $K+1$ gradient evaluations of the predictor. This is much cheaper than exact discrete Shapley, whose cost grows exponentially in the number of input features. In Alg. 1, $\odot$ denotes coordinatewise multiplication.

5.1 Evaluation questions

We organize experiments around four questions. These questions are meant to test the transport path-selection hypothesis, not to establish a universal ranking of attribution methods. Does the numerical integral satisfy efficiency as the step count increases? Does Reflow reduce the geometry gap relative to less rectified flows? Does attribution error track vector-field error as predicted by Theorem 4.6? On real images, do transport-consistent paths give structured explanations without destroying deletion faithfulness?

The evaluation uses CUB-200 for the numerical completeness study [14], controlled flow checkpoints for stability analysis, and CIFAR-10 plus CelebA-HQ for image benchmarks [15, 16]. We compare with SmoothGrad [17], Guided Backpropagation [18], GradientSHAP and KernelSHAP-style baselines [3], Integrated Gradients [5], and DDIM generative paths [19]. We report standard deletion metrics together with path diagnostics. We use the term Flow Consistency Error only for dynamical consistency with the learned vector field.

5.2 Numerical completeness

Efficiency requires the attribution sum to match the score change. For N evaluated examples, we measure the residual

$$R_{\text{eff}} = \frac{1}{N} \sum_{j=1}^{N} \left| \sum_{i=1}^{d} \hat{\Psi}_i^{(j)} - \left(f_c(\boldsymbol{x}_1^{(j)}) - f_c(\hat{\boldsymbol{x}}_0^{(j)}) \right) \right| .$$

Table 1 shows convergence as the number of integration steps increases. The default $K = 50$ gives a practical balance, while $K = 100$ or $K = 200$ can be used when tighter efficiency residuals are desired.

Table 1: Completeness residual for the discrete estimator in Eq. (11). The residual decreases as the integration grid becomes finer.

Steps K	MAE ↓	Std. dev.	SEM	Relative error ↓
10	1.483	1.346	0.177	19.30%
20	0.895	0.778	0.102	11.65%
50	0.411	0.317	0.042	5.34%
100	0.229	0.181	0.024	2.98%
200	0.101	0.079	0.010	1.34%

Table 2: Path action and attribution stability across seeds. Lower-action reflowed paths produce more stable maps and more consistent feature rankings.

Method	Action $\widehat{\mathcal{A}}$ ↓	Pixel var. ↓	SSIM ↑	Rank corr. ↑
1-RF	3179.2 ± 295.8	0.0032 ± 0.0021	0.716 ± 0.080	0.662 ± 0.087
2-RF	$\mathbf{3006.9 \pm 340.4}$	$\mathbf{0.0010 \pm 0.0008}$	$\mathbf{0.911 \pm 0.051}$	$\mathbf{0.882 \pm 0.061}$

5.3 Path geometry and stability

We next ask whether a more geodesic transport path produces a more stable attribution. We compare a one-step Rectified Flow baseline with a reflowed model over three seeds. We estimate discrete kinetic action by

$$\widehat{\mathcal{A}} = \sum_{k=0}^{K-1} \frac{\|\hat{\boldsymbol{x}}_{k+1} - \hat{\boldsymbol{x}}_k\|_2^2}{\Delta t}.$$

We measure attribution stability with pixelwise variance, SSIM between attribution maps, and rank correlation of feature scores. Table 2 shows that Reflow moderately lowers action and substantially improves stability. We interpret this as evidence that path geometry matters for explanation stability. We avoid saying that this proves strict manifold adherence.

The stability bound in Theorem 4.6 predicts that attribution error should decrease when the learned vector field approaches a stronger oracle field. We treat a converged flow checkpoint as an oracle and compare earlier checkpoints against it. Let $\boldsymbol{\Psi}^{\mathrm{ref}}$ and $\boldsymbol{v}_{t_k}^{\mathrm{ref}}$ denote the attribution and velocity field of this reference checkpoint. We use $\|\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}^{\mathrm{ref}}\|_2 / (\|\boldsymbol{\Psi}^{\mathrm{ref}}\|_2 + 10^{-12})$ as the relative attribution error and

$$E_{\mathrm{field}} = \frac{\left(\sum_{k=0}^{K-1} \|\hat{\boldsymbol{v}}_{t_k}(\hat{\boldsymbol{x}}_k) - \boldsymbol{v}_{t_k}^{\mathrm{ref}}(\hat{\boldsymbol{x}}_k)\|_2^2\right)^{1/2}}{\left(\sum_{k=0}^{K-1} \|\boldsymbol{v}_{t_k}^{\mathrm{ref}}(\hat{\boldsymbol{x}}_k)\|_2^2\right)^{1/2} + 10^{-12}}$$

as the empirical flow approximation error. Across checkpoints, these two quantities show an approximately linear relation, with median Pearson correlation above 0.95. This supports the narrower claim that generative flow quality controls the reliability of the transport-based attribution we study. We keep the full scatter plots for the appendix so that the main text can focus on the central evidence in Table 2.

5.4 Image benchmarks

For CIFAR-10 and CelebA-HQ, we report both path diagnostics and standard explanation metrics. The geometric path straightness score is

$$\mathrm{GPS} = \frac{\sum_{k=0}^{K-1} \|\hat{\boldsymbol{x}}_{k+1} - \hat{\boldsymbol{x}}_k\|_2}{\|\hat{\boldsymbol{x}}_K - \hat{\boldsymbol{x}}_0\|_2}.$$

A value near one indicates a nearly straight discrete path in ambient space. The Flow Consistency Error is

$$\mathrm{FCE} = \frac{1}{K} \sum_{k=0}^{K-1} \left\| \frac{\hat{\boldsymbol{x}}_{k+1} - \hat{\boldsymbol{x}}_k}{\Delta t} - \hat{\boldsymbol{v}}_{t_k}(\hat{\boldsymbol{x}}_k) \right\|_2^2. \tag{12}$$

Table 3: CIFAR-10 and CelebA-HQ benchmark results. We report path diagnostics only for methods that define a continuous path. Our method has much lower FCE than IG and DDIM, while deletion faithfulness remains competitive rather than uniformly best.

Data	Method	GPS	FCE ↓	SATV ↓	EAS ↑	Del. zero ↓	Del. blur ↓
CIFAR-10	SmoothGrad	-	-	0.022 ± 0.008	0.363 ± 0.117	0.525 ± 0.923	1.216 ± 1.202
	GuidedBackprop	-	-	0.040 ± 0.018	0.439 ± 0.128	0.966 ± 0.978	1.627 ± 1.064
	GradientSHAP	-	-	0.669 ± 0.408	0.116 ± 0.146	0.691 ± 0.975	1.978 ± 1.050
	Integrated Gradients	1.000 ± 0.000	$(1.0 \pm 0.1) \times 10^4$	0.643 ± 0.404	0.114 ± 0.134	0.670 ± 0.965	1.931 ± 1.105
	DDIM	1.059 ± 0.017	$(3.1 \pm 0.5) \times 10^3$	0.073 ± 0.062	0.119 ± 0.160	$\mathbf{0.444 \pm 0.940}$	1.360 ± 1.143
	Transport Flow	1.023 ± 0.008	$\mathbf{0.005 \pm 0.003}$	$\mathbf{0.062 \pm 0.048}$	$\mathbf{0.119 \pm 0.138}$	0.456 ± 0.967	$\mathbf{1.311 \pm 1.177}$
CelebA-HQ	SmoothGrad	-	-	0.001 ± 0.000	0.232 ± 0.105	0.216 ± 0.356	0.631 ± 0.483
	GuidedBackprop	-	-	0.001 ± 0.000	0.385 ± 0.097	0.222 ± 0.313	0.475 ± 0.434
	GradientSHAP	-	-	0.011 ± 0.004	0.061 ± 0.114	0.188 ± 0.305	1.107 ± 0.663
	Integrated Gradients	1.000 ± 0.000	$(1.3 \pm 0.1) \times 10^6$	0.010 ± 0.003	0.060 ± 0.114	0.188 ± 0.304	1.108 ± 0.662
	DDIM	1.047 ± 0.006	$(2.8 \pm 0.1) \times 10^5$	$\mathbf{0.003 \pm 0.001}$	$\mathbf{0.147 \pm 0.097}$	$\mathbf{0.175 \pm 0.321}$	$\mathbf{0.897 \pm 0.579}$
	Transport Flow	1.011 ± 0.005	$\mathbf{1.780 \pm 0.557}$	$\mathbf{0.003 \pm 0.001}$	0.091 ± 0.101	0.184 ± 0.320	0.926 ± 0.587

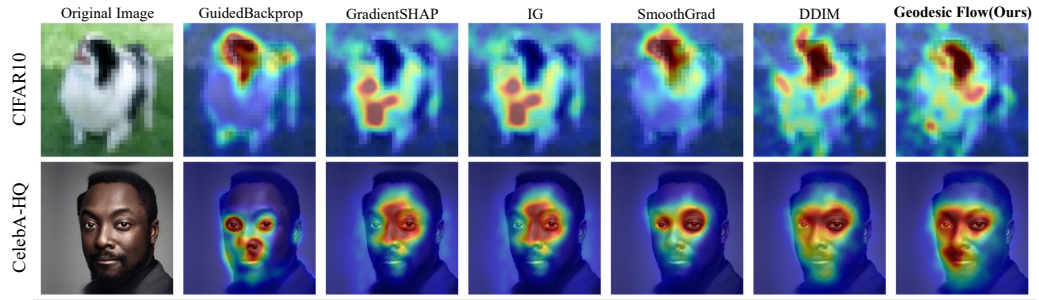

Figure 1: Qualitative visualization on CIFAR-10 and CelebA-HQ. Transport Flow explanations are spatially coherent and avoid some high-frequency artifacts of straight-line Integrated Gradients. The figure should be read together with Table 3; visual clarity alone is not a complete faithfulness metric.

