FROM HEATMAPS TO RELIANCE FIELDS: DIAGNOSING VISUAL EXPLANATIONS UNDER UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual explanation methods are widely consumed under the assumption that spatial heatmaps directly expose model reliance. However, what an explanation highlights, what functionally supports a prediction under intervention, how evidence propagates across internal representations, and what is semantically meaningful are fundamentally non-interchangeable. We prove in **Theorem 1** that finite-order local attribution signals do not uniquely determine finite-intervention reliance. This conflation obscures critical failure modes, including highlighted-but-unused evidence, hidden reliance, and potential shortcut cues. To overcome this limitation, we introduce *Reliance Uncertainty Fields* (RUF), which formulate latent functional reliance as a layer-wise posterior distribution over an Energy-Based Factor Graph. RUF couples local preference, finite-intervention influence, representation flow, stability, and optional semantic references into a unified probabilistic field. Building on RUF, we propose *RIFT* (*Reliance Inference over Flow-based Traces*), a diagnostic framework that computes posterior reliance means, marginal entropy (uncertainty), preference–function disagreement, flow attenuation, and shortcut candidates. RIFT reframes visual explanation from generating plausible heatmaps to inferring operational functional reliance under a specified intervention family and explicit uncertainty.

1 INTRODUCTION

What does a visual explanation actually explain? For over a decade, visual explanation methods in Computer Vision have assigned spatial scores to pixels or feature maps. Input-level methods compute differential sensitivity Simonyan et al. (2013); Sundararajan et al. (2017), feature-map methods visualize class-discriminative activation prominence Zhou et al. (2016); Selvaraju et al. (2017); Wang et al. (2020); Ramaswamy et al. (2020), and perturbation methods measure output changes under spatial masking Fong & Vedaldi (2017); Petsiuk et al. (2018). Despite their procedural differences, these outputs are overwhelmingly interpreted as a singular, monolithic quantity: *visual model reliance* Zhou et al. (2022); Kares et al. (2025).

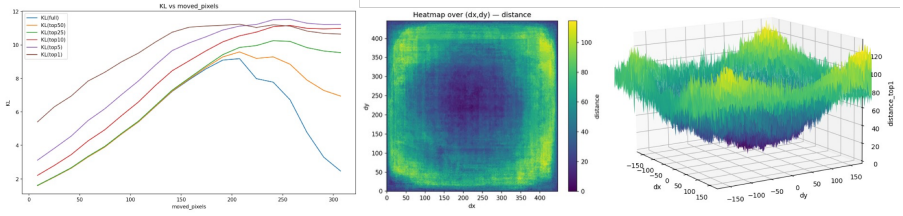

Figure 1: **Highlighted regions do not imply stable functional reliance.** Spatially displacing regions selected by an explanation map yields non-monotonic and direction-dependent KL divergence responses.

We argue that this monolithic treatment is fundamentally flawed. A high spatial score can reflect four distinct explanatory targets: (1) *Model-Perceived Importance*, spatial emphasis by gradients or activations Simonyan et al. (2013); Selvaraju et al. (2017); (2) *Learning-Influential Importance*, training samples or loss factors shaping decision boundaries Koh & Liang (2017); Ilyas et al. (2022);

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

(3) *Prediction-Influential Importance*, representation units functionally supporting a test-time output under intervention Fong & Vedaldi (2017); Petsiuk et al. (2018); and (4) *Semantic Importance*, image regions aligning with human task concepts Bau et al. (2017); Kim et al. (2018). RIFT focuses on test-time explanation reliability and excludes training-time influence. Here, functional reliance denotes prediction-stage influence defined relative to a specified intervention family, baseline distribution, and representation graph, without assuming an intervention-independent ground-truth map. While these notions may coincide in simple cases, they routinely diverge in complex visual architectures Adebayo et al. (2018); Hooker et al. (2019); Yeh et al. (2019); Tomsett et al. (2020). Conflating them creates severe blind spots: a highlighted region may fail to support predictions (*highlighted-but-unused evidence*), a model may rely on unhighlighted background regions (*hidden reliance*), or depend on non-semantic shortcuts Lapuschkin et al. (2019); Geirhos et al. (2020). We establish in **Theorem 1** that finite-intervention reliance is non-identifiable from finite-order local attribution signals alone. As empirically illustrated in Fig. 1, perturbing highly ranked attribution regions produces non-monotonic and direction-dependent responses, demonstrating that visual highlightedness alone cannot predict functional reliance. Furthermore, reliance is not static: representations evolve across network layers, amplifying, suppressing, or redistributing evidence Zeiler & Fergus (2014); Yosinski et al. (2015). Existing single-layer explanation methods fail to capture whether highlighted evidence survives downstream representation flow or decays prior to prediction.

To resolve this challenge, we shift the paradigm of visual explanation from *deterministic heatmap generation* to *probabilistic posterior inference*. We introduce **Reliance Uncertainty Fields (RUF)**, an Energy-Based Factor Graph / Markov Random Field that models latent functional reliance $R_i^c \in \{0, 1\}$ over a layer-wise representation graph $G = (V, E)$. RUF integrates local preference evidence, finite-intervention influence, pairwise representation flow, baseline stability, and optional semantic references into a unified probabilistic field. Building on RUF, we propose **RIFT** (*Reliance Inference over Flow-based Traces*), a diagnostic framework that infers posterior reliance means, marginal entropy (uncertainty), preference–function disagreement, flow attenuation, and shortcut candidates. To summarize, our key contributions are:

1. **A Four-Way Decomposition of Visual Importance and a Local Non-Identifiability Result.** We distinguish four conceptually distinct notions of visual importance and establish in **Theorem 1** that finite-intervention reliance cannot, in general, be uniquely identified from finite-order local attribution signals alone.
2. **An Energy-Based Framework for Reliance Inference and Diagnosis.** We introduce the Reliance Uncertainty Field (RUF), a layer-wise energy-based factor graph that integrates node-level evidence with pairwise potentials encoding representation flow and downstream attenuation. Building on this formulation, RIFT provides explicit diagnostic readouts of inferred reliance, marginal entropy, preference–reliance disagreement, flow reliability, hidden reliance, highlighted-but-unused evidence, and potential shortcut reliance.
3. **A Multi-Domain Evaluation of Explanation Reliability.** We evaluate RIFT across image classification, corruption robustness, dense prediction, medical imaging, and visual tracking, assessing response calibration, failure-mode detection, architectural generality, robustness to modeling choices, and sparse-probing efficiency.

2 PRELIMINARIES

We introduce the notation and formal objects used throughout the paper. Let $f : \mathcal{X} \to \mathbb{R}^K$ be a trained vision model mapping an input image $x \in \mathcal{X} \subset \mathbb{R}^{C\times H\times W}$ to logit scores for K classes, and let $c \in \{1, \dots, K\}$ be a target class. We denote the class logit by $s_c(x) = [f(x)]_c$. For an intermediate layer $\ell \in \mathcal{L} = \{1, \dots, L\}$, let $F^{(\ell)}(x) \in \mathbb{R}^{C_\ell\times H_\ell\times W_\ell}$ be its feature activation tensor, and let $u \in \mathcal{U}_\ell = \{1, \dots, H_\ell W_\ell\}$ index a spatial location or token index. We write $i = (\ell, u)$ for a unique representation unit in the network.

Observable explanation signals. Existing visual explanation methods provide observable spatial signals about model behavior. Gradient-based methods compute local sensitivity $\nabla_x s_c(x)$ Simonyan et al. (2013); Sundararajan et al. (2017); Smilkov et al. (2017); perturbation-based methods evaluate score drops under spatial masking Fong & Vedaldi (2017); Petsiuk et al. (2018); Samek et al. (2016); and feature-map methods visualize class-discriminative activation prominence Zhou et al.

(2016); Selvaraju et al. (2017); Wang et al. (2020). We define these normalized explanatory outputs as *preference evidence* $A_E^c(x;i) \in [0,1]$. A preference score reflects what an explanation algorithm emphasizes, but does not by itself establish that unit i is functionally relied upon under the intervention family.

Functional intervention response. For a representation unit $i = (\ell, u)$, let $\widetilde{s}_c(x; i \leftarrow b_i)$ be the target class logit obtained after intervening on unit i by replacing feature vector $F_u^{(\ell)}(x)$ with a baseline feature $b_i \sim q_i$. We define the signed finite-logit response as:

$$d_i^c(b_i) = s_c(x) - \widetilde{s}_c(x; i \leftarrow b_i). \tag{1}$$

More generally, finite-intervention responses can be evaluated via non-negative discrepancy measures $\Delta(s_c(x), \widetilde{s}_c(x; i \leftarrow b_i))$ such as logit drop, margin drop, or loss shift.

Representation graph and flow. To model how evidence propagates through intermediate layers, we construct a representation graph $G = (V, E)$ with node set $V = \{(\ell, u) : \ell \in \mathcal{L}, u \in \mathcal{U}_\ell\}$. Edges $(i, j) \in E$ connect upstream node $i = (\ell, u)$ to downstream-reachable node $j = (m, v)$ $(m > \ell)$. For convolutional backbones, edges are governed by spatial receptive-field reachability; for vision transformers, edges reflect patch-token connectivity and attention pathways. For any edge $(i, j) \in E$, we define intervention-induced representation flow as:

$$\Phi_{i \to j}^c = \mathbb{E}_{b_i \sim q_i} \left[d_F \left(F_v^{(m)}(x), \widetilde{F}_v^{(m)}(x; i \leftarrow b_i) \right) \right], \tag{2}$$

where $d_F(a, b) = \|a - b\|_2 / (\|a\|_2 + \epsilon)$ is a normalized feature-space distance.

Definition 1 (Latent Functional Reliance)**.** *For each representation unit $i = (\ell, u) \in V$ and class c, we define* Latent Functional Reliance *as an unobserved binary state variable $R_i^c \in \{0, 1\}$, where $R_i^c = 1$ indicates that representation unit i functionally supports the prediction $s_c(x)$ under the intervention family and downstream propagation, and $R_i^c = 0$ otherwise. Since R_i^c cannot be directly read from static activations, observable explanation maps, intervention responses, flow measurements, and stability estimates serve as noisy evidence about R_i^c.*

Non-identifiability from preference alone. The following theorem formalizes why a static attribution map is fundamentally insufficient to determine functional reliance.

Theorem 1 (Non-Identifiability of Functional Reliance from Local Preference)**.** *Let i be a representation unit and let $A_E^c(x; i)$ be an attribution functional determined solely by the local k-jet of f at i—i.e., by the activation value, gradient, and higher-order derivatives up to order k evaluated at i. There exist pairs of functions (f_1, f_2) with continuous derivatives through order $(k+1)$ such that $A_{E,1}^c(x; i) = A_{E,2}^c(x; i)$ (identical local preference at i) yet the signed finite-logit responses satisfy $|d_{i,1}^c(b_i) - d_{i,2}^c(b_i)| > C$ for any prescribed $C > 0$. Consequently, the finite-intervention response is non-identifiable from finite-order local attribution evidence alone. Furthermore, any binary reliance state $R_i^c \in \{0, 1\}$ derived via thresholding over finite-intervention responses is also non-identifiable without additional structural assumptions.*

Intuition. Local preference signals (gradients, activations) capture the tangent behavior of f at the evaluation point, but finite intervention displaces the representation to a baseline that may lie far from this point. A nonlinear f can exhibit identical local derivatives yet arbitrarily different global curvature, causing matched preference scores to accompany divergent intervention responses on the *same* unit. Theorem 1 is a negative identifiability result for finite-order local attribution signals. It motivates the need for additional intervention and structural evidence, but does not by itself establish that RIFT recovers a unique or ground-truth reliance field. The formal proof is given in Appendix C.2.1.

Remark 1. *Theorem 1 covers gradient-magnitude, activation-magnitude, and higher-order-jet attribution methods. Path-integrated methods (e.g., Integrated Gradients) accumulate gradients along the full intervention path and may carry partial functional information; they are not covered by this result. Perturbation-based methods directly measure intervention responses and thus incorporate functional evidence by construction.*

3 METHODOLOGY

We propose **RIFT** (*Reliance Inference over Flow-based Traces*), illustrated in Fig. 2, a probabilistic framework for diagnosing visual explanation reliability by inferring a posterior field of functional reliance over a representation graph $G = (V, E)$.

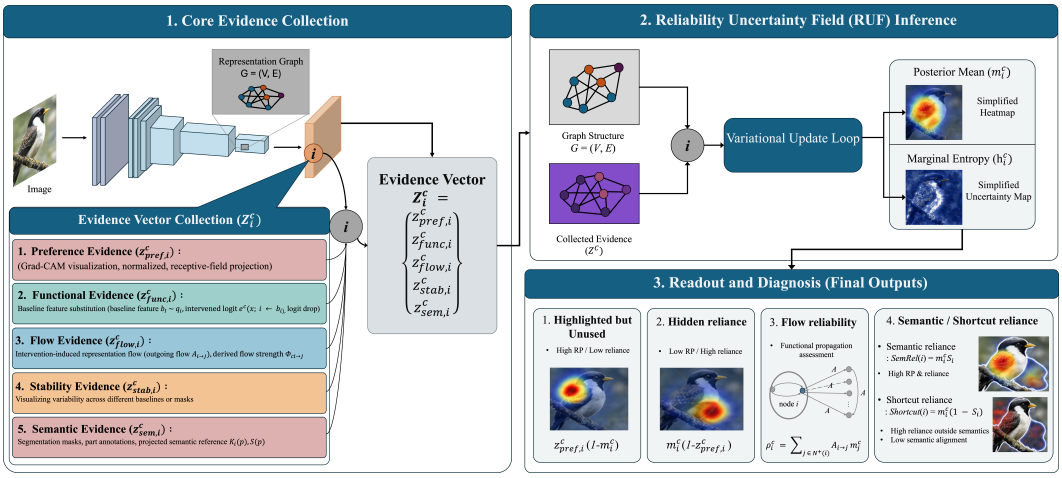

Figure 2: **RIFT method overview. (1)** RIFT collects preference, finite-intervention, flow, stability, and optional semantic evidence over a representation graph. **(2)** RIFT integrates these signals to infer the RUF, providing node-wise reliance estimates and uncertainty. **(3)** These posterior readouts support explanation-reliability and failure-mode diagnostics.

Reliance Uncertainty Field (RUF) Given image x, target class c, and representation graph $G = (V, E)$, RIFT models the full configuration of latent reliance variables $R^c = \{R_i^c \in \{0, 1\} : i \in V\}$ as an Energy-Based Factor Graph / Markov Random Field:

$$p(R^c \mid Z^c, G, x) = \frac{1}{\mathcal{Z}} \exp\left(-E(R^c; Z^c, G, x)\right), \tag{3}$$

where $\mathcal{Z} = \sum_{R^c} \exp(-E(R^c))$ is the partition function, and $Z^c = \{Z_i^c : i \in V\}$ represents the collection of evidence vectors across all nodes. We call $p(R^c \mid Z^c, G, x)$ the *Reliance Uncertainty Field* (RUF). RUF is an energy-based posterior belief induced by the specified unary evidence terms and pairwise flow potentials. It should not be interpreted as a posterior over a uniquely observable ground-truth explanation. Accordingly, its marginal entropy quantifies uncertainty relative to the adopted evidence model, intervention family, graph construction, and fitted coefficients. The potential coefficients are selected on held-out validation subsets and are therefore interpreted as calibrated evidence weights rather than parameters of a known data-generating process.

Evidence Construction For node i, RIFT constructs a four-dimensional node evidence vector $Z_i^c = \{z^c_{\text{pref},i}, z^c_{\text{func},i}, z^c_{\text{stab},i}, z^c_{\text{sem},i}\}$.

Preference evidence. For explainer E, preference evidence is $z^c_{\text{pref},i} = \text{Norm}(A_E^c(x; i))$. When A_E^c is defined in image space, it is aggregated to node i via projection kernel $K_i(p)$.

Functional evidence. By sampling feature replacement $b_i \sim q_i$, functional evidence is computed as $z^c_{\text{func},i} = \mathbb{E}_{b_i\sim q_i}[\Delta(s_c(x), \widetilde{s}_c(x; i \leftarrow b_i))]$. This measures prediction-stage influence under finite intervention.

Edge flow evidence. For each edge $(i, j) \in E$, RIFT additionally observes edge-level flow evidence $A_{i\to j} = \Phi^c_{i\to j}/(\sum_{j'\in\mathcal{N}^+(i)} \Phi^c_{i\to j'} + \epsilon)$, quantifying the fraction of intervention disturbance at i that propagates to downstream node j.

Stability and semantic evidence. Stability evidence $z^c_{\text{stab},i} = 1 - \text{Norm}\left(\text{Var}_{b_i}[\Delta^c_i(b_i)]\right)$ rewards invariant responses across baseline samples. Optional semantic evidence $z^c_{\text{sem},i} = \sum_{p\in\Omega_x} K_i(p)S(p)$ incorporates external reference masks $S(p)$ when available.

Held-Out Node Partitioning. Functional intervention responses on V_{eval} are never used during posterior construction, candidate selection, graph smoothing, or hyperparameter selection. Evaluation nodes may receive non-functional observable evidence, but their held-out intervention responses are accessed strictly after inference for metric computation. All reported functional-predictiveness results use disjoint probe and evaluation nodes ($V_{\text{probe}} \cap V_{\text{eval}} = \emptyset$). This node-level partition prevents direct reuse of evaluation-node intervention responses. However, because probe and evaluation nodes belong to the same image and representation graph, this protocol evaluates held-out node prediction and sparse spatial generalization rather than fully independent cross-image generalization.

Energy-Based Factor Graph Posterior The energy function $E(R^c; Z^c, G, x)$ decomposes into unary node energy terms and binary flow interaction potentials:

$$E(R^c; Z^c, G, x) = -\sum_{i\in V} \phi_i(R^c_i; Z^c_i) - \sum_{(i,j)\in E} \phi_{ij}(R^c_i, R^c_j), \tag{4}$$

yielding the factorized Gibbs posterior field over the representation graph:

$$p(R^c \mid Z^c, G, x) = \frac{1}{\mathcal{Z}} \prod_{i\in V} \psi_i(R^c_i; Z^c_i) \prod_{(i,j)\in E} \psi_{ij}(R^c_i, R^c_j). \tag{5}$$

The unary potential $\psi_i(R^c_i) = \exp(R^c_i \eta^c_i)$ models local evidence sufficiency:

$$\eta^c_i = \beta_{\text{pref}} z^c_{\text{pref},i} + \beta_{\text{func}} z^c_{\text{func},i} + \beta_{\text{stab}} z^c_{\text{stab},i} + \beta_{\text{sem}} z^c_{\text{sem},i}, \tag{6}$$

where $\beta_{\text{sem}} = 0$ when no semantic reference is available. Crucially, RUF goes beyond a naive point-wise weighted sum by incorporating the pairwise flow potential $\psi_{ij}(R^c_i, R^c_j)$:

$$\begin{aligned}\psi_{ij}(R^c_i, R^c_j) = \exp\Big(&\lambda_{\text{flow}} A_{i\to j} R^c_i R^c_j \\ &- \lambda_{\text{cut}} A_{i\to j} R^c_i (1 - R^c_j)\Big).\end{aligned} \tag{7}$$

The term $\lambda_{\text{flow}} A_{i\to j} R^c_i R^c_j$ reinforces co-reliance whenever intervention disturbances at upstream node i propagate along verified flow edges to downstream node j. Conversely, the cut penalty $-\lambda_{\text{cut}} A_{i\to j} R^c_i (1 - R^c_j)$ explicitly suppresses reliance when an upstream feature appears important in isolation but fails to survive downstream computation. This structural graph coupling enables RIFT to model layer-wise representation propagation, attenuation, and redistribution—a property not explicitly represented by static, single-layer attribution maps or point-wise evidence fusion.

