On Pitfalls of *RemOve-And-Retrain*: A Data Processing Inequality Perspective

Junhwa Song [1] **Keumgang Cha** [2] **Junghoon Seo** [1]

Abstract

The RemOve-And-Retrain (ROAR) benchmark is widely used to evaluate feature attribution methods, yet its validity remains underexplored from an information-theoretic perspective. We show that model- and data-agnostic post-processing of attribution maps (transformations that, by the data processing inequality, *cannot* add information about the decision function) can often improve ROAR scores. This means that an improved ROAR ranking is not, by itself, evidence that an attribution map carries more information about the model. We trace this failure mode to a bias toward spatially blurry masks. Experiments on CIFAR-10, SVHN, and CUB-200 show a consistent association between blurriness and ROAR performance, a pattern that also appears in the ROAD variant. We provide guidelines for more cautious removal-based benchmarking, with implications for validating mechanistic understanding of neural network internals.

1. Introduction

Mechanistic interpretability seeks to understand the internal computations of neural networks, from identifying task-relevant circuits (Wang et al., 2023) and validating hypotheses about learned representations (Bricken et al., 2023) to reverse-engineering decision-making processes (Sharkey et al., 2025). Feature attribution methods, which measure the contribution of each input feature to a model's decisions, are foundational building blocks in this enterprise. They underpin circuit discovery, concept validation, and safety-oriented model auditing, making their *reliable evaluation* a first-order concern for the field. Without sound benchmarks, we cannot distinguish methods that genuinely reveal decision-relevant structure from those that merely produce visually appealing but uninformative saliency maps (Samek et al., 2021).

Despite significant efforts to develop quantitative evaluation methods (Samek et al., 2016; Dabkowski & Gal, 2017; Adebayo et al., 2018b; Hooker et al., 2019; Adebayo et al., 2020), selecting the most effective metric remains contentious (Duan et al., 2024). One influential method is the RemOve-And-Retrain (ROAR) protocol (Hooker et al., 2019), widely adopted for benchmarking attribution methods (Schramowski et al., 2020; Liang et al., 2020; Kim et al., 2019a; Yang et al., 2020; O'Shaughnessy et al., 2020; Khakzar et al., 2021; Hartley et al., 2021; Chefer et al., 2021; Ismail et al., 2021; Meng et al., 2022; Hong et al., 2025) and for developing novel evaluation metrics (Bhatt et al., 2020; Ismail et al., 2020; DeYoung et al., 2020; Zhang et al., 2021). Additionally, (Kim et al., 2019b) extended ROAR to measure fixed attribution interpretability of various model functions $f(\cdot;\theta)$.

In this study, we question the reliability of the ROAR metric. As attribution evaluation directly affects which methods are trusted in downstream mechanistic analyses and safety-relevant model auditing, understanding the failure modes of widely-used benchmarks is essential. Our key contributions are:

- We prove, via the data processing inequality, that model/data-agnostic post-processing of attributions can *improve* ROAR scores while carrying *less* information about the decision function — a concrete hypothesis that we test empirically.
- We provide a structural causal model of the ROAR data-generation process, a formal counterexample, and experiments on real-world datasets.
- We reveal a persistent blurriness bias in ROAR (and ROAD) metrics, and offer practical guidelines for more cautious benchmarking of interpretability methods.

[1]PIT-IN Corp., Anyang-si, South Korea [2]SI Analytics Co., Ltd., Daejeon-si, South Korea. Correspondence to: Junghoon Seo <sjh@pitin-ev.com>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

Algorithm 1 **R**em**O**ve-**A**nd-**R**etrain (ROAR) with attribution post-processing

Require: $f(x;\theta)$: classifier; $e(x, f(\cdot;\theta), y)$: attribution; $k(a)$: post-processing; D_{train}, D_{test}; T: drop rates.
Ensure: V: {drop rate $\rightarrow$ ROAR accuracy}.

1: $\hat{\theta} \leftarrow \arg\min_\theta \mathbb{E}_{(x,y)\in D_{train}}[L(f(x;\theta), y)]$; $V \leftarrow \{\}$
2: $A_{tr} \leftarrow [e(x, f(\cdot;\hat{\theta}), \arg\max_i f_i(x;\hat{\theta}))$ for $(x, y) \in D_{train}]$
3: $A_{te} \leftarrow [e(x, f(\cdot;\hat{\theta}), \arg\max_i f_i(x;\hat{\theta}))$ for $(x, y) \in D_{test}]$
4: $A_{tr} \leftarrow [k(a)$ for $a \in A_{tr}]$; $A_{te} \leftarrow [k(a)$ for $a \in A_{te}]$
5: **for** $t \in T$ **do**
6: $\quad M_{tr} \leftarrow [\{a > \text{Pctl}(a, t)\}$ for $a \in A_{tr}]$
7: $\quad M_{te} \leftarrow [\{a > \text{Pctl}(a, t)\}$ for $a \in A_{te}]$
8: $\quad D'_{tr} \leftarrow [((1 - m) \odot x, y)$ for $((x, y), m) \in \text{zip}(D_{train}, M_{tr})]$
9: $\quad D'_{te} \leftarrow [((1 - m) \odot x, y)$ for $((x, y), m) \in \text{zip}(D_{test}, M_{te})]$
10: $\quad \hat{\theta}_{new} \leftarrow \arg\min_\theta \mathbb{E}_{(x',y)\in D'_{tr}}[L(f(x'), y)]$
11: $\quad Acc \leftarrow \mathbb{E}_{(x',y)\in D'_{te}}[\mathbb{1}_{\arg\max_i f_i(x';\hat{\theta}_{new})=y}]$
12: $\quad V[t] \leftarrow Acc$
13: **end for**

