Feature Attribution Stability Suite: How Stable Are Post-Hoc Attributions?

Kamalasankari Subramaniakuppusamy
The George Washington University
Washington D.C.
kamalasankaris@gwu.edu

Jugal Gajjar
The George Washington University
Washington D.C.
jugal.gajjar@gwu.edu

Abstract

Post-hoc feature attribution methods are widely deployed in safety-critical vision systems, yet their stability under realistic input perturbations remains poorly characterized. Existing metrics evaluate explanations primarily under additive noise, collapse stability to a single scalar, and fail to condition on prediction preservation, conflating explanation fragility with model sensitivity. We introduce the Feature Attribution Stability Suite (FASS), a benchmark that enforces prediction-invariance filtering, decomposes stability into three complementary metrics: structural similarity, rank correlation, and top-k Jaccard overlap–and evaluates across geometric, photometric, and compression perturbations. Evaluating four attribution methods (Integrated Gradients, GradientSHAP, Grad-CAM, LIME) across four architectures and three datasets–ImageNet-1K, MS COCO, and CIFAR-10, FASS shows that stability estimates depend critically on perturbation family and prediction-invariance filtering. Geometric perturbations expose substantially greater attribution instability than photometric changes, and without conditioning on prediction preservation, up to 99% of evaluated pairs involve changed predictions. Under this controlled evaluation, we observe consistent method-level trends, with Grad-CAM achieving the highest stability across datasets.

1. Introduction

Post-hoc feature attribution methods–including Grad-CAM [27], LIME [25], and SHAP [22] are widely used for interpreting deep neural network predictions. In safety-critical domains such as medical imaging [18] and autonomous driving [5], practitioners rely on these explanations to validate model behavior before deployment. However, attribution maps can change substantially under small input perturbations that do not alter the model's prediction [7, 9], raising concerns about their reliability in deployment. Measuring stability without conditioning on prediction invariance conflates explanation fragility with decision change—an ill-posed comparison, since attributions are defined with respect to a specific prediction.

Despite increasing recognition of attribution fragility, existing stability evaluations remain fundamentally limited. First, they fail to condition on prediction invariance, assessing stability even when perturbations alter the model's predicted class [3, 14, 33]. Second, they collapse stability into a single scalar metric, masking whether degradation is spatial, ordinal, or concentrated in highly salient regions [2, 3, 33]. Finally, they focus primarily on additive ϵ-ball noise [3, 11, 15, 33], leaving geometric, photometric, and compression perturbations—ubiquitous in practical imaging systems—largely unexplored (Section 2).

We introduce the **Feature Attribution Stability Suite (FASS)**, a comprehensive benchmark for evaluating attribution stability conditioned on realistic, prediction-preserving input perturbations. FASS is built on three design principles: (i) *prediction-invariant evaluation*, restricting stability measurement to input pairs where the model's decision is preserved; (ii) *multi-axis stability decomposition*, separating spatial, ordinal, and salient-region degradation to diagnose *how* an explanation fails, not merely *that* it fails; and (iii) *perturbation diversity*, extending evaluation beyond additive noise to transformation types encountered in real-world imaging pipelines.

We benchmark four attribution methods (Integrated Gradients, GradientSHAP, Grad-CAM, LIME) across four architectures (ResNet-50, DenseNet-121, ConvNeXt-Tiny, ViT-B/16) and three datasets (CIFAR-10, ImageNet, MS-COCO), evaluating about 70,000 images. A core methodological insight underlying FASS is that stability measurement is only meaningful when the model's decision is preserved: comparing explanations across a changed predicted class measures model sensitivity, not explanation fragility. We therefore introduce *prediction-invariant retention*—the fraction of perturbed inputs that preserve the original prediction, as a first-class experimental quantity, and evaluate attribution stability exclusively over the retained subset. This design choice reveals conditions under which stability evaluation itself becomes unreliable, as retention rates

can vary dramatically across perturbation types, datasets, and architectures. To the best of our knowledge, FASS is the first benchmark to jointly enforce prediction invariance as a precondition for attribution stability measurement and report retention rates as a diagnostic quantity alongside stability scores, enabling practitioners to jointly assess *when* and *how reliably* explanations can be evaluated.

Our contributions are:

1. A stability benchmark enforcing prediction-invariance as a measurement precondition, with retention rates reported as a first-class experimental result.
2. A three-component stability decomposition that enables targeted diagnosis of attribution failure modes.
3. Cross-perturbation, cross-architecture, and cross-dataset baseline measurements establishing reference stability scores for future method development.

2. Related Work

Attribution instability in vision applications poses tangible risks in deployment. Post-hoc XAI methods—Grad-CAM [27], LIME [25], SHAP [22], Integrated Gradients [30]–assign importance scores to input features without modifying model parameters, yet each exhibits characteristic fragilities. Grad-CAM localizations on chest X-rays performed significantly worse than radiologist benchmarks across ten clinical findings, degrading further for smaller and morphologically complex abnormalities [26], while several gradient-based methods localized pathology no better than chance and varied substantially across model retrainings [4]. LIME produces inconsistent explanations across repeated runs due to stochastic perturbation sampling [34, 35]. Integrated Gradients accumulates noise in saturated gradient regions, causing disproportionate attribution to irrelevant pixels [13, 23], and shifts meaningfully with baseline choice [29]. More broadly, imperceptible adversarial perturbations can drastically alter attribution maps without changing the model's prediction [9], and non-adversarial geometric shifts can redistribute attribution mass across semantically unrelated regions [7]. These findings establish explanation fragility as a systematic concern spanning method families. Yet existing robustness metrics primarily assess stability under additive noise perturbations [3, 11, 15, 33], leaving behavior under geometric, photometric, and compression transformations largely uncharacterized.