Inference and Readouts We approximate $p(R^c \mid Z^c, G, x)$ via a mean-field variational distribution $q(R^c) = \prod_{i\in V} q_i(R^c_i)$, where $q_i(R^c_i = 1) = m^c_i \in [0, 1]$. The fixed-point update is

$$\begin{aligned} m^c_i \leftarrow \sigma\Bigg(&\eta^c_i \\ &+ \sum_{j\in\mathcal{N}^+(i)} \Big[\lambda_{\text{flow}} A_{i\to j} m^c_j \\ &\quad - \lambda_{\text{cut}} A_{i\to j}(1 - m^c_j)\Big] \\ &+ \sum_{k\in\mathcal{N}^-(i)} (\lambda_{\text{flow}} + \lambda_{\text{cut}}) A_{k\to i} m^c_k\Bigg), \end{aligned} \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function.

The two fundamental readouts of RUF are the posterior reliance mean $m^c_i = \mathbb{E}_{q_i}[R^c_i]$ and the reliance marginal entropy $h^c_i = H(q_i) = -m^c_i \log m^c_i - (1 - m^c_i)\log(1 - m^c_i)$. High marginal entropy h^c_i signifies high model-relative reliance uncertainty, reflecting ambiguity or disagreement among the evidence sources represented in RUF.

Six Fine-Grained Diagnostic Maps From the RUF posterior distribution $q(R^c)$, RIFT derives six explicit diagnostic outputs, transforming black-box visual explanation into actionable reliance auditing. First, the **Posterior Reliance Map** ($m_i^c = \mathbb{E}_{q_i}[R_i^c] \in [0,1]$) captures the inferred reliance score of representation unit i for class c under the specified evidence model, integrating preference, intervention, and flow. Second, the **Reliance Uncertainty Map** ($h_i^c = H(q_i) = -m_i^c \log(m_i^c + \epsilon) - (1 - m_i^c)\log(1 - m_i^c + \epsilon)$) measures marginal entropy, where high entropy indicates severe conflict or instability among evidence sources. Third, the **Highlighted-but-Unused Diagnostic** ($\mathrm{HighlightedUnused}(i) = z_{\mathrm{pref},i}^c(1 - m_i^c)$) isolates false alarms—regions heavily emphasized by attribution maps that carry low functional reliance. Fourth, the **Hidden Reliance Diagnostic** ($\mathrm{HiddenReliance}(i) = m_i^c(1 - z_{\mathrm{pref},i}^c)$) flags candidate regions with low preference scores but high inferred intervention-based reliance. Fifth, the **Flow Attenuation Diagnostic** ($\mathrm{Atten}(i) = m_i^c - \sum_{j \in \mathcal{N}^+(i)} A_{i \to j} m_j^c$) detects local features whose influence fails to survive downstream representation propagation. Finally, the **Shortcut Candidate Diagnostic** ($\mathrm{ShortcutCandidate}(i) = m_i^c(1 - S_i)$) identifies reliance outside the supplied semantic reference S_i, serving as a shortcut candidate rather than definitive evidence of spurious reliance. Implementation details and projection operators are elaborated in Appendix D.

4 Experiments

We evaluate RIFT across vision benchmarks spanning classification, robustness, dense prediction, medical imaging, and tracking. Our primary metric for functional predictiveness is the image-averaged Spearman rank correlation (ρ) between inferred reliance scores m_i^c and held-out intervention logit drops on V_{eval}. Reliance Calibration Error (RCE) measures whether posterior reliance means align numerically with normalized held-out intervention responses, evaluating relative within-instance calibration. For deterministic baselines, uncertainty is estimated using stochastic

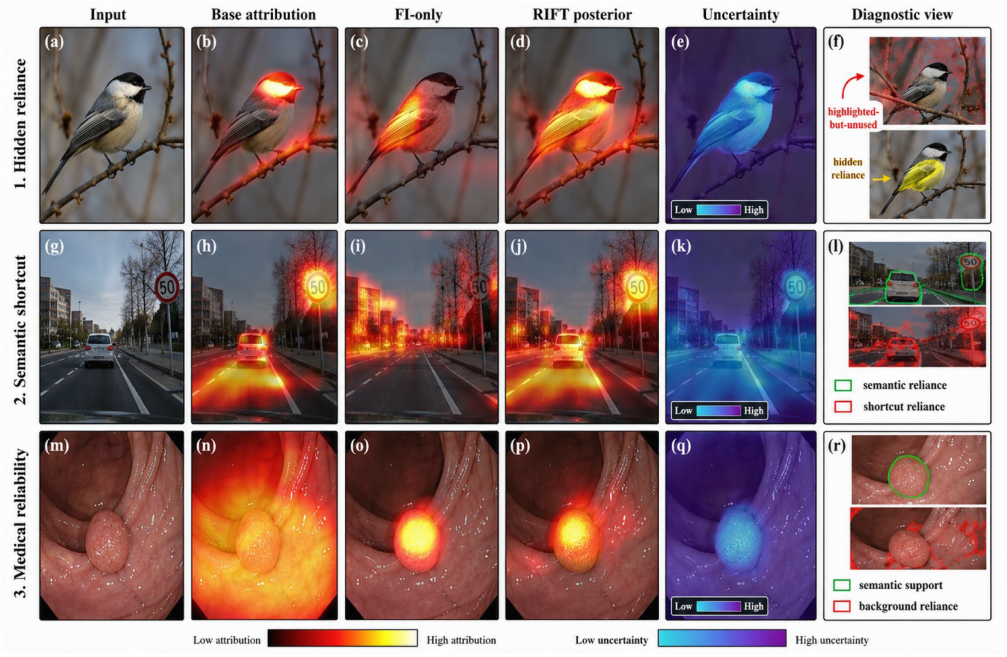

Figure 3: **Representative qualitative diagnostics.** Each row shows the input, base attribution, FI-only response, RIFT posterior, uncertainty, and diagnostic view. The final column provides schematic interpretations of the diagnosed failure mode and is not an additional output for the same input. **Row 1** shows hidden reliance missed by the base explainer. **Row 2** separates semantic reliance from reference-external reliance (shortcut candidate). **Row 3** shows medical localization, in which RIFT focuses on the clinically relevant region while revealing residual reliance on background and uncertainty.

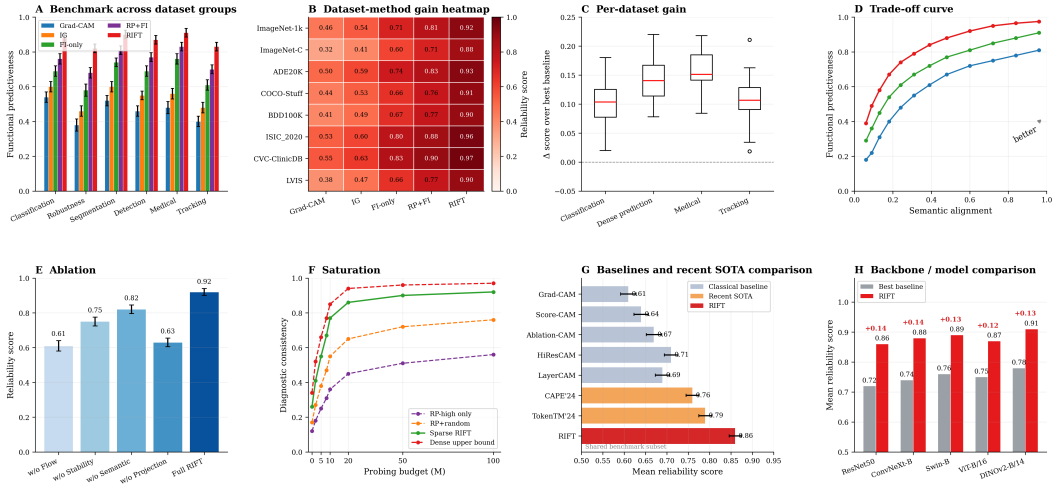

Figure 4: **Main quantitative results. (A)** RIFT improves functional predictiveness across dataset groups. **(B)** Dataset–method heatmap shows consistent gains over attribution, intervention-only, and RP+FI baselines. **(C)** Per-dataset gains over the best baseline are positive across task domains. **(D)** RIFT improves semantic alignment–functional predictiveness trade-off. **(E)** Component ablations show that flow, stability, semantic evidence, and projection-aware localization each contribute to reliability. **(F)** Sparse RIFT approaches dense probing by graph-smoothing probed evidence over unmeasured nodes. **(G)** RIFT outperforms classical baselines and recent explanation methods. **(H)** Gains persist across various backbones.

input perturbation variance under the same sample budget and normalization as RIFT stability evidence for fair error-prediction comparison.

The main evaluation uses ImageNet-1k Russakovsky et al. (2015), ImageNet-C Hendrycks & Dietterich (2019), ADE20K Zhou et al. (2017), COCO-Stuff Caesar et al. (2018), BDD100K Yu et al. (2020), ISIC 2020 Rotemberg et al. (2021), CVC-ClinicDB Bernal et al. (2015), LVIS Gupta et al. (2019), and LaSOT Fan et al. (2019). Other datasets, including ImageNet-128 Russakovsky et al. (2015), iNaturalist Van Horn et al. (2018), AFHQ Choi et al. (2020), FFHQ Karras et al. (2019), CelebA-H Liu et al. (2015), Mapillary Vistas Neuhold et al. (2017), MS-COCO-2017 Lin et al. (2014), Objects365 Shao et al. (2019), Open Images V7 Kuznetsova et al. (2020), TrackingNet Muller et al. (2018), and GOT-10K Huang et al. (2019), are reported in Appendix E. Images are resized to 224×224 for classification models, while dense prediction and tracking tasks follow their standard preprocessing protocols. We use ResNet-50 and ResNet-101 He et al. (2016) as convolutional backbones, and evaluate ConvNeXt-B Liu et al. (2022), Swin-B Liu et al. (2021), ViT-B/16 Dosovitskiy et al. (2020), and DINOv2 ViT-B/14 Oquab et al. (2023) to test architectural generality. RIFT is compared with representative gradient-, CAM-, perturbation-, and transformer-based explanation methods, including Grad-CAM Selvaraju et al. (2017), Grad-CAM++ Chattopadhay et al. (2018), XGrad-CAM Fu et al. (2020), Score-CAM Wang et al. (2020), Ablation-CAM Ramaswamy et al. (2020), HiResCAM Draelos & Carin (2020), LayerCAM Jiang et al. (2021), Integrated Gradients Sundararajan et al. (2017), SmoothGrad Smilkov et al. (2017), RISE Petsiuk et al. (2018), Attention Rollout Abnar & Zuidema (2020), TokenTM Wu et al. (2024), and CAPE Chowdhury et al. (2024). Architecture-specific methods are evaluated only on compatible backbones, and shared-subset comparisons are reported.

Qualitative Diagnostics Fig. 3 illustrates how the posterior readouts translate into diagnostic explanations. In the hidden-reliance example, the base attribution focuses on a visually salient region, but FI-only and RIFT posterior maps reveal additional regions that functionally affect the prediction. In the semantic-shortcut example, RIFT assigns substantial posterior mass outside the semantic reference, exposing reliance on contextual cues. In the medical example, RIFT localizes high-confidence reliance on the lesion or polyp core while preserving uncertainty over ambiguous surrounding tissue. These examples show that RIFT is not intended to produce only a sharper

heatmap; it provides a structured diagnosis of where explanations align with, miss, or mischaracterize functional reliance.

Quantitative Results Fig. 4 summarizes the main quantitative results. Semantic-reference results are reported as reference-assisted diagnostics ('RIFT+Sem') and are separated from semantic-free functional-predictiveness evaluation ('RIFT'). RIFT achieves consistent and substantial gains across dataset groups, with improvements appearing not only on ImageNet-style classification but also under corruption, dense prediction, medical imaging, object-centric recognition, and tracking. The dataset–method heatmap further shows that the improvement is consistent across individual benchmarks rather than driven by a single domain. Bootstrap confidence intervals on the shared benchmark subset indicate stable improvements across resampled image sets with non-overlapping 95% confidence intervals for the principal comparisons. RIFT also achieves a better trade-off between semantic alignment and functional , indicating that the posterior field not only sharpens attribution maps but also better captures regions that are both functionally influential and semantically meaningful when semantic references are available.

We next examine the reasons for these gains. The ablation results in Fig. 4E show that removing flow coupling or projection-aware localization causes the largest drops, supporting the need to model both downstream survival and the distinction between representation-space reliance and naive image-space heatmap upsampling. The probing-budget curve in Fig. 4F shows that sparse RIFT rapidly approaches the dense upper bound, while probing only high-attribution regions remains substantially worse. Finally, Fig. 4H shows that gains are observed across the evaluated architecture families.

Qualitative Stress Tests on Natural Images We further examine whether qualitative diagnostic behavior extends to realistic visual inputs. Fig. 5 shows two stress tests: corruption stress and context shift. Under snow-like corruption, base attribution is distracted by background artifacts, whereas RIFT relies more on stable object evidence and assigns higher uncertainty to corrupted regions. Under context shift, base attribution spreads over weather, reflection, and scene-context cues, whereas RIFT separates semantic support for the target and drivable region from context-driven reliance (shortcut candidate). These examples show that RIFT decomposes predictive evidence into reliance, uncertainty, and diagnostic-failure modes rather than simply producing sharper heatmaps.

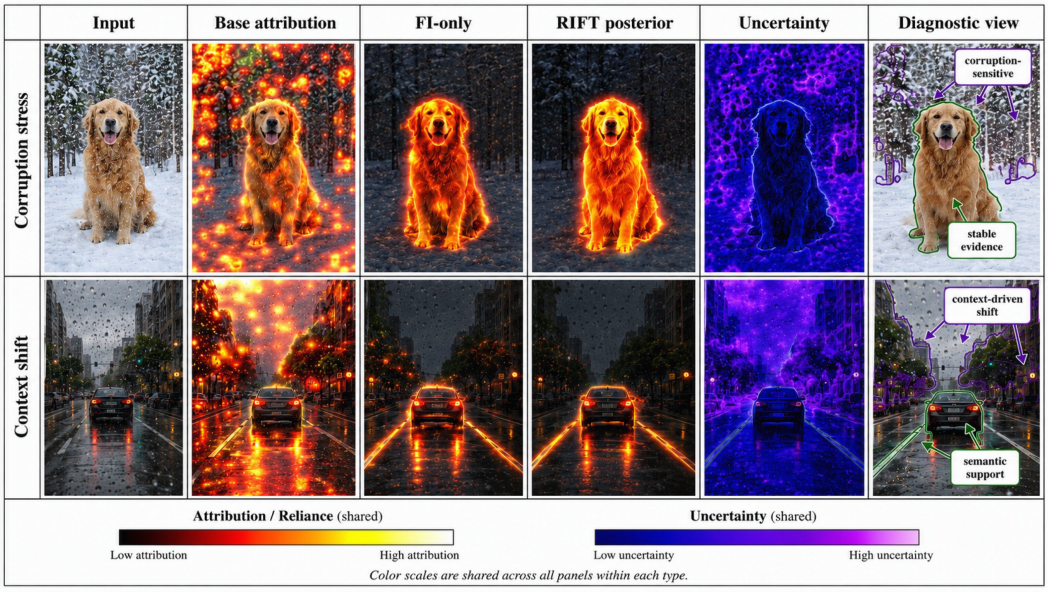

Figure 5: **Qualitative stress-test diagnostics on GPT-4-generated controlled, realistic images.** Each row shows the input, base attribution, FI-only response, RIFT posterior, uncertainty, and diagnostic view under corruption or context shifts. Red–yellow maps denote attribution or reliance strength; blue–purple maps denote uncertainty.

Table 1: **Overall comparison on the shared benchmark subset.** RIFT is evaluated in its default semantic-free setting ($\beta_{\text{sem}} = 0$) and compared with classical attribution methods, recent explanation methods, and structural variants of RIFT. Higher is better for functional predictiveness, mismatch AUROC, and uncertainty–error correlation; lower is better for RCE and relative runtime.

Method	Family	Evidence / structure modeled					Diagnostic performance				
		Pref.	FI	Flow	Stab.	Sem.	Func. pred. ↑	RCE ↓	Mismatch AUROC ↑	Unc.–err ρ ↑	Rel. time ↓
Classical attribution and CAM baselines											
Grad-CAM Selvaraju et al. (2017)	CAM	✓	✗	✗	✗	✗	0.61	0.184	0.66	0.21	1.0×
Score-CAM Wang et al. (2020)	CAM / forward	✓	✗	✗	✗	✗	0.64	0.169	0.68	0.24	7.8×
Ablation-CAM Ramaswamy et al. (2020)	CAM / ablation	✓	✓	✗	✗	✗	0.67	0.153	0.71	0.27	5.6×
HiResCAM Draelos & Carin (2020)	CAM / gradient	✓	✗	✗	✗	✗	0.71	0.141	0.73	0.29	1.2×
LayerCAM Jiang et al. (2021)	CAM / multi-layer	✓	✗	✗	✗	✗	0.69	0.148	0.72	0.31	1.8×
Recent explanation comparators											
CAPE Chowdhury et al. (2024)	Probabilistic CAM ensemble	✓	✗	✗	✓	✗	0.76	0.126	0.77	0.35	3.4×
TokenTM Wu et al. (2024)	ViT explanation	✓	✗	✓	✗	✗	0.79	0.119	0.79	0.38	2.9×
RIFT variants & semantic extension											
RP+FI	Evidence fusion	✓	✓	✗	✗	✗	0.81	0.108	0.82	0.42	4.1×
RIFT w/o Flow	Posterior field	✓	✓	✗	✓	✗	0.75	0.132	0.78	0.39	4.8×
RIFT w/o Stability	Posterior field	✓	✓	✓	✗	✗	0.82	0.101	0.83	0.43	5.1×
RIFT w/o RF-aware projection	Posterior field	✓	✓	✓	✓	✗	0.78	0.124	0.80	0.44	5.0×
RIFT (Semantic-free)	**Posterior reliance field**	✓	✓	✓	✓	✗	**0.86**	**0.083**	**0.87**	**0.56**	5.3×
RIFT+Sem (Reference-assisted)	Posterior reliance field	✓	✓	✓	✓	✓	0.87	0.081	0.88	0.58	5.5×

Table 1 reports the shared-subset comparison underlying Fig. 4G. Classical CAM and gradient methods provide preference evidence but do not directly model finite-intervention responses, downstream flow, or uncertainty. Ablation-CAM and related perturbation-style methods move closer to evidence of intervention, but still do not infer a posterior field over the representation flow. Recent methods such as CAPE and TokenTM improve over classical attribution baselines, but they remain specialized explanation mechanisms rather than a unified diagnostic framework. In contrast, RIFT jointly incorporates preference, finite-intervention response, representation flow, stability, and optional semantic reference, achieving the strongest results among the evaluated compatible methods on the shared benchmark subset.