2. Preliminaries

2.1. Notions and Notations

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input space and class label space, respectively. We consider a C-class multi-class pre-softmax classifier $f(x;\theta) : \mathcal{X} \times \Theta \rightarrow \mathbb{R}^C$ with parameters θ, and define $\mathcal{F}$ as the set of all possible functions $f(\cdot;\theta) : \mathcal{X} \rightarrow \mathbb{R}^C$ for a given θ. A feature importance measure (or explainer), also referred to as an attribution method (Ancona et al., 2018), is a function $e(x, f(\cdot;\theta), y) : \mathcal{X} \times \mathcal{F} \times \mathcal{Y} \rightarrow \tilde{\mathcal{X}}$ that identifies which input features are important in determining a class decision $y \in \mathcal{Y}$ given an input x and a function $f(\cdot;\theta)$ (Lundberg & Lee, 2017; Hooker et al., 2019). As an example, input-gradient (Baehrens et al., 2010; Simonyan et al., 2014), a widely recognized basic feature importance measure, is defined as $e(x, f(\cdot;\theta), y) = \frac{\partial f(\mathrm{x};\theta)y}{\partial \mathrm{x}}\Big|_{\mathrm{x}=x}$, where $f(\mathrm{x};\theta)_y$ is the y-th indexed value of $f(\mathrm{x};\theta)$. In this paper, the output of the explainer is referred to as an "attribution map." For consistency, variables are denoted using uppercase letters, while functions or values are denoted using lowercase letters, with some exceptions.

2.2. RemOve-And-Retrain (ROAR)

To make our work self-contained, we present a brief introduction here (Hooker et al., 2019). Algorithm 1 outlines the pseudo-code for the ROAR procedure in Python style. The proposal of ROAR was motivated by the limitations of existing modification-based evaluation methods (Bach et al., 2015; Samek et al., 2016). At the time, the dominant approach for modification-based evaluation was a type of sequential procedure that only used the test dataset D_{test}. This approach involves applying attribution methods (line 3), sorting feature importance estimates and modifying features with high attribution values (lines 8 and 10), and measuring the performance drops of the trained classifier (lines 12 and 13). It should be noted that, in this case, $\hat{\theta}_{new} \leftarrow \hat{\theta}$ because the parameters of the classifier are never changed. However, (Dabkowski & Gal, 2017; Fong & Vedaldi, 2017) have pointed out that it is difficult to determine whether the performance drops are due to the significance of the feature importance estimates or to the out-of-distribution nature of the input samples (Zheng et al., 2025).

To address the issue of a distribution gap between training data and testing data, the ROAR approach employs a re-training strategy. This involves applying data processing techniques to the training data, D_{train}, in a similar manner to how they are applied to the testing data in the previously described modification-based evaluation. Specifically, after sorting the attribution for each instance based on feature importance (Line 8), a new training dataset, D'_{train}, is created by dropping a certain percentage of features (as determined by the parameter t, see Line 6). The model is then re-trained on this synthesized training dataset to produce new parameters, $\hat{\theta}_{new}$ (Line 10). These new parameters are used to evaluate model performance, rather than the original parameters, $\hat{\theta}$, in order to alleviate the distribution mismatch between the training and testing datasets (Line 11).

The recent work (Rong et al., 2022) has demonstrated that mutual information (MI) between data variable and class variable can be utilized as a surrogate for attainable accuracy in pixel removal strategies that involve retraining, as higher MI generally leads to higher accuracy (Hellman & Raviv, 1970; Schramowski et al., 2020). In the ROAR approach, MI between the modified data variable and the class variable is particularly important, as it determines the obtainable accuracy. In this context, low MI between the modified data variable and the class variable is desirable, as it results in a decrease in accuracy and improved benchmarking results.

3. Sanity Check in View of DPI

3.1. Sketch of Our Argument

A central intuition behind ROAR is that, after re-training on the modified dataset, the attainable test accuracy at a given drop rate is governed by how much class-relevant information remains in the modified input. In particular, Rong et al. (2022) argue (under mild regularity assumptions) that larger mutual information between the modified input and the label typically allows higher re-trained accuracy, and conversely that lower mutual information implies lower attainable accuracy. Motivated by this, we view the ROAR score at drop rate t as a proxy for $I(X'_t; Y)$, where X'_t denotes the random variable of the ROAR-modified input at

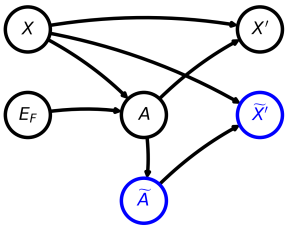


Figure 1. A structural causal model for the generation of modified data variables.

drop rate t. Hence, *a lower ROAR accuracy corresponds to a smaller* $I(X'_t; Y)$, which is "better" in the ROAR sense.

What ROAR is supposed to measure. Conceptually, ROAR is intended to rank explainers by how well their attributions reflect the model's decision mechanism: if an explainer highlights features that the model truly relies on for predicting Y, then removing those features (and retraining) should destroy predictive information and reduce accuracy. The potential pitfall we stress is that ROAR *only observes the modified data* X' (equivalently, the mask), and thus it is not immediate that a low $I(X'; Y)$ must be caused by an attribution that is informative about the model/decision function.