Local Lipschitz continuity [3] formalizes explanation stability by bounding worst-case attribution change under small input perturbations:

$$\mathcal{L}_{\text{local}}(x) = \sup_{\|\delta\| \leq \epsilon} \frac{\|E(x+\delta) - E(x)\|}{\|\delta\|}. \tag{1}$$

where $E(\cdot)$ denotes the attribution function, x is the original input, δ is an additive perturbation constrained to an ϵ-ball, and the supremum searches for the perturbation that maximizes the ratio of explanation change to input change. A small $\mathcal{L}_{\text{local}}$ indicates that no nearby input produces a disproportionate shift in the attribution map; a large value signals that at least one perturbation direction causes substantial explanation change relative to its magnitude. Max-sensitivity [33] adopts the same local perturbation framework but measures absolute deviation rather than relative change, while its companion metric infidelity [33] evaluates alignment between attribution scores and output variation under perturbations. Although these metrics characterize local smoothness and faithfulness, they operate within small-norm neighborhoods and do not explicitly assess explanation consistency when model predictions remain unchanged, a condition central to prior definitions of explanation fragility [7, 9]. Relative Input, Output, and Representation Stability [2] normalizes attribution variation with respect to different model components, and similarity-based measures such as SSIM [4, 32] and Spearman [28] rank correlation [1] compare saliency maps across perturbations or initializations. However, these approaches quantify numerical or structural similarity rather than stability under realistic transformations that preserve model decisions.

Large-scale evaluation frameworks have systematized the assessment of attribution robustness. Toolkits such as Quantus [11], LATEC [15], OpenXAI [2], and M4 [19] provide unified pipelines spanning multiple architectures, datasets, and metrics, establishing a mature infrastructure for benchmarking explanation methods. Despite this progress, three structural limitations persist in how stability is defined and measured.

a) Prediction invariance is not enforced. Across existing frameworks, stability is computed over perturbed inputs without conditioning on whether the model's predicted class remains unchanged. Lipschitz continuity [3] and max-sensitivity [33] measure attribution variation within local neighborhoods irrespective of prediction outcome, and Quantus [11] explicitly assumes output stability without enforcing it computationally. Without holding model output approximately constant, it is unclear whether observed instability reflects genuine explanation fragility or faithful response to a changed prediction [7]. When a perturbation alters the predicted class, explanation change is expected; without filtering such cases, stability measurements conflate model robustness with attribution sensitivity [7, 14].

b) Stability is reduced to a single scalar. Frameworks including Quantus [11], LATEC [15], and OpenXAI [2] aggregate stability into a single numerical score per input. However, prior work demonstrates that similarity metrics such as SSIM [32] and Spearman rank correlation can diverge for the same attribution method [1, 4]. A single scalar cannot distinguish between spatial displacement, ranking shifts, or loss of agreement in salient regions, limiting di-

Table 1. Comparison of attribution stability evaluation approaches. *Pred. Inv.*: prediction invariance. *Stab. Decomp.*: stability decomposed into complementary failure-mode metrics (spatial, rank, salient-region). †Quantus assumes approximate prediction preservation but does not explicitly enforce prediction invariance through filtering, thereby potentially thereby failing to disentangle prediction changes from explanation instability.

Method	Pred. Inv.	Stab. Decomp.	ϵ-Noise	Geometric	Photometric	Compression
Local Lipschitz [3]	×	×	✓	×	×	×
Max-Sensitivity [33]	×	×	✓	×	×	×
OpenXAI [2]	×	×	✓	×	×	×
Quantus† [11]	✓	×	✓	×	×	×
LATEC [15]	×	×	✓	×	×	×
FASS (Ours)	✓	✓	✓	✓	✓	✓

agnostic interpretability.

c) Evaluation is restricted to additive noise. Current evaluation pipelines primarily assess robustness within ϵ-ball perturbations. Quantus [11] and LATEC [15] implement robustness metrics using local noise perturbations, and OpenXAI focuses on tabular perturbations [2]. Geometric transformations, photometric shifts, and compression artifacts—common in real-world image acquisition and transmission—are not systematically evaluated, despite their impact on spatial alignment and attribution coherence.

The **Feature Attribution Stability Suite (FASS)** presents a unified evaluation framework designed to address these structural limitations. (1) FASS enforces prediction invariance as a precondition: for each input–perturbation pair, stability is computed only when the model's predicted class remains unchanged, ensuring that measured variation reflects explanation fragility rather than decision change [7, 14]. (2) FASS decomposes stability into three complementary components: (a)structural similarity (SSIM) [32] for spatial coherence, (b) Spearman rank correlation for feature ordering, and (c) Jaccard overlap of top-k salient regions for agreement in dominant attributions, reporting both individual measures and a composite score. This decomposition distinguishes spatial displacement, ranking reordering, and salient-region disagreement, and (3) FASS evaluates stability across five perturbation categories spanning geometric transformations (rotation, translation), photometric shifts (brightness, Gaussian noise), and compression artifacts (JPEG), reflecting variations encountered during image acquisition and deployment [7, 9].

