FRInGe: Distribution-Space Integrated Gradients with Fisher–Rao Geometry

Gabriele Martino
Faculty of Computer Science
Doctoral School Computer Science
University of Vienna, Vienna, Austria
gabriele.martino@univie.ac.at

Sebastian Tschiatschek
Faculty of Computer Science
ds:UniVie
University of Vienna, Vienna, Austria
sebastian.tschiatschek@univie.ac.at

Abstract

Although gradient-based attribution methods are model-faithful and scalable, a prominent variant, Integrated Gradients (IG), suffers from high sensitivity to design choices, particularly in baseline selection, path formulation, and discretization, as well as gradient saturation. To address these limitations, we propose Fisher–Rao Integrated Gradients (FRInGe), which defines both the reference and integration paths in the predictive distribution space. FRInGe replaces input baselines with a maximum-entropy predictive reference and follows a Fisher–Rao geodesic on the probability simplex. The corresponding input-space trajectory is realized through the pullback Fisher metric and stabilized by Kullback–Leibler (KL) and Euclidean trust regions; attributions are obtained by integrating input gradients along this trajectory. We further derive FRInGe-B, a variant that binarizes the predictive probability distribution, removing the computationally expensive Fisher products of FRInGe. Across six ImageNet architectures, FRInGe most clearly improves Magnitude Aligned Scoring (MAS), while remaining competitive on perturbation area-under-the-curve (AUC) metrics and infidelity.

1 Introduction

Deep neural networks are increasingly deployed in high-stakes domains such as medical imaging and robotics [21, 20, 36], yet the rationale behind individual predictions often remains opaque [14]. This opacity has motivated research in explainable AI (XAI), where *feature attribution* methods seek to identify the input components most relevant to a model's output [13]. Among these, gradient-based approaches are widely adopted because they leverage exact model gradients directly and can be implemented efficiently via automatic differentiation.

A prominent representative of this family is Integrated Gradients (IG) [27]. Given an input x and a baseline x', IG attributes importance to each feature by integrating the gradient of the model output along a path connecting x' to x. Under mild regularity assumptions, this construction satisfies desirable properties such as implementation invariance, sensitivity, and completeness [27].

Despite its theoretical appeal, IG can be sensitive to several design choices. First, the *baseline* is intended to represent an absence of information, but common choices such as an all-zero image or noise can induce qualitatively different explanations [25]. Second, the *integration path*, typically chosen as a straight line in input space, can traverse regions where gradients are poorly conditioned or dominated by artifacts [5, 10, 17, 37]. Recent work has therefore explored replacing straight-line paths with geometry-aware trajectories [19]. Third, IG approximates the path integral by a finite Riemann sum, and the resulting discretization error can become substantial when gradients vary sharply along the path. These issues are further compounded by *saturation*: along large portions of the standard path, the model output may change rapidly early on and then plateau, so informative

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

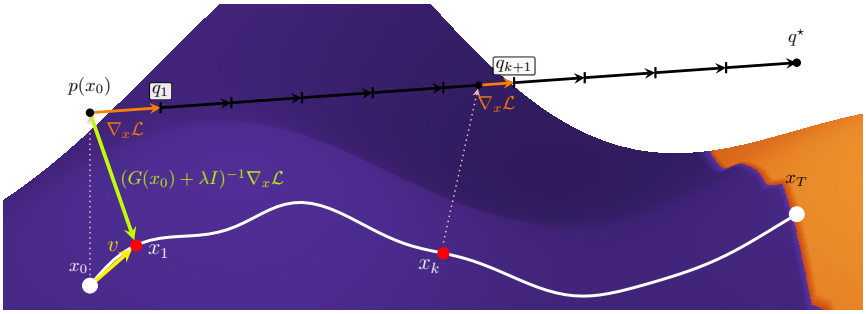

Figure 1: FRInGe constructs a Fisher–Rao geodesic from the initial prediction to a maximum-entropy reference in predictive distribution space and computes an input trajectory whose predictions approximately track its waypoints.

gradients are concentrated in a narrow segment while later integration steps contribute a mostly noisy or weakly informative signal [31].

In this paper, we address these issues by defining both the reference and the interpolation schedule in *predictive distribution space*. Rather than selecting a heuristic input-space baseline, we use a *distributional* reference point: the maximum-entropy predictive distribution, which operationalizes "no information" directly at the model output level. We then interpolate from the model's predictive distribution for a given input to this reference along a Fisher–Rao geodesic on the probability simplex. For categorical predictions—distributions retaining a separate probability for every class—this Fisher–Rao geometry is canonical rather than a tunable design choice: it is uniquely determined by fundamental statistical principles. Formally, up to an overall scale, it is the unique Riemannian metric invariant under statistically sufficient transformations [6, 3]. Moreover, assuming strictly positive endpoints (as guaranteed by our maximum-entropy reference and softmax predictive distributions), the square-root great-circle construction gives the unique constant-speed shortest path. To realize this trajectory in input space, we use the Fisher–Rao *pullback metric*, which induces a local geometry on the input space through the model's mapping from inputs to predictive distributions. This yields a model-aware direction field that avoids committing to a Euclidean straight-line interpolation. To discretize and integrate this path stably, we further use a trust-region step rule combining KL control in predictive space with a Euclidean step-size control in input space. Feature attributions are then computed by integrating input gradients along the resulting path.

To the best of our knowledge, FRInGe is the first attribution method that (i) defines the IG reference in predictive distribution space, (ii) constructs the integration schedule as a Fisher–Rao geodesic in that space, and (iii) induces the corresponding input-space trajectory via the Fisher–Rao pullback metric. In contrast to prior path variants that design trajectories directly in input space, or methods that rely on an explicit generative model of the data manifold, our path is defined by the geometry of the model's predictive distribution and then mapped back to input space. Defining the reference in predictive distribution space also reduces dependence on data-type-specific baseline design.

We additionally derive FRInGe-B, a specialization that collapses the categorical prediction into the target class versus all remaining classes. Whereas full FRInGe must iteratively solve a linear system using model-dependent Fisher matrix-vector products at each waypoint, FRInGe-B induces a rank-one pullback metric that eliminates these costly inner-loop evaluations, trading within-rest class geometry for a substantial speedup and competitive attribution quality.

2 Related Work

2.1 Baseline choice in Integrated Gradients

A recurring practical challenge in IG is that the baseline is not a benign implementation detail: different baselines can yield qualitatively different explanations. Prior work has therefore sought to reduce baseline arbitrariness either by analyzing baseline sensitivity or by replacing a single reference with a distribution or a set of baselines. Sturmfels et al. [25] provide a detailed analysis of baseline sensitivity and argue that baselines encode different notions of "missingness," motivating the use of

baseline distributions rather than a single reference input. Expected Gradients operationalizes this idea by averaging IG over baselines sampled from a reference distribution, typically the data distribution, thereby reducing dependence on any specific baseline and improving stability in practice [7]. A complementary line of work connects baselines to Shapley-style attributions. Sundararajan and Najmi [26] formalize how different Shapley operationalizations correspond to different reference conventions and propose *Baseline Shapley* as a principled construction. Building on this perspective, Liu et al. [11] propose Shapley Integrated Gradients (SIG), which constructs a *set* of baselines via proportional sampling to better approximate Shapley coalitions. Other approaches tailor baselines to domain semantics; for example, Bardhan et al. [5] show in particle-physics settings that averaging baselines from background events can yield more meaningful attributions than a zero reference. Recent work further argues that candidate baselines should not be treated as equally appropriate and proposes selecting or weighting them according to suitability criteria [30]. Overall, these approaches mitigate baseline dependence by averaging across baselines, constructing baseline sets, or weighting candidate references. In contrast, Tan et al. [29] argue that a principled notion of "no information" corresponds to *maximum entropy* in the model's output distribution. We adopt this perspective by defining the reference point in *predictive distribution space*, thereby avoiding the need to specify a heuristic reference directly in input space.

2.2 Finding a better path

Beyond the baseline, a substantial body of work improves IG by modifying the integration path to avoid regions where gradients are unstable, saturated, or dominated by artifacts. Guided Integrated Gradients (GuidedIG) [10] attributes much of IG's noise to the fixed Euclidean straight-line path and proposes an adaptive, model-dependent trajectory that advances preferentially along features with small partial derivatives, helping avoid high-curvature and saturation-dominated segments. Blur Integrated Gradients [34] replaces linear interpolation with a scale-space path defined by progressively blurring the input, yielding cleaner and more structured attributions through controlled information removal. Other variants modify the path more aggressively to counteract saturation or align the trajectory with decision-relevant directions. IG^2 [38] constructs a counterfactual-guided integration path and baseline through iterative representation matching, while Adversarial Gradient Integration (AGI) [15] integrates gradients along adversarially constructed trajectories that follow steepest-ascent directions in input space; related work further refines such adversarial paths to better align gradient accumulation with decision boundaries [18]. More explicitly geometric approaches have also emerged. Geometrically Guided Integrated Gradients (GGIG) [17] perturbs interpolants by ascending the loss landscape to emphasize discriminative directions, while Manifold Integrated Gradients (MIG) [37] integrates along geodesics of a learned data manifold induced by a deep generative model. MIG consequently depends on both the classifier and an auxiliary generator, whereas FRInGe is a model-only post-hoc method. Most closely related to our perspective, Salek and Enguehard [19] propose Geodesic Integrated Gradients (GeoIG), which replaces the Euclidean straight-line interpolation with geodesics in input space under a model-induced metric defined from the Jacobian, approximated either via shortest paths on a k-NN graph or via an energy-based sampling procedure. In contrast, our method does not define the path directly in input space: it defines both the reference and the interpolation schedule in *predictive distribution space* under Fisher–Rao geometry, and then realizes the corresponding input-space trajectory through the pullback metric together with trust-region stabilization. Because of this close conceptual overlap, geometry-aware path methods such as GGIG and GeoIG form especially important empirical comparators for FRInGe.

3 Background

Let $F\colon \mathbb{R}^n \to \mathbb{R}^C$ be a differentiable model, let $x \in \mathbb{R}^n$ denote an input, let C be the number of classes, and let $x' \in \mathbb{R}^n$ be a baseline in input space. Fix a target class $t \in \{1, \dots, C\}$ and define the corresponding scalar score $F^t : \mathbb{R}^n \to \mathbb{R}$, for example the t-th logit.

Integrated Gradients [27]. Let $\gamma\colon [0,1] \to \mathbb{R}^n$ be a C^1 curve with $\gamma(0) = x'$ and $\gamma(1) = x$. IG along γ is defined component-wise by

$$\big[\text{IG}_\gamma(x; x')\big]_i = \int_0^1 \frac{\partial F^t}{\partial x_i}\big(\gamma(\alpha)\big)\, \frac{d\gamma_i(\alpha)}{d\alpha}\, d\alpha, \tag{1}$$

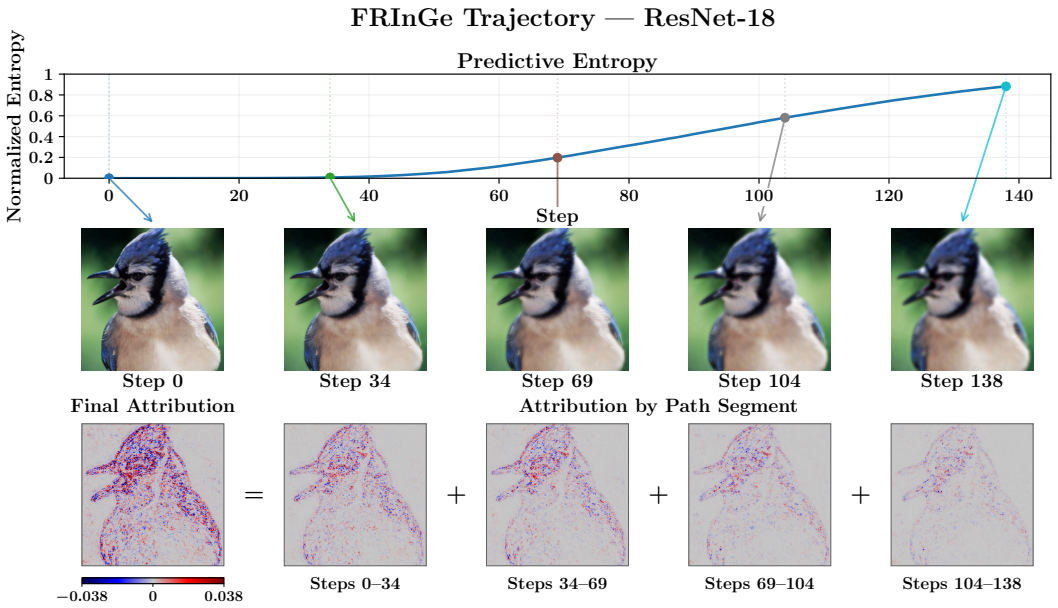

Figure 2: Illustrative FRInGe trajectory on ResNet-18 (protocol in Appendix B). *Top*: normalized predictive entropy. *Middle*: five intermediate inputs. *Bottom*: final attribution (left) and signed contributions accumulated over four consecutive path segments. All maps share one symmetric colour scale, with grey denoting zero.

or equivalently in vector form $\mathrm{IG}_\gamma(x;x') = \int_0^1 \nabla F^t\big(\gamma(\alpha)\big) \odot \gamma'(\alpha)\, d\alpha$. By the fundamental theorem for line integrals, IG satisfies *completeness*:

$$\sum_{i=1}^{n} \big[\mathrm{IG}_\gamma(x;x')\big]_i = F^t(x) - F^t(x'). \tag{2}$$

In practice, the integral is approximated numerically, and the resulting discretization error can become severe when gradients vary sharply along the path or when the realized trajectory undergoes abrupt directional changes between successive steps.

Fisher–Rao geometry and geodesics on the simplex. Let $F(x) \in \mathbb{R}^C$ denote the logits and let $p(x) = \mathrm{softmax}(F(x)) \in \Delta^{C-1}$ be the predictive distribution, where $\Delta^{C-1} = \big\{p \in \mathbb{R}^C_{\geq 0} : \sum_{c=1}^C p_c = 1\big\}$. For categorical distributions, the Fisher–Rao geometry is conveniently represented through the elementwise square-root map $\psi(p) = \sqrt{p}$ which embeds the simplex into the positive orthant of the unit sphere $\mathbb{S}_+^{C-1} = \big\{u \in \mathbb{R}^C_{\geq 0} : \|u\|_2 = 1\big\}$. For this embedding, Fisher–Rao geodesics on the simplex correspond to great-circle geodesics on the sphere, up to a constant rescaling of arc length. Hence, given two predictive distributions $p_0, p_1 \in \Delta^{C-1}$, we define

$$\psi_0 = \sqrt{p_0}, \qquad \psi_1 = \sqrt{p_1}, \qquad \theta = \arccos(\langle \psi_0, \psi_1 \rangle).$$

So a constant-speed shortest path on the sphere from ψ_0 to ψ_1 is given by the spherical linear interpolation [23]:

$$s(\alpha) = \frac{\sin((1-\alpha)\theta)}{\sin\theta}\,\psi_0 + \frac{\sin(\alpha\theta)}{\sin\theta}\,\psi_1, \qquad \alpha \in [0,1]. \tag{3}$$

Mapping this path back to the simplex yields $p(\alpha) = s(\alpha) \odot s(\alpha)$, which is the Fisher–Rao geodesic between p_0 and p_1. The corresponding Fisher–Rao distance is $d_{\mathrm{FR}}(p_0, p_1) = 2\arccos\big(\langle \sqrt{p_0}, \sqrt{p_1} \rangle\big)$.

Pullback Fisher metric in input space. Let $J_F(x) = \partial F(x)/\partial x \in \mathbb{R}^{C\times n}$ denote the Jacobian of logits with respect to the input. For a softmax model, the Fisher information in logit coordinates is

$$S(x) = \mathrm{diag}\big(p(x)\big) - p(x)p(x)^\top \in \mathbb{R}^{C\times C}. \tag{4}$$

The induced *pullback Fisher metric* [4] in the input space is

$$G(x) = J_F(x)^\top \, S(x) \, J_F(x) \in \mathbb{R}^{n \times n}. \tag{5}$$

This matrix is positive semidefinite and may be rank-deficient or ill-conditioned. It provides a local quadratic approximation of the KL divergence between nearby predictive distributions:

$$\mathrm{KL}\big(p(x + \mathrm{d}x) \,\|\, p(x)\big) \approx \tfrac{1}{2} \, \mathrm{d}x^\top G(x) \, \mathrm{d}x. \tag{6}$$

4 Method

Given an input x and a fixed target class t, FRInGe computes attributions of the target logit F^t along an input trajectory guided by predictive geometry. We first choose a reference distribution $q^\star$ and construct a Fisher–Rao geodesic connecting the initial prediction $p(x)$ to this reference. We discretize that geodesic into predictive waypoints and define a spherical alignment loss to guide their approximate tracking in input space. We then derive damped natural-gradient directions using the pullback Fisher tensor. Next, we introduce a KL-based step-size bound using a local quadratic approximation and couple the number of tracking steps to the predictive budget. A complementary Euclidean displacement cap controls the lengths of the input-space segments used for numerical integration. For the image models studied here, we further regularize spatial smoothness of both the update directions and the inputs. Finally, we use trapezoidal quadrature to accumulate target-logit gradients along the realized trajectory, where completeness holds relative to the reached endpoint x_T and is evaluated separately from whether $p(x_T)$ reaches $q^\star$ (Section 4 and Appendix A.6).

Unless stated otherwise, FRInGe denotes the full C-class formulation with both smoothness terms active (penalizing high-frequency variation in the update direction and in the current input). Variants disabling either term are reported as ablations.

Reference distribution. The reference distribution $q^\star \in \Delta^{C-1}$ represents a problem-dependent notion of "no evidence" or "uninformative prediction." In this work, we adopt the maximum-entropy choice, i.e.,

$$q^\star = \tfrac{1}{C}\mathbf{1}, \tag{7}$$

motivated by the maximum entropy principle [29]. This yields a model-agnostic reference specified directly in predictive distribution space. Uniformity represents no predictive preference among classes. Although we use the uniform reference, other choices may be appropriate depending on domain knowledge or experimental design. Beyond the interpretation as "no evidence", the reference can also encode a targeted contrast, for example by favoring a selected alternative class or assigning equal maximal probability to two selected classes to represent a tie in the classification prediction.

Inducing an input-space trajectory that tracks predictive waypoints. Specifying a reference in predictive distribution space avoids the need to prescribe a particular input baseline. Computing path-integrated attributions requires an input trajectory along which target-logit gradients can be evaluated. We therefore seek an input-space trajectory whose predictions track successive waypoints along the Fisher–Rao geodesic from $p(x)$ to $q^\star$.

We initialize the input trajectory at the original input, $x_0 = x$. Let $p_0 = p(x)$ denote the predictive distribution of the input x, and let $s(\alpha)$ be the Fisher–Rao geodesic in square-root coordinates defined in Eq. (3). For $\alpha_k = k/T$, $k = 0, \ldots, T$, we define the waypoint sequence $s_k = s(\alpha_k)$. Defining $q_k = s_k \odot s_k$, we have $q_0 = p_0$ and $q_T = q^\star$. At state x_k, $k = 0, \ldots, T-1$, we target the *next* waypoint using the spherical alignment loss

$$\mathcal{L}_{k+1}(x) = 1 - \Big\langle \sqrt{p(x)}, \, s_{k+1} \Big\rangle, \tag{8}$$

which is minimized when the current predictive distribution aligns with that target. This indexing ensures that $s_0 = \sqrt{p(x_0)}$ is the starting prediction, whereas $s_T = \sqrt{q^\star}$ is the final predictive-space target. How we choose the number of steps T is described in the paragraphs below.

A Euclidean descent direction for $\mathcal{L}_{k+1}$ would track the predictive waypoints, but it ignores the model-induced geometry and can enter poorly conditioned input-space regions. We instead use the pullback Fisher metric to obtain geometry-aware directions, as described in the paragraph below; this optimization only realizes the path, while attribution is still computed by path-integrated gradients.

Geometry-aware steps via a damped natural gradient. Let $G(x_k)$ denote the pullback Fisher tensor defined in Eq. (5). Since it may be singular, it defines the Fisher seminorm

$$\|v\|^2_{G(x_k)} := v^\top G(x_k)\, v. \tag{9}$$

At state x_k, let

$$g_k = \nabla_x \mathcal{L}_{k+1}(x)|_{x=x_k}.$$

For a candidate displacement $-v$, the first-order change in the tracking loss is $-g_k^\top v$, while $\frac{1}{2} v^\top G(x_k) v$ approximates the KL divergence induced by the displacement, cf. Eq. (6). We therefore balance predicted loss reduction against local predictive change using the natural-gradient construction [2].