Limitations RIFT has three principal limitations. First, its conclusions are conditional on the intervention baseline and replacement operator. Second, flow-based diagnostics depend on representation-graph construction. Third, sparse probing may miss small or weakly sampled hidden-reliance regions. RIFT addresses these limitations through stability evidence, graph ablations, and sparse-versus-dense evaluations, but does not eliminate them. Accordingly, RIFT should be used as a diagnostic framework for evidence-conditioned reliance analysis rather than as a universal ground-truth explanation method.

5 Conclusion

We introduced RIFT, a diagnostic framework for visual explanations based on Reliance Uncertainty Fields (RUF). This paper delivers four primary contributions: (1) a four-way visual importance decomposition establishing **Theorem 1** (Non-Identifiability from finite-order local preference alone); (2) an Energy-Based Factor Graph / MRF formulation coupling local evidence with pairwise representation flow; (3) six explicit diagnostic readouts isolating reliance, uncertainty, false highlights, hidden reliance, flow attenuation, and shortcut candidates; and (4) a multi-domain diagnostic evaluation suite spanning classification, corruption, dense prediction, medical imaging, and tracking. RIFT shifts the paradigm of XAI from deterministic heatmap rendering to probabilistic reliance inference under explicit uncertainty.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4190–4197, 2020.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. *arXiv preprint arXiv:1807.08024*, 2018.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 782–791, 2021.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8188–8197, 2020.

Townim Faisal Chowdhury, Kewen Liao, Vu Minh Hieu Phan, Minh-Son To, Yutong Xie, Kevin Hung, David Ross, Anton Van Den Hengel, Johan W Verjans, and Zhibin Liao. Cape: Cam as a probabilistic ensemble for enhanced dnn interpretation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11072–11081, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020.

Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5374–5383, 2019.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2950–2958, 2019.

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

Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312*, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5356–5364, 2019.

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

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

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

Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022.

Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE transactions on image processing*, 30:5875–5888, 2021.

Felix Kares, Timo Speith, Hanwei Zhang, and Markus Langer. What makes for a good saliency map? comparing strategies for evaluating saliency maps in explainable ai (xai). *arXiv preprint arXiv:2504.17023*, 2025.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, explaining and visualizing deep learning*, pp. 267–280. Springer, 2019.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 300–317, 2018.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pp. 4990–4999, 2017.

Weili Nie, Yang Zhang, and Ankit Patel. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *International conference on machine learning*, pp. 3809–3818. Pmlr, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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

Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 983–991, 2020.

Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):34, 2021.

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

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

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

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8430–8439, 2019.

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

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

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

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

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

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25, 2020.

Junyi Wu, Bin Duan, Weitai Kang, Hao Tang, and Yan Yan. Token transformation matters: Towards faithful post-hoc explanation for vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10926–10935, 2024.

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*, 32, 2019.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

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

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 9623–9633, 2022.

A Introduction

This appendix provides an expanded conceptual motivation for RIFT. While the main manuscript introduces the central framework, this section develops the underlying motivation in greater detail. In particular, we clarify why visual highlightedness, functional reliance, representation flow, and semantic relevance should not be treated as interchangeable notions of importance. This expanded discussion also explains why RIFT models explainability as posterior inference over latent functional reliance rather than as a direct evaluation of a single attribution map.

A.1 VISUAL EXPLANATIONS DO NOT DIRECTLY SPECIFY FUNCTIONAL RELIANCE

Visual explanation methods are commonly used as if they answer a simple question: which image regions did the model use for its prediction? In practice, however, most explanation methods answer a narrower question determined by their underlying signal. Gradient-based methods identify regions of local sensitivitySimonyan et al. (2013); Sundararajan et al. (2017), feature-map-based methods identify spatially prominent or class-discriminative representationsZhou et al. (2016); Selvaraju et al. (2017); Wang et al. (2020); Ramaswamy et al. (2020), and perturbation-based methods measure output changes under a chosen masking or replacement schemeFong & Vedaldi (2017); Petsiuk et al. (2018). Each signal provides useful evidence about model behavior, but none of them alone uniquely determines the evidence on which the model functionally relies.

This distinction matters because visual explanations are often evaluated by apparent plausibility. A heatmap that overlaps with an object, lesion, part, or annotated region may appear faithful, but such overlap does not establish that the highlighted region causally supports the prediction. The highlighted region may reflect image edges, texture statistics, architectural biases, class-discriminative activation, or properties of the explanation method itself. Conversely, a model may rely on regions that are weakly highlighted or visually unintuitive. Thus, the central failure mode is not merely that some heatmaps are noisy or low-resolution, but that the quantity being visualized is often underspecified.

A.2 FOUR NOTIONS OF VISUAL IMPORTANCE

To make this ambiguity explicit, we distinguish four notions of visual importance. *Model-perceived importance* describes what a model-derived explanatory signal emphasizes, such as gradients, activations, attention weights, or class-discriminative responses. *Learning-influential importance* refers to factors that shape the learned decision rule during training. *Prediction-influential importance* denotes evidence that functionally affects a particular test-time output under intervention. *Semantic importance* refers to regions, concepts, or structures that are meaningful with respect to a human or task-level reference.

These notions may align in simple cases, but their disagreement is often the most informative evidence regarding the reliability of explanations. A region can be visually emphasized without being functionally used. A region can functionally support the prediction while receiving a low attribution score. A semantically plausible explanation can still hide reliance on contextual or spurious cues. Therefore, explanation reliability should not be understood as an intrinsic property of a heatmap alone, but as a relationship among multiple evidence sources.

A.3 FAILURE MODES INDUCED BY CONFLATING IMPORTANCE

The above decomposition reveals several common failure modes. A region with high model-perceived importance but low prediction-influential importance corresponds to *highlighted but unused* evidence. A region with low model-perceived importance but high prediction-influential importance corresponds to *hidden reliance*. A region with high prediction-influential importance outside a semantic reference may indicate *shortcut reliance*. A region with unstable intervention response or strong disagreement among explanation signals should be treated as uncertain, even if it appears visually plausible.

These cases motivate treating explanation outputs as observations rather than conclusions. An attribution map should not be interpreted as a direct measurement of functional reliance. Instead, it should be treated as preference evidence, and its reliability assessed using additional sources of information, such as intervention response, representation flow, measurement stability, and optional semantic reference.

A.4 RELIANCE EVOLVES THROUGH REPRESENTATION FLOW

A further complication is that reliance is not static across the network. Deep visual models transform evidence through a sequence of intermediate representations. A signal that appears strong in an early or middle layer may be amplified, suppressed, redistributed, recombined, or ignored downstream.

Likewise, a signal that is weak at the input level may become functionally important only after interaction with later representations.

This representation-level evolution is difficult to capture with explanations computed at a single layer or projected directly to image space. For example, a feature-map explanation may highlight a spatial region in a selected layer, but it does not show whether the corresponding evidence survives downstream computation. Similarly, an input-level attribution map may indicate local sensitivity without revealing how that sensitivity is transformed across internal representations. A reliability framework should therefore model not only where evidence appears, but also how evidence propagates, attenuates, or re-emerges across layers.

A.5 From explanation maps to Reliance Uncertainty Fields

RIFT addresses these issues by reframing the reliability of visual explanations as inference over latent functional reliance. For each representation unit, RIFT introduces a latent reliance variable indicating whether that unit functionally supports the target prediction. Explanation maps, intervention responses, representation flow, stability estimates, and semantic references are then treated as noisy evidence about this latent variable.

The resulting object is the *Reliance Uncertainty Field* (RUF), a posterior field over a layer-wise representation graph. Each node corresponds to a representation unit, and each edge captures downstream reachability or empirically measured propagation. Node evidence captures local preferences, functional influences, stability, and optional semantic alignment, whereas flow-based coupling indicates whether reliance is supported by downstream representations. This posterior formulation allows RIFT to estimate not only reliance, but also uncertainty, disagreement, flow attenuation, hidden reliance, highlighted-but-unused evidence, and shortcut reliance.

Importantly, the goal of RIFT is not to replace existing explanation methods with another heatmap. Rather, RIFT asks when and where explanation signals should be trusted as evidence of functional reliance. This changes the role of attribution maps: they become one evidence source in a broader reliability inference problem.

A.6 Organization of the appendix

The remainder of the appendix supports this formulation. Appendix B situates RIFT relative to gradient-based, perturbation-based, feature-map-based, reliability-oriented, and semantic explanation methods. Appendix C provides additional background and formal statements showing why preference evidence alone cannot identify functional reliance. Appendix D gives implementation details for evidence normalization, intervention design, flow estimation, posterior inference, projection to image space, sparse probing, and diagnostic scores. Appendix E discusses assumptions, limitations, and practical scope.

Together, these sections expand the main manuscript's thesis: visual explanation should be treated not as static heatmap visualization, but as inference over functional reliance under uncertainty.

B Related Work

This appendix provides additional discussion of prior work related to RIFT. We focus on visual explanation methods that produce spatial attribution or localization maps, faithfulness evaluation protocols, and semantic analyses of model reliance. Our goal is not to provide an exhaustive survey, but to clarify how existing methods relate to the evidence sources used in our framework.

B.1 Gradient-based visual explanations

Gradient-based explanation methods estimate visual importance by measuring the sensitivity of a model's output with respect to the input. Representative methods include Saliency MapsSimonyan et al. (2013), Guided BackpropagationSpringenberg et al. (2014), Integrated GradientsSundararajan et al. (2017), and SmoothGradSmilkov et al. (2017). Although these methods differ in how gradients are modified, accumulated, or averaged, they share the basic idea that locally sensitive input dimensions are informative for explaining a prediction.

A general gradient-based attribution can be written as

$$S^{\mathrm{grad}}(x, c) = \phi\left(\frac{\partial f_c(x)}{\partial x}\right), \tag{9}$$

where $f_c(x)$ denotes the class-specific score and $\phi(\cdot)$ may represent absolute value, path integration, smoothing, or noise averaging. Saliency MapsSimonyan et al. (2013), for example, use

$$S_i^{\mathrm{sal}}(x) = \left|\frac{\partial f_c(x)}{\partial x_i}\right|. \tag{10}$$

This score measures local differential sensitivity. However, local sensitivity does not necessarily imply a large output change under finite intervention. The first-order approximation

$$f_c(x + \delta) \approx f_c(x) + \nabla_x f_c(x)^\top \delta \tag{11}$$

is valid only for sufficiently small perturbations. Thus, high gradient magnitude may not correspond to prediction-level reliance under realistic masking, replacement, or spatial transformation. This gap between sensitivity and fidelity has been analyzed in prior workYeh et al. (2019); Kindermans et al. (2019).

Guided Backpropagation modifies the backward pass by suppressing negative gradientsSpringenberg et al. (2014). Although this often produces visually sharp maps, the modified backward rule no longer corresponds to the true model gradient. Sanity checks have shown that some saliency maps can remain visually plausible even when model parameters or labels are randomized, suggesting that they may capture input structure rather than model-dependent decision evidenceAdebayo et al. (2018); Nie et al. (2018). Integrated Gradients addresses saturation by accumulating gradients along a path from a baseline to the inputSundararajan et al. (2017), but the resulting attribution remains sensitive to baseline and path choicesKindermans et al. (2019). SmoothGrad averages gradients over noisy inputs to reduce visual noiseSmilkov et al. (2017), but it still measures local sensitivity rather than direct functional reliance.

In RIFT, gradient-based explanations are therefore treated as *preference evidence*. They indicate where a local differential signal is strong, but they do not by themselves identify latent functional reliance.

B.2 Perturbation-based explanations and faithfulness evaluation

Perturbation-based methods estimate importance by modifying parts of the input and measuring the resulting change in model output. Early occlusion-style analyses replace or mask image patches and observe score variationZeiler & Fergus (2014). Later methods optimize deletion or preservation masksFong & Vedaldi (2017), sample random masks to estimate pixel importancePetsiuk et al. (2018), or evaluate explanations through deletion, insertion, and retraining-based protocolsSamek et al. (2016); Hooker et al. (2019).

A general perturbation score can be written as

$$S^{\mathrm{pert}}(r, c) = \Delta\left(f_c(x), f_c(x \odot (1 - m_r) + b \odot m_r)\right), \tag{12}$$

where m_r is a mask for region r, b is a baseline or replacement value, and Δ measures a change in logit, probability, margin, loss, or output distribution. Compared with gradient-based explanations, perturbation methods are closer to functional influence because they evaluate finite changes to the inputFong & Vedaldi (2017); Petsiuk et al. (2018); Samek et al. (2016).

However, perturbation estimates can depend strongly on mask design, mask granularity, baseline choice, replacement distribution, and regularizationFong & Vedaldi (2017); Chang et al. (2018); Hooker et al. (2019). Perturbations can also produce out-of-distribution inputs, making the observed output change partly reflect distribution shift rather than genuine feature relianceChang et al. (2018); Hooker et al. (2019). In addition, optimized masks or dense random-mask sampling can be computationally expensiveFong & Vedaldi (2017); Petsiuk et al. (2018). Faithfulness metrics such as deletion and insertion are useful for evaluating whether high-scoring regions affect the outputSamek et al. (2016); Hooker et al. (2019), but they often summarize explanation quality as a final-output response and do not explicitly model how evidence is transformed across intermediate layers.

RIFT uses perturbation responses as *intervention evidence* about latent reliance. Unlike standalone perturbation maps or deletion scores, RIFT combines intervention evidence with representation preference, flow, stability, and optional semantic reference to infer a posterior reliance field.

B.3 FEATURE-MAP-BASED LOCALIZATION METHODS

Feature-map-based methods construct spatial explanations from intermediate activations, especially in convolutional neural networks. A common formulation is

$$M_c(x) = \sum_k w_k^c(x) A^k(x), \tag{13}$$

where $A^k(x)$ is the k-th feature map and $w_k^c(x)$ is a class-specific weight. Since $M_c(x)$ is usually lower-resolution than the input, it is projected back to image space:

$$\widehat{M_c}(x) = \mathcal{R}(M_c(x)). \tag{14}$$

These maps are often interpreted as localizing the evidence used by the model. However, they are derived from feature-space aggregation followed by projection, and therefore primarily reflect representation prominence rather than direct input-level functional reliance.

Class Activation Mapping (CAM) computes class-specific localization by applying classifier weights to feature maps before global average poolingZhou et al. (2016):

$$M_c(x) = \sum_k w_k^c A^k(x). \tag{15}$$

This assumes that each feature map contributes to the class score through a spatially uniform weight and is limited to architectures with compatible pooling and classifier structureZhou et al. (2016). Grad-CAM generalizes CAM by using gradients of the class score with respect to feature mapsSelvaraju et al. (2017):

$$w_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial f_c(x)}{\partial A_{ij}^k}. \tag{16}$$

Although Grad-CAM applies to a broader range of architectures, spatially averaging gradients into a single channel weight can discard fine-grained information about where the feature map affects the output. This can yield visually plausible but not fully faithful mapsDraelos & Carin (2020).

Grad-CAM++ incorporates higher-order gradient terms to better handle multiple object instancesChattopadhay et al. (2018), while XGrad-CAM normalizes activation–gradient interactionsFu et al. (2020):

$$w_k^c = \sum_{i,j} \alpha_{ij}^k \frac{\partial f_c(x)}{\partial A_{ij}^k}, \tag{17}$$

$$w_k^c = \sum_{i,j} \frac{A_{ij}^k}{\sum_{i,j} A_{ij}^k} \frac{\partial f_c(x)}{\partial A_{ij}^k}. \tag{18}$$

These variants improve localization or consistency in specific settingsChattopadhay et al. (2018); Fu et al. (2020), but they still operate through feature-space weighting.

Score-CAM avoids gradients by using forward passes to estimate feature-map importanceWang et al. (2020):

$$w_k^c = f_c(x \odot \mathrm{mask}(A^k)). \tag{19}$$

Ablation-CAM removes individual feature maps and measures output changeRamaswamy et al. (2020):

$$w_k^c = f_c(x) - f_c(x; A^k \leftarrow 0). \tag{20}$$

These methods move closer to intervention-based reasoning by evaluating changes in predictionsWang et al. (2020); Ramaswamy et al. (2020). Nevertheless, the intervention is performed at the feature-map level or through feature-derived masks. Since feature maps do not correspond one-to-one with localized input regions, the resulting attribution remains an indirect approximation of input-level influence.

HiResCAM preserves spatial gradient–activation interactions by replacing Grad-CAM's spatial gradient averaging with element-wise activation–gradient multiplicationDraelos & Carin (2020):

$$M_c(x) = \sum_k A^k(x) \odot \frac{\partial f_c(x)}{\partial A^k(x)}. \tag{21}$$

This can improve faithfulness under certain architectural assumptionsDraelos & Carin (2020). However, the attribution is still computed in feature space and must be projected to the input domain.

RIFT treats feature-map-based explanations as evidence about *representation preference*. Rather than assuming that a selected-layer heatmap directly identifies reliance, RIFT asks whether the highlighted representation units are functionally supported by interventions and whether their evidence survives downstream flow.

B.4 Sanity checks and explanation reliability

Several studies have questioned whether visually plausible explanations faithfully reflect model behavior. Sanity checks for saliency maps show that some explanation methods are insensitive to model parameters or training labelsAdebayo et al. (2018). Other work has studied explanation infidelity and sensitivity, showing that explanation quality depends strongly on the evaluation criterionYeh et al. (2019); Tomsett et al. (2020). Benchmarking work further shows that explanations can behave differently under deletion, insertion, retraining, and perturbation-based protocolsHooker et al. (2019); Samek et al. (2016).

These findings motivate a central premise of RIFT: explanation reliability should not be reduced to the visual plausibility of a heatmap. A heatmap can be sharp, class-discriminative, or semantically plausible, yet fail to capture functional reliance. Conversely, a model may rely on evidence that is weakly highlighted. RIFT formalizes such cases by accounting for posterior uncertainty and disagreement among evidence sources.

B.5 Semantic reference and shortcut reliance

Semantic interpretability studies examine whether learned representations align with human-interpretable concepts, object parts, or task-relevant regions. Network Dissection quantifies the interpretability of hidden units by measuring alignment with semantic conceptsBau et al. (2017). TCAV evaluates whether user-defined concepts influence model predictions through concept activation vectorsKim et al. (2018). Such methods highlight the importance of semantic reference in understanding model behavior.

At the same time, models may rely on contextual or non-semantic shortcuts. Clever Hans predictors and shortcut learning show that high predictive performance can be supported by spurious cues rather than task-relevant evidenceLapuschkin et al. (2019); Geirhos et al. (2020). In visual explanation, this creates an important distinction: a model's functional reliance may be high outside human-annotated semantic regions.

RIFT uses semantic annotations only as optional external references. We do not assume that semantic masks define the ground-truth importance of models. Instead, RIFT compares posterior reliance with semantic reference to diagnose semantic reliance and shortcut reliance.

B.6 Positioning of RIFT

The methods above provide complementary signals about model behavior. Gradient-based methods primarily measure local sensitivitySimonyan et al. (2013); Sundararajan et al. (2017); Smilkov et al. (2017); perturbation-based methods estimate output response under finite interventionsFong & Vedaldi (2017); Petsiuk et al. (2018); Samek et al. (2016); Hooker et al. (2019); feature-map-based methods visualize class-discriminative representation prominenceZhou et al. (2016); Selvaraju et al. (2017); Wang et al. (2020); Ramaswamy et al. (2020); and semantic methods provide external references for human-meaningful conceptsBau et al. (2017); Kim et al. (2018). These signals are often interpreted as revealing the same object: the evidence used by the model. However, sensitivity, intervention effect, representation preference, and semantic relevance need not coincideAdebayo et al. (2018); Yeh et al. (2019); Hooker et al. (2019); Tomsett et al. (2020).

RIFT reframes these signals as noisy evidence about latent functional reliance. The Reliance Uncertainty Field represents reliance as a posterior field over the model's representation graph. Preference evidence captures what an explanation appears to emphasize; intervention evidence captures prediction-level influence; flow evidence captures how signals propagate across layers; stability evidence captures measurement uncertainty; and semantic reference supports shortcut diagnostics. This

posterior formulation enables RIFT to diagnose highlighted but unused evidence, hidden reliance, uncertain reliance, flow attenuation, redistribution, and semantically spurious reliance.

C Detailed Background and Proofs

This appendix provides additional background on existing visual explanation signals and gives formal support for the non-identifiability statement in Section 2. The goal is not to provide an exhaustive survey, but to clarify how existing explanation families relate to the evidence sources used by RIFT.

C.1 Detailed background on visual explanation signals

Existing visual explanation methods are commonly interpreted as identifying the evidence used by a model. However, different explanation families measure different quantities. This motivates our treatment of existing explanation outputs as noisy evidence about latent functional reliance rather than as direct measurements of reliance.

C.1.1 Gradient-based methods

Gradient-based methods identify important input features by measuring the sensitivity of a model output with respect to the input. Representative methods include Saliency MapsSimonyan et al. (2013), Guided BackpropagationSpringenberg et al. (2014), Integrated GradientsSundararajan et al. (2017), and SmoothGradSmilkov et al. (2017). Although these methods differ in how gradients are transformed, accumulated, or averaged, they share the principle that visual importance can be approximated by local differential sensitivity.

A general gradient-based attribution can be written as

$$S^{\mathrm{grad}}(x, c) = \phi\left(\frac{\partial f_c(x)}{\partial x}\right), \tag{22}$$

where $f_c(x)$ denotes the class-specific score and $\phi(\cdot)$ represents a transformation such as absolute value, path integration, smoothing, or noise averaging. This formulation measures how locally sensitive the output is to infinitesimal input perturbations.

For example, Saliency MapsSimonyan et al. (2013) define importance as

$$S_i^{\mathrm{sal}}(x) = \left|\frac{\partial f_c(x)}{\partial x_i}\right|. \tag{23}$$

This score captures local variation in the model response. However, it does not necessarily determine the effect of a finite intervention. The first-order approximation

$$f_c(x + \delta) \approx f_c(x) + \nabla_x f_c(x)^\top \delta \tag{24}$$

is valid only locally and for sufficiently small perturbations. Thus, a large gradient magnitude does not guarantee a large prediction change under realistic masking, replacement, or spatial displacement. This gap between local sensitivity and explanation fidelity has been discussed in prior workYeh et al. (2019).

Guided BackpropagationSpringenberg et al. (2014) modifies the backward pass by suppressing negative gradients:

$$R_l = R_{l+1} \cdot \mathbf{1}(a_l > 0) \cdot \mathbf{1}(R_{l+1} > 0). \tag{25}$$

This often produces visually sharp maps, but the modified backward rule breaks correspondence with the true model gradient. Such maps can remain visually plausible even when model parameters are randomized, suggesting that they may reflect input structure rather than model-dependent decision evidenceAdebayo et al. (2018); Nie et al. (2018).

Integrated GradientsSundararajan et al. (2017) accumulates gradients along a path from a baseline x' to the input:

$$S_i^{\mathrm{IG}}(x) = (x_i - x_i') \int_0^1 \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha. \tag{26}$$

This improves over raw gradients by integrating local sensitivities along a path and satisfies useful axiomatic properties. Nevertheless, the attribution depends on the baseline and path, and it does not directly model whether a feature is necessary for a prediction under interventionKindermans et al. (2019).

SmoothGradSmilkov et al. (2017) reduces visual noise by averaging gradients over noisy input samples:

$$S^{\mathrm{SG}}(x) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2 I)}\left[|\nabla_x f(x+\epsilon)|\right]. \tag{27}$$

This can improve the visual stability of saliency maps, but it still measures local sensitivity rather than prediction-level reliance.

Taken together, gradient-based explanations provide useful sensitivity evidence. In our formulation, they are best viewed as preference evidence: they indicate where a local differential signal is strong, but they do not directly identify latent functional reliance.

C.1.2 Perturbation-based methods

Perturbation-based methods estimate importance by modifying parts of the input and measuring the resulting change in the model output. Early occlusion approaches mask or replace image patches and observe score variationZeiler & Fergus (2014). Later methods search for minimal deletion or preservation masksFong & Vedaldi (2017), sample random masks to estimate pixel importancePetsiuk et al. (2018), or evaluate attribution faithfulness through deletion, insertion, and retraining-based protocolsSamek et al. (2016); Hooker et al. (2019).

A general perturbation score can be written as

$$S^{\mathrm{pert}}(r,c) = \Delta\left(f_c(x), f_c(x\odot(1-m_r)+b\odot m_r)\right), \tag{28}$$

where m_r is a mask for region r, b is a baseline or replacement value, and Δ measures a change in logit, probability, margin, loss, or output distribution. Compared with gradient methods, perturbation methods are closer to functional influence because they evaluate finite interventionsFong & Vedaldi (2017); Petsiuk et al. (2018); Samek et al. (2016).

However, perturbation estimates depend on the mask shape, mask granularity, baseline choice, replacement distribution, and regularization strategyFong & Vedaldi (2017); Chang et al. (2018); Hooker et al. (2019). Perturbed images may also fall off the natural image manifold, so the observed score change may partly reflect distribution shift rather than genuine relianceChang et al. (2018); Hooker et al. (2019). In addition, repeated optimization or dense sampling can be computationally expensiveFong & Vedaldi (2017); Petsiuk et al. (2018).

Most importantly for our setting, perturbation methods typically summarize final output responses. They do not by themselves explain how evidence is transformed, suppressed, redistributed, or compensated across intermediate representations. RIFT therefore uses intervention responses as one evidence source for latent reliance rather than treating a perturbation score as a complete explanation.

C.1.3 Feature-map-based methods

Feature-map-based methods construct spatial explanations from intermediate activations, especially in convolutional neural networks. A common form is

$$M_c(x) = \sum_k w_k^c(x) A^k(x), \tag{29}$$

where $A^k(x)$ is the k-th feature map and $w_k^c(x)$ is a class-specific weight. Since $M_c(x)$ is usually defined at a lower spatial resolution than the input, it is projected to image space by an upsampling operator:

$$\widehat{M}_c(x) = \mathcal{R}(M_c(x)). \tag{30}$$

These maps are often interpreted as showing image regions that contribute to the prediction. However, they are derived from feature-space aggregation followed by projection, so they primarily reflect representation prominence rather than direct input-level functional reliance.

CAMZhou et al. (2016) computes localization using classifier weights:

$$M_c(x) = \sum_k w_k^c A^k(x). \tag{31}$$

This assumes that each feature map contributes to the class score through a spatially uniform weight. Grad-CAMSelvaraju et al. (2017) replaces classifier weights with spatially averaged gradients:

$$w_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial f_c(x)}{\partial A_{ij}^k}. \tag{32}$$

Although Grad-CAM applies to a broader range of architectures, spatially averaging gradients into a single channel weight can discard fine-grained information about where a feature affects the output. As a result, the heatmap may be visually plausible but not fully faithful to individual spatial contributionsDraelos & Carin (2020).

Grad-CAM++Chattopadhay et al. (2018) and XGrad-CAMFu et al. (2020) refine the weighting scheme using higher-order gradients or activation-gradient normalization:

$$w_k^c = \sum_{i,j} \alpha_{ij}^k \frac{\partial f_c(x)}{\partial A_{ij}^k}, \tag{33}$$

$$w_k^c = \sum_{i,j} \frac{A_{ij}^k}{\sum_{i,j} A_{ij}^k} \frac{\partial f_c(x)}{\partial A_{ij}^k}. \tag{34}$$

These methods improve localization or consistency in particular settings, but they still rely on feature-space weighting. Thus, feature-space prominence or sensitivity need not correspond to input-level causal influence or semantic relevance.

Score-CAMWang et al. (2020) avoids gradients by using forward passes:

$$w_k^c = f_c(x \odot \mathrm{mask}(A^k)). \tag{35}$$

Ablation-CAMRamaswamy et al. (2020) estimates importance by removing feature maps:

$$w_k^c = f_c(x) - f_c(x; A^k \leftarrow 0). \tag{36}$$

These methods are closer to intervention-based reasoning because they evaluate prediction changes. Still, the interventions operate at the feature-map level or through feature-derived masks, and feature maps do not correspond one-to-one to localized input regions.

HiResCAMDraelos & Carin (2020) preserves spatial gradient-activation interactions:

$$M_c(x) = \sum_k A^k(x) \odot \frac{\partial f_c(x)}{\partial A^k(x)}. \tag{37}$$

This can improve faithfulness under certain architectural assumptions. However, it still computes attribution in feature space and requires projection to the input domain.

Overall, feature-map-based methods are useful for visualizing class-discriminative representation preference. They do not directly characterize how input evidence functionally supports a prediction, nor do they model how reliance emerges, attenuates, or redistributes across layers.

C.2 Formal statements

We now state and prove formal claims supporting the preliminary discussion in Section 2.

C.2.1 Preference does not identify reliance

Proof of Theorem 1. We construct a counterexample in which preference at unit i is identical across two functions, yet the finite-intervention response at the *same* unit i diverges arbitrarily.

Let $u \in \mathbb{R}$ represent the scalar activation of unit i, evaluated at point u_0, with intervention baseline $b \neq u_0$. Define two differentiable functions:

$$f_1(u) = u, \qquad f_2(u) = u + \alpha(u - u_0)^2, \tag{38}$$

where $\alpha > 0$ is a free parameter.

Step 1: Identical local preference. The local derivatives at the evaluation point u_0 are:

$$f_1'(u_0) = 1, \qquad f_2'(u_0) = 1 + 2\alpha(u_0 - u_0) = 1. \tag{39}$$

Similarly, $f_1(u_0) = u_0$ and $f_2(u_0) = u_0$. More generally, for any k-jet-based preference functional depending on $f(u_0), f'(u_0), \ldots, f^{(k)}(u_0)$, one can choose $f_2(u) = u + \alpha(u - u_0)^{k+1}$ so that all derivatives up to order k match at u_0 while f_2 differs from f_1 away from u_0. For simplicity, we present the first-order case ($k = 1$).

Thus any attribution functional $A_E(i)$ determined by the local 1-jet satisfies $A_{E,1}(i) = A_{E,2}(i)$.

Step 2: Divergent intervention response. Now intervene on unit i by replacing u_0 with b. The finite-intervention responses are:

$$\Delta_{i,1} = f_1(u_0) - f_1(b) = u_0 - b, \tag{40}$$

$$\begin{aligned}\Delta_{i,2} &= f_2(u_0) - f_2(b) = (u_0 - b) - \alpha(b - u_0)^2 \\ &= (u_0 - b)(1 + \alpha(u_0 - b)).\end{aligned} \tag{41}$$

The difference is $|\Delta_{i,2} - \Delta_{i,1}| = \alpha(u_0 - b)^2$, which can be made arbitrarily large by choosing α sufficiently large. Since the preference scores at unit i are identical while the intervention response at the *same* unit i can be made to differ by any prescribed $C > 0$, functional reliance R_i^c is non-identifiable from local preference evidence alone. □

C.2.2 Local sensitivity and finite intervention can diverge

The previous proof uses two functions. We can also show that, even for a single function, local sensitivity at a point need not predict finite intervention response.

Proposition 1 (Local sensitivity does not determine finite intervention)**.** *For a differentiable non-linear function f, the local derivative $\nabla f(x)$ at a point does not generally determine the finite-intervention response $f(x) - f(x')$ for a baseline or intervened point x'.*

Proof. By the fundamental theorem of calculus along the line segment from x' to x,

$$f(x) - f(x') = \int_0^1 \nabla f(x' + t(x - x'))^\top (x - x')\, dt. \tag{42}$$

The finite response depends on gradients along the entire path from x' to x, not only on $\nabla f(x)$. Therefore, two functions can share the same local derivative at x but differ in their derivatives along the intervention path, yielding different finite responses. Hence, local sensitivity at the original point does not determine finite-intervention response. □

C.2.3 Feature-space maps do not uniquely define image-space reliance

Feature-map-based explanations require a projection from the representation space to the image space. The following proposition formalizes why this projection is not unique without additional assumptions.

Proposition 2 (Projection ambiguity)**.** *Let $M \in \mathbb{R}^{H_\ell \times W_\ell}$ be a feature-space explanation map and let $\mathcal{R}$ be an upsampling operator producing an image-space map $\widehat{M} = \mathcal{R}(M)$. In general, $\widehat{M}$ is not uniquely determined by functional image-space reliance. Different image-space contribution fields can induce the same low-resolution feature-space map under aggregation.*

Proof. Let Ω_x be the image domain and let $\{K_i(p)\}_{i \in V}$ be projection or receptive-field kernels from image pixels to feature nodes. Suppose a feature-space score at node i is produced by aggregation

$$M_i = \sum_{p \in \Omega_x} K_i(p) C(p), \tag{43}$$

where $C(p)$ is an image-space contribution field. This is a linear map from C to M. Since the image domain usually has much higher dimension than the feature grid, the linear map has a nontrivial null space. Therefore, there exist distinct contribution fields $C_1 \neq C_2$ such that

$$\sum_p K_i(p) C_1(p) = \sum_p K_i(p) C_2(p) \qquad \text{for all } i. \tag{44}$$

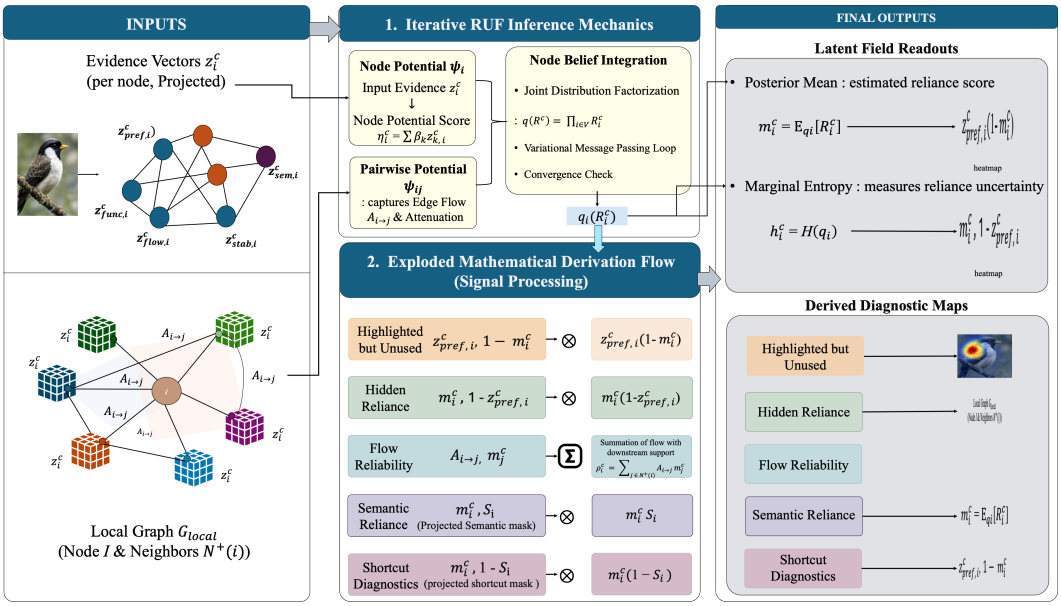

Figure 6: **Detailed RUF inference and diagnostic construction.** Projected node evidence and intervention-induced flow define the RUF posterior potentials. Variational message passing yields node-wise reliance and uncertainty, which are combined with preference, flow, and optional semantic references to produce the diagnostic maps.

Thus, the same feature-space map can correspond to multiple image-space contribution fields. A naive upsampling operator $\mathcal{R}$ selects one visualization but does not identify the true image-space reliance. □

C.3 Implications for RIFT

The above discussion supports three design choices in RIFT. First, attribution maps are treated as preference evidence rather than direct reliance estimates. Second, finite interventions are incorporated as prediction-influential evidence, but they are not used alone because they depend on baselines, masks, and replacement distributions. Third, representation flow is modeled explicitly because neither input-level perturbation nor selected-layer feature attribution explains how evidence propagates, attenuates, or redistributes across intermediate layers.

RIFT therefore infers a posterior Reliance Uncertainty Field over the representation graph by combining preference evidence, intervention response, flow, stability, and optional semantic reference. This posterior view allows explanation reliability to be diagnosed through uncertainty, disagreement, hidden reliance, highlighted-but-unused evidence, and shortcut reliance.

D Additional Method Details

This appendix provides additional details for RIFT, including evidence construction, intervention choices, flow estimation, posterior inference, image-space projection, sparse probing, and diagnostic scores. Figure 6 summarizes how these components interact. Node-level evidence is combined through local potentials, while intervention-induced flow defines pairwise couplings over the representation graph. Iterative variational inference then produces node-wise posterior marginals, from which reliance, uncertainty, and the derived diagnostic maps are computed.

D.1 EVIDENCE NORMALIZATION

All evidence scores are normalized per image and target class before being used in the posterior. For a nonnegative score z_i, we use

$$\widehat{z}_i = \frac{z_i - \min_{j \in V} z_j}{\max_{j \in V} z_j - \min_{j \in V} z_j + \epsilon}. \tag{45}$$

For signed quantities, such as logit change, we use either the positive drop or the absolute change depending on the diagnostic target. Unless otherwise stated, class-specific reliance uses positive logit drop:

$$\Delta_{\text{drop}} = \max\{0, s_c(x) - \widetilde{s}_c(x)\}. \tag{46}$$

All normalized evidence values lie in $[0, 1]$.

D.2 PREFERENCE EVIDENCE

Preference evidence represents what an explanation method appears to highlight. Given an explanation method E, let $A_E^c(x)$ denote its class-specific explanation map.

If $A_E^c(x)$ is already defined on the same representation grid as node $i = (\ell, u)$, we set

$$z^c_{\text{pref},i} = \text{Norm}\left(A_E^c(x; i)\right). \tag{47}$$

If $A_E^c(x)$ is defined in image space, we aggregate it to the representation node using a projection kernel $K_i(p)$:

$$z^c_{\text{pref},i} = \text{Norm}\left(\sum_{p \in \Omega_x} K_i(p) A_E^c(x; p)\right). \tag{48}$$

Here Ω_x denotes the image coordinate domain, and $K_i(p)$ represents the receptive-field or effective-receptive-field contribution of pixel p to representation node i.

When multiple explanation methods $\{E_k\}_{k=1}^K$ are used, we define the mean preference evidence as

$$z^c_{\text{pref},i} = \frac{1}{K} \sum_{k=1}^{K} z^{c,k}_{\text{pref},i}, \tag{49}$$

and the preference disagreement as

$$v^c_{\text{pref},i} = \text{Var}_k\left[z^{c,k}_{\text{pref},i}\right]. \tag{50}$$

This disagreement is later used as part of the stability evidence.