Equation (12) measures consistency with the learned dynamics. It does not certify membership in the true data distribution. For spatial structure, we use a Structure-Aware Total Variation diagnostic,

$$\mathrm{SATV}(\boldsymbol{\phi}) = \sum_{(u,v)\in\Omega} \|\nabla_{\mathrm{img}} \boldsymbol{\phi}_{u,v}\|_1 \exp\left(-\alpha \|\nabla_{\mathrm{img}} I_{u,v}\|_2\right), \tag{13}$$

where Ω is the pixel grid, I is the input image, ∇_{img} denotes finite differences on the image grid, and $\alpha = 10$. Equation (13) penalizes high-frequency saliency variation in flat image regions while allowing changes near image edges. We also report the Edge Alignment Score (EAS), which measures whether attribution-map edges align with image edges, and deletion scores under zero and blur replacement.

The main conclusion from Table 3 is not that one method dominates every metric. On CIFAR-10, Transport Flow improves FCE by several orders of magnitude relative to IG and DDIM and has the best blur-deletion score among the path-based methods. On CelebA-HQ, DDIM has slightly better deletion scores, while Transport Flow keeps comparable deletion performance and much lower FCE. These results support the paper's intended claim: transport-consistent path selection improves the geometry and stability of this family of explanations while preserving competitive faithfulness. They should not be read as a claim that Transport Flow is uniformly better than all Riemannian or perturbation-based attribution methods.

The qualitative examples in Figure 1 match the quantitative trend. Straight-line IG often produces scattered patterns because it evaluates gradients along a path that is easy to compute but not adapted to the generative transition. DDIM gives a structured generative path but can be curved and sensitive to discretization. Our flow path is not a proof of true manifold membership, yet it is dynamically consistent with the learned transport field and empirically gives structured attributions.

6 Conclusion

We reformulate feature attribution as a path-selection problem: fixed-path Aumann-Shapley axioms determine how score change is allocated, while a least-action generative transport from a reference distribution to the data distribution provides a complementary transport-geodesic path whose Rectified Flow/Reflow approximation yields more stable and dynamically consistent attributions with competitive faithfulness, without claiming strict manifold membership or universal superiority across metrics.

References

[1] Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, 1953.

[2] Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18, 2010.

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

[4] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

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

[6] 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.

[7] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5050–5058, 2021.

[8] Eslam Zaher, Maciej Trzaskowski, Quan Nguyen, and Fred Roosta. Manifold integrated gradients: Riemannian geometry for feature attribution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 58090–58104. PMLR, 2024.

[9] Sina Salek and Joseph Enguehard. Using the path of least resistance to explain deep networks. *arXiv preprint arXiv:2502.12108*, 2025.

[10] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[11] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

[12] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[13] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[14] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200-2011. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[15] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. URL http://arxiv.org/abs/1710.10196.

[17] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017. URL http://arxiv.org/abs/1706.03825.

[18] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[19] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020. URL https://arxiv.org/abs/2010.02502.

[20] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations Workshop*, 2014.

[21] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[22] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[23] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017.

[24] 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*, pages 1135–1144, 2016.

[25] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[26] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[27] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017.

[28] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[29] Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in neural information processing systems*, 33:1229–1239, 2020.

[30] Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. *arXiv preprint arXiv:2006.01272*, 2020.

[31] Muhammad Faaiz Taufiq, Patrick Blöbaum, and Lenon Minorics. Manifold restricted interventional shapley values. In *International Conference on Artificial Intelligence and Statistics*, pages 5079–5106. PMLR, 2023.

[32] Robert J Aumann and Lloyd S Shapley. *Values of non-atomic games*. Princeton University Press, 2015.

[33] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

[34] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

[35] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6021–6029, 2020.

Appendix Contents

This appendix gives the details that are not needed for the main narrative but are useful for checking the claims. We keep the main paper focused on the path-selection story. The appendix contains full proofs, the mathematical background behind the transport path, implementation details, metric definitions, additional diagnostics, and limitations.

A Related Work

Feature attribution has several classic lines. Gradient saliency methods explain a prediction by differentiating the output with respect to the input or an internal representation. Early saliency maps, Grad-CAM, SmoothGrad, and Guided Backpropagation are representative examples [20, 21, 17, 18]. These methods are efficient, but a local gradient does not by itself describe the finite change from a reference state to the input. Perturbation methods instead mask or modify input regions and measure the output change. Occlusion, meaningful perturbation, LIME, RISE, and SHAP-style methods belong to this family [22, 23, 24, 25, 3]. They connect more directly to counterfactual reasoning, but they must choose how to replace missing content. Backpropagation decomposition methods such as Layer-wise Relevance Propagation, Deep Taylor decomposition, and DeepLIFT propagate relevance through network layers using conservation or difference rules [26, 27, 28]. These methods motivate conservation of total relevance, but they do not solve the global path-selection problem studied here.

Shapley-value explanations inherit their appeal from cooperative game theory [1]. The Shapley value is unique for a finite game once the coalition value function is fixed. In model explanation, however, the coalition value function is not fixed by the predictor alone. Interventional, conditional, causal, and asymmetric Shapley variants encode different assumptions about feature absence and feature dependence [2, 3, 4, 29]. Manifold-restricted Shapley methods use generative or conditional models to keep perturbations closer to the data distribution [30, 31]. These works identify the missing-feature problem. We take a different route. We do not define all subset values. We formulate a continuous path game and ask how to choose the path that reveals the input.

Path attribution methods show that continuous paths can replace discrete coalitions. Integrated Gradients is the most widely used example and can be viewed as an Aumann-Shapley line integral along the straight path from a baseline to the input [5, 32]. Expected Gradients averages Integrated Gradients over a background distribution [6]. Guided Integrated Gradients and related adaptive methods alter the route in input space to reduce visual artifacts [7]. Recent geometric variants make the role of the path even more explicit. Manifold Integrated Gradients and Geodesic Integrated Gradients replace the straight path with a Riemannian geodesic under a learned data geometry or a model-induced geometry [8, 9]. We view these methods as close and complementary, not as targets that our method must dominate. Their path-selection object is an instance-level geodesic in a prescribed input-space metric, and in Geodesic Integrated Gradients this metric is induced by the explained model. Our path-selection object is different: it is a characteristic curve of a distribution-level least-action transport process from p_0 to p_1. Thus, we separate the geometry that chooses the path from the classifier whose score is explained. The generative transport model chooses the path, and the classifier gradient only allocates score change along that path. This changes the question from how to choose a better curve between two endpoints to how to select a generative transport process between two distributions.

Generative models provide structured transitions from simple priors to data-like samples. Diffusion models and DDIM define reverse-time generative trajectories, while continuous normalizing flows and flow matching define time-dependent vector fields [19, 13]. Rectified Flow learns straighter

trajectories by regressing velocities between coupled reference and data samples, and Reflow further rectifies the learned coupling [12]. These models are useful for attribution because they provide more structured trajectories than pixel-space interpolation. Still, a generative path is not automatically a principled explanation path. We use optimal transport to specify which generative path we want.

Optimal transport gives a geometric way to compare probability distributions [11]. In the dynamic Benamou-Brenier formulation, the squared Wasserstein-2 distance is the minimum kinetic action among all density paths and velocity fields that move one distribution to another [10]. This formulation matches our attribution problem. If explaining an input means measuring the model score along a transition from a reference distribution p_0 to a data distribution p_1, then the transition path should not be arbitrary. We use the Wasserstein-2 geodesic as the least-action path-selection principle.

Evaluation of attributions remains difficult. Deletion, insertion, faithfulness correlation, and sanity checks measure different aspects of explanation quality [33, 34, 35]. These metrics can favor noisy maps or penalize smooth maps depending on the perturbation protocol. We therefore report standard faithfulness metrics together with path-geometry diagnostics. Our Flow Consistency Error measures consistency with the learned vector field. It is not a certificate of true manifold membership.

B Proofs of main results

This section proves the formal statements used in the main text. We keep the assumptions visible because the main claim is a representation claim. The paper does not say that every attribution rule is forced without restrictions. It says that, once we restrict attention to coordinatewise path rules that use only the corresponding coordinate trace and that obey efficiency, linearity, dummy, and reparameterization invariance, the line integral is forced.

Lemma B.1 (Coordinate-trace representation)**.** *Fix a path γ and a coordinate i. Under the coordinate-trace determined part of Assumption 3.2, there is a functional L_i on the set of admissible scalar traces*

$$\mathcal{H}_i(\gamma) = \left\{ t \mapsto \frac{\partial f(\gamma(t))}{\partial x_i} \frac{\mathrm{d}\gamma_i(t)}{\mathrm{d}t} : f \in \mathcal{F} \right\}$$

such that

$$A_i(f, \gamma) = L_i\left(t \mapsto \frac{\partial f(\gamma(t))}{\partial x_i} \frac{\mathrm{d}\gamma_i(t)}{\mathrm{d}t} \right).$$

Moreover, if the attribution rule is linear in f, then L_i is linear on $\mathcal{H}_i(\gamma)$. If the attribution rule is continuous under uniform convergence of coordinate traces, then L_i is continuous for the uniform norm.

