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

Cenwei Zhang[*] [†]
The Chinese University of Hong Kong
Hong Kong, 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 therefore produce different explanations. We study this path ambiguity as a modeling problem. Our central question is whether the path can be chosen by a distribution-level 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 fixed-path axioms and coordinate-trace separability. 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 instantiate this framework using Rectified Flow and derive stability bounds quantifying how vector-field approximation errors affect the resulting attributions. Experiments show that Reflow improves cross-seed attribution stability, while learned transport paths yield spatially coherent explanations with competitive deletion faithfulness among the evaluated path-based methods. 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

[*]Equal contribution.
[†]Work done while at Shanghai Jiao Tong University.
[‡]Work done while at Technical University of Denmark.
[§]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

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 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 distribution-level transport 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, together with coordinate-trace separability, 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 for which we adopt the least-action optimal transport principle. The resulting transport geodesic is defined in Wasserstein distribution space, rather than on a prescribed input-space data manifold. Together, these components define our transport-geodesic attribution framework.

The practical method follows the same principle. The ideal Wasserstein geodesic is not available in high-dimensional image spaces. To instantiate this framework, we use Rectified Flow to learn a time-dependent velocity field transporting samples from a reference distribution toward the data distribution [12, 13]. Given an input $\boldsymbol{x}$, we integrate the learned flow backward to recover an input-conditioned reference endpoint, then evaluate the predictor's gradients along the corresponding characteristic trajectory. Accumulating coordinatewise gradient contributions along this path provides a practical approximation to 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 fixed-path axioms and coordinate-trace separability.
- Then, we formulate distribution-level attribution path selection through kinetic-action minimization in dynamic optimal transport.

- Finally, we instantiate the framework with Rectified Flow, derive attribution stability bounds, and demonstrate improved cross-seed stability and spatial structure with competitive deletion faithfulness.

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 $\mathrm{d}\boldsymbol{x}_t/\mathrm{d}t$. For a flow transporting p_0 to p_1, let ρ_t denote the marginal law of the solution of Eq. (2) initialized with $\boldsymbol{x}_0 \sim p_0$. Mass conservation and the prescribed endpoint marginals are expressed as

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

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 separability)**.** For a fixed path γ and coordinate i, let

$$H_i(\gamma) = \{h_i^f : f \in \mathcal{F}\}$$

denote the set of admissible coordinate traces. For every $h \in H_i(\gamma)$, there exists $g \in \mathcal{F}$ such that

$$h_i^g = h, \qquad h_j^g = 0 \quad \text{for all } j \neq i.$$

This condition requires separation only of realizable coordinate traces, rather than arbitrarily prescribed traces. For self-intersecting paths, repeated spatial points impose consistency constraints on the gradients of a smooth score function, so separability need not hold. Nevertheless, the Aumann–Shapley integral remains well defined even when separability fails. For smooth embedded paths and sufficiently rich score classes, separability follows from suitable extensions of first-order information along the curve. This includes the nondegenerate straight displacement paths arising under ideal quadratic optimal transport.

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. Under its assumptions, fixing the path also fixes the credit-allocation rule. 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 a hand-specified reference-and-path rule with a distribution-level transport coupling and its induced characteristic.

4 Transport-geodesic path selection by optimal generative flows

We now address the path-selection problem. Given the reference distribution p_0 and data distribution p_1, many possible couplings and transport paths remain. We adopt kinetic-action minimization as a predictor-independent variational principle for resolving this ambiguity. Specifically, among all admissible flows transporting p_0 to p_1, we select the one minimizing the time-integrated expected squared velocity, independently of the predictor's gradients used for credit allocation. By the dynamic formulation of quadratic optimal transport, this yields a Wasserstein-2 geodesic and an associated optimal endpoint coupling, whose characteristics define the attribution paths. [5]

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

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

[5] We further discuss the scope and limitations of this path-selection principle in Section 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 transport-geodesic. The path of distributions is shortest in the Wasserstein geometry. A single sample trajectory is a characteristic curve of this distributional flow; under ideal quadratic transport, it is the straight displacement interpolation associated with the optimal coupling. In the ideal quadratic OT setting, the resulting attribution therefore coincides with Integrated Gradients from an optimally coupled, input-conditioned reference endpoint.

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 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. 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, and the selected characteristic path is regular. 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.*

Under the stated assumptions, Theorem 4.4 combines two sources of uniqueness: the fixed-path axioms and coordinate-trace separability determine the line integral, while the kinetic-action principle selects the unique distributional path.

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 $f(\boldsymbol{x}) = b + \sum_i f_i(x_i)$, 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 bounded and uniformly Lipschitz in $\boldsymbol{x}$ on $[0,1] \times \mathcal{K}$. 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 uniform bounds and spatial Lipschitz constants of the vector fields, and the gradient bound and gradient Lipschitz constant of f_c, such that for every coordinate i,*

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

Algorithm 1 Transport-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 $\hat{\boldsymbol{x}}_0, \ldots, \hat{\boldsymbol{x}}_K$
set $\hat{\boldsymbol{x}}_K \leftarrow \boldsymbol{x}_1$
initialize $\hat{\boldsymbol{\Psi}} = \mathbf{0}$
for $k = 0, \ldots, K-1$ **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 $\hat{\boldsymbol{\Psi}}$

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 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 refines the coupling by first generating trajectories with a learned flow and then retraining on the induced endpoint pairs [12]. We use Rectified Flow and Reflow to construct tractable transport paths and assess their approximation to the quadratic OT target through kinetic-action and oracle-based diagnostics. Standard Reflow does not in general guarantee recovery of the quadratic optimal transport coupling. 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 $\hat{\boldsymbol{x}}_0, \ldots, \hat{\boldsymbol{x}}_K$. Before accumulating attributions, we set the final stored state to the observed input, $\hat{\boldsymbol{x}}_K \leftarrow \boldsymbol{x}_1$.Throughout the experiments, K denotes the number of attribution intervals, $t_k = k/K$, $\Delta t = 1/K$, 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 attribution intervals for path-based baselines whenever possible. The discrete estimator requires K predictor gradient evaluations, while generative path construction adds solver-dependent ODE evaluations. 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. 1) Does the numerical integral satisfy efficiency as the step count increases? 2) Does Reflow reduce the geometry gap relative to less rectified flows? 3) Does attribution error track vector-field error as predicted by Theorem 4.6? 4) On real images, do transport-consistent paths give structured explanations without destroying deletion faithfulness?

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 (2-RF) 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	$\mathbf{3006.9 \pm 340.4}$	$\mathbf{0.0010 \pm 0.0008}$	$\mathbf{0.911 \pm 0.051}$	$\mathbf{0.882 \pm 0.061}$

5.2 Experimental setup

We use CUB-200-2011 for the numerical completeness study [14], controlled flow checkpoints for the stability analysis, and CIFAR-10 and CelebA-HQ for the image benchmarks [15, 16].

We compare with SmoothGrad [17], Guided Backpropagation [18], GradientSHAP baselines [3], Integrated Gradients [5], and DDIM generative paths [19]. We report standard deletion metrics together with path diagnostics.

We use 32×32 inputs for CUB-200-2011 and CIFAR-10 and 256×256 inputs for CelebA-HQ, with ResNet-18 predictors throughout. Additional implementation details are provided in Appendix D.1.

5.3 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 attribution intervals increases. The default $K = 50$ gives a practical balance, while $K = 100$ or $K = 200$ can be used when tighter efficiency residuals are desired.

5.4 Path geometry and stability

We next ask whether a more geodesic transport path produces a more stable attribution. We compare a one-round Rectified Flow model (1-RF) with its reflowed counterpart (2-RF) over three seeds. We estimate the mean discrete kinetic action over N evaluated trajectories for each seed by

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

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.

Motivated by Theorem 4.6, we compare earlier flow checkpoints with a converged reference checkpoint. This measures agreement with the reference model rather than error against the unknown ideal OT field. Let $\boldsymbol{\Psi}^{\text{ref}}$ and $\boldsymbol{v}_{t_k}^{\text{ref}}$ denote the attribution and velocity field of this reference checkpoint. We

Table 3: CIFAR-10 and CelebA-HQ benchmark results. We report path diagnostics only for methods that define a continuous path. FCE measures agreement with our learned transport field and is not used to rank attribution methods; deletion scores provide perturbation-based faithfulness evidence.

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	0.005 ± 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	1.780 ± 0.557	$\mathbf{0.003 \pm 0.001}$	0.091 ± 0.101	0.184 ± 0.320	0.926 ± 0.587

use $\|\hat{\mathbf{\Psi}} - \mathbf{\Psi}^{\mathrm{ref}}\|_2 / (\|\mathbf{\Psi}^{\mathrm{ref}}\|_2 + 10^{-12})$ 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 field mismatch relative to this reference. Across checkpoints, the median per-sample Pearson correlation between field mismatch and attribution discrepancy exceeds 0.95. This supports an empirical association between the two reference-relative quantities. Full scatter plots are reported in the appendix.

5.5 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

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

Equation (12) measures consistency with the learned transport field, not attribution faithfulness or membership in the true data distribution. Since Transport Flow uses this same field to generate its paths, FCE is an internal path-dynamics diagnostic and is not used to rank attribution methods. 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, and $\boldsymbol{\phi}$ is the channel-aggregated and normalized attribution map defined in Appendix E.3. Here, ∇_{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.

On CIFAR-10, Transport Flow achieves the lowest SATV and blur-deletion score among the evaluated path-based methods, while DDIM achieves the lowest zero-deletion score. On CelebA-HQ, both obtain similar SATV scores, with DDIM achieving slightly better deletion performance. Overall, Transport Flow produces spatially coherent attributions with competitive deletion faithfulness among

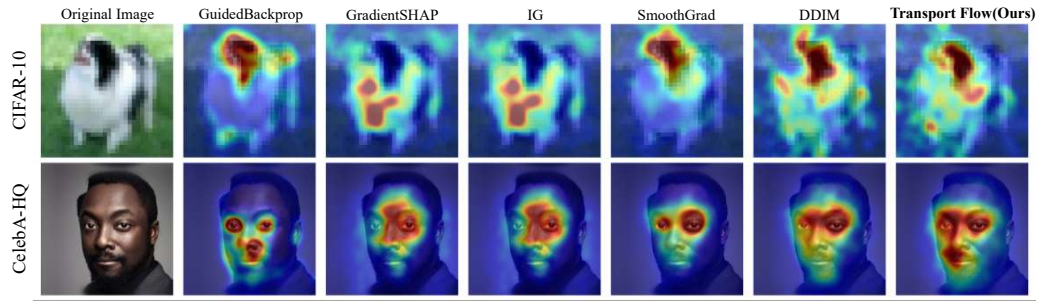


Figure 1: Qualitative visualization on CIFAR-10 and CelebA-HQ. Transport Flow explanations are spatially coherent and avoid some high-frequency artifacts of straight-line Integrated Gradients. The figure should be read together with Table 3; visual clarity alone is not a complete faithfulness metric.

path-based methods on both datasets (Table 3). The additive sanity check in Appendix F.1 shows numerical agreement with analytical coordinate contributions. This checks the fixed-path attribution estimator rather than the transport mechanism.

The qualitative examples in Figure 1 match the quantitative trend. Compared with conventional IG, our method uses an input-conditioned reference recovered through the learned transport flow, rather than a fixed reference. DDIM provides an alternative generative trajectory, while our transport paths exhibit consistency with the learned vector field, as reflected in the FCE results in Table 3.

To distinguish reference selection from trajectory effects, we compare the 1-RF characteristic with straight-line IG using the same endpoint pair $(\hat{x}_0, x_1)$, where $\hat{x}_0$ is recovered by 1-RF. On a CIFAR-10 subset, both achieve comparable faithfulness, with slightly better deletion scores for IG and insertion scores for 1-RF. This is consistent with the ideal quadratic OT formulation, where transport-geodesic attribution coincides with IG using the optimally coupled reference endpoint. Full results are provided in Appendix F.5.

6 Discussion

Scope of least-action transport. Least-action transport is not the only possible principle for path selection. We adopt it as a mathematically grounded criterion for resolving coupling ambiguity independently of the predictor's gradients. De Lara et al. [20] study transport-based counterfactual models and establish conditions under which optimal transport can recover structural counterfactuals, providing related motivation for distribution-level counterfactual matching. We do not claim that least-action transport is universally the best choice for attribution. Rather, it provides a well-defined path-selection principle conditional on the specified reference distribution and ground cost, while alternative transport geometries may be appropriate for different explanation objectives.

Extensions to other modalities. Although our experiments focus on vision, the fixed-path attribution result applies more broadly to differentiable scalar predictors along sufficiently regular paths. Extending the transport component additionally requires meaningful reference and target distributions and a continuous flow. Continuous tabular data is a natural next step, while categorical features may require constrained flows or continuous relaxations. For NLP, a possible extension is to operate in embedding space and map the resulting attributions back to tokens. We leave empirical validation of these extensions to future work.

7 Conclusion

We reformulate feature attribution by separating fixed-path credit allocation from path selection. Building on the Aumann–Shapley attribution principle, we introduce transport-geodesic attribution based on least-action optimal transport to select paths between reference and data distributions. Our experiments show that Reflow improves cross-seed attribution stability, while learned transport paths yield spatially coherent explanations with competitive deletion faithfulness among the evaluated path-based methods. Together, these results support distribution-level transport as a principled approach to reference and path selection for feature attribution.

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] Lucas De Lara, Alberto González-Sanz, Nicholas Asher, Laurent Risser, and Jean-Michel Loubes. Transport-based counterfactual models. *Journal of Machine Learning Research*, 25 (136):1–59, 2024. URL http://jmlr.org/papers/v25/21-1440.html.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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 provides related work, proofs of the main results, optimal transport background, implementation details, evaluation protocols, additional experiments, and limitations.

Appendix A: Related Work . . . 13
Appendix B: Proofs of main results . . . 14
Appendix C: Optimal transport details and the meaning of transport-geodesic . . . 16
Appendix D: Implementation details . . . 19
Appendix E: Evaluation metrics and protocols . . . 20
Appendix F: Additional results and visualizations . . . 22
Appendix G: Limitations and scope of the claims . . . 31

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 [21, 22, 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 [23, 24, 25, 26, 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 [27, 28, 29]. 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, 30]. Manifold-restricted Shapley methods use generative or conditional models to keep perturbations closer to the data distribution [31, 32]. 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, 33]. 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 [34, 35, 36]. 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 fixed-path uniqueness, transport-geodesic characterization, and flow-approximation stability results. It also establishes the discrete efficiency bound and the attribution formula for additive scores.

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. □

Proof of Theorem 4.4. Assumption 4.1 gives a unique optimal dynamic plan from p_0 to p_1, determining the distributional path and its characteristic curves under the stated regularity conditions. Every rule in the specified class therefore selects the same characteristic $\gamma^\star_{\boldsymbol{x}_1}$. This path is regular by the theorem's hypothesis, so Theorem 3.5 applies. 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) selects this optimal transport path and 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 $B_v = \sup_{t\in[0,1],\,\boldsymbol{x}\in\mathcal{K}} \max\{\|\boldsymbol{v}_t^\star(\boldsymbol{x})\|_2, \|\hat{\boldsymbol{v}}_t(\boldsymbol{x})\|_2\}$. 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

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

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

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

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 uniform bounds and spatial Lipschitz constants of the vector fields, and the gradient bound and gradient Lipschitz constant 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 every line segment $[\hat{\boldsymbol{x}}_k, \hat{\boldsymbol{x}}_{k+1}]$, $k = 0, \dots, K-1$. 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 transport-geodesic

This section connects optimal endpoint couplings, displacement interpolation, and kinetic action, and explains their roles in path attribution.

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 is

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

An optimal coupling $\pi^\star$ induces the displacement 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. Under standard regularity conditions, this path of laws is the constant-speed Wasserstein geodesic. 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 $T(\boldsymbol{x}_0)$.

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 transport-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. □

An ideal W_2 geodesic combines an optimal endpoint coupling with constant-speed straight displacement paths. Straightness describes the individual trajectories, whereas transport optimality also depends on the coupling. We use kinetic-action and oracle-based diagnostics to assess learned transport paths.

C.3 Transport geometry and reference selection

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.

Separating reference selection from trajectory effects. The same-endpoint ablation in Table 5 compares the learned 1-RF characteristic with the straight path

$$\gamma_{\text{line}}(t) = (1-t)\hat{x}_0 + tx_1,$$

where $\hat{x}_0$ is recovered by the same 1-RF model. The input, recovered reference, integration step count, and attribution and evaluation protocols are held fixed between these two methods. The unmatched-reference IG row provides an additional reference-selection comparison. The point estimates favor same-endpoint IG on both deletion metrics and the learned characteristic on insertion. Thus, this experiment identifies no systematic faithfulness advantage from the residual curvature of the learned 1-RF trajectory across the reported metrics. It isolates trajectory effects for a given learned reference; optimality of the reference coupling remains a separate question.

The role of quadratic transport. As described in Appendices C.1–C.2, ideal quadratic transport selects an optimal endpoint coupling and connects each matched pair by straight displacement interpolation. For an optimally matched pair $(x_0^\star, x_1)$,

$$\gamma_{x_1}^\star(t) = (1-t)x_0^\star + tx_1.$$

Substituting this path into the attribution integral gives

$$\begin{aligned}\Psi_i(f_c, x_1) &= (x_{1,i} - x_{0,i}^\star)\int_0^1 \partial_i f_c\big((1-t)x_0^\star + tx_1\big)\,\mathrm{d}t \\ &= \mathrm{IG}_i(f_c; x_0^\star, x_1).\end{aligned}$$

Consequently, the ideal construction coincides with Integrated Gradients from an optimally coupled, input-conditioned reference. Its nontrivial transport contribution is the distribution-level coupling and reference selection. Rectified Flow and Reflow provide scalable approximations to this endpoint-conditioned transport. Differences between a learned characteristic and same-endpoint IG reflect residual trajectory geometry together with finite ODE and quadrature effects.

C.4 Coordinate-trace separability and self-intersecting paths.

Assumption 3.3 requires separation only of realizable coordinate traces. Regularity of a path does not exclude self-intersections. If $\gamma(t_a) = \gamma(t_b)$ for $t_a \neq t_b$, every C^1 score g must satisfy

$$\nabla g(\gamma(t_a)) = \nabla g(\gamma(t_b)).$$

Consistency of values and gradients at repeated spatial points can prevent coordinate-trace separation, even when the trace to be retained is realizable. Separability therefore remains an explicit assumption for the fixed-path uniqueness characterization. The Aumann–Shapley integral itself remains well defined for C^1 scores along C^1 paths, including self-intersecting paths.

For smooth embedded paths, separability follows from extending compatible first-order information along the curve when $\mathcal{F}$ contains the required extensions. This includes the nondegenerate straight displacement paths

$$\gamma(t) = (1-t)x_0 + tx_1, \qquad x_0 \neq x_1,$$

arising under ideal quadratic optimal transport, subject to the same score-class requirement.

C.5 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.

Numerical integration errors can cause the backward-forward trajectory to end at a point different from $\boldsymbol{x}_1$. We set $\hat{\boldsymbol{x}}_K \leftarrow \boldsymbol{x}_1$ before accumulating attributions. The resulting stored path connects the recovered reference $\hat{\boldsymbol{x}}_0$ to the observed input $\boldsymbol{x}_1$. 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

Rectified Flow fits a time-dependent vector field $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ using Eq. (10). We write $\hat{\boldsymbol{v}}_t(\boldsymbol{x}) = \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ for the learned field. The training pairs $(\boldsymbol{z}_0, \boldsymbol{z}_1)$ are drawn from a coupling π, typically the independent coupling between the reference prior and data distribution for 1-RF. For Reflow, the learned flow maps prior samples to generated endpoints, and the induced endpoint pairs are used to train a new vector field. This procedure refines the training coupling and is used to straighten transport trajectories [12].

Experimental setup. We use ResNet-18 predictors with approximately 57% accuracy on CUB-200-2011 at 32×32, approximately 98% on CIFAR-10 at 32×32, and above 98% on the CelebA-HQ gender attribute task at 256×256. For CUB-200-2011, generative models are fine-tuned from pretrained checkpoints. The main CIFAR-10 and CelebA-HQ image benchmarks use official pretrained RF and diffusion checkpoints. The numerical completeness study uses 2-RF, while the main image benchmarks use 1-RF. The theoretical guarantees apply under the stated regularity assumptions, including the smoothness conditions on the target score. We evaluate completeness and attribution stability empirically for the experimental predictors.

Implementation and sample selection. Experiments are implemented in PyTorch on a single NVIDIA RTX 3090 GPU with 24 GB memory. Flow trajectories are integrated with the `rk45` solver, and the default attribution grid uses $K = 50$ intervals. For classifier-based evaluations, we select correctly classified samples with prediction confidence above 0.6 for the axiomatic and stability studies and above 0.8 for the image benchmarks.

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 $\Delta t = 1/K$. The backward solve ties the recovered reference to the observed input through the learned vector field. This differs from Expected Gradients or GradientSHAP, where the reference is sampled independently from a background distribution.

Finite solver accuracy can produce a terminal mismatch after the backward-forward integration. We set $\hat{\boldsymbol{x}}_K \leftarrow \boldsymbol{x}_1$ before evaluating Eq. (11). Completeness is evaluated against $f_c(\boldsymbol{x}_1) - f_c(\hat{\boldsymbol{x}}_0)$ using this endpoint-corrected stored trajectory.

The discrete kinetic action, GPS, and FCE are evaluated using the same endpoint-corrected stored states.

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 use it because it is simple and matches the finite-difference view of adding small path increments. For sufficiently smooth scores and paths, midpoint quadrature can achieve second-order accuracy by evaluating gradients at intermediate states. We compare both quadrature rules in Appendix F.1. 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 integration intervals requires storing $K + 1$ path states. Both left Riemann and midpoint quadrature require K target-score gradient evaluations, although midpoint quadrature requires intermediate path states. 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.

GPS and FCE are reported only for methods with a specified continuous path; the other entries in Table 3 are marked with a dash. FCE is evaluated against our learned RF field and serves as a dynamics diagnostic, separately from the perturbation-based faithfulness scores.

D.5 Computational cost

Exact discrete Shapley computation for a general feature-level game requires an exponential number of coalition evaluations. Practical explainers reduce this cost through coalition sampling or feature grouping.

For Eq. (11), attribution accumulation requires K predictor gradient evaluations and scales linearly with the number of attribution intervals. Constructing the learned path adds backward and forward ODE integration. The ODE function-evaluation count is solver-dependent and is distinct from the attribution grid size K. Overall computational cost also depends on the input resolution and the predictor and vector-field architectures. Compared with straight-line IG at matched gradient evaluation counts, the additional work is the generative path construction.

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. Integrated Gradients has GPS exactly one for a straight path. GPS characterizes trajectory geometry, while transport optimality additionally depends on the endpoint coupling.

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 indicates agreement with the learned vector-field family $\hat{\boldsymbol{v}}_t$. Since Transport Flow uses this same field to generate its paths, FCE serves as an internal dynamics diagnostic, not as a faithfulness metric or a basis for ranking attribution methods.

For methods without a specified continuous path, such as SmoothGrad or Guided Backpropagation, FCE is not applicable. For IG and DDIM, it measures the mismatch between their paths and the same learned RF field.

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 summarizes spatial structure and is interpreted alongside deletion metrics and completeness checks. Completeness is evaluated using the original signed, unnormalized coordinate attributions.

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

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

The reported score is the normalized weighted average

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

A higher score means that attribution-map edges are more aligned with image edges. 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.

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 and visualizations on numerical accuracy, transport geometry, and attribution behavior.

F.1 Synthetic additive sanity check

We first check the additive case in Proposition B.3. We sample 50,000 inputs in $d = 10$ dimensions with independent coordinates uniformly distributed over $[-\pi, \pi]$ 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 approximate the Aumann–Shapley integral along the straight path from $\boldsymbol{x}_0$ to $\boldsymbol{x}$, without using a learned flow.

We compare the left Riemann estimator in Eq. (11) with midpoint quadrature using $K \in \{10, 20, 50, 100, 200\}$ integration intervals. Both estimators use the same sampled inputs.

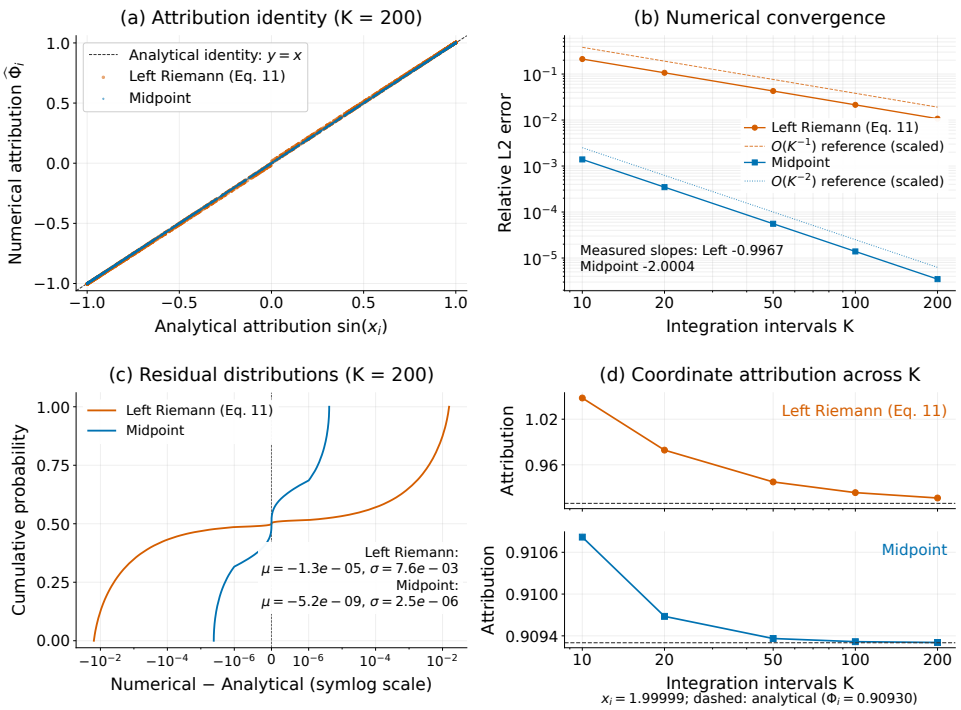


Figure 2: Synthetic additive sanity check. Both estimators converge to the analytical additive contributions. The left Riemann estimator exhibits first-order convergence, while midpoint quadrature achieves second-order convergence.

Figure 2 shows near-identity alignment between analytical and numerical attributions. The measured convergence slopes are -0.9967 for left Riemann and -2.0004 for midpoint quadrature, consistent with their expected convergence rates. These results numerically support the additive allocation property for the tested straight paths, consistent with Proposition B.3.

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-round Rectified Flow (1-RF) and Reflow.

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

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

The reported action gap is the relative excess action

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

The oracle column reports $\Delta A = 0$ using the analytical action $\mathcal{A}^\star$ as the reference. 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,

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

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.

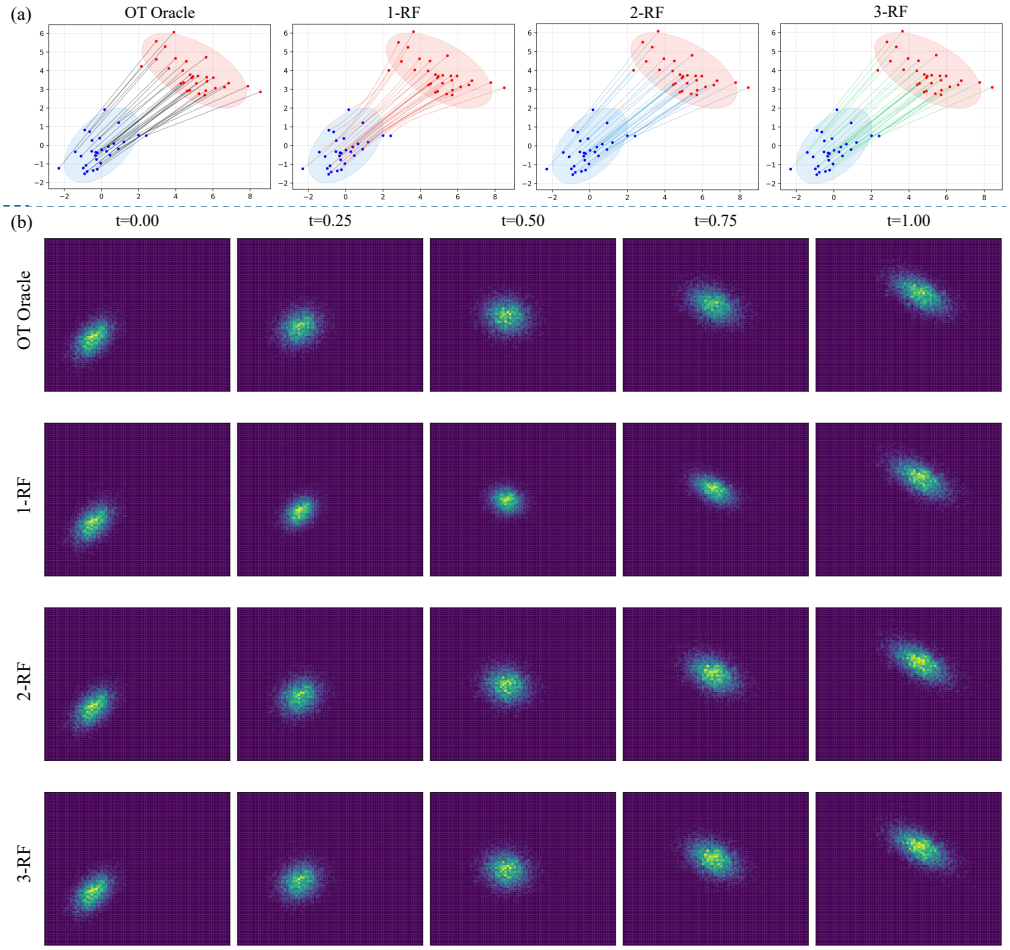


Figure 3: Controlled two-dimensional Gaussian transport. The oracle displacement interpolation gives the reference geometry. One-round Rectified Flow can be visibly more curved, while Reflow straightens the trajectories and makes the intermediate densities closer to the oracle interpolation.

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-round Rectified Flow.

Table 4 supports the same conclusion quantitatively. The one-round 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 transport-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 field mismatch is the relative field error RFE_m defined above. For checkpoint comparisons, we replace $\boldsymbol{v}_t^\star$ with the reference checkpoint's velocity field, yielding E_{field} as defined in Section 5.4. The Gaussian experiment therefore uses the known OT field, while the checkpoint experiment measures agreement with a learned reference.

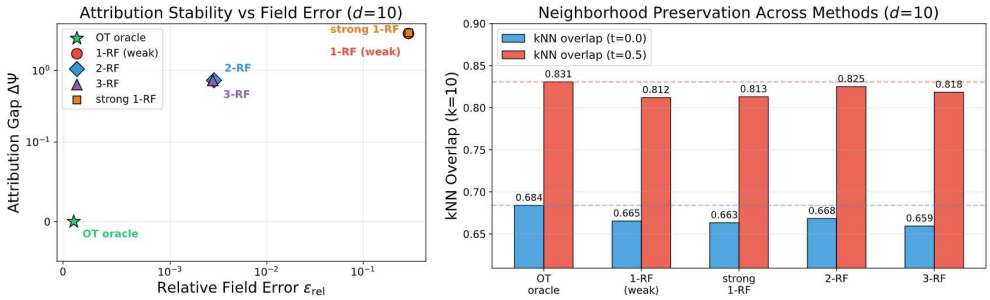


Figure 4: High-dimensional controlled diagnostic. The left panel relates relative attribution error to relative field mismatch. The right panel reports neighborhood preservation along intermediate transport states. The result supports the stability view: better field alignment gives smaller attribution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

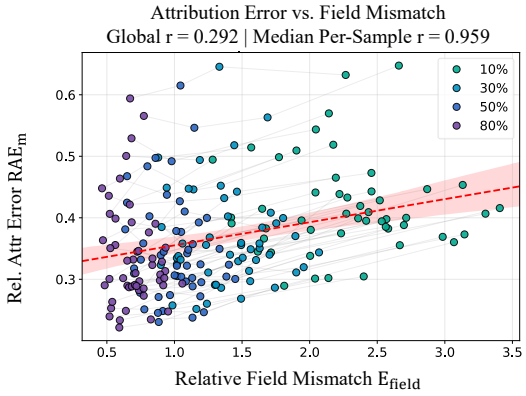


Figure 5: Attribution convergence across flow checkpoints. The horizontal axis measures normalized field mismatch E_{field}, while the vertical axis measures relative attribution error RAE, both evaluated against the reference checkpoint.

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.

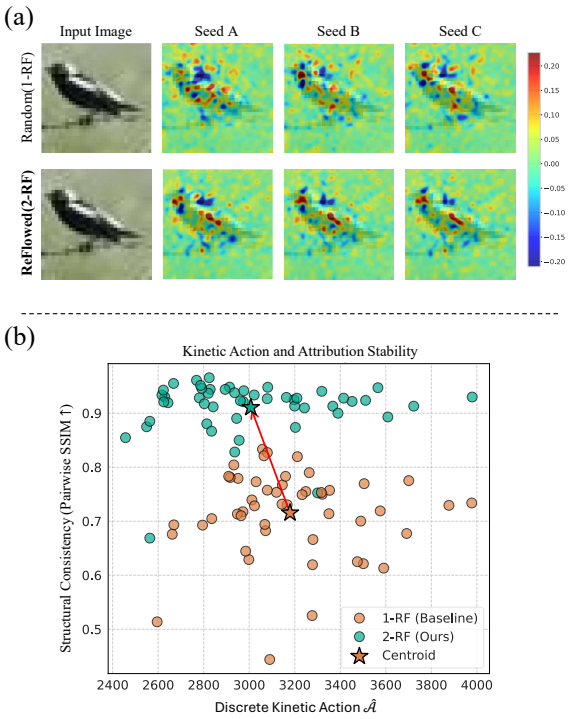


Figure 6: Kinetic action and attribution stability across flow models. (a) Representative attribution maps from three random seeds for 1-RF and 2-RF. (b) Relationship between discrete kinetic action and pairwise attribution SSIM, with stars indicating group centroids. Reflow is associated with lower action and higher attribution stability.

F.5 Same-endpoint path ablation

To distinguish the contribution of reference selection from that of the learned trajectory, we compare the 1-RF characteristic with straight-line Integrated Gradients (IG) using the same recovered endpoint pair $(\hat{x}_0, x_1)$. Both methods use the same number of integration steps and identical attribution and evaluation protocols. We additionally include conventional IG with its original reference choice. The experiment is conducted on a subset of CIFAR-10.

Table 5: Same-endpoint path ablation on a CIFAR-10 subset. Straight IG (same endpoint) and 1-RF use the same recovered reference $\hat{x}_0$ and target input x_1.

Method	Del.-Black ↓	Del.-Blur ↓	Insertion ↑
IG (original reference)	0.1825	0.3866	0.4179
1-RF path	0.1795	0.3134	**0.4303**
Straight IG (same endpoint)	**0.1727**	**0.3010**	0.4286

Table 5 allows us to examine reference selection and trajectory effects separately. Compared with conventional IG, straight-line IG using the 1-RF-recovered reference achieves lower deletion scores and a higher insertion score on the evaluated CIFAR-10 subset. This suggests that the learned reference selection contributes to attribution faithfulness.

With the reference endpoints fixed, straight-line IG and the 1-RF characteristic achieve comparable faithfulness. Straight-line IG yields slightly lower deletion scores, whereas the 1-RF characteristic achieves a slightly higher insertion score. These results indicate no consistent advantage from the residual trajectory curvature of the learned flow.

This observation is consistent with the ideal quadratic optimal transport formulation. For fixed endpoints, the constant-speed straight segment minimizes kinetic action. Under exact quadratic optimal transport, the optimal coupling determines the reference endpoint, and the corresponding characteristic follows straight displacement interpolation. Consequently, ideal transport-geodesic attribution coincides with IG using the OT-selected reference.

In practice, the learned Rectified Flow provides a tractable transport coupling and its associated characteristic trajectories without guaranteeing exact optimality. Differences between the learned characteristic and same-endpoint IG therefore reflect residual trajectory geometry and numerical approximation effects.

F.6 Additional qualitative examples

The following figures show additional randomly selected attribution examples on CIFAR-10 and CelebA-HQ, illustrating the spatial structure of the resulting maps.

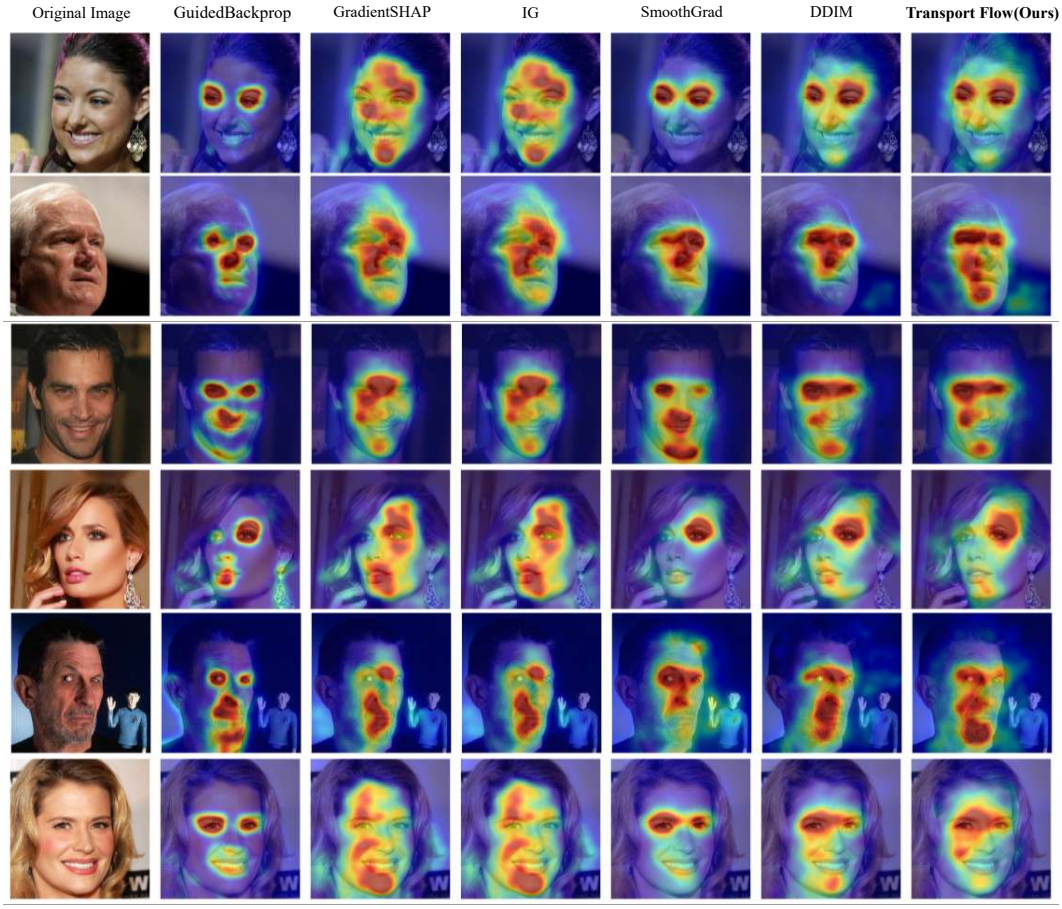


Figure 7: Additional CelebA-HQ examples, part I.



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

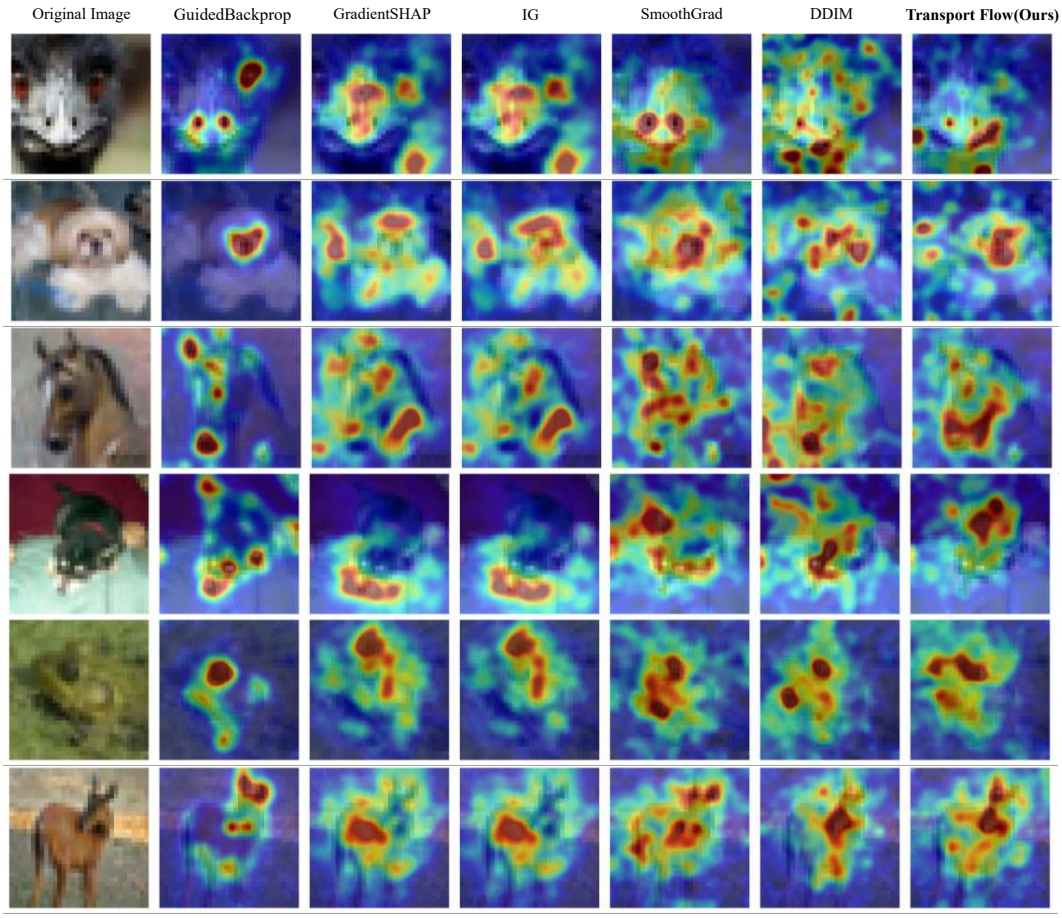


Figure 9: Additional CIFAR-10 examples, part I.

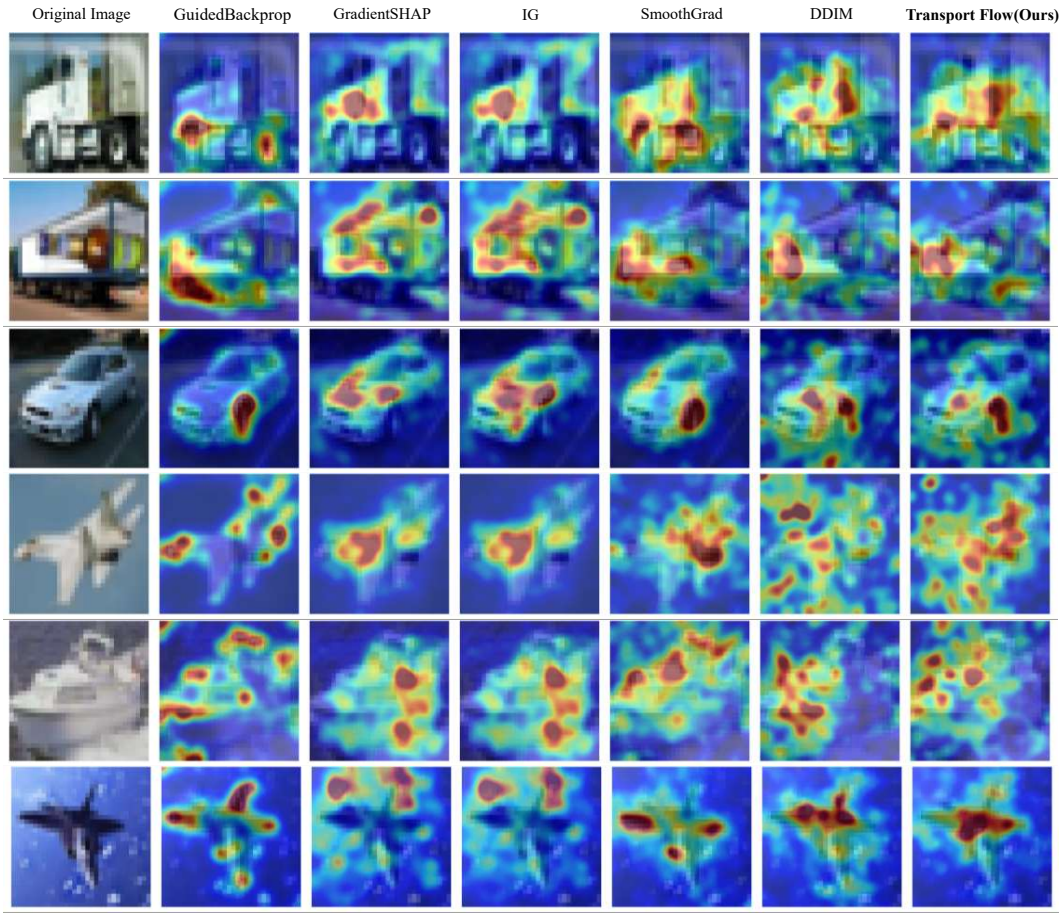


Figure 10: Additional CIFAR-10 examples, part II.

G Limitations and scope of the claims

This appendix section states what the paper does not claim. This section discusses limitations arising from learned transport, computational cost, and evaluation.

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.

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 reduces the relative action gap and vector-field mismatch against the OT oracle. This provides empirical evidence of improved transport approximation in the evaluated setting.

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, insertion, SATV, EAS, GPS, and FCE each measure only one aspect of explanation behavior. Deletion and insertion depend on the perturbation 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 is evaluated against the same learned field used to generate Transport Flow paths. It measures dynamical consistency rather than independent attribution faithfulness, density support, or transport optimality.

The synthetic additive sanity check in Appendix F.1 assesses numerical recovery of analytical coordinate contributions by the fixed-path estimator, separately from learned path selection. The reported Reflow comparisons associate lower action with greater attribution stability, while the image benchmarks provide evidence of structured maps and competitive deletion faithfulness. The empirical conclusions concern the evaluated datasets, predictors, and perturbation protocols.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the intended scope: the paper studies path ambiguity, separates fixed-path credit allocation from path selection, and proposes a transport-geodesic path-selection principle. The claims are intentionally limited to stability, path geometry, and competitive faithfulness rather than universal superiority.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the conclusion and in Appendix G, including no strict data-manifold certificate, dependence on the learned generator, approximate optimal transport, computational cost, and limits of attribution metrics.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The main text states the assumptions for the fixed-path uniqueness theorem and the stability result, including coordinate-trace richness and regularity of the transport flow. Complete proofs are provided in Appendix B, with auxiliary transport details in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: The paper gives the code, algorithm, objectives, datasets, metrics, and evaluation protocol in abstract, Section 5 and Appendices D–E. However, the submission does not yet include trained checkpoints, and exact low-level implementation settings needed for exact numerical reproduction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments use public benchmark datasets and cite the corresponding sources, and the current submission provide an open anonymous code. The method is described algorithmically, and the authors plan to release the last full-specified code after review if permitted by the submission policy.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: The paper specifies the core estimator, Rectified Flow/Reflow objectives, datasets, baselines, path-step notation, and metrics. It does not yet provide every exact engineering detail such as all network architectures, optimizer settings, hardware-dependent solver settings, and full random seed configuration.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The quantitative tables report mean values with standard deviations for the benchmark metrics and for controlled Gaussian transport diagnostics. The stability experiments also compare independently trained flows across seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix D.5 analyzes algorithmic cost and explains that one explanation scales linearly with the number of path steps. The paper does not yet report the exact hardware type, memory, wall-clock time, or total compute used for each experimental run.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is a methodological study of model explanations using standard public image benchmarks and does not involve deception, intervention on users, or sensitive deployment. The paper also includes an impact statement and discusses limitations and responsible use.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The impact statement discusses potential positive uses for model auditing and scientific or medical analysis, as well as risks such as overinterpreting attributions as causal guarantees and inheriting biases from learned generative models.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk pretrained model, scraped dataset, or dual-use generative system. It studies an attribution method evaluated on existing public benchmarks, so special release safeguards are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper credits the existing datasets and baseline methods through citations, including CUB-200, CIFAR-10, and CelebA-HQ. The current manuscript does not yet explicitly list the licenses or terms of use for each external asset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a new dataset, benchmark, pretrained model release, or other standalone asset. It introduces a method and reports experiments on existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or new research with human participants. It uses existing public image datasets and does not collect new human-subject data.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not conduct new human-subject experiments, crowdsourcing studies, or interventions with participants. Therefore IRB or equivalent approval is not applicable to the reported work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the proposed method, theory, or experiments. Any language-model assistance, if used for writing or formatting, does not affect the core methodology or scientific claims.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.