Because $G(x_k)$ may be singular or ill-conditioned, we additionally penalize the Euclidean size of the displacement through Tikhonov damping [12]. Without any additional regularization, the damped natural-gradient vector is defined, for $\lambda > 0$, by

$$v_k = \arg\min_{v \in \mathbb{R}^n} \left\{ \frac{1}{2} v^\top G(x_k)\, v + \frac{\lambda}{2} \|v\|_2^2 - g_k^\top v \right\}, \tag{10}$$

whose first-order optimality condition is

$$\big(G(x_k) + \lambda I\big)\, v_k = g_k. \tag{11}$$

FRInGe extends this system with both spatial regularization terms, as specified below in Eq. (17). We solve the resulting system using conjugate gradients (CG), evaluating Fisher matrix-vector products implicitly through Jacobian-Vector products (JVPs) and Vector-Jacobian products (VJPs). Implementation details are given in Appendix A.3.

The resulting update has the form

$$x_{k+1} = x_k - \eta_k v_k, \tag{12}$$

where η_k is selected using the step-size bounds below.

KL-based step-size bound. Following the use of KL constraints in natural-gradient and trust-region methods [12, 22], we introduce a hyperparameter $\tau > 0$ specifying the desired upper bound on the KL divergence between successive predictive distributions:

$$\mathrm{KL}\big(p(x_k - \eta v_k) \,\|\, p(x_k)\big) \le \tau.$$

Thus, τ controls the permitted predictive change per update. Smaller values prescribe more conservative updates, whereas larger values permit greater predictive changes but may reduce the accuracy of the local quadratic approximation. Applying this constraint to the quadratic model in Eq. (6) gives the approximate step-size cap

$$\eta_{\mathrm{KL}} = \sqrt{\frac{2\tau}{\|v_k\|^2_{G(x_k)} + \varepsilon}}, \tag{13}$$

where $\varepsilon > 0$ prevents division by zero. For a fixed search direction, this rule limits the predictive change estimated by the local Fisher quadratic model. Because the approximation is local, the resulting cap does not by itself guarantee satisfaction of the actual KL constraint.

Matching waypoint spacing to the KL budget. The KL budget τ controls the permitted predictive change per update, while the number of tracking steps T determines the spacing of the predictive waypoints. We couple these choices so that the distance between consecutive waypoints is consistent with the predictive change allowed by the local step-size rule.

Let $D_{\mathrm{total}} = d_{\mathrm{FR}}\big(p(x), q^\star\big)$ denote the Fisher–Rao distance from the initial prediction to the reference distribution. Under the local quadratic approximation, a KL budget of τ corresponds to a nominal Fisher–Rao displacement of $\sqrt{2\tau}$. For $D_{\mathrm{total}} > 0$, we therefore set

$$T = \left\lceil \frac{D_{\mathrm{total}}}{\sqrt{2\tau}} \right\rceil. \tag{14}$$

This choice couples the waypoint discretization to the KL budget, allocating more tracking steps when the reference distribution is farther from the initial prediction or when a smaller predictive change is permitted per update.

Euclidean displacement cap. The KL-based step-size rule controls local predictive change but does not directly bound the change in the input space. In directions of low predictive sensitivity, it may permit large input updates, across which the gradients used for attribution can vary substantially. To control the spacing of the input-space samples used for numerical integration, we introduce a hyperparameter $\delta_{\mathrm{euc}} > 0$ specifying the maximum Euclidean displacement per update. This gives the additional step-size bound

$$\eta_{\mathrm{Euc}} = \frac{\delta_{\mathrm{euc}}}{\|v_k\|_2 + \varepsilon}, \tag{15}$$

where also here $\varepsilon > 0$ prevents division by zero. The resulting update satisfies $\|x_{k+1} - x_k\|_2 \leq \delta_{\mathrm{euc}}$, thereby bounding the length of each segment of the integration path. For a fixed number of updates, a restrictive displacement cap may reduce progress toward the predictive waypoints, creating a trade-off between tracking progress and the resolution of the trajectory used for attribution. Appendices A.5 and A.6 analyze how this cap affects waypoint tracking and numerical integration error.

Bounding the step size. We combine the Euclidean displacement bound with the predictive-space bound and a maximum step size $\eta_{\max} > 0$ to select

$$\eta_k = \min\{\eta_{\max},\, \eta_{\mathrm{KL}},\, \eta_{\mathrm{Euc}}\}\,. \tag{16}$$

Spatial regularization of the input trajectory. Adversarial examples demonstrate that neural-network predictions can be altered through visually imperceptible perturbations [28], while convolutional networks can exploit high-frequency image components that are weakly aligned with human perception [33]. These observations motivate introducing a spatial smoothness prior into the predictive-waypoint tracking procedure.

Let L be a linear spatial differential operator, such as a discrete Laplacian. We use nonnegative coefficients $\gamma_{\mathrm{step}}, \gamma_{\mathrm{prior}} \geq 0$. Combining the pullback Fisher geometry, Euclidean damping, and spatial regularization yields the FRInGe update direction used in our experiments, defined by

$$\big(G(x_k) + \lambda I + \gamma_{\mathrm{step}} L^\top L\big) v_k = g_k + \gamma_{\mathrm{prior}} L^\top L x_k. \tag{17}$$

The coefficient γ_{step} controls the penalty $\|Lv_k\|_2^2$ on spatial variation in the update direction. Defining the input roughness penalty as $R(x) = \frac{1}{2}\|Lx\|_2^2$, the right-hand side combines the alignment gradient g_k with $\gamma_{\mathrm{prior}} \nabla_x R(x_k)$. The update therefore incorporates both predictive alignment and smoothness of update directions and inputs. For a symmetric discrete Laplacian, $L^\top L = L^2$ is a discrete biharmonic operator, corresponding to the squared-Laplacian smoothness energy [9].

In Section 6, we evaluate FRInGe with both spatial terms active on convolutional neural networks. Figure 1 illustrates the overall FRInGe construction.

Attribution computation. We integrate target-logit gradients along the piecewise-linear input trajectory through $\{x_k\}_{k=0}^{T}$. Starting from $A_{\mathrm{path}} = 0$, we apply the trapezoidal rule on each segment:

$$A_{\mathrm{path}} \leftarrow A_{\mathrm{path}} + \frac{1}{2}\Big(\nabla_x F^t(x_k) + \nabla_x F^t(x_{k+1})\Big) \odot (x_{k+1} - x_k). \tag{18}$$

We return $A = -A_{\mathrm{path}}$ to express the attributions in the conventional endpoint-to-input orientation. The attributions explain the change in the target logit between the original input x and the endpoint $x_T(x)$ reached by the algorithm. Exact integration along any piecewise smooth trajectory recovers this change, while numerical quadrature introduces a residual in the completeness relation $\sum_i A_i \approx F^t(x) - F^t(x_T(x))$. This relation concerns the endpoint actually reached, regardless of whether its predictive distribution matches $q^\star$. We therefore evaluate numerical completeness and predictive-target convergence separately, reporting both the completeness residual and the mismatch between $p(x_T(x))$ and $q^\star$. Appendix A.1 gives pseudocode for the full procedure. Figure 2 illustrates a representative FRInGe trajectory.

5 Computational cost and target-versus-rest specialization

FRInGe requires a CG solve at each tracking step. Each CG iteration evaluates a Fisher matrix-vector product using one JVP and one VJP. Appendix A.7 details the total computational cost.

To reduce this cost, we introduce FRInGe-B, a target-versus-rest specialization. It applies the FRInGe construction to the aggregated predictive distribution $\widetilde{p}_t(x) = (p_t(x), 1 - p_t(x))^\top$ and reference $\widetilde{q}_t^\star = (1/C, 1 - 1/C)^\top$. Defining the target-versus-rest log-odds as

$$\delta_t(x) = F^t(x) - \log\Big(\sum_{j \neq t} \exp F^j(x)\Big),$$

the corresponding pullback Fisher tensor is

$$G_{\mathrm{B}}(x) = p_t(x)(1 - p_t(x))\, \nabla_x \delta_t(x) \nabla_x \delta_t(x)^\top. \tag{19}$$

This tensor has rank at most one, independently of the number of classes. Substituting G_{B} into Eq. (11) therefore gives a closed-form unsmoothed direction, eliminating the repeated Fisher JVP/VJP products and CG solve. The spatially regularized variant uses solves with the fixed matrix $\lambda I + \gamma_{\mathrm{step}} L^\top L$, together with a rank-one correction. Unless stated otherwise, FRInGe-B also keeps both spatial terms active: the suffix B in FRInGe-B identifies target-versus-rest aggregation, not the presence or absence of spatial regularization.

The aggregation of the predictive distribution discards distinctions among non-target classes, so its trajectories can differ from those of FRInGe. Appendix F provides the detailed derivation, algorithm, and analysis of attribution quality and computational cost of FRInGe-B.

6 Experiments

Experimental setup. We evaluate FRInGe and its target-versus-rest specialization, FRInGe-B, on ImageNet validation images. For each image, we explain the model's top-1 predicted class. We report results on Inception-v3 (Inc-v3), ResNet-18 (RN18), ResNet-50, ResNet-101, ResNet-152, and VGG-19, all using standard pretrained weights and their default preprocessing pipelines. FRInGe and FRInGe-B are tuned independently. Image cohorts and reference-pool exclusions are specified in Appendix C.

We compare against ten attribution baselines: Integrated Gradients (IG) [27], IG^2 [38], SmoothGrad (SG) [24], BlurIG [34], GuidedIG [10], Adversarial IG (AGI) [15], Expected Gradients (EG) [7], GGIG [17], Geodesic Integrated Gradients (GeoIG) [19], and Manifold Integrated Gradients (MIG) [37]. Table 1 shows results for FRInGe, FRInGe-B, and a subset of these methods; Appendix D.2 reports the complete comparison for all six architectures.

To assess the contributions of pullback geometry and spatial regularization, we additionally report ablations comparing Euclidean distribution tracking, Fisher–Rao tracking without spatial regularization, a Fisher-aware prior-only variant, and FRInGe with two-term spatial regularization. Implementation details and hyperparameter settings are provided in the appendix. Code is available at `https://github.com/GabMartino/FRInGe`.

Evaluation protocol and metrics. We evaluate each attribution map using four complementary criteria. First, for *perturbation faithfulness*, we report insertion (Ins-AUC) and deletion (Del-AUC) area under the curve under blur-based perturbations [16], together with Infidelity [35], which measures agreement between attribution-weighted input perturbations and the corresponding changes in model output. Second, for *attribution calibration*, we report Magnitude Aligned Scoring for insertion (MAS-Ins) and deletion (MAS-Del) [32], which assesses whether cumulative attribution mass aligns with changes in model confidence. Perturbation AUC primarily evaluates feature ranking, while MAS additionally accounts for attribution magnitudes. Third, for *local robustness*, we report Max Sensitivity [35] under small input perturbations in the appendix. Fourth, for *explanation focus*, we report Gini sparseness [8] as a descriptive measure of attribution concentration. Full metric definitions and perturbation details are given in Appendix C.4.

6.1 Comparison against baseline methods

Table 1 and the full benchmark in Appendix D.2 (Tables 8–13) highlight FRInGe's strong performance on attribution magnitude alignment. Against the ten external baselines, FRInGe achieves the highest mean MAS-Ins on five of six architectures. Its gains over geometric competitors are particularly consistent: FRInGe improves both mean MAS-Ins and MAS-Del over GGIG and GeoIG on every

Table 1: Attribution performance on representative architectures (mean ± half-width of the 95% bootstrap interval). The best mean score in each row is **bolded**; methods whose displayed mean ± half-width ranges overlap with that of the best are **bold**.

Metric	Arch.	FRInGe (Ours)	FRInGe-B (Ours)	IG	SG	BlurIG	GGIG	GeoIG
MAS-Ins ↑	RN18	**0.597±.015**	**0.601±.014**	0.443±.015	0.488±.014	**0.577±.014**	0.426±.016	0.453±.014
	Inc-v3	**0.761±.009**	0.743±.008	0.673±.010	0.660±.009	0.735±.009	0.631±.011	0.647±.009
MAS-Del ↓	RN18	**0.264±.010**	**0.259±.008**	0.359±.011	0.325±.004	0.276±.008	0.373±.008	0.359±.007
	Inc-v3	0.444±.020	0.416±.018	0.485±.019	**0.373±.008**	**0.386±.017**	0.473±.015	0.469±.016
Ins-AUC ↑	RN18	0.492±.016	**0.541±.015**	0.415±.014	0.477±.015	**0.563±.015**	0.404±.016	0.430±.014
	Inc-v3	0.709±.014	0.723±.013	0.647±.014	0.690±.013	**0.758±.013**	0.658±.015	0.668±.014
Del-AUC ↓	RN18	0.130±.008	**0.095±.006**	0.178±.010	0.122±.007	**0.096±.007**	0.164±.009	0.135±.008
	Inc-v3	0.240±.012	0.208±.011	0.271±.013	0.206±.011	**0.174±.011**	0.215±.011	0.216±.011

architecture. Figure 3 illustrates selected qualitative comparisons against these baselines, highlighting differences in attribution concentration and boundary detail.

Among the external baselines, BlurIG leads Ins-AUC and Del-AUC. A plausible reason is the close alignment between its Gaussian scale-space path and the blur perturbations used for evaluation, which may favor its feature ordering under this protocol. FRInGe's leading MAS-Ins results under the same perturbation protocol highlight its strength when evaluation additionally accounts for attribution magnitudes.

FRInGe-B extends FRInGe with a computationally simpler target-versus-rest formulation. When independently tuned, it improves mean MAS-Del and both Ins-AUC and Del-AUC over FRInGe on five architectures; FRInGe retains higher mean MAS-Ins on five architectures (Table 18; see Figure 20 for qualitative comparisons).

6.2 Ablation study

We compare Euclidean tracking, Fisher–Rao tracking without spatial regularization ($\gamma_{\text{step}} = \gamma_{\text{prior}} = 0$), a prior-only variant ($\gamma_{\text{step}} = 0$, keeping all other hyperparameters fixed to FRInGe's tuned values in Table 6), and FRInGe with two-term spatial regularization. Table 7 in the appendix reports quantitative results; trajectory diagnostics and additional smoothing ablations appear in Appendix B.

Among these variants, FRInGe achieves the best mean MAS-Ins, MAS-Del, and Ins-AUC on all six architectures. Del-AUC results are mixed between FRInGe and prior-only. The Euclidean and Fisher variants without spatial regularization produce sparser maps but weaker MAS and AUC results. Trajectory diagnostics further show that two-term spatial regularization suppresses speckled input transformations and yields smoother, more spatially coherent changes, supporting its role in the tracking procedure.

6.3 Hyperparameter sensitivity and local robustness

We assess FRInGe's hyperparameter sensitivity through one-factor-at-a-time sweeps over δ_{euc}, τ, $\eta_{\max}$, λ, and a tied spatial regularization weight $\gamma_{\text{step}} = \gamma_{\text{prior}} = \gamma$, while keeping the remaining hyperparameters fixed (Appendix C.7). We additionally examine how the Euclidean cap affects numerical completeness and predictive-space tracking, and evaluate local robustness against IG on 64 images per architecture using Max Sensitivity [35], which measures the maximum relative change in ℓ_2-normalized attributions under small input perturbations (Appendix C.4). The selected architecture-specific hyperparameters are held fixed across images.

Over the tested ranges, the tuning objective (a weighted harmonic mean of Ins-AUC, $1 -$ Del-AUC, and Infidelity; Appendix C) is comparatively insensitive to $\eta_{\max}$ and stable across a range of τ values, with weaker, architecture-dependent effects from λ. The strongest sensitivities arise for δ_{euc} and γ, where moderate nonzero spatial regularization ($\gamma > 0$) generally improves this score. Activating the Euclidean cap typically reduces the realized-path completeness residual, but increases endpoint mismatch and slightly worsens waypoint tracking (Appendix A.6).

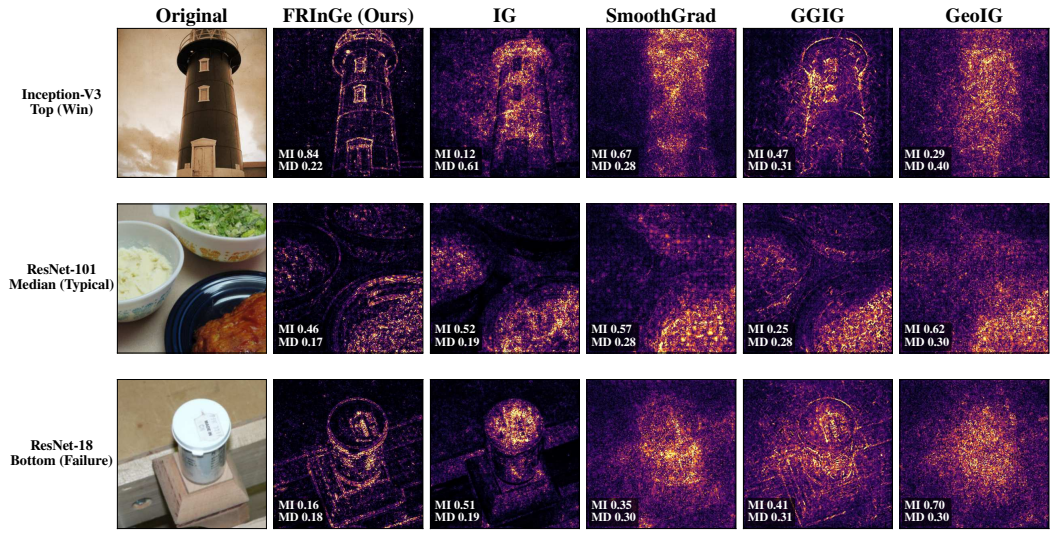

Figure 3: Selected qualitative comparisons on Inception-v3 (top), ResNet-101 (middle), and ResNet-18 (bottom). The ResNet-18 example illustrates reduced boundary detail associated with smoothing. Each example reports MAS-Insertion (MI) and MAS-Deletion (MD).

Local robustness varies between architectures. FRInGe achieves lower median Max Sensitivity than IG on ResNet-18 and VGG-19, while IG achieves lower medians on the other four architectures (Appendix D.3, Table 14).

7 Discussion and Conclusion

In this paper, we proposed FRInGe which introduces a predictive-space formulation of path-based attribution, replacing a prescribed baseline input with an explicit reference distribution. It constructs a Fisher–Rao geodesic toward a maximum-entropy reference and computes an input-space trajectory whose predictions approximately track its waypoints on the path. The pullback Fisher geometry guides this realization, while step caps and spatial regularization stabilize and shape the trajectory.

The clearest empirical advantage of FRInGe appears in MAS-Ins: FRInGe achieves the highest mean among ten external baselines on five of six architectures. It also improves both mean MAS-Ins and MAS-Del over GGIG and GeoIG across all six architectures we benchmarked it on. These results support stronger alignment between attribution magnitudes and model confidence under the tested protocols. BlurIG leads Ins-AUC and Del-AUC, an advantage that may partly reflect the alignment between its Gaussian scale-space path and the evaluation perturbations.

FRInGe-B provides a computationally simpler specialization through target-versus-rest aggregation. Its rank-one pullback removes repeated model-dependent Fisher matrix-vector products from the inner solve. The two formulations offer distinct performance profiles: FRInGe retains information about the complete class distribution and achieves higher MAS-Ins and sparseness, while FRInGe-B reduces the computational burden of constructing a target-specific trajectory and improves MAS-Del, Ins-AUC, and Del-AUC on most architectures.

Several limitations remain. Finite-step tracking can leave a predictive endpoint mismatch, and completeness applies to the realized input endpoint, up to numerical integration error. Damping, spatial regularization, and Euclidean displacement caps also make the realized trajectory dependent on the input representation. Applications beyond the image models evaluated here require appropriate regularizers and separate validation.

More broadly, FRInGe and FRInGe-B establish two instances of a general distribution-guided attribution framework. In this framework, the target predictive reference and class aggregation become explicit design choices, while the corresponding input endpoint emerges from the trajectory tracking procedure. Alternative predictive references, class groupings, and modality-specific regularizers provide concrete directions for extending our work.

