# Rethinking Explanation Evaluation Under the Retraining Scheme

**Yi Cai, Thibaud Ardoin, Mayank Gulati, Gerhard Wunder**

Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany
{yi.cai, thibaud.ardoin, mayank.gulati, g.wunder}@fu-berlin.de

## Abstract

Feature attribution has gained prominence as a tool for explaining model decisions, yet evaluating explanation quality remains challenging due to the absence of ground-truth explanations. To circumvent this, explanation-guided input manipulation has emerged as an indirect evaluation strategy, measuring explanation effectiveness through the impact of input modifications on model outcomes during inference. Despite the widespread use, a major concern with inference-based schemes is the distribution shift caused by such manipulations, which undermines the reliability of their assessments. The retraining-based scheme ROAR overcomes this issue by adapting the model to the altered data distribution. However, its evaluation results often contradict the theoretical foundations of widely accepted explainers. This work investigates this misalignment between empirical observations and theoretical expectations. In particular, we identify the *Sign* issue as a key factor responsible for residual information that ultimately distorts retraining-based evaluation. Based on the analysis, we show that a straightforward reframing of the evaluation process can effectively resolve the identified issue. Building on the existing framework, we further propose novel variants that jointly structure a comprehensive perspective on explanation evaluation. These variants largely improve evaluation efficiency over the standard retraining protocol, thereby enhancing practical applicability for explainer selection and benchmarking. Following our proposed schemes, empirical results across various data scales provide deeper insights into the performance of carefully selected explainers, revealing open challenges and future directions in explainability research.

**Code** — https://github.com/caiy0220/KAFT-C
**Extended version** — https://arxiv.org/abs/2511.08281

## 1 Introduction

Explainable AI (XAI) has gained significant attention over the past decade, motivated by the blooming real-world applications of data-driven models. These models learn decision rules implicitly from data; however, the self-learned decision processes — shaped by complex and non-linear model architectures — are often opaque, raising concerns about transparency and trustworthiness. As a foundational step towards model explainability, feature attribution aims at quantifying the contribution of input features to inquired predictions, thereby identifying the most influential evidence during inference. Encouraged by the growing demand for transparency, numerous attribution methods have been proposed, each claiming improved performance and distinct advantages. Despite this progress, an objective assessment of explanation quality has long been a challenge in XAI. Unlike standard machine learning tasks, feature attribution naturally lacks ground truth that exactly describes model behavior (Hedström et al. 2023). Moreover, the premise that humans have limited insight to model internals prohibits the application of human judgment in explanation evaluation. This circumstance creates a **paradox of explanation evaluation**: assessing explanation quality requires precise knowledge of model behavior, yet explainability is pursued because that behavior is unknown. In practice, the optimal explainer choice and its hyperparameter configuration vary across use cases, affected by factors such as input modality, model architecture, and feature space dimensionality. The challenge in explanation evaluation enhances the difficulty in selecting and configuring explainers for a given scenario.

Compromising with the explanation evaluation paradox, input manipulation has become a commonly adopted strategy for indirectly evaluating explanations by quantifying their effectiveness in affecting model outcomes. Intuitively, an effective explainer should identify influential features whose absence triggers significant prediction changes. We refer to such evaluation strategies as **inference schemes**, since they investigate explanation impacts at inference time. While inference schemes offer an efficient means of explanation evaluation, concerns have been raised regarding the reliability of resulting assessments. Hooker et al. (2019) point out that the performance changes may stem from distribution shifts caused by input manipulations, rather than the absence of influential features. Wang and Wang (2024) further emphasize this issue by showing that the evaluation outcome of inference schemes can be treated as an optimization objective, delivering manipulation sequences that maximize prediction change but deviate from the intent of explainability — to reflect model behaviors rather than to optimize the manipulation process.

To address distribution shift during input manipulation, Hooker et al. (2019) proposed ROAR, arguing that retraining the tested model is mandatory after feature occlusion. Specif-

ically, the **retraining scheme** creates manipulated copies of the dataset by removing the most influential features, then retrains the model on these modified copies. The performance degradation of the retrained model is interpreted as an indicator of explanation effectiveness. Surprisingly, ROAR reports that many popular explanation methods perform no better than random (Hooker et al. 2019). This striking conflict between empirical observations and the well-founded theoretical backgrounds of feature attribution methods (Sundararajan, Taly, and Yan 2017; Sundararajan and Najmi 2020; Erion et al. 2021) motivates the investigation presented in this work. The contributions of this paper are: 1. We investigate the catastrophic explainer performance reported by ROAR and identify an implicit, unrealistic assumption as the primary source of evaluation distortion; 2. We further strengthen this claim by highlighting the *Sign* issue as a concrete violation of ROAR's assumption; 3. We address the distortion by reframing the *retraining scheme* and propose additional evaluation variants for computational efficiency, promoting their practical applicability; 4. Our empirical results reveal limitations in certain attribution methods, exposing open challenges and future research directions in explainability research.

## 2 Related Work

The use of inference schemes for explanation evaluation spans from the early stages of XAI studies (Samek et al. 2016; Montavon, Samek, and Müller 2018) to recent developments (Cai and Wunder 2024; Muzellec et al. 2024). The core idea is to assess explanation quality indirectly by examining whether the attribution scores assigned to input features correspond to their actual impact on model predictions (Zeiler and Fergus 2014; Bach et al. 2015). For explainers that correctly capture relevant evidence, removing the highly attributed features should lead to substantial drops in model confidence. Samek et al. (2016) formalized this idea by recursively removing features in descending order of attributions and quantifying explanation quality with the area over the perturbation curve (AOPC). Subsequent studies extended this approach by altering the feature removal order (Petsiuk, Das, and Saenko 2018; Brocki and Chung 2023) and normalizing AOPC scores to mitigate the sensitivity of the measure to the original prediction confidence (Cai, Ardoin, and Wunder 2025). Similarly, Bhatt, Weller, and Moura (2020) and Yeh et al. (2019) evaluated explainer performance by measuring the correlation between attribution scores and prediction changes under random feature removal.

Although inference schemes are widely adopted for their efficiency, there has been criticism about the validity of their evaluation results due to the out-of-distribution (OOD) concern (Hooker et al. 2019; Jain et al. 2022). The difficulty in disentangling the effects of feature removal (intentional) from the consequences of distribution shift (unintentional) undermines the trustworthiness of inference-based evaluation results. To address this, Hooker et al. (2019) proposed incorporating model retraining after input manipulation as part of the evaluation process, introducing remove and retrain (ROAR). ROAR retrains the target model on manipulated inputs and interprets the retraining accuracy as a proxy for explanation quality. While retraining mitigates the OOD issue, the assessments by ROAR contradict theoretical expectations and may appear misleading (Lundberg et al. 2020).

Rong et al. (2022) and Park et al. (2023) sought to improve the retraining protocol by refining the choice of replacement values during feature removal. However, since most feature attribution methods are baseline-dependent (Lundberg and Lee 2017; Sundararajan, Taly, and Yan 2017), decoupling the removal process from the baseline value — which provides the necessary context to interpret the derived explanations — can introduce evaluation biases. In particular, the use of generative models for feature removal (Park et al. 2023) loses precise control over the manipulation process, challenging the transparency of the evaluation framework. Beyond concerns of bias, the intensive computational overhead associated with exhaustive retraining remains unaddressed, limiting the broader impact of retraining-based evaluation schemes.

## 3 Distortion in Retraining Scheme

To investigate the source of distortion in retraining-based evaluation, we focus on ROAR — the standard protocol in this category. Under the ROAR scheme, an effective explainer is expected to induce greater performance degradation after retraining on manipulated data with the highest attributed features removed (Hooker et al. 2019). This expectation implicitly poses the following evaluation question:

**Question 1** (ROAR)**.** *Does the tested explainer identify all task-relevant features?*

### 3.1 Distortion by Residual Information

Let[1] $\boldsymbol{N} = \{x_1, \ldots, x_n\}$ be the full feature set and $\boldsymbol{S} \subset \boldsymbol{N}$ represent the subset of the most important features for the model function $\boldsymbol{f}(\boldsymbol{N})$. Borrowing the notion of mutual information from information theory, let $I(\boldsymbol{N}; y)$ represent the utility of the input features for predicting the target label $y$. The core expectation of Question 1 can be formalized as:

$$I(\boldsymbol{N}; y) \gg 0 \overset{\text{manipulate}}{\Longrightarrow} \tilde{I}(\boldsymbol{N} \backslash \boldsymbol{S}; y) \approx 0 \tag{1}$$

where $\tilde{I}$ denotes the remaining utility after distribution shift caused by explanation-guided input manipulation. While this expectation appears reasonable — the most important features $\boldsymbol{S}$ should be identified and removed — a subtle gap between model knowledge and data information can lead to deviations from the expected manipulation results.

Specifically, feature attribution reflects model knowledge in solving a specific task, which does not necessarily align with the true utility of input features as represented by the underlying data distribution. This misalignment can arise because standard machine learning theory does not guarantee that a trained model will capture and exploit *all relevant features* in accordance with their ground-truth importance. As a result, residual information can persist in the manipulated data and distort evaluation outcomes, reflected as:

$$\tilde{I}(\boldsymbol{N} \backslash \boldsymbol{S}; y) \gg 0 \tag{2}$$

[1]Throughout the paper, vectors and sets are typeset in boldface, whereas scalars are presented in plain font.

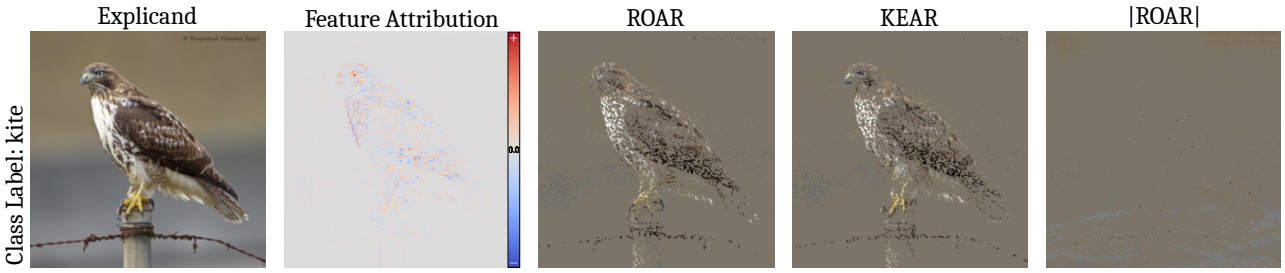


Figure 1: Example of the *Sign* issue. The first two columns show the original input and the attributions derived by IG. The subsequent columns show manipulated inputs after removing $90\%$ of features following three evaluation schemes: ROAR, KEAR, and |ROAR|. In ROAR, the target-overlapping negative features leak information about the target object, leading to evaluation distortion. Despite the visual similarity, retraining following KEAR achieves over $20\%$ higher accuracy, indicating that the tested explainer correctly identifies contributing features for the queried decision. As a control group, |ROAR| removes all relevant features and behaves as expected from Equation (1), yielding the lowest accuracy, $6\%$ below random removal.

Retraining after manipulation fits the model to the residual information, producing unexpectedly high accuracies and leading to an underestimation of explanation quality. Practically, residual information can arise from various sources, such as model underfitting, which overlooks relevant features, or feature redundancy, which renders some informative features completely unused. While some factors can be controlled through careful experimental setup, we particularly highlight the *Sign* issue, which can provably harm evaluation validity under mild conditions.

### 3.2 The *Sign* Issue

Under the highest-first occlusion strategy, features with negative attributions survive the manipulation process. Despite their negative scores, features with large attribution magnitudes are often highly task-relevant. During retraining, such "*negative*" features can leak substantial task-relevant information, leading to unexpectedly high performance of retrained models. We refer to the distortion caused by negatively attributed features as the "*Sign*" issue.

In contrast to positively attributed features that support the target decision $y$, negative attributions arise when features serve as evidence for an alternative class $y^* \neq y$ while being shared with $y$. We refer to such shared features as *secondary evidence* for $y$, formally defined below.

**Definition 1** (Secondary Evidence)**.** Two disjoint feature subsets $\boldsymbol{S_1}, \boldsymbol{S_2} \subseteq \boldsymbol{N}$ are called the *primary* and *secondary* evidence, respectively, for class $y$ if they satisfy:

$$I(\boldsymbol{S_1}; y) > I(\boldsymbol{S_2}; y) \gg 0 \;\text{ and }\; I(\boldsymbol{S_2}; y^*) > I(\boldsymbol{S_2}; y)$$

The second condition indicates the greater utility of $\boldsymbol{S_2}$ for $y^*$, which can lead to negative attributions w.r.t. $y$. When removing features in descending attribution order, these negatively scored secondary features will remain and turn into primary evidence for predicting $y$, as formalized in Theorem 1 (see Appendix A.1 for the detailed proof).

**Theorem 1** (Increasing Utility of Secondary Features)**.** *The mutual information between $\boldsymbol{S_2}$ and $y$ increases due to distribution shift after input manipulation:*

$$\tilde{I}(\boldsymbol{S_2}; y) > I(\boldsymbol{S_2}; y) \gg 0$$

The increase in utility pinpoints the emergence of residual information, as described in Equation (2). Notably, the presence of shared features across multiple classes is common in classification tasks, where the targets often involve overlapping low-level concepts. The statement in Theorem 1 is related to the *information swapping* issue raised by Lundberg et al. (2020) in the context of binary classification; our discussion formalizes and generalizes this concern to a broader multi-class setting.

### 3.3 Weak Positive Contributor

Figure 1 qualitatively illustrates information leakage caused by the *Sign* issue under the highest-first occlusion strategy. Although Theorem 1 establishes the utility of negatively attributed features, it may still appear counterintuitive when regions representing the target object receive negative attribution scores. This subsection concretizes a case where even positively correlated features can receive negative attributions, thereby demonstrating that the target-overlapping negatives are not necessarily indicative of poor explanation quality.

Let $\boldsymbol{o}(\boldsymbol{N})$ be the output of the final dense layer in a classifier (i.e., the classification head) and $o_y(\boldsymbol{N})$ be the activation corresponding to class $y$. A common practice in classification tasks is to apply a softmax function to normalize the raw model outcomes at inference time: $\boldsymbol{f}(\boldsymbol{N}) = \boldsymbol{\sigma}(\boldsymbol{o}(\boldsymbol{N}))$. However, this softmax normalization can unintentionally result in target-overlapping negative attributions. In certain cases, even features that positively activate the output node $o_y(\boldsymbol{N})$ can dramatically receive negative attributions for the final prediction $f_y(\boldsymbol{N})$. We refer to such features as *weak positive contributors*.

**Definition 2** (Weak Positive Contributor)**.** A feature $x_i$ is called a *weak positive contributor* to class $y$ if its attribution $\xi_i^{o_y}$ to the output logit $o_y$ satisfies:

$$\underbrace{1 < \exp(\xi_i^{o_y})}_{\text{Positive Contribution}} \overbrace{< \mathbb{E}_{y^* \neq y}\left[\exp(\xi_i^{o_{y^*}})\right]}^{\text{Weak Contribution}} \tag{3}$$

By writing the expectation $\mathbb{E}_{y^* \neq y}$, we interpret the final

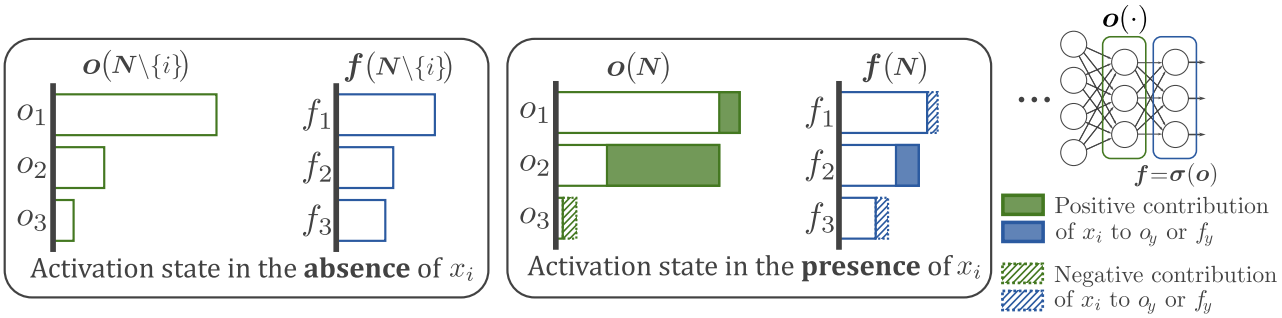


Figure 2: Example of a weak positive contributor. The feature $x_i$ is a weak contributor to class 1: it positively influences $o_1$ but yields a negative contribution to $f_1$.

model output $\boldsymbol{f}(\boldsymbol{N})$, normalized by a softmax layer, as a probability distribution over labels conditioned on $y^* \neq y$. The first inequality in (3) indicates the positive contribution to $o_y$, whereas the second implies that this contribution is weak relative to its expected exponential contributions to the remaining classes. Please note that $\xi_i^{o_y}$ refers to the ground-truth attribution, i.e., the quantity that feature attribution methods seek to estimate. Therefore, the discussion of weak positive contributors is independent of any specific explanation method and reflects intrinsic properties of the model's predictive behavior.

**Theorem 2** (Negative Attribution)**.** *The attribution $\xi_i^{f_y}$ of a weak positive contributor w.r.t. the final prediction is negative, despite its positive attribution to $o_y$.*

Figure 2 showcases that a positively associated feature can negatively affect the prediction confidence of the corresponding class, as stated in Theorem 2 (see Appendix A.2 for proof and further details).

## 4 Keep and Fine-tune (KAFT)

Based on the previous discussion, a direct remedy for the *Sign* issue is to perform feature manipulation in descending order of attribution magnitudes. This variant is denoted by |ROAR|, where the absolute value operator implies that only attribution magnitudes are used to rank features. We refer to this occlusion strategy as *relevant*-first occlusion, since high-magnitude attributions reflect feature relevance, regardless of their signs. However, |ROAR| remains sensitive to residual information arising from redundant features, which are informative but receive low attribution scores with their significance masked by redundancies.

An alternative is to reframe the occlusion strategy completely, thereby relieving the evaluation scheme from the over-restrictive expectation of $\tilde{I}(\boldsymbol{N}\backslash\boldsymbol{S}; y) \approx 0$. Instead of removing the most important features, we argue that retaining the top-ranked features for retraining yields more reliable assessments of explanation quality. Keep and retrain (KEAR) reframes the evaluation question as follows:

**Question 2** (KEAR)**.** *Does the tested explanation method correctly identify influential features for model decisions?*

The lowest-first occlusion strategy was first mentioned by Hooker et al. (2019). Appendix A.3 further discusses the role of KEAR in prior work and interprets the differences in empirical observations. By reversing the occlusion priority, an effective explanation method should capture relevant information learned by the target model, resulting in less performance degradation after retraining with the same proportion of retained features. Compared to ROAR, Question 2 loosens the assumption of exact model-data alignment and shifts the focus towards model behavior.

### 4.1 Retraining? Fine-tuning!

Parallel to distortion in evaluation results, the heavy computational cost also limits the practical use of retraining schemes for explainer selection and benchmarking. The evaluation costs are twofold: the expense of deriving explanations for numerous training data, and the overhead of the multiple retraining epochs. Unlike inference schemes, which typically require explanations for only the test set or even a subset of it, retraining schemes must derive explanations for the entire dataset, including training, validation, and test splits, to reproduce the training environment. Additionally, these manipulated entries are visited repeatedly during retraining to refine model performance. While such costs are manageable in small-scale settings, efficiency concerns become significant as training pipelines become more complicated. This challenge is pronounced when benchmarking explainers on downstream models fine-tuned from pretrained versions, where reproducing the entire training process becomes infeasible.

Given that retraining serves to adapt the target model to the disrupted data distribution, we propose replacing full retraining with model fine-tuning. When the manipulated dataset is viewed as representing a different distribution that preserves partial information overlap with the original data, fine-tuning offers a lightweight alternative for model adaptation under distribution shift. By leveraging the explained model as initialization, fine-tuning requires fewer training samples and epochs. More specifically, keep and fine-tuning (KAFT) creates manipulated copies of a subset of training samples and, instead of training from scratch, fine-tunes the explained model on the manipulated data. As in the original retraining scheme, the performance of the fine-tuned model serves as an indicator of explanation quality.

Following this direction, we further simplify the evaluation scheme by restricting keep and fine-tuning to the

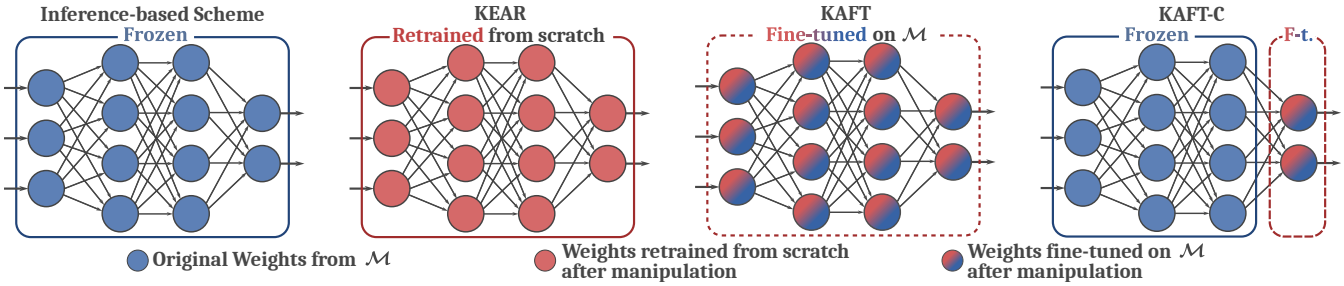


Figure 3: Visualization of model updates under different evaluation schemes. The inference scheme is sensitive to distribution shifts as it does not update the tested model. KAFT and its variants overcome this limitation while mitigating distortions inherent in other retraining schemes. Particularly, KAFT-C evaluates explanation quality by assessing the utility of preserved features through restricted updates, striking a balance between resolving distribution shifts and maintaining focus on model behavior.

classification head (KAFT-C)[2]. This simplification is inspired by the observation that the hidden layers of a model function primarily as feature extractors, deriving and summarizing high-level representations that support the final decisions at the classification head. An effective explainer should capture input features that contribute to these informative high-level representations (from the model perspective) and ensure that they are preserved during explanation-guided manipulation. Although input perturbation may degrade model performance by disrupting contextual dependencies, the feature extraction performed by the hidden layers should remain at least partially effective if the most relevant input components are preserved. Consequently, adapting the classification head allows the model to reuse the extracted representations under the altered context, achieving higher classification accuracy without full model updates. From the efficiency perspective, when fine-tuning with $10\%$ of the data and $10\%$ of the training epochs used for full retraining, KAFT-C achieves a $100\times$ speedup over KEAR. Figure 3 provides an overview of the three retraining alternatives and compares them to the inference-based evaluation scheme.

Compared to full retraining or unrestricted fine-tuning, fine-tuning only the classification head further simplifies the retraining process and strengthens the connection to the explained model. This modification better aligns the evaluation process with the goal of feature attribution methods. By freezing the hidden layers of the tested model, the evaluation concentrates on the original behavior of the target model without completely altering its internal characteristics. This restriction further mitigates the distortion caused by feature redundancy when applying *relevant-first* occlusion, as unused features are no longer reorganized through full retraining. Following this direction, we propose another alternative evaluation scheme $|$RAFT-C$|$ (remove and fine-tuning on the classification head following attribution magnitude). Compared to KAFT-C, which reflects the utility of the highest-attributed features, $|$RAFT-C$|$ assesses the ability of an explainer to identify irrelevant features, providing a complementary perspective for assessing explanation quality.

[2]Abbreviation naming rules are detailed in Appendix A.4.

# 5 Experiments

To demonstrate the validity of the proposed evaluation schemes and their differences from ROAR, we begin with small-scale experiments with a carefully selected set of explainers. These explainers are chosen based on their different theoretical backgrounds, serving as a reference for explanation quality. Following the small-scale tests, we adopt KAFT-C and $|$RAFT-C$|$, the most efficient retraining variants, for benchmarking an extended set of attribution methods in large-scale settings.

Across all tests, full retraining (ROAR and KEAR) is performed on the complete dataset, whereas $20\%$ and $10\%$ of the training samples are used for fine-tuning and classification-head-only fine-tuning, respectively. All reported values are averages over 5 repetitions for better reliability of the results.

## 5.1 Small-scale Experiments

**Datasets and Classifiers** The small-scale experiments are conducted on three publicly available datasets: MNIST (LeCun et al. 1998), CIFAR10 (Krizhevsky and Hinton 2009), and STL10 (Coates, Ng, and Lee 2011), covering input types with varying color models and resolutions. A simple CNN, WideResNet (Zagoruyko and Komodakis 2016), and EfficientNet-B0 (Tan and Le 2019) are trained on MNIST, CIFAR10, and STL10, respectively. All classifiers are trained from scratch for efficient reproduction of the exact training environments during retraining.

**Feature Attribution Methods** Three gradient-based attribution methods with various theoretical backgrounds are selected: Vanilla Gradient (**VG**) (Simonyan, Vedaldi, and Zisserman 2014), SMOOTHGRAD (**SG**) (Smilkov et al. 2017), and Integrated Gradients (**IG**) (Sundararajan, Taly, and Yan 2017). When configured with identical query budgets, SG and IG share the same complexity. However, IG is distinguished by explicitly modeling feature absence with a baseline. This baseline, which also defines feature absence during input manipulation, anchors the attribution process, allowing more accurate quantification of how feature presence contributes relative to absence. Founded on its theoretical strength, IG is expected to outperform SG, which in turn should outperform VG by reducing sensitivity to gradient noise with smoothing. Appendix A.4 provides additional experimental details

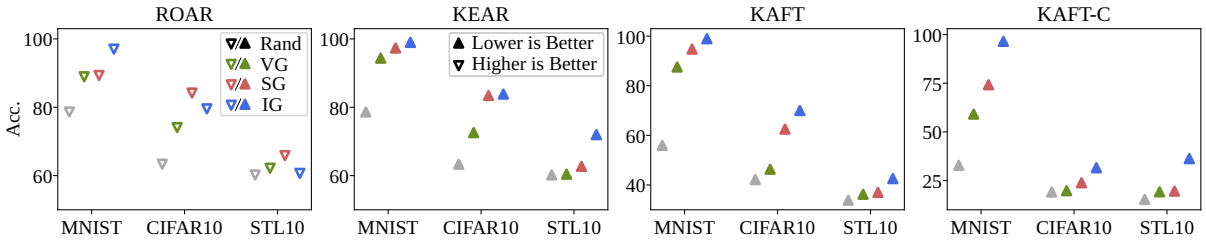


Figure 4: Explainer performance reflected by different schemes with 90% of features removed.

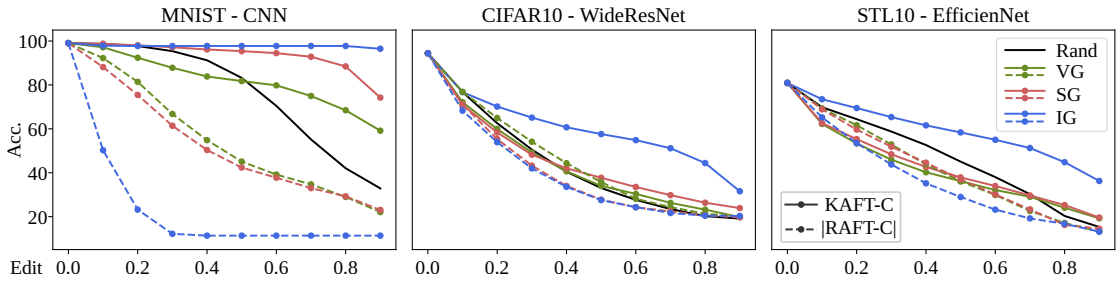


Figure 5: Explanation quality in *small-scale* settings tested by KAFT-C and |RAFT-C|.

| ΔAcc. | VG | SG | IG |
|---|---|---|---|
| MNIST-CNN | 26.01 | 39.48 | **72.48** |
| CIFAR10-WRN | -1.44 | 5.24 | **20.03** |
| STL-EfficientNet | -0.41 | 0.92 | **21.67** |

Table 1: Quantification of explainer performance in the small-scale settings. ΔAcc. represents the difference in prediciton power, computed as the area between thede gradtion curves produced by KAFT-C and |RAFT-C| as presented in Figure 5.

about the explainers and models, including implementation specifics and configurations for (re-)training and fine-tuning.

**Small-scale Results** We start with empirically setting the manipulation ratio to 90% and retrain the models on the manipulated datasets. Figure 4 illustrates the retraining performances[3]. As an indicator of explanation quality, the interpretation of retraining accuracies depends on deletion priority: for KEAR, higher accuracy suggests better explanation quality, as it reflects the preservation of influential features; for ROAR, lower accuracy is preferred, which indicates the effective removal of task-relevant features.

First, we concentrate on the relative ranking of competitors within each evaluation scheme. Our results reproduce the findings of Hooker et al. (2019), showing that all competitors perform worse than random when tested under the ROAR scheme, as demonstrated by the limited performance degradation. As discussed in Section 3, these misleading assessments arise from residual information, where the remaining negatively attributed features lead to information leakage and, consequently, unexpectedly high retraining accuracies. Moreover, the variability in explainer rankings across test cases — for example, IG ranking highest with STL10-EfficientNet but lowest with MNIST-CNN — suggests that ROAR's assumption (inherited from Question 1) is highly sensitive to the characteristics of tested models. This sensitivity leads to varying degrees of distortion depending on the test scenario. In contrast, the results by KAFT and other variants of lowest-first occlusion match the theoretically grounded expectation: IG consistently outperforms other explainers due to its explicit use of a baseline. Regardless of how a model is updated, the KAFT family delivers coherent assessments of explanation quality, as reflected by the consistent competitor rankings.

To further demonstrate the effectiveness of gradient-based approaches and to show that explainers capture not only influential features but also distinguish them from irrelevant ones, we adopt KAFT-C and |RAFT-C|, reporting evaluation results across manipulation ratios ranging from 0.1 to 0.9 in increments of 0.1. Figure 5 shows model performance degradation curves as the manipulation ratio increases. The solid and dashed lines correspond to performance degradation under different manipulation priorities. For an effective explainer, the solid line should lie above the random baseline (black), while the dashed line should fall below. Among the competitors, only IG exhibits curves that are well-separated by the random baseline across all three test cases, indicating the effectiveness of its explanations. Complementing the line plots, explainer performance is further quantified by the area between the solid and dashed lines associated with each explainer. This area reflects the difference in prediction power between irrelevant-first and relevant-first manipulations. Table 1 confirms the superiority of IG with the largest ΔAcc., highlighting its capability in identifying influential features to model decisions. Appendix A.5 provides additional qualitative examples that visualize the manipulated

[3]Error bars are not presented for visual clarity. Please refer to Appendix A.5 for the complete results and raw numerical values.

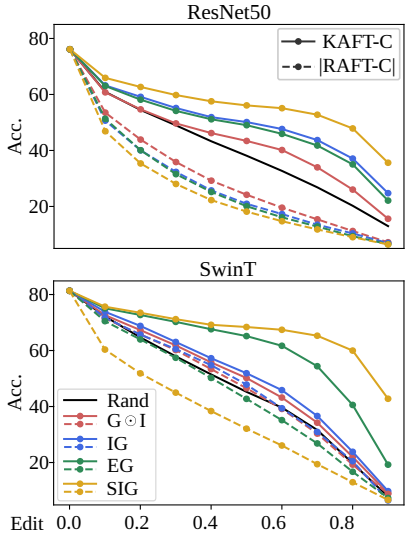


Figure 6: Explanation quality in *large-scale* settings tested by KAFT-C and |RAFT-C|. The plots show a selected subset of competitors for clarity. Table 2 reports ΔAcc. for all tested explainers.

| ΔAcc. | VG | SG | G⊙I | IG | EG | SIG |
|---|---|---|---|---|---|---|
| ResNet50 | 2.84 | 9.58 | 13.04 | 21.50 | 20.73 | **30.07** |
| SwinT | -5.02 | 14.43 | 2.14 | 3.31 | 15.59 | **30.05** |

Table 2: Quantification of explainer performmance with ΔAcc. for the large-scale settings on ImageNet (Figure 6).

inputs, emphasizing the *Sign* issue.

### 5.2 Large-scale Experiments

**Additional Competitors** We extended the competitor set with Gradient⊙Input (**G⊙I**) (Shrikumar, Greenside, and Kundaje 2017), Expected Gradients (**EG**) (Erion et al. 2021), and Smoothed Integrated Gradients (**SIG**). The last approach is an ensemble of IG and SG, combining theoretical grounding with denoising effects. All selected competitors are widely used and generalizable feature attribution methods.

**Dataset and Classifiers** We benchmark the competitors on classifiers trained with ImageNet1k (Russakovsky et al. 2015). For architectural diversity, we consider *ResNet50* (He et al. 2016), a CNN-based model, and *SwinT* (Liu et al. 2021), a transformer-based model. For both architectures, pretrained versions are used.

**Large-scale Results** We repeat the KAFT-C and |RAFT-C| evaluations in the large-scale settings, with results presented in Figure 6 and Table 2. For ResNet50, attribution methods that explicitly incorporate a baseline generally outperform those that do not. Notably, G⊙I, which can be viewed as a special case of IG with a single observation on the interpolation path, improves explanation quality by amplifying raw gradients with the corresponding input. In both cases, limited performance degradation is achieved by the best-performing explainer (SIG) under the restrictive fine-tuning setting. This observation supports the use of fine-tuning as an effective and substantially more efficient alternative to full retraining for handling distribution shifts.

On the other hand, IG exhibits a performance collapse when tested on SwinT. As shown in the right panel of Figure 6, the degradation curves produced by IG are nearly indistinguishable from the random baseline in both manipulation orders, indicating a failure to identify features relevant to model decisions. We consider two possible causes for the failure of IG on SwinT: 1. a long plateaux of gradient saturation, and 2. the cancellation effect due to feature interactions. First, a key motivation of the integration in IG is to resolve gradient saturation (Sundararajan, Taly, and Yan 2017). However, when the interpolation path between an explicand and the chosen baseline traverses a prolonged saturation region, the evenly interpolated instances will be dominated by the saturated gradients. As a result, the integrated gradients become biased towards the saturation, failing to address the issue. The similar performance of IG and G⊙I on SwinT supports this interpretation — a long saturation plateau undermines the effectiveness of integration over an even interpolation, reducing IG to a single-sample explanation, as in the case of G⊙I. Second, the straightline integration path can cause underestimation of feature importance due to cancellation effects from feature interactions. Consider the simple function $f(x_1, x_2) = (x_1 - x_2)^2$ as a concrete example. When explaining the decision at $(1, 1)$ relative to the baseline $(0, 0)$, IG assigns $0$ attribution to both features, as their opposing contributions cancel each other out along the interpolation path, despite both being influential to the outcome. This limitation can be mitigated by averaging over multiple integration paths that deviate from the straightline, which decomposes interactions into individual contributions. This simple solution also underlies the improved performance of SIG: by adding noise to the interpolated points, SIG implicitly averages over observations from diverse paths, thereby revealing interaction effects and recovering the performance of IG.

## 6 Conclusion

This work addresses the gap between the empirical assessments by retraining-based schemes and the theoretical foundations of gradient-based approaches. By identifying the cause of evaluation distortion, we propose several alternatives that reframe the evaluation objective to resolve this issue. Among them, the fine-tuning variant KAFT-C largely reduces the computational overhead associated with retraining-based evaluation. Our empirical results generally align with theoretical expectations for the tested explainer — except for the test on SwinT, where IG exhibits a performance collapse. We discuss two possible causes of the collapse: saturation plateaux and cancellation effects, both linked to the integration path. The discussion on the limitation highlights the need for further investigation into path selection and allocation of feature interactions. Overall, this work provides a new toolset for explanation evaluation, which differs from traditional inference-based schemes, and offers a complementary perspective for benchmarking attribution methods.

## Acknowledgements

Yi Cai, Thibaud Ardoin, and Gerhard Wunder were supported by the Federal Ministry of Education and Research of Germany (BMBF) in the program of "Souverän. Digital. Vernetzt.", joint project "AIgenCY: Chances and Risks of Generative AI in Cybersecurity", project identification number 16KIS2013. Mayank Gulati and Gerhard Wunder were supported by BMBF joint project "6G-RIC: 6G Research and Innovation Cluster", project identification number 16KISK020K.

## References

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*, 10(7): e0130140.

Bhatt, U.; Weller, A.; and Moura, J. M. 2020. Evaluating and aggregating feature-based model explanations. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, 3016–3022. International Joint Conferences on Artificial Intelligence Organization.

Brocki, L.; and Chung, N. C. 2023. Feature perturbation augmentation for reliable evaluation of importance estimators in neural networks. *Pattern Recognition Letters*, 176: 131–139.

Cai, Y.; Ardoin, T.; and Wunder, G. 2025. GEFA: A general feature attribution framework using proxy gradient estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, 6165–6192. PMLR.

Cai, Y.; and Wunder, G. 2024. On gradient-like explanation under a black-box setting: When black-box explanations become as good as white-box. In *Proceedings of the 41st International Conference on Machine Learning*, 5360–5382. PMLR.

Coates, A.; Ng, A.; and Lee, H. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 215–223. PMLR.

Erion, G.; Janizek, J. D.; Sturmfels, P.; Lundberg, S. M.; and Lee, S.-I. 2021. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3(7): 620–631.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. IEEE.

Hedström, A.; Bommer, P.; Wickstrøm, K. K.; Samek, W.; Lapuschkin, S.; and Höhne, M. M.-C. 2023. The meta-evaluation problem in explainable AI: Identifying reliable estimators with MetaQuantus. *Transactions on Machine Learning Research*.

Hooker, S.; Erhan, D.; Kindermans, P.-J.; and Kim, B. 2019. A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Jain, S.; Salman, H.; Wong, E.; Zhang, P.; Vineet, V.; Vemprala, S.; and Madry, A. 2022. Missingness bias in model debugging. In *International Conference on Learning Representations*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical Report, University of Toronto.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022. IEEE.

Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and Lee, S.-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1): 56–67.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30: 4768–4777.

Montavon, G.; Samek, W.; and Müller, K.-R. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73: 1–15.

Muzellec, S.; Fel, T.; Boutin, V.; Andéol, L.; Vanrullen, R.; and Serre, T. 2024. Saliency strikes back: How filtering out high frequencies improves white-box explanations. In *Proceedings of the 41st International Conference on Machine Learning*, 37041–37075. PMLR.

Park, Y.-H.; Seo, J.; Park, B.; Lee, S.; and Jo, J. 2023. Geometric remove-and-retrain (GOAR): Coordinate-invariant explainable AI assessment. In *XAI in Action: Past, Present, and Future Applications*.

Petsiuk, V.; Das, A.; and Saenko, K. 2018. RISE: Randomized input sampling for explanation of black-box models. In *Proceeedings of the British Machine Vision Conference 2018*. British Machine Vision Association.

Rong, Y.; Leemann, T.; Borisov, V.; Kasneci, G.; and Kasneci, E. 2022. A consistent and efficient evaluation strategy for attribution methods. In *Proceedings of the 39th International Conference on Machine Learning*, 18770–18795. PMLR.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115: 211–252.

Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; and Müller, K.-R. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11): 2660–2673.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, 3145–3153. PMLR.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations*.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. SmoothGrad: Removing noise by adding noise. In *Proceedings of the 2017 ICML Workshop on Visualization for Deep Learning*.

Sundararajan, M.; and Najmi, A. 2020. The many Shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning*, 9269–9278. PMLR.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, 3319–3328. PMLR.

Tan, M.; and Le, Q. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 6105–6114. PMLR.

Wang, Y.; and Wang, X. 2024. Benchmarking deletion metrics with the principled explanations. In *Proceedings of the 41st International Conference on Machine Learning*, 51569–51595. PMLR.

Yeh, C.-K.; Hsieh, C.-Y.; Suggala, A.; Inouye, D. I.; and Ravikumar, P. K. 2019. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. In *Proceeedings of the British Machine Vision Conference 2016*. British Machine Vision Association.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.