Proof. The coordinate-trace determined condition says exactly that two admissible scores with the same coordinate trace for coordinate i must receive the same coordinate-i attribution. Hence we can define $L_i(h)$ as $A_i(f, \gamma)$ for any $f \in \mathcal{F}$ whose coordinate trace equals h. This definition does not depend on which such f we choose. If h_f and h_g are traces generated by f and g, then the trace generated by $af + bg$ is $ah_f + bh_g$. Linearity of A_i in the score gives

$$L_i(ah_f + bh_g) = A_i(af + bg, \gamma) = aA_i(f, \gamma) + bA_i(g, \gamma) = aL_i(h_f) + bL_i(h_g).$$

This proves linearity of L_i. The continuity statement follows directly from the continuity clause in Assumption 3.2. □

Proof of Theorem 3.5. For each coordinate i, Lemma B.1 gives a linear functional L_i on the coordinate-trace space. We compare it with the ordinary integral functional $I(h) = \int_0^1 h(t)\,\mathrm{d}t$. Let $h \in \mathcal{H}_i(\gamma)$ be any admissible coordinate trace. By Assumption 3.3, there is an admissible score g whose coordinate-i trace is h and whose coordinate-j trace is zero for every $j \neq i$. The dummy property gives $A_j(g, \gamma) = L_j(0) = 0$ for every $j \neq i$. Efficiency applied to g gives

$$L_i(h) = \sum_{j=1}^{d} A_j(g, \gamma) = g(\gamma(1)) - g(\gamma(0)).$$

The chain rule along γ gives

$$g(\gamma(1)) - g(\gamma(0)) = \int_0^1 \nabla g(\gamma(t))^\top \frac{\mathrm{d}\gamma(t)}{\mathrm{d}t}\,\mathrm{d}t = \sum_{j=1}^{d} \int_0^1 \frac{\partial g(\gamma(t))}{\partial x_j} \frac{\mathrm{d}\gamma_j(t)}{\mathrm{d}t}\,\mathrm{d}t.$$

All terms except the i-th term vanish by the trace separation property, so

$$g(\gamma(1)) - g(\gamma(0)) = \int_0^1 h(t)\,\mathrm{d}t.$$

Therefore $L_i(h) = I(h)$ for every admissible coordinate trace h. Applying this identity to the trace $h_i^f(t) = \frac{\partial f(\gamma(t))}{\partial x_i} \frac{\mathrm{d}\gamma_i(t)}{\mathrm{d}t}$ of an arbitrary admissible score f, we obtain

$$A_i(f,\gamma) = L_i(h_i^f) = \int_0^1 \frac{\partial f(\gamma(t))}{\partial x_i} \frac{\mathrm{d}\gamma_i(t)}{\mathrm{d}t}\,\mathrm{d}t = \Phi_i(f,\gamma).$$

This proves uniqueness. □

The proof shows why we included Assumption 3.3. Without some richness or separation condition, efficiency can identify only the sum of all coordinate attributions. The theorem needs enough admissible scores to test each coordinate trace separately. This is common in axiomatic representation arguments, and here we state it explicitly rather than hiding it in the proof.

Proof of Theorem 4.4. Assumption 4.1 gives a unique kinetic-action minimizer $(\rho_t^\star, \boldsymbol{v}_t^\star)$ among admissible flows from p_0 to p_1. Therefore any rule in the class described by the theorem must choose the same distributional path and the same characteristic curve $\gamma^\star_{\boldsymbol{x}_1}$ for the endpoint $\boldsymbol{x}_1$. Once this path has been chosen, Theorem 3.5 applies to the fixed path $\gamma^\star_{\boldsymbol{x}_1}$. Hence the coordinate attribution must equal

$$\int_0^1 \frac{\partial f_c(\gamma^\star_{\boldsymbol{x}_1}(t))}{\partial x_i} \frac{\mathrm{d}\gamma^\star_{\boldsymbol{x}_1,i}(t)}{\mathrm{d}t}\,\mathrm{d}t,$$

which is Eq. (8). Conversely, the rule in Eq. (8) first selects the unique kinetic-action minimizer and then applies the Aumann-Shapley fixed-path rule, so it belongs to the stated class and satisfies the fixed-path axioms. This proves the characterization. □

Proof of Theorem 4.6. Let $\gamma(t)$ denote the ideal backward characteristic $\gamma^\star_{\boldsymbol{x}_1}(t)$, and let $\hat\gamma(t)$ denote the learned backward characteristic with the same terminal endpoint $\hat\gamma(1) = \gamma(1) = \boldsymbol{x}_1$. Write L_v for a common Lipschitz constant of $\boldsymbol{v}_t^\star$ and $\hat{\boldsymbol{v}}_t$ in $\boldsymbol{x}$ on the compact set $\mathcal{K}$. Write $B_g = \sup_{\boldsymbol{x}\in\mathcal{K}} \|\nabla f_c(\boldsymbol{x})\|_2$, L_g for the Lipschitz constant of ∇f_c on $\mathcal{K}$, and $B_v = \sup_{t,\boldsymbol{x}\in\mathcal{K}} \max\{\|\boldsymbol{v}_t^\star(\boldsymbol{x})\|_2, \|\hat{\boldsymbol{v}}_t(\boldsymbol{x})\|_2\}$. These constants are finite by Assumption 4.5 and compactness. We write $\boldsymbol{v}^\star_{t,i}(\boldsymbol{x})$ and $\hat{\boldsymbol{v}}_{t,i}(\boldsymbol{x})$ for the i-th coordinates of the ideal and learned vector fields.

For $t \le 1$, the two terminal-value ODEs give

$$\gamma(t) = \boldsymbol{x}_1 - \int_t^1 \boldsymbol{v}_s^\star(\gamma(s))\,\mathrm{d}s, \qquad \hat\gamma(t) = \boldsymbol{x}_1 - \int_t^1 \hat{\boldsymbol{v}}_s(\hat\gamma(s))\,\mathrm{d}s.$$

Taking the difference and using the triangle inequality gives

$$\|\gamma(t) - \hat\gamma(t)\|_2 \le \int_t^1 \|\boldsymbol{v}_s^\star(\gamma(s)) - \boldsymbol{v}_s^\star(\hat\gamma(s))\|_2\,\mathrm{d}s + \int_t^1 \|\boldsymbol{v}_s^\star(\hat\gamma(s)) - \hat{\boldsymbol{v}}_s(\hat\gamma(s))\|_2\,\mathrm{d}s.$$

The Lipschitz condition and the uniform vector-field error bound imply

$$\|\gamma(t) - \hat\gamma(t)\|_2 \le \int_t^1 L_v \|\gamma(s) - \hat\gamma(s)\|_2\,\mathrm{d}s + (1-t)\varepsilon.$$

The backward form of Gronwall's inequality gives

$$\sup_{t\in[0,1]} \|\gamma(t) - \hat\gamma(t)\|_2 \le e^{L_v}\varepsilon.$$

This is the only place where we use an ODE stability result. It says that a uniformly small error in the vector field produces a uniformly small error in the characteristic curve over a finite time interval.

For a fixed coordinate i, subtract the two attribution integrals:

$$|\Psi_i - \hat{\Psi}_i| \le \int_0^1 \left| \frac{\partial f_c(\gamma(t))}{\partial x_i} \boldsymbol{v}^\star_{t,i}(\gamma(t)) - \frac{\partial f_c(\hat{\gamma}(t))}{\partial x_i} \hat{\boldsymbol{v}}_{t,i}(\hat{\gamma}(t)) \right| \mathrm{d}t.$$

The integrand is bounded by

$$B_g \|\boldsymbol{v}^\star_t(\gamma(t)) - \hat{\boldsymbol{v}}_t(\hat{\gamma}(t))\|_2 + B_v \|\nabla f_c(\gamma(t)) - \nabla f_c(\hat{\gamma}(t))\|_2.$$

Using Lipschitz continuity and the vector-field error bound, we have

$$\|\boldsymbol{v}^\star_t(\gamma(t)) - \hat{\boldsymbol{v}}_t(\hat{\gamma}(t))\|_2 \le L_v \|\gamma(t) - \hat{\gamma}(t)\|_2 + \varepsilon \le (L_v e^{L_v} + 1)\varepsilon.$$

We also have

$$\|\nabla f_c(\gamma(t)) - \nabla f_c(\hat{\gamma}(t))\|_2 \le L_g \|\gamma(t) - \hat{\gamma}(t)\|_2 \le L_g e^{L_v} \varepsilon.$$

Combining the last three displays and integrating over a time interval of length one gives

$$|\Psi_i - \hat{\Psi}_i| \le \left[B_g (L_v e^{L_v} + 1) + B_v L_g e^{L_v} \right] \varepsilon.$$

The constant in brackets depends only on $\mathcal{K}$, the time horizon, the Lipschitz constants of the vector fields, and the first two derivative bounds of f_c. This proves Eq. (9). □

Proposition B.2 (Discrete efficiency residual)**.** *Let f have Hessian norm bounded by M on a compact set containing the discrete path $\hat{\boldsymbol{x}}_0, \dots, \hat{\boldsymbol{x}}_K$. Let $\Delta\hat{\boldsymbol{x}}_k = \hat{\boldsymbol{x}}_{k+1} - \hat{\boldsymbol{x}}_k$, and define the discrete attribution by Eq.* (11)*. Then*

$$\left| \sum_{i=1}^{d} \hat{\Psi}_i - (f(\hat{\boldsymbol{x}}_K) - f(\hat{\boldsymbol{x}}_0)) \right| \le \frac{M}{2} \sum_{k=0}^{K-1} \|\Delta\hat{\boldsymbol{x}}_k\|_2^2.$$

If the path increments satisfy $\|\Delta\hat{\boldsymbol{x}}_k\|_2 \le V/K$, then the residual is at most $MV^2/(2K)$.