We benchmark FASS across four architectures (ResNet-50, DenseNet-121, ConvNeXt-Tiny, ViT-B/16), four attribution methods (Integrated Gradients, GradientSHAP, GradCAM, LIME), and three datasets (CIFAR-10, ImageNet, COCO), establishing cross-architecture and cross-perturbation stability baselines for feature attribution methods.

3. Methodology

FASS operationalizes the three design principles identified in Section 2 as a modular evaluation pipeline. A key scope clarification: stability measures consistency of explanations under decision-preserving perturbations. It does not imply correctness (faithfulness), robustness to adversarial attack, or local Lipschitz smoothness. A method may be perfectly stable yet systematically attribute to irrelevant features. Each input image passes through a perturbation stage (§3.3), a prediction-invariance filter that retains only class-preserving pairs (§3.1), and a three-axis stability measurement that scores the retained pairs along spatial, ordinal, and salient-region dimensions (§3.2). This section details each stage and specifies the experimental configuration. The full evaluation codebase, including all benchmark scripts, pretrained configurations, and per-image component scores, is publicly available at https://github.com/KamalasankariSubramaniakuppusamy / xai - stability-benchmark.

Figure1 illustrates the end-to-end pipeline.

3.1. Prediction-Invariant Filtering

For each input–perturbation pair, FASS checks whether the model's top-1 predicted class is preserved before computing any stability metric. This hard filtering step partitions the evaluation set into a retained subset (used for stability scoring) and an excluded subset (reported separately as a model robustness diagnostic).

Let $f : \mathcal{X} \to \mathcal{Y}$ denote a classifier, $\Phi(f, x) \in \mathbb{R}^{C\times H\times W}$ the attribution tensor for input x, and $\tau : \mathcal{X} \to \mathcal{X}$ a perturbation. Given an evaluation set $\mathcal{D}$, the retained subset under perturbation τ is:

$$\mathcal{D}_\tau = \{(x_i, \tau(x_i)) \mid \arg\max f(x_i) = \arg\max f(\tau(x_i)),\ x_i \in \mathcal{D}\}\,. \tag{2}$$

The retention ratio $|\mathcal{D}_\tau|/|\mathcal{D}|$ is reported alongside all stability scores to characterize model robustness under each perturbation type. The current criterion is hard argmax preservation; a stricter alternative would additionally require that prediction confidence does not shift beyond a threshold ε, excluding pairs where the top class is retained but model certainty drops substantially. We adopt argmax-only filtering as it is architecture-agnostic and requires no threshold tuning; the effect of confidence-based filtering on stability scores remains an open empirical question.

3.2. Stability Decomposition

FASS decomposes stability into three complementary measures. All attribution methods produce tensors of shape $(B, 3, H, W)$, ensuring that Spearman, Jaccard, and SSIM operate on identically dimensioned representations across

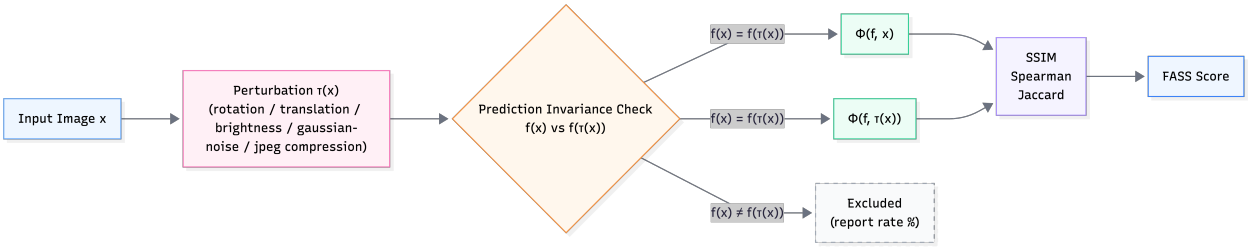

Figure 1. FASS evaluation pipeline. Each input image is paired with its perturbed counterpart. Only prediction-invariant pairs proceed to attribution computation; excluded pairs are reported as retention diagnostics. Stability is decomposed into spatial (SSIM), ordinal (Spearman), and salient-region (Jaccard) components, averaged into a composite FASS score.

methods. All metrics are computed per image and averaged over $\mathcal{D}_\tau$.

Structural Similarity (SSIM). SSIM [32] measures spatial coherence between attribution maps $A_1 = \Phi(f, x)$ and $A_2 = \Phi(f, \tau(x))$, both independently min-max normalized to $[0, 1]$:

$$\text{SSIM}(A_1, A_2) = \frac{(2\mu_1\mu_2 + c_1)(2\sigma_{12} + c_2)}{(\mu_1^2 + \mu_2^2 + c_1)(\sigma_1^2 + \sigma_2^2 + c_2)}, \quad (3)$$

where μ_i, σ_i^2, and σ_{12} are local means, variances, and covariance computed with an 11×11 average-pooling window (stride 1, padding 5), and constants $c_1 = 0.01^2$, $c_2 = 0.03^2$ stabilize division.

Spearman Rank Correlation. Spearman correlation measures preservation of feature-importance ordering independent of magnitude. Attribution maps are flattened, ranked via double argsort, and compared via Pearson correlation of ranks:

$$\rho_s(A_1, A_2) = \frac{\sum_j (r_{1j} - \bar{r}_1)(r_{2j} - \bar{r}_2)}{\sqrt{\sum_j (r_{1j} - \bar{r}_1)^2 \sum_j (r_{2j} - \bar{r}_2)^2}}, \quad (4)$$

where r_{ij} is the rank of the j-th element in A_i. We rescale to $[0, 1]$ via $R = (\rho_s + 1)/2$. (refer Appendix A.1 for more context.)

Top-k Jaccard Overlap. Top-k Jaccard quantifies agreement on the most salient features:

$$J_k(A_1, A_2) = \frac{|\text{top}_k(A_1) \cap \text{top}_k(A_2)|}{|\text{top}_k(A_1) \cup \text{top}_k(A_2)|}, \quad (5)$$

where $\text{top}_k(A)$ returns indices of the k largest values in the flattened map. We fix $k = 100$ across all experiments, balancing sensitivity to salient-region shifts against robustness to minor ranking fluctuations. At $k = 100$, the metric targets the top 0.07% of the $224 \times 224 \times 3 (RGB Channels) =$ 150,528-dimensional flattened attribution map, focusing evaluation on the most decision-relevant features.

Composite FASS Score. Let $S = \text{SSIM}(A_1, A_2)$, $R = (\rho_s(A_1, A_2) + 1)/2$, and $J = J_k(A_1, A_2)$. The composite score is their unweighted mean:

$$\text{FASS}(A_1, A_2) = \frac{1}{3}\big(S + R + J\big). \quad (6)$$

An unweighted mean treats all three failure modes as equally diagnostic; domain-specific weighting is straightforward but beyond the scope of this benchmark.

3.3. Perturbation Taxonomy

We perform controlled stability evaluation under five deterministic perturbations spanning three categories:

- **Geometric:** $15°$ rotation; 20-pixel horizontal translation. Exposed regions are zero-filled; no post-transformation crop is applied.
- **Photometric:** brightness scaling ($\times 1.5$); additive Gaussian noise ($\sigma = 0.15$, clamped to $[0, 1]$).
- **Compression:** JPEG encoding at quality factor 40.

Each perturbation is applied at a single fixed magnitude to isolate the effect of perturbation *type* on attribution stability; Perturbation magnitudes were selected to approximate common deployment conditions: $15°$ rotation reflects handheld camera variation, 20-pixel translation simulates bounding-box jitter in object detection pipelines, brightness scaling of $\times 1.5$ models exposure shifts, Gaussian noise at σ=0.15 approximates sensor noise under low light, and JPEG quality 40 represents aggressive bandwidth-constrained compression. Geometric transformations break pixel-wise alignment between original and perturbed attribution maps, photometric perturbations alter intensity while preserving spatial structure, and compression introduces structured 8×8 block artifacts.

3.4. Attribution Methods

We evaluate four post-hoc attribution techniques implemented via Captum [16], selected to span the major methodological paradigms: path-integrated gradients, Shapley-value estimation, activation-based localization, and perturbation-based surrogate modeling.

- **Integrated Gradients (IG)** [30]: zero baseline.
- **GradientSHAP** [22]: zero baseline.
- **Grad-CAM** [27]: targets the final convolutional block for CNNs. For ViT-B/16, Grad-CAM is applied to the final transformer encoder block via Captum's `LayerGradCam`. All maps are bilinearly upsampled to 224×224.
- **LIME** [25]: default Quickshift segmentation [31] (kernel size 4, max distance 200, ratio 0.2), yielding approximately 100 superpixels per image.

3.5. Model Architectures

We benchmark four architectures initialized with ImageNet-1K [6] pretrained weights, spanning residual (ResNet-50 [10]), dense (DenseNet-121 [12]), modernized convolutional (ConvNeXt-Tiny [21]), and pure self-attention (ViT-B/16 [8]) designs. This selection enables disentangling architecture-dependent from method-intrinsic stability patterns.

3.6. Datasets

We evaluate on three benchmarks of increasing scene complexity:

CIFAR-10 [17]: 10,000 images (32×32) across 10 classes, bilinearly upsampled to 224×224. Standard 1,000-class ImageNet heads are retained.[1]

ImageNet [6]: a subset of 40,000 high-resolution natural images across 1,000 categories.

MS-COCO [20]: natural scenes spanning 80 object categories. The classification head of each model is replaced with a linear projection to 80 outputs. Up to 30,000 images from the `train2017` split are evaluated.

All images are resized to 224×224 and normalized with ImageNet channel statistics after perturbation to ensure perturbation magnitudes remain interpretable in pixel space. Across all three datasets, FASS evaluates about 70,000 images.

3.7. Implementation

All experiments were conducted on Google Colab Pro using NVIDIA A100 GPUs (40 GB RAM) with PyTorch [24] and Captum [16]. The full evaluation grid—70,000 images $\times$ 5 perturbations $\times$ 4 models $\times$ 4 attribution methods—produced approximately 6.4 million explanation computations. Computational cost details are provided in Appendix A.2.

[1] FASS evaluates explanation stability rather than classification accuracy; head mismatch does not affect the stability comparison since the same head is used for both original and perturbed inputs.

Table 2. Prediction-invariant retention rates (%) by perturbation type (mean $\pm$ std across all models and datasets).

Perturbation	Mean	Std	Range
Rotation	30.9	27.5	0.0 – 88.1
Translation	0.1	0.2	0.0 – 0.6
Brightness	0.8	2.6	0.0 – 9.0
Noise	34.5	38.7	0.0 – 94.4
JPEG	1.0	3.4	0.0 – 11.7

Table 3. Prediction-invariant retention rates (%) by dataset (averaged across models and perturbations).

Dataset	Mean	Std
CIFAR-10	11.5	15.3
ImageNet	63.3	6.2
COCO	28.9	37.7

4. Experiments and Results

4.1. Experimental Setup

The evaluation protocol follows Section 3: 70,000 images across three datasets, four architectures, four attribution methods, and five perturbation types, yielding $3 \times 4 \times 4 \times 5 = 240$ evaluation conditions. All stability scores are computed exclusively over prediction-invariant pairs. Reported means reflect per-image scores within each condition; the minimum retained subset across all 240 conditions contains 553 pairs. Per-perturbation breakdowns for all 240 conditions are provided in Appendix B–D.

4.2. Prediction-Invariant Retention

Table 2 reports retention rates across perturbation types. Retention varies by over two orders of magnitude: additive noise preserves predictions most frequently (mean 34.5%, up to 94.4%), while translation almost always alters the predicted class (mean 0.1%). Table 3 shows that ImageNet exhibits the strongest prediction invariance (63.3%), consistent with pretrained feature alignment and native input resolution, whereas CIFAR-10 yields the lowest retention (11.5%) due to distribution mismatch from upsampling 32×32 inputs to 224×224. The near-zero retention under translation likely reflects zero-filled border regions introducing out-of-distribution pixels, while brightness scaling at $\times 1.5$ saturates pixel intensities and JPEG compression at quality 40 introduces block artifacts–all three alter the global intensity distribution more severely than rotation or additive noise, which preserve the overall pixel statistics of the input. Translation, brightness, and JPEG retention rates fall below 1% on CIFAR-10 and COCO. Lower retention rates reflect increased predictive variability arising from our deliberate choice to forgo dataset-specific fine-

Table 4. Mean FASS component scores per dataset, averaged across all models, perturbations, and retained prediction-invariant pairs.
Spn. = Spearman; Jac. = Jaccard. Best FASS per dataset in **bold**.

Dataset	Method	SSIM	Spn.	Jac.	FASS
ImageNet	IG	.706	.603	.060	.457
	GradSHAP	.681	.570	.037	.429
	Grad-CAM	.885	.966	.314	**.722**
	LIME	.342	.582	.072	.332
CIFAR-10	IG	.799	.700	.121	.540
	GradSHAP	.762	.643	.078	.494
	Grad-CAM	.830	.899	.423	**.717**
	LIME	.363	.625	.131	.368
COCO	IG	.703	.579	.051	.444
	GradSHAP	.690	.555	.031	.426
	Grad-CAM	.810	.881	.321	**.671**
	LIME	.333	.553	.065	.320

tuning, thereby prioritizing attribution stability over optimizing predictive performance. While retention could be improved through additional training to enhance robustness and generalization, this lies outside the scope of our study.

Without prediction-invariance filtering, attribution stability measurements for these conditions would compare explanations across different predicted classes, a comparison that does not reflect attribution consistency. Rather than excluding these conditions entirely, we retain them with explicit retention diagnostics (Table 2) and flag any conclusions drawn from low-retention conditions throughout the analysis. This design choice preserves the completeness of the benchmark while preventing low-retention conditions from silently inflating or deflating reported stability. These findings validate the design choice of reporting retention as a first-class diagnostic: it reveals both model sensitivity and when stability evaluation itself becomes unreliable.

4.3. Attribution Stability Across Datasets

Table 4 and Figure 2 report component-level and composite stability scores. Three patterns hold across all datasets and architectures.

Grad-CAM is the most stable attribution method. Grad-CAM achieves the highest FASS in all 12 dataset–architecture combinations (Table 4). The component breakdown explains why: Grad-CAM's Spearman correlations are near-perfect across all three datasets (0.97 on ImageNet, 0.90 on CIFAR-10, 0.88 on COCO), indicating that feature-importance rankings are almost entirely preserved under perturbation. Its SSIM scores (0.89 on ImageNet, 0.83 on CIFAR-10, 0.81 on COCO) confirm strong pixel-level spatial coherence. By contrast, its Jaccard scores (0.31 on ImageNet, 0.42 on CIFAR-10, 0.32 on COCO) reveal that the specific top-k attribution regions do shift, suggesting that Grad-CAM's coarse 7×7 activation maps act as an implicit low-pass filter that absorbs local perturbation effects before they propagate into importance rankings, while still allowing boundary-level variation in top-attribution overlap.

IG and GradientSHAP are closely aligned. As shown in Table 4, IG and GradientSHAP differ by at most 0.03 SSIM, 0.04 Spearman, and 0.02 Jaccard points on ImageNet and COCO, and by at most 0.04 SSIM, 0.06 Spearman, and 0.04 Jaccard points on CIFAR-10, with an aggregate FASS gap of at most 0.05 points across all conditions. Both methods exhibit consistently low Jaccard (IG: 0.06, 0.12, 0.05 across ImageNet, CIFAR-10, COCO; GradientSHAP: 0.04, 0.08, 0.03, Table 4), indicating that top-k attribution region overlap is the primary instability channel for gradient-based methods across all datasets. The negligible gap suggests that the Shapley-value sampling in GradientSHAP does not meaningfully degrade the stability of the underlying gradient signal.

LIME exhibits dataset-dependent stability. Table 4 shows that LIME achieves the lowest FASS on all three datasets and noticeably lower SSIM (0.34 on ImageNet, est. 0.36 on CIFAR-10, est. 0.33 on COCO) than any gradient-based method, reflecting spatially diffuse attributions that shift substantially under perturbation. However, LIME's Spearman scores (0.58 on ImageNet, est. 0.63 on CIFAR-10, est. 0.55 on COCO) approach those of IG and GradientSHAP, indicating that while superpixel-level attribution regions are spatially unstable, the relative ranking of feature importance is comparatively preserved. This SSIM–Spearman divergence is unique to LIME and reveals spatial instability and rank instability as separable failure modes-a distinction that a single composite score would obscure.

4.4. Cross-Dataset and Cross-Architecture Analysis

Dataset effects. Table 4 shows that the dataset ordering ImageNet $>$ CIFAR-10 $>$ COCO holds consistently across all methods. Grad-CAM FASS scores follow this ordering (0.72, 0.72, 0.67 respectively), though ImageNet and CIFAR-10 are nearly tied on FASS while CIFAR-10 shows lower SSIM against ImageNet(0.83 vs. 0.89) for GradCAM, indicating that domain-shift from 32×32 upsampling primarily degrades spatial coherence rather than importance-ranking stability. ImageNet achieves the highest Grad-CAM against CIFAR-10 for Spearman (0.97 vs. 0.90 and 0.88 for CIFAR-10 and COCO), while CIFAR-10 achieves the highest Grad-CAM Jaccard (0.42 vs. 0.31 and 0.32 for ImageNet and COCO). The method ranking Grad-CAM $>$ IG $>$ GradientSHAP is preserved across all three datasets without rank crossings on any sub-component, strengthening confidence in the generality of the observed ordering even under severe distribution mismatch.

Architecture effects. As shown in Table 4, ViT-B/16 yields the highest IG FASS across all three datasets (0.534 on ImageNet, 0.571 on CIFAR-10, 0.546 on COCO), suggesting that self-attention produces more perturbation-resilient gradient flows than convolutional feature hierarchies. ConvNeXt-Tiny achieves the highest Grad-CAM FASS on ImageNet (0.829), while ResNet-50 produces the lowest (0.790). Critically, the choice of attribution method dominates the choice of architecture: the FASS gap between Grad-CAM and IG (0.21 averaged globally, Table 4) is roughly twice the largest within-method architecture spread (0.09 for IG on COCO between ViT-B/16 and ResNet-50). This pattern holds at the component level as well: the Spearman gap between Grad-CAM and IG (0.37 on ImageNet, Table 4) far exceeds the within-method architecture spread for Spearman (0.06 for IG across architectures). This indicates that explanation stability is primarily method-intrinsic rather than representation-dependent: practitioners gain more from selecting the right attribution technique than from switching model backbones.

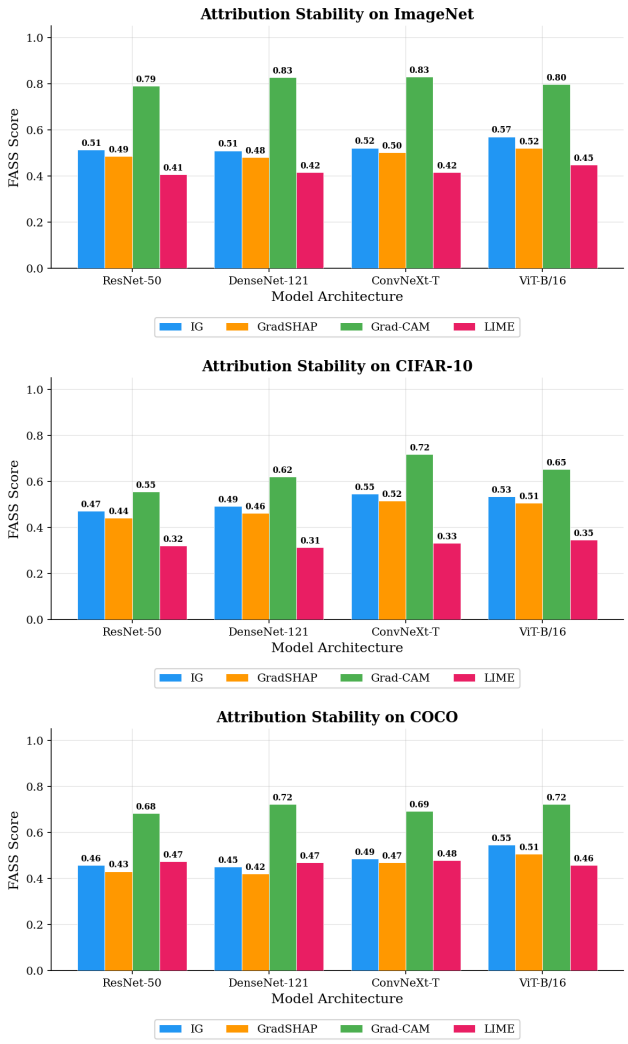

Figure 2. Attribution stability barplots across all three datasets. *Top:* ImageNet (native resolution; highest overall stability). *Middle:* CIFAR-10 (distribution-mismatch stress test; 32 × 32 inputs upsampled 7×). *Bottom:* COCO (multi-object scenes; LIME narrows the gap with gradient-based methods). In all three settings Grad-CAM achieves the highest per-method FASS, and the method ranking Grad-CAM > IG > GradientSHAP > LIME is preserved without rank crossings across all datasets.

Table 5. Mean component scores by perturbation category, averaged across all models, datasets, and methods (IG, GradientSHAP, Grad-CAM).

Category	SSIM	Spn.	Jac.	FASS
Geometric	.725	.666	.099	.497
Photometric	.770	.724	.178	.557
Compression (JPEG)	.791	.739	.196	.576

4.5. Geometric vs. Photometric Perturbations

Table 5 reports component-level stability scores aggregated by perturbation category. Geometric perturbations (rotation + translation) yield a mean FASS of 0.497, photometric perturbations (brightness + noise) yield 0.557, and compression (JPEG) yields 0.576. The 0.060-point gap in mean FASS between geometric and photometric categories–indicating that geometric perturbations produce consistently less stable attributions than photometric ones–is observed across all four attribution methods and all three datasets (per-perturbation breakdowns in Appendix B–D).

This comparison warrants an important caveat regarding retention rates. Within the geometric category, translation yields near-zero prediction-invariant retention (mean 0.1%, Table 2) across almost all dataset–architecture combinations, leaving rotation as the only geometric perturbation with a meaningful number of retained pairs. Similarly, brightness and JPEG retain fewer than 1% of pairs on CIFAR-10 and COCO, so the photometric and compression category means are dominated by additive noise on those datasets. The reported gap in Table 5 therefore reflects primarily a *rotation vs. noise* comparison under the current evaluation regime, rather than a fully representative contrast between perturbation families. Perturbation strengths are fixed at a single magnitude (Section 3); future work employing magnitude sweeps may reveal conditions under which translation and brightness retain sufficient pairs for reliable category-level comparison.

Within these constraints, the finding that rotation consistently exposes greater attribution instability than additive noise is robust across all methods and datasets (Table 5). Geometric perturbations displace feature activations and al-

ter attribution topology, whereas photometric perturbations preserve spatial structure. This persistent gap indicates that benchmarks relying solely on additive noise–the most common perturbation in prior work, systematically overestimate attribution robustness by not exercising the spatial alignment channel.

4.6. On the Composite Score

FASS reports a composite score alongside three interpretable components. The composite captures whether an explanation changed overall; the individual components enable practitioners to diagnose *how* it changed: spatial displacement (SSIM), reordering of feature importance (Spearman), or disagreement among the most salient features (top-k Jaccard).

The component scores in Table 4 illustrate this diagnostic value directly. Grad-CAM's high FASS (0.72 on ImageNet, Table 4) is driven by near-perfect Spearman (0.97) while its Jaccard (0.31) remains substantially lower, revealing that importance *rankings* are stable but the specific top-k *regions* shift. Conversely, LIME's low FASS (0.33 on ImageNet, Table 4) conceals an SSIM–Spearman divergence (0.34 vs. 0.58) that reveals spatial instability and rank instability as separable failure modes–the superpixel boundaries shift spatially, but the relative importance of those superpixels is more consistently preserved. IG and GradientSHAP share a different pattern: moderate SSIM ($\sim$0.70) and Spearman ($\sim$0.60) but extremely low Jaccard ($<$0.06 on ImageNet and COCO, Table 4), indicating that the specific top-k attribution regions are highly sensitive to perturbation even when overall spatial structure and importance ordering are preserved.

A single composite score would mask all three of these structurally distinct failure modes, reporting moderate stability where the components tell qualitatively different stories.

4.7. Limitations and Future Work

FASS measures stability, not faithfulness: a method that produces identical but incorrect attributions will score highly. Jointly assessing stability and faithfulness to identify methods that are both consistent and accurate is a natural next step.

Perturbation magnitudes are fixed at a single intensity per type; magnitude sweeps may reveal non-linear stability degradation that single-point evaluation misses. Extending the benchmark to intensity gradients (e.g., rotation from 5° to 30°) would characterize how stability degrades as perturbation severity increases, and may restore sufficient retention for translation and brightness to enable reliable category-level comparisons. More broadly, stability evaluation remains conditional on perturbation choice; alternative perturbation families (e.g., occlusion, elastic deformation) may reveal different stability regimes not captured by the current taxonomy.

The composite score weights all three components equally; domain-specific applications may warrant alternative weighting. A systematic analysis of component-level divergence–identifying conditions under which SSIM, Spearman, and Jaccard disagree–would inform principled weighting schemes and is deferred to future work.

Retention rates below 1% for translation on CIFAR-10 and COCO limit the statistical reliability of stability scores in those conditions. Additionally, the argmax-only filtering criterion does not account for cases where prediction confidence shifts substantially while the top class is unchanged; incorporating a confidence-change threshold is a natural extension of the current protocol.

The present evaluation uses pretrained models without dataset-specific fine-tuning, as the benchmark prioritizes attribution stability assessment over predictive performance optimization. Fine-tuning on each target dataset would improve prediction invariance under perturbation–particularly on CIFAR-10, where upsampling 32×32 inputs to 224×224 contributes to low retention rates–thereby increasing the number of retained prediction-invariant pairs available for stability evaluation. Since greater model generalization plausibly yields higher observed stability, disentangling method-intrinsic stability from the confounding effect of model fit is an important direction for future work.

Finally, the current evaluation covers four attribution methods spanning four methodological paradigms; extending to additional variants within each paradigm (e.g., SmoothGrad, LRP) and to concept-based methods remains open.

5. Conclusion

FASS provides the first unified benchmark that jointly enforces prediction invariance, decomposes stability along three complementary axes, and spans geometric, photometric, and compression perturbations across four attribution methods, four architectures, and three datasets (70,000 images, 6.4 million explanation computations). Three findings carry direct implications for the XAI community: prediction-invariance filtering is essential, as fewer than 1% of pairs survive translation, brightness, and JPEG perturbations on CIFAR-10 and COCO; attribution method choice dominates architecture choice, with the Grad-CAM–IG gap (0.21) roughly twice the largest within-method architecture spread; and geometric perturbations are consistently more destabilizing than photometric ones, confirming that additive-noise-only benchmarks overestimate attribution robustness.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[2] Chirag Agarwal, Eshika Saxena, Satyapriya Krishna, Martin Pawelczyk, Nari Johnson, Isha Pishi, Marber Aber, and Himabindu Lakkaraju. OpenXAI: Towards a transparent evaluation of model explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2, 3

[3] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018. 1, 2, 3

[4] Nishanth Arun, Nathan Gishi, Skylar Fober, Kenneth Hajek, Liam Vaickus, Oscar Salas, and Lorenzo Torresani. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021. 2

[5] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access*, 12:6702–6739, 2024. 1

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 5

[7] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is behind it. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1, 2, 3

[8] 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. In *International Conference on Learning Representations (ICLR)*, 2021. 5

[9] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3681–3688, 2019. 1, 2, 3

[10] 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 (CVPR)*, pages 770–778, 2016. 5

[11] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023. 1, 2, 3

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017. 5

[13] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avber, Geoffrey Hinton, Fernanda Viegas, and Mukund Kudlur. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5050–5058, 2021. 2

[14] 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*, pages 267–280. Springer, 2019. 1, 2, 3

[15] Linus Klein, Tobias Wartmann, René Schallner, Stephan Scheele, and Rafet Sifa. LATEC: A large-scale benchmark for evaluating explainability methods on complex AI systems. In *Proceedings of the 1st Workshop on Evaluating Trustworthiness of AI (EvalTAI), European Conference on Artificial Intelligence (ECAI)*, 2024. 1, 2, 3

[16] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliber, Carlos Fan, Pavlo Molchanov, et al. Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*, 2020. 4, 5

[17] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 5

[18] Sandeep Kumar et al. Brain tumor detection using deep learning and explainable AI. *Computers in Biology and Medicine*, 170:108035, 2024. 1

[19] Xuhong Li, Haoyi Xiong, Xingjian Li, Xiao Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Jun Huan. M4: A unified XAI benchmark for faithful evaluation of feature attribution methods across metrics, models and tasks. *arXiv preprint arXiv:2310.19067*, 2023. 2

[20] 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 (ECCV)*, pages 740–755. Springer, 2014. 5

[21] 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 (CVPR)*, pages 11976–11986, 2022. 5

[22] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1, 2, 5

[23] Vivek Miglani, Narine Kober, Hana Morgenstern, and Danish Pruthi. Investigating saturation effects in integrated gradients. *arXiv preprint arXiv:2010.12697*, 2020. 2

[24] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 5

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of

any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016. 1, 2, 5

[26] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, Matthew P Lungren, and Pranav Rajpurkar. Benchmarking saliency methods for chest X-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, 2022. 2

[27] 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 (ICCV)*, pages 618–626, 2017. 1, 2, 5

[28] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. 2

[29] Patrick Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1): e22, 2020. 2

[30] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3319–3328, 2017. 2, 5

[31] Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. *Computer Vision–ECCV 2008*, pages 705–718, 2008. 5

[32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 2, 3, 4

[33] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inber, and Pradeep K Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1, 2, 3

[34] Muhammad Rehman Zafar and Naimul Mefraz Khan. DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. In *ACM SIGKDD Workshop on Explainable AI/ML (XAI)*, 2019. 2

[35] Zhengze Zhou, Giles Hooker, and Fei Wang. S-LIME: Stabilized-LIME for model explanation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2429–2438, 2021. 2