D.3 FUNCTIONAL INTERVENTION EVIDENCE

Functional evidence measures whether the prediction changes when a representation unit is intervened upon. For node $i = (\ell, u)$, we replace the feature vector $F_u^{(\ell)}(x)$ with a baseline vector $b_i \sim q_i$:

$$\widetilde{F}_v^{(\ell)}(x; i \leftarrow b_i) = \begin{cases} b_i, & v = u, \\ F_v^{(\ell)}(x), & v \neq u. \end{cases} \tag{51}$$

The modified representation is passed through the downstream subnetwork:

$$\widetilde{s}_c(x; i \leftarrow b_i) = h_{\ell \to \text{out}, c}\left(\widetilde{F}^{(\ell)}(x; i \leftarrow b_i)\right). \tag{52}$$

The functional response is

$$\Delta_i^c(b_i) = \Delta\left(s_c(x), \widetilde{s}_c(x; i \leftarrow b_i)\right). \tag{53}$$

The functional evidence and its sampling variance are

$$z^c_{\text{func},i} = \mathbb{E}_{b_i \sim q_i}\left[\Delta_i^c(b_i)\right], \tag{54}$$

$$v^c_{\text{func},i} = \text{Var}_{b_i \sim q_i}\left[\Delta_i^c(b_i)\right]. \tag{55}$$

Baseline choices. We consider the following baseline distributions:

1. **Zero baseline**: $b_i = 0$.
2. **Dataset mean baseline**: $b_i = \mathbb{E}_{x'}[F_i(x')]$.
3. **Sampled feature baseline**: $b_i = F_i(x')$, where x' is sampled from the dataset.
4. **Local noise baseline**: $b_i = F_i(x) + \xi$, where ξ is sampled from a small isotropic distribution.

The default implementation uses the dataset mean or sampled feature baseline and reports variance across baseline samples.

Discrepancy functions. The discrepancy function Δ can be instantiated in several ways:

$$\Delta_{\text{logit}} = s_c(x) - \widetilde{s}_c(x; i \leftarrow b_i), \tag{56}$$

$$\Delta_{\text{abs}} = \left| s_c(x) - \widetilde{s}_c(x; i \leftarrow b_i) \right|, \tag{57}$$

$$\Delta_{\text{prob}} = p_c(x) - \widetilde{p}_c(x; i \leftarrow b_i), \tag{58}$$

$$\Delta_{\text{margin}} = \left(s_c(x) - \max_{k \neq c} s_k(x) \right) - \left(\widetilde{s}_c(x) - \max_{k \neq c} \widetilde{s}_k(x) \right), \tag{59}$$

$$\Delta_{\text{KL}} = D_{\text{KL}} \left(p(y \mid x) \,\|\, p(y \mid \widetilde{x}_i) \right). \tag{60}$$

Class-specific reliance uses logit drop or margin drop. Distribution-level reliability uses KL divergence.

D.4 Flow estimation

RIFT measures whether the effect of an upstream representation unit propagates to downstream representations. For edge $(i, j) \in E$, where $i = (\ell, u)$ and $j = (m, v)$ with $m > \ell$, we define intervention-induced flow as

$$\Phi^c_{i \to j}(x) = \mathbb{E}_{b_i \sim q_i} \left[d_F \left(F_j(x), \widetilde{F}_j(x; i \leftarrow b_i) \right) \right], \tag{61}$$

where

$$d_F(a, b) = \frac{\|a - b\|_2}{\|a\|_2 + \epsilon}. \tag{62}$$

The normalized outgoing flow from node i is

$$A_{i \to j} = \frac{\Phi^c_{i \to j}}{\sum_{j' \in \mathcal{N}^+(i)} \Phi^c_{i \to j'} + \epsilon}. \tag{63}$$

Here $\mathcal{N}^+(i)$ is the downstream neighborhood of i.

We also define the outgoing flow mass

$$M_{\text{flow}}(i) = \sum_{j \in \mathcal{N}^+(i)} \Phi^c_{i \to j}, \tag{64}$$

and the flow entropy

$$H_{\text{flow}}(i) = - \sum_{j \in \mathcal{N}^+(i)} A_{i \to j} \log(A_{i \to j} + \epsilon). \tag{65}$$

Low flow entropy indicates concentrated propagation, while high flow entropy indicates redistribution across downstream units.

D.5 NODE AND PAIRWISE POTENTIALS

The RUF posterior in the main text is

$$p(R^c \mid Z^c, G, x) = \frac{1}{\mathcal{Z}} \prod_{i \in V} \psi_i(R_i^c; Z_i^c) \prod_{(i,j) \in E} \psi_{ij}(R_i^c, R_j^c). \tag{66}$$

As illustrated in Figure 6, node potentials encode local evidence, while pairwise potentials couple reliance beliefs along intervention-induced flow edges.

Node potential. The node potential combines local evidence:

$$\begin{aligned} \psi_i(R_i^c; Z_i^c) = \exp\Big(& R_i^c \eta_i^c - \lambda_{\text{var}} R_i^c v_{\text{func},i}^c \\ & - \lambda_{\text{pref}} R_i^c v_{\text{pref},i}^c \Big), \end{aligned} \tag{67}$$

where

$$\eta_i^c = \beta_{\text{pref}} \widehat{z}_{\text{pref},i}^c + \beta_{\text{func}} \widehat{z}_{\text{func},i}^c + \beta_{\text{stab}} \widehat{z}_{\text{stab},i}^c + \beta_{\text{sem}} \widehat{z}_{\text{sem},i}^c. \tag{68}$$

The variance penalties reduce reliance confidence when intervention or preference evidence is unstable. If no semantic reference is available, we set $\beta_{\text{sem}} = 0$.

Flow pairwise potential. For a directed edge (i, j), we use

$$\begin{aligned} \psi_{ij}^{\text{flow}}(R_i^c, R_j^c) = \exp\Big(& \lambda_{\text{flow}} A_{i \to j} R_i^c R_j^c \\ & - \lambda_{\text{cut}} A_{i \to j} R_i^c (1 - R_j^c) \Big). \end{aligned} \tag{69}$$

The first term encourages reliance to propagate along strong flow edges. The second term captures attenuation or rejection when an upstream node appears relied upon but downstream nodes do not support that reliance.

Optional spatial smoothness. For neighboring nodes within the same layer, we may add a smoothness potential:

$$\psi_{ij}^{\text{spatial}}(R_i^c, R_j^c) = \exp\left(-\lambda_{\text{sp}} w_{ij} (R_i^c - R_j^c)^2\right), \tag{70}$$

where w_{ij} reflects spatial adjacency or feature similarity. This term is optional and is not used when evaluating the pure contribution of flow coupling.

D.6 MEAN-FIELD INFERENCE

Exact inference in the RUF posterior is generally intractable for dense graphs. We therefore use a factorized variational approximation:

$$q(R^c) = \prod_{i \in V} q_i(R_i^c). \tag{71}$$

In the implementation, each marginal is represented as a Bernoulli relaxation:

$$q_i(R_i^c = 1) = m_i^c, \qquad q_i(R_i^c = 0) = 1 - m_i^c. \tag{72}$$

The variational objective is

$$\mathcal{L}_{\text{MF}} = \mathbb{E}_q \left[\log p(R^c, Z^c \mid G, x)\right] - \mathbb{E}_q \left[\log q(R^c)\right]. \tag{73}$$

Equivalently, we minimize

$$D_{\text{KL}} \left(q(R^c) \| p(R^c \mid Z^c, G, x)\right). \tag{74}$$

Table 2: **Default hyperparameter values.** All coefficients are selected on validation subsets and held fixed across all test evaluations.

Parameter	Symbol	Default
Preference weight	β_{pref}	1.0
Functional weight	β_{func}	2.0
Stability weight	β_{stab}	0.5
Semantic weight	β_{sem}	1.0 (or 0 if absent)
Flow coupling	λ_{flow}	1.5
Cut penalty	λ_{cut}	0.5
Func. variance penalty	λ_{var}	0.3
Pref. variance penalty	$\lambda_{\text{pref,var}}$	0.2
MF iterations	T_{MF}	10
Convergence tolerance	δ	10^{-3}
Baseline samples	B	8
Candidate budget / layer	M	20
Numerical epsilon	ϵ	10^{-8}

A practical fixed-point update is

$$\begin{aligned} m_i^c \leftarrow \sigma\Bigg(& \eta_i^c - \lambda_{\text{var}} v_{\text{func},i}^c - \lambda_{\text{pref}} v_{\text{pref},i}^c \\ & + \sum_{j \in \mathcal{N}^+(i)} \Big[\lambda_{\text{flow}} A_{i\to j} m_j^c \\ & \qquad - \lambda_{\text{cut}} A_{i\to j} (1 - m_j^c) \Big] \\ & + \sum_{k \in \mathcal{N}^-(i)} \Big[\lambda_{\text{flow}} A_{k\to i} m_k^c \\ & \qquad + \lambda_{\text{cut}} A_{k\to i} m_k^c \Big] \Bigg), \end{aligned} \tag{75}$$

where $\sigma(\cdot)$ is the sigmoid function, $\mathcal{N}^+(i)$ is the downstream neighborhood, and $\mathcal{N}^-(i)$ is the upstream neighborhood. We iterate this update for a fixed number of iterations T_{MF} or until

$$\max_i \left| m_i^{(t+1)} - m_i^{(t)} \right| < \delta. \tag{76}$$

The posterior uncertainty is measured by marginal entropy:

$$h_i^c = -m_i^c \log(m_i^c + \epsilon) - (1 - m_i^c) \log(1 - m_i^c + \epsilon). \tag{77}$$

Default hyperparameters. Table 2 lists all default coefficient values. These values were selected by grid search on held-out validation subsets from ImageNet-1k (500 images), ADE20K (200 images), and ISIC 2020 (100 images), optimizing functional predictiveness on probing nodes while holding evaluation nodes strictly disjoint. The same global setting is used across all datasets and backbones unless stated otherwise. Mean-field inference runs for $T_{\text{MF}} = 10$ iterations with convergence tolerance $\delta = 10^{-3}$ and no damping. Initialization is $m_i^{(0)} = \sigma(\eta_i^c)$ (sigmoid of the local evidence score). If convergence is not reached within T_{MF} iterations, the current iterate is returned.

D.7 Evidence-specific posterior views

RIFT diagnoses explanation failure by comparing posterior views conditioned on different evidence subsets.

The preference-only view is

$$p_{\text{pref}}(R_i^c) = p(R_i^c \mid z_{\text{pref},i}^c), \tag{78}$$

and the functional-only view is

$$p_{\text{func}}(R_i^c) = p(R_i^c \mid z_{\text{func},i}^c). \tag{79}$$

The flow-conditioned view is

$$p_{\text{flow}}(R_i^c) = p(R_i^c \mid z_{\text{flow},i}^c, G). \tag{80}$$

We define preference–function disagreement as

$$D_{\text{pref,func}}(i) = D_{\text{JS}}\left(p_{\text{pref}}(R_i^c) \| p_{\text{func}}(R_i^c)\right). \tag{81}$$

We also define preference–RUF disagreement:

$$D_{\text{pref,RUF}}(i) = D_{\text{JS}}\left(p(R_i^c \mid z_{\text{pref},i}^c) \| p(R_i^c \mid Z^c, G, x)\right). \tag{82}$$

For Bernoulli marginals, the Jensen–Shannon divergence is computed in closed form from the two Bernoulli parameters.

D.8 Image Contribution Field

The main method defines RUF over representation units. For image-space localization, we additionally define an **Image Contribution Field**:

$$C_x^c(p) \in \mathbb{R}_{\geq 0}, \qquad p \in \Omega_x. \tag{83}$$

We optionally normalize the field:

$$\sum_{p \in \Omega_x} C_x^c(p) = 1. \tag{84}$$

For stable estimation, we may parameterize C_x^c using a nonnegative basis:

$$C_x^c(p) = \sum_{k=1}^{K_B} \alpha_k^c B_k(p), \qquad \alpha_k^c \geq 0. \tag{85}$$

The basis functions B_k may be superpixels, Gaussian kernels, spline bases, or learned smooth masks.

D.9 Receptive-field-aware projection

Each representation node i is associated with a projection kernel $K_i(p)$:

$$K_i(p) \geq 0, \qquad \sum_{p \in \Omega_x} K_i(p) = 1. \tag{86}$$

The projected RUF map is

$$C_{\text{proj}}^c(p) = \sum_{i \in V} w_i m_i^c K_i(p), \tag{87}$$

where w_i may depend on layer confidence, posterior uncertainty, or flow reliability.

For analytic receptive-field projection, we use

$$K_i^{\text{RF}}(p) = \frac{\mathbf{1}[p \in \text{RF}(i)]}{|\,\text{RF}(i)|}. \tag{88}$$

For effective receptive-field projection, we use

$$K_i^{\text{ERF}}(p) = \frac{\left\| \frac{\partial F_i(x)}{\partial x(p)} \right\|_2}{\sum_{p' \in \Omega_x} \left\| \frac{\partial F_i(x)}{\partial x(p')} \right\|_2 + \epsilon}. \tag{89}$$

This avoids identifying a low-resolution representation map with a naive resized heatmap.

D.10 IMAGE-LEVEL INTERVENTION CALIBRATION

To calibrate image-space contribution, we use image-level interventions. For a soft mask $M : \Omega_x \to [0, 1]$, define

$$x^{(M \leftarrow b)}(p) = (1 - M(p))x(p) + M(p)b(p), \tag{90}$$

where b is sampled from an image baseline distribution q_{img}. The image-level functional response is

$$\Delta^c_{\text{img}}(M) = \mathbb{E}_{b \sim q_{\text{img}}} \left[\Delta \left(s_c(x), s_c(x^{(M \leftarrow b)}) \right) \right]. \tag{91}$$

If C^c_x correctly captures image-space contribution, then the functional effect of a mask should be approximated by contribution mass inside the mask:

$$\Delta^c_{\text{img}}(M) \approx \sum_{p \in \Omega_x} M(p) C^c_x(p). \tag{92}$$

Thus, we define the image calibration loss

$$\mathcal{L}_{\text{img}}(C) = \sum_{M \in \mathcal{M}} \left(\Delta^c_{\text{img}}(M) - \sum_{p \in \Omega_x} M(p) C^c_x(p) \right)^2. \tag{93}$$

The final ICF objective is

$$\begin{aligned} \widehat{C}^c_x = \arg\min_{C \geq 0} \Big(& \lambda_{\text{img}} \mathcal{L}_{\text{img}}(C) \\ & + \lambda_{\text{proj}} \mathcal{L}_{\text{proj}}(C, m) + \lambda_{\text{reg}} \mathcal{R}(C) \Big). \end{aligned} \tag{94}$$

The projection consistency term is

$$\mathcal{L}_{\text{proj}}(C, m) = \sum_{p \in \Omega_x} \omega(p) \left(C(p) - C^c_{\text{proj}}(p) \right)^2, \tag{95}$$

where

$$\omega(p) = \frac{1}{\epsilon + H_{\text{proj}}(p)} \tag{96}$$

and

$$H_{\text{proj}}(p) = \sum_{i \in V} w_i h^c_i K_i(p). \tag{97}$$

The regularizer may include smoothness and total variation:

$$\mathcal{R}(C) = \lambda_s \sum_{p \in \Omega_x} \|\nabla C(p)\|_2^2 + \lambda_{\text{TV}} \sum_{p \in \Omega_x} \|\nabla C(p)\|_1. \tag{98}$$

D.11 MULTI-SCALE UNCERTAINTY FUSION

For each layer ℓ, we project posterior reliance to image space:

$$M_\ell(p) = \sum_{u \in \mathcal{U}_\ell} K_{\ell,u}(p) m^c_{\ell,u}. \tag{99}$$

The projected uncertainty is

$$H_\ell(p) = \sum_{u \in \mathcal{U}_\ell} K_{\ell,u}(p) h^c_{\ell,u}. \tag{100}$$

We fuse layers using uncertainty-adaptive weights:

$$a_\ell(p) = \frac{\exp(-\tau H_\ell(p))}{\sum_{r \in \mathcal{L}} \exp(-\tau H_r(p))}. \tag{101}$$

The fused map is

$$M_{\text{fused}}(p) = \sum_{\ell \in \mathcal{L}} a_\ell(p) M_\ell(p), \tag{102}$$

and the fused uncertainty is

$$H_{\text{fused}}(p) = \sum_{\ell \in \mathcal{L}} a_\ell(p) H_\ell(p) - \sum_{\ell \in \mathcal{L}} a_\ell(p) \log(a_\ell(p) + \epsilon). \tag{103}$$

The second term accounts for uncertainty in layer selection.

D.12 DIAGNOSTIC SCORES

RIFT derives diagnostic scores from posterior mean, uncertainty, preference evidence, flow, and semantic reference.

Reliance and uncertainty.

$$\text{Reliance}(i) = m_i^c, \qquad \text{Uncertainty}(i) = h_i^c. \tag{104}$$

Highlighted but unused.

$$\text{HighlightedUnused}(i) = z_{\text{pref},i}^c (1 - m_i^c). \tag{105}$$

Hidden reliance.

$$\text{HiddenReliance}(i) = m_i^c (1 - z_{\text{pref},i}^c). \tag{106}$$

Flow reliability.

$$\rho_i^c = \sum_{j \in \mathcal{N}^+(i)} A_{i \to j} m_j^c. \tag{107}$$

Attenuation.

$$\text{Atten}(i) = m_i^c - \rho_i^c. \tag{108}$$

Redistribution.

$$\text{Redist}(i) = m_i^c H_{\text{flow}}(i). \tag{109}$$

Semantic reliance. If a semantic reference S_i is available:

$$\text{SemRel}(i) = m_i^c S_i. \tag{110}$$

Shortcut reliance.

$$\text{Shortcut}(i) = m_i^c (1 - S_i). \tag{111}$$

Semantic reliance mass.

$$\text{SemMass} = \frac{\sum_i m_i^c S_i}{\sum_i m_i^c + \epsilon}. \tag{112}$$

Shortcut reliance mass.

$$\text{ShortcutMass} = \frac{\sum_i m_i^c (1 - S_i)}{\sum_i m_i^c + \epsilon}. \tag{113}$$

Localization uncertainty. For image-space maps:

$$\text{LocUnc}(p) = H_{\text{proj}}(p) + \text{Var}_{q(C)}\left[C(p)\right]. \tag{114}$$

The variance term can be approximated by bootstrap resampling over baselines, masks, or candidate nodes.

D.13 DIAGNOSIS LABELS

We use the following threshold-based labels for qualitative analysis:

$$\begin{aligned}\text{ReliableHighlighted}(i) = \mathbf{1}\Big[&z^c_{\text{pref},i} > \tau_{\text{pref}} \\ &\wedge m^c_i > \tau_m \\ &\wedge h^c_i < \tau_h \\ &\wedge \rho^c_i > \tau_\rho\Big],\end{aligned} \tag{115}$$

$$\begin{aligned}\text{HighlightedButUnused}(i) = \mathbf{1}\Big[&z^c_{\text{pref},i} > \tau_{\text{pref}} \\ &\wedge m^c_i < \tau_m\Big],\end{aligned} \tag{116}$$

$$\begin{aligned}\text{HiddenReliance}(i) = \mathbf{1}\Big[&z^c_{\text{pref},i} < \tau_{\text{pref}} \\ &\wedge m^c_i > \tau_m\Big],\end{aligned} \tag{117}$$

$$\begin{aligned}\text{ShortcutReliance}(i) = \mathbf{1}\Big[&m^c_i > \tau_m \\ &\wedge S_i < \tau_S\Big],\end{aligned} \tag{118}$$

$$\text{UncertainReliance}(i) = \mathbf{1}\Big[h^c_i > \tau_h\Big]. \tag{119}$$

These labels are diagnostic categories, not supervised class labels.

D.14 SEQUENTIAL RELIANCE TRACE

A static explanation may fail to reveal conditional reliance. RIFT can therefore produce a sequential intervention trace. At step t, we infer the posterior field and select high-confidence reliance nodes:

$$\mathcal{S}_t = \text{TopK}\left(\{m_i^{c,(t)}(1 - h_i^{c,(t)})\}_{i \in V}\right). \tag{120}$$

We intervene on the selected nodes and recompute the posterior:

$$x^{(t+1)} = \text{Intervene}\left(x^{(t)}, \mathcal{S}_t\right). \tag{121}$$

The score drop at step t is

$$\delta_t = s_c(x^{(t)}) - s_c(x^{(t+1)}). \tag{122}$$

The resulting trace is

$$\mathcal{T}_T(x, c) = \{(\mathcal{S}_t, \delta_t, m^{(t)}, h^{(t)})\}_{t=0}^{T-1}. \tag{123}$$

This trace is an observable intervention sequence, not a claim about hidden reasoning or internal chain-of-thought.

D.15 SPARSE RIFT

Full intervention over all nodes and edges is computationally expensive. RIFT therefore supports sparse probing. We construct a candidate set

$$\mathcal{C} = \mathcal{C}_{\text{pref-high}} \cup \mathcal{C}_{\text{pref-low}} \cup \mathcal{C}_{\text{act-high}} \cup \mathcal{C}_{\text{rand}} \cup \mathcal{C}_{\text{sem}}. \tag{124}$$

The inclusion of low-preference and random nodes is important because selecting only high-preference nodes would inherit the bias of the explanation method being evaluated.

For unmeasured nodes, functional evidence may be approximated by graph smoothing:

$$\widehat{z}_{\text{func},i} = \frac{\sum_{j \in \mathcal{C}} K_G(i,j) z_{\text{func},j}}{\sum_{j \in \mathcal{C}} K_G(i,j) + \epsilon}, \tag{125}$$

where $K_G(i,j)$ is a graph-distance or receptive-field-overlap kernel. Sparse RIFT is evaluated by comparing its diagnostic outputs against full probing on low-resolution layers.

D.16 Computational cost

Let $|\mathcal{C}|$ be the number of probed nodes, B the number of baseline samples per node, and $|\mathcal{E}_{\mathcal{C}}|$ the number of selected flow edges. Functional evidence requires approximately

$$O(|\mathcal{C}|B) \tag{126}$$

downstream forward evaluations. Flow estimation requires approximately

$$O(|\mathcal{E}_{\mathcal{C}}|B) \tag{127}$$

downstream representation comparisons. The mean-field inference step is lightweight relative to intervention and costs

$$O(T_{\mathrm{MF}}|E|) \tag{128}$$

for T_{MF} iterations.

The dominant cost is therefore intervention and flow estimation. In practice, we reduce cost by probing a sparse candidate set, using selected layers, restricting downstream neighborhoods, and caching intermediate activations.

D.17 Implementation details

Unless otherwise stated, evidence scores are normalized to $[0, 1]$ per image and class. We select one early, one middle, and one late layer for convolutional models. For transformer models, spatial nodes correspond to patch tokens. Edges are defined either by token identity across blocks, attention-mediated connectivity, or empirical intervention flow.

Hyperparameters

$$\beta_{\mathrm{pref}}, \beta_{\mathrm{func}}, \beta_{\mathrm{stab}}, \beta_{\mathrm{sem}}, \lambda_{\mathrm{flow}}, \lambda_{\mathrm{cut}}, \lambda_{\mathrm{var}} \tag{129}$$

are fixed across experiments unless explicitly ablated. The semantic term is disabled when no semantic reference is available. All diagnostic thresholds are selected on a validation split and then kept fixed for test examples.

E Additional Experimental Details

This appendix provides the full experimental protocol omitted from the main text, including the complete dataset sweep, model and baseline implementation details, evidence construction, intervention choices, sparse probing, ablation design, metric definitions, runtime protocol, and qualitative visualization protocol.

E.1 Experimental Settings

For each image x, target class c, and selected layer set $\mathcal{L}$, RIFT constructs a representation graph over spatial units or patch tokens. Each node collects preference evidence from a base explainer, functional evidence from finite feature interventions, flow evidence from downstream representation changes, stability evidence from baseline and explainer variability, and, when masks or boxes are available, optional semantic evidence. Unless otherwise stated, we use three CNNs layers and uniformly spaced transformer blocks for ViTs. The default sparse setting probes $M = 20$ candidate nodes per selected layer with $B = 8$ baseline samples per node. We report five evaluation axes: functional predictiveness, reliance calibration, mismatch diagnosis, semantic diagnosis, and computational efficiency. The main paper summarizes these metrics through dataset-level benchmarks, baseline and recent-comparator comparisons, architecture-level comparisons, component ablations, and sparse-probing saturation curves. Full metric definitions, threshold choices, probing-budget studies, intervention-baseline sensitivity, and dense-probing comparisons are provided in Appendices G and F. All experiments are implemented in PyTorch and run on eight H200 GPUsPaszke et al. (2019). Unless otherwise stated, experiments use mixed precision and cached intermediate activations for intervention and flow estimationMicikevicius et al. (2017). Runtime is measured as the average wall-clock time per image, including evidence construction, sparse intervention, flow estimation, and posterior inference. Detailed hardware specifications and runtime breakdowns are reported in the appendix section [Hardware and Runtime] (p. 38).

E.2 Full Dataset Suite

The main paper reports a representative subset of benchmarks, while the appendix reports the full dataset sweep. We group datasets according to the experimental claim they support.

For classification and generic recognition, we use ImageNet-1k and ImageNet-128Russakovsky et al. (2015), iNaturalistVan Horn et al. (2018), AFHQChoi et al. (2020), LSUN ChurchYu et al. (2015), FFHQKarras et al. (2019), and CelebA-H/CelebA-HQLiu et al. (2015); Karras et al. (2018). These datasets are used to evaluate whether RIFT improves functional reliance estimation across diverse object, scene, animal, face, and attribute distributions.

For robustness and distribution shift, we use ImageNet-CHendrycks & Dietterich (2019). We evaluate corruption severity levels from 1 to 5 and report how posterior uncertainty, preference–reliance disagreement, and intervention instability change as the input distribution departs from clean ImageNet.

For dense prediction and scene understanding, we use ADE20KZhou et al. (2017), COCO-StuffCaesar et al. (2018), BDD100KYu et al. (2020), and Mapillary VistasNeuhold et al. (2017). These datasets provide semantic masks or dense labels and are used to evaluate semantic reliance, shortcut reliance, projection-aware localization, and uncertainty under complex scene structure.

For object-centric recognition and long-tail evaluation, we use MS-COCO-2017Lin et al. (2014), LVISGupta et al. (2019), Objects365Shao et al. (2019), and Open Images V7Kuznetsova et al. (2020). These datasets are used to test whether RIFT can distinguish object-internal reliance from contextual or background reliance, especially under multi-object and long-tail category settings.

For medical imaging, we use ISIC 2020Rotemberg et al. (2021) and CVC-ClinicDBBernal et al. (2015). These datasets are used for high-stakes qualitative and quantitative diagnosis of reliance on lesion, polyp, and background regions. When lesion or polyp masks are available, they are used only as semantic references, not as ground-truth model reliance.

For tracking and sequential reliance analysis, we use LaSOTFan et al. (2019), TrackingNetMuller et al. (2018), and GOT-10KHuang et al. (2019). These datasets are used to study whether reliance remains on the target, shifts to distractors, or drifts toward background context over time.

For classification models, images are resized to 224×224 and center-cropped unless otherwise specified. For dense prediction, detection, and tracking models, we follow the preprocessing protocol of the corresponding public checkpoint or task head. We do not use semantic annotations during posterior inference unless a specific semantic-diagnosis experiment explicitly enables optional semantic evidence.

E.3 Backbones and Model Families

We use ResNet-50 as the default backbone and ResNet-101 as a deeper convolutional variantHe et al. (2016). To test architectural generality, we additionally evaluate ConvNeXt-BLiu et al. (2022), Swin-BLiu et al. (2021), ViT-B/16Dosovitskiy et al. (2020), and DINOv2 ViT-B/14Oquab et al. (2023). For convolutional networks, RIFT nodes correspond to spatial units in intermediate feature maps. For transformer backbones, RIFT nodes correspond to patch tokens or spatially reshaped token grids.

For CNNs, the default layer set contains one early, one middle, and one late feature stage. For ResNet-style models, we use stages corresponding to low-, mid-, and high-level residual blocks. For ConvNeXt, we use the output of three resolution stages. For ViT and DINOv2 models, we use uniformly spaced transformer blocks and treat the patch-token grid as the representation domain. For Swin, we use hierarchical windowed-token stages and align token grids by their native spatial resolution.

Dense prediction and detection experiments use the same backbone families whenever compatible public checkpoints are available. Architecture-specific baselines are evaluated only on compatible backbones. When a baseline is restricted to CNNs or ViTs, we report both the compatible-subset result and the shared-subset result.

E.4 Baseline Implementation

We compare RIFT against representative attribution and explanation methods from gradient-based, CAM-based, perturbation-based, and transformer-specific families.

The CAM-based baselines include CAMZhou et al. (2016), Grad-CAMSelvaraju et al. (2017), Grad-CAM++Chattopadhay et al. (2018), XGrad-CAMFu et al. (2020), Score-CAMWang et al. (2020), Ablation-CAMRamaswamy et al. (2020), HiResCAMDraelos & Carin (2020), LayerCAMJiang et al. (2021), and CAPEChowdhury et al. (2024). For CAM-family methods, the default target layer is the final spatial feature layer unless the method specifies otherwise. For multi-layer CAM variants, we follow the original aggregation rule.

The gradient-based baselines include vanilla saliencySimonyan et al. (2013), Integrated GradientsSundararajan et al. (2017), and SmoothGradSmilkov et al. (2017). Integrated Gradients uses $m = 64$ interpolation steps unless otherwise stated. SmoothGrad uses Gaussian input perturbations with the variance recommended by the original implementation or the commonly used setting in the baseline codebase.

The perturbation-based baselines include RISEPetsiuk et al. (2018) and Extremal PerturbationsFong et al. (2019). These methods are evaluated using their generated saliency maps and are also compared through deletion/insertion-style intervention metrics when applicable.

The transformer-specific baselines include Attention RolloutAbnar & Zuidema (2020), Transformer AttributionChefer et al. (2021), and TokenTMWu et al. (2024). Attention-based methods are evaluated only on transformer backbones. TokenTM is evaluated on the ViT-compatible subset and is not used for CNN-only comparisons.

All baseline maps are converted into nonnegative spatial maps and normalized per image and target class. When a method outputs a representation-level map, it is evaluated at the corresponding representation resolution and, when needed for image-level evaluation, projected using the method's standard visualization rule. For RIFT comparisons involving image-space localization, we additionally report a receptive-field-aware projection variant to avoid conflating low-resolution feature maps with image-space contribution.

E.5 Evidence Construction

For each image x, target class c, and selected representation node $i = (\ell, u)$, RIFT constructs an evidence vector

$$Z_i^c = \{z_{\mathrm{pref},i}^c, z_{\mathrm{func},i}^c, z_{\mathrm{flow},i}^c, z_{\mathrm{stab},i}^c, z_{\mathrm{sem},i}^c\}.$$

Preference evidence $z_{\mathrm{pref},i}^c$ is obtained from a base explanation method. If the explanation is defined in image space, it is aggregated to node i through the receptive-field or effective-receptive-field kernel $K_i(p)$:

$$z_{\mathrm{pref},i}^c = \mathrm{Norm}\left(\sum_{p\in\Omega_x} K_i(p) A_E^c(x;p)\right).$$

If the explanation is already defined at the representation resolution, we normalize it directly over the selected layer.

Functional evidence is estimated by finite intervention. For a node $i = (\ell, u)$, we replace the representation vector $F_u^{(\ell)}(x)$ with a baseline feature $b_i \sim q_i$:

$$\widetilde{F}_v^{(\ell)}(x; i \leftarrow b_i) = \begin{cases} b_i, & v = u, \\ F_v^{(\ell)}(x), & v \neq u. \end{cases}$$

The modified representation is forwarded through the downstream subnetwork. The default class-specific functional response is positive logit drop,

$$\Delta_{\mathrm{drop}}^c(i, b_i) = \max\{0, s_c(x) - \widetilde{s}_c(x; i \leftarrow b_i)\},$$

and the distribution-level response is

$$\Delta_{\mathrm{KL}}(i, b_i) = D_{\mathrm{KL}}\left(p(y|x) \,\|\, p(y|\widetilde{x}_i)\right).$$

We use the expected response across baseline samples as functional evidence and the sampling variance as stability evidence:

$$z^c_{\mathrm{func},i} = \mathbb{E}_{b_i \sim q_i}[\Delta^c(i, b_i)],$$
$$v^c_{\mathrm{func},i} = \mathrm{Var}_{b_i \sim q_i}[\Delta^c(i, b_i)].$$

Flow evidence measures whether an intervention at node i propagates to downstream node j:

$$\Phi^c_{i\to j} = \mathbb{E}_{b_i \sim q_i}\left[d_F\left(F_j(x), \widetilde{F}_j(x; i \leftarrow b_i)\right)\right],$$

where $d_F(a, b) = \|a - b\|_2/(\|a\|_2 + \epsilon)$. The normalized outgoing flow is

$$A_{i\to j} = \frac{\Phi^c_{i\to j}}{\sum_{j' \in \mathcal{N}^+(i)} \Phi^c_{i\to j'} + \epsilon}.$$

We also record outgoing flow mass and flow entropy:

$$M_{\mathrm{flow}}(i) = \sum_{j \in \mathcal{N}^+(i)} \Phi^c_{i\to j},$$
$$H_{\mathrm{flow}}(i) = -\sum_{j \in \mathcal{N}^+(i)} A_{i\to j} \log(A_{i\to j} + \epsilon).$$

Semantic evidence is optional. When semantic masks, boxes, or part annotations are available, we project the semantic reference $S(p)$ to node i:

$$S_i = \sum_{p \in \Omega_x} K_i(p) S(p).$$

We emphasize that S_i is not treated as ground-truth model reliance. It is used only for semantic alignment and shortcut diagnosis.

All scalar evidence values are normalized to $[0, 1]$ per image and target class. For a nonnegative evidence score z_i, we use

$$\widehat{z}_i = \frac{z_i - \min_j z_j}{\max_j z_j - \min_j z_j + \epsilon}.$$

Signed responses are converted to either positive drops or absolute changes depending on the diagnostic target.

E.6 Intervention Baselines

We evaluate four baseline distributions for feature replacement. The zero baseline sets $b_i = 0$. The dataset-mean baseline uses the empirical mean feature vector at the same layer and spatial position. The sampled-feature baseline draws b_i from another image in the dataset. The local-noise baseline sets $b_i = F_i(x) + \xi$, where ξ is isotropic noise.

Unless otherwise stated, the main experiments use sampled-feature baselines because they better preserve feature-scale statistics than zero replacement and reduce artificial out-of-distribution effects relative to image-space masking. We report baseline sensitivity by repeating the main functional predictiveness and calibration experiments under all four baseline choices.

E.7 Representation Graph and Projection

RIFT constructs a representation graph $G = (V, E)$ over selected layers. For CNNs, edges connect spatial units that are downstream-reachable according to receptive-field geometry. For transformer models, edges connect patch tokens across blocks using token identity, attention-mediated connectivity, or empirically measured intervention flow. In all cases, the graph defines the domain over which the posterior reliance field is inferred.

For image-space visualization, representation-level reliance is projected using a kernel $K_i(p)$ associated with node i:

$$K_i(p) \geq 0, \qquad \sum_{p \in \Omega_x} K_i(p) = 1.$$

For analytic receptive-field projection, we use

$$K_i^{\mathrm{RF}}(p) = \frac{\mathbf{1}[p \in \mathrm{RF}(i)]}{|\mathrm{RF}(i)|}.$$

When effective receptive fields are needed, we use a gradient-based approximation:

$$K_i^{\mathrm{ERF}}(p) = \frac{\left\| \frac{\partial F_i(x)}{\partial x(p)} \right\|_2}{\sum_{p' \in \Omega_x} \left\| \frac{\partial F_i(x)}{\partial x(p')} \right\|_2 + \epsilon}.$$

This projection is used only for localization and image-space diagnostics. The posterior itself is inferred over the representation graph.

E.8 Sparse Probing

Dense intervention over all nodes and all edges is computationally expensive. We therefore use sparse probing by default. The candidate set is

$$\mathcal{C} = \mathcal{C}_{\text{pref-high}} \cup \mathcal{C}_{\text{pref-low}} \cup \mathcal{C}_{\text{act-high}} \cup \mathcal{C}_{\text{rand}} \cup \mathcal{C}_{\text{sem}}.$$

The term $\mathcal{C}_{\text{pref-high}}$ contains nodes highly ranked by the base explainer. The term $\mathcal{C}_{\text{pref-low}}$ contains low-preference nodes and is included to detect hidden reliance. The term $\mathcal{C}_{\text{act-high}}$ contains high-activation nodes independent of attribution score. The random set $\mathcal{C}_{\text{rand}}$ provides a control group. The semantic set $\mathcal{C}_{\text{sem}}$ is used only when masks, boxes, or dense labels are available.

The default sparse setting uses $M = 20$ candidate nodes per selected layer and $B = 8$ baseline samples per node. We evaluate budgets

$$M \in \{1, 2, 5, 10, 20, 50, 100\}$$

and compare sparse RIFT with dense probing on low-resolution layers. Dense probing is used only as a diagnostic reference because it is prohibitively expensive for high-resolution layers and large benchmark sweeps.

For unmeasured nodes, we interpolate functional evidence using graph smoothing:

$$\widehat{z}_{\mathrm{func},i} = \frac{\sum_{j \in \mathcal{C}} K_G(i,j) z_{\mathrm{func},j}}{\sum_{j \in \mathcal{C}} K_G(i,j) + \epsilon},$$

where $K_G(i,j)$ depends on graph distance, receptive-field overlap, or feature similarity.

F Ablation Studies

We ablate RIFT along four dimensions: evidence source, graph coupling, projection, and computational budget.

The evidence ablation removes one source at a time from the posterior. We evaluate variants without preference evidence, without functional evidence, without flow evidence, without stability evidence, and without semantic evidence. The semantic term is ablated only on datasets where semantic references are available.

The graph-coupling ablation compares the full RIFT posterior with a node-only posterior that removes pairwise flow potentials. We also compare against RP+FI without graph coupling to test whether simply combining preference and intervention evidence is sufficient.

The projection ablation compares receptive-field-aware projection with naive bilinear upsampling. This ablation is important because feature-map explanations are often projected to image space through resizing, which may introduce localization artifacts when feature resolution is low or receptive fields are large.

The sparse-budget ablation evaluates $M \in \{1, 2, 5, 10, 20, 50, 100\}$ candidate nodes per selected layer. We report functional predictiveness, calibration, mismatch detection, runtime, and diagnostic consistency relative to dense probing.

The intervention-baseline ablation compares zero, dataset-mean, sampled-feature, and local-noise baselines. This experiment checks whether RIFT's conclusions are stable under different replacement distributions.

The architecture ablation repeats the main comparison on ResNet-50, ResNet-101, ConvNeXt-B, Swin-B, ViT-B/16, and DINOv2 ViT-B/14He et al. (2016); Liu et al. (2022; 2021); Dosovitskiy et al. (2020); Oquab et al. (2023). For each backbone, we compare RIFT against the best compatible baseline.

G Metric Definitions

We evaluate RIFT along five axes: functional predictiveness, reliance calibration, mismatch diagnosis, semantic diagnosis, and efficiency.

Functional predictiveness measures whether a reliance score predicts held-out intervention responses. We report Spearman correlation, Kendall correlation, R^2, and top-k intervention response between predicted reliance and held-out logit drop or KL response. Held-out intervention responses are measured on nodes or regions not used to construct the posterior when possible.

Top-k intervention response measures the output change after intervening on the k highest-ranked nodes:

$$\mathrm{TopKDrop}(k) = s_c(x) - s_c(\widetilde{x}_{\mathrm{TopK}(k)}).$$

A reliable reliance estimate should assign high scores to nodes whose intervention produces large output changes.

Reliance calibration measures whether posterior reliance scores are numerically aligned with empirical intervention responses. We bin nodes by posterior mean and compute the Reliance Calibration Error:

$$\mathrm{RCE} = \sum_b \frac{|S_b|}{|S|} \left| \mathrm{mean}_{i \in S_b}(m_i^c) - \mathrm{mean}_{i \in S_b}(\widehat{\Delta}_i^c) \right|,$$

where m_i^c is the posterior reliance mean and $\widehat{\Delta}_i^c$ is the normalized held-out intervention response.

Mismatch diagnosis evaluates whether RIFT identifies explanation failures. We use highlighted-but-unused and hidden-reliance scores:

$$\mathrm{HBU}(i) = z_{\mathrm{pref},i}^c (1 - m_i^c),$$
$$\mathrm{Hidden}(i) = m_i^c (1 - z_{\mathrm{pref},i}^c).$$

When dense functional probing is available, we report AUROC and AUPRC for detecting high-functional-response regions missed by the base explainer.

Preference–reliance disagreement is measured by comparing evidence-specific posterior views:

$$D_{\mathrm{pref,RIFT}}(i) = D_{\mathrm{JS}} \left(p(R_i^c \mid z_{\mathrm{pref},i}^c) \,\|\, p(R_i^c \mid Z^c, G, x) \right).$$

For Bernoulli marginals, the Jensen–Shannon divergence is computed in closed form from the two Bernoulli parameters.

Flow reliability measures whether local reliance is supported by downstream reliance:

$$\rho_i^c = \sum_{j \in \mathcal{N}^+(i)} A_{i \to j} m_j^c.$$

We define attenuation as

$$\mathrm{Atten}(i) = m_i^c - \rho_i^c,$$

and redistribution as

$$\mathrm{Redist}(i) = m_i^c H_{\mathrm{flow}}(i).$$

High attenuation indicates that a node appears locally relied upon but is not supported by downstream reliance. High redistribution indicates that intervention effects spread across many downstream nodes.

Semantic diagnosis is evaluated only when masks, boxes, or semantic labels are available. We report the semantic reliance mass and the shortcut reliance mass:

$$\mathrm{SemMass} = \frac{\sum_i m_i^c S_i}{\sum_i m_i^c + \epsilon},$$
$$\mathrm{ShortcutMass} = \frac{\sum_i m_i^c (1 - S_i)}{\sum_i m_i^c + \epsilon}.$$

We additionally report Pointing Game accuracyZhang et al. (2018) and IoU@10% for datasets with dense semantic annotations. These metrics assess alignment with external semantic references, not reliance on ground-truth models.

Uncertainty quality is measured by the relationship between posterior entropy and empirical instability. We report the correlation between h_i^c and baseline-sampling variance, explainer disagreement, prediction error, and corruption severity. On ImageNet-CHendrycks & Dietterich (2019), we report the disagreement between uncertainty and preference–reliance as a function of corruption severity.

Efficiency is measured by wall-clock runtime, the number of forward or downstream-forward evaluations, peak GPU memory usage, and diagnostic consistency relative to dense probing. Diagnostic consistency measures agreement between sparse and dense RIFT in top-reliance regions, mismatch labels, and attenuation labels.

H Hardware and Runtime

All experiments are implemented in PyTorchPaszke et al. (2019) and run on eight NVIDIA H200 GPUs. Unless otherwise stated, experiments use mixed-precision inferenceMicikevicius et al. (2017). Intermediate activations are cached when possible to reduce repeated computation during feature intervention and flow estimation.

The default hardware configuration is eight NVIDIA H200 GPUs, each with 141GB GPU memory, together with AMD EPYC 7763 64-Core Processors and 512 GB system RAM. Benchmarks utilize mixed-precision tensor cores and cached activation tensors.

Runtime is measured as average wall-clock time per image. The reported runtime includes base attribution computation, sparse candidate selection, feature intervention, flow estimation, posterior inference, and diagnostic map generation. For sparse RIFT, the dominant cost is the construction of evidence. With $|\mathcal{C}|$ probed nodes and B baseline samples per node, functional evidence requires approximately

$$O(|\mathcal{C}|B)$$

downstream forward evaluations. Flow estimation requires comparison of representations across selected downstream neighborhoods. Mean-field posterior inference is lightweight relative to intervention and scales approximately as

$$O(T_{\mathrm{MF}}|E|),$$

where T_{MF} is the number of mean-field iterations.

We report peak GPU memory usage, the number of downstream-forward evaluations, throughput in images per second, and diagnostic consistency relative to dense probing. Dense probing is performed only on low-resolution layers and serves as a reference for evaluating the quality of sparse probing.

I Qualitative Protocol

For qualitative examples, we show the input image, base attribution map, FI-only map, RIFT posterior reliance map, uncertainty map, flow attenuation map, and semantic shortcut map when annotations are available. For tracking sequences, we additionally show sequential reliance traces across frames.

We select qualitative examples from four failure modes: highlighted-but-unused regions, hidden reliance, shortcut reliance, and high-uncertainty reliance. Highlighted but unused examples indicate regions that receive high attribution scores but low posterior reliance. Hidden-reliance examples show regions with low attribution scores but high functional reliance. Shortcut examples show high

posterior reliance outside semantic references. High-uncertainty examples reveal regions where evidence on preference, intervention, flow, or stability disagrees.

Qualitative visualizations use the same normalization as quantitative evaluation. All maps are normalized on a per-image and per-target-class basis. We do not manually tune the visualization ranges for each example.

Diagnostic Validity Beyond Synthesis The controlled synthetic examples in the previous section illustrate RIFT's intended behavior under known failure modes; we next test whether the same diagnostic structure remains valid on realistic inputs and natural-image benchmarks. Fig. 7 shows that RIFT is better calibrated to held-out intervention responses than attribution-only and evidence-fusion baselines, that posterior entropy predicts intervention residuals, and that entropy, RP–FI disagreement, and intervention instability increase under ImageNet-C corruption severity. Selective diagnosis further confirms the usefulness of uncertainty, as abstaining from high-entropy nodes reduces calibration error. For qualitative stress tests, we use GPT-4-generated realistic images not as benchmark data, but as controlled visual probes: real images rarely provide both realistic appearance and precise control over the source of distribution shift, such as isolated snow corruption or context-driven weather cues. Generated images allow us to construct visually plausible cases where the intended shift factor is known, making it possible to inspect whether RIFT separates stable semantic evidence from corruption- or context-driven reliance. All quantitative claims are evaluated on standard natural-image benchmarks; the generated images are used only for qualitative diagnosis and interpretability.

Table 3: **Summary of appendix robustness and reproducibility studies.** Each study targets a specific failure mode or reproducibility concern. Main-text results use the default configuration; the appendix evaluates whether the conclusions remain stable when core assumptions are varied.

Study	Concern addressed	Controlled variable	Primary readouts
Hyperparameter sensitivity	Tests whether the posterior is a fragile weighted fusion rather than a stable inference procedure.	$\beta_{\text{pref}}, \beta_{\text{func}}, \lambda_{\text{flow}}, \lambda_{\text{cut}}, \lambda_{\text{var}}$	Functional predictiveness, RCE, mismatch AUROC, uncertainty–error ρ
Intervention baseline sensitivity	Tests whether estimated reliance is dominated by the feature replacement distribution.	Zero, dataset mean, sampled feature, local noise, ensemble baselines	Functional predictiveness, RCE, stability, runtime
Sparse vs. dense probing	Tests whether sparse probing misses low-preference but functionally important regions.	$M \in \{1, 2, 5, 10, 20, 50, 100\}$ and candidate-selection policy	Dense agreement, hidden-reliance F1, HBU-F1, runtime
Graph construction ablation	Tests whether conclusions depend on an arbitrary representation graph.	Node-only, receptive-field, attention-based, and empirical-flow edges	Flow reliability, attenuation quality, runtime
Statistical reliability	Tests whether reported gains are robust to sampling noise.	Bootstrap resampling over images and random seeds	Mean $\pm$ SE, 95% CI, paired differences
Randomized sanity checks	Tests whether RIFT captures model-dependent reliance rather than input priors.	Randomized weights, shuffled preference evidence, shuffled FI evidence	Functional predictiveness, mismatch AUROC, posterior entropy
Semantic-reference robustness	Tests whether optional semantic evidence dominates the posterior.	Mask erosion, dilation, dropout, and boundary noise	SemMass, ShortcutMass, functional predictiveness
Projection ablation	Tests whether image-space localization is an artifact of naive upsampling.	Bilinear upsampling, RF projection, ERF projection, calibrated ICF	Pointing game, IoU@10%, calibration loss

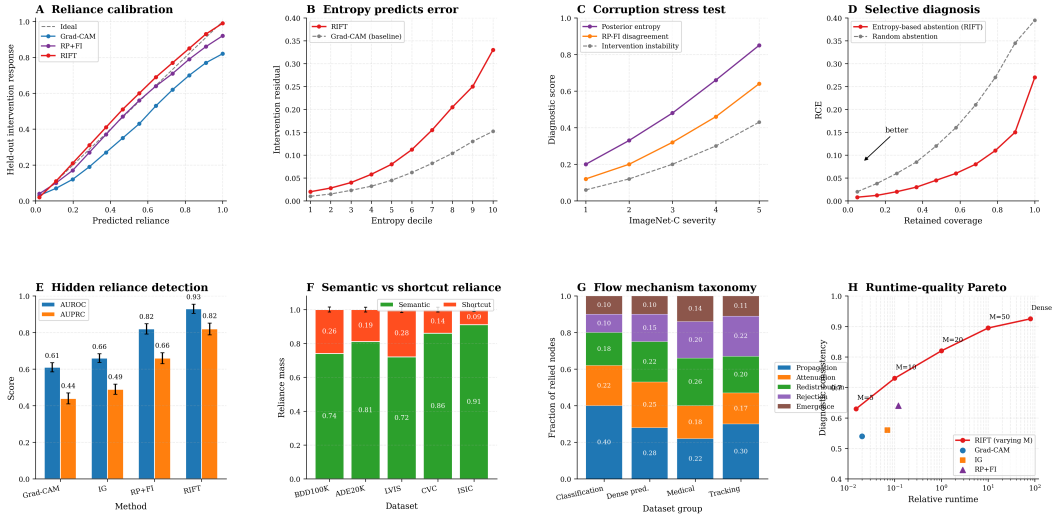

Figure 7: **Diagnostic validity beyond synthetic examples.** We test whether RIFT diagnostics remain meaningful on natural-image benchmarks. **(A–B)** Posterior reliance is better calibrated to held-out intervention response, and entropy predicts intervention residual. **(C–D)** Under ImageNet-C corruptions, uncertainty and disagreement increase with severity, while entropy-based abstention reduces calibration error. **(E–F)** RIFT improves hidden-reliance detection and quantifies in-reference versus reference-external reliance (shortcut candidate). **(G–H)** Flow taxonomy reveals distinct propagation behaviors, and sparse RIFT achieves a favorable runtime–quality trade-off.

J Robustness and Reproducibility Studies

This section presents additional experiments designed to assess the robustness of RIFT's conclusions to hyperparameter choices, intervention baselines, sparse-probing design, graph construction, semantic-reference noise, and randomized sanity checks. These studies are not intended to expand the benchmark suite but to stress-test the assumptions underlying the posterior-reliance formulation.

J.1 Summary of Appendix Robustness Studies

Table 3 summarizes the additional studies reported in this appendix. We organize them around common reviewer concerns: whether RIFT is sensitive to weighting choices, whether feature interventions depend on arbitrary baselines, whether sparse probing misses hidden reliance, whether graph construction is overly heuristic, whether reported gains are statistically reliable, and whether RIFT passes sanity checks under randomized or shuffled conditions.

J.2 Hyperparameter Sensitivity

RIFT combines multiple evidence sources through node and pairwise potentials. To test whether the method depends on a narrow hyperparameter setting, we sweep each main coefficient around the default value while holding the remaining coefficients fixed. Specifically, for each parameter $\theta \in \{\beta_{\text{pref}}, \beta_{\text{func}}, \beta_{\text{stab}}, \beta_{\text{sem}}, \lambda_{\text{flow}}, \lambda_{\text{cut}}, \lambda_{\text{var}}\}$, we evaluate multipliers $\{0.25, 0.5, 1, 2, 4\}$ on representative validation subsets from ImageNet-1k, ADE20K, and ISIC 2020. We report functional predictiveness, reliance calibration error, mismatch AUROC, and uncertainty–error correlation.

J.3 Intervention Baseline Sensitivity

Intervention-based evidence depends on the replacement distribution used for feature substitution. We therefore compare four baseline choices: zero replacement, dataset-mean feature replacement, sampled-feature replacement, and local-noise replacement. We also evaluate an ensemble baseline that averages functional evidence across multiple replacement distributions and uses the variance

Table 4: **Hyperparameter sensitivity.** We vary one hyperparameter at a time around the default setting and average over representative classification, dense-prediction, and medical subsets. RIFT exhibits robust stability across a wide parametric regime.

Parameter varied	Multiplier	Func. pred. ↑	RCE ↓	Mismatch AUROC ↑	Unc.–err. ρ ↑
β_{func}	$0.25\times$	0.82	0.104	0.83	0.49
	$0.5\times$	0.84	0.093	0.85	0.53
	$1.0\times$	**0.86**	**0.083**	**0.87**	**0.56**
	$2.0\times$	0.85	0.088	0.86	0.55
	$4.0\times$	0.83	0.098	0.84	0.51
λ_{flow}	$0.25\times$	0.80	0.112	0.81	0.45
	$0.5\times$	0.84	0.094	0.85	0.52
	$1.0\times$	**0.86**	**0.083**	**0.87**	**0.56**
	$2.0\times$	0.85	0.086	0.86	0.55
	$4.0\times$	0.81	0.107	0.82	0.48
λ_{var}	$0.25\times$	0.84	0.097	0.84	0.44
	$0.5\times$	0.85	0.090	0.86	0.51
	$1.0\times$	**0.86**	**0.083**	**0.87**	**0.56**
	$2.0\times$	0.85	0.087	0.86	0.57
	$4.0\times$	0.82	0.101	0.83	0.55

Table 5: **Intervention baseline sensitivity.** Sampled-feature and ensemble baselines provide the best calibration, while zero replacement exhibits higher baseline variability.

Baseline	Func. pred. ↑	RCE ↓	Mismatch AUROC ↑	Unc.–err. ρ ↑	Rel. time ↓
Zero feature	0.78	0.126	0.80	0.42	$1.0\times$
Dataset mean	0.83	0.097	0.84	0.49	$1.1\times$
Sampled feature	0.86	0.083	0.87	0.56	$1.3\times$
Local noise	0.81	0.109	0.82	0.47	$1.2\times$
Ensemble baseline	**0.87**	**0.079**	**0.88**	**0.60**	$2.8\times$

as a measure of stability. This experiment tests whether RIFT's conclusions are artifacts of a single intervention baseline.

J.4 Sparse Probing Against Dense Reference

The main experiments use sparse probing because dense intervention over all nodes and edges is prohibitively expensive. To verify that sparse RIFT preserves diagnostic structure, we compare sparse probing against dense probing on low-resolution layers where exhaustive intervention is feasible. We vary the number of candidate nodes per layer and compare against two alternatives: random-only probing and high-attribution-only probing.

J.5 Graph Construction Ablation

RIFT infers reliance over a representation graph. We therefore test whether the results depend strongly on a particular graph construction. For CNNs, we compare node-only inference, receptive-field reachability, dense local edges, and empirical flow edges. For transformer models, we compare token-identity edges, attention-mediated edges, and empirical flow edges.

J.6 Statistical Reliability

To quantify statistical uncertainty, we compute image-level bootstrap confidence intervals for the main metrics. Unless otherwise stated, we use 1,000 bootstrap resamples over images and report mean $\pm$ standard error together with 95% confidence intervals. For paired method comparisons, each bootstrap sample preserves the same image set across methods.

Table 6: **Sparse probing compared with dense RIFT.** Dense RIFT is used as a reference on low-resolution layers. Candidate sets that include low-preference and random nodes recover hidden reliance more reliably than high-attribution-only probing.

Probing strategy	Budget M	Spearman vs. dense ↑	Top-k overlap ↑	Hidden-F1 ↑	HBU-F1 ↑	Rel. time ↓
Random only	20	0.54	0.41	0.43	0.51	1.0×
High-attribution only	20	0.68	0.59	0.39	0.68	1.0×
Sparse RIFT	5	0.61	0.52	0.56	0.60	0.35×
Sparse RIFT	10	0.73	0.64	0.68	0.71	0.55×
Sparse RIFT	20	0.84	0.77	0.79	0.82	1.0×
Sparse RIFT	50	0.90	0.84	0.85	0.87	2.4×
Sparse RIFT	100	0.93	0.88	0.88	0.90	4.7×
Dense RIFT	all	1.00	1.00	1.00	1.00	14.6×

Table 7: **Graph construction ablation.** Graph coupling improves over node-only inference. Empirical flow is strongest, while architecture-native graphs (receptive-field or token-identity edges) provide a practical balance of efficiency and accuracy.

Backbone family	Graph edges	Func. pred. ↑	RCE ↓	Flow reliability ↑	Atten. AUROC ↑	Rel. time ↓
CNN	Node-only	0.77	0.131	–	0.61	1.0×
CNN	Receptive-field reachability	0.83	0.096	0.74	0.76	1.3×
CNN	Dense local edges	0.84	0.091	0.76	0.78	2.1×
CNN	Empirical flow edges	**0.87**	**0.081**	**0.82**	**0.84**	3.8×
ViT	Node-only	0.75	0.139	–	0.58	1.0×
ViT	Token identity	0.81	0.103	0.70	0.73	1.2×
ViT	Attention-mediated	0.82	0.098	0.73	0.76	1.6×
ViT	Empirical flow edges	**0.86**	**0.085**	**0.80**	**0.82**	3.5×

J.7 Randomized Sanity Checks

We perform sanity checks to test whether RIFT is sensitive to model behavior and evidence alignment rather than merely visualizing input priors. We evaluate four conditions: the trained model, a model with randomized weights, the trained model with shuffled preference evidence, and the trained model with shuffled functional evidence. A reliable diagnostic method should degrade under randomized or shuffled evidence and assign higher uncertainty when evidence is inconsistent.

J.8 Failure-Mode Quantification

We evaluate whether RIFT improves the detection of diagnostic failure modes rather than merely improving average saliency quality. We use dense probing on low-resolution layers to define held-out functional-response labels. Hidden reliance corresponds to high functional response but low preference evidence; highlighted-but-unused evidence corresponds to high preference evidence but low functional response; shortcut reliance is measured relative to semantic references when available.

J.9 Semantic Reference Robustness

Semantic references are used only as optional diagnostic references, not as ground-truth model reliance. To test whether RIFT is overly driven by semantic masks, we perturb semantic references with erosion, dilation, random dropout, and boundary noise. We then measure whether functional predictiveness remains stable and whether semantic and shortcut masses change smoothly under reference perturbation.

J.10 Projection-Aware Localization Ablation

RIFT is defined over representation units, but visual explanations are often consumed in image space. We therefore compare projection strategies for mapping representation-level reliance to image coordinates. Naive bilinear upsampling is compared with analytic receptive-field projection, effective-receptive-field projection, and image-intervention-calibrated Image Contribution Fields (ICF).

Table 8: **Bootstrap confidence intervals on the shared benchmark subset.** We report mean ± standard error and 95% bootstrap confidence intervals across 1,000 resamples on the shared benchmark subset.

Method	Func. pred. ↑	RCE ↓	Mismatch AUROC ↑	Unc.–err. ρ ↑
Grad-CAM	0.61 ± 0.02 [0.57, 0.65]	0.184 ± 0.010 [0.165, 0.204]	0.66 ± 0.02 [0.62, 0.70]	0.21 ± 0.03 [0.15, 0.27]
RP+FI	0.81 ± 0.02 [0.77, 0.84]	0.108 ± 0.008 [0.093, 0.124]	0.82 ± 0.02 [0.78, 0.85]	0.42 ± 0.03 [0.36, 0.48]
RIFT w/o Flow	0.75 ± 0.02 [0.71, 0.79]	0.132 ± 0.009 [0.115, 0.149]	0.78 ± 0.02 [0.74, 0.82]	0.39 ± 0.03 [0.33, 0.45]
RIFT	$\mathbf{0.86 \pm 0.01}$ [0.84, 0.88]	$\mathbf{0.083 \pm 0.006}$ [0.071, 0.095]	$\mathbf{0.87 \pm 0.01}$ [0.85, 0.90]	$\mathbf{0.56 \pm 0.02}$ [0.52, 0.61]

Table 9: **Randomized sanity checks.** RIFT degrades under randomized models and shuffled evidence while posterior entropy increases, confirming model-dependency.

Condition	Func. pred. ↑	Mismatch AUROC ↑	Mean entropy ↑	Top-k drop ↑
Trained model	**0.86**	**0.87**	0.24	**0.71**
Randomized weights	0.23	0.51	**0.62**	0.18
Shuffled preference evidence	0.58	0.60	0.48	0.49
Shuffled functional evidence	0.41	0.55	0.56	0.32
Shuffled flow edges	0.66	0.68	0.44	0.54

J.11 Failure Cases and Practical Limitations

We include representative failure cases to clarify the practical scope of RIFT. First, sparse probing can miss small hidden-reliance regions when they are not sampled by the candidate policy. Second, intervention responses can remain unstable when replacement features are far from the data manifold, even when multiple baselines are used. Third, semantic shortcut labels may be ambiguous when context outside a mask is genuinely task-relevant. Fourth, graph construction can affect flow-based diagnostics, particularly for transformer models, where attention-mediated connectivity and empirical flow may differ. These cases do not invalidate the RUF formulation, but they indicate that RIFT should be interpreted as a diagnostic framework whose conclusions depend on the chosen evidence sources, baselines, graph, and semantic reference.

K Extended Discussion

This section provides an extended discussion of the conceptual implications, assumptions, limitations, and practical scope of RIFT. The main manuscript presents RIFT as a framework for inferring Reliance Uncertainty Fields (RUF) from multiple evidence sources. Here, we discuss why this formulation is useful, what it does and does not claim, and how it should be interpreted in practice.

K.1 Why visual importance should not be collapsed into a single heatmap

A central motivation of this work is that visual importance is not a single quantity. Existing visual explanation methods are often consumed as if they answer the same question: *which visual evidence did the model use?* However, different methods expose different aspects of model behavior. Gradient-based explanations measure local sensitivity; feature-map-based explanations visualize class-discriminative representation prominence; perturbation-based explanations measure finite output response; semantic references describe human- or task-relevant regions. These signals can agree, but agreement is not guaranteed.

The four-way decomposition introduced in the main paper separates visual importance into *model-perceived*, *learning-influential*, *prediction-influential*, and *semantic* importance. This decomposition is not intended to introduce four unrelated explanation tasks. Rather, it clarifies why a single attribution map cannot generally serve as a complete explanation. A highlighted region may reflect the model's perceived preference without being functionally necessary for the prediction. Conversely, a region may be functionally important even if it is weakly highlighted by an attribution method. A prediction may also depend on evidence outside a semantic reference, indicating shortcut reliance or non-semantic contextual reliance.

Table 10: **Failure-mode detection.** RIFT achieves significant gains in detecting hidden reliance, highlighted-but-unused evidence, shortcut reliance, and flow attenuation.

Method	Hidden AUROC ↑	HBU AUROC ↑	Shortcut calib. ↓	Unc.–resid. ρ ↑	Atten. AUROC ↑
Grad-CAM	0.61	0.64	0.172	0.21	0.55
Score-CAM	0.64	0.67	0.158	0.24	0.58
Ablation-CAM	0.67	0.70	0.141	0.27	0.61
FI-only	0.76	0.72	0.133	0.34	0.60
RP+FI	0.82	0.80	0.112	0.42	0.68
RIFT w/o Flow	0.78	0.77	0.121	0.39	0.63
RIFT	**0.93**	**0.88**	**0.084**	**0.56**	**0.84**

Table 11: **Robustness to semantic-reference perturbations.** Functional predictiveness remains stable under moderate mask perturbations, while SemMass and ShortcutMass vary smoothly with reference quality.

Semantic reference	Func. pred. ↑	SemMass ↑	ShortcutMass ↓	RCE ↓
Clean mask	0.86	0.81	0.19	0.083
Erosion 10%	0.85	0.77	0.23	0.087
Dilation 10%	0.86	0.84	0.16	0.085
Boundary noise 10%	0.84	0.78	0.22	0.092
Random dropout 20%	0.83	0.74	0.26	0.097
No semantic reference	0.84	–	–	0.094

Table 12: **Projection-aware localization ablation.** Projection-aware variants improve localization accuracy and image-level calibration over naive bilinear resizing.

Projection method	Pointing game ↑	IoU@10% ↑	Img. calib. loss ↓	Loc. unc. ↓
Naive bilinear resize	0.62	0.31	0.184	0.41
Analytic RF projection	0.69	0.37	0.142	0.34
Effective RF projection	0.72	0.40	0.128	0.31
Calibrated ICF	**0.76**	**0.44**	**0.101**	**0.27**

Under this view, explanation failure is not merely a case in which one heatmap appears visually poor. It is a mismatch among different notions of importance. RIFT is designed to make such a mismatch measurable. Instead of asking whether a heatmap is visually plausible, RIFT asks whether the preference evidence agrees with intervention evidence, whether the signal survives through representation flow, whether the estimate is stable, and whether posterior reliance aligns with optional semantic reference.

K.2 RUF AS A POSTERIOR FIELD RATHER THAN A NEW ATTRIBUTION MAP

RUF should not be interpreted as simply another heatmap. The Reliance Uncertainty Field is a posterior distribution over latent functional reliance variables defined on a representation graph. Its posterior mean can be visualized as a reliance map, but the mean is only one readout of the field. The posterior entropy, evidence-specific disagreements, flow reliability, attenuation, redistribution, hidden reliance, highlighted-but-unused evidence, and shortcut reliance are all derived from the same posterior object.

This distinction is important. Many explanation methods directly output a deterministic score map. Such maps often lack an explicit account of uncertainty: if two evidence sources disagree, the final map may simply average or suppress the conflict. RUF, by contrast, treats disagreement as diagnostically meaningful. High uncertainty can indicate that preference evidence, intervention evidence, stability, and flow structure do not support a confident assignment of reliance. Similarly, preference–reliance disagreement can indicate that an explanation highlights a region that is not functionally used, or misses a region that contributes to the prediction.

The posterior formulation also clarifies the role of intervention. RIFT does not claim that intervention evidence alone is the final explanation. Intervention responses are one source of evidence for latent reliance. They can be noisy, baseline-dependent, and computationally expensive. By combining intervention evidence with preference, flow, stability, and optional semantic reference, RIFT aims to infer a more structured diagnostic object than a standalone perturbation map.

K.3 Why representation flow matters

A static explanation map cannot reveal how evidence evolves through the model. In deep visual networks, intermediate representations are transformed across layers. A signal that is strong at an early or middle layer may be amplified, suppressed, redistributed, recombined with other evidence, or ignored by the classifier head. Conversely, a signal that is weak in a selected explanation layer may become important downstream.

RIFT models this process through flow-based coupling. The flow term measures whether an intervention on an upstream representation unit changes downstream representations. However, flow alone is not equivalent to reliance. A representation perturbation may propagate through the network without affecting the target prediction, or it may affect downstream features that are not relevant to class c. Therefore, RIFT combines flow with posterior reliance: a node is considered flow-reliable when its propagated influence reaches downstream nodes with high posterior reliance.

This design allows RIFT to distinguish several mechanisms that are collapsed by static heatmaps:

- **Propagation**: a relied-upon signal affects downstream representations that also support the prediction.
- **Attenuation**: a locally highlighted or locally relied-upon signal weakens downstream and does not support the final prediction.
- **Redistribution**: an upstream signal spreads across many downstream units, making localization ambiguous.
- **Downstream rejection**: an intervention affects downstream representations but does not support class-specific reliance.
- **Emergent reliance**: a weakly highlighted signal becomes functionally important at later layers.

These mechanisms are difficult to observe from a single attribution map, but they are central to diagnosing why an explanation succeeds or fails.

K.4 Interpretation of semantic reference and shortcut reliance

RIFT uses semantic reference only as an optional external reference. A semantic mask, bounding box, part annotation, or concept label should not be interpreted as a measure of model importance. Models may legitimately rely on contextual cues, shape cues, texture cues, or background evidence depending on the task and training distribution. Therefore, high posterior reliance outside a semantic mask does not automatically imply that the model is wrong.

Instead, RIFT defines semantic and shortcut diagnostics relative to the provided reference. If posterior reliance is high inside the semantic reference, RIFT identifies semantic reliance. If posterior reliance is high outside the reference, RIFT identifies shortcut or non-semantic reliance. This diagnostic should be interpreted as a signal that the model may be using evidence not captured by the semantic annotation. Whether this is harmful depends on the application context. In medical imaging, autonomous perception, or scientific discovery, reliance outside the expected semantic region may require careful inspection. In other visual recognition settings, contextual reliance may be expected or even useful.

This treatment avoids a common pitfall in explanation evaluation: treating human annotations as complete ground-truth explanations of model behavior. RIFT instead asks whether the model's inferred reliance aligns with a chosen semantic reference and reports mismatch as diagnostic information.

K.5 Uncertainty is not merely visualization noise

The uncertainty produced by RIFT should be interpreted as uncertainty over latent functional reliance, not as aesthetic uncertainty in the heatmap. In RIFT, uncertainty can arise from multiple sources:

- disagreement between preference evidence and intervention evidence;
- high variance across baseline samples or perturbation choices;
- inconsistent signals across multiple explanation methods;
- weak or diffuse representation flow;
- ambiguity in projecting representation-level reliance to image space;
- conflict between posterior reliance and semantic reference.

Thus, high uncertainty is itself an informative diagnostic output. It indicates that the available evidence does not support a confident claim that a region is or is not functionally relied upon. In reliability-sensitive settings, such regions may warrant additional analysis rather than being treated as low- or high-importance.

K.6 Practical scope of RIFT

RIFT is designed as a diagnostic framework rather than a lightweight attribution method. It is most useful when the goal is to understand whether an explanation is faithful, where it fails, and how reliance evolves across representations. This makes it suitable for settings where explanation reliability matters more than producing a fast visualization, such as model auditing, scientific analysis, high-stakes vision systems, dataset bias analysis, and explanation-method evaluation.

RIFT is less suitable as a drop-in replacement for fast saliency methods across all deployment settings. Computing intervention evidence and flow evidence requires additional forward passes, access to intermediate activations, and design choices regarding baselines and graph construction. In applications where only a quick visual explanation is needed, standard attribution methods may be more practical. RIFT is intended for deeper reliability analysis, not necessarily for real-time explanation generation.

K.7 Computational limitations

The main practical limitation of RIFT is computational cost. Functional evidence requires intervening on the representation units and forwarding the downstream model. Flow estimation requires measuring the effects of these interventions on downstream representations. If all nodes and all edges are probed densely, the cost can become prohibitive for large models or high-resolution feature maps.

We address this through sparse probing. Instead of intervening on every node, RIFT can probe a candidate set containing high-preference regions, low-preference contrast regions, high-activation regions, random controls, and optional semantic regions. The inclusion of low-preference and random candidates is important: if RIFT probes only high-preference regions, it may inherit the bias of the explanation method under evaluation. Sparse RIFT can then interpolate or regularize evidence over nearby graph nodes and evaluate whether diagnostic labels are preserved relative to denser probing at lower-resolution layers.

Nevertheless, sparse probing introduces an approximation. It may miss small but functionally important regions if they are not sampled. It may also underestimate hidden reliance when low-preference candidates are insufficiently covered. Therefore, sparse candidate design is an important practical component of RIFT, and its effect should be evaluated empirically.

K.8 Dependence on baselines and intervention operators

Intervention-based evidence depends on the choice of baseline and perturbation operator. Replacing a feature vector with a zero vector, the dataset mean, a sampled feature from another image, or a

noisy local feature can yield different responses. Similarly, image-level interventions can introduce artifacts if the replacement distribution is unrealistic. This is a general limitation of perturbation-based explanation and faithfulness evaluation.

RIFT does not eliminate baseline dependence. Instead, it explicitly represents baseline sensitivity via evidence of stability. If an intervention response varies strongly across baselines, RIFT can assign higher uncertainty to the corresponding reliance estimate. This is preferable to reporting a single intervention score without acknowledging its instability. In practice, we recommend reporting sensitivity to baseline choice or using multiple baselines when computationally feasible.

K.9 Dependence on graph construction

RIFT requires a representation graph. For convolutional networks, graph edges can be defined using receptive-field reachability or empirically measured flow. For vision transformers, nodes may correspond to patch tokens, and edges may be defined by token identity, attention-mediated connectivity, or empirical intervention response. Each choice reflects a different assumption about how information propagates.

This means that RIFT's diagnostics can depend on graph construction. A sparse graph may miss long-range interactions, while a dense graph may introduce noisy or weakly meaningful couplings. Attention weights may not always correspond to causal information flow, while empirical flow can be more faithful but more expensive. Therefore, graph construction should be selected based on the model architecture and the diagnostic goal. The most conservative approach is to treat graph structure as part of the experimental design and evaluate ablations over alternative graph definitions.

K.10 Projection from representation space to image space

RUF is defined over representation units, while many visual explanations are consumed in image space. Projecting representation-level reliance onto image coordinates is inherently ambiguous because a low-resolution feature unit can correspond to a broad, overlapping receptive field. Naive upsampling can create visually sharp maps that suggest a degree of localization precision not supported by the representation.

For this reason, RIFT distinguishes representation-space reliance from image-space contribution. When image-space diagnostics are needed, RIFT uses receptive-field-aware or effective-receptive-field-aware projection, optionally calibrated by image-level intervention response. The resulting image-space map should be interpreted as a calibrated contribution estimate, not as a directly resized feature map. Localization uncertainty should be reported when projection ambiguity is high.

K.11 Relationship to existing explanation methods

RIFT is compatible with existing explanation methods. Grad-CAM, saliency, Integrated Gradients, attention rollout, Score-CAM, Ablation-CAM, and related methods can all provide evidence of preference. Perturbation and deletion metrics can provide evidence for interventions. Semantic annotations can provide external reference evidence. RIFT does not replace these methods; rather, it provides a probabilistic framework for combining and diagnosing them.

This perspective changes the role of existing explanations. Instead of asking which heatmap is the final explanation, RIFT asks how much each signal supports latent functional reliance and where signals disagree. A visually plausible heatmap may still be useful as preference evidence, even if it is not fully faithful on its own. Similarly, a perturbation score may be useful as functional evidence, even if it is unstable or incomplete. The posterior field provides a common space in which these signals can be compared.

K.12 What RIFT does not claim

RIFT makes several specific claims, but it also makes claims it does not make.

First, RIFT does not claim to recover a unique ground-truth explanation. Functional reliance is latent and evidence-dependent. The posterior field represents the inferred reliance under the chosen evidence sources, graph structure, baselines, and model-access assumptions.

Second, RIFT does not claim that semantic annotations define the correct explanation. Semantic reference is optional and used only to diagnose alignment or shortcut reliance relative to that reference.

Third, RIFT does not claim that intervention evidence is always causal in the strongest sense. Interventions may induce distribution shifts, and representation replacement may not correspond to a natural input manipulation. RIFT treats the intervention response as prediction-stage evidence and accounts for its instability by modeling uncertainty.

Fourth, RIFT does not claim to be as computationally lightweight as standard attribution methods. Its purpose is reliability diagnosis, not fast visualization.

Finally, RIFT does not claim to explain hidden reasoning or internal chain-of-thought. The sequential reliance trace is an observable intervention sequence showing how posterior reliance changes as evidence is removed or modified. It should not be interpreted as a literal reasoning process of the model.

K.13 Broader implications

The broader implication of RIFT is that the reliability of explanations should be studied as an inference problem rather than as a heatmap-ranking problem. A single visual explanation can be misleading because it compresses distinct notions of importance into a single spatial score. By introducing latent reliance variables and modeling multiple evidence sources, RIFT enables separation of preference, influence, propagation, uncertainty, and semantic alignment.

This view also suggests a shift in evaluation. Instead of evaluating explanations only by overlap with semantic masks or deletion curves, we can evaluate whether the highlighted evidence in an explanation is functionally used, whether reliance is hidden, whether reliance is stable under perturbation choices, whether reliance propagates through the model, and whether high-reliance regions align with semantic reference. These diagnostics are especially important for auditing models that may rely on shortcuts or spurious correlations.

K.14 Summary

The four-way decomposition of visual importance clarifies why explanation reliability cannot be reduced to the evaluation of heatmaps. Model-perceived, learning-influential, prediction-influential, and semantic importance may coincide in some cases, but they are not interchangeable. RUF formalizes this idea by representing latent functional reliance as a posterior field over the representation graph. RIFT operationalizes the posterior by combining preference, intervention, flow, stability, and optional semantic evidence.

The resulting framework is best understood as a process-level diagnostic tool. It identifies where explanations align with functional reliance, where they fail, and how failure arises through uncertainty, attenuation, redistribution, hidden reliance, highlighted-but-unused evidence, or shortcut reliance. Its main limitations are computational cost and dependence on choices such as baselines, intervention operators, graph construction, and semantic references. These limitations are not incidental; they reflect the fact that functional reliance is not directly observed and must be inferred from imperfect evidence.