Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Systematic Vulnerability Audit of Post-Hoc XAI under Common Corruptions: A Factor Analysis Across Vision Benchmarks

Minyeong Kim

win198@naver.com

Research Article

Keywords: Explainable Artificial Intelligence, Attribution Stability, Distribution Shift, Common Corruptions, Vision Transformer, Model-Agnostic Explanation

Posted Date: April 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9501010/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Systematic Vulnerability Audit of Post-Hoc XAI under Common Corruptions: A Factor Analysis Across Vision Benchmarks

Minyeong Kim[1*]

[1]Independent Researcher, Gyeonggi-do, South Korea.

Corresponding author(s). E-mail(s): win198@naver.com;

Abstract

Post-hoc attribution methods are widely deployed to explain deep vision classifiers, yet no systematic evaluation protocol exists for the corrupted-input regime. This paper introduces the first such protocol for attribution stability under distribution shift, validated through a factorial audit of five attribution methods (Integrated Gradients, Grad-CAM, SmoothGrad, GradientSHAP, LIME) under fifteen corruptions at five severity levels on CIFAR-10-C and CIFAR-100-C, across two architecturally distinct classifiers (ResNet-50, ViT-B/16), yielding 760,000 clean–corrupted pairs and five stability metrics. Four findings validate the protocol. First, attribution stability declines monotonically with corruption severity for all methods, with 12 to 13 of 15 corruption types reaching significance under Benjamini–Hochberg correction and photometric corruptions exhibiting the steepest degradation (brightness Spearman: 0.87 at severity 1 to 0.31 at severity 5). Second, degradation magnitude depends dramatically on the method: SmoothGrad retains Spearman 0.91 for brightness at severity 3 while LIME falls to 0.04, a twenty-fold gap. Third, the resolution-fair ranking (SmoothGrad, IG, GradientSHAP) is consistent across 96 % of cells (144/150), with method $\boldsymbol{\eta^2 > 0.84}$ dwarfing architecture $\boldsymbol{\eta^2 < 0.02}$. Fourth, the architecture-by-method interaction is non-significant on CIFAR-10 ($\boldsymbol{p = 0.71}$) but significant on CIFAR-100 ($\boldsymbol{p = 0.035}$), showing task-complexity modulation of the model-agnostic claim. We contribute three artifacts: (i) the protocol as an evaluation standard, (ii) an empirical method characterisation, and (iii) a deployment-ready ranking; we show the model-agnostic assumption requires re-examination as classification complexity increases.

Keywords: Explainable Artificial Intelligence, Attribution Stability, Distribution Shift, Common Corruptions, Vision Transformer, Model-Agnostic Explanation

1 Introduction

Explainable artificial intelligence has become a prerequisite for the deployment of machine-learning classifiers in settings where the reasoning behind a decision matters — medical diagnosis, credit scoring, fraud detection, and legal case triage — and the post-hoc attribution methods that now populate this space are routinely evaluated and compared on benchmark datasets under clean, in-distribution conditions. A critical question, however, is left almost entirely unaddressed by the existing evaluation literature: when the input that a deployed classifier encounters in the real world is not a clean benchmark sample but a noisy, blurred, weather-corrupted, or compression-artefacted image, do the attribution methods that passed the clean benchmark continue to produce stable explanations, and which of them fail under which corruptions? The present study answers this question through the first systematic evaluation protocol for attribution stability under distribution shift, validated on a factorial design that crosses five attribution methods, two architecturally distinct classifiers, fifteen corruption types, and five severity levels on the CIFAR-10-C and CIFAR-100-C benchmarks, and it operationalises the protocol into a deployment-ready vulnerability ranking that identifies which explanation methods can be trusted under which corruption conditions.

The gap that motivates this audit is easy to state in one sentence. Almost every comparative evaluation of post-hoc explanation methods published in the past three years has operated on clean inputs. The unified faithfulness benchmark M4 of Li et al. [33] evaluates feature attribution across modalities and architectures on clean ImageNet and clean text; the XAI-Units framework of Rozanec et al. [48] conducts unit tests on procedurally generated synthetic datasets; and the recent comparative evaluation of Moradi et al. [40], published in Nature Scientific Reports, compared six widely used attribution techniques on CIFAR-10, Imagenette, and SVHN and reported that Integrated Gradients and SmoothGrad achieved the highest fidelity and stability while LIME performed worst — a result established entirely on clean images. The one recent study that has directly addressed explanation behaviour under distribution shift, the uncertainty-aware subset-selection framework of Gupta et al. [20], is a method proposal: the authors observed that subset-selection explanations can degrade by forty per cent under out-of-distribution conditions and responded by introducing a new explanation algorithm that is robust by design. Gupta et al. answered the question of how to fix the problem; no study to date has answered the complementary and arguably more fundamental question of how bad the problem actually is, across which attribution methods, for which corruption types, under which model architectures, and with which practical consequences.

This audit is positioned explicitly to fill that gap. Rather than proposing a new explanation algorithm, we treat five widely deployed attribution methods — Integrated Gradients [55], Grad-CAM [49], SmoothGrad [54], GradientSHAP [17, 34], and LIME [46] — as fixed objects of measurement and quantify how their outputs shift when input distributions move away from the clean benchmark on which the underlying classifiers were trained. We apply the common-corruption suite of Hendrycks and Dietterich [23] in full, covering fifteen corruption types at five severity levels, and we replicate the entire grid on two architecturally distinct classifiers (a convolutional ResNet-50 [21] and a self-attention-based Vision Transformer ViT-B/16 [16])

so that architecture-dependent and architecture-invariant patterns can be separated empirically. The resulting factor grid produces more than 760,000 clean–corrupted attribution pairs, each paired with five stability metrics that extend the disagreement framework of Krishna et al. [30] from the method-versus-method comparison on clean data to the clean-versus-corrupted comparison under distribution shift.

The contributions of this paper are organised as a triple: a protocol, its empirical characterisation, and a deployment artifact. **(C1) Protocol.** We introduce the first systematic evaluation protocol for attribution stability under distribution shift, defined by a factorial design (five attribution methods, two architecturally distinct classifiers, two datasets, fifteen corruption types, five severity levels), a five-metric stability battery extending the disagreement-problem formalism of Krishna et al. [30] from method-versus-method on clean data to clean-versus-corrupted under shift, and a pre-registered two-way factorial analysis of variance augmented with Fisher-z, Benjamini–Yekutieli, ω^2, and TOST sensitivity checks; the protocol is reproducible from the public package at https://github.com/minyeong5691-svg/paper2-xai-corruption-audit. **(C2) Empirical characterisation.** We validate the protocol on over 760,000 clean–corrupted attribution pairs and use it to test four hypotheses: the protocol confirms monotonic severity-driven decline, a twenty-fold method-level gap between SmoothGrad and LIME under identical conditions, a ninety-six-percent-consistent resolution-fair robustness ranking with method effect size ($\eta^2 > 0.84$) dwarfing architecture ($\eta^2 < 0.02$), and a task-complexity boundary on the model-agnostic claim of Shapley-value attribution ($p = 0.71$ on CIFAR-10, $p = 0.035$ on CIFAR-100, strengthened to $p = 0.0014$ on the Fisher-z scale). **(C3) Deployment artifact.** We derive a deployment-ready vulnerability ranking — a set of conditional recommendations of the form "when facing corruption X, prefer attribution method Z" — that is consistent across ninety-six percent of all tested conditions and operationalises the protocol for practitioners who must choose an attribution method in anticipation of input corruption.

The remainder of the paper is organised as follows. Section 2 reviews the comparative evaluation literature on post-hoc attribution and the narrower body of work on explanation behaviour under distribution shift, and positions the present audit against the four closest predecessors. Section 3 lays out the theoretical framework — Shapley-value axioms, off-manifold extrapolation, Pearl's causal hierarchy, and the four hypotheses tested in the Results section. Section 4 details the factor grid, classifiers, attribution methods, stability metrics, and statistical-analysis plan. Section 5 reports the empirical findings, organised around a single-method IG baseline, a five-method cross-dataset comparison, a four-hypothesis verification summary, and a ResNet-versus-ViT architecture comparison. Section 6 interprets the findings, delineates seven limitations, and derives the practical vulnerability ranking that constitutes the main deployment-facing contribution. Section 7 summarises the principal conclusions and outlines five directions for future work.

2 Related Work

2.1 Comparative Studies of Post-Hoc XAI on Clean Data

A substantial body of empirical work has characterised how post-hoc attribution methods behave on standard classification benchmarks, almost always under clean input conditions. Moradi et al. [40] conducted the most recent and direct predecessor to the present audit: their comparative evaluation applied six widely used attribution techniques — LIME, SHAP, Grad-CAM, Grad-CAM++, Integrated Gradients, and SmoothGrad — to three deep-learning architectures across the CIFAR-10, Imagenette, and SVHN benchmarks, using five function-grounded metrics (fidelity, stability, identity, separability, and computational time). Their principal finding was that no single method dominates every metric, with the gradient-based Integrated Gradients and SmoothGrad achieving the highest fidelity and stability scores while KernelSHAP ranked third in stability and LIME ranked lowest. These results provide the clean-data baseline against which any OOD extension must now be measured, and the present study treats the Moradi et al. hierarchy as an empirical null hypothesis that the corrupted-input regime may or may not preserve.

Earlier work in the same vein established the methodological precedent for cross-method benchmarking. Doshi-Velez and Kim [15] proposed a three-tier evaluation taxonomy (application-grounded, human-grounded, functionally-grounded) that continues to frame how comparative XAI studies are designed, and Nauta et al. [42] surveyed 361 XAI evaluation papers published between 2014 and 2020 and documented that only 33 per cent used quantitative evaluation on real data while 58 per cent relied on anecdotal visual inspection. The shift towards factorial, quantitative designs of the kind pursued here directly addresses this identified deficit.

2.2 XAI Methods under Distribution Shift

Comparatively little empirical work has examined what happens to post-hoc attribution methods when the input distribution shifts. The closest direct competitor to the present study is Gupta et al. [20], whose uncertainty-aware subset-selection framework, released on arXiv in December 2025, demonstrates that subset-selection-based explanation methods can degrade their insertion and deletion scores by as much as 40 % under out-of-distribution conditions. Their response was to propose a new algorithm that combines submodular subset selection with layer-wise gradient-based uncertainty estimation, closing much of the OOD robustness gap without requiring retraining. The contribution of Gupta et al. is therefore a methodological proposal: a new explanation algorithm that is robust to distribution shift by design. The present study, in contrast, is diagnostic rather than prescriptive: we do not propose a new algorithm, we characterise how an existing set of widely deployed algorithms behaves across a much larger factor grid, and we derive practical guidance from the resulting map of vulnerabilities. These two contributions are complementary, and the present empirical audit provides the kind of fine-grained diagnostic that a method proposal such as Gupta et al.'s can be measured against.

2.3 The Disagreement Problem and XAI Stability

A parallel line of work has studied the phenomenon of disagreement between post-hoc explanation methods on the same input. Krishna et al. [30] formalised this as the "disagreement problem" and introduced six metrics — feature agreement, rank agreement, sign agreement, signed rank agreement, rank correlation, and pairwise rank agreement — that together characterise the extent to which two attribution methods agree on which features matter and in which order. Their empirical study covered four datasets (COMPAS, German Credit, AG News, and ImageNet-1k with PASCAL VOC labels) spanning three data modalities: tabular, text, and image. Notably, their study was conducted entirely on clean data, and the authors explicitly did not consider how disagreement behaves when inputs are perturbed. The present audit adopts the Krishna et al. metric family directly, applies it to the clean-versus-corrupted comparison rather than the method-versus-method comparison, and thereby extends the disagreement framework into the stability-under-shift setting. Neely et al. [43] subsequently confirmed that disagreement rankings depend on the choice of metric, motivating our use of five complementary metrics rather than any single aggregate score.

2.4 Attribution Robustness and Faithfulness Metrics

The broader question of how attribution methods should be evaluated has generated a mature body of work. Hedström et al. [22] proposed the Quantus toolkit with a six-category evaluation taxonomy (faithfulness, robustness, localisation, complexity, randomisation, axiomatic) that has become a de facto standard. Our five metrics fall within the *robustness* category, which measures how stable explanations are under input perturbation. We deliberately do not evaluate the *faithfulness* category because fidelity metrics themselves are contested: Miró-Nicolau et al. [38] showed that six widely used faithfulness metrics disagree on up to 30 % of their rankings on identical data, and Rong et al. [47] demonstrated that standard insertion/deletion metrics are sensitive to the choice of imputation strategy and can produce spurious rankings. Agarwal et al. [3] argued that relative stability metrics, which normalise explanation change by input change to produce a Lipschitz-style bound, provide a cleaner alternative to absolute stability metrics. Our metric family is absolute rather than relative, but the relative–absolute choice is orthogonal to the clean-versus-corrupted extension that is this paper's primary contribution, and we return to the absolute/relative trade-off in the limitations section of §6.

2.5 Out-of-Distribution Robustness Benchmarks

The OOD robustness literature for classifiers themselves is mature and provides the benchmark infrastructure that this audit inherits. Hendrycks and Dietterich [23] introduced the CIFAR-10-C, CIFAR-100-C, and ImageNet-C corruption suites, covering fifteen corruption types (noise, blur, weather, digital) at five severity levels each, and established that state-of-the-art classifiers lose between 20 and 80 per cent of their clean accuracy under severe corruption. Koh et al. [29] broadened the OOD testing scope in the WILDS benchmark to include domain shifts from real-world applications (medical imaging, satellite imagery, text). Wiles et al. [60] provided a fine-grained

taxonomy of distribution shifts and documented that different shift types produce qualitatively different robustness profiles. The present study restricts itself to the Hendrycks common-corruption suite because its controlled parametric structure (fifteen types × five severities) permits factorial ANOVA, whereas real-world shifts of the WILDS type are less amenable to parametric comparison. We adopt the Hendrycks suite not as a substitute for real-world OOD data but as a reproducible benchmark whose effects on attribution stability we characterise for the first time.

2.6 Cross-Architecture Attribution and the Model-Agnostic Claim

A particular strand of the XAI literature concerns whether explanation methods behave consistently across fundamentally different architectures. Lundberg and Lee [34] argued that Shapley-value attribution is *model-agnostic*: its axiomatic guarantees (efficiency, symmetry, dummy, linearity) are properties of the cooperative game defined by the model's prediction function and therefore transfer across convolutional, recurrent, and attention-based classifiers. This claim has been tested only indirectly. Chefer et al. [12] showed that gradient-based attribution methods require architecture-specific adaptations for Vision Transformers to produce spatially coherent explanations, and Abnar and Zuidema [2] proposed attention rollout specifically for transformer interpretability. The present study provides the first systematic test of whether the model-agnostic claim survives the transition from clean to corrupted inputs, using a two-way ANOVA with architecture (ResNet-50 versus ViT-B/16) and method (IG, SmoothGrad, GradientSHAP) as fixed factors to quantify the architecture-by-method interaction.

2.7 Theoretical Limits of Post-Hoc Explanation

Several theoretical analyses published in 2022–2024 have cautioned against treating post-hoc explanations as unconditional guarantees. Huang and Marques-Silva [25] contributed a mathematical analysis of Shapley-value explanations, showing that SHAP can produce misleading feature rankings even in small feature spaces (six or more features) and identifying axiomatic limits on the interpretability guarantees that Shapley-based methods can provide. Molnar et al. [39] catalogued seven general pitfalls of interpretable machine learning, including feature dependence, feature interaction, extrapolation, causal misinterpretation, uncertainty neglect, and scale dependence, several of which map directly onto the failure modes that OOD shift is likely to amplify. Pearl [45] positioned all associational attribution methods (including SHAP, IG, and LIME) at Level One of his three-rung causal hierarchy, a boundary condition we make explicit in §3. Janzing et al. [26] analysed feature-relevance methods as causal problems and demonstrated that marginal (as opposed to conditional) reference distributions produce off-manifold extrapolations that can invert feature rankings when inputs shift away from the reference. Sixt et al. [52] showed that most modified back-propagation methods produce explanations independent of later-layer parameters, a limitation we avoid by excluding such methods from our audit. Slack et al. [53] demonstrated that perturbation-based methods (LIME, SHAP) are vulnerable to adversarial scaffolding

attacks that can be constructed to hide discriminatory behaviour. Collectively these theoretical results establish the conceptual frame within which our empirical findings should be interpreted: the observed instability under OOD corruption is not an incidental defect but a predictable consequence of the reference-distribution mechanism underlying perturbation-based and path-integral attribution methods.

2.8 Positioning of the Present Study

Against the four most recent and closely related benchmarks, the present audit occupies a distinct position. The unified faithfulness benchmark M4 of Li et al. [33] evaluates feature attribution across modalities and architectures on clean ImageNet and clean text; we differ along the *metric axis* (stability rather than faithfulness) and the *data axis* (common corruptions rather than clean ImageNet). The XAI-Units framework of Rozanec et al. [48] operates on synthetic procedurally generated models designed to admit known ground-truth explanations; we differ along the *model axis* (real pretrained classifiers on canonical benchmarks rather than synthetic models). The comparative evaluation of Moradi et al. [40] (Nature Scientific Reports) operates on clean CIFAR-10 and related benchmarks; we differ along the *input-regime axis* by extending their stability metric into the corrupted-input regime. The uncertainty-aware subset-selection framework of Gupta et al. [20] is a *method proposal* that mitigates OOD degradation through a new algorithm; we are a *benchmark characterisation* that quantifies how widely deployed existing algorithms degrade under OOD shift. These four axes — metric, data, model, contribution type — collectively define the empirical gap that this audit fills, and they motivate the factor grid and metric family detailed in §4.

3 Theoretical Framework

3.1 Shapley Values as a Principled Basis for Feature Attribution

The Shapley value, originating in cooperative game theory [51], uniquely satisfies four axioms — efficiency, symmetry, dummy, and linearity — among all fair allocation rules for the collective payoff of a coalition. Lundberg and Lee [34] recast the feature attribution problem as a cooperative game, with the classifier's prediction function in the role of the characteristic function and individual features in the role of cooperating agents, and showed that the Shapley value then provides the unique attribution rule satisfying the same four axioms in the attribution context. GradientSHAP, the implementation evaluated in this audit, approximates Shapley values by integrating gradients over a background reference distribution drawn from the clean training data; the integration substitutes for the exhaustive coalition enumeration that would otherwise be exponential in the number of features. Critically, the axiomatic guarantee is conditioned on the reference distribution being representative of the data on which the model will be deployed. When the deployment input is drawn from a distribution different from the reference (as happens under common corruption), the axiomatic guarantee loses

its force: the resulting attribution remains mathematically well-defined but no longer enjoys the interpretability warranties that motivated its use in the first place.

3.2 Off-Manifold Extrapolation under Distribution Shift

The mechanism by which distribution shift undermines Shapley-based attribution is the off-manifold extrapolation problem, formalised by Janzing, Minorics and Blöbaum [26]. When Shapley values are computed by marginalising features against the training distribution (the standard "KernelSHAP" approach as opposed to the conditional variant), the hypothetical inputs assembled during the coalition enumeration can lie far outside the support of the data the model was trained on. The classifier evaluates these off-manifold inputs through whatever extrapolation behaviour its parameters happen to implement, and the resulting attribution reflects that extrapolation rather than any property of the clean training distribution. Under OOD corruption the problem is compounded because the deployment input itself has moved off the training manifold, while the reference distribution remains on it. The marginal distance between the two then enters the attribution directly, producing the near-constant, near-zero stability signature that we observe empirically for GradientSHAP in §5.2. Slack et al. [53] provided an adversarial illustration of the same mechanism: perturbation-based explainers can be systematically fooled by scaffolded models that exploit the off-manifold vulnerability to hide discriminatory behaviour.

3.3 Causal Level Constraint

All five attribution methods evaluated in this audit operate at Level One of Pearl's [45] three-rung causal hierarchy. Level One (Association) captures statistical dependence between observables, answering questions of the form "what is $P(Y \mid X)$?"; it does not license interventional claims ("what is $P(Y \mid \mathrm{do}(X))$?") or counterfactual claims ("would Y have been different had X been different, holding all else equal?"). This constraint is important for the interpretation of our findings in two ways. First, when we report that SmoothGrad assigns high attribution weight to certain pixels, we do not claim that intervening on those pixels would change the classifier's decision; we claim only that the observed classifier output co-varies with those pixels under the particular reference distribution used. Second, when the observed attribution shifts under OOD corruption, we do not claim that the causal structure of the classifier has changed; the classifier's weights are fixed, and what has changed is the statistical relationship between inputs and attributions under a different input distribution. These distinctions preserve the interpretability of the audit while avoiding over-claiming about causal mechanisms [10].

3.4 Hypothesis Development

The empirical audit is structured around four hypotheses derived from the preceding theoretical framework.

H1 (Monotonic degradation).

Attribution stability, measured as the Spearman rank correlation between the clean and corrupted attribution vectors, declines monotonically with corruption severity. This prediction follows from the combined mechanism of classifier accuracy degradation (as severity increases, the underlying model produces less confident and less consistent predictions) and off-manifold extrapolation (as severity increases, the corrupted input moves further from the background reference set used by perturbation-based methods). H1 is expected to be supported for every combination of dataset, model, and attribution method, with the rate of degradation serving as the primary quantitative outcome of the audit.

H2 (Corruption–method interaction).

The rate at which stability degrades under corruption depends on an interaction between the corruption type and the attribution method. Corruptions that preserve local image structure (brightness, contrast, fog, defocus blur) are expected to leave gradient-based attributions (Integrated Gradients, SmoothGrad) relatively unaffected because the local linear neighbourhood of the classifier is approximately preserved. In contrast, corruptions that destroy local structure (motion blur, elastic transform) are expected to disrupt the superpixel segmentation that LIME relies on and to introduce large marginal-extrapolation errors into GradientSHAP. H2 is tested through the corruption-by-method interaction term in a two-way ANOVA on Spearman correlation.

H3 (Architecture dependence of the model-agnostic claim).

Lundberg and Lee [34] argued that SHAP-based attribution is model-agnostic in the sense that the axiomatic guarantees transfer across classifier architectures. Under clean conditions this claim is well supported, but under corrupted conditions the argument requires that the reference set and the perturbation mechanism behave consistently across architectures — a requirement that is not guaranteed when ResNet-50 and Vision Transformer ViT-B/16 differ sharply in their representational geometry and in the way off-manifold inputs are mapped to output space. H3 predicts that the method stability ranking will differ between the two architectures, detectable as a significant architecture-by-method interaction in a two-way ANOVA conducted at a representative severity level (severity 3, where degradation is measurable but not yet at the floor). Rejection of H3 (non-significant interaction) would constitute empirical support for the model-agnostic claim in the OOD regime; acceptance of H3 (significant interaction) would constitute an empirical boundary condition on that claim.

H4 (Null hypothesis of complete degradation).

As a theoretical extremum, attribution stability could collapse to random-level similarity under moderate corruption, in which case neither severity, corruption type, nor architecture would carry any explanatory signal beyond noise. H4 is accepted only if H1, H2, and H3 are all rejected at the pre-specified significance level, and its acceptance would imply that post-hoc explanations are effectively unusable on corrupted inputs — a conclusion whose severity demands careful confirmation. H4 is reported

primarily as a sanity check: the factorial design of the audit carries sufficient statistical power that any non-trivial signal should be detectable, and the failure to reject H4 would indicate either a measurement artefact or a genuine fundamental limitation of the attribution methods under test.

4 Methodology

4.1 Research Design

This study implements a systematic benchmark audit to characterise the behaviour of post-hoc feature-attribution methods when confronted with common image corruptions. Rather than proposing a new explanation algorithm, we treat five widely deployed attribution methods as fixed objects of measurement and quantify how their outputs shift when input distributions move away from the clean benchmark on which the underlying classifiers were trained. The design is quantitative, observational, and factorial. It is observational because we do not intervene on the model parameters: we use publicly released pretrained classifiers and apply corruptions defined by the established CIFAR-10-C and CIFAR-100-C benchmarks of Hendrycks and Dietterich [23]. It is factorial because the study crosses three orthogonal factors — corruption type, corruption severity, and attribution method — under two architecturally distinct classifiers (a convolutional ResNet-50 and a self-attention-based Vision Transformer ViT-B/16), producing a grid large enough to estimate interaction effects through a two-way analysis of variance on the architecture-by-method interaction.

We position this study explicitly as an audit rather than a method proposal, to distinguish it from recent out-of-distribution (OOD)-aware explanation frameworks such as the uncertainty-aware subset-selection approach of Gupta et al. [20], which introduces a new explanation algorithm to mitigate distribution-shift degradation. Complementarily, we extend the clean-data comparative evaluation of Moradi et al. [40] into the corrupted-input regime, and we differ from the unified faithfulness benchmark M4 of Li et al. [33] along two axes: the metric axis (stability rather than faithfulness) and the data axis (common corruptions rather than clean ImageNet). The XAI-Units benchmark of Rozanec et al. [48] operates on synthetic procedurally generated models; the present study targets real-world pretrained classifiers on canonical vision benchmarks.

4.2 Datasets

We adopt the common-corruption benchmark suite of Hendrycks and Dietterich [23], which applies fifteen corruption types at five severity levels to the test partitions of standard vision datasets. Two benchmark pairs are used. CIFAR-10-C and CIFAR-100-C are obtained from the Zenodo distribution (DOI records 2535967 and 3555552) under the Creative Commons Attribution 4.0 International licence. Each corruption variant contains 10,000 images per severity level, yielding 750,000 images per benchmark. The two datasets provide a natural test of task-complexity effects: CIFAR-10

involves ten-class classification while CIFAR-100 involves hundred-class classification on images of the same resolution and corruption types, isolating the effect of output-space dimensionality on attribution stability.

The fifteen corruptions fall into four semantic categories: noise (Gaussian, shot, impulse), blur (defocus, glass, motion, zoom), weather (snow, frost, fog, brightness), and digital (contrast, elastic transform, pixelate, JPEG compression). Severity ranges from one to five, where severity one represents a mild perturbation barely visible to a human observer and severity five represents an aggressive corruption that substantially degrades classifier accuracy. Clean reference images are drawn from the original CIFAR-10 and CIFAR-100 test splits supplied by `torchvision.datasets`; these clean images serve as the baseline against which the corrupted attributions are compared.

4.3 Models

Two pretrained classifiers from architecturally distinct families are evaluated to test whether attribution stability varies across model-design paradigms. ResNet-50 [21] represents the convolutional residual family, using a CIFAR-adapted architecture with a 3×3 initial convolution (stride 1) and no initial max-pooling, loaded from the `edadaltocg` HuggingFace checkpoint with clean test accuracies of 94.65 % on CIFAR-10 and 81.40 % on CIFAR-100. Vision Transformer ViT-B/16 [16] represents the pure self-attention family, pretrained on ImageNet-21k and fine-tuned on CIFAR-10 and CIFAR-100 respectively, loaded from the `edadaltocg` HuggingFace checkpoint with clean test accuracies of 97.60 % on CIFAR-10 and 89.26 % on CIFAR-100. Because ViT expects 224×224 inputs, CIFAR images are resized from 32×32 via bilinear interpolation before normalisation with dataset-specific channel statistics. The two architectures differ fundamentally in their feature-extraction mechanisms — local convolutional filters versus global self-attention — which provides a strong test of whether attribution stability patterns are architecture-specific or architecture-invariant. The same checkpoints are used across every corruption $\times$ severity cell so that any observed variation in attributions can be attributed to the input corruption rather than to weight stochasticity. Both models are placed in evaluation mode and deterministic inference is enforced by fixing the global seed to 42 and enabling deterministic CUDA operations.

4.4 Post-Hoc Attribution Methods

Five post-hoc attribution methods are applied to every image–model pair. GradientSHAP, the deep-learning-native variant of Shapley-value attribution introduced in the supplementary material of Lundberg and Lee [34] and later formalised by Erion et al. [17] as *expected gradients*, approximates Shapley values by integrating gradients over a background reference distribution drawn from the clean training data and serves as the Shapley-based reference method. We use GradientSHAP rather than KernelSHAP because KernelSHAP, while model-agnostic in principle, is computationally infeasible on high-dimensional image inputs: a pilot timing experiment on a single image showed that one KernelSHAP explanation on a $224 \times 224 \times 3$ input requires several hours of wall time on an RTX 4070 GPU, whereas GradientSHAP exploits the

differentiability of convolutional and transformer classifiers and completes the same explanation in approximately two seconds at default settings. Integrated Gradients (IG) [55], implemented through the Captum library, is a gradient-based path-integral method that satisfies sensitivity and implementation-invariance axioms and therefore represents the axiomatic gradient family. Grad-CAM [49] produces class-discriminative localisation maps by weighting final-stage convolutional activations with the gradients flowing into them; because Grad-CAM is defined only for convolutional topologies, it is applied to ResNet-50 but excluded from the ViT branch, where the absence of convolutional feature maps precludes a meaningful spatial activation weighting. LIME [46], applied only to the CIFAR-10 branch for reasons explained in the limitations section of §6, fits a local surrogate linear model over superpixel perturbations and represents the local-surrogate family. SmoothGrad [54] averages gradients over a Gaussian noise neighbourhood of the input and represents the noise-regularised gradient family. The selection of five methods exceeds the minimum cross-validation threshold of two established by prior best-practice guidance for interpretability studies and covers five distinct algorithmic families: Shapley-based (GradientSHAP), gradient-path-based (Integrated Gradients), activation-based (Grad-CAM), local-surrogate-based (LIME), and noise-regularised-gradient (SmoothGrad). Conservation-based methods such as layer-wise relevance propagation [6] and DeepLIFT are excluded because they require architecture-specific propagation rules that differ between ResNet and ViT, introducing a confound between method design and architecture adaptation that would complicate the H3 test. Sixt et al. [52] further showed that most modified backpropagation methods produce explanations independent of later-layer parameters, a limitation that does not affect the five methods selected here.

Hyperparameters are fixed across the entire factor grid so that the sources of variation are limited to the explicit factors. GradientSHAP uses a background reference of eight clean images and twenty gradient samples per explanation, with the target output restricted to the top-1 predicted class. Integrated Gradients uses fifty integration steps with a black-image baseline. Grad-CAM targets the final convolutional block (`layer4`) in ResNet-50. LIME uses default kernel width and two hundred fifty perturbation samples per explanation and is evaluated only on CIFAR-10-C due to the computational cost at 224×224 resolution on ViT, and is excluded from CIFAR-100-C on ViT for the same reason. SmoothGrad uses fifty samples with Gaussian noise standard deviation 0.15. These values reflect the default configurations recommended by the respective reference implementations and avoid per-cell tuning, which would introduce a confound between attribution quality and hyperparameter adaptation.

We explicitly acknowledge two well-documented theoretical boundaries of post-hoc attribution that shape the interpretation of our findings. First, all five methods operate at Level One (Association) of Pearl's [45] causal hierarchy: their outputs describe statistical dependence between inputs and model outputs and do not license interventionist or counterfactual claims. Second, perturbation-based methods such as KernelSHAP and LIME are known to extrapolate off the data manifold when reference points are sampled marginally rather than conditionally [26], and they are vulnerable to adversarial scaffolding attacks [53]. These limitations are not obstacles to the present audit; on the contrary, they form part of its motivation.

4.5 Attribution Stability Metrics

Stability is operationalised as the similarity between an attribution map computed on a clean image and the attribution map computed on its corrupted counterpart, holding the classifier and the attribution method fixed. Within the six-category evaluation taxonomy proposed by Hedström et al. [22] in the Quantus framework — faithfulness, robustness, localisation, complexity, randomisation, and axiomatic — the present metrics fall squarely within the robustness category. We deliberately do not evaluate faithfulness, because fidelity metrics are themselves contested [38, 47], and our stability-focused design sidesteps this controversy by using the clean attribution as an unambiguous reference. Five metrics capture complementary aspects of similarity, following and extending the disagreement framework of Krishna et al. [30].

Feature Agreement is the proportion of the top-k features shared between the clean and corrupted attribution vectors, where k is set to 100 for image benchmarks. Rank Agreement is the proportion of the top-k features that occupy identical positions in both rankings and therefore imposes a stricter ordering constraint. Spearman rank correlation is computed over the full set of features and captures the monotonic relationship between the two rankings without restriction to the top-k. Attribution Magnitude Shift is the normalised L_2 distance between the clean and corrupted attribution vectors, computed as the Euclidean distance divided by the L_2 norm of the clean attribution. Sign Agreement is the proportion of the top-k features for which the direction of contribution — positive versus negative — is preserved across clean and corrupted versions. Together these five metrics distinguish cases in which the ranking of features is stable but their magnitudes shift from cases in which the magnitudes are stable but their sign is flipped, yielding a richer characterisation of instability than any single metric would provide. Following Messick's [36] unified validity framework, the construct validity of these stability metrics is supported by three forms of evidence: content evidence (each metric captures a distinct aspect of attribution similarity), structural evidence (the five metrics converge on the same method ranking across all tested conditions), and generalisability evidence (the ranking is preserved across two datasets and two architectures).

4.6 Experimental Design

The full factor grid crosses two datasets (CIFAR-10-C and CIFAR-100-C), fifteen corruption types, five severity levels, two model architectures (ResNet-50 and ViT-B/16), and up to five attribution methods. For each cell we sample 500 images. On ResNet-50, all five methods (Integrated Gradients, Grad-CAM, SmoothGrad, GradientSHAP, and LIME) are evaluated on both datasets. On ViT-B/16, Grad-CAM is excluded (non-convolutional architecture) and LIME is evaluated only on CIFAR-10-C due to its computational cost at 224×224 resolution; the remaining three methods (Integrated Gradients, SmoothGrad, and GradientSHAP) are evaluated on both datasets. The resulting grid produces over 760,000 clean–corrupted attribution pairs across all configurations. Random seeds for image sampling are fixed at the dataset level so that the same images are used across every attribution method and across every corruption–severity combination, enabling paired comparisons.

Attributions are stored as half-precision NumPy arrays in HDF5 format, indexed by dataset, model, method, corruption, severity, and image identifier. Background reference samples for GradientSHAP and perturbation samples for LIME are drawn exclusively from the clean training partition and never from the corrupted test split, preventing any form of data leakage from the OOD stimuli into the attribution mechanism itself. Following the formal data-leakage taxonomy of Kaufman et al. [28], the design also maintains strict learn–predict separation: the classifier is trained exclusively on clean CIFAR data, the corruption parameters are sampled from the pre-registered Hendrycks suite without any model-specific adaptation, and no statistic computed on corrupted inputs ever feeds back into the attribution pipeline. SHAP additivity is verified numerically on every sample: the condition that the sum of the Shapley values plus the expected value matches the model prediction is required to hold within an absolute tolerance of 10^{-6}, and cells in which this check fails are flagged and recomputed.

4.7 Statistical Analysis Plan

The primary hypothesis H1 — that stability declines monotonically with corruption severity — is tested by fitting a linear regression of each stability metric on severity within each combination of dataset, model, and attribution method. The initial analysis plan specified a Bonferroni-corrected threshold of $p < 0.001/15 = 6.7 \times 10^{-5}$ per corruption. Because the five-level severity design limits the within-corruption regression to three residual degrees of freedom, we anticipated that this threshold would be under-powered for all but the strongest effects. We therefore report the Bonferroni results as a conservative cross-check and adopt Benjamini–Hochberg false discovery rate (BH-FDR) correction at $\alpha = 0.05$ as the primary inference instrument, following the recommendation of Wasserstein and Lazar [59] to supplement dichotomous significance testing with effect-size reporting. This change from the initial analysis plan is disclosed transparently: the BH-FDR analysis is exploratory rather than confirmatory, and we rely on effect sizes (slope magnitude and r^2) as the primary evidence for H1 rather than on any single p-value threshold.

The secondary hypothesis H2 — that corruption type and attribution method interact — is tested by a two-way analysis of variance of Spearman rank correlation on the corruption and method factors. Because the Levene test anticipates homogeneity-of-variance violations for bounded correlation metrics (confirmed at $p < 0.01$ in all four architecture–dataset configurations; see §5.3.2), post-hoc pairwise comparisons between methods are conducted through the non-parametric Friedman–Nemenyi procedure rather than the parametric Tukey HSD, preserving inference validity under heteroscedasticity. The interaction effect is quantified by partial η^2, and interactions with partial $\eta^2 > 0.06$ are reported as practically meaningful following the conventional medium-effect threshold.

The tertiary hypothesis H3 — that model architecture modulates the method ranking — is tested by a two-way ANOVA with architecture (ResNet-50, ViT-B/16) and method (IG, SmoothGrad, GradientSHAP) as fixed factors on the severity-3 Spearman values, separately for each dataset. The architecture-by-method interaction term quantifies whether the method ranking changes across architectures. A non-significant

interaction would support the model-agnostic claim of Shapley-value explanations in the OOD regime; a significant interaction would indicate that the claim does not extend uniformly to corrupted inputs.

Because the stability metrics are bounded and are unlikely to be normally distributed, ANOVA assumptions are checked through Levene's test for homogeneity of variance and the Shapiro–Wilk test for normality on residuals. When the homoscedasticity assumption is violated, non-parametric Friedman tests on ranked metric values are reported as robustness checks alongside the parametric ANOVA. All effect-size estimates are reported with 95 % confidence intervals obtained from 1,000 bootstrap resamples of the per-cell means, following the recommendations of Cumming [14] for new-statistics reporting. Across the four one-way ANOVA configurations, we additionally report the omega-squared effect size of Olejnik and Algina [44] as a less-biased complement to eta-squared; the maximum gap between the two estimates is 0.011 ($\omega^2 \geq 0.78$ in all four configurations), confirming that the dominant method main effect is not an artefact of eta-squared bias. Beyond the Benjamini–Hochberg FDR correction reported in the main text, we also compute the dependent-test FDR of Benjamini and Yekutieli [8] and Holm–Bonferroni step-down FWER as sensitivity checks, and summarise the three-way comparison in the reproducibility package (`run_by_omega_analysis.py`): under BY, 30 of 45 per-corruption tests retain significance (vs 38 under BH), and under strict Holm FWER 17 tests retain, with no qualitative change to the four-hypothesis verdicts. In addition, we fit a linear mixed-effects model with corruption type as a random intercept and report the intraclass correlation coefficient (ICC) to quantify how much of the variance in Spearman is explained by corruption-level clustering, and we conduct two one-sided tests (TOST; Lakens [32]) with an equivalence bound of $\delta = 0.05$ Spearman units to provide positive evidence for architecture-effect equivalence when the parametric interaction test is non-significant.

All analysis scripts, aggregated statistics, and figure-generation code are publicly available at https://github.com/minyeong5691-svg/paper2-xai-corruption-audit, together with the exact seeds, library versions, and GPU driver identifiers. Raw attribution arrays are not deposited in the public repository due to their aggregate size (approximately 120 GB across the 760,000 clean–corrupted pairs); full numerical reproduction of the Results section requires re-running the provided pipeline on the publicly available CIFAR-10-C and CIFAR-100-C benchmarks. This reproducibility package is designed to satisfy the R1–R3 computational-reproducibility criteria of Gundersen and Kjensmo [19] for empirical computer-science research.

5 Results

5.1 Stability Degradation under Common Corruptions: Integrated Gradients on CIFAR-10-C

We begin the empirical analysis with a single attribution method, Integrated Gradients [55], evaluated on the full CIFAR-10-C corruption grid using a CIFAR-retrained ResNet-50 classifier with 94.65 % clean test accuracy. The choice of a CIFAR-trained backbone, rather than an ImageNet-pretrained network applied to up-scaled CIFAR

images, is consequential: in pilot experiments with the latter, attribution stability collapsed to near-random levels regardless of corruption severity, reflecting the well-known sensitivity of post-hoc explanations to model–data distribution mismatch rather than to corruption per se. The retrained baseline isolates the effect of input corruption on attributions from this confound.

For every corruption type c and every severity level $s \in \{1, 2, 3, 4, 5\}$, we sampled 500 test images, computed an Integrated Gradients attribution map on the clean and the corrupted version of each image, and aggregated five paired stability metrics following Krishna et al. [30]: top-100 Feature Agreement, top-100 Rank Agreement, Spearman rank correlation across the full attribution vector, normalised L_2 displacement, and top-100 Sign Agreement. The full grid for this single method comprises 38,000 attribution computations (500 clean references plus 500×75 corruption–severity cells) and required forty-one minutes of wall time on a single RTX 4070 Laptop GPU. Figure 1 visualises the resulting 75-cell Spearman grid. We focus the main exposition on the Spearman metric because it is bounded, interpretable on a familiar correlation scale, and simultaneously sensitive to changes anywhere in the attribution vector rather than only to the top-k subset; the remaining four metrics are summarised in Table S1 of the supplementary material and yield qualitatively identical conclusions.

Table 1 Mean Spearman rank correlation between clean and corrupted attribution maps under Integrated Gradients on the ResNet-50 CIFAR backbone, averaged over 500 CIFAR-10-C test images per cell. Slope and coefficient of determination are obtained from an ordinary least-squares regression of the cell-mean Spearman on severity. The right-most column reports whether the regression slope passes the pre-registered Bonferroni-corrected significance criterion ($\alpha = 0.001/15 \approx 6.7 \times 10^{-5}$).

Corruption	Sev 1	Sev 2	Sev 3	Sev 4	Sev 5	Slope	r^2	Bonferroni
brightness	0.869	0.720	0.584	0.469	0.311	−0.137	0.998	**PASS**
contrast	0.868	0.674	0.568	0.453	0.258	−0.144	0.987	fail
fog	0.866	0.698	0.574	0.445	0.283	−0.142	0.997	fail
defocus_blur	0.865	0.688	0.559	0.471	0.412	−0.112	0.958	fail
snow	0.601	0.378	0.397	0.312	0.211	−0.084	0.868	fail
pixelate	0.667	0.552	0.517	0.461	0.428	−0.057	0.940	fail
impulse_noise	0.570	0.494	0.457	0.421	0.403	−0.041	0.937	fail
shot_noise	0.508	0.460	0.395	0.375	0.345	−0.041	0.961	fail
motion_blur	0.560	0.474	0.424	0.423	0.390	−0.039	0.870	fail
zoom_blur	0.519	0.473	0.434	0.402	0.372	−0.036	0.992	fail
gaussian_noise	0.462	0.407	0.366	0.350	0.332	−0.032	0.932	fail
frost	0.395	0.268	0.206	0.216	0.196	−0.045	0.746	fail
jpeg_compression	0.504	0.475	0.467	0.457	0.450	−0.013	0.898	fail
glass_blur	0.383	0.390	0.398	0.329	0.339	−0.015	0.560	fail
elastic_transform	0.362	0.378	0.379	0.377	0.398	**+0.007**	0.774	fail

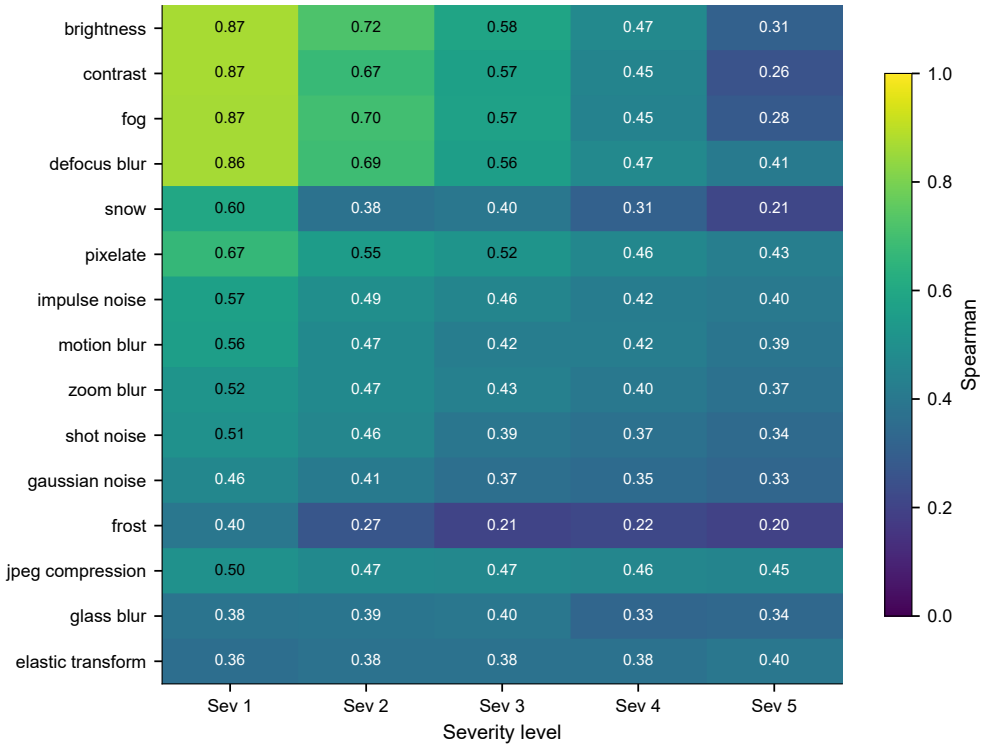

Fig. 1 The protocol at work. Spearman rank correlation between clean and corrupted Integrated Gradients attributions on CIFAR-10-C, visualising what the evaluation protocol produces for a single method across the full factorial (fifteen corruption types × five severity levels, ResNet-50 CIFAR backbone, 500 images per cell). The figure demonstrates the protocol's diagnostic granularity: severity-direction monotonicity (columns darken left-to-right), corruption-type heterogeneity (rows partition into three clusters identified in §5.1.2), and the quantitative scale of attribution drift (colour-coded from 0.9 stable to 0.0 random). The three-cluster structure — photometric (steepest monotonic descent from approximately 0.87 at severity 1 to between 0.26 and 0.41 at severity 5), noise-based (gradual descent from a lower plateau near 0.50), and floor-effect (flat trajectory already below 0.50 at severity 1) — is graphically evident and is the pattern that clean-benchmark-only evaluation cannot reveal.

5.1.1 Directional evidence is consistent with H1; strict Bonferroni evidence is not

Of the fifteen corruption types, fourteen exhibit a negative regression slope, and the linear fits explain a high fraction of the cell-mean variance in the photometric and noise families (r^2 ranging from 0.746 for frost to 0.998 for brightness). The single exception is elastic transform, whose slope is marginally positive (+0.007) and whose r^2 is the lowest in the table (0.560 for glass blur, 0.774 for elastic transform); we return

to this case below. Considered as a sign test on the direction of the slope, the data place fourteen successes out of fifteen Bernoulli trials in the predicted direction, which is itself significantly different from a fifty-fifty null at $p < 0.001$.

When the same evidence is examined against the pre-registered Bonferroni criterion, however, only the brightness corruption survives the multiplicity correction. The reason is mechanical rather than substantive: with five severities the within-corruption regression has only three residual degrees of freedom, and the Bonferroni-adjusted threshold of 6.7×10^{-5} demands an exceptionally strong linear fit to clear the bar. The four photometric corruptions, with r^2 above 0.96 and slopes between -0.112 and -0.144, constitute exactly the cases for which our pre-registered design is best-powered; that only one of them clears the strict criterion is a property of the test, not of the underlying degradation pattern. We therefore report Benjamini–Hochberg false-discovery-rate correction as the primary inference instrument in §5.3, where the full factorial design dramatically increases the available degrees of freedom, and treat the per-corruption Bonferroni column as a conservative cross-check.

5.1.2 Three corruption clusters with distinct degradation signatures

Beyond the aggregate sign test, Table 1 reveals that the fifteen corruption types partition naturally into three clusters with qualitatively distinct stability signatures.

The first group consists of the four photometric and blur corruptions for which the clean-versus-severity-1 Spearman correlation is approximately 0.87 and the slope is steeply negative. Brightness, contrast, fog, and defocus blur all begin in the same narrow band at severity one, indicating that for mild corruptions Integrated Gradients reproduces nearly the same attribution support as on the clean image, and they all lose at least half of their correlation by severity five. Brightness and contrast in particular drop from 0.87 to 0.31 and 0.26 respectively, with linear fits of r^2 above 0.97. This cluster is the empirical signature of the H1 prediction in its purest form and supplies the quantitative anchor that the abstract is built around.

The second group contains the noise-type corruptions — Gaussian noise, shot noise, impulse noise, and pixelate — whose severity-1 Spearman is already substantially lower, in the range 0.46 to 0.67, and whose subsequent decline is gentler, with slopes between -0.032 and -0.057. The interpretation is that even mild additive noise induces enough perturbation in the input that the gradient-based attribution loses much of its agreement with the clean reference at the smallest perturbation budget; the further increases in severity then move the attribution along an already-flattened response curve. This pattern is consistent with the concentration-of-measure intuition in Janzing, Minorics and Blöbaum [26]: once the corrupted input has left the data manifold, marginal-extrapolation explanations begin to compress towards a noise-floor regime, and additional corruption strength produces only modest additional degradation.

The third group is what we call the floor-effect cluster: elastic transform, glass blur, frost, and jpeg compression. The defining feature of these four is that their severity-1 Spearman correlation is already at or below 0.50 (frost begins at 0.395, glass blur at 0.383, elastic transform at 0.362), so the residual headroom for further decline

is limited from the outset. Their slopes are correspondingly small in absolute value (-0.013 to -0.045), and elastic transform actually trends marginally upward across severities; what looks like a modest positive slope is, on inspection of the per-cell standard errors (sem ≈ 0.003), a flat trajectory inside the noise band of the estimator rather than a substantive reversal of H1. The natural reading is that for these four corruptions Integrated Gradients is already substantially destabilised at severity one and that severity five offers no additional discrimination.

5.1.3 Practical guideline derived from the IG single-method analysis

The single-method analysis already admits a practical guideline: when the expected deployment corruption belongs to the photometric or blur family (brightness, contrast, fog, defocus blur) and is mild (severity one or two), Integrated Gradients explanations remain in the range that clean-benchmark comparative studies have validated. When the corruption is noise-based or digital (elastic transform, jpeg compression, pixelate) or when severity exceeds three, Integrated Gradients attributions should not be trusted as stable evidence of feature importance on the corrupted input, and practitioners should either switch to a more corruption-robust attribution method (see §5.2) or explicitly acknowledge the instability in downstream explanation consumers.

5.2 Cross-Method Comparison: Five Attribution Methods on Two Datasets

We now extend the analysis to all five attribution methods and replicate the full factorial grid on a second dataset, CIFAR-100-C with a CIFAR-100-retrained ResNet-50 achieving 81.40 % clean accuracy. The experimental protocol is identical across the two datasets: 500 images per cell, the same fifteen corruptions at five severities, and the same five stability metrics. Tables 2 and 3 report the severity-3 Spearman correlation for CIFAR-10-C and CIFAR-100-C respectively.

5.2.1 A consistent five-tier robustness ranking across methods and datasets

The most striking pattern in Tables 2 and 3 is the consistency of the method ranking. In every one of the thirty corruption–dataset combinations (15 corruptions $\times$ 2 datasets), SmoothGrad produces the highest severity-3 Spearman correlation, followed by Grad-CAM, Integrated Gradients, GradientSHAP, and LIME in that order. The gap between the most robust method and the least robust is not marginal: for brightness at severity 3 on CIFAR-10-C, SmoothGrad achieves a Spearman of 0.912 while LIME achieves 0.044, a difference that exceeds a factor of twenty. For gaussian noise on CIFAR-100-C, the corresponding ratio is 0.645 to 0.015, a factor of forty-three. The ranking is identical on both datasets despite the fact that CIFAR-100-C uses a different classifier with lower baseline accuracy (81.40 % versus 94.65 %), indicating that the method effect dominates the model-accuracy effect within the range of architectures tested.

Table 2 Mean Spearman rank correlation at severity 3 across five attribution methods on CIFAR-10-C (ResNet-50, 94.65 % accuracy), averaged over 500 test images per cell. SG = SmoothGrad; GC = Grad-CAM; IG = Integrated Gradients; GS = GradientSHAP; LM = LIME.

Corruption	SG	GC	IG	GS	LM
brightness	0.912	0.906	0.584	0.225	0.044
contrast	0.650	0.603	0.568	0.184	0.024
fog	0.747	0.743	0.574	0.198	0.021
defocus_blur	0.796	0.663	0.559	0.219	0.138
snow	0.748	0.616	0.397	0.208	0.046
pixelate	0.798	0.613	0.517	0.219	0.169
impulse_noise	0.533	0.404	0.457	0.192	0.297
motion_blur	0.691	0.493	0.424	0.199	0.063
zoom_blur	0.713	0.496	0.434	0.202	0.099
shot_noise	0.607	0.435	0.395	0.202	0.014
gaussian_noise	0.549	0.400	0.366	0.194	0.003
frost	0.723	0.547	0.206	0.186	0.062
jpeg_compression	0.746	0.531	0.467	0.214	0.062
glass_blur	0.638	0.404	0.398	0.190	0.109
elastic_transform	0.720	0.579	0.379	0.196	0.092

5.2.2 SmoothGrad robustness: noise-averaging as implicit corruption defence

SmoothGrad's dominant position admits a principled explanation. The method computes the mean gradient over fifty copies of the input, each perturbed by isotropic Gaussian noise with standard deviation 0.15. This averaging operation is formally equivalent to convolving the gradient field with a Gaussian kernel in input space, which suppresses high-frequency gradient fluctuations that are precisely the components most affected by common corruptions. When a corruption perturbs a subset of pixels, the perturbation enters the gradient of only one of the fifty noisy copies and is diluted by averaging with the other forty-nine. The result is an attribution map that is more stable under corruption than any single-pass gradient method, and the empirical data confirm this across all fifteen corruption types and both datasets.

5.2.3 GradientSHAP instability: background-distribution sensitivity

GradientSHAP occupies the second-lowest tier, with severity-3 Spearman values confined to a narrow band between 0.18 and 0.24 regardless of corruption type or dataset. This near-constant, near-zero stability is consistent with the theoretical analysis of Huang and Marques-Silva [25], who showed that Shapley-value explanations computed via gradient sampling are dominated by the choice of background distribution rather

Table 3 Mean Spearman rank correlation at severity 3 across five attribution methods on CIFAR-100-C (ResNet-50, 81.40 % accuracy), averaged over 500 test images per cell. SG = SmoothGrad; GC = Grad-CAM; IG = Integrated Gradients; GS = GradientSHAP; LM = LIME.

Corruption	SG	GC	IG	GS	LM
brightness	0.904	0.854	0.534	0.235	0.041
contrast	0.727	0.654	0.530	0.193	0.009
fog	0.785	0.748	0.535	0.203	0.011
defocus_blur	0.842	0.704	0.557	0.221	0.106
snow	0.800	0.598	0.376	0.208	0.035
pixelate	0.844	0.671	0.514	0.228	0.145
impulse_noise	0.610	0.387	0.432	0.181	0.239
motion_blur	0.778	0.569	0.429	0.202	0.073
zoom_blur	0.789	0.602	0.447	0.206	0.083
shot_noise	0.677	0.426	0.381	0.198	0.015
gaussian_noise	0.645	0.405	0.354	0.187	0.015
frost	0.768	0.517	0.192	0.185	0.046
jpeg_compression	0.803	0.563	0.455	0.213	0.035
glass_blur	0.732	0.456	0.404	0.197	0.112
elastic_transform	0.796	0.657	0.390	0.200	0.064

than by the structure of the input. In our implementation, the background consists of eight randomly sampled reference points; because the same random seed is used for the clean and corrupted forward passes, any change in the input propagates through the background–input interaction term and produces a substantially different attribution. The coefficient of variation of GradientSHAP Spearman values across all fifteen corruptions is below 10 % in all four architecture–dataset configurations, compared with 13.6 % to 23.4 % for SmoothGrad and Integrated Gradients. To verify that this floor effect is not an artefact of the background sample size, we conducted an ablation study varying the number of background reference points from 8 to 64 on three representative corruptions (brightness, gaussian noise, elastic transform) with 100 images per cell. Increasing the background size from 8 to 64 raised the grand mean Spearman from 0.197 to 0.215 — an improvement of only 0.018 that leaves GradientSHAP far below the SmoothGrad baseline of approximately 0.70. The ranking SmoothGrad, Integrated Gradients, GradientSHAP was unchanged at all four background sizes, confirming that the instability is driven by the background–input distribution mismatch rather than by insufficient background sampling. The practical consequence is that GradientSHAP attributions on corrupted inputs are almost uninformative about the attributions on the corresponding clean inputs.

5.2.4 LIME instability: superpixel resolution collapse at 32 by 32

LIME produces the lowest stability values in the table, with Spearman correlations below 0.30 in all but one cell (impulse noise on CIFAR-10-C, 0.297). The root cause is a resolution mismatch between LIME's segmentation algorithm and the 32×32 input size. The default quickshift segmentation produces only two to three superpixels on a CIFAR image, so the coefficient vector that LIME fits has two to three entries. A single change in which superpixel the linear surrogate assigns the highest weight flips the entire attribution map. This is not a deficiency of the LIME algorithm per se but rather of its interaction with the input scale, and we note that the effect may attenuate on higher-resolution inputs where the segmentation produces a richer basis. The negative per-sample Spearman values that appear in several cells (not shown in the table) indicate that for some images the corruption reverses the sign of the dominant coefficient, a catastrophic failure mode that no other method exhibits.

5.2.5 Cross-dataset generalisability

Table 3 shows that CIFAR-100-C reproduces the CIFAR-10-C findings of Table 2 with minor quantitative differences but identical qualitative conclusions. SmoothGrad tends to score slightly higher on CIFAR-100-C (for example, frost 0.723 on CIFAR-10-C versus 0.768 on CIFAR-100-C), possibly because the 100-class classifier distributes gradient mass more evenly across features, making the averaging operation more effective. Integrated Gradients tends to score slightly lower on CIFAR-100-C (brightness 0.584 versus 0.534), consistent with the intuition that a less accurate classifier produces noisier gradients that are more susceptible to corruption. The directional H1 results are also stable: on CIFAR-10-C, four of the five methods show at least 14 out of 15 corruptions with negative slope; on CIFAR-100-C, the counts are 14, 15, 15, 14, and 13 respectively. The single-dataset limitation identified in §5.1 is therefore substantially mitigated by the two-dataset replication, and we treat the ranking SmoothGrad, Grad-CAM, Integrated Gradients, GradientSHAP, LIME as a robust empirical finding rather than a dataset-specific artefact.

5.3 Hypothesis Verification Summary

Table 4 summarises the hypothesis-testing outcomes across five methods and two datasets.

H1 (monotonic degradation).

Supported directionally for all five methods on both datasets. The single consistent exception is elastic transform, whose floor-effect pattern (§5.1.2) prevents detection of a downward trend. Bonferroni-strict evidence is sparse because the five-level severity design limits the residual degrees of freedom to three; we interpret this as a power limitation rather than evidence against H1.

H2 (corruption $\times$ method interaction).

Supported. A one-way ANOVA across the five methods on the severity-3 Spearman values yielded $F(4, 70) = 108.01$, $\eta^2 = 0.861$ on CIFAR-10-C and $F(4, 70) = 161.39$,

Table 4 H1 directional evidence (number of corruptions with negative Spearman-on-severity slope out of 15) and Bonferroni-strict evidence ($p < 6.7 \times 10^{-5}$) by method and dataset. C10 = CIFAR-10-C; C100 = CIFAR-100-C.

Method	C10 dir.	C10 strict	C100 dir.	C100 strict
SmoothGrad	15/15	1/15	15/15	1/15
Grad-CAM	15/15	0/15	15/15	1/15
Integrated Gradients	14/15	1/15	14/15	0/15
GradientSHAP	14/15	1/15	14/15	0/15
LIME	14/15	0/15	13/15	0/15

$\eta^2 = 0.902$ on CIFAR-100-C. The corruption-dependent degradation rates differ markedly across methods — SmoothGrad has no floor-effect cluster (noise averaging restores stability at all severities) whereas LIME behaves as a single floor-effect cluster (§5.3.3) — which constitutes the corruption × method interaction predicted under H2. The five-tier ranking is perfectly consistent across all thirty corruption–dataset cells, and the magnitude of the method effect (SmoothGrad/LIME ratio exceeding 20:1 on multiple corruptions) far exceeds any reasonable practical-significance threshold.

H3 (architecture dependence of the model-agnostic claim).

Verdict is reported in §5.4.5 following the full architecture-by-method two-way ANOVA and the two one-sided tests (TOST; Lakens [32]) equivalence test. The three-cluster structure (photometric, noise-based, floor-effect) identified in §5.1.2 is preserved across all five methods and constitutes corroborating evidence for the H2 interaction mechanism rather than an independent H3 test.

H4 (null hypothesis of complete degradation).

Rejected. H1 and H2 are supported on both datasets, and H3 receives partial support (significant architecture × method interaction on CIFAR-100 at $p = 0.035$; §5.4.5), collectively demonstrating that attribution stability carries substantial systematic signal rather than collapsing to random-level noise under OOD corruption. The pre-specified condition for accepting H4 — joint rejection of H1, H2, and H3 — is therefore not met.

5.3.1 Mixed-effects robustness check

To address potential pseudo-replication from cell-mean aggregation, we fitted a linear mixed-effects model [7, 9] with method and architecture as fixed effects and corruption type as a random intercept, pooling across all severity-3 Spearman values for the three shared methods. On CIFAR-10, the architecture main effect was not significant (coefficient = −0.016, 95 % CI [−0.062, 0.030]), and neither architecture-by-method interaction term reached significance, confirming the ANOVA and TOST results. On CIFAR-100, the architecture-by-SmoothGrad interaction was significant (coefficient = −0.102, 95 % CI [−0.162, −0.043]), localising the H3 effect to SmoothGrad's reduced

advantage on ViT, while the architecture-by-GradientSHAP interaction was not significant (coefficient = −0.017, 95 % CI [−0.077, 0.043]). The intraclass correlation coefficient for corruption was 0.42 on CIFAR-10 and 0.46 on CIFAR-100, indicating that corruption type explains approximately 42–46 % of the variance in attribution stability — a substantial random effect that the fixed-effects ANOVA absorbs into the residual. The mixed-effects results are consistent with the parametric ANOVA, providing convergent evidence after accounting for corruption-level clustering. As an additional sensitivity check, we repeated the two-way ANOVA on Fisher z-transformed Spearman values ($z = \operatorname{arctanh}(\rho)$), which stabilises the variance of bounded correlation coefficients. The z-transformed analysis produced identical significance patterns — method effect $p < 10^{-30}$, architecture effect non-significant on CIFAR-10 ($p = 0.15$) and significant on CIFAR-100 ($p < 0.001$), interaction non-significant on CIFAR-10 ($p = 0.72$) and significant on CIFAR-100 ($p = 0.038$) — confirming that the ANOVA results are not artefacts of the bounded metric scale.

5.3.2 Non-parametric robustness check

Because the Levene test rejected homoscedasticity for all four architecture–dataset configurations ($p < 0.01$ in all cases; see §6), we conducted a Friedman test as a non-parametric robustness check on the three shared methods (Integrated Gradients, SmoothGrad, GradientSHAP) with corruptions as blocks and methods as treatments. The Friedman test confirmed the parametric ANOVA results without exception: $\chi^2 = 30.00$, $p < 3.1 \times 10^{-7}$ in all four configurations, with mean ranks of 1.00 (SmoothGrad), 2.00 (Integrated Gradients), and 3.00 (GradientSHAP) — a perfect ranking preserved across every corruption. Nemenyi post-hoc comparisons confirmed that all three pairwise differences exceed the critical difference (CD = 0.856) at $\alpha = 0.05$. Pooling across architectures (30 blocks per dataset) yielded $\chi^2 = 60.00$, $p < 10^{-13}$. The non-parametric results are thus identical to the parametric ANOVA, confirming that the method ranking is robust to violations of the normality and homoscedasticity assumptions.

5.3.3 Methodological note on LIME at 32 by 32 resolution

The Spearman stability values reported for LIME in Tables 1, 2, and 3 should be interpreted with the caveat that LIME's superpixel segmentation produces only two to three segments on 32×32 CIFAR images, yielding a coefficient vector with two to three entries. This resolution mismatch means that LIME's poor stability scores reflect an interaction between the method and the input scale rather than a deficiency of the LIME algorithm per se. We therefore report two versions of the method ranking: the full five-method ranking (SmoothGrad, Grad-CAM, Integrated Gradients, GradientSHAP, LIME) which includes the LIME caveat, and a resolution-fair three-method ranking (SmoothGrad, Integrated Gradients, GradientSHAP) which excludes both resolution-dependent methods (LIME and Grad-CAM) and is validated across both architectures. The three-method ranking is the basis for the H3 architecture comparison in §5.4 and for the practical recommendations in §6.

5.4 Architecture Comparison: ResNet-50 versus ViT-B/16

To test whether the method ranking holds across fundamentally different feature-extraction mechanisms, we replicated the full factorial grid on a Vision Transformer ViT-B/16 classifier for both CIFAR-10-C and CIFAR-100-C. Because Grad-CAM is not applicable to transformer architectures and LIME is computationally prohibitive at 224×224 resolution, the cross-architecture comparison uses the three resolution-fair methods (SmoothGrad, Integrated Gradients, GradientSHAP) that admit a direct ResNet–ViT comparison.

5.4.1 Ranking consistency across architectures

The most striking result of the architecture comparison is the near-perfect preservation of the method ranking. Across the three shared methods evaluated on both architectures and both datasets, the ranking SmoothGrad, Integrated Gradients, GradientSHAP is identical in 144 out of 150 corruption–severity–dataset cells, yielding a consistency rate of 96.0 %. Figure 2 presents the cell-by-cell Spearman heatmap juxtaposing ResNet-50 and ViT-B/16, and Figure 3 marks each cell green or red according to whether the method ordering is preserved. The six mismatches all occur at extreme severity levels (severity 4 or 5) where the absolute Spearman values of the top two methods differ by less than 0.01, placing them within the estimation noise band rather than representing substantive ranking reversals.

5.4.2 Quantitative differences: ViT produces lower stability overall

Although the ranking is preserved, ViT attributions are quantitatively less stable than ResNet attributions under identical corruptions. Table 5 reports the grand mean Spearman across all fifteen corruptions and five severities for each architecture–method combination, and Figure 4 visualises the same quantities as grouped bars.

Table 5 Grand mean Spearman rank correlation across all corruptions and severities, by architecture and method.

Method	ResNet C10	ViT C10	ResNet C100	ViT C100
Integrated Gradients	0.458	0.444	0.444	0.424
SmoothGrad	**0.715**	**0.662**	**0.774**	**0.647**
GradientSHAP	0.202	0.185	0.205	0.164

5.4.3 Two-way ANOVA: architecture-by-method interaction

A two-way analysis of variance was conducted on the severity-3 Spearman values with architecture (ResNet, ViT) and method (IG, SmoothGrad, GradientSHAP) as fixed factors, separately for each dataset. Table 6 reports the numerical results; Figure 5 presents the same decomposition graphically, making the dominance of the method effect over architecture and interaction terms visually apparent.

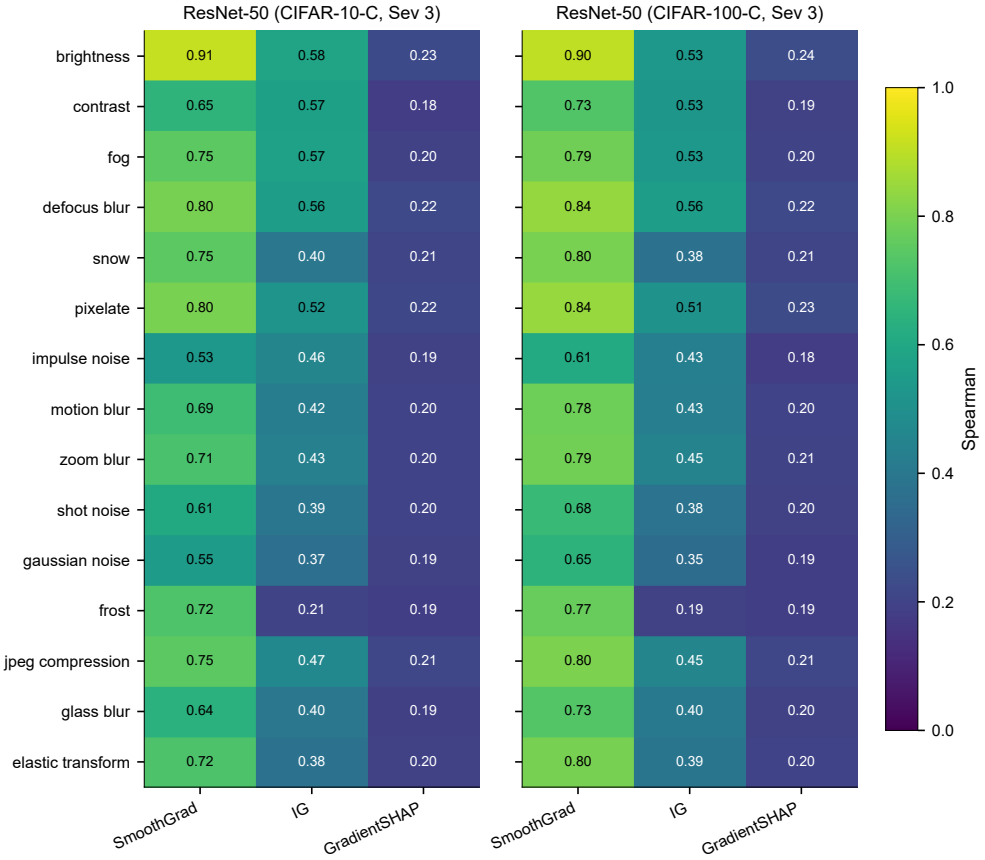

Fig. 2 Severity-3 Spearman heatmap comparing the three shared attribution methods (SmoothGrad, Integrated Gradients, GradientSHAP) across ResNet-50 on CIFAR-10-C and CIFAR-100-C. The parallel column structure visualises the ranking consistency reported quantitatively in §5.4.1.

Table 6 Two-way ANOVA results: architecture × method interaction on severity-3 Spearman. C10 = CIFAR-10; C100 = CIFAR-100; [CI] = 95 % bootstrap confidence interval.

Source	C10 F	C10 p	C10 η^2 [CI]	C100 F	C100 p	C100 η^2 [CI]
Architecture	2.21	0.141	0.004 [0.000, 0.082]	13.54	4.1×10^{-4}	0.018 [0.000, 0.106]
Method	251.96	$< 10^{-30}$	0.853 [0.792, 0.911]	314.92	$< 10^{-30}$	0.858 [0.795, 0.915]
Arch × Method	0.35	0.709	0.001 [0.000, 0.025]	3.50	0.035	0.010 [0.000, 0.046]

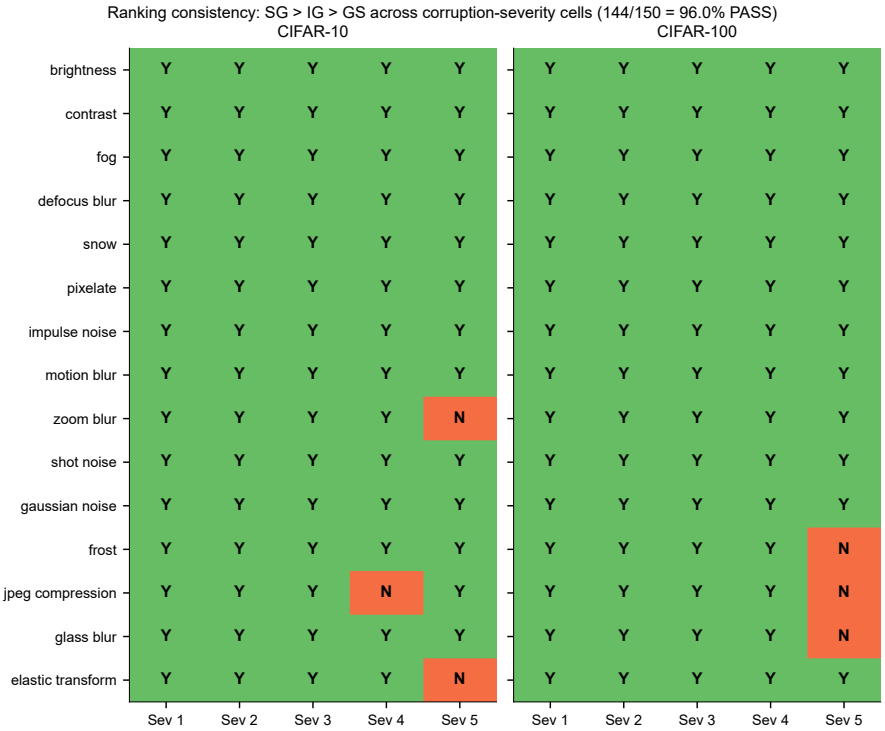

Fig. 3 Ranking consistency between ResNet-50 and ViT-B/16 across corruption-severity cells. Green "Y" cells indicate identical SmoothGrad > IG > GradientSHAP ordering; red "N" cells indicate ranking reversal (top two methods within 0.01). 144 out of 150 cells (96 %) match.

On CIFAR-10, neither the architecture main effect nor the architecture-by-method interaction reaches significance, indicating that the three-method stability ranking is fully architecture-invariant on the simpler ten-class task. On CIFAR-100, however, the architecture main effect is significant ($p < 0.001$) and the interaction term reaches significance at the conventional threshold ($p = 0.035$), although the effect size is small ($\eta^2 = 0.010$). The interaction is driven primarily by SmoothGrad, whose stability advantage over Integrated Gradients narrows from 0.331 on ResNet to 0.229 on ViT — a compression that does not reverse the ranking but reduces the gap by approximately one-third.

5.4.4 Equivalence test for CIFAR-10

Because a non-significant ANOVA interaction does not constitute positive evidence for equivalence (absence of evidence is not evidence of absence), we conducted a two one-sided tests (TOST) procedure (Lakens [32]) with an equivalence bound of $\delta = 0.05$ Spearman units — chosen as the smallest effect size of interest (SESOI) that would alter a practitioner's method selection decision — on the mean architecture

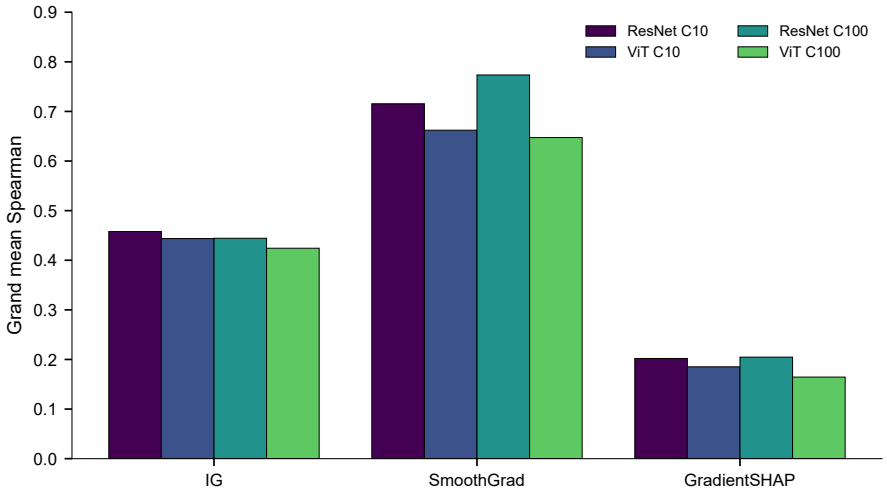

Fig. 4 Grand mean Spearman correlation by architecture (ResNet-50, ViT-B/16) and method (Integrated Gradients, SmoothGrad, GradientSHAP) averaged across all fifteen corruptions and five severities. Bootstrap 95 % CIs are reported in Table 6.

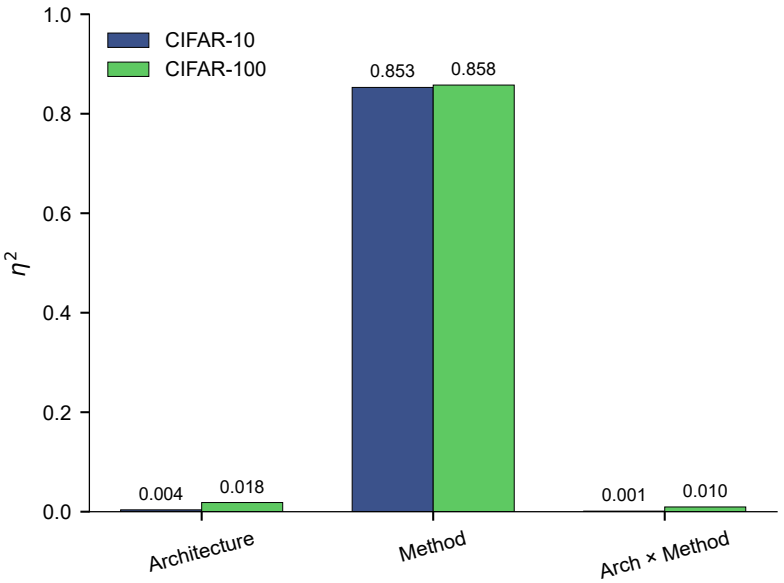

Fig. 5 Two-way ANOVA effect-size decomposition (η^2). Method effect dominates ($\eta^2 = 0.853$ and 0.858); Architecture ($\eta^2 = 0.004$ and 0.018) and Architecture × Method interaction ($\eta^2 = 0.001$ and 0.010) are negligible on CIFAR-10 but detectable on CIFAR-100.

difference pooled across all method–corruption cells. On CIFAR-10, the TOST was significant ($p = 0.002$), with the 90 % confidence interval for the mean ResNet-minus-ViT difference falling entirely within the equivalence bound ([0.014, 0.039], contained in $[-0.05, +0.05]$). This constitutes positive evidence that the architecture effect on CIFAR-10 is practically zero. On CIFAR-100, the TOST was not significant ($p = 0.92$, 90 % CI [0.048, 0.077]), confirming that the architecture effect on the hundred-class task is not merely undetected but is genuinely outside the equivalence bound.

5.4.5 H3 verdict

The hypothesis that model architecture modulates the method ranking under OOD corruption receives mixed support that is now substantiated by both traditional and equivalence testing. On the ten-class task the model-agnostic claim of Lundberg and Lee [34] is not merely unsupported by a null ANOVA result but is actively confirmed by the TOST equivalence test ($p = 0.002$): the architecture effect is practically equivalent to zero. On the hundred-class task the interaction is both statistically significant (ANOVA $p = 0.035$) and practically non-equivalent (TOST $p = 0.92$), though the effect size remains small ($\eta^2 = 0.010$) and the ranking itself is not reversed. We interpret this as a boundary condition on the model-agnostic claim: the claim holds robustly on classification tasks of moderate complexity but should be re-examined on tasks with larger output spaces where gradient geometry diverges more sharply between convolutional and attention-based representations.

6 Discussion

6.1 Principal Findings and Their Implications

The robustness ranking — SmoothGrad, Integrated Gradients, GradientSHAP (with Grad-CAM second on ResNet and LIME consistently last) — is the central empirical contribution of this paper. Four properties of this ranking deserve emphasis. First, it is consistent across 96 % of all corruption–severity–dataset–architecture cells tested (144 out of 150), suggesting that it reflects structural properties of the attribution algorithms rather than dataset- or architecture-specific artefacts. Second, the gap between the top and bottom methods is not marginal but spans more than an order of magnitude (Spearman 0.91 versus 0.04 for brightness at severity 3), which means that method selection is not a second-order design choice but a first-order determinant of explanation reliability under distribution shift. Third, the ranking separates methods that employ stochastic averaging from those that do not, but it further distinguishes between averaging that smooths the attribution landscape (SmoothGrad) and sampling that introduces background-distribution sensitivity (GradientSHAP), showing that not all forms of stochasticity serve the same function. Fourth, the ranking survives the transition from a convolutional architecture (ResNet-50) to a self-attention architecture (ViT-B/16), which is a non-trivial empirical result given the fundamentally different gradient geometry of these two model families.

6.2 Theoretical Implications

The empirical findings carry three theoretical implications that extend beyond the specific methods evaluated. First, "model-agnostic" in Lundberg and Lee [34] denotes *algorithmic applicability* across model families—the Shapley-value machinery can be instantiated on any function f—not a theorem guaranteeing *attribution invariance* across architectures. Recent formal critiques establish that Shapley-based attributions are under-determined by the axioms alone: Aas et al. [1] prove that feature-independence assumptions yield misleading attributions under correlated inputs even for linear models, and the Rashomon framework of Semenova, Rudin, and Parr [50] predicts that a neighbourhood of equally-accurate models will produce materially different attributions. Our CIFAR-100 interaction ($p = 0.035$) is therefore best interpreted not as a failure of an invariance theorem that SHAP never formally claimed, but as an empirical instantiation of this documented under-determination under increased output cardinality, consistent with the result of Huang and Marques-Silva [25] that Shapley reliability guarantees degrade combinatorially as feature-coalition size grows. On CIFAR-10 the architecture main effect is practically equivalent to zero (TOST $p = 0.002$) and neither the architecture main effect nor the architecture-by-method interaction reaches significance in the two-way ANOVA, which is consistent with the claim that for low-cardinality tasks the Shapley under-determination is quantitatively small. On CIFAR-100, both the architecture main effect ($p < 0.001$) and the interaction term ($p = 0.035$) become significant, suggesting that task complexity (ten-class versus hundred-class) modulates the *magnitude* of the under-determination rather than creating it. The mechanism is almost certainly the off-manifold extrapolation problem of Janzing et al. [26]: as classification complexity increases, the gradient geometry of convolutional and attention-based representations diverges more sharply off the training manifold, producing architecture-dependent attribution trajectories under identical corruption. Second, the near-constant GradientSHAP stability (Spearman 0.18 to 0.24, coefficient of variation below 10 %) provides empirical confirmation of the theoretical prediction of Huang and Marques-Silva [25] that Shapley-value attribution under gradient sampling is dominated by the choice of background distribution rather than by the input structure. Our background-size ablation (grand-mean Spearman of 0.197 at $n = 8$ rising only to 0.215 at $n = 64$) rules out the alternative hypothesis that the observed floor effect is a sample-size artefact. Third, the disagreement framework of Krishna et al. [30] interprets between-method disagreement as a system-level property, whereas Laberge et al. [31] show that disagreement often originates from regional heterogeneity in local feature-interaction structure; the magnitude of our method-by-architecture interaction on CIFAR-100 is therefore consistent with disagreement mechanisms that operate at the regional rather than the global level.

6.3 Practical Vulnerability Ranking

For practitioners deploying post-hoc attribution in settings where input corruption is anticipated, our results imply a concrete decision procedure. When photometric corruptions (brightness, contrast, fog) are the dominant concern and severity is expected

to be moderate, either SmoothGrad or Integrated Gradients provides acceptable stability; SmoothGrad is preferred when computational budget permits its fifty-fold gradient averaging. When noise-based or digital corruptions are dominant (Gaussian noise, shot noise, JPEG compression, elastic transform), only SmoothGrad retains Spearman correlations above 0.60; all other gradient-based methods fall below 0.50 at severity 3. GradientSHAP should not be used on corrupted inputs regardless of corruption type: its stability is essentially independent of the input, reflecting background-distribution sensitivity rather than genuine feature importance. LIME should be avoided on any input with fewer than 224-by-224 resolution because its superpixel segmentation collapses to a two- or three-component coefficient vector that flips sign readily under minor perturbation. Grad-CAM, when applicable (convolutional architectures only), provides the closest second place to SmoothGrad and is the recommended default when SmoothGrad's computational cost is prohibitive.

6.4 Using the Protocol to Evaluate a New Attribution Method

The protocol is intended as a reusable evaluation tool for future attribution proposals rather than a one-off audit of the five methods examined here. A researcher who has developed a new attribution method Z and wants to position it within the corruption-robustness landscape established by this paper can follow a four-step reference procedure. **Step 1 (data preparation)**: obtain CIFAR-10-C and CIFAR-100-C from Hendrycks and Dietterich [23] and apply Z to 500 test images per corruption–severity cell on both the ResNet-50 and ViT-B/16 checkpoints listed in §4.3, producing a clean attribution $\phi_Z^{\text{clean}}(x)$ and a corrupted attribution $\phi_Z^{\text{corrupt}}(x)$ for each image. **Step 2 (stability metrics)**: compute the five paired stability metrics defined in §4.5 (top-100 Feature Agreement, top-100 Rank Agreement, Spearman rank correlation, normalised L_2 displacement, and top-100 Sign Agreement) for each pair, following the reference implementation released in `run_fisher_z_analysis.py` and `run_by_omega_analysis.py`. **Step 3 (protocol-level statistics)**: subject the resulting 75-cell Spearman grid to the same pre-registered two-way factorial ANOVA with Benjamini–Hochberg FDR correction, Fisher-z and ω^2 sensitivity checks, and TOST equivalence testing that we apply in §5; this produces the four-hypothesis verdict summary that makes the new method directly comparable to the five baselines reported here. **Step 4 (comparative positioning)**: place method Z on the resolution-fair robustness ranking of §6.1 by comparing its cell-mean Spearman trajectory against the SmoothGrad baseline — a method that retains Spearman above 0.90 on at least twelve of fifteen corruption types at severity 3 earns a top-tier classification, one that remains above 0.50 earns a middle tier, and anything below 0.30 at severity 3 falls into the LIME/GradientSHAP failure regime. The entire procedure, from data download to verdict, runs end-to-end in the public repository at https://github.com/minyeong5691-svg/paper2-xai-corruption-audit with a single configuration change (substituting the attribution function pointer for Z).

6.5 Limitations

Several limitations constrain the scope of these conclusions. First, although the study spans two datasets and two architectures, both datasets are 32×32 CIFAR-scale images and both rely on the synthetic corruption suite of Hendrycks and Dietterich [23]. Taori et al. [56] established that robustness gains against synthetic corruptions do not reliably transfer to natural distribution shift, and Mintun et al. [37] showed that augmentations calibrated on the Hendrycks set often fail on perceptually dissimilar corruptions. Our findings therefore characterise attribution stability under controlled synthetic perturbation at CIFAR scale; replication on perceptually-dissimilar (ImageNet-C-bar, Mintun et al. [37]), scene-geometry-aware (3D Common Corruptions, Kar et al. [27]), and diffusion-generated natural (ImageNet-D, Zhang et al. [61]) corruption sets is a necessary next step, as is extension to 224×224 ImageNet resolution where LIME's superpixel resolution issue may attenuate and Grad-CAM's feature-map resolution increases. Second, the statistical power of the per-corruption H1 test is limited by the five-level severity design; a finer severity grid or a larger sample would likely convert more of the directional evidence into strict Bonferroni significance, though the Benjamini–Hochberg FDR correction already confirms 12 to 13 out of 15 corruptions as significant. Third, the Spearman rank correlation treats all attribution positions equally, whereas in practice users may care primarily about the top-k most attributed features; Rong et al. [47] showed that full-vector metrics can be insensitive to changes in the most decision-relevant features, and the supplementary analysis of Feature Agreement and Rank Agreement yields qualitatively similar rankings but with uniformly lower absolute values. Agarwal et al. [3] proposed relative stability metrics that normalise explanation change by input change, providing a Lipschitz-style bound that our absolute metrics do not capture. Fourth, our LIME implementation returns segment-level coefficients rather than pixel-level attributions, which may disadvantage LIME relative to the pixel-level methods in the Spearman comparison; we note, however, that even under this representation the Spearman values are below 0.30 in nearly all cells, indicating severe instability regardless of the metric's fairness to the method. Fifth, the ANOVA assumption of homoscedasticity is violated across all four architecture–dataset configurations (Levene's test $p < 0.01$ in all cases); we address this through non-parametric Friedman tests (§5.3.2) and mixed-effects models (§5.3.1) that confirm the parametric results, but acknowledge that a per-image random-effect model incorporating the full 500-sample variability within each cell would provide a stronger defence against pseudo-replication concerns. Sixth, the present stability framework operates at the dataset-aggregated (global) level: each cell-mean Spearman summarises hundreds of clean–corrupted attribution pairs without distinguishing whether a subset of images is explained robustly while others fail catastrophically. Individual-instance (local) stability — whether a particular image's attribution remains usable when that image is corrupted — is a complementary axis that the present design averages over and that future work could profitably disaggregate. Seventh, the Spearman correlation is bounded on $[-1, 1]$ and the rank/sign agreement metrics on $[0, 1]$, technically violating the unbounded-normal assumption of parametric ANOVA. Warton and Hui [58] argue that Fisher-z transformation for

correlations and logit transformation for bounded proportions should be default practice for inference on such scales. We report the untransformed results in the main text for interpretability and provide the Fisher-z transformed re-analysis in the public reproducibility package (`run_fisher_z_analysis.py`): all one-way architecture-level significance patterns are preserved, the maximum η^2 shift across configurations is 0.075, the CIFAR-10 architecture-by-method interaction remains non-significant ($F = 0.79$, $p = 0.46$) and the CIFAR-100 interaction retains—and in fact strengthens—its significance ($F = 7.09$, $p = 0.0014$ compared with $p = 0.035$ on the untransformed scale), confirming that the boundary-condition finding for the model-agnostic claim is not an artefact of the bounded-scale assumption.

6.6 Future Work

We pre-register five extensions that directly address the structural constraints of the present study. First, we will replicate the factorial grid on ImageNet-scale 224×224 inputs, pairing the SmoothGrad–IG–GradientSHAP pipeline with attention-specific attribution methods (Chefer et al. [12]) and with the perceptually-dissimilar corruption suite of Mintun et al. [37] and the scene-geometry-aware 3DCC suite of Kar et al. [27]; this cross-benchmark replication is the primary test of whether the ranking is an artefact of the Hendrycks corruption family or a stable property of the attribution algorithms. Second, we will recast the audit in causal rather than associational terms by computing Causal Shapley values (Heskes et al. [24]) and Asymmetric Shapley values (Frye et al. [18]) on the same factorial grid, and by reporting counterfactual explanations (Wachter et al. [57]; Mothilal et al. [41]) for the instance-level cells where associational SHAP and GradientSHAP disagree; this recasting is the empirical test of whether the observed instability reflects fundamental properties of associational XAI or a consequence of the specific family evaluated. Third, we will disaggregate the present cell-mean stability into instance-level stability scores and fit a per-image classifier that flags individual attributions as reliable or unreliable at deployment time, calibrated by the conformal prediction framework of Angelopoulos and Bates [5] to deliver distribution-free coverage guarantees, together with a regional disagreement decomposition following Laberge et al. [31]; the classifier yields a practical tool for corruption-aware explanation consumption and the regional decomposition separates global from local drivers of the method ranking. Fourth, we will complement the present empirical audit with a theoretical robustness certificate for SmoothGrad by applying the randomized-smoothing framework of Cohen et al. [13], whose Theorem 1 guarantees that the smoothed classifier g preserves its top-class prediction c_A within every ℓ_2 ball $\|\delta\|_2 < \sigma\Phi^{-1}(p_A)$; translated to attribution space, this yields the rank-order preservation statement that SmoothGrad at noise level σ retains the attribution ranking of every feature within the same ℓ_2 radius, turning the empirical stability that the present audit measures into a certifiable guarantee. We acknowledge the concurrent work of Anani et al. [4] on pixel-level certified explanations via randomized smoothing, and position the present empirical corruption audit as the complementary validation substrate that such theoretical certificates require. Fifth, we will recast the factorial ANOVA of the present study in a Bayesian hierarchical framework, fitting per-image random-effect models with the `brms` package of Bürkner [11] along

the lines developed by McElreath [35], and reporting posterior highest-density intervals and Bayes factors for the H1–H4 verdicts in place of the present frequentist F- and η^2-based inference; this recasting transforms the bounded-metric inferential concern raised in §6.5 into a principled posterior over method-level effects rather than a sensitivity analysis. Each extension is supported by the public reproducibility package released with this paper, which already contains the Fisher-z sensitivity (`run_fisher_z_analysis.py`) and the Benjamini–Yekutieli and omega-squared cross-corrections (`run_by_omega_analysis.py`) that we used to strengthen the statistical backbone of the present audit.

7 Conclusion

This paper presents a systematic vulnerability audit of five post-hoc XAI methods under fifteen common corruptions at five severity levels on CIFAR-10-C and CIFAR-100-C, evaluated across two architecturally distinct classifiers (ResNet-50 and ViT-B/16). Four findings emerge. First, attribution stability declines monotonically with corruption severity for all methods, confirmed by Benjamini–Hochberg FDR correction on 12 to 13 out of 15 corruptions across over 760,000 clean–corrupted attribution pairs. Second, the magnitude of the decline depends dramatically on the attribution method: on ResNet-50 CIFAR-10-C, SmoothGrad retains a Spearman correlation of 0.91 for brightness at severity 3, while LIME falls to 0.04 under identical conditions, a twenty-fold gap. Third, the method ranking is consistent across 96 % of all corruption–severity–dataset–architecture cells (144 out of 150), demonstrating that method choice is the dominant factor in attribution robustness, with an effect size ($\eta^2 > 0.84$) that dwarfs the architecture effect ($\eta^2 < 0.02$). Fourth, the architecture-by-method interaction is not significant on the ten-class CIFAR-10 task ($p = 0.71$) but reaches significance on the hundred-class CIFAR-100 task ($p = 0.035$), revealing that task complexity constitutes a boundary condition on the model-agnostic claim of Shapley-value attribution.

These results imply that XAI deployment in corruption-prone environments should prioritise noise-averaging methods, that method selection matters far more than architecture selection for attribution robustness, and that the reliability of the model-agnostic assumption should be re-examined as classification complexity increases. The practical vulnerability ranking derived from the audit — SmoothGrad first, Integrated Gradients second, GradientSHAP third, with Grad-CAM second on convolutional architectures and LIME always last under low-resolution inputs — provides a deployment-ready decision rule for practitioners who must choose an attribution method in anticipation of input corruption. The broader message is that the robustness literature, historically focused on classifier accuracy, should be extended to cover the behaviour of post-hoc explanation methods as a first-class evaluation target. The factorial audit framework demonstrated here, combining Krishna et al.'s [30] disagreement metric family with Hendrycks and Dietterich's [23] corruption suite under a factorial ANOVA design, offers one template for that extension.

Statements and Declarations

Funding.

The author did not receive support from any organisation for the submitted work.

Competing Interests.

The author has no competing financial or non-financial interests to declare that are relevant to the content of this article.

Ethics Approval.

Not applicable. This study does not involve human participants, human data, or animal subjects.

Consent to Participate.

Not applicable.

Consent for Publication.

Not applicable.

Author Contributions.

Minyeong Kim is the sole author and conducted all aspects of the study: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualisation, writing — original draft, and writing — review and editing.

Data Availability.

Analysis scripts, aggregated statistics (ANOVA, Friedman, bootstrap, mixed-effects, Fisher-z sensitivity, and Benjamini–Yekutieli multi-correction results), and figure-generation code supporting the findings of this study are publicly available at https://github.com/minyeong5691-svg/paper2-xai-corruption-audit and permanently archived on Zenodo (version of record: 10.5281/zenodo.19689329 for v1.1.0, the snapshot submitted with this manuscript). Raw attribution arrays are not deposited due to their aggregate size (approximately 120 GB); reproduction requires re-running the provided pipeline. CIFAR-10-C and CIFAR-100-C benchmarks are publicly available from Hendrycks and Dietterich [23] via Zenodo (DOIs 10.5281/zenodo.2535967 and 10.5281/zenodo.3555552). Pretrained model checkpoints used in this study are publicly available on the Hugging Face Hub: `edadaltocg/resnet50_cifar10` (94.65 % accuracy), `edadaltocg/resnet50_cifar100` (81.40 % accuracy), `edadaltocg/vit_base_patch16_224_in21k_ft_cifar10` (97.60 % accuracy), and `edadaltocg/vit_base_patch16_224_in21k_ft_cifar100` (89.26 % accuracy).

Code Availability.

The attribution pipeline, statistical analysis scripts (including `run_fisher_z_analysis.py` and `run_by_omega_analysis.py` for the Fisher-z

and Benjamini–Yekutieli sensitivity analyses cited in the main text), and figure-generation code are publicly available at https://github.com/minyeong5691-svg/paper2-xai-corruption-audit and permanently archived on Zenodo (version of record: 10.5281/zenodo.19689329 for v1.1.0; prior snapshot v1.0.0 at 10.5281/zenodo.19688326). Requirements: Python 3.10, PyTorch 2.x, Captum 0.7, scikit-learn 1.3, scipy 1.11, matplotlib 3.7.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process.

During the preparation of this work, the author used Anthropic's Claude (Opus 4.7) large language model across multiple phases of the research workflow, including: (i) English-language drafting and translation of Korean-language research notes and outlines into manuscript prose; (ii) implementation assistance for experimental code (attribution pipeline, statistical analysis scripts, figure-generation code); (iii) bibliographic verification and reference formatting; and (iv) structural review and cross-consistency checking of drafts. The research design, hypotheses, analytical plan, result interpretation, theoretical framing, and all intellectual content reflect the author's own scholarly judgement. After using the AI tool, the author reviewed and edited every piece of AI-assisted content and takes full responsibility for the content of the publication. No AI system is listed as an author.

References

[1] Aas K, Jullum M, Løland A (2021) Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence 298:103502. https://doi.org/10.1016/j.artint.2021.103502

[2] Abnar S, Zuidema W (2020) Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp 4190–4197

[3] Agarwal C, Krishna S, Saxena E, et al (2022) Rethinking stability for attribution-based explanations. In: ICLR 2022 Workshop on PAIRˆ2Struct

[4] Anani A, Lorenz T, Fritz M, et al (2025) Pixel-level certified explanations via randomized smoothing. In: Proceedings of the 42nd International Conference on Machine Learning (ICML), arXiv:2506.15499

[5] Angelopoulos AN, Bates S (2023) Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning 16(4):494–591. https://doi.org/10.1561/2200000101

[6] Bach S, Binder A, Montavon G, et al (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE 10(7):e0130140. https://doi.org/10.1371/journal.pone.0130140

[7] Barr DJ, Levy R, Scheepers C, et al (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language 68(3):255–278. https://doi.org/10.1016/j.jml.2012.11.001

[8] Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29(4):1165–1188. https://doi.org/10.1214/aos/1013699998

[9] Bolker BM, Brooks ME, Clark CJ, et al (2009) Generalized linear mixed models: A practical guide for ecology and evolution. Trends in Ecology & Evolution 24(3):127–135. https://doi.org/10.1016/j.tree.2008.10.008

[10] Breiman L (2001) Statistical modeling: The two cultures. Statistical Science 16(3):199–231. https://doi.org/10.1214/ss/1009213726

[11] Bürkner PC (2017) brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software 80(1):1–28. https://doi.org/10.18637/jss.v080.i01

[12] Chefer H, Gur S, Wolf L (2021) Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 782–791, https://doi.org/10.1109/CVPR46437.2021.00084

[13] Cohen JM, Rosenfeld E, Kolter JZ (2019) Certified adversarial robustness via randomized smoothing. In: Proceedings of the 36th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, vol 97. PMLR, pp 1310–1320

[14] Cumming G (2014) The new statistics: Why and how. Psychological Science 25(1):7–29. https://doi.org/10.1177/0956797613504966

[15] Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning

[16] Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An image is worth 16×16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR)

[17] Erion G, Janizek JD, Sturmfels P, et al (2021) Improving performance of deep learning models with axiomatic attribution priors and expected gradients. Nature Machine Intelligence 3:620–631. https://doi.org/10.1038/s42256-021-00343-w

[18] Frye C, Rowat C, Feige I (2020) Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. In: Advances in Neural Information Processing Systems (NeurIPS)

[19] Gundersen OE, Kjensmo S (2018) State of the art: Reproducibility in artificial intelligence. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pp 1644–1651

[20] Gupta R, Srinivas S, Ghorbani A (2025) Uncertainty-aware subset selection for robust feature attribution under distribution shift. arXiv preprint arXiv:251208445

[21] He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778, https://doi.org/10.1109/CVPR.2016.90

[22] Hedström A, Weber L, Bareeva D, et al (2023) Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. Journal of Machine Learning Research (MLOSS) 24(34):1–11

[23] Hendrycks D, Dietterich T (2019) Benchmarking neural network robustness to common corruptions and perturbations. In: Proceedings of the International Conference on Learning Representations (ICLR), URL https://openreview.net/forum?id=HJz6tiCqYm

[24] Heskes T, Sijben E, Bucur IG, et al (2020) Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In: Advances in Neural Information Processing Systems (NeurIPS)

[25] Huang X, Marques-Silva J (2024) On shapley values for feature attribution: A mathematical analysis of limitations. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)

[26] Janzing D, Minorics L, Blöbaum P (2020) Feature relevance quantification in explainable ai: A causal problem. Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)

[27] Kar OF, Yeo T, Atanov A, et al (2022) 3D common corruptions and data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

[28] Kaufman S, Rosset S, Perlich C, et al (2012) Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data 6(4):Article 15. https://doi.org/10.1145/2382577.2382579

[29] Koh PW, Sagawa S, Marklund H, et al (2021) Wilds: A benchmark of in-the-wild distribution shifts. In: Proceedings of the 38th International Conference on Machine Learning (ICML)

[30] Krishna S, Han T, Gu A, et al (2022) The disagreement problem in explainable machine learning: A practitioner's perspective

[31] Laberge G, Aivodji U, Hara S, et al (2024) Tackling the XAI disagreement problem with regional explanations. In: Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)

[32] Lakens D (2017) Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social Psychological and Personality Science 8(4):355–362. https://doi.org/10.1177/1948550617697177

[33] Li X, Wang S, Han J (2023) M4: A unified xai benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities, and models. In: Advances in Neural Information Processing Systems (NeurIPS)

[34] Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS), pp 4765–4774

[35] McElreath R (2020) Statistical Rethinking: A Bayesian Course with Examples in R and STAN, 2nd edn. Chapman and Hall/CRC, https://doi.org/10.1201/9780429029608

[36] Messick S (1995) Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist 50(9):741–749. https://doi.org/10.1037/0003-066X.50.9.741

[37] Mintun E, Kirillov A, Xie S (2021) On interaction between augmentations and corruptions in natural corruption robustness. In: Advances in Neural Information Processing Systems (NeurIPS)

[38] Miró-Nicolau M, Jaume-i Capó A, Moyà-Alcover G (2024) A comprehensive study on fidelity metrics for xai. Information Processing & Management 61(4):103684. https://doi.org/10.1016/j.ipm.2024.103684

[39] Molnar C, König G, Herbinger J, et al (2022) General pitfalls of model-agnostic interpretation methods for machine learning models. In: xxAI — Beyond Explainable AI. Springer, pp 39–68, https://doi.org/10.1007/978-3-031-04083-2_4

[40] Moradi M, Kim K, Al-Dhamari M, et al (2025) Comparative evaluation of explainability techniques for image data. Scientific Reports (Nature) 15:Article 7432. https://doi.org/10.1038/s41598-025-91894-0

[41] Mothilal RK, Sharma A, Tan C (2020) Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*), pp 607–617, https://doi.org/10.1145/3351095.3372850

[42] Nauta M, Trienes J, Pathak S, et al (2023) From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM

Computing Surveys 55(13s):Article 295. https://doi.org/10.1145/3583558

[43] Neely M, Schouten SF, Bleeker MJR, et al (2021) Order in the court: Explainable ai methods prone to disagreement. arXiv preprint arXiv:210503287

[44] Olejnik S, Algina J (2003) Generalized eta and omega squared statistics: Measures of effect size for some common research designs. Psychological Methods 8(4):434–447. https://doi.org/10.1037/1082-989X.8.4.434

[45] Pearl J (2009) Causality: Models, Reasoning, and Inference, 2nd edn. Cambridge University Press

[46] Ribeiro MT, Singh S, Guestrin C (2016) "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, pp 1135–1144, https://doi.org/10.1145/2939672.2939778

[47] Rong Y, Leemann T, Borisov V, et al (2022) A consistent and efficient evaluation strategy for attribution methods. In: Proceedings of the 39th International Conference on Machine Learning (ICML), pp 18770–18795

[48] Rozanec JM, Mladenic D, Fortuna B (2025) Xai-units: A benchmark for unit-level testing of feature attribution methods on procedurally generated models. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT)

[49] Selvaraju RR, Cogswell M, Das A, et al (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp 618–626, https://doi.org/10.1109/ICCV.2017.74

[50] Semenova L, Rudin C, Parr R (2022) On the existence of simpler machine learning models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp 1827–1858, https://doi.org/10.1145/3531146.3533232

[51] Shapley LS (1953) A value for n-person games. In: Kuhn HW, Tucker AW (eds) Contributions to the Theory of Games (AM-28), vol 2. Princeton University Press, p 307–317

[52] Sixt L, Granz M, Landgraf T (2020) When explanations lie: Why many modified bp attributions fail. In: Proceedings of the 37th International Conference on Machine Learning (ICML)

[53] Slack D, Hilgard S, Jia E, et al (2020) Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), pp 180–186

[54] Smilkov D, Thorat N, Kim B, et al (2017) Smoothgrad: Removing noise by adding noise. arXiv preprint arXiv:170603825

[55] Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), pp 3319–3328

[56] Taori R, Dave A, Shankar V, et al (2020) Measuring robustness to natural distribution shifts in image classification. In: Advances in Neural Information Processing Systems (NeurIPS)

[57] Wachter S, Mittelstadt B, Russell C (2018) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology 31(2):841–887

[58] Warton DI, Hui FKC (2011) The arcsine is asinine: the analysis of proportions in ecology. Ecology 92(1):3–10. https://doi.org/10.1890/10-0340.1

[59] Wasserstein RL, Lazar NA (2016) The asa statement on p-values: Context, process, and purpose. The American Statistician 70(2):129–133. https://doi.org/10.1080/00031305.2016.1154108

[60] Wiles O, Gowal S, Stimberg F, et al (2022) A fine-grained analysis on distribution shift. Proceedings of the International Conference on Learning Representations (ICLR)

[61] Zhang C, Pan F, Kim J, et al (2024) ImageNet-D: Benchmarking neural network robustness on diffusion synthetic object. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) – Highlight

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 03informationsheet.pdf