Variables and the data-generation graph. We formalize the dependency structure using the structural causal model in Fig. 1. Here X is the original input, Y is the label, and A is the attribution map. To reason about "information about the model/decision function", we introduce an abstract random variable

$$Z \triangleq (F, E), \tag{1}$$

where F denotes the (possibly random) decision function (e.g., trained parameters) and E denotes the explainer identity (or, more generally, any model-side information accessible to the explainer). For example, in the empirical setting below, F can be the trained ResNet classifier and E can indicate whether the explainer is Input-Gradient, Integrated Gradients, or Grad-CAM; conditioning on a fixed input $X = x$, $I(Z; A \mid X = x)$ captures the extent to which the produced map still varies with such model-side choices. The attribution is generated as

$$A = e_E(X, F) \quad \text{(possibly with internal randomness)}, \tag{2}$$

and ROAR constructs a binary mask M_t by thresholding A at drop rate t (cf. Algorithm 1), then forms the modified input

$$X'_t = \phi(X, M_t) = (1 - M_t) \odot X. \tag{3}$$

Model/data-agnostic post-processing. We consider a post-processing map $k(\cdot)$ applied to the attribution:

$$\widetilde{A} = k(A, U), \tag{4}$$

where U is an independent random seed (optional; deterministic k is the special case without U). Crucially, "agnostic" means that k does not access (X, Y, F, E) except through A. ROAR is then run with $\widetilde{A}$ in place of A, producing $\widetilde{M}_t$ and $\widetilde{X}'_t$.

Why conditioning on X matters. In Fig. 1, A is influenced by both X and Z, so marginal statements like $I(Z; A)$ can be confounded by the shared cause/collider structure. To isolate the information carried by the attribution *about* Z for a fixed input instance, the right object is the conditional mutual information $I(Z; A \mid X)$.

Theorem 3.1 (Conditional data processing for agnostic post-processing)**.** *Assume* $\widetilde{A} = k(A, U)$ *where* $U \perp (X, Y, Z, A)$ *and* k *uses* (A, U) *only. Then, for the causal graph in Fig. 1,*

$$I(Z; \widetilde{A} \mid X) \le I(Z; A \mid X). \tag{5}$$

Moreover, for any (possibly randomized) measurable mapping ψ*,*

$$I\big(Z; \psi(X, \widetilde{A}) \mid X\big) \le I(Z; \widetilde{A} \mid X), \tag{6}$$

and in particular,

$$I(Z; \widetilde{M}_t \mid X) \le I(Z; \widetilde{A} \mid X), \tag{8a}$$

$$I(Z; \widetilde{X}'_t \mid X) \le I(Z; \widetilde{A} \mid X). \tag{8b}$$

The proof is deferred to Appendix A.

Interpretation. Theorem 3.1 states a one-way guarantee: *agnostic post-processing cannot increase the information the attribution carries about the model-side variable* Z *for a fixed input*. This is the correct DPI-based sanity check. Importantly, DPI *does not* imply any ordering such as $I(Z; X'_t \mid X) \ge I(Z; \widetilde{X}'_t \mid X)$, because $X'_t = \phi(X, h_t(A))$ and $\widetilde{X}'_t = \phi(X, h_t(\widetilde{A}))$ are generally *two different* nonlinear transformations of A (one does not have to be a processing of the other). Therefore, a ROAR improvement after applying k is not ruled out by DPI.

Theorem 3.2 (ROAR can be improved while destroying model/explainer information)**.** *There exist random variables* (X, Y, Z)*, an attribution rule producing* A*, and an agnostic post-processing* k *such that*

$$I(Z; \widetilde{A} \mid X) = 0 \quad \textit{but}$$
$$I(\widetilde{X}'_t; Y) < I(X'_t; Y) \tag{7}$$

for a fixed drop rate t*. Consequently, an attribution pipeline can obtain a* strictly better *ROAR score (i.e., smaller* $I(\cdot; Y)$*) even though the post-processed attribution contains* no *information about* Z*.*

Proof. Let $X = (X_1, X_2)$ with $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(1/2)$ and let $Y = X_1$. Let $Z \in \{1, 2\}$ be uniform and independent of X (think of Z as indexing "what the explainer/model claims is important"). Define the attribution map (two "pixels") by

$$A = \begin{cases} (2,1) & \text{if } Z = 1, \\ (1,2) & \text{if } Z = 2. \end{cases} \tag{8}$$

Thus A perfectly reveals Z, so $I(Z; A \mid X) = H(Z) = 1$ bit.

Let $t = 50\%$ and define ROAR's mask operator h_t to select the coordinate whose attribution exceeds the empirical 50-th percentile (no ties occur below). Then $M_t = h_t(A)$ removes feature Z and keeps the other one, hence

$$X'_t = \begin{cases} (0, X_2) & \text{if } Z = 1, \\ (X_1, 0) & \text{if } Z = 2. \end{cases} \tag{9}$$

When $Z = 2$, X'_t retains $X_1 = Y$ and therefore $I(X'_t; Y \mid Z = 2) = H(Y) = 1$ bit; in particular $I(X'_t; Y) > 0$.

Now define an agnostic post-processing k (deterministic) by

$$\widetilde{A} = k(A) \triangleq (A_1 + A_2,\ A_1 + A_2 - \varepsilon), \tag{10}$$

for any fixed $\varepsilon \in (0, 1)$. Since $A_1 + A_2 = 3$ for both possible values of A, we have $\widetilde{A} \equiv (3, 3 - \varepsilon)$ deterministically, hence $I(Z; \widetilde{A} \mid X) = 0$.

Because $\widetilde{A}_1 > \widetilde{A}_2$, the 50%-drop mask $\widetilde{M}_t = h_t(\widetilde{A})$ always removes the first coordinate, so

$$\widetilde{X}'_t = (0, X_2). \tag{11}$$

But $Y = X_1$ is independent of X_2, so $I(\widetilde{X}'_t; Y) = 0$. Therefore $I(\widetilde{X}'_t; Y) < I(X'_t; Y)$ while $I(Z; \widetilde{A} \mid X) = 0$, proving the claim. □

Consequence for ROAR. Theorem 3.2 shows that, even under the strongest DPI sanity check (post-processing cannot add information about Z), the ROAR objective $I(X'_t; Y)$ can strictly decrease after applying a Z-agnostic transformation k. This means that a *better* ROAR score does not necessarily certify that the attribution is *more informative* about the model/decision function; it may instead reflect how the mask-generation pipeline interacts with the data distribution.