Acknowledgments and Disclosure of Funding

This work has been funded in parts by the Vienna Science and Technology Fund (WWTF) [10.47379/LS23053].

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631. ACM, 2019.

[2] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[3] Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer Japan, 2016.

[4] Georgios Arvanitidis, Miguel González-Duque, Alison Pouplin, Dimitris Kalatzis, and Søren Hauberg. Pulling back information geometry. *arXiv preprint arXiv:2106.05367*, 2021.

[5] Jai Bardhan, Cyrin Neeraj, Mihir Rawat, and Subhadip Mitra. Constructing sensible baselines for integrated gradients. *arXiv preprint arXiv:2412.13864*, 2024.

[6] N. N. Čencov. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1982.

[7] 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.

[8] Niall Hurley and Scott Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

[9] Alec Jacobson, Ilya Baran, Jovan Popovic, and Olga Sorkine. Bounded biharmonic weights for real-time deformation. *ACM Trans. Graph.*, 30(4):78, 2011.

[10] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5050–5058. IEEE, 2021.

[11] Shuyang Liu, Zixuan Chen, Ge Shi, Ji Wang, Changjie Fan, Yu Xiong, Runze Wu Yujing Hu, Ze Ji, and Yang Gao. A new baseline assumption of integated gradients based on shaply value. *arXiv preprint arXiv:2310.04821*, 2023.

[12] James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

[13] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15, 2018.

[14] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

[15] Deng Pan, Xin Li, and Dongxiao Zhu. Explaining deep neural network models with adversarial gradient integration. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[16] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, page 151. BMVA Press, 2018.

[17] Md Mahfuzur Rahman, Noah Lewis, and Sergey Plis. Geometrically guided saliency maps. In *ICLR 2022 Workshop on PAIR^{2}Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022.

[18] Yuchen Ren, Zhengyu Zhao, Chenhao Lin, Bo Yang, Lu Zhou, Zhe Liu, and Chao Shen. Improving integrated gradient-based transferable adversarial examples by refining the integration path. In *Conference on Artificial Intelligence (AAAI)*, volume 39, pages 6731–6739, 2025.

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

[20] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, 2022.

[21] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, Shawn Xu, Scott Barb, Anthony Joseph, Michael Shumski, Jesse Smith, Arjun B. Sood, Greg S. Corrado, Lily Peng, and Dale R. Webster. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*, 126(4):552–564, 2019.

[22] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. Pmlr, 2015.

[23] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, 1985.

[24] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

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

[26] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International Conference on Machine Learning (ICML)*, pages 9269–9278, 2020.

[27] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, pages 3319–3328, 2017.

[28] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[29] Hanxiao Tan. Maximum entropy baseline for integrated gradients. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2023.

[30] Kien Tran Duc Tuan, Tam Nguyen Trong, Son Nguyen Hoang, Khoat Than, and Anh Nguyen Duc. Enhancing visual feature attribution via weighted integrated gradients. *arXiv preprint arXiv:2505.03201*, 2025.

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

[32] Chase Walker, Dominic Simon, Kenny Chen, and Rickard Ewetz. Attribution quality metrics with magnitude alignment. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 530–538, 2024.

[33] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8681–8691. IEEE, 2020.

[34] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9680–9689. IEEE, 2020.

[35] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in)fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[36] Wanming Yu, Chuanyu Yang, Christopher McGreavy, Eleftherios Triantafyllidis, Guillaume Bellegarda, Milad Shafiee, Auke Jan Ijspeert, and Zhibin Li. Identifying important sensory feedback for learning locomotion skills. *Nature Machine Intelligence*, 5(8):919–932, 2023.

[37] Eslam Zaher, Maciej Trzaskowski, Quan Nguyen, and Fred Roosta. Manifold integrated gradients: Riemannian geometry for feature attribution. In *International Conference on Machine Learning (ICML)*, pages 58090–58104, 2024.

[38] Yue Zhuo and Zhiqiang Ge. IG^2: Integrated gradient on iterative gradient path for feature attribution. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(11):7173–7190, 2024.

Appendix Contents

A FRInGe Methodological Details **15**

A.1 FRInGe Pseudocode . . . 15

A.2 Matrix-Free Pullback Fisher Products . . . 15

A.3 Spatial Regularization and Numerical Solution . . . 16

A.4 Trust-Region Step Selection . . . 16

A.5 Geometric Effect of the Euclidean Trust-Region Cap . . . 16

A.6 Quantitative Effect of the Euclidean Trust-Region Cap . . . 18

A.7 Computational Cost and Measured Latency . . . 19

B Ablation Study Configurations **20**

B.1 Shared Protocol and Configuration Map . . . 20

B.2 Geometry and Joint Spatial Regularization . . . 21

B.3 Separating the Two Spatial Terms . . . 22

C Experimental Protocol and Reproducibility **24**

C.1 Dataset, Models, and Target Selection . . . 24

C.2 Attribution Baselines . . . 24

C.3 Evaluation Pipeline and Statistical Reporting . . . 25

C.4 Metric Definitions . . . 25

C.5 Hyperparameter Selection Protocol . . . 27

C.6 Architecture-Specific Hyperparameters . . . 28

C.7 Hyperparameter Sensitivity Across Architectures . . . 28

D Additional Quantitative Results **29**

D.1 Ablation Study . . . 30

D.2 Comparison across attribution methods . . . 30

D.3 Robustness under small input perturbations . . . 33

E Additional Qualitative Results **33**

F Target-vs-Rest Fisher–Rao Specialization (FRInGe-B) **37**

F.1 Exact Binary Predictive Geometry . . . 37

F.2 Rank-One Direction and Spatial Regularization . . . 38

F.3 Information Discarded by Aggregation . . . 39

F.4 FRInGe-B Algorithm . . . 39

F.5 Computational Cost . . . 40

F.6 Empirical Comparisons . . . 40

F.7 Endpoint Diagnostics and Limitations . . . 42

F.8 Scope and Modality-Specific Priors . . . 43

A FRInGe Methodological Details

This appendix explains how the predictive-space construction in Section 4 is realized numerically and how to interpret the resulting trajectory. We begin with the complete algorithm, then explain matrix-free Fisher products, the distinct roles of the spatial penalties, and the approximate solution of the regularized system. We next examine how the predictive and Euclidean step-size constraints interact, using trajectory profiles and quantitative diagnostics to distinguish accuracy of the path integral from agreement with the prescribed predictive waypoints. Finally, we give the leading computational cost and measured latency of categorical FRInGe. Together, these details explain which approximations affect the update direction, the predictive endpoint, and the accumulated attributions.

A.1 FRInGe Pseudocode

Algorithm 1 combines waypoint construction, input-space tracking, and attribution accumulation. The target class t remains fixed throughout. At each input state, the direction is computed toward the next predictive waypoint, and the step size is then restricted before the target-logit gradients are integrated along the resulting input segment. The trajectory runs from the original input toward the realized endpoint; the final sign reversal gives attributions oriented from that endpoint back to the original input, consistently with Eq. (2).

Algorithm 1 FRInGe

Require: Input x, target class t, logits F; damping λ, KL budget τ, step cap $\eta_{\max}$, Euclidean cap δ_{euc}; spatial operator L and coefficients $\gamma_{\text{step}}, \gamma_{\text{prior}}$
Returns: Attributions A

Compute T using Eq. (14)
$p_0 \leftarrow \mathrm{softmax}(F(x))$, $q^\star \leftarrow \frac{1}{C}\mathbf{1}$
Build Fisher–Rao waypoints $\{s_k\}_{k=0}^{T}$ between $\sqrt{p_0}$ and $\sqrt{q^\star}$
$x_0 \leftarrow x$, $A_{\text{path}} \leftarrow 0$
for $k = 0, \ldots, T-1$ **do**
  $p_k \leftarrow \mathrm{softmax}(F(x_k))$
  $\mathcal{L}_{k+1} \leftarrow 1 - \langle \sqrt{p_k}, s_{k+1} \rangle$
  $g_k \leftarrow \nabla_x \mathcal{L}_{k+1}(x)|_{x_k}$
  Solve $(G_k + \lambda I + \gamma_{\text{step}} L^\top L) v_k = g_k + \gamma_{\text{prior}} L^\top L x_k$ for v_k by CG
  $n_k^2 \leftarrow v_k^\top G_k v_k$
  $\eta_k \leftarrow \min\Big(\eta_{\max}, \sqrt{\frac{2\tau}{n_k^2+\varepsilon}}, \frac{\delta_{\text{euc}}}{\|v_k\|_2+\varepsilon}\Big)$
  $x_{k+1} \leftarrow x_k - \eta_k v_k$
  $A_{\text{path}} \leftarrow A_{\text{path}} + \frac{1}{2}(\nabla_x F^t(x_k) + \nabla_x F^t(x_{k+1})) \odot (x_{k+1} - x_k)$
end for
return $A = -A_{\text{path}}$

A.2 Matrix-Free Pullback Fisher Products

The pullback tensor in Eq. (5) measures local predictive sensitivity, not Euclidean input distance. Its quadratic form can be small even for a large input displacement if the classifier is locally insensitive in that direction. This distinction motivates both damping in the linear system and the separate input-space cap below.

The numerical solver needs only the action of this tensor on a vector, not the dense $n \times n$ matrix. Using the logit Jacobian $J_F(x)$ and the softmax Fisher matrix $S(x)$ defined in the background, we evaluate

$$G(x)v = J_F(x)^\top \big(S(x)(J_F(x)v)\big).$$

A Jacobian-vector product maps an input direction into logit space, the softmax Fisher matrix weights its predictive effect, and a vector-Jacobian product maps the result back to input space. This avoids storing the dense input-space tensor, although every such product still requires model-dependent differentiation. That repeated derivative work, rather than matrix storage alone, determines much of the computational cost.

A.3 Spatial Regularization and Numerical Solution

At state x_k, the regularized direction satisfies

$$\big(G(x_k) + \lambda I + \gamma_{\text{step}} L^\top L\big) v_k = g_k + \gamma_{\text{prior}} L^\top L x_k,$$

where

$$g_k = \nabla_x \mathcal{L}_{k+1}(x)|_{x=x_k},$$

$\lambda > 0$ supplies Euclidean damping, and L measures spatial roughness. For the image models, L is a discrete Laplacian, so the quadratic penalty acts through $L^\top L$. These spatial terms express an image-specific prior; they are not required by the predictive Fisher–Rao construction.

The two nonnegative coefficients serve different purposes. γ_{step} penalizes rough update directions through the left-hand side, whereas γ_{prior} introduces the current input's roughness gradient on the right-hand side. With the update $x_{k+1} = x_k - \eta_k v_k$, the latter term encourages removal of spatial roughness from the current iterate. Consequently, smoothing the direction and encouraging a smoother input trajectory are distinct choices. Setting $\gamma_{\text{step}} = \gamma_{\text{prior}} = \gamma$ is a tied-coefficient special case, not a requirement of the method.

We approximate the direction using preconditioned conjugate gradients (PCG). In the spatially regularized setting, the empirical preconditioner attenuates high-frequency components by Gaussian blurring and scaling:

$$M^{-1}(v) \approx \frac{\text{Blur}(v)}{\lambda}.$$

This is a numerical choice for the iterative solve, not an exact inverse of $L^\top L$ or a replacement for the spatial penalties. When smoothing is disabled, it reduces to diagonal scaling,

$$M^{-1}(v) = \frac{v}{\lambda}.$$

The preceding waypoint's direction initializes the next solve. A finite iteration ceiling and residual tolerances limit the work per waypoint, so the computed direction need not solve the displayed system exactly. Linear-solve accuracy, predictive-waypoint tracking, and path-integral accuracy are therefore separate numerical considerations.

A.4 Trust-Region Step Selection

Once v_k is computed, Eq. (16) selects the smallest of the global step-size ceiling, the local KL-based bound, and the Euclidean displacement bound. These constraints act on different quantities. The KL-based bound limits predictive change according to the local Fisher quadratic model; it is not an exact constraint on the nonlinear model's KL divergence. The Euclidean bound instead limits the length of the input segment used for numerical integration. It can be restrictive even when the predicted change in the output distribution is small.

The active constraint can change along a trajectory. In regions of low predictive sensitivity, the Euclidean cap may limit the input displacement; elsewhere, the Fisher-based bound or the global ceiling may determine the step. Shorter input segments can improve the resolution of the path integral, but they need not improve tracking of the prescribed waypoints within a fixed number of updates. In particular, T is chosen from the nominal predictive arc length before tracking begins. Neither that choice nor the step-size rule guarantees that the realized endpoint predicts the reference distribution.

A.5 Geometric Effect of the Euclidean Trust-Region Cap

Figures 4, 5, and 6 illustrate the interaction between the two step constraints through three example trajectories: standard IG, FRInGe without the Euclidean cap, and FRInGe with both bounds. Each profile shows the local Fisher step, the Euclidean input displacement, and normalized predictive entropy. These examples illustrate possible trajectory regimes rather than establishing their prevalence across images.

For the straight-line IG path in Fig. 4, equal Euclidean steps produce highly uneven predictive changes. The Fisher step stays close to zero over much of the path and rises sharply near its endpoint, while entropy increases mainly late in the trajectory. The displayed orientation runs from the original

input to the baseline, opposite to the usual attribution-integration direction. This example shows why uniform sampling in input space need not resolve predictive changes uniformly.

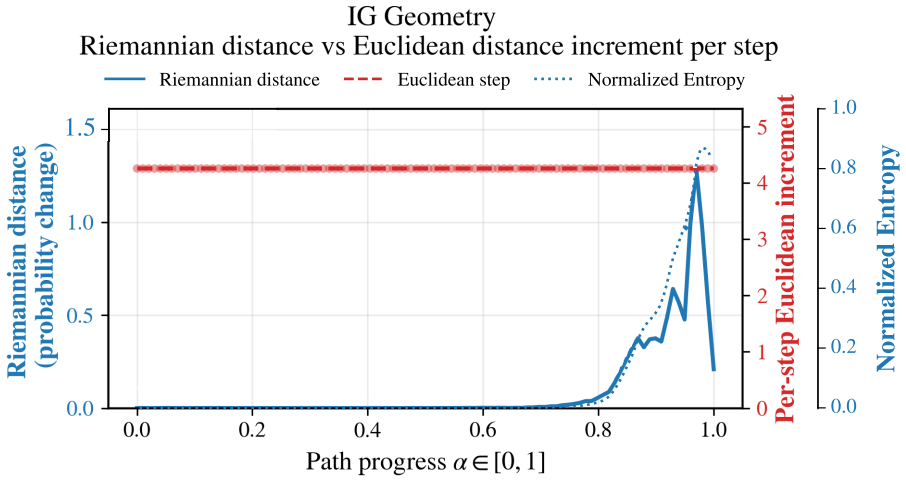

Figure 4: Geometric profile of standard IG. Constant Euclidean steps induce highly non-uniform predictive-space progress, with entropy increasing mainly near the end of the displayed path.

Without the Euclidean cap, Fig. 5 shows a more uniform local Fisher step and a smoother entropy profile. However, maintaining the predictive-space step in regions of low sensitivity can require large input displacements. The Euclidean step norm reaches approximately 50 in this example. Such long segments can leave substantial variation in target-logit gradients unresolved by the quadrature, even when the local predictive step appears controlled.

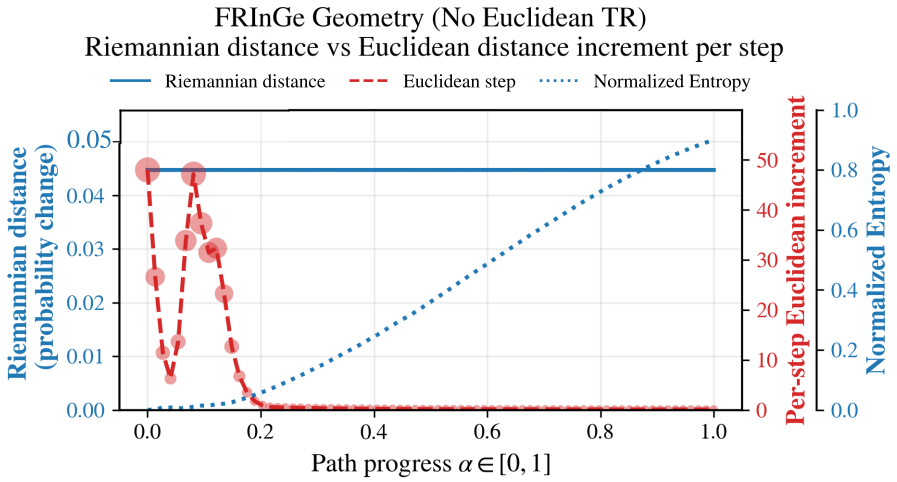

Figure 5: Geometric profile of FRInGe without the Euclidean cap. Approximately constant Fisher–Rao steps coexist with large early displacements in input space; normalized predictive entropy rises more smoothly along the trajectory. Larger Euclidean steps are required in flat geometry regions.

With the Euclidean cap, Fig. 6 initially limits input displacements to approximately 10. Later, the KL-based bound becomes more restrictive, and the local Fisher increment is approximately constant instead. The profile thus illustrates a transition between input-space and predictive-space step control.

The cap's purpose is to stabilize the numerical integration path, not to guarantee closer waypoint tracking. The quantitative comparison in Appendix A.6 examines this distinction across models.

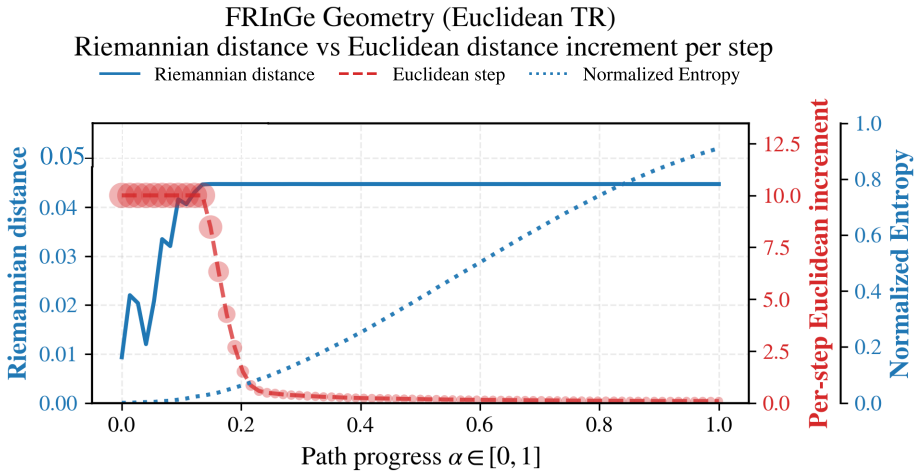

Figure 6: Geometric profile of FRInGe with the Euclidean cap. The cap limits early input-space displacements to approximately 10; later, the KL constraint limits the step and the Fisher–Rao increment becomes approximately constant.

A.6 Quantitative Effect of the Euclidean Trust-Region Cap

The Euclidean trust-region radius δ_{euc} introduces a trade-off between numerical integration accuracy and predictive-space agreement. We assess these separately through realized-path completeness, endpoint mismatch, and waypoint tracking.

First, we evaluate the *realized-path completeness residual*, measured with respect to the endpoint actually reached by the discrete trajectory:

$$\varepsilon_{\text{comp}} = \frac{\left|\sum_i A_i - \left(F^t(x_0) - F^t(x_T)\right)\right|}{|F^t(x_0) - F^t(x_T)| + \varepsilon}.$$

Here $\varepsilon > 0$ prevents division by zero. This residual measures numerical integration error relative to the endpoint actually reached, not agreement with the desired predictive reference. We also report the endpoint $\mathrm{KL}\big(p(x_T) \,\|\, u\big)$, with u denoting the uniform reference. A small completeness residual can therefore coexist with a large endpoint mismatch.

Second, we quantify predictive-space waypoint tracking by measuring the Fisher–Rao distance between the realized predictive distribution and the intended geodesic waypoint:

$$\varepsilon_{\text{track},k} = d_{\text{FR}}(p(x_k), q_k).$$

We summarize this tracking error using both its mean and its maximum over the trajectory.

Tables 2 and 3 show a consistent pattern across the five models with a restrictive Euclidean cap (Inception-v3 is omitted as its tuned cap $\delta_{\text{euc}} = 24.18$ is rarely active). Activating δ_{euc} often reduces the completeness residual of the realized input-space integral, but it also tends to increase the final endpoint KL and to worsen predictive-space waypoint tracking. The Euclidean cap should therefore be interpreted as a stabilizer of the realized integration path rather than as a mechanism for exact nonlinear shooting toward the reference distribution.

Table 2: Effect of the Euclidean trust-region cap on realized-path completeness and endpoint mismatch. Lower completeness residual is better. Lower endpoint KL indicates closer agreement with the ideal maximum-entropy target.

Model	Completeness error δ_{euc} on	Completeness error δ_{euc} off	Endpoint KL δ_{euc} on	Endpoint KL δ_{euc} off
ResNet-18	$\mathbf{0.011 \pm 0.003}$	0.078 ± 0.034	3.697 ± 0.605	$\mathbf{0.148 \pm 0.030}$
ResNet-50	$\mathbf{0.417 \pm 0.174}$	0.535 ± 0.168	2.749 ± 0.473	$\mathbf{2.435 \pm 0.474}$
ResNet-101	$\mathbf{0.184 \pm 0.061}$	0.313 ± 0.108	2.096 ± 0.530	$\mathbf{1.439 \pm 0.383}$
ResNet-152	$\mathbf{0.208 \pm 0.070}$	0.239 ± 0.071	1.235 ± 0.328	$\mathbf{1.161 \pm 0.288}$
VGG-19	$\mathbf{0.014 \pm 0.005}$	0.258 ± 0.360	4.707 ± 0.603	$\mathbf{0.188 \pm 0.088}$

Table 3: Effect of the Euclidean trust-region cap on predictive-space waypoint tracking. Lower values indicate closer agreement between the realized trajectory and the intended Fisher–Rao waypoint sequence.

Model	Mean tracking error δ_{euc} on	Mean tracking error δ_{euc} off	Max tracking error δ_{euc} on	Max tracking error δ_{euc} off
ResNet-18	0.197 ± 0.022	$\mathbf{0.046 \pm 0.008}$	0.977 ± 0.015	$\mathbf{0.850 \pm 0.088}$
ResNet-50	0.154 ± 0.017	$\mathbf{0.141 \pm 0.017}$	0.563 ± 0.083	$\mathbf{0.540 \pm 0.087}$
ResNet-101	0.145 ± 0.035	$\mathbf{0.124 \pm 0.035}$	0.859 ± 0.073	$\mathbf{0.817 \pm 0.086}$
ResNet-152	0.093 ± 0.014	$\mathbf{0.090 \pm 0.013}$	0.711 ± 0.099	$\mathbf{0.700 \pm 0.101}$
VGG-19	0.254 ± 0.024	$\mathbf{0.054 \pm 0.014}$	0.917 ± 0.035	$\mathbf{0.556 \pm 0.119}$

A.7 Computational Cost and Measured Latency

Leading-order model evaluations. Let C_{fwd} and C_{bwd} denote the costs of a model forward and reverse-mode pass. A pullback-Fisher product requires one JVP and one VJP, with costs comparable to C_{fwd} and C_{bwd}, respectively. Let N be the number of IG integration steps and T the number of FRInGe tracking steps. Let M denote the average number of pullback-Fisher products evaluated per categorical waypoint, including the products required by CG and the trust-region norm. Thus $M = O(K)$ when K is the CG iteration ceiling, but M need not equal K exactly.

Table 4 separates classifier-dependent derivative work from structured spatial-regularizer operations. The accounting treats a gradient evaluation as its forward–reverse pair and omits lower-order setup, solver bookkeeping, and opportunities for graph reuse or vectorization.

Table 4: Leading classifier-dependent work for one attribution. Spatial regularizer operations are accounted for separately in the text.

Method	Steps	Work per step	Approximate classifier cost
IG	N	target-score gradient	$N(C_{\mathrm{fwd}} + C_{\mathrm{bwd}})$
FRInGe	T	tracking and target-score gradients; M Fisher products	$T(M+2)(C_{\mathrm{fwd}} + C_{\mathrm{bwd}})$

Under the rough approximation $C_{\mathrm{fwd}} \approx C_{\mathrm{bwd}}$, this gives

$$\frac{\mathcal{C}_{\mathrm{FRInGe}}}{\mathcal{C}_{\mathrm{IG}}} \approx \frac{T}{N}(M+2). \tag{20}$$

The ratio counts classifier-dependent evaluations, not wall-clock time. With spatial smoothing, FRInGe also applies the fixed regularizer inside CG and to the current input; these operations do not evaluate the classifier.

Nominal waypoint bound. The largest Fisher–Rao distance to the uniform reference is approached by a one-hot prediction:

$$D_{\max} = 2 \arccos\left(\frac{1}{\sqrt{C}}\right) < \pi. \tag{21}$$

The nominal update count is

$$T_{\max} = \left\lceil \frac{2\arccos(1/\sqrt{C})}{\sqrt{2\tau}} \right\rceil \leq \left\lceil \frac{\pi}{\sqrt{2\tau}} \right\rceil . \tag{22}$$

The expression inside the first ceiling approaches the bound as $C \to \infty$. Thus the nominal count is bounded with respect to the number of classes, but still scales as $\tau^{-1/2}$. The realized trajectory may fail to reach the nominal endpoint when the Euclidean cap is active; this affects endpoint error, not the precomputed number of waypoints.

Runtime–fidelity controls. The KL budget τ controls waypoint resolution, while K bounds the categorical Fisher solve. Smaller τ increases path resolution and runtime. Larger solver ceilings can reduce linear-system error but do not guarantee better attribution quality. The benchmark solver uses a previous-waypoint warm start and stopping tolerances, so the ceiling need not equal the realized iteration count. The waypoint resolution and linear-solve accuracy thus control different sources of approximation; neither increased runtime nor a larger nominal iteration budget alone establishes better attributions.

FRInGe versus IG-300. Table 5 compares categorical FRInGe with 300-step Gauss–Legendre IG on the same 64 ImageNet inputs. Both methods process one image at a time on an NVIDIA H100 GPU under matched preprocessing conditions. Gradient evaluations for IG's integration steps can be batched, whereas FRInGe's tracking steps are sequential. The measured latency therefore reflects both derivative work and the available parallelism; it is a comparison of the evaluated configurations, not a universal speed ratio.

Table 5: Per-image attribution latency under batch-size-one execution. Values are means ± 95% confidence intervals across 64 images. The slowdown is the ratio of mean latencies.

Model	FRInGe (s)	IG-300 (s)	Slowdown
ResNet-18	55.12 ± 0.80	2.90 ± 0.01	19.0 ×
ResNet-50	45.00 ± 0.41	9.53 ± 0.03	4.7 ×
ResNet-101	228.21 ± 1.58	11.59 ± 0.04	19.7 ×
ResNet-152	319.21 ± 2.89	15.28 ± 0.01	20.9 ×
Inception-v3	159.80 ± 3.58	14.94 ± 0.03	10.7 ×
VGG-19	43.65 ± 0.58	1.95 ± 0.01	22.3 ×

The table confirms that the categorical pullback solve is a substantial practical cost. Appendix F examines how the binary specialization reduces this cost.

B Ablation Study Configurations

The ablations ask which parts of the input-space update are needed in addition to a prescribed predictive path. We first compare Euclidean waypoint tracking, damped Fisher tracking without spatial penalties, and FRInGe with two-term spatial regularization. We then separate the direction penalty from the current-input prior and include a non-geometric prior control. The six trajectories illustrate the consequences on one ResNet-18 input; the dataset-level comparison in Table 7 assesses whether those visual differences correspond to improved metric values. All variants retain predictive-waypoint tracking, so this study does not ablate the presence of that tracking mechanism.

B.1 Shared Protocol and Configuration Map

The reconstructed trajectories in Figures 7–12 use a shared ResNet-18 hyperparameters, changing the indicated geometry and spatial terms. Parameters associated with a disabled operator are inactive. These illustrations should be distinguished from the dataset-level ablation: its three core configurations use their architecture-specific settings in Table 6, whereas the additional prior-only configuration inherits FRInGe's settings and disables only γ_{step}, without retuning.

Configuration	Fisher geometry	Direction penalty	Input prior
Euclidean tracking	No	No	No
Fisher–Rao without spatial regularization	Yes	No	No
FRInGe	Yes	Yes	Yes
Direction-only FRInGe	Yes	Yes	No
Prior-only FRInGe	Yes	No	Yes
Euclidean tracking with prior	No	No	Yes

Each figure uses the same visual layout: normalized predictive entropy, selected intermediate inputs, spatial blur-change maps relative to the initial input, and the final signed attribution. Grey denotes zero attribution, blue negative values, and red positive values. Blur-change maps share a colour scale across the displayed steps within each figure, but not across ablations. Their spatial pattern can therefore be compared within a trajectory; colour intensity alone is not a common measure of change across configurations.

B.2 Geometry and Joint Spatial Regularization

Euclidean distribution tracking. This control follows the spherical-loss gradient directly,

$$v_k = g_k,$$

without a Fisher solve, damping, or spatial penalty. Figure 7 shows that predictive entropy can change substantially while the intermediate inputs remain visually close to the original. The blur-change maps and final attribution are speckled in this example. Predictive-waypoint tracking alone therefore need not produce spatially coherent image transformations; visual similarity of successive inputs is not evidence that their predictions are unchanged.

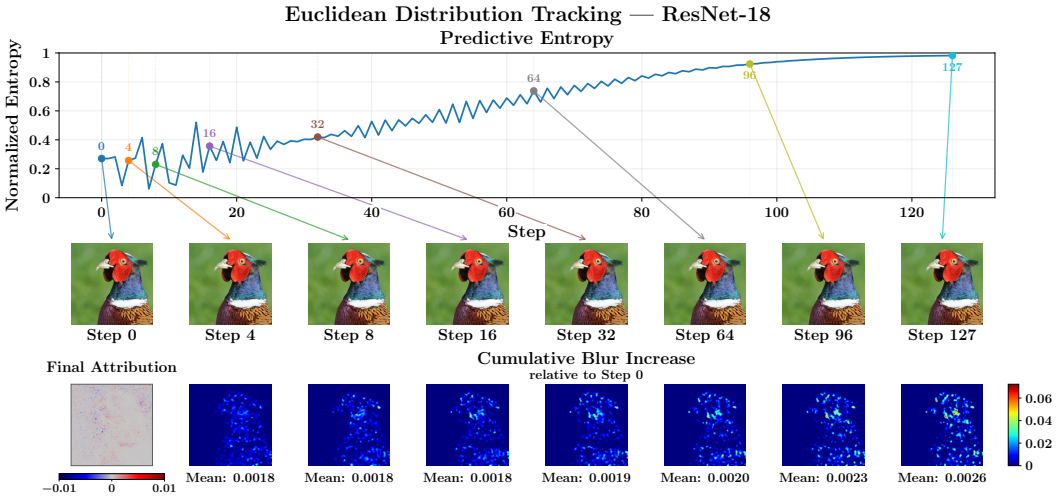

Figure 7: Euclidean predictive-waypoint tracking on one ResNet-18 input. The configuration omits Fisher conditioning and both spatial penalties.

Fisher–Rao tracking without spatial regularization. This variant retains damping and predictive step control but removes both spatial penalties:

$$\big(G(x_k) + \lambda I\big)v_k = g_k.$$

In Figure 8, entropy evolves more smoothly than in the Euclidean example, but the spatial changes and final map remain finely speckled. The example separates a smoother predictive profile from a smoother input-space transformation: the former does not by itself ensure the latter. It does not establish a general quality advantage for Fisher conditioning alone.

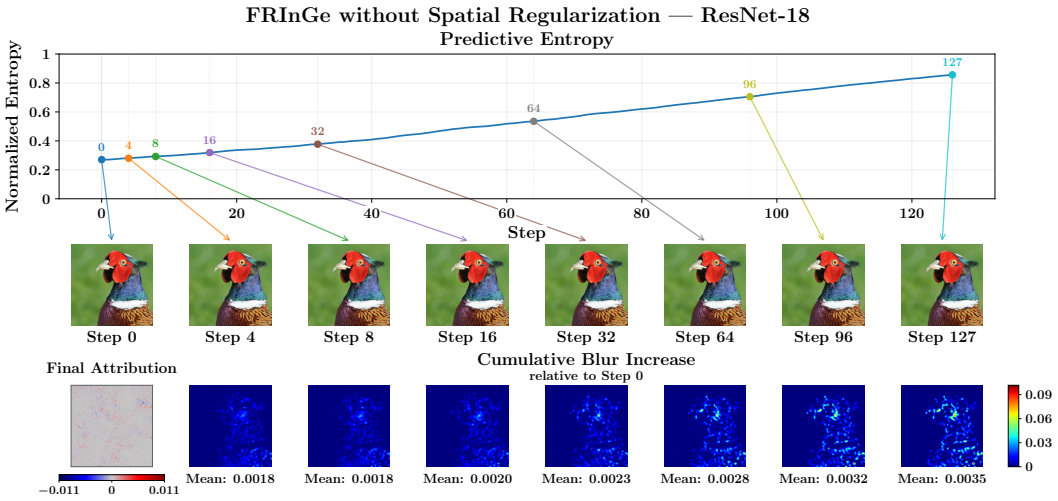

Figure 8: Damped Fisher–Rao tracking without spatial regularization on the same ResNet-18 input.

FRInGe with two-term spatial regularization. The direction satisfies

$$\big(G(x_k) + \lambda I + \gamma_{\text{step}} L^\top L\big) v_k = g_k + \gamma_{\text{prior}} L^\top L x_k, \tag{23}$$

with the empirical blur preconditioner described in Appendix A.3. In Figure 9, the blur-change maps are more spatially organized and the final attribution is less speckled. The changes concentrate on parts of the depicted object while entropy increases. This is consistent with the spatial prior shaping the numerical input path, but the appearance of one map does not establish faithfulness; that interpretation must be assessed using the quantitative metrics.

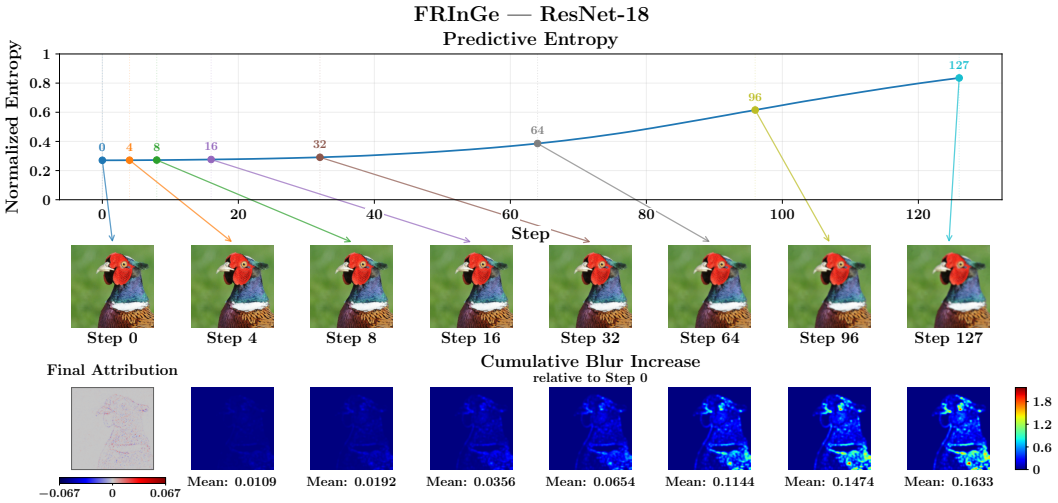

Figure 9: FRInGe on the same ResNet-18 input, combining Fisher geometry with direction regularization and a current-input prior.

B.3 Separating the Two Spatial Terms

The coefficients affect different parts of the update. With $\gamma_{\text{prior}} = 0$, the direction-only configuration solves

$$\big(G(x_k) + \lambda I + \gamma_{\text{step}} L^\top L\big) v_k = g_k. \tag{24}$$

It penalizes spatial variation in the direction rather than directly encouraging a smoother current input. Figure 10 shows a smooth entropy profile but weak, speckled blur changes and a noisy final

attribution for this input. Thus, a regularized direction field need not yield the same transformation as FRInGe with both spatial terms active.

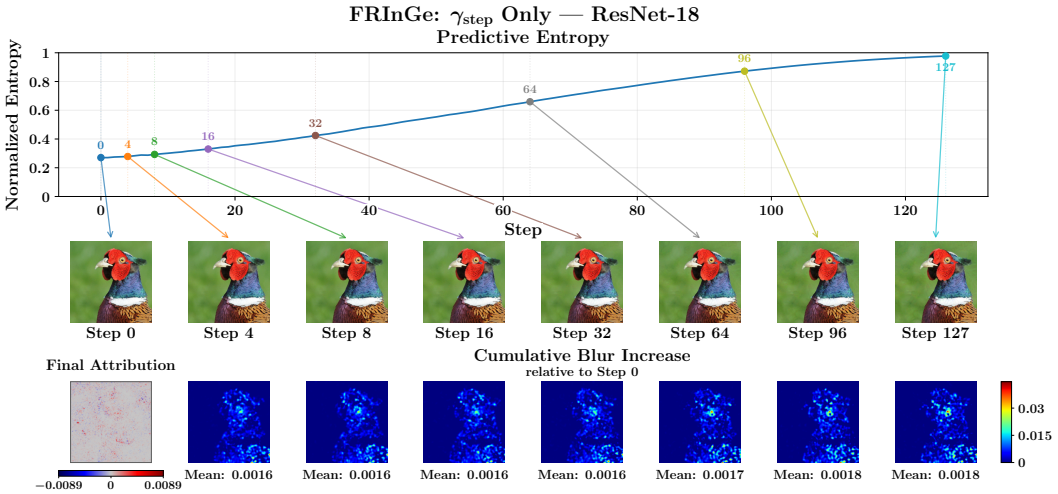

Figure 10: Direction-only FRInGe on the shared ResNet-18 input: $\gamma_{\text{step}} > 0$ and $\gamma_{\text{prior}} = 0$. The entropy profile is smooth, while spatial changes remain speckled.

With $\gamma_{\text{step}} = 0$, the Fisher-aware prior-only configuration instead solves

$$(G(x_k) + \lambda I)v_k = g_k + \gamma_{\text{prior}} L^\top L x_k. \tag{25}$$

Under the update $x_{k+1} = x_k - \eta_k v_k$, the additional right-hand-side term encourages removal of input roughness. In Figure 11, the spatial changes become concentrated, but normalized entropy increases by only approximately 0.01. This contrast illustrates why neither spatial coherence nor entropy progress alone is sufficient to assess an attribution method.

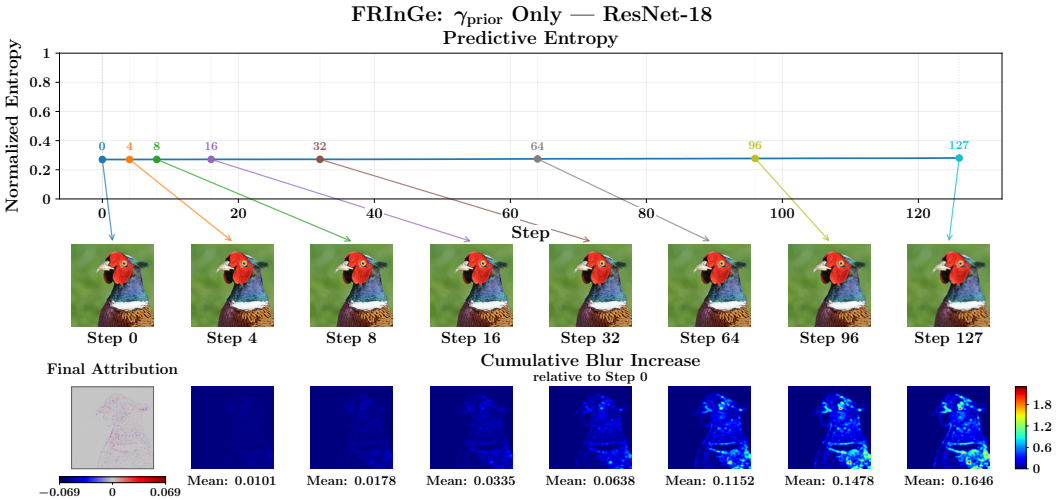

Figure 11: Prior-only FRInGe on the shared ResNet-18 input: $\gamma_{\text{step}} = 0$ and $\gamma_{\text{prior}} > 0$. Spatial changes concentrate despite limited entropy progress.

Finally, the non-geometric prior control uses

$$v_k = \nabla_x \mathcal{L}_{k+1}(x)|_{x=x_k} + \gamma_{\text{prior}} L^\top L x_k.$$

Figure 12 again shows more organized blur-change maps than Euclidean tracking without the prior, but a speckled final attribution. This control illustrates that structured intermediate transformations

need not produce a comparably structured integrated explanation. Together, the examples motivate examining the interaction of geometry and both spatial terms, rather than attributing FRInGe's behavior to smoothing alone.

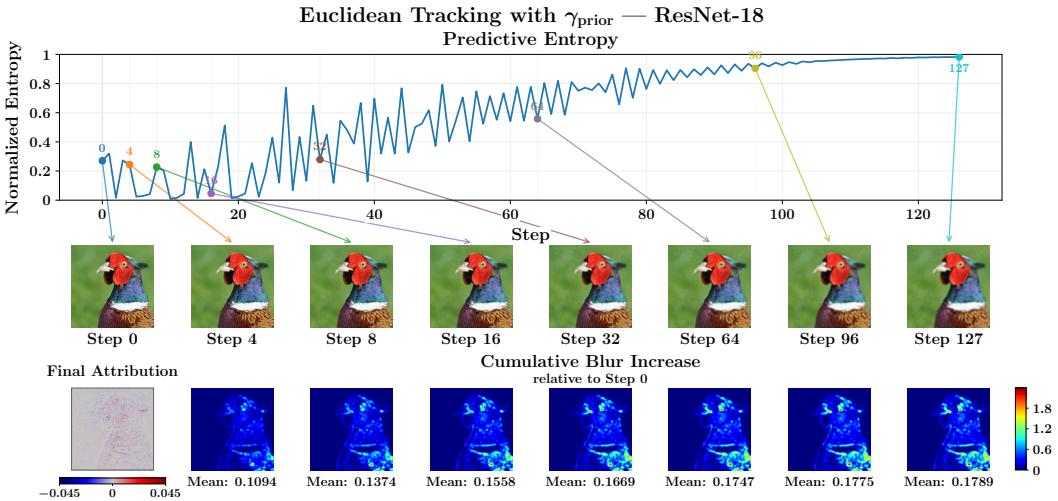

Figure 12: Euclidean waypoint tracking with a current-input prior on the shared ResNet-18 input. More organized blur-change maps coexist with a noisy final attribution.

Table 7 reports the corresponding dataset-level evidence for the core configurations and Fisher-aware prior-only variant. The single-image controls above explain how the components can affect a trajectory; they do not isolate a universal contribution of Fisher geometry under fully matched spatial regularization.

C Experimental Protocol and Reproducibility

This section specifies the choices needed to interpret and reproduce the reported comparisons. We describe the image cohorts, pretrained models, fixed explanation targets, and baseline configurations before defining the metric-specific responses, attribution reductions, and perturbation references. We then explain the hyperparameter-selection objective, list the selected architecture-specific settings, and examine their local sensitivity. The distinctions matter: FRInGe explains a target logit, but a perturbation metric may evaluate its target probability, and a magnitude map used for ranking is not the signed tensor used by every metric.

C.1 Dataset, Models, and Target Selection

The original benchmark contains 1,001 ImageNet validation images and six pretrained architectures: ResNet-18, ResNet-50, ResNet-101, ResNet-152, Inception-v3, and VGG-19, each with its official preprocessing pipeline. For each image x, the explained target is the original top-1 class, $t = \arg\max_c F^c(x)$, held fixed throughout attribution and evaluation. Methods compared on that image therefore address the same target rather than following changing predictions along a perturbation path.

C.2 Attribution Baselines

The baselines vary reference selection, path construction, and gradient averaging. IG integrates along a straight-line path, whereas EG averages eight 16-step Gauss–Legendre estimates using references sampled from the fixed pool. SG averages gradients at noise-perturbed inputs.

GuidedIG adapts feature updates along its path, AGI uses adversarial trajectories, and IG^2 constructs counterfactual-guided paths through representation matching. BlurIG uses a 100-point Gaussian scale-space path from $\sigma = 10$ to the input.

GGIG adjusts interpolants using gradient ascent, GeoIG constructs paths under a classifier-induced input geometry, and MIG uses a learned generative manifold. The tested MIG configuration uses a frozen pretrained VAE and a 12-point latent path refined for at most three sweeps. Its endpoints are VAE reconstructions, and it averages target-probability gradients rather than computing a segment-wise integral of the target logit.

All results refer to the tested configurations, evaluated using the metric conventions below.

C.3 Evaluation Pipeline and Statistical Reporting

For each image and method, the attribution is a tensor $A \in \mathbb{R}^{3\times H\times W}$. The following overview distinguishes the quantities derived from it. Perturbation scores assess the response to a spatial ranking, MAS additionally assesses magnitude alignment, infidelity compares predicted and observed local score changes, and sparseness describes concentration.

Metric	Model response	Attribution representation	Perturbation/reference
Ins/Del AUC	Target probability	Maximum absolute channel value	Average-blurred image
MAS	Normalized target probability	Same ranking and cumulative mass	Gaussian-blurred image
Infidelity	Target-logit change	Original signed tensor	Gaussian input noise
Sparseness	Not applicable	All absolute tensor entries	No perturbation
Max Sensitivity	Not applicable	ℓ_2-normalized tensor	Uniform input noise

Mean-score tables report the mean plus or minus half the 95% bootstrap interval width; percentile intervals need not be symmetric about the mean. Max Sensitivity instead uses a median and half-IQR, a dispersion summary rather than a confidence interval. Higher MAS-Ins and Ins-AUC are preferred; lower MAS-Del, Del-AUC, Infidelity, and Max Sensitivity are preferred. Higher sparseness means greater concentration, not necessarily greater faithfulness.

C.4 Metric Definitions

C.4.1 Spatial Saliency Reduction

Metrics based on progressive perturbations require a spatial ranking rather than a channel-wise attribution tensor. Given $A \in \mathbb{R}^{3\times H\times W}$, we define the spatial saliency map

$$S_{ij} = \max_{c\in\{1,2,3\}} |A_{cij}|,$$

where S_{ij} is the saliency assigned to pixel (i, j). Pixels are ranked in descending order of S_{ij}, and this ranking is used in all perturbation-based evaluations.

C.4.2 Causal Faithfulness: Blur-Based Insertion and Deletion AUC

We follow the insertion/deletion protocol of Petsiuk et al. [16].

The reference x_{blur} is a local average of the input with an odd window size proportional to image resolution:

$$k = \max\left(3,\ \mathrm{odd}\left(\left\lfloor \frac{\min(H, W)}{10} \right\rfloor\right)\right).$$

Here odd rounds upward to the next odd integer, including its argument when that argument is already odd.

At perturbation step s, let x_s be the perturbed image and

$$c_s = \mathrm{softmax}(F(x_s))_t$$

the target-class confidence. Insertion starts from the blurred reference and restores pixels in saliency order; deletion starts from the original and replaces them by their reference values in the same order. Their AUCs measure how quickly confidence is recovered or removed as the fraction of replaced pixels increases.

The normalized confidence curve is

$$\widetilde{c}_s = \mathrm{clip}\left(\frac{c_s - p_{\mathrm{blur}}}{\max\big(p_{\mathrm{orig}} - p_{\mathrm{blur}}, \varepsilon\big)}, 0, 1\right),$$

where

$$p_{\mathrm{orig}} = \mathrm{softmax}(F(x))_t, \qquad p_{\mathrm{blur}} = \mathrm{softmax}(F(x_{\mathrm{blur}}))_t.$$

We integrate this curve using the trapezoidal rule over the replaced-pixel fraction. The anchors are the target confidences on the original and blurred inputs. This rescales the curve by their confidence gap and clips it to $[0, 1]$; $\varepsilon > 0$ prevents a zero denominator. The blur here is distinct from the Gaussian reference used by MAS below.

C.4.3 Magnitude Aligned Scoring (MAS)

MAS [32] asks whether attribution magnitudes, not only their ordering, agree with the classifier's confidence response. It uses the same spatial saliency S_{ij} but a Gaussian-blurred reference with a 15×15 kernel and standard deviation $\sigma = 3.0$. The target is again the original top-1 class. An image is eligible only when its original target confidence exceeds the Gaussian-reference confidence by more than 10^{-6}. This eligibility condition depends on the model, image, and baseline, not on the attribution method.

Let $\mathcal{I}_s$ be the pixels restored or removed after step s, with $s = 0, \ldots, N$. The density responses represent the fraction of attribution mass restored by insertion or remaining after deletion:

$$DR_s^{\mathrm{ins}} = \frac{\sum_{(i,j)\in\mathcal{I}_s} S_{ij}}{\max\Big(\sum_{i,j} S_{ij}, \varepsilon\Big)}, \qquad DR_s^{\mathrm{del}} = 1 - DR_s^{\mathrm{ins}}.$$

The model response first uses the confidence-gap normalization defined above, with the Gaussian MAS reference in place of the average-blurred reference. Its insertion curve is replaced by its cumulative maximum, and its deletion curve by its cumulative minimum. Denote these monotone normalized responses by MR_s^{ins} and MR_s^{del}. Their discrepancy from the density response is

$$AP_s^m = \left|MR_s^m - DR_s^m\right|, \qquad m \in \{\mathrm{ins}, \mathrm{del}\}.$$

The pointwise scores are

$$z_s^{\mathrm{ins}} = \mathrm{clip}\big(MR_s^{\mathrm{ins}} - AP_s^{\mathrm{ins}}, 0, 1\big), \qquad z_s^{\mathrm{del}} = \mathrm{clip}\big(MR_s^{\mathrm{del}} + AP_s^{\mathrm{del}}, 0, 1\big).$$

For either mode, the reported MAS is the trapezoidal area under this score curve over the perturbation steps:

$$MAS^m = \frac{1}{N}\left(\frac{z_0^m + z_N^m}{2} + \sum_{s=1}^{N-1} z_s^m\right).$$

Higher MAS-Insertion rewards rapid confidence recovery with aligned attribution mass; lower MAS-Deletion rewards rapid removal with aligned remaining mass. These are attribution magnitude-alignment scores, not measures of classification-probability calibration. Summaries exclude ineligible images, so their sample sizes can differ from those of the other metrics even within the same cohort.

C.4.4 Local Robustness: Infidelity

Infidelity [35] measures how well an attribution predicts the change in target score under small random perturbations. For attribution $\Phi(x)$ and perturbation δ, it is defined as

$$\mathbb{E}_\delta\Big[\big(\delta^\top \Phi(x) - \big(F^t(x) - F^t(x-\delta)\big)\big)^2\Big].$$

The first term, $\delta^\top \Phi(x)$, is the attribution-predicted score variation under perturbation δ, while the second term is the actual change in model output. Lower infidelity indicates better local faithfulness. We estimate this expectation by Monte Carlo sampling with 50 Gaussian perturbations per image, each drawn with standard deviation $\sigma = 0.02$.

C.4.5 Explanation Focus: Sparseness

We quantify explanation focus using the Gini index of all absolute attribution entries [8]. For $A \in \mathbb{R}^{3\times H\times W}$, write

$$a = \big(|A_1|, \dots, |A_n|\big) \in \mathbb{R}^n_{\geq 0}, \qquad n = 3 \cdot H \cdot W.$$

Let $a_{(1)} \leq a_{(2)} \leq \cdots \leq a_{(n)}$ denote these values sorted in ascending order. The sparseness score is then computed as

$$\mathrm{Sparseness}(A) = \frac{2\sum_{i=1}^{n} i\, a_{(i)}}{n\sum_{i=1}^{n} a_{(i)} + \varepsilon} - \frac{n+1}{n},$$

with the reported score clipped to $[0, 1]$.

This score is the Gini index of the attribution magnitude distribution. Values near 0 indicate a diffuse or nearly uniform attribution map, whereas values near 1 indicate that attribution mass is concentrated on a small number of entries. This metric is reported only as a complementary descriptive statistic and is not used directly in the causal-faithfulness or calibration evaluations above.

C.4.6 Local Robustness: Max Sensitivity

We additionally report Max Sensitivity [35] as a robustness diagnostic under small input perturbations. For a given input x, fixed target class t, and attribution method Φ, we compute the attribution map and normalize it by its ℓ_2 norm so that the score reflects changes in explanation structure rather than overall scale:

$$\widehat{\Phi}(x) = \frac{\Phi(x)}{\|\Phi(x)\|_2 + \varepsilon}.$$

We then sample perturbations $\delta^{(1)}, \dots, \delta^{(M)}$ independently, with entries drawn uniformly from $[-r, r]$, and recompute the normalized attribution for each perturbed input. In our experiments, $r = 0.02$ and $M = 10$. The reported score is

$$\mathrm{MaxSens}(x) = \max_{m=1,\dots,M} \frac{\left\|\widehat{\Phi}(x) - \widehat{\Phi}(x + \delta^{(m)})\right\|_2}{\|\delta^{(m)}\|_2 + \varepsilon}.$$

Lower values indicate greater local robustness. Since the maximum is taken over a finite Monte Carlo sample, this is a sampled approximation of local sensitivity rather than an exact worst-case bound.

C.5 Hyperparameter Selection Protocol

Architecture-specific settings for the three core configurations were selected using Optuna [1]. The search includes the trust-region parameters τ, $\eta_{\max}$, and δ_{euc}, damping λ, and, where applicable, the two spatial coefficients. The added prior-only ablation is not a separately optimized method: it inherits FRInGe's tuned hyperparameters and sets $\gamma_{\mathrm{step}} = 0$.

The optimization objective was a weighted harmonic mean of three quantities: blur-based insertion AUC, inverted blur-based deletion AUC, and an infidelity-derived fidelity score. Let InsAUC and DelAUC denote the insertion and deletion AUC values on the tuning subset, and let Inf denote the infidelity score. We define $\mathrm{Fid} = \exp(-k\,\mathrm{Inf}), k = 2$ and optimize

$$\mathrm{Score} = \frac{w_{\mathrm{ins}} + w_{\mathrm{del}} + w_{\mathrm{fid}}}{\frac{w_{\mathrm{ins}}}{\mathrm{InsAUC}+\varepsilon} + \frac{w_{\mathrm{del}}}{(1-\mathrm{DelAUC})+\varepsilon} + \frac{w_{\mathrm{fid}}}{\mathrm{Fid}+\varepsilon}}, \tag{26}$$

with weights $w_{\mathrm{ins}} = 1, w_{\mathrm{del}} = 1, w_{\mathrm{fid}} = 0.5$, and $\varepsilon = 10^{-6}$. The harmonic mean discourages a high score on one component from compensating for poor performance on another. MAS and sparseness are not components of this objective, so the selected configuration is not directly optimized for every reported metric. Because hyperparameters were tuned on a subset of the validation cohort rather than a disjoint split, the reported results should be interpreted as benchmark comparisons rather than a held-out evaluation of hyperparameter generalization.

C.6 Architecture-Specific Hyperparameters

Table 6 lists the selected hyperparameters for the three core configurations. Euclidean tracking bypasses the linear system, so damping, Fisher-solve, and spatial-penalty parameters are inapplicable. Fisher–Rao tracking without spatial regularization (*Unreg. FR*) retains the damped Fisher system but disables both spatial terms; FRInGe uses both terms in the system in Appendix A.3. The prior-only variant inherits the FRInGe row and sets only $\gamma_{\mathrm{step}} = 0$, so those otherwise duplicated rows are omitted. In the main benchmark runs for FRInGe, each conjugate-gradient solve allows at most 25 iterations per tracking step. Metric settings are held fixed within each comparison and are distinct from these optimization parameters.

Table 6: Architecture-specific ablation hyperparameters. Dashes indicate inapplicable parameters.

Model	**Variant**	τ	$\eta_{\max}$	δ_{euc}	λ	γ_{step}	γ_{prior}
ResNet-18	Euclidean	–	29.58	–	–	–	–
	Unreg. FR	3.03e-4	1.98	34.52	2.77e-11	–	–
	FRInGe	3.03e-4	13.56	0.61	4.77e-11	9.97e-3	9.75e-4
ResNet-50	Euclidean	–	29.70	–	–	–	–
	Unreg. FR	3.21e-4	1.13	18.32	1.28e-10	–	–
	FRInGe	2.70e-3	1.18	5.01	2.58e-4	9.94e-3	9.49e-4
ResNet-101	Euclidean	–	29.13	–	–	–	–
	Unreg. FR	6.43e-4	1.02	14.48	1.12e-12	–	–
	FRInGe	3.69e-4	1.53	3.85	8.02e-8	9.98e-3	6.08e-4
ResNet-152	Euclidean	–	0.19	–	–	–	–
	Unreg. FR	3.29e-4	1.06	26.35	3.15e-10	–	–
	FRInGe	4.13e-4	1.75	3.32	7.55e-5	6.06e-3	9.16e-4
Inception-v3	Euclidean	–	6.28	–	–	–	–
	Unreg. FR	1.97e-3	2.81	3.28	9.79e-3	–	–
	FRInGe	5.34e-4	16.92	24.18	7.17e-6	9.73e-3	3.43e-4
VGG-19	Euclidean	–	0.02	–	–	–	–
	Unreg. FR	3.22e-4	1.11	9.39	2.17e-12	–	–
	FRInGe	3.04e-4	2.75	0.70	3.03e-11	5.13e-3	9.62e-4

C.7 Hyperparameter Sensitivity Across Architectures

The selected hyperparameters do not show how sensitive the method is to nearby choices. Figure 13 therefore varies one parameter at a time while holding the others fixed and reports the same weighted harmonic-mean objective used for selection. The spatial sweep is an explicit exception: it ties $\gamma_{\mathrm{step}} = \gamma_{\mathrm{prior}} = \gamma$ and varies both coefficients jointly. It does not measure their separate effects.

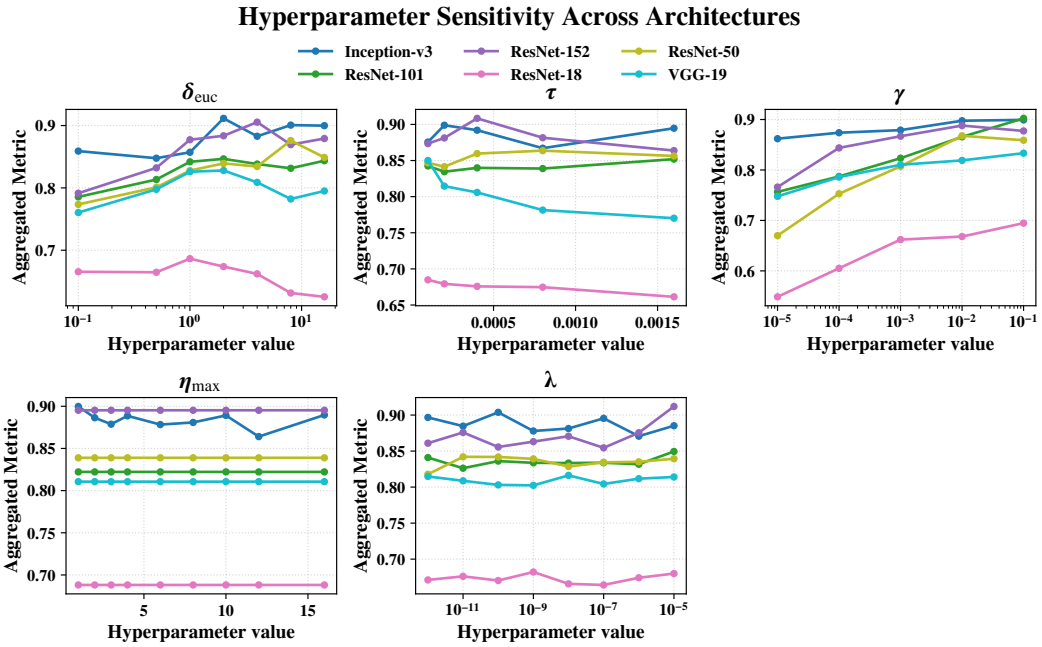

Figure 13: Sensitivity analysis of FRInGe hyperparameters across all evaluated architectures. Each panel varies one hyperparameter while keeping the others fixed and reports the aggregated tuning objective. For the smoothing sweep, we vary the regularization strength jointly, setting $\gamma_{\text{step}} = \gamma_{\text{prior}} = \gamma$.

Over the displayed ranges, the objective is relatively insensitive to $\eta_{\max}$ and more architecture-dependent for τ and λ. A permissive maximum step size can be inactive when another cap is tighter, whereas τ also changes the nominal waypoint count. Damping affects the computed direction and the numerical solve; its role should not be inferred from a single architecture's curve.

The more pronounced changes occur for the Euclidean radius and tied spatial strength. These parameters directly alter input-segment lengths and the spatial character of the path. Small radii can restrict progress within the available updates, while large radii can leave gradient variation unresolved by quadrature. The smoothing sweep illustrates the benefit and eventual saturation of the tested low-frequency prior, without establishing that larger regularization is universally better.

These are conditional, one-factor sweeps, not a complete analysis of parameter interactions or a new selection rule. In particular, the tied spatial sweep cannot replace the separate-component controls in Appendix B.

D Additional Quantitative Results

This section complements the main results in three stages. The component ablation compares the roles of geometry and spatial regularization; the architecture-specific tables extend the attribution-method comparison; and a separate perturbation study assesses local explanation stability. These answer different questions, so concentration, magnitude alignment, perturbation faithfulness, and robustness are interpreted separately.

Unless a table states otherwise, entries are means plus or minus half the 95% image-bootstrap interval width. The best displayed mean within each comparison is bold and underlined; bold overlap highlighting compares the displayed ranges and is descriptive, not a test of equivalence or a paired significance test. A symmetric displayed range need not reproduce the endpoints of an asymmetric percentile interval. Max Sensitivity uses median plus or minus half-IQR instead. The metric definitions and normalization conventions are given in Appendix C.4.

D.1 Ablation Study

Table 7 evaluates the three core configurations and the added Fisher-aware prior-only variant across all six architectures. The latter retains FRInGe's damping, step controls, prior coefficient, and preconditioner, changing only $\gamma_{\text{step}} = 0$; it is not separately retuned. The core Euclidean tracking, Fisher tracking without spatial regularization, and FRInGe use their respective selected hyperparameters. Their comparison therefore evaluates complete configurations, not geometry alone under matched regularization.

FRInGe has the highest displayed MAS-Insertion and lowest MAS-Deletion mean on each architecture. The prior-only variant improves Ins-AUC and Del-AUC over the unregularized configurations and has the lower Del-AUC mean on ResNet-50 and ResNet-152, while FRInGe has the higher insertion-AUC mean throughout. Fisher tracking without spatial regularization is more sparse without having better MAS. This contrast shows why concentration cannot substitute for attribution-quality evaluation. The illustrative trajectories and component definitions appear in Appendix B.

Table 7: Ablation study of FRInGe components across evaluated architectures. **Euclid.** ignores manifold curvature, **Unreg. FR** retains damping and pullback geometry but disables both spatial terms, and **Prior-only** fixes $\gamma_{\text{step}} = 0$ using FRInGe's tuned hyperparameters without separate retuning. **FRInGe** uses both spatial terms. Best results are in **bold underline**; methods with overlapping confidence intervals are in **bold**.

	ResNet-18				ResNet-50			
Metric	**Euclid.**	**Unreg. FR**	**Prior-only**	**FRInGe**	**Euclid.**	**Unreg. FR**	**Prior-only**	**FRInGe**
MAS-Ins ↑	0.414±.017	0.399±.017	0.461±.016	**0.597±.015**	0.491±.018	0.477±.018	0.581±.014	**0.648±.014**
MAS-Del ↓	0.319±.012	0.330±.012	0.324±.010	**0.264±.010**	0.361±.014	0.372±.014	0.326±.010	**0.302±.012**
Ins-AUC ↑	0.358±.014	0.347±.013	0.409±.014	**0.492±.016**	0.415±.014	0.403±.014	**0.560±.015**	**0.581±.016**
Del-AUC ↓	0.195±.010	0.207±.010	0.165±.009	**0.130±.008**	0.246±.012	0.261±.012	**0.152±.009**	**0.161±.009**
Sparseness ↑	0.764±.003	**0.785±.003**	0.661±.002	0.706±.002	0.763±.003	**0.791±.004**	0.605±.003	0.702±.002
Infidelity ↓	**0.058±.008**	**0.058±.008**	**0.058±.008**	**0.058±.008**	**0.149±.050**	**0.149±.050**	**0.152±.050**	**0.150±.050**
	ResNet-101				**ResNet-152**			
MAS-Ins ↑	0.568±.017	0.554±.018	0.598±.014	**0.679±.014**	0.591±.017	0.581±.018	0.624±.013	**0.691±.013**
MAS-Del ↓	0.421±.017	0.432±.017	0.389±.014	**0.351±.015**	0.451±.018	0.464±.018	**0.403±.015**	**0.392±.017**
Ins-AUC ↑	0.499±.015	0.488±.015	0.595±.015	**0.641±.017**	0.531±.015	0.519±.015	**0.666±.015**	**0.681±.016**
Del-AUC ↓	0.309±.014	0.321±.014	**0.228±.012**	**0.228±.012**	0.351±.014	0.362±.014	**0.240±.012**	**0.265±.013**
Sparseness ↑	0.772±.003	**0.795±.004**	0.634±.003	0.706±.002	0.779±.003	**0.799±.004**	0.589±.003	0.703±.002
Infidelity ↓	**0.121±.022**	**0.121±.022**	**0.125±.022**	**0.121±.022**	**0.111±.020**	**0.111±.020**	**0.113±.020**	**0.111±.020**
	Inception-v3				**VGG-19**			
MAS-Ins ↑	0.688±.015	0.681±.015	0.676±.010	**0.761±.009**	0.427±.017	0.413±.017	0.478±.016	**0.588±.015**
MAS-Del ↓	**0.467±.019**	**0.483±.019**	**0.464±.017**	**0.444±.020**	0.319±.012	0.329±.012	0.309±.010	**0.255±.008**
Ins-AUC ↑	0.580±.013	0.557±.013	0.672±.013	**0.709±.014**	0.389±.015	0.382±.015	0.444±.015	**0.503±.016**
Del-AUC ↓	0.300±.013	0.315±.013	**0.241±.012**	**0.240±.012**	0.193±.010	0.204±.010	0.157±.008	**0.119±.007**
Sparseness ↑	0.763±.003	**0.801±.003**	0.617±.003	0.723±.003	**0.807±.003**	**0.814±.006**	0.691±.003	0.707±.003
Infidelity ↓	**0.020±.005**	**0.020±.005**	**0.026±.005**	**0.020±.005**	**0.343±.064**	**0.343±.064**	**0.344±.064**	**0.343±.064**

D.2 Comparison across attribution methods

Tables 8–13 compare FRInGe with all ten attribution baselines across six architectures. Method settings, cohort differences, and metric conventions are specified in Appendix C. AGI denotes Adversarial IG and SG denotes SmoothGrad.

The comparison reveals distinct strengths rather than a single uniformly best method. FRInGe has the highest mean MAS-Insertion on most architectures, while MAS-Deletion favors different methods depending on the architecture. Among the methods in these tables, BlurIG leads mean Ins-AUC and Del-AUC, whereas IG^2 produces the most concentrated maps. Infidelity intervals overlap for many methods, so their displayed rankings should not be interpreted as clear statistical separation. Together, these results distinguish attribution magnitude alignment from feature ordering and concentration; an advantage in one does not establish superiority under the others.

FRInGe's spatial regularization and BlurIG's Gaussian scale-space path share a low-frequency inductive bias, although their trajectories differ. This may help explain their competitive results on these image models. FRInGe additionally conditions the trajectory on the classifier's pullback Fisher geometry and predictive waypoints, consistent with its strong MAS-Insertion results. BlurIG's blur-based AUC advantage may also partly reflect alignment between its path and the evaluation perturbations. These are plausible interpretations, not established causal explanations; testing the latter would require other perturbation references.

Table 8: Quantitative evaluation on Inception-v3. Means ± half-widths of 95% image-bootstrap intervals; descriptive highlighting.

	Attribution calibration		*Causal faithfulness*		*Focus and robustness*	
Method	**MAS-Ins ↑**	**MAS-Del ↓**	**Ins-AUC ↑**	**Del-AUC ↓**	**Sparseness ↑**	**Infidelity ↓**
FRInGe	**0.761±0.009**	0.444±0.020	0.709±0.014	0.240±0.012	0.723±0.003	**0.020±0.005**
AGI	0.605±0.011	0.520±0.017	0.573±0.014	0.321±0.013	0.514±0.006	**0.021±0.005**
GGIG	0.631±0.011	0.473±0.015	0.658±0.015	0.215±0.011	0.471±0.004	0.411±0.018
GeoIG	0.647±0.009	0.469±0.016	0.668±0.014	0.216±0.011	0.529±0.002	4,760±209
GuidedIG	0.701±0.010	0.452±0.019	0.675±0.013	0.242±0.013	0.665±0.003	**0.027±0.005**
IG	0.673±0.010	0.485±0.019	0.647±0.014	0.271±0.013	0.624±0.003	0.032±0.006
IG^2	0.691±0.013	0.455±0.019	0.596±0.013	0.289±0.013	**0.739±0.003**	**0.020±0.005**
SG	0.660±0.009	**0.373±0.008**	0.690±0.013	0.206±0.011	0.540±0.003	**0.021±0.005**
BlurIG	0.735±0.009	**0.386±0.017**	**0.758±0.013**	**0.174±0.011**	0.655±0.004	**0.022±0.006**
EG	0.648±0.011	0.481±0.018	0.634±0.015	0.267±0.014	0.557±0.002	**0.022±0.006**
MIG	0.671±0.012	0.467±0.018	0.644±0.016	0.281±0.014	0.622±0.004	**0.021±0.006**

Table 9: Quantitative evaluation on ResNet-18. Means ± half-widths of 95% image-bootstrap intervals; descriptive highlighting.

	Attribution calibration		*Causal faithfulness*		*Focus and robustness*	
Method	**MAS-Ins ↑**	**MAS-Del ↓**	**Ins-AUC ↑**	**Del-AUC ↓**	**Sparseness ↑**	**Infidelity ↓**
FRInGe	**0.597±0.015**	**0.264±0.010**	0.492±0.016	0.130±0.008	0.706±0.002	**0.058±0.008**
AGI	0.387±0.014	0.395±0.010	0.347±0.013	0.212±0.010	0.504±0.007	**0.057±0.008**
GGIG	0.426±0.016	0.373±0.008	0.404±0.016	0.164±0.009	0.450±0.002	0.432±0.017
GeoIG	0.453±0.014	0.359±0.007	0.430±0.014	0.135±0.008	0.497±0.002	3,810±132
GuidedIG	0.477±0.016	0.320±0.010	0.448±0.014	0.147±0.008	0.649±0.002	0.077±0.008
IG	0.443±0.015	0.359±0.011	0.415±0.014	0.178±0.010	0.608±0.003	0.104±0.009
IG^2	0.421±0.016	0.314±0.011	0.369±0.013	0.179±0.010	**0.714±0.002**	**0.057±0.008**
SG	0.488±0.014	0.325±0.004	0.477±0.015	0.122±0.007	0.513±0.002	**0.057±0.008**
BlurIG	**0.577±0.014**	**0.276±0.008**	**0.563±0.015**	**0.096±0.007**	0.634±0.004	**0.061±0.009**
EG	0.423±0.016	0.368±0.009	0.398±0.015	0.165±0.010	0.530±0.002	**0.063±0.009**
MIG	0.440±0.016	0.358±0.011	0.405±0.016	0.197±0.011	0.590±0.004	**0.058±0.009**

Table 10: Quantitative evaluation on ResNet-50. Means ± half-widths of 95% image-bootstrap intervals; descriptive highlighting.

	Attribution calibration		*Causal faithfulness*		*Focus and robustness*	
Method	**MAS-Ins ↑**	**MAS-Del ↓**	**Ins-AUC ↑**	**Del-AUC ↓**	**Sparseness ↑**	**Infidelity ↓**
FRInGe	**0.648±0.014**	**0.302±0.012**	0.581±0.016	0.161±0.009	0.702±0.002	**0.150±0.050**
AGI	0.463±0.015	0.424±0.011	0.423±0.014	0.257±0.012	0.490±0.006	**0.147±0.048**
GGIG	0.491±0.015	0.410±0.011	0.477±0.016	0.213±0.010	0.463±0.003	2.004±0.101
GeoIG	0.523±0.014	0.379±0.009	0.527±0.016	0.159±0.009	0.507±0.002	5,430±198
GuidedIG	0.551±0.016	0.347±0.012	0.527±0.015	0.183±0.010	0.656±0.003	**0.175±0.047**
IG	0.516±0.015	0.388±0.013	0.481±0.014	0.226±0.012	0.621±0.003	**0.232±0.054**
IG^2	0.492±0.017	0.355±0.013	0.424±0.014	0.233±0.012	**0.716±0.002**	**0.148±0.050**
SG	0.562±0.013	0.323±0.005	0.582±0.016	**0.144±0.008**	0.539±0.003	**0.147±0.048**
BlurIG	0.619±0.014	**0.303±0.010**	**0.621±0.015**	**0.128±0.008**	0.639±0.004	**0.163±0.055**
EG	0.507±0.015	0.390±0.012	0.482±0.016	0.205±0.012	0.542±0.002	**0.165±0.055**
MIG	0.527±0.017	0.380±0.013	0.491±0.016	0.238±0.012	0.609±0.004	**0.159±0.056**

Table 11: Quantitative evaluation on ResNet-101. Means ± half-widths of 95% image-bootstrap intervals; descriptive highlighting.

	Attribution calibration		*Causal faithfulness*		*Focus and robustness*	
Method	**MAS-Ins ↑**	**MAS-Del ↓**	**Ins-AUC ↑**	**Del-AUC ↓**	**Sparseness ↑**	**Infidelity ↓**
FRInGe	**0.679±0.014**	**0.351±0.015**	0.641±0.017	0.228±0.012	0.706±0.002	**0.121±0.022**
AGI	0.520±0.014	0.475±0.014	0.511±0.015	0.315±0.014	0.492±0.006	**0.119±0.021**
GGIG	0.555±0.014	0.437±0.013	0.572±0.017	0.241±0.011	0.482±0.003	1.681±0.084
GeoIG	0.567±0.013	0.414±0.012	0.607±0.016	0.199±0.010	0.506±0.002	6,030±207
GuidedIG	0.605±0.014	0.397±0.015	0.602±0.016	0.243±0.013	0.660±0.003	**0.149±0.021**
IG	0.577±0.014	0.433±0.015	0.570±0.015	0.278±0.013	0.620±0.003	0.189±0.022
IG^2	0.569±0.016	0.416±0.016	0.505±0.015	0.300±0.014	**0.722±0.002**	**0.120±0.022**
SG	0.605±0.012	**0.341±0.006**	0.646±0.016	0.187±0.010	0.537±0.003	**0.119±0.021**
BlurIG	**0.662±0.013**	**0.338±0.014**	**0.694±0.016**	**0.166±0.010**	0.635±0.004	**0.126±0.024**
EG	0.562±0.015	0.439±0.015	0.572±0.018	0.261±0.014	0.545±0.002	**0.129±0.024**
MIG	0.591±0.016	0.425±0.015	0.578±0.018	0.288±0.014	0.613±0.004	**0.124±0.024**

Table 12: Quantitative evaluation on ResNet-152. Means ± half-widths of 95% image-bootstrap intervals; descriptive highlighting.

	Attribution calibration		*Causal faithfulness*		*Focus and robustness*	
Method	**MAS-Ins ↑**	**MAS-Del ↓**	**Ins-AUC ↑**	**Del-AUC ↓**	**Sparseness ↑**	**Infidelity ↓**
FRInGe	**0.691±0.013**	0.392±0.017	**0.681±0.016**	0.265±0.013	0.703±0.002	**0.111±0.020**
AGI	0.537±0.014	0.500±0.015	0.538±0.016	0.358±0.015	0.496±0.006	**0.111±0.019**
GGIG	0.563±0.015	0.469±0.015	0.592±0.017	0.301±0.013	0.490±0.003	2.539±0.134
GeoIG	0.588±0.012	0.432±0.014	0.651±0.016	**0.229±0.012**	0.510±0.002	6,290±225
GuidedIG	0.619±0.015	0.419±0.016	0.631±0.015	0.275±0.013	0.661±0.003	**0.142±0.020**
IG	0.588±0.015	0.463±0.016	0.587±0.015	0.324±0.015	0.624±0.003	0.192±0.021
IG^2	0.590±0.016	0.442±0.018	0.539±0.015	0.343±0.014	**0.726±0.002**	**0.110±0.019**
SG	0.620±0.012	**0.349±0.007**	**0.682±0.015**	**0.215±0.011**	0.546±0.003	**0.111±0.019**
BlurIG	**0.666±0.012**	0.379±0.016	**0.704±0.015**	**0.206±0.012**	0.639±0.004	**0.120±0.023**
EG	0.575±0.014	0.458±0.016	0.596±0.018	0.296±0.015	0.546±0.002	**0.121±0.023**
MIG	0.608±0.015	0.456±0.016	0.609±0.018	0.333±0.015	0.614±0.004	**0.116±0.023**

Table 13: Quantitative evaluation on VGG-19. Means ± half-widths of 95% image-bootstrap intervals; descriptive highlighting.

	Attribution calibration		*Causal faithfulness*		*Focus and robustness*	
Method	**MAS-Ins ↑**	**MAS-Del ↓**	**Ins-AUC ↑**	**Del-AUC ↓**	**Sparseness ↑**	**Infidelity ↓**
FRInGe	**0.588±0.015**	**<u>0.255±0.008</u>**	0.503±0.016	0.119±0.007	0.707±0.003	**0.343±0.064**
AGI	0.440±0.014	0.405±0.009	0.448±0.016	0.192±0.010	0.433±0.006	**0.346±0.064**
GGIG	0.460±0.015	0.357±0.008	0.441±0.016	0.159±0.008	0.497±0.003	0.862±0.065
GeoIG	0.487±0.014	0.335±0.006	0.490±0.015	0.118±0.007	0.529±0.002	4,270±186
GuidedIG	0.496±0.015	0.305±0.009	0.468±0.015	0.148±0.008	0.674±0.003	**0.387±0.063**
IG	0.479±0.015	0.337±0.010	0.450±0.014	0.170±0.009	0.637±0.003	**0.419±0.066**
IG^2	0.448±0.017	0.290±0.010	0.419±0.015	0.159±0.009	**<u>0.765±0.003</u>**	**<u>0.342±0.064</u>**
SG	0.523±0.014	0.304±0.004	0.522±0.016	0.113±0.007	0.547±0.003	**<u>0.342±0.064</u>**
BlurIG	**<u>0.596±0.014</u>**	**0.266±0.008**	**<u>0.567±0.015</u>**	**<u>0.098±0.007</u>**	0.649±0.004	**0.361±0.073**
EG	0.439±0.016	0.357±0.009	0.423±0.016	0.165±0.010	0.560±0.003	**0.367±0.072**
MIG	0.475±0.016	0.345±0.010	0.433±0.016	0.190±0.010	0.620±0.004	**0.356±0.072**

D.3 Robustness under small input perturbations

We additionally report Max Sensitivity as a direct robustness diagnostic under small input perturbations, evaluated on a sample of 64 images for each architecture. Because this metric is defined through a maximum over sampled perturbations and can therefore be skewed by a small number of difficult examples, we summarize it here using the median together with the interquartile half-width, i.e., $(Q_3 - Q_1)/2$. This yields a robust summary of central tendency and dispersion that is less sensitive to outliers than the mean.

Table 14 shows lower median sensitivity for FRInGe on ResNet-18 and VGG-19, and for IG on the other four architectures. FRInGe's half-IQR is smaller on five architectures, but a narrower spread does not offset a higher median or establish greater robustness. The results therefore do not support a general local-stability advantage for FRInGe. This diagnostic is complementary to MAS and AUC: an attribution can align well with a model's perturbation response without being less sensitive to small input changes.

Table 14: Max Sensitivity (↓) across architectures for FRInGe and standard IG, evaluated on 64 images per model. Values are reported as $\mathrm{median} \pm \mathrm{half\text{-}IQR}$, where $\mathrm{half\text{-}IQR} = (Q_3 - Q_1)/2$. The best median in each column is shown in **<u>bold underline</u>**; methods whose $\mathrm{median} \pm \mathrm{half\text{-}IQR}$ intervals overlap with the best are shown in **bold** as a descriptive near-tie heuristic.

Method	**ResNet-18**	**ResNet-50**	**ResNet-101**	**ResNet-152**	**Inception-v3**	**VGG-19**
FRInGe (Ours)	**<u>0.0758±.0122</u>**	**0.1139±.0141**	0.1225±.0138	0.1383±.0170	0.1081±.0131	**<u>0.0665±.0094</u>**
Standard IG	**0.0833±.0171**	**<u>0.0917±.0313</u>**	**<u>0.0829±.0209</u>**	**<u>0.0900±.0313</u>**	**<u>0.0381±.0093</u>**	**0.0678±.0126**

E Additional Qualitative Results

For each architecture, images are ranked by FRInGe's average signed improvement over the seven displayed baselines, with weights $(1, 1, 0.1, 0.1)$ for MAS-Insertion, MAS-Deletion, insertion AUC, and deletion AUC. Differences are oriented so that positive values favor FRInGe. We show two highest-ranked images, two near the median, and two lowest-ranked images; these are stratified examples, not a random sample.

Maps show absolute attributions summed over channels, clipped at the 99th percentile, and independently normalized to $[0, 1]$. MI/MD denote MAS-Insertion/Deletion and I/D normalized insertion/deletion AUC; *Ours* denotes FRInGe and GIG denotes GeoIG.



Figure 14: ResNet-18 magnitude maps for top-, median-, and bottom-ranked examples under the selection rule in Appendix E.



Figure 15: ResNet-50 magnitude maps under the same stratified selection and display conventions.

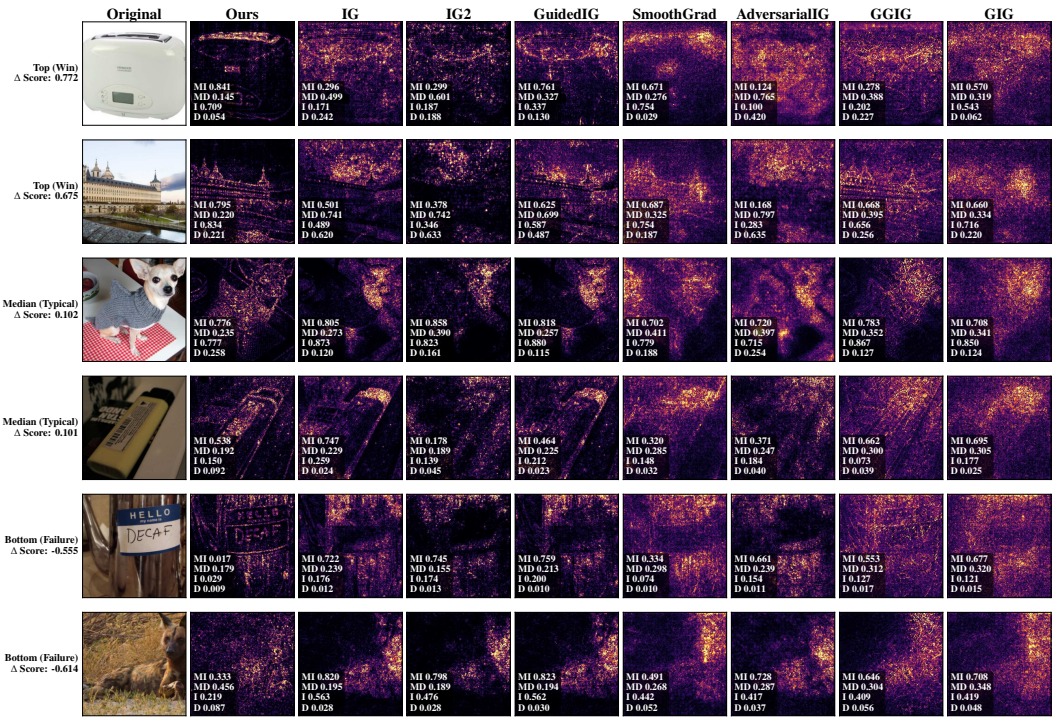

Figure 16: ResNet-101 magnitude maps under the same stratified selection and display conventions.

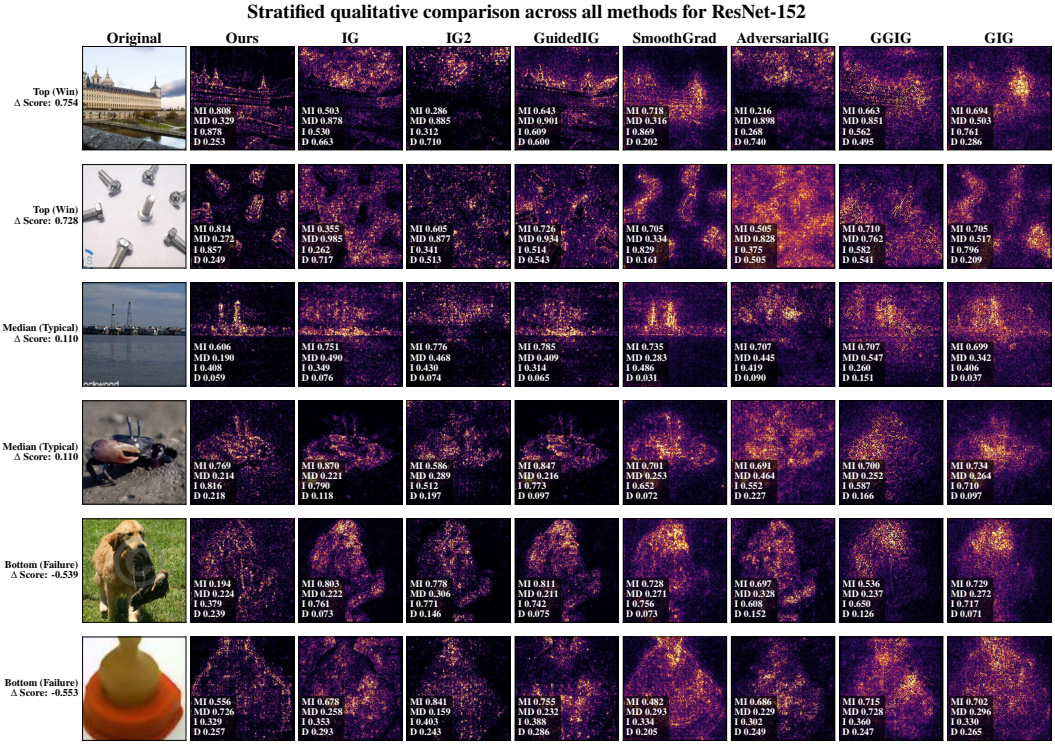

Figure 17: ResNet-152 magnitude maps under the same stratified selection and display conventions.



Figure 18: Qualitative comparison across the displayed attribution methods on Inception-v3 under the same stratified selection and display conventions.

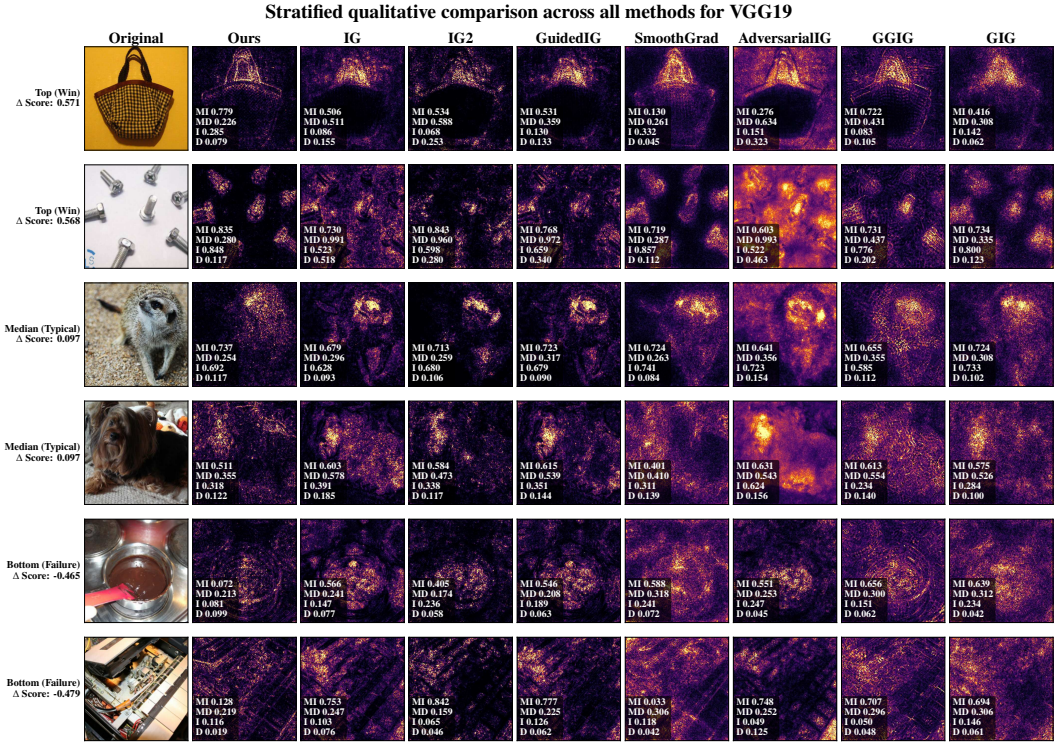

Figure 19: VGG-19 magnitude maps under the same stratified selection and display conventions.

F Target-vs-Rest Fisher–Rao Specialization (FRInGe-B)

FRInGe-B changes the predictive object before solving the geometric update. It coarse-grains the categorical prediction into the probability of the explained class versus the probability of all remaining classes. Within this Bernoulli model, its Fisher–Rao geometry and rank-one pullback are exact. Relative to categorical FRInGe, the deliberate modeling simplification is the removal of variation *within* the non-target group. FRInGe-B is therefore a lower-cost specialization, not a numerical approximation that converges to the categorical formulation as its solver tolerance is tightened.

We first define the chance-preserving binary reference and its Fisher–Rao waypoints, then derive the rank-one direction and the spatially regularized correction. A metric decomposition identifies exactly which categorical information is discarded, and the algorithm combines these ingredients with the same attribution quadrature as categorical FRInGe. The empirical part separates a matched 64-image diagnostic pilot, independently tuned full-cohort comparisons, and shared-setting qualitative examples. We conclude with the endpoint and modality limitations needed to interpret the specialization. Exactness below concerns the binary predictive model and algebraic identities; finite linear solves, input-space tracking, and quadrature remain approximate.

Table 15: FRInGe-B at a glance. *Exact* refers to the coarse-grained Bernoulli model, not to the original C-class geometry.

Predictive state	Target probability $p_t(x)$ versus rest probability $1 - p_t(x)$
Reference	$(1/C,\, 1 - 1/C)$, the image of the categorical uniform distribution
Pullback geometry	Exact Bernoulli Fisher metric; rank at most one
Classifier-dependent solve	One target-vs-rest log-odds gradient; no Fisher JVP/VJP products
Information retained	Changes that move probability mass between class t and its complement
Information discarded	Redistribution of mass among the $C - 1$ non-target classes

F.1 Exact Binary Predictive Geometry

FRInGe-B applies the same predictive-space construction as FRInGe after grouping the classes into the explained class t and all remaining classes. With $p_t(x) = \mathrm{softmax}(F(x))_t$, the resulting binary prediction is

$$\widetilde{p}_t(x) = \begin{bmatrix} p_t(x) \\ 1 - p_t(x) \end{bmatrix}, \tag{27}$$

where the second entry is the total probability assigned to non-target classes. Applying this grouping to the uniform categorical reference gives

$$\widetilde{q}_t^{\star} = \begin{bmatrix} 1/C \\ 1 - 1/C \end{bmatrix}. \tag{28}$$

This reference preserves the original chance probability $1/C$ for class t. For $C > 2$, it therefore differs from the unconstrained binary maximum-entropy distribution $(1/2, 1/2)$. When $C = 2$, the grouping preserves the full predictive distribution.

Starting from $x_0 = x$, we construct the Fisher–Rao path from $\widetilde{p}_t(x_0)$ to $\widetilde{q}_t^{\star}$. Let $\widetilde{s}(\alpha)$ denote its square-root representation, obtained using the spherical interpolation in Eq. (3). As in FRInGe, we discretize this path into waypoints:

$$\widetilde{s}_k = \widetilde{s}(\alpha_k) = \begin{bmatrix} \widetilde{s}_{k,t} \\ \widetilde{s}_{k,\mathrm{rest}} \end{bmatrix}, \qquad \alpha_k = \frac{k}{T_{\mathrm{B}}}, \qquad k = 0, \ldots, T_{\mathrm{B}},$$

where T_{B} is the number of binary tracking steps. The index k identifies the waypoint, while t and rest identify its two coordinates. Specifically, $\widetilde{s}_{k,t}^2$ is the prescribed target-class probability and $\widetilde{s}_{k,\mathrm{rest}}^2$ is the prescribed probability of all other classes combined, with $\widetilde{s}_{k,t}^2 + \widetilde{s}_{k,\mathrm{rest}}^2 = 1$. These waypoint probabilities need not equal those attained at the realized input x_k. At state x_k, we target the next waypoint $\widetilde{s}_{k+1}$, following the same indexing convention as FRInGe.

To derive the binary pullback metric, define the target-vs-rest log-odds

$$\delta_t(x) = F^t(x) - \log\left(\sum_{j\neq t} \exp F^j(x)\right), \qquad p_t(x) = \sigma(\delta_t(x)), \tag{29}$$

where σ is the logistic sigmoid, and let $u_t(x) = \nabla_x \delta_t(x) \in \mathbb{R}^n$. The Bernoulli Fisher information in log-odds coordinates is $p_t(x)(1-p_t(x))$. Its exact pullback to input space is therefore

$$G_{\mathrm{B}}(x) = p_t(x)(1-p_t(x))\, u_t(x) u_t(x)^\top \in \mathbb{R}^{n\times n}. \tag{30}$$

This metric has rank at most one, independently of the number of original classes.

The spherical alignment loss toward the next waypoint is

$$\begin{aligned}\widetilde{\mathcal{L}}_{k+1}(x) &= 1 - \left\langle \sqrt{\widetilde{p}_t(x)}, \widetilde{s}_{k+1}\right\rangle \\ &= 1 - \sqrt{p_t(x)}\,\widetilde{s}_{k+1,t} - \sqrt{1-p_t(x)}\,\widetilde{s}_{k+1,\mathrm{rest}}.\end{aligned}$$

The waypoint is held fixed when differentiating this loss. Since its dependence on x is entirely through $\delta_t(x)$, the gradient is collinear with $u_t(x)$. Defining

$$p_{t,k} = p_t(x_k), \qquad u_k = u_t(x_k),$$

we obtain the same per-step gradient structure used by FRInGe:

$$\begin{aligned} g_k &= \nabla_x \widetilde{\mathcal{L}}_{k+1}(x)|_{x=x_k} = \beta_k u_k, \\ \beta_k &= \frac{1}{2}\left[p_{t,k}\sqrt{1-p_{t,k}}\,\widetilde{s}_{k+1,\mathrm{rest}} - (1-p_{t,k})\sqrt{p_{t,k}}\,\widetilde{s}_{k+1,t}\right]. \end{aligned} \tag{31}$$

Thus, β_k combines the current target probability $p_{t,k}$ with the prescribed coordinates of the next waypoint $\widetilde{s}_{k+1}$.

F.2 Rank-One Direction and Spatial Regularization

Without spatial smoothing, the damped FRInGe-B system is

$$\left(\lambda I + p_{t,k}(1-p_{t,k}) u_k u_k^\top\right) v_k = g_k = \beta_k u_k.$$

Because the right-hand side lies in the span of u_k, its exact solution is

$$v_k = \frac{\beta_k}{\lambda + p_{t,k}(1-p_{t,k})\|u_k\|_2^2}\, u_k. \tag{32}$$

The damping $\lambda > 0$ keeps the denominator positive when the target probability is saturated or u_k is small. Equation (32) requires one reverse-mode derivative of the target-vs-rest log-odds and no pullback-Fisher products or Fisher conjugate-gradient solve. This is the main source of the computational reduction.

The benchmarked image-model variant retains both spatial terms of categorical FRInGe. Define

$$H = \lambda I + \gamma_{\mathrm{step}} L^\top L, \qquad b_k = g_k + \gamma_{\mathrm{prior}} L^\top L x_k.$$

Here H is fixed across waypoints, whereas b_k depends on the current state. The direction solves $\left(H + p_{t,k}(1-p_{t,k}) u_k u_k^\top\right) v_k = b_k$. The Sherman–Morrison identity gives

$$v_k = H^{-1} b_k - \frac{p_{t,k}(1-p_{t,k}) H^{-1} u_k\, u_k^\top H^{-1} b_k}{1 + p_{t,k}(1-p_{t,k}) u_k^\top H^{-1} u_k}. \tag{33}$$

This formula is exact when H^{-1} is applied exactly. In the evaluated spatially regularized variant, the two inverse applications are approximated by iterative solves involving the fixed spatial operator. Thus, only the unsmoothed direction is closed form; the smoothed variant still has numerical solve error. These spatial iterations do not evaluate the classifier, which is the computational distinction from the categorical Fisher solve.

The Fisher norm used by the trust-region rule also becomes scalar:

$$\|v_k\|^2_{G_{\mathrm{B}}(x_k)} = p_{t,k}(1-p_{t,k})(u_k^\top v_k)^2. \tag{34}$$

The KL and Euclidean caps, trapezoidal attribution quadrature, sign convention, and realized-endpoint completeness relation are otherwise unchanged.

F.3 Information Discarded by Aggregation

The rank-one metric in Eq. (30) is exactly the between-group component of the categorical pullback. The following decomposition identifies the omitted information.

Proposition F.1 (Categorical metric decomposition)**.** *Let $h_j(x) = \nabla_x F^j(x)$ and define the conditional non-target weights and mean gradient*

$$r_j(x) = \frac{p_j(x)}{1 - p_t(x)}, \qquad \bar{h}_r(x) = \sum_{j \neq t} r_j(x) h_j(x).$$

Then $u_t(x) = h_t(x) - \bar{h}_r(x)$ and

$$G(x) = G_{\mathrm{B}}(x) + \big(1 - p_t(x)\big) \sum_{j \neq t} r_j(x) \big(h_j(x) - \bar{h}_r(x)\big) \big(h_j(x) - \bar{h}_r(x)\big)^\top. \tag{35}$$

Consequently, $G_{\mathrm{B}}(x) \preceq G(x)$ in the positive-semidefinite order, with equality for binary classification or whenever the conditional non-target logit gradients are identical.

Proof. The categorical pullback is the covariance of the logit gradients: $G = \sum_j p_j (h_j - \bar{h})(h_j - \bar{h})^\top$, where $\bar{h} = \sum_j p_j h_j$. Applying the law of total covariance to the partition $\{t\}$ versus $\{j : j \neq t\}$ gives the between-group term $p_t(1 - p_t)(h_t - \bar{h}_r)(h_t - \bar{h}_r)^\top = G_{\mathrm{B}}$ and the conditional within-rest covariance in Eq. (35). The latter is positive semidefinite. □

Thus, two categorical predictions with the same p_t but different allocations among the remaining classes are indistinguishable to FRInGe-B. The specialization trades this within-rest information for a rank-one, target-aware geometry. Removing that information changes the direction and step control; the decomposition does not predict that either specialization will yield better attributions. That is an empirical question addressed by the comparisons below.

F.4 FRInGe-B Algorithm

Algorithm 2 makes the connection to the categorical procedure explicit. The predictive state and Fisher action are replaced by their binary forms, while the fixed explanation target, step constraints, trapezoidal accumulation, and sign convention are retained. The distinction between exact and iterative direction computations is shown in the two branches.

Algorithm 2 Target-vs-Rest FRInGe (FRInGe-B)

Require: Input x, target t, logits F, class count C; λ, τ, $\eta_{\max}$, δ_{euc}; optional L, γ_{step}, γ_{prior}
Returns: Attributions A

$p_{t,0} \leftarrow \mathrm{softmax}(F(x))_t$, $\widetilde{q}_t^\star \leftarrow (1/C, 1 - 1/C)^\top$
Compute T_{B} from Eq. (14) using $d_{\mathrm{FR}}(\widetilde{p}_t(x), \widetilde{q}_t^\star)$
Build binary Fisher–Rao waypoints $\{\widetilde{s}_j\}_{j=0}^{T_{\mathrm{B}}}$ from $(\sqrt{p_{t,0}}, \sqrt{1 - p_{t,0}})^\top$ to $\sqrt{\widetilde{q}_t^\star}$
$x_0 \leftarrow x$, $A_{\mathrm{path}} \leftarrow 0$
for $k = 0, \ldots, T_{\mathrm{B}} - 1$ **do**
    Compute $p_{t,k}$, $\delta_t(x_k)$, and $u_k \leftarrow \nabla_x \delta_t(x_k)$
    Compute β_k from Eq. (31) using $\widetilde{s}_{k+1}$
    if $\gamma_{\mathrm{step}} = \gamma_{\mathrm{prior}} = 0$ **then**
        $v_k \leftarrow \beta_k u_k / \big(\lambda + p_{t,k}(1 - p_{t,k}) \|u_k\|_2^2\big)$
    else
        Compute v_k from Eq. (33)
    end if
    $n_k^2 \leftarrow p_{t,k}(1 - p_{t,k})(u_k^\top v_k)^2$
    $\eta_k \leftarrow \min\Big(\eta_{\max}, \sqrt{2\tau/(n_k^2 + \varepsilon)}, \delta_{\mathrm{euc}}/(\|v_k\|_2 + \varepsilon)\Big)$
    $x_{k+1} \leftarrow x_k - \eta_k v_k$
    $A_{\mathrm{path}} \leftarrow A_{\mathrm{path}} + \frac{1}{2}(\nabla F^t(x_k) + \nabla F^t(x_{k+1})) \odot (x_{k+1} - x_k)$
end for
return $A = -A_{\mathrm{path}}$

F.5 Computational Cost

Classifier and spatial work. Let T_{B} be the number of binary tracking steps and let N, T, and M have the meanings given in Appendix A.7. Each binary step needs one log-odds gradient and one target-score gradient, but no classifier-dependent Fisher JVP/VJP products. With the same leading-order gradient accounting used for FRInGe, its classifier work is

$$\mathcal{C}_{\mathrm{B}}^{\mathrm{model}} \approx 2T_{\mathrm{B}}(C_{\mathrm{fwd}} + C_{\mathrm{bwd}}), \qquad \frac{\mathcal{C}_{\mathrm{B}}^{\mathrm{model}}}{\mathcal{C}_{\mathrm{IG}}^{\mathrm{model}}} \approx \frac{2T_{\mathrm{B}}}{N}, \qquad \frac{\mathcal{C}_{\mathrm{B}}^{\mathrm{model}}}{\mathcal{C}_{\mathrm{FRInGe}}^{\mathrm{model}}} \approx \frac{2T_{\mathrm{B}}}{T(M+2)}.$$

At equal tracking-step counts, the leading classifier-work ratio of FRInGe to FRInGe-B is therefore approximately $(M+2)/2$. These expressions do not compare total runtimes: the methods can take different numbers of steps, and their sequential steps and batching patterns differ.

For the spatially regularized binary direction, each step also needs two approximate applications of H^{-1}, with H defined above. If each application takes at most K_H iterations and applying the fixed spatial operator costs C_L, the unbatched-equivalent spatial work is $O(2T_{\mathrm{B}}K_H C_L)$. These fixed-operator iterations never evaluate the classifier. The evaluated binary solver starts from zero at each waypoint, whereas the categorical solver uses the previous direction as its initial guess. Initialization and truncation can therefore affect the attained accuracy even with equal iteration ceilings. Removing classifier-dependent Fisher products does not by itself guarantee a wall-clock speedup over IG.

F.6 Empirical Comparisons

The following studies serve different purposes. The pilot controls the input cohort and operating point to compare measured quality and latency. The full-cohort comparison assesses separately tuned pipelines over the larger image collection. The gallery holds scalar settings fixed to illustrate differences in spatial appearance. Their protocols should not be merged into a single geometry-only or speed–quality claim.

Study	Images and architectures	Parameter and solver policy
Paired quality–latency pilot	64 images, Inception-v3	Shared $\tau = 0.05$, damping, spatial coefficients, and two-iteration ceilings; IG with 100/200/300 steps
Full-cohort benchmark comparison	Original cohort, six models; matched-target exclusions	Independently tuned methods and their retained numerical configurations
Matched-setting gallery	Three shared images, six models	FRInGe scalar settings for both methods; 20-iteration ceilings; no binary retuning

F.6.1 Matched 64-Image Quality–Latency Pilot

In Table 16 we compare all methods in one run on the same 64 ImageNet images and top-1 targets using Inception-v3, timing attribution for one image at a time under matched conditions. Both FRInGe variants use $\tau = 0.05$, the same damping and spatial regularization, and a two-iteration ceiling for their respective solves; IG uses 100, 200, or 300 Gauss–Legendre integration steps. These settings define diagnostic operating points, not the architecture-tuned configurations of the main benchmark.

Table 16: Matched Inception-v3 runtime pilot ($n = 64$). All methods were measured in the same run. Latency is the median per-image attribution time with a 95% image-bootstrap interval, excluding initialization and metric evaluation. Steps, Fisher products, and relative completeness error are means. Steps denote IG integration steps or FRInGe tracking steps. Completeness is measured with respect to each method's own path endpoints. Speedup is the FRInGe median divided by each method's median.

Method	Steps	Fisher products	Latency (ms) [95% CI]	Speedup vs FRInGe	Comp. error ↓
IG-100	100.0	—	813.9 [805.5, 846.4]	8.0×	0.036
IG-200	200.0	—	1632.3 [1606.1, 1679.1]	4.0×	0.015
IG-300	300.0	—	2513.8 [2431.2, 2571.6]	2.6×	0.010
FRInGe-B	8.9	0.0	588.4 [559.1, 606.6]	11.1×	0.478
FRInGe	9.7	47.4	6529.1 [6282.9, 6774.5]	1.0×	0.055

FRInGe-B has the lowest median latency in this pilot, consistent with removing classifier-dependent Fisher products. However, it also has a larger completeness residual than FRInGe and IG. The speed advantage therefore does not establish equivalent path-integration accuracy at the tested operating point. Step counts alone are not a measure of computational cost, since the methods perform different work per step.

Quality across IG integration budgets. Table 17 tests whether increasing the IG integration budget changes the quality comparison on the same images and targets.

Table 17: Paired attribution-quality pilot on the same 64 Inception-v3 images. All methods come from the same run. Entries are means ± half-widths of 95% image-bootstrap confidence intervals; the original percentile intervals can be asymmetric. Bold and underlined entries have the best mean at the displayed precision.

Method	MAS-Ins ↑	MAS-Del ↓	Ins-AUC ↑	Del-AUC ↓	Infidelity ↓	Sparseness ↑
IG-100	0.692±0.031	0.438±0.065	0.651±0.050	0.256±0.051	0.0306±0.0135	0.625±0.012
IG-200	0.693±0.031	0.433±0.062	0.653±0.050	0.256±0.051	0.0306±0.0135	0.625±0.012
IG-300	0.691±0.031	0.429±0.062	0.652±0.050	0.256±0.051	0.0306±0.0134	0.625±0.012
FRInGe-B	**0.740±0.029**	**0.370±0.069**	**0.715±0.049**	**0.213±0.046**	**0.0202±0.0118**	0.681±0.012
FRInGe	0.721±0.039	**0.370±0.069**	0.649±0.049	0.253±0.046	**0.0202±0.0119**	**0.718±0.012**

Paired bootstrap comparisons favor FRInGe-B over IG-300 on the reported metrics. FRInGe is favored on MAS, Infidelity, and Sparseness, while its Ins-AUC and Del-AUC differences are inconclusive. Comparisons between IG step budgets also remain inconclusive. Together with the small changes in IG means, this suggests that integration resolution alone does not explain the observed method differences in this pilot. The intervals are descriptive and unadjusted for multiple comparisons.

Direct FRInGe versus FRInGe-B comparison. Within-image comparisons between FRInGe-B and FRInGe on the same 64-image pilot yield paired intervals favoring FRInGe-B on both Ins-AUC and Del-AUC and FRInGe on Sparseness; they include zero for both MAS scores and Infidelity. Sparseness measures concentration rather than faithfulness. Thus, the observed quality profile is mixed even when the two methods share the same regularizer, step constraints, and two-iteration solver ceiling. Equal ceilings do not imply equal direction accuracy, since the underlying linear systems and initialization policies differ.

F.6.2 Independently Tuned Architecture Comparisons

Table 18 complements this controlled pilot with the full-cohort benchmark results across all six architectures. Unlike the pilot, these rows use the independently tuned configuration of each method; they compare the two deployed pipelines and do not isolate the binary predictive geometry. FRInGe-B has better insertion and deletion AUC sample means on the four ResNets and Inception-v3, whereas FRInGe has both better AUC means on VGG-19 and is consistently sparser. The MAS and Infidelity

comparisons are mixed. This architecture dependence rules out a blanket quality-equivalence claim from the available results.

Table 18: Full-cohort comparison of independently tuned FRInGe and FRInGe-B configurations. Entries are means ± half-widths of 95% image-bootstrap percentile intervals (2,000 resamples; seed 42). Bold and underlined entries have the best mean within each architecture at the displayed precision, without implying statistical superiority.

Model	Method	MAS-Ins ↑	MAS-Del ↓	Ins-AUC ↑	Del-AUC ↓	Infidelity ↓	Sparseness ↑
ResNet-18	FRInGe	0.598±0.016	0.264±0.010	0.492±0.016	0.130±0.008	**0.058±0.008**	**0.706±0.002**
	FRInGe-B	**0.601±0.014**	**0.259±0.008**	**0.541±0.015**	**0.095±0.006**	0.059±0.008	0.660±0.003
ResNet-50	FRInGe	**0.648±0.014**	0.302±0.012	0.581±0.016	0.161±0.009	**0.150±0.050**	**0.702±0.002**
	FRInGe-B	0.635±0.013	**0.295±0.010**	**0.618±0.015**	**0.122±0.007**	0.151±0.051	0.643±0.003
ResNet-101	FRInGe	**0.679±0.014**	0.351±0.015	0.641±0.017	0.228±0.012	**0.121±0.022**	**0.706±0.002**
	FRInGe-B	0.660±0.012	**0.342±0.013**	**0.687±0.015**	**0.165±0.010**	0.123±0.022	0.626±0.003
ResNet-152	FRInGe	**0.691±0.013**	0.392±0.017	0.681±0.016	0.265±0.013	**0.111±0.020**	**0.703±0.002**
	FRInGe-B	0.669±0.011	**0.375±0.015**	**0.708±0.015**	**0.205±0.012**	0.113±0.019	0.625±0.003
Inception-v3	FRInGe	**0.761±0.009**	0.444±0.020	0.709±0.014	0.240±0.012	**0.020±0.005**	**0.723±0.003**
	FRInGe-B	0.743±0.008	**0.416±0.018**	**0.723±0.013**	**0.208±0.011**	0.021±0.005	0.676±0.003
VGG-19	FRInGe	**0.588±0.015**	**0.255±0.008**	**0.503±0.016**	**0.119±0.007**	0.343±0.064	**0.707±0.003**
	FRInGe-B	0.512±0.015	0.281±0.008	0.485±0.015	0.130±0.008	**0.341±0.063**	0.672±0.003

F.6.3 Qualitative Comparison under Matched Settings

Figure 20 uses three shared examples, ear, bubble, and drake, across all six architectures. The examples are selected before attribution computation from a fixed shuffled ordering. Eligibility requires the original target confidence to exceed both evaluation-baseline confidences by more than 10^{-4} on every model. No attribution map or method score enters selection. All displayed MAS scores are therefore defined, although the gallery remains conditional on baseline-confidence eligibility rather than representing a random sample of the full benchmark.

Each pair explains the same original top-1 logit under identical model-specific preprocessing. Both methods use the FRInGe scalar coefficients in Table 6, including damping, the two spatial coefficients, and step constraints, with a common 20-iteration solver ceiling. FRInGe-B is not retuned. This differs from the two-iteration pilot and the main benchmark runs with a 25-iteration ceiling; it is a controlled illustration, not an exact rerun of either quantitative study.

Metrics use the original signed tensors, while the displayed maps use absolute-channel sums, 99th-percentile clipping, and independent normalization, as in Appendix E. Finite-budget solve residuals are appreciable at these settings, so the panels compare the evaluated numerical procedures, not isolated exact geometries. Three shared images cannot establish a population-level quality ranking.

F.7 Endpoint Diagnostics and Limitations

FRInGe-B controls only the target probability. Even exact arrival at $p_t(x_T) = 1/C$ does not imply that the categorical prediction is uniform, because the remaining probability mass may be concentrated on a small number of non-target classes. The relevant binary endpoint quantities are

$$\left| p_t(x_T) - \frac{1}{C} \right|, \qquad \mathrm{KL}(\widetilde{p}_t(x_T) \,\|\, \widetilde{q}_t^{\star}) \,, \tag{36}$$

whereas the original categorical reference is assessed through

$$\mathrm{KL}\big(p(x_T) \,\|\, \tfrac{1}{C}\mathbf{1}\big) \,. \tag{37}$$

The first pair measures target neutralization; the latter measures proximity to the original maximum-entropy reference. These quantities describe different endpoint requirements; they are not interchangeable quality scores. Realized-path completeness remains $\sum_i A_i \approx F^t(x) - F^t(x_T)$ and is logically separate from either endpoint criterion. Finally, for an arbitrary non-top target with $p_t(x) < 1/C$, the reference path increases rather than decreases its probability. The present evaluation avoids this ambiguity by explaining the top-1 predicted class.

F.8 Scope and Modality-Specific Priors

The spatial operator L in Eq. (17) was chosen for the vision architectures evaluated here. It supplies an image-specific prior; Gaussian smoothing is not required by the predictive Fisher–Rao construction. Other modalities could use regularizers on suitable continuous input representations while retaining a predictive reference and geometric tracking. This separates the choice of input prior from the choice of predictive geometry, including target-versus-rest aggregation.

For example, an LLM extension could operate on token embeddings with a language-appropriate prior and a specified predictive target. Such an extension requires separate validation: embedding trajectories need not correspond to discrete text, and the predictive target and reference must be defined for the task. BlurIG's image-space Gaussian path has no direct counterpart on discrete tokens, so adapting it would also require a new corruption or interpolation path. The present image experiments establish neither method's effectiveness in that setting.

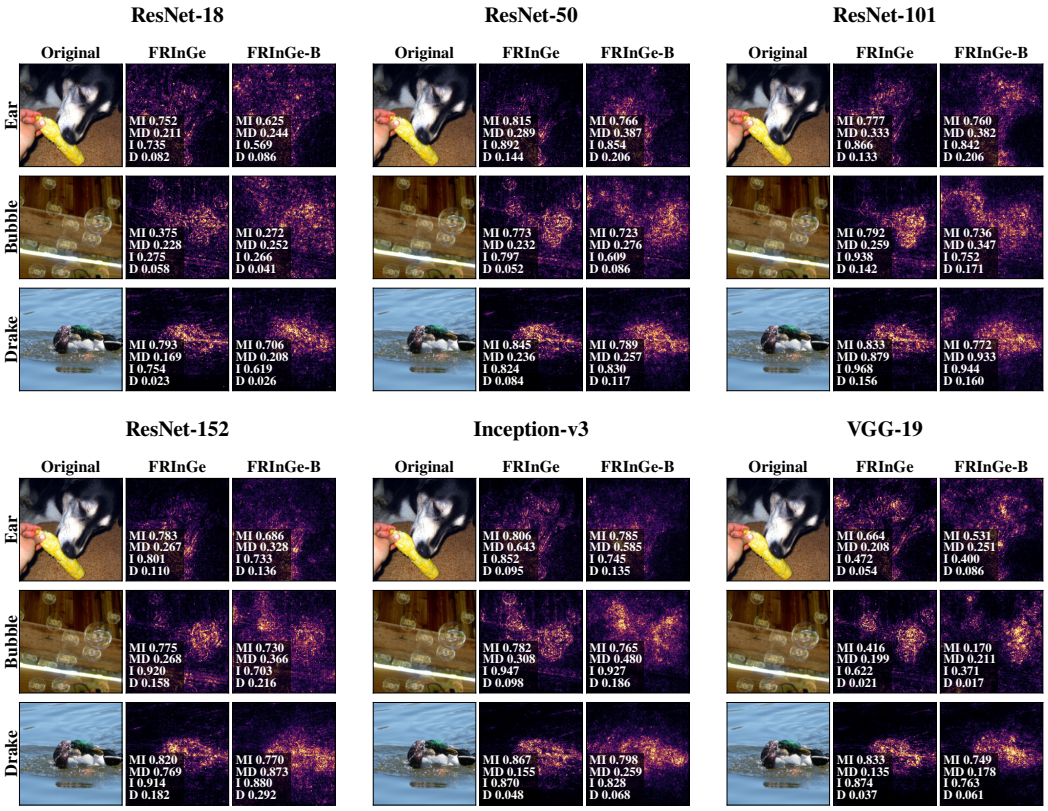

Figure 20: FRInGe and FRInGe-B on six architectures using shared FRInGe scalar settings and 20-iteration ceilings. Each model group contains Original, FRInGe, and FRInGe-B columns, with ear, bubble, and drake inputs in the same row order. All MAS scores are defined. MI/MD denote MAS-Insertion/Deletion; I/D denote normalized insertion/deletion AUC. Magnitude maps are independently normalized.

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 that FRInGe defines the reference and interpolation schedule in predictive distribution space and realizes the path through Fisher–Rao pullback geometry. The experimental claims are supported by the main quantitative results in Section 6 and the full tables in Appendix D, while trade-offs are discussed in the Discussion and Appendix.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses computational cost, the runtime overhead relative to standard IG, the effect of the Euclidean trust-region cap, and cases where FRInGe is not uniformly strongest across all diagnostics. Additional limitations and trade-offs are reported in the hyperparameter, robustness, and realized-path completeness analyses in the appendix.

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: Mathematical definitions, geometric derivations, complexity estimates, and the realized-path completeness relation are stated in Sections 3–5 and Appendix A, and Proposition F.1 (categorical metric decomposition) is stated with a complete proof in Appendix F.

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: [Yes]

   Justification: The paper specifies the dataset, evaluated models, target selection rule, attribution baselines, metric definitions, hyperparameter selection protocol, architecture-specific hyperparameters, and statistical reporting procedure in Appendix C. The FRInGe algorithm is also provided as pseudocode in Appendix A.1.

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: We provide the code repository containing the FRInGe and FRInGe-B implementations, evaluation scripts, hyperparameter configurations, and instructions for obtaining the ImageNet validation data to reproduce both the quantitative benchmarks and qualitative examples.

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: [Yes]

   Justification: The paper evaluates pretrained models and therefore does not involve model training. The experimental setup, ImageNet validation sampling, target selection, hyperparameter search procedure, selected hyperparameters, perturbation protocols, and metric definitions are specified in Section 6 and Appendix C.

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 main and appendix tables report dataset-level means with 95% bootstrap confidence intervals computed over images. The statistical reporting procedure is described in Appendix C.

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: [Yes]

   Justification: Appendices A.7 and F.6 report empirical runtime measurements for FRInGe, FRInGe-B, and standard IG, including wall-clock times on 64 ImageNet images. The experiments were run on NVIDIA H100 GPUs, and the supplemental material/code release provides the software environment, GPU memory, and execution commands needed to reproduce the reported experiments.

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 uses standard pretrained models and the ImageNet validation benchmark for post-hoc interpretability evaluation. It does not involve human-subject experiments, private data, surveillance data, or deployment decisions affecting individuals.

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 work may improve transparency of deep learning systems by producing attribution maps better aligned with model confidence dynamics. A potential negative impact is over-reliance on attribution maps in high-stakes settings; the paper therefore emphasizes empirical trade-offs, metric limitations, and the need to interpret explanations as diagnostic tools rather than guarantees of model correctness.

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 high-risk pretrained models, image generators, language models, or newly scraped datasets. The released material consists solely of the implementation and evaluation code for our post-hoc attribution method.

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: [Yes]

    Justification: The paper cites all attribution baselines, the ImageNet benchmark, and pretrained TorchVision models (BSD-3-Clause license) in Section 6 and Appendix C, and the code repository documentation specifies all software and data licenses.

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 released code package documents the FRInGe and FRInGe-B implementations, evaluation scripts, hyperparameter files, and commands needed to reproduce the reported results. No new dataset or pretrained model is introduced.

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 research with human subjects.

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 involve human-subject research and therefore does not require IRB or equivalent approval.

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 were not used as part of the proposed method, mathematical formulation, experimental design, evaluation protocol, or scientific contribution. They were used only for routine assistance with code drafting/debugging and manuscript editing.