Proof. Taylor's theorem with remainder gives, for each segment of the discrete path,

$$f(\hat{\boldsymbol{x}}_{k+1}) = f(\hat{\boldsymbol{x}}_k) + \nabla f(\hat{\boldsymbol{x}}_k)^\top \Delta\hat{\boldsymbol{x}}_k + r_k, \qquad |r_k| \le \frac{M}{2} \|\Delta\hat{\boldsymbol{x}}_k\|_2^2.$$

Summing this identity from $k = 0$ to $K - 1$ makes the left side telescope:

$$f(\hat{\boldsymbol{x}}_K) - f(\hat{\boldsymbol{x}}_0) = \sum_{k=0}^{K-1} \nabla f(\hat{\boldsymbol{x}}_k)^\top \Delta\hat{\boldsymbol{x}}_k + \sum_{k=0}^{K-1} r_k.$$

The first sum is exactly the sum over coordinates of Eq. (11). Taking absolute values and applying the remainder bound gives the first claim. If $\|\Delta\hat{\boldsymbol{x}}_k\|_2 \le V/K$, then $\sum_k \|\Delta\hat{\boldsymbol{x}}_k\|_2^2 \le K(V/K)^2 = V^2/K$, which gives the second claim. □

Proposition B.3 (Additive scores)**.** *Suppose $f(\boldsymbol{x}) = b + \sum_{i=1}^d f_i(x_i)$ and the path γ connects $\boldsymbol{x}_0$ to $\boldsymbol{x}_1$. Then the Aumann-Shapley attribution along any continuously differentiable path is*

$$\Phi_i(f, \gamma) = f_i(x_{1,i}) - f_i(x_{0,i}).$$

Proof. For an additive score, $\partial_i f(\boldsymbol{x}) = f_i'(x_i)$. Therefore

$$\Phi_i(f, \gamma) = \int_0^1 f_i'(\gamma_i(t)) \frac{\mathrm{d}\gamma_i(t)}{\mathrm{d}t} \, \mathrm{d}t.$$

The one-dimensional chain rule gives $\frac{\mathrm{d}}{\mathrm{d}t} f_i(\gamma_i(t)) = f_i'(\gamma_i(t)) \frac{\mathrm{d}\gamma_i(t)}{\mathrm{d}t}$. Hence

$$\Phi_i(f, \gamma) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} f_i(\gamma_i(t)) \, \mathrm{d}t = f_i(\gamma_i(1)) - f_i(\gamma_i(0)).$$

Since $\gamma_i(1) = x_{1,i}$ and $\gamma_i(0) = x_{0,i}$, the claim follows. □

C Optimal transport details and the meaning of geodesic

This section explains the transport objects used in the main text. The purpose is not to develop optimal transport from scratch. We only need the dynamic viewpoint that connects a probability path, a velocity field, and kinetic action.

C.1 From static couplings to dynamic paths

A coupling between p_0 and p_1 is a joint law π on pairs $(\boldsymbol{x}_0, \boldsymbol{x}_1)$ whose first marginal is p_0 and whose second marginal is p_1. Let $\Pi(p_0, p_1)$ denote the set of all such couplings, and let $W_2(p_0, p_1)$ denote the quadratic Wasserstein distance. The quadratic-cost optimal transport problem searches over this set:

$$W_2^2(p_0, p_1) = \inf_{\pi \in \Pi(p_0, p_1)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\boldsymbol{x}_1 - \boldsymbol{x}_0\|_2^2 \, \mathrm{d}\pi(\boldsymbol{x}_0, \boldsymbol{x}_1).$$

When an optimal coupling $\pi^\star$ is available, it induces the linear interpolation

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1, \qquad (\boldsymbol{x}_0, \boldsymbol{x}_1) \sim \pi^\star.$$

The law of $\boldsymbol{x}_t$ is a distribution ρ_t. This path of laws is the constant-speed Wasserstein geodesic under standard regularity conditions. In the deterministic Monge case, where $\boldsymbol{x}_1 = T(\boldsymbol{x}_0)$ for an optimal transport map T, each sample moves along a straight segment from $\boldsymbol{x}_0$ to $T(\boldsymbol{x}_0)$. In general, the geodesic statement is distributional, not a statement that every point along every segment is a natural image.

C.2 Kinetic action and the Benamou-Brenier view

The dynamic formulation used in Eq. (6) replaces couplings by probability-law paths and velocity fields. A pair $(\rho, \boldsymbol{v})$, with $\rho = (\rho_t)_{t\in[0,1]}$ and $\boldsymbol{v} = (\boldsymbol{v}_t)_{t\in[0,1]}$, is admissible when it satisfies the continuity equation in Eq. (3). This equation means that probability mass moves according to the velocity field and is neither created nor destroyed. The action

$$\mathcal{A}(\rho, \boldsymbol{v}) = \int_0^1 \int_{\mathbb{R}^d} \|\boldsymbol{v}_t(\boldsymbol{x})\|_2^2 \, \mathrm{d}\rho_t(\boldsymbol{x}) \, \mathrm{d}t$$

measures the average squared speed of the mass over time. The Benamou-Brenier formula says that the minimum action over all admissible pairs equals $W_2^2(p_0, p_1)$ [10, 11]. This is why we call the selected path a geodesic path: it is the least-action path between distributions in the Wasserstein geometry.

Proposition C.1 (Action of an optimal displacement interpolation)**.** *Assume that the optimal coupling is induced by a map T, so $\boldsymbol{x}_1 = T(\boldsymbol{x}_0)$ and $\boldsymbol{x}_0 \sim p_0$. Define $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + tT(\boldsymbol{x}_0)$. Then the associated constant velocity along each particle is $T(\boldsymbol{x}_0) - \boldsymbol{x}_0$, and the kinetic action equals the quadratic transport cost:*

$$\int_0^1 \mathbb{E}\|T(\boldsymbol{x}_0) - \boldsymbol{x}_0\|_2^2 \, \mathrm{d}t = \mathbb{E}\|T(\boldsymbol{x}_0) - \boldsymbol{x}_0\|_2^2.$$

If T is optimal, this value equals $W_2^2(p_0, p_1)$.

Proof. For each fixed starting point $\boldsymbol{x}_0$, differentiating $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + tT(\boldsymbol{x}_0)$ with respect to t gives $\mathrm{d}\boldsymbol{x}_t/\mathrm{d}t = T(\boldsymbol{x}_0) - \boldsymbol{x}_0$. This velocity is constant in time along that particle. Therefore the particle action over $[0, 1]$ is

$$\int_0^1 \left\| \frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} \right\|_2^2 \mathrm{d}t = \int_0^1 \|T(\boldsymbol{x}_0) - \boldsymbol{x}_0\|_2^2 \, \mathrm{d}t = \|T(\boldsymbol{x}_0) - \boldsymbol{x}_0\|_2^2.$$

Taking expectation over $\boldsymbol{x}_0 \sim p_0$ gives the displayed identity. If T is the optimal quadratic-cost transport map, the expected squared displacement is the definition of $W_2^2(p_0, p_1)$ in the Monge formulation. □

This proposition clarifies why straightness matters. A W2 geodesic has constant-speed displacement interpolation at the ideal level. Rectified Flow and Reflow are useful because they try to learn trajectories that are closer to such straight displacement paths than generic generative trajectories. This connection is approximate in high-dimensional image spaces. We therefore evaluate path action and stability empirically rather than claiming that the neural flow exactly recovers the optimal map.

C.3 Why geodesic does not mean strictly on-manifold

The phrase "on-manifold" can mean different things. A strict version would assume a low-dimensional set $\mathcal{M} \subset \mathbb{R}^d$, a data distribution supported on $\mathcal{M}$, and a certificate that every path point $\gamma(t)$ belongs to $\mathcal{M}$. We do not make this assumption and we do not prove such a certificate.

Our statement is distributional. We choose a flow whose time marginals ρ_t move from p_0 to p_1. At intermediate times, ρ_t is generally neither p_0 nor p_1. If p_0 is a Gaussian prior and p_1 is an image distribution, intermediate states are generated states along a transport bridge. They may look more structured than arbitrary pixel interpolation, but this visual fact is not a mathematical proof of data-manifold membership.

For this reason, the main text uses the terms transport-consistent and generative. Flow Consistency Error checks whether a numerical path follows the learned vector field. It does not check whether the true data density is high at every intermediate point. This distinction makes the claim weaker but much more accurate: we replace heuristic paths with a variationally selected transport path, not with a certified manifold path.

C.4 Endpoint conditioning

The ideal definition in Eq. (7) conditions on a target endpoint $\boldsymbol{x}_1$. In an exact deterministic transport map, this endpoint has a unique preimage $\boldsymbol{x}_0$ under the flow map, except on sets where the map is not invertible. In a learned ODE model, we approximate this preimage by integrating the learned vector field backward from $\boldsymbol{x}_1$ to time zero. We then integrate forward from the obtained $\hat{\boldsymbol{x}}_0$ to store a stable numerical trajectory. This backward-forward procedure keeps the endpoint tied to the input being explained and avoids sampling an unrelated reference point.

When the learned flow is imperfect, the backward-forward trajectory may not return exactly to $\boldsymbol{x}_1$. In implementation we either use the stored backward trajectory in reverse order or correct the final point by setting $\hat{\boldsymbol{x}}_K = \boldsymbol{x}_1$ before the last gradient accumulation. Both choices preserve the intended interpretation: the attribution explains the score difference between the recovered reference endpoint and the observed input. The completeness residual in Table 1 reports the remaining numerical integration error.

D Implementation details

This section describes how we instantiate the ideal transport-geodesic attribution with a learned flow. The main point is that all baselines are evaluated through the same path-integral form whenever they define a path. This keeps the comparison focused on the path rather than on a different allocation formula.

D.1 Learned vector field and Reflow

We train a time-dependent vector field $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ by the Rectified Flow objective in Eq. (10). As in the main text, $\hat{\boldsymbol{v}}_t(\boldsymbol{x}) = \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ denotes the learned field when we view it as a time-indexed vector field. The objective uses pairs $(\boldsymbol{z}_0, \boldsymbol{z}_1)$ drawn from a coupling π. For the first Rectified Flow, π is usually the independent coupling between the reference prior and the data distribution. For Reflow, we first run the learned flow from prior samples to generated endpoints, then use the induced pairs to train a new vector field. This changes the coupling used by the regression problem.

We use Reflow because the first independent coupling can contain crossings and unnecessary displacement. Reflow tends to reduce the transport cost of the induced coupling and to straighten trajectories [12]. In our paper, this is an approximation strategy rather than an exact theorem that the learned model reaches the optimal transport map. The ideal object remains the kinetic-action minimizer in Eq. (6); the neural vector field is the computable approximation.

D.2 Reference endpoint for a given input

For an observed input $\boldsymbol{x}_1$, we need a reference endpoint tied to this input. We obtain it by solving the learned ODE backward from time one to time zero:

$$\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t), \qquad \boldsymbol{x}_{t=1} = \boldsymbol{x}_1.$$

The result is $\hat{\boldsymbol{x}}_0$. We then integrate forward from $\hat{\boldsymbol{x}}_0$ and store the states $\hat{\boldsymbol{x}}_0, \ldots, \hat{\boldsymbol{x}}_K$ on the uniform grid $t_k = k/K$, with $\Delta t = 1/K$. Using the same learned field in both directions keeps the reference and the input connected by one trajectory. This differs from Expected Gradients or GradientSHAP, where the reference is sampled independently from a background distribution.

The backward integration can be numerically imperfect when the learned vector field is stiff or inaccurate. To avoid making the attribution depend on a small terminal mismatch, we compute the reported score difference using the actual endpoints used by the stored trajectory. When we compare completeness against $f_c(\boldsymbol{x}_1) - f_c(\hat{\boldsymbol{x}}_0)$, we set the last stored point to the observed input. This choice makes the explanation target explicit.

D.3 Numerical integration

The discrete estimator in Eq. (11) is a left Riemann estimator for the path integral on the grid $t_k = k/K$. We chose it because it is simple and matches the finite-difference view of adding small path increments. A midpoint rule can reduce the quadrature error when we can afford extra gradient evaluations. Proposition B.2 shows that the completeness residual is controlled by the squared path increments when the predictor has bounded Hessian on the visited region.

In practice, one explanation with K steps requires storing $K + 1$ states and computing K or $K + 1$ gradients of the target score, depending on the quadrature rule. We use the target logit before the softmax rather than the probability after the softmax. Logits avoid saturation effects and are standard in attribution experiments. We aggregate pixel-level RGB attributions by summing or taking the channelwise absolute value depending on the visualization protocol. For deletion metrics, we use the signed attribution score to rank features when the target is a score increase and use absolute scores only when the baseline method is defined as unsigned.

D.4 Baselines

We compare against local gradient smoothing methods, backpropagation decomposition methods, Shapley-style randomized baselines, straight-line path methods, and generative path methods. SmoothGrad averages gradients under small input noise. Guided Backpropagation modifies the backward pass through ReLU layers. GradientSHAP samples references and interpolation coefficients. Integrated Gradients uses the straight path from a fixed reference to the input. DDIM supplies a deterministic generative trajectory from a diffusion model. Our method uses the learned Rectified Flow or Reflow trajectory and then applies the same Aumann-Shapley coordinate integral.

For methods that do not define a continuous path, GPS and FCE are not meaningful, so the corresponding entries are marked with a dash in Table 3. For methods that define a path but not through the Rectified Flow vector field, FCE measures mismatch with our learned flow. This is useful as a path-dynamics diagnostic but should not be interpreted as a universal measure of explanation quality.

D.5 Computational cost

Exact discrete Shapley values require evaluating an exponential number of coalitions if each input coordinate is treated as a player. Practical Shapley explainers reduce this cost by sampling coalitions or grouping features, but they still need a missing-feature model. Our method has cost linear in the number of path steps. The main cost is the repeated gradient evaluation of the fixed predictor along the path. The generative ODE cost is separate and depends on the solver and the vector field architecture.

This cost profile makes the method closer to Integrated Gradients than to exact Shapley. The key difference is not the line integral itself. The key difference is the path used by the line integral. Integrated Gradients chooses a straight pixel path. We choose a learned transport path that approximates the kinetic-action principle.

E Evaluation metrics and protocols

Attribution metrics measure different properties, and no single metric proves explanation quality. We therefore separate four questions. First, we check whether a sampled path is geometrically short in the ambient space. Second, we check whether the sampled path follows the learned vector field. Third, we check whether the attribution map is visually structured. Fourth, we check whether important pixels affect the target score under deletion. Throughout this section, the discrete path is $\hat{\boldsymbol{x}}_0, \ldots, \hat{\boldsymbol{x}}_K$, the grid is uniform with $t_k = k/K$, and $\Delta t = 1/K$. We write $\hat{\boldsymbol{v}}_{t_k}(\hat{\boldsymbol{x}}_k) = \boldsymbol{v}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_k, t_k)$ for the learned velocity evaluated at the sampled state.

E.1 Geometric path straightness

For a discrete path $\hat{\boldsymbol{x}}_0, \ldots, \hat{\boldsymbol{x}}_K$, the Geometric Path Straightness score is

$$\mathrm{GPS} = \frac{\sum_{k=0}^{K-1} \|\hat{\boldsymbol{x}}_{k+1} - \hat{\boldsymbol{x}}_k\|_2}{\|\hat{\boldsymbol{x}}_K - \hat{\boldsymbol{x}}_0\|_2}.$$

The denominator is the endpoint displacement and the numerator is the discrete path length. The score is at least one by the triangle inequality whenever $\hat{\boldsymbol{x}}_K \neq \hat{\boldsymbol{x}}_0$. A value close to one means that the sampled path is close to a straight segment in ambient Euclidean space. This does not by itself imply optimal transport. It only checks one geometric consequence of low-curvature displacement interpolation.

Integrated Gradients has GPS exactly one when the path is the straight line. This is why GPS must be interpreted together with FCE and faithfulness metrics. A straight pixel path can have excellent GPS while still being a poor generative transition.

E.2 Flow Consistency Error

The Flow Consistency Error is

$$\mathrm{FCE} = \frac{1}{K} \sum_{k=0}^{K-1} \left\| \frac{\hat{\boldsymbol{x}}_{k+1} - \hat{\boldsymbol{x}}_k}{\Delta t} - \hat{\boldsymbol{v}}_{t_k}(\hat{\boldsymbol{x}}_k) \right\|_2^2.$$

This metric compares the finite-difference velocity of the sampled path with the velocity predicted by the learned flow. Low FCE means that the path is dynamically consistent with the learned vector-field family $\hat{\boldsymbol{v}}_t$. It does not mean that the path lies on a true data manifold. If a method uses the same vector field to generate the path, FCE can be very small. We therefore treat it as a diagnostic for path dynamics, not as the main faithfulness metric.

For methods that do not define a learned vector field, such as SmoothGrad or Guided Backpropagation, FCE is not applicable. For straight-line Integrated Gradients, we can still evaluate the straight path against the learned vector-field family $\hat{\boldsymbol{v}}_t$ to show that the straight path is not a characteristic curve of the learned transport field. This is why the main table reports large FCE values for IG.

E.3 Structure-aware total variation

Let Ω be the two-dimensional pixel grid of an image. Let I denote the input image, and let $\boldsymbol{\phi}$ denote the attribution map after summing or averaging over color channels. We write ∇_{img} for finite differences on the image grid. For a pixel $(u, v) \in \Omega$, $\nabla_{\mathrm{img}} \boldsymbol{\phi}_{u,v}$ is the local finite-difference gradient of the attribution map and $\nabla_{\mathrm{img}} I_{u,v}$ is the corresponding image gradient. The Structure-Aware Total Variation diagnostic is

$$\mathrm{SATV}(\boldsymbol{\phi}) = \sum_{(u,v)\in\Omega} \|\nabla_{\mathrm{img}} \boldsymbol{\phi}_{u,v}\|_1 \exp\left(-\alpha \|\nabla_{\mathrm{img}} I_{u,v}\|_2\right),$$

where $\alpha > 0$ controls how strongly the penalty is reduced near image edges. We set $\alpha = 10$ in all experiments. The exponential weight reduces the penalty near image edges and keeps the penalty high in flat regions. As a result, SATV penalizes high-frequency attribution noise where the image itself has little structure.

SATV is only a structure diagnostic. A very smooth but unfaithful attribution can score well under SATV. We therefore read SATV together with deletion metrics and completeness checks rather than using it as a standalone measure of explanation quality.

E.4 Edge alignment score

The Edge Alignment Score measures whether changes in the attribution map occur near image edges. We define image edge strength and attribution edge strength by

$$e_{u,v} = \|\nabla_{\text{img}} I_{u,v}\|_2, \qquad s_{u,v} = \|\nabla_{\text{img}} \boldsymbol{\phi}_{u,v}\|_1.$$

The reported score is the normalized weighted average

$$\text{EAS}(\boldsymbol{\phi}, I) = \frac{\sum_{(u,v)\in\Omega} s_{u,v} e_{u,v}}{\sum_{(u,v)\in\Omega} s_{u,v} + 10^{-12}}.$$

A higher score means that attribution-map edges are more aligned with image edges. This is again a structural diagnostic rather than a causal test. Smooth maps can have high edge alignment if they concentrate changes near object boundaries, while noisy maps can have low edge alignment because their gradients are spread across flat regions.

E.5 Deletion metrics

Deletion evaluates whether removing highly attributed pixels reduces the target score. We sort pixels by attribution magnitude or signed positive attribution, depending on the method's output convention. We then replace the top-ranked pixels progressively with either a zero value or a blurred value. At each deletion fraction, we evaluate the target logit. The deletion score is the area under the resulting score curve. Lower is better because a faithful positive attribution should identify pixels whose removal quickly decreases the target score.

Deletion depends on the perturbation operator. Zero deletion can create unnatural black patches, while blur deletion can preserve low-frequency image statistics but still change the data distribution. We report both variants because agreement between them is more informative than either one alone. We do not claim that deletion is a complete evaluation of explanation quality.

E.6 Completeness residual

For a path attribution rule, completeness means that the attribution sum equals the score difference. In continuous time, the Aumann-Shapley integral satisfies completeness exactly by the chain rule. In discrete time, numerical quadrature creates a residual. For N evaluated examples, we report

$$R_{\text{eff}} = \frac{1}{N} \sum_{j=1}^{N} \left| \sum_{i=1}^{d} \hat{\Psi}_i^{(j)} - \left(f_c(\boldsymbol{x}_1^{(j)}) - f_c(\hat{\boldsymbol{x}}_0^{(j)}) \right) \right|.$$

Here $\hat{\Psi}_i^{(j)}$ is the discrete attribution for coordinate i on example j, $\boldsymbol{x}_1^{(j)}$ is the observed endpoint, and $\hat{\boldsymbol{x}}_0^{(j)}$ is the learned reference endpoint obtained by backward integration. Table 1 reports this residual for different integration step counts. Proposition B.2 explains why the residual decreases when the path discretization becomes finer.

E.7 Stability metrics

We measure stability across random seeds by comparing attribution maps generated by independently trained flows. Pixel variance is the average pointwise variance of normalized attribution maps. SSIM measures perceptual similarity between pairs of normalized maps. Rank correlation measures whether the feature ordering induced by attribution scores is stable. These metrics answer a different question from deletion. A method can be faithful but unstable if small changes to the path alter the feature ranking. Our path-stability experiments test the hypothesis that lower-action and lower-curvature paths reduce this variability.

F Additional results and visualizations

This section reports supplementary experiments. We include them to support the main claims without changing the main message. The paper does not claim that a learned path is certified to stay on a data manifold. The message is that a better transport path gives a more stable and more structured attribution integral.

F.1 Synthetic additive sanity check

We first check the additive case in Proposition B.3. We sample $d = 10$ dimensional inputs with independent coordinates from $[-\pi, \pi]^d$ and use

$$f(\boldsymbol{x}) = \sum_{i=1}^{d} \sin(x_i), \qquad \boldsymbol{x}_0 = \mathbf{0}.$$

The exact coordinate contribution is $\sin(x_i)$. We compute the path integral along the straight path from $\boldsymbol{x}_0$ to $\boldsymbol{x}$. This experiment does not use a learned flow. It isolates the numerical accuracy of the Aumann-Shapley estimator.

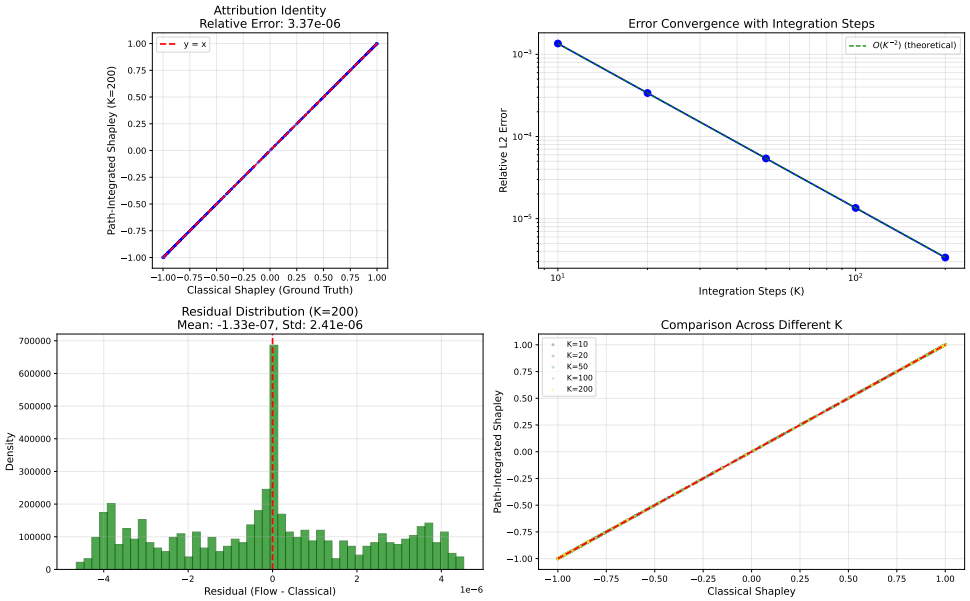

Figure 2: Synthetic additive sanity check. The path-integral estimator recovers the analytical additive Shapley values, and the error decreases as the quadrature grid becomes finer.

Figure 2 shows near-identity alignment between analytical values and numerical attributions. This supports the claim that our method reduces to the classical additive allocation when interactions vanish.

F.2 Controlled Gaussian transport experiment

We next use Gaussian-to-Gaussian transport because the quadratic optimal transport map is available in closed form. This gives an oracle against which we can compare one-step Rectified Flow and Reflow. We use this experiment only as a controlled diagnostic. It does not prove that high-dimensional image flows exactly recover optimal transport.

Let $\mathcal{A}^\star = W_2^2(p_0, p_1)$ denote the oracle kinetic action. For a learned method m, let $\widehat{\mathcal{A}}_m$ be its empirical discrete action,

$$\widehat{\mathcal{A}}_m = \sum_{k=0}^{K-1} \frac{\|\hat{\boldsymbol{x}}_{k+1}^m - \hat{\boldsymbol{x}}_k^m\|_2^2}{\Delta t}.$$

The reported action gap is the relative excess action

$$\Delta A_m = \frac{\widehat{\mathcal{A}}_m - \mathcal{A}^\star}{\mathcal{A}^\star + 10^{-12}}.$$

For the oracle, $\Delta A = 0$. Let $\boldsymbol{v}_t^\star$ be the oracle velocity field and let $\hat{\boldsymbol{v}}_t^m$ be the learned velocity field for method m. The relative field error is

$$\mathrm{RFE}_m = \frac{\left(\sum_{k=0}^{K-1} \|\hat{\boldsymbol{v}}_{t_k}^m(\hat{\boldsymbol{x}}_k^m) - \boldsymbol{v}_{t_k}^\star(\hat{\boldsymbol{x}}_k^m)\|_2^2\right)^{1/2}}{\left(\sum_{k=0}^{K-1} \|\boldsymbol{v}_{t_k}^\star(\hat{\boldsymbol{x}}_k^m)\|_2^2\right)^{1/2} + 10^{-12}}.$$

The curvature proxy is the squared finite-difference acceleration integrated over the path,

$$\mathrm{Curv}_m = \sum_{k=1}^{K-1} \left\| \frac{\hat{\boldsymbol{x}}_{k+1}^m - 2\hat{\boldsymbol{x}}_k^m + \hat{\boldsymbol{x}}_{k-1}^m}{\Delta t^2} \right\|_2^2 \Delta t.$$

These definitions make the table interpretable. ΔA measures excess kinetic cost, RFE measures vector-field mismatch against the oracle, and Curv measures unnecessary bending of the sampled path.

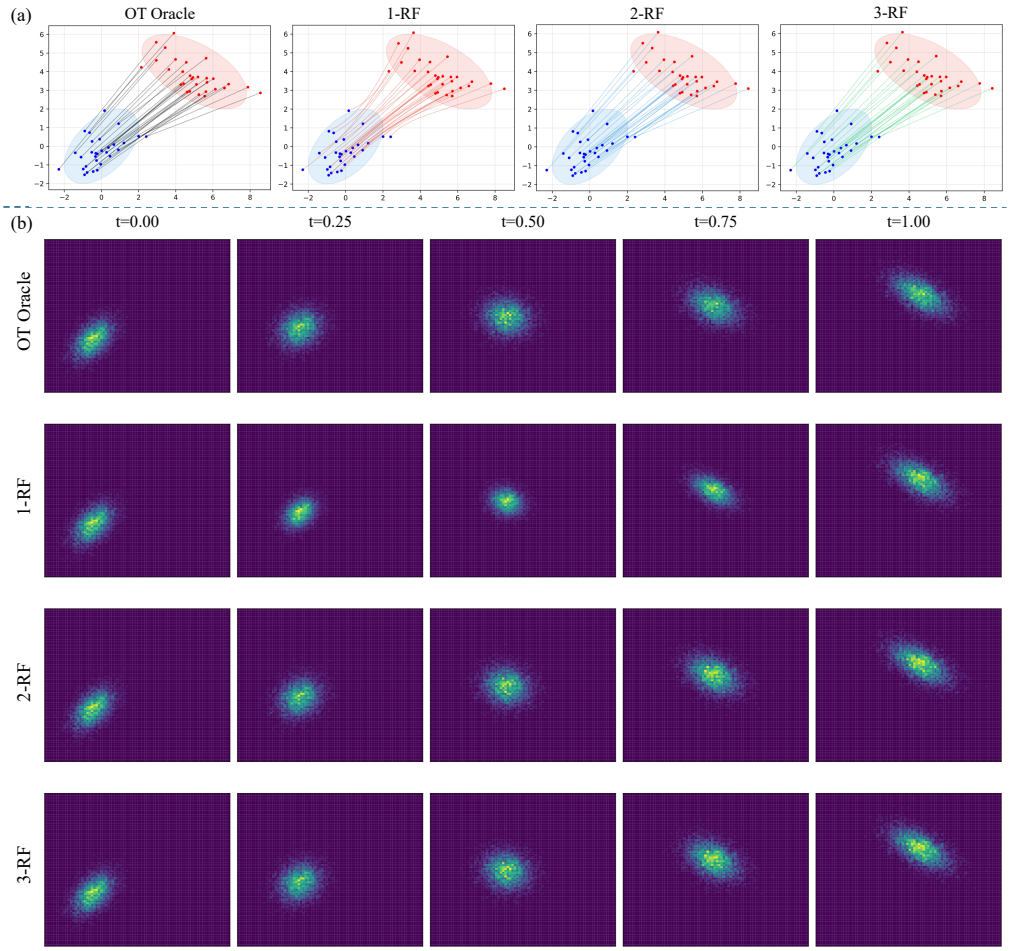

Figure 3: Controlled two-dimensional Gaussian transport. The oracle displacement interpolation gives the reference geometry. One-step Rectified Flow can be visibly more curved, while Reflow straightens the trajectories and makes the intermediate densities closer to the oracle interpolation.

The qualitative result in Figure 3 explains why Reflow is useful for attribution. Since the attribution integral samples gradients along the path, unnecessary path curvature can expose the integral to unstable regions of the predictor. Reflow reduces this curvature in the controlled setting.

Metric	OT oracle	1-RF	2-RF	3-RF
$W_2^2(p_0, p_1)$	27.8976	-	-	-
Action gap ΔA	0	0.2205 ± 0.0033	0.00189 ± 0.00144	0.00199 ± 0.00141
Relative field error	0	0.2933 ± 0.0050	0.00284 ± 0.00014	0.00276 ± 0.00008
Curvature proxy	$3.45\times10^{-5} \pm 2.63\times10^{-7}$	0.0763 ± 0.00077	$4.85\times10^{-4} \pm 1.50\times10^{-5}$	$4.35\times10^{-4} \pm 1.31\times10^{-5}$

Table 4: Controlled Gaussian transport in $d = 10$, reported as mean $\pm$ standard deviation over five seeds. Reflow reduces the relative action gap, relative field error, and curvature proxy compared with one-step Rectified Flow.

Table 4 supports the same conclusion quantitatively. The one-step flow has a nontrivial action gap and field mismatch relative to the oracle. After one Reflow iteration, these gaps decrease by roughly two orders of magnitude in this controlled setting. This is the empirical basis for using Reflow as a closer approximation to the geodesic ideal.

F.3 Attribution error versus field mismatch

Theorem 4.6 predicts that attribution error should decrease when the learned field approaches the ideal field on the relevant region. We test this prediction in the Gaussian setting where the oracle is known, and also across flow checkpoints where a strong checkpoint serves as a practical oracle.

Let $\mathbf{\Psi}^{\mathrm{ref}}$ be the attribution produced by the oracle or reference checkpoint, and let $\hat{\mathbf{\Psi}}^m$ be the attribution produced by method or checkpoint m. The relative attribution error is

$$\mathrm{RAE}_m = \frac{\|\hat{\mathbf{\Psi}}^m - \mathbf{\Psi}^{\mathrm{ref}}\|_2}{\|\mathbf{\Psi}^{\mathrm{ref}}\|_2 + 10^{-12}}.$$

When an oracle vector field is available, the empirical flow mismatch is the relative field error RFE_m defined above. When only checkpoints are available, we replace $\boldsymbol{v}_t^\star$ by the velocity field of the reference checkpoint in the same formula. This distinction matters because the checkpoint experiment measures convergence to a practical reference, while the Gaussian experiment measures error against the known OT field.

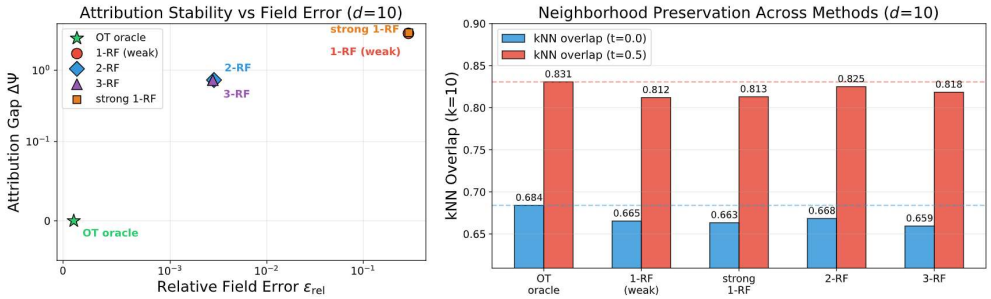

Figure 4: High-dimensional controlled diagnostic. The left panel relates relative attribution error to relative field mismatch. The right panel reports neighborhood preservation along intermediate transport states. The result supports the stability view: better field alignment gives smaller attribution discrepancy.

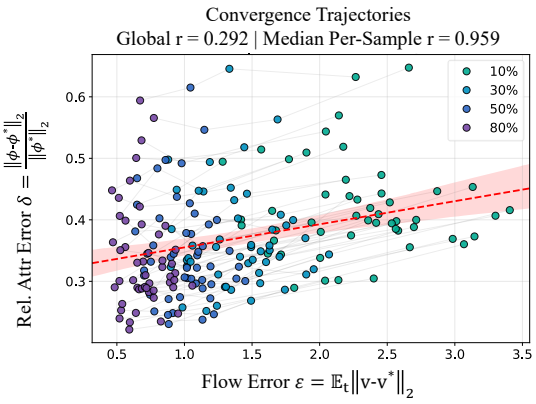

Figure 5: Attribution convergence across flow checkpoints. The trend is consistent with the stability theorem: as the learned vector field approaches the reference checkpoint, the attribution gap decreases.

Figures 4 and 5 are not meant to show a universal linear law. They show that, in our controlled diagnostics, attribution error tracks vector-field mismatch in the direction predicted by Theorem 4.6.

F.4 Action diagnostics

We also visualize representative action estimates along sampled paths. Lower action does not automatically imply better explanation under every metric, but high unnecessary action often means that the path moves more than needed before reaching the same endpoint. Such movement can make the line integral more sensitive to local gradient variation.

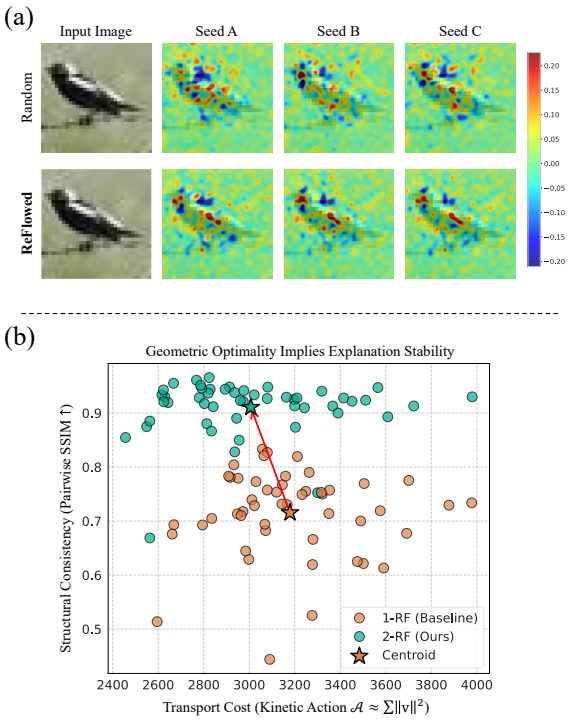

Figure 6: Representative action diagnostic for sampled paths. The plot illustrates how path movement and action are measured along a numerical trajectory.

F.5 Additional qualitative examples

The following figures show additional randomly selected attribution examples. We include them to make the qualitative claim easier to inspect. The figures should not be read as proof of faithfulness. They are visual evidence that the learned transport path tends to give more spatially coherent maps than straight-line interpolation on the shown samples.

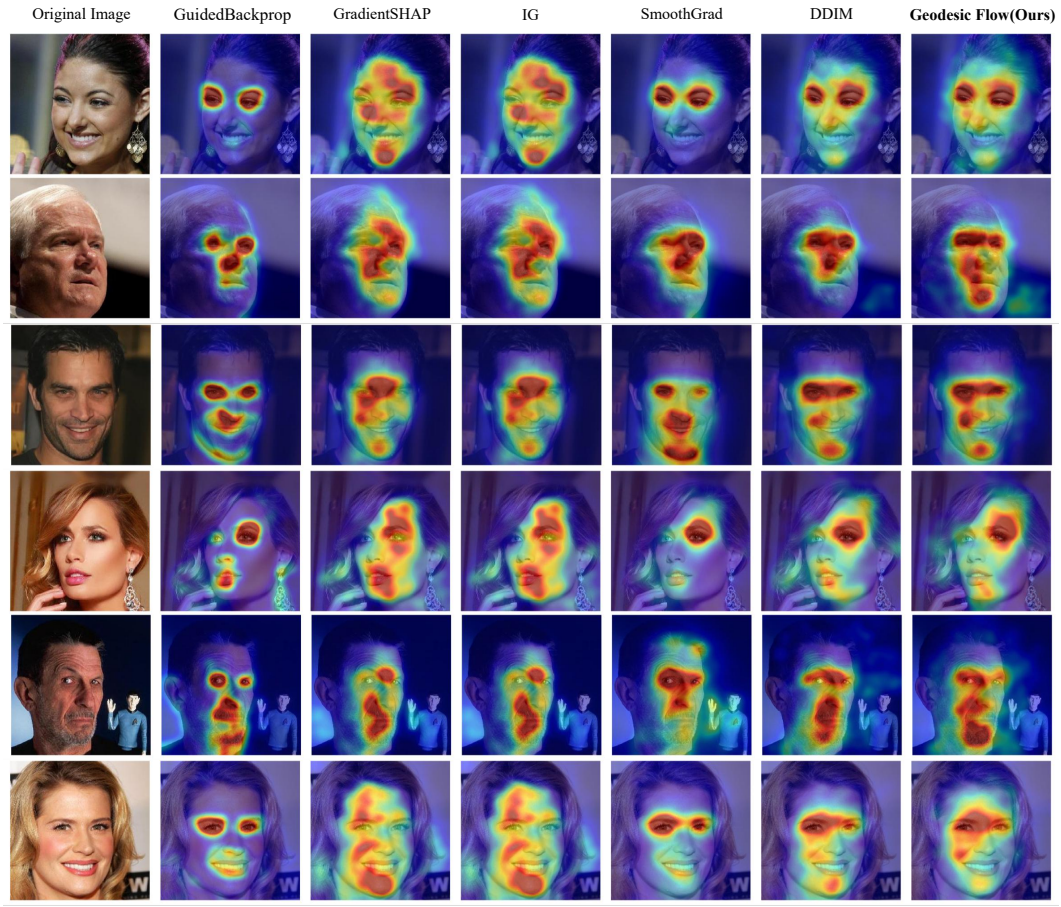

Figure 7: Additional CelebA-HQ examples, part I.

Figure 8: Additional CelebA-HQ examples, part II.

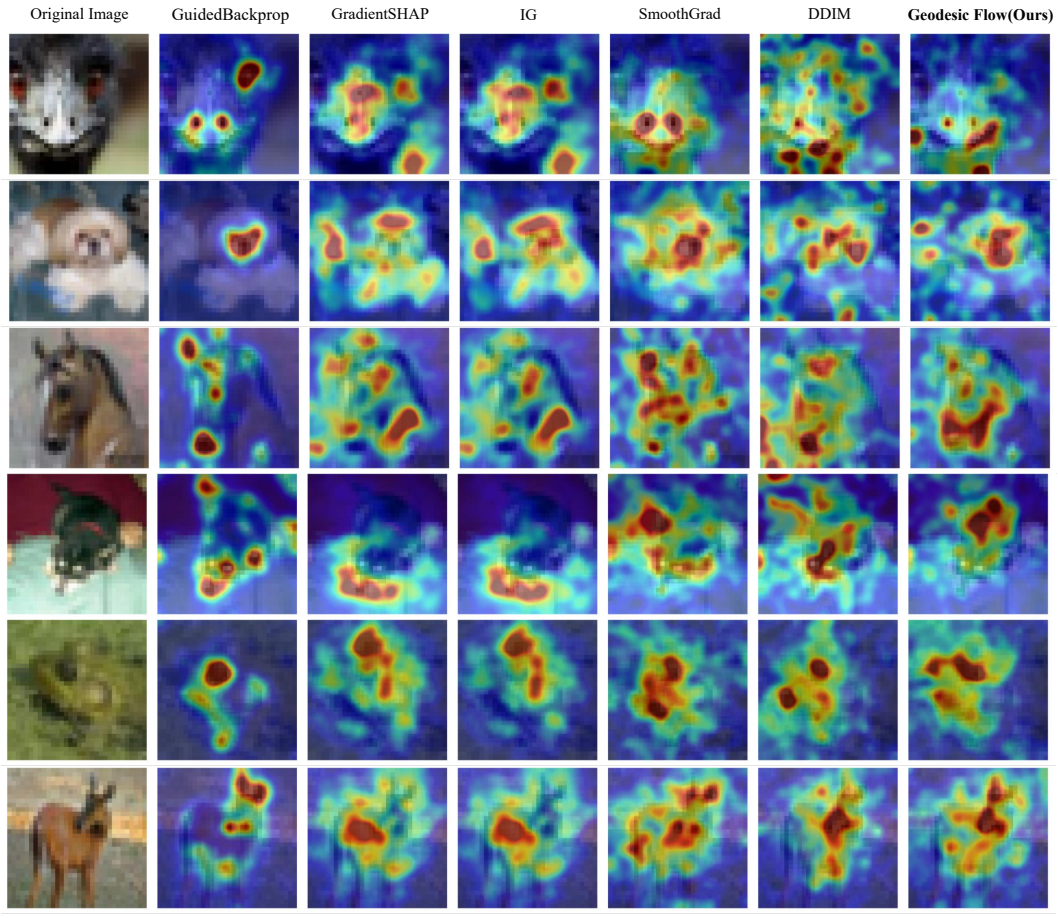

Figure 9: Additional CIFAR-10 examples, part I.

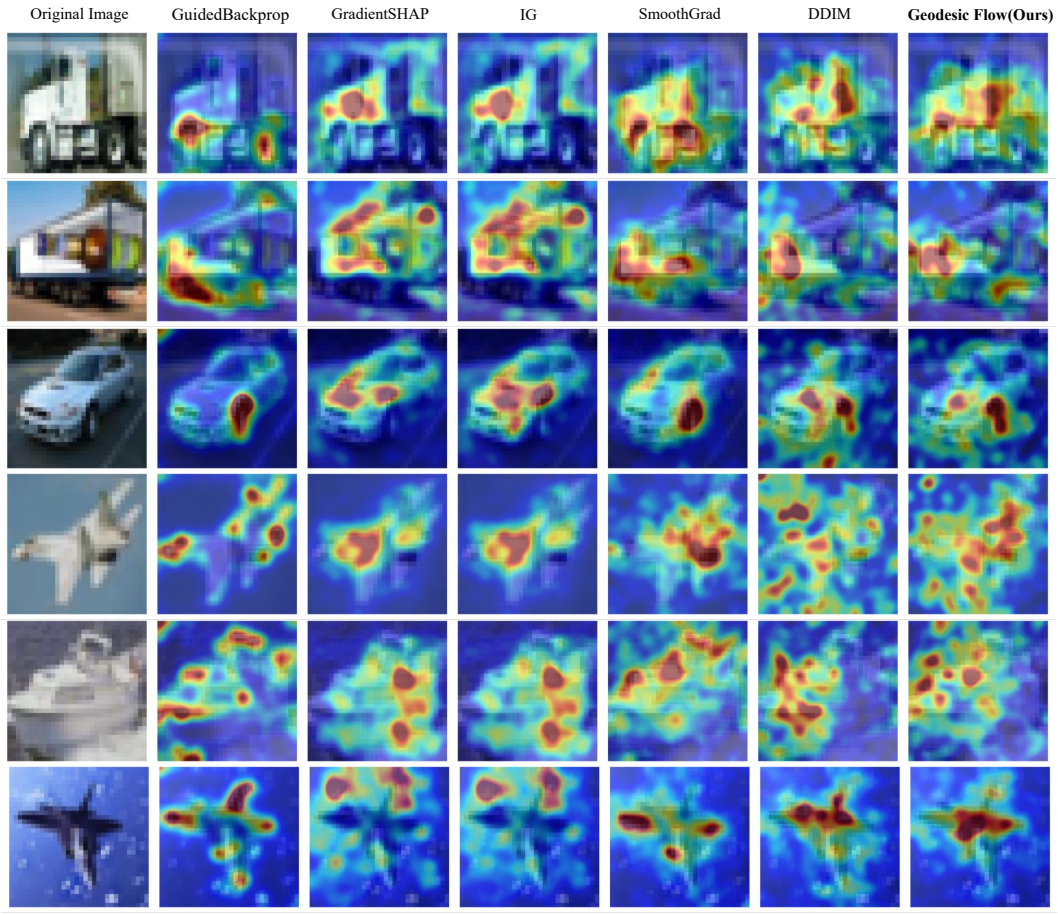

Figure 10: Additional CIFAR-10 examples, part II.

G Limitations and scope of the claims

This appendix section states what the paper does not claim. We include it because the earlier wording around on-manifold attribution can be misleading. The revised paper deliberately avoids that overclaim.

G.1 No strict manifold certificate

We do not prove that the learned path stays on a true data manifold. A strict manifold statement would require a defined manifold, a membership criterion, and a proof that the learned ODE trajectory satisfies that criterion for every time. Our theory instead uses a path of probability distributions that transports p_0 to p_1. The intermediate laws are transport marginals. This is enough for the attribution principle, because the principle needs a selected counterfactual transition, not a manifold certificate.

This limitation changes how one should read FCE. FCE close to zero means that a numerical path follows the learned vector field. It does not mean that the true data density is high at every point. We therefore use FCE as a dynamics diagnostic and not as evidence of strict on-manifold behavior.

G.2 Dependence on the learned generator

The method inherits errors from the learned flow. If $\boldsymbol{v_\theta}$ fails to approximate the target transport field near the trajectory for $\boldsymbol{x}_1$, then the attribution may follow a poor path. Theorem 4.6 explains this dependence by bounding attribution error in terms of vector-field error under regularity assumptions. The theorem does not remove the need for a good generator. It only tells us how generator error propagates once the assumptions hold.

This also means that datasets with poor generative coverage or strong spurious correlations can produce misleading transport paths. In such cases, the attribution may reflect the generator's learned biases together with the classifier's behavior. We recommend checking generator quality and using the method together with sanity checks rather than treating it as a standalone certificate.

G.3 Approximate optimal transport

The ideal path is the kinetic-action minimizer. In high-dimensional image experiments, we do not solve the exact Benamou-Brenier problem. Rectified Flow and Reflow give a scalable approximation. The controlled Gaussian experiments show that Reflow can move the learned path much closer to an OT oracle, but this evidence is empirical and problem dependent. The correct claim is that Reflow reduces the geometry gap in our diagnostics, not that it always reaches exact optimal transport.

The uniqueness language in the main theorem should also be read relative to the ideal problem. If the OT dynamic plan is not unique, then the path-selection problem can have multiple minimizers. Assumption 4.1 rules out that ambiguity for the formal theorem. In applications, different trained flows can approximate different near-minimizing paths, which is why we report stability across seeds.

G.4 Computational cost

The method is more expensive than one backward pass. It needs ODE integration and repeated gradients along the path. This is still linear in the number of integration steps and much cheaper than exact feature-level Shapley values, but it is slower than SmoothGrad with a small number of noise samples or a single Grad-CAM pass. The cost is most relevant for high-resolution images and large predictors.

Several engineering choices can reduce the cost. We can use fewer path steps when approximate completeness is acceptable, use midpoint or adaptive quadrature to improve accuracy per step, cache path states, or group pixels into superpixels before deletion evaluation. These choices change the numerical estimator but not the continuous attribution principle.

G.5 Evaluation limitations

Deletion, blur deletion, SATV, EAS, GPS, and FCE each measure only one aspect of explanation behavior. Deletion depends on the replacement operator. SATV can reward smooth maps even when

they are not faithful. EAS measures visual alignment with image edges but not causal relevance. GPS measures path straightness but not distributional correctness. FCE measures consistency with a learned vector field but not true density support. We therefore interpret the experiments as a collection of evidence rather than as a proof that one method is universally best.

The most defensible empirical conclusion is the one stated in the main text. More transport-consistent and lower-action paths tend to produce more stable and more structured attribution maps, while deletion faithfulness remains competitive. This conclusion is narrower than a universal superiority claim, but it is aligned with what the theory and experiments actually support.