Our target phenomenon. Motivated by Theorem 3.1 and Theorem 3.2, we test for the following behavior in realistic vision settings: there exists a simple, model/data-agnostic post-processing k that improves ROAR (reduces $I(X'_t; Y)$, hence reduces re-trained accuracy) while, by DPI, never increasing the information about Z.

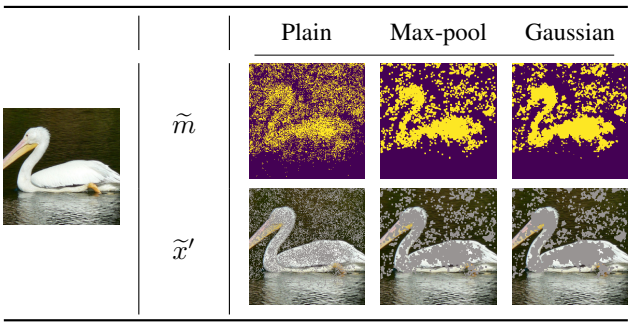


Figure 2. Model/data-agnostic attribution post-processings. The term 'Plain' indicates no post-processing, i.e., $k(a) = a$. The leftmost image represents the original input image x. The feature importance measure used in this illustration is input-gradient.

Conjecture 3.3. *There exists a post-processing function k such that, for some drop rate t,*

$$I(\widetilde{X}'_t; Y) < I(X'_t; Y), \tag{17a}$$

$$\textit{while} \quad I(Z; \widetilde{A} \mid X) \leq I(Z; A \mid X). \tag{17b}$$

3.2. Instantiation of Post-processing

We now instantiate $k(\cdot)$ with two simple post-processings that are *agnostic* in the sense of Theorem 3.1: (i) a Gaussian filter and (ii) a max-pooling (maximum) filter, both applied directly to the attribution map a. These operations only transform the attribution values on the attribution grid and do not access the input x, the decision function $f(\cdot; \theta)$, or the explainer identity beyond the produced map itself. Examples are shown in Fig. 2. Thus, the plain pipeline uses A, whereas the processed pipeline uses $\widetilde{A} = k(A)$, exactly matching the setup of Theorem 3.1. When the processed pipeline obtains lower ROAR/ROAD accuracy, it is the empirical analogue of Eq. (17a), while Eq. (17b) holds by construction of the agnostic k. We do not estimate these mutual informations directly; re-trained accuracy is used as the observable ROAR/ROAD proxy.

The purpose of these choices is not to propose a better explainer, but to operationalize Conjecture 3.3: if such k consistently improves ROAR/ROAD scores, then the benchmark can be optimized by procedures that, by DPI, cannot add model-side information.

Rationale of choice: from PixelRandom to BlockRandom. Our choice is motivated by a simple observation about mask geometry. Consider two model-agnostic baselines: *PixelRandom* erases uniformly random pixels covering t percent of the mask area, while *BlockRandom* erases a single random rectangle covering t percent of the area (Fig. 3). Although both are uninformative about the model,

Figure 3. (a) Input Image, (b) PixelRandom, and (c) BlockRandom. Examples of *PixelRandom* and *BlockRandom*.

Table 1. The results of the ROAR method applied to the CUB-200 dataset using PixelRandom and BlockRandom. Each column represents the attribution drop rate $t \in T$.

	10%	30%	50%	70%	90%
PixelRandom	0.7556	0.7269	0.7099	0.6772	0.6195
BlockRandom	0.6586	0.4384	0.2815	0.1705	0.0533

BlockRandom can remove more structured content in natural images, leading to a larger reduction in class-relevant information. This intuition is reflected by the ROAR results in Table 1.

This motivates testing blur-inducing transformations (Gaussian smoothing and max-pooling) that tend to produce more spatially coherent ("block-like") masks than the unprocessed attribution. If ROAR rewards such mask shapes, then these agnostic transformations may improve ROAR/ROAD scores despite (by Theorem 3.1) never increasing information about Z.

4. Experiments

4.1. Experimental Settings

We perform a series of experiments on three image classification datasets as (Hooker et al., 2019; Rong et al., 2022) do: CIFAR-10 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011), and CUB-200 (Welinder et al., 2010). For our neural network implementation, we utilize the PyTorch model zoo implementation of ResNet18 (He et al., 2016) with a customized kernel size for CIFAR-10 and SVHN, and ResNet50 for CUB-200.

We compare the ROAR performance of several feature importance estimates with a fixed neural network architecture on each dataset. These estimates include Input-Gradient (Baehrens et al., 2010; Simonyan et al., 2014) (Grad), Grad * Input (Shrikumar et al., 2017) (GI), Integrated Gradient (Sundararajan et al., 2017) (IG), SmoothGrad (Smilkov et al., 2017) (SG), VarGrad (Adebayo et al., 2018a) (VG), and Grad-CAM (Selvaraju et al., 2017) (GC). Following (Hooker et al., 2019), we primarily report a squared version of these attribution methods, abbreviated as Grad^2, GI^2, IG^2, SG^2, VG, and GC^2. It should be noted that squaring does not apply to VG, as it is itself a squaring method. Additionally, we include SG-SQ, the pre-squared version of SmoothGrad (Hooker et al., 2019), as a comparative methodology.

Table 2. Number of mean comparisons in which post-processing lowers the final accuracy relative to the plain attribution map. Lower accuracy is better under ROAR/ROAD. Each cell aggregates 3 datasets $\times$ 9 attribution/baseline maps $\times$ 3 drop rates, for 81 comparisons.

Benchmark	Max-pooling	Gaussian
ROAR	74/81	76/81
ROAD	76/81	78/81

For reference, we also report the plain, non-squared version of Input-Gradient (Grad), PixelRandom (Rand), and the squared version of Sobel edge detection (Sobel^2) as baselines. To simplify our results, we use attribution drop rates of 10%, 30%, and 50% in the ROAR protocol. To account for trial variability, all results are reported as the average of five trials. Further experimental details can be found in Appendix B.

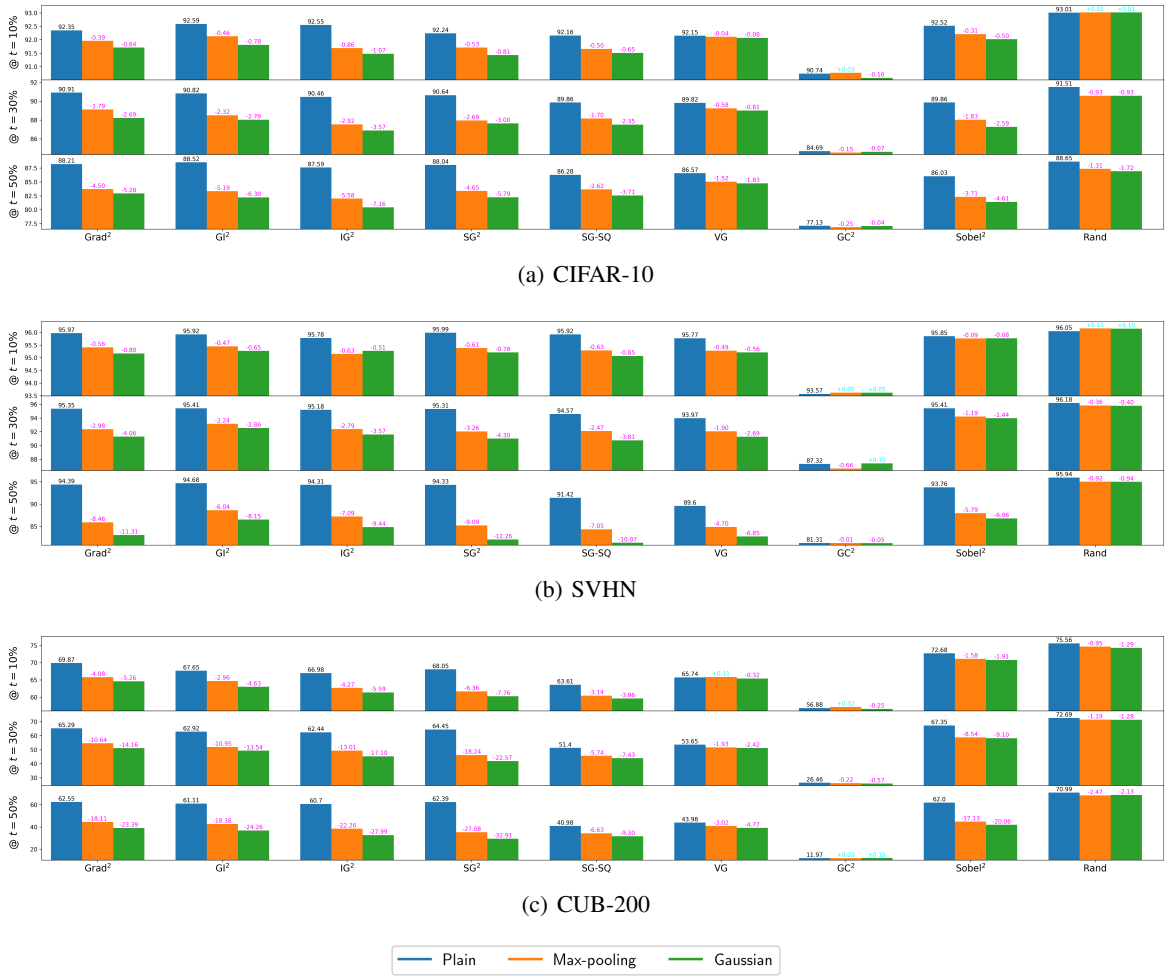
4.2. Effects of Agnostic Post-processing

To examine the effect of agnostic post-processing on the ROAR metric, we analyze its impact across datasets and attribution methods. The results in Figure 4 show that max-pooling and Gaussian filtering frequently lower the re-trained accuracy, which corresponds to a better ROAR score.

The benefits of post-processing are not uniform and depend on the dataset, attribution method, and drop rate. In our experiments, post-processing was particularly effective on the CUB-200 dataset compared to the other two datasets. Max-pooling and Gaussian filtering generally yielded favorable ROAR scores, while the GC^2 method produced inconclusive results for all datasets and drop rates. The unusual behavior of GC^2 is covered in Section 4.4. These results support Conjecture 3.3: agnostic transformations can obtain better ROAR scores without adding information about the model-side variable Z. Because individual bar differences can be small, Table 2 summarizes the directional pattern over the plotted mean accuracies.

4.3. ROAD Analysis

We conducted a similar experiment on the ROAD benchmark (Figure 5), a variation of ROAR designed to reduce confounds in removal-based evaluation (Rong et al., 2022). ROAD changes the debiasing/removal procedure, but it still constructs its score from attribution-derived masks and the resulting modified inputs. Therefore, the agnostic post-processing intervention considered here occurs upstream of the ROAD scoring step, and the DPI argument for $A \mapsto \widetilde{A} = k(A)$ applies in the same way.



(a) CIFAR-10

(b) SVHN

(c) CUB-200

Figure 4. The effect of Gaussian filtering and max-pooling on the ROAR metric. The labels 'P', 'M', and 'G' refer to 'Plain', 'Max-pooling', and 'Gaussian filter', respectively. The numbers at the top of each column indicate the attribution drop rate. For ease of interpretation, the results of max-pooling and Gaussian filtering are expressed as the difference from the plain method. A decrease in model accuracy is indicated by a magenta '$-$' sign, while an increase is represented by a cyan '$+$' sign. For optimal viewing, it is recommended to zoom in.

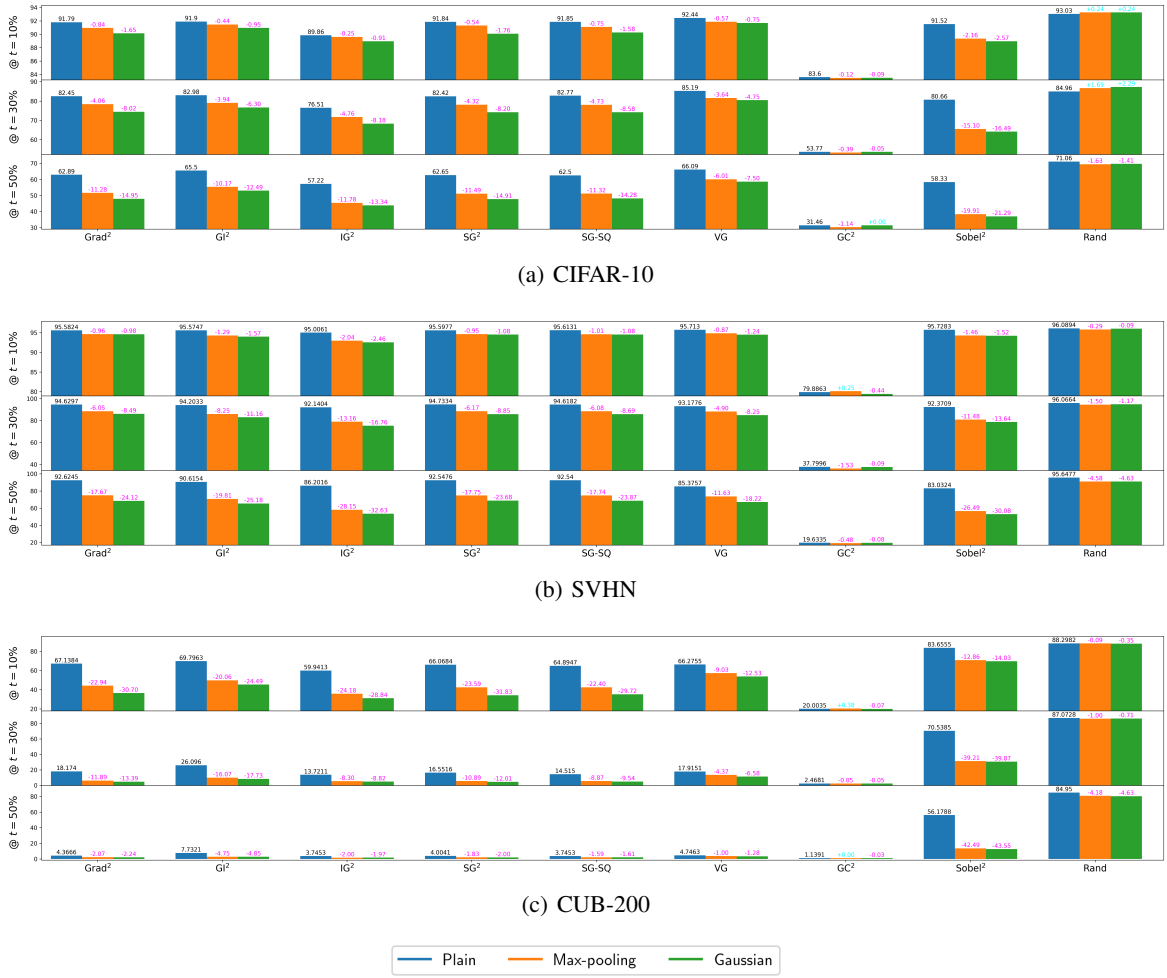
As summarized in Table 2, max-pooling and Gaussian filtering lower the mean ROAD accuracy in most comparisons. This suggests that ROAD's debiasing step does not by itself eliminate the mask-geometry sensitivity studied in this paper, although the exact magnitude remains method- and dataset-dependent.

4.4. Delving into Relationship between ROAR Benchmark and Blurriness

Going further, we examine the relationship between the ROAR metric and attribution blurriness, as discussed in Section 3.2. We measure blurriness through the average total variation (TV) of attribution masks for the training set A_{train} in each ROAR trial; lower TV corresponds to blurrier masks. The results in Figure 6 show that lower-TV attributions tend to yield lower final accuracy in the ROAR protocol. Results on CIFAR-10 and SVHN are reported in Figures 8 and 9 in Appendix C.

Our findings indicate that for most attribution methods and drop rates, max-pooling and Gaussian filtering reduce both the total variation and the model's performance. However, this trend does not hold for the GC^2 method, as its unique approach (derived from gradients on intermediate feature maps rather than back-propagation on inputs) results in attribution maps with a smaller resolution, which are less affected by the post-processing $k(\cdot)$. Similar patterns were observed in the CIFAR-10 and SVHN datasets (Figures 8 and 9).

Next, we examine the relationship between model performance and total variation without post-processing. Figure 7 presents a scatter plot of all attribution methods' results for



(a) CIFAR-10

(b) SVHN

(c) CUB-200

Figure 5. The effect of Gaussian filtering and max-pooling on the ROAD metric. The conventions used in this figure are the same as those in Figure 4, except that this figure relates to the ROAD benchmark instead of the ROAR benchmark.

Table 3. Coefficient of determination (R^2) for the linear relationship between total variation of squared attribution masks and final ROAR test accuracy in Figure 7.

Dataset	$t = 10\%$	$t = 30\%$	$t = 50\%$
CIFAR-10	0.84	0.85	0.82
SVHN	0.92	0.90	0.95
CUB-200	0.91	0.90	0.92

each dataset and drop rate. It shows a strong linear association between total variation and model accuracy, excluding results after post-processing. Table 3 summarizes the corresponding R^2 values, and the non-squared attribution results in Figure 10 show a similar pattern. Thus, because lower accuracy is treated as better in ROAR, blurrier/lower-TV masks can be favored even when the blurring is introduced by an agnostic post-processing step.

5. Related Works

Evaluating the quality of attribution methods is an active area (Nauta et al., 2022). (Doshi-Velez & Kim, 2017) categorize evaluation into functionally-grounded, human-grounded, and application-grounded approaches; (Zhou et al., 2021) further divide functionally-grounded evaluations into clarity, broadness, simplicity, completeness, and soundness. Our focus is on the ROAR protocol (Hooker et al., 2019), a widely-used soundness-oriented methodology.

ROAR and its variants have been used across image tasks (Chefer et al., 2021; Yang et al., 2020), NLP (Ismail et al., 2021; Zhang et al., 2021), graph analysis (Funke et al., 2022), and time-series (Liang et al., 2020; Ismail et al., 2020), and ROAR has been adopted as the primary metric in (Meng et al., 2022)'s standard benchmark. Our research focuses on the intriguing and unintended behavior of the ROAR protocol.

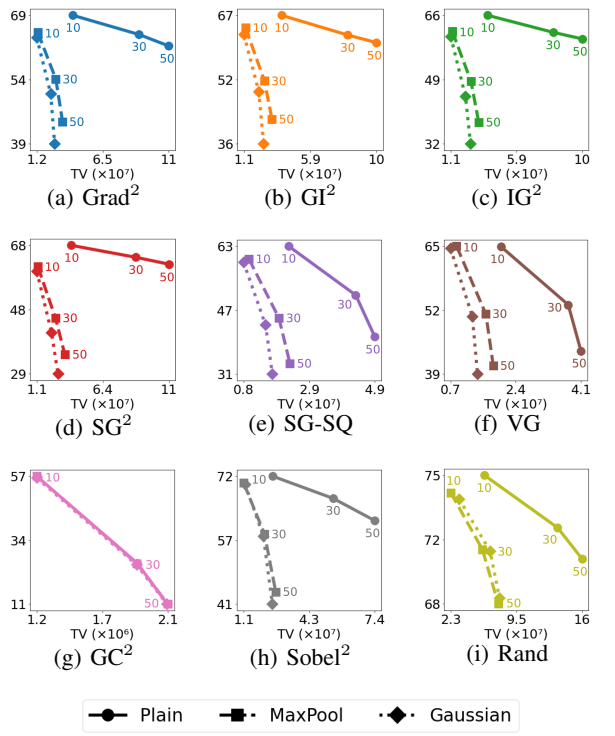


Figure 6. Effects of Gaussian filtering and max-pooling on the total variation of attribution masks (TV) and model accuracy on the CUB-200 dataset. The number displayed above each point indicates the attribution drop rate. The y-axis represents the final test accuracy (%).

Existing Discussions on the Limitations of ROAR. (Hooker et al., 2019) noted that feature redundancy can prevent ROAR from distinguishing meaningless attributions from redundant features, though no such case was observed in practice. (Rong et al., 2022) identified information leakage through mask shape and proposed RemOve-And-Debias (ROAD) to reduce confounding factors; coordinate/geometry sensitivity was further studied by Park et al. (2023). Our findings address a different issue, namely the dependency of the data generation process, rather than mask-based information leakage. As discussed in Section 4.3, ROAD still relies on attribution-derived masks and modified inputs, and our experiments suggest that its debiasing step does not eliminate sensitivity to agnostic post-processing. More recently, Bora et al. (2025) proposed frameworks to estimate the inconsistency of saliency metrics themselves.

6. Discussion

Practical Takeaways. Our work does not aim to discourage the use of ROAR (or ROAD) in future research, but rather to emphasize the importance of exercising caution when employing and interpreting these methods. We present two main practical takeaways.

- Ranking attribution methods solely by ROAR scores can be misleading (Canha et al., 2025). We recommend jointly reporting factors that contribute to ROAR bias, such as attribution total variation (Fig. 7).
- ROAR trends are partly influenced by data structure, independent of model and explainer quality. Similar relationships between data structure and feature redundancy have been reported in time series (Radovic et al., 2017) and graph structures (Liu et al., 2022).

Implications for Mechanistic Interpretability. Attribution methods increasingly serve as components in broader mechanistic interpretability pipelines, including circuit analysis (Wang et al., 2023), sparse dictionary validation (Bricken et al., 2023), and safety-oriented monitoring of deployed models. An evaluation metric with such confounds, if used as the sole criterion for selecting attributions, may propagate biases into downstream mechanistic conclusions. As Sharkey et al. (2025) note, validating interpretability methods using benchmarks remains an important open problem. Our findings contribute a concrete diagnostic result clarifying what ROAR-style benchmarks can and cannot measure, and highlight the need for metrics that assess information about the *decision function* more directly rather than relying only on proxy accuracy drops that may be confounded by mask geometry.

Extension and Limitations. The core issue, namely confounds in the data generation process, extends to other perturbation-based evaluation methods (Samek et al., 2016; Yeh et al., 2019; Bhatt et al., 2020; Ismail et al., 2020). Our findings are primarily validated on image tasks where spatial structure is strongly expressed. Biases may manifest differently in text or graph domains, and exploring these settings remains an important direction for scaling interpretability evaluation. Our theoretical analysis assumes the idealized causal model in Fig. 1. In practice, retraining dynamics may introduce additional confounds beyond what DPI captures.

7. Conclusion

This paper highlights a limitation of using the ROAR metric and its variant as stand-alone evaluations of feature importance estimates. We show theoretically that attributions carrying *less* information about the model can achieve better ROAR scores, and empirically observe that simple agnostic post-processing often improves ROAR/ROAD scores through a blurriness-related bias. As the mechanistic interpretability community increasingly relies on attribution methods to validate hypotheses about model internals and to support safety-relevant applications, careful benchmark interpretation becomes important. We hope our results encourage the development of evaluation protocols.

References

Adebayo, J., Gilmer, J., Goodfellow, I., and Kim, B. Local explanation methods for deep neural networks lack sensitivity to parameter values. In *ICLR Workshop*, 2018a.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. *NeurIPS*, 2018b.

Adebayo, J., Muelly, M., Liccardi, I., and Kim, B. Debugging tests for model explanations. In *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/075b051ec3d22dac7b33f788da631fd4-Abstract.html.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. In *ICLR*, 2018. URL https://openreview.net/forum?id=Sy21R9JAW.

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

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. How to explain individual classification decisions. *JMLR*, 11(61):1803–1831, 2010. URL http://jmlr.org/papers/v11/baehrens10a.html.

Bhatt, U., Weller, A., and Moura, J. M. Evaluating and aggregating feature-based model explanations. In *IJCAI*, 2020.

Bora, R. P., Terhörst, P., Veldhuis, R., Ramachandra, R., and Raja, K. Fries: Framework for inconsistency estimation of saliency metrics. *Pattern Recognition*, pp. 112136, 2025.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Canha, D., Kubler, S., Främling, K., and Fagherazzi, G. A functionally-grounded benchmark framework for xai methods: Insights and foundations from a systematic literature review. *ACM Computing Surveys*, 2025.

Chefer, H., Gur, S., and Wolf, L. Transformer interpretability beyond attention visualization. In *CVPR*, pp. 782–791, 2021.

Contributors, M. Openmmlab's image classification toolbox and benchmark. https://github.com/open-mmlab/mmclassification, 2020.

Dabkowski, P. and Gal, Y. Real time image saliency for black box classifiers. In *NeurIPS*, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., and Wallace, B. C. ERASER: A benchmark to evaluate rationalized NLP models. In *ACL*, 2020.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Duan, J., Li, H., Zhang, H., Jiang, H., Xue, M., Sun, L., Song, M., and Song, J. On the evaluation consistency of attribution-based explanations. In *European Conference on Computer Vision*, pp. 206–224. Springer, 2024.

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, pp. 3429–3437, 2017.

Funke, T., Khosla, M., Rathee, M., and Anand, A. Zorro: Valid, sparse, and stable explanations in graph neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Hartley, T., Sidorov, K., Willis, C., and Marshall, D. Swag: Superpixels weighted by average gradients for explanations of cnns. In *WACV*, pp. 423–432, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Hellman, M. and Raviv, J. Probability of error, equivocation, and the chernoff bound. *IEEE Transactions on Information Theory*, 16(4):368–372, 1970.

Hong, J.-H., Kim, H.-J., Jeon, K.-S., and Lee, S.-W. Comprehensive information bottleneck for unveiling universal attribution to interpret vision transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25166–25175, 2025.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. A benchmark for interpretability methods in deep neural networks. In *NeurIPS*, pp. 9737–9748, 2019.

Ismail, A. A., Gunady, M., Bravo, H. C., and Feizi, S. Benchmarking deep learning interpretability in time series predictions. In *NeurIPS*, 2020.

Ismail, A. A., Corrada Bravo, H., and Feizi, S. Improving deep learning interpretability by saliency guided training. *NeurIPS*, 34, 2021.

Khakzar, A., Baselizadeh, S., Khanduja, S., Rupprecht, C., Kim, S. T., and Navab, N. Neural response interpretation through the lens of critical pathways. In *CVPR*, pp. 13528–13538, 2021.

Kim, B., Seo, J., Jeon, S., Koo, J., Choe, J., and Jeon, T. Why are saliency maps noisy? cause of and solution to noisy saliency maps. In *ICCV Workshop*, pp. 4149–4157. IEEE, 2019a.

Kim, B., Seo, J., and Jeon, T. Bridging adversarial robustness and gradient interpretability. *ICLR Workshop*, 2019b.

Krizhevsky, A. et al. Learning multiple layers of features from tiny images. 2009.

Liang, J., Bai, B., Cao, Y., Bai, K., and Wang, F. Adversarial infidelity learning for model interpretation. In *KDD*, pp. 286–296, 2020. ISBN 9781450379984. URL https://doi.org/10.1145/3394486.3403071.

Liu, X., Xiong, X., Yan, M., Xue, R., Pan, S., Ye, X., and Fan, D. Rethinking efficiency and redundancy in training large-scale graphs. *arXiv preprint arXiv:2209.00800*, 2022.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *NeurIPS*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Meng, C., Trinh, L., Xu, N., Enouen, J., and Liu, Y. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1): 1–28, 2022.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *arXiv preprint arXiv:2201.08164*, 2022.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

O'Shaughnessy, M., Canal, G., Connor, M., Davenport, M., and Rozell, C. Generative causal explanations of black-box classifiers. *NeurIPS*, 2020.

Park, Y.-H., Seo, J., Park, B., Lee, S., and Jo, J. Geometric remove-and-retrain (goar): Coordinate-invariant explainable ai assessment. *NeurIPS Workshop*, 2023.

Radovic, M., Ghalwash, M., Filipovic, N., and Obradovic, Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC bioinformatics*, 18(1):1–14, 2017.

Rong, Y., Leemann, T., Borisov, V., Kasneci, G., and Kasneci, E. A consistent and efficient evaluation strategy for attribution methods. In *ICML*, pp. 18770–18795. PMLR, 2022.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. Evaluating the visualization of what a deep neural network has learned. *TNNLS*, 2016.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K.-R. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 2021.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.-G., Mahlein, A.-K., and Kersting, K. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2020.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *CVPR*, pp. 618–626, 2017.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *ICML*, 2017.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *ICML Workshop*, 2017.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *ICML*, pp. 3319–3328, 2017.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Yang, Y., Qiu, J., Song, M., Tao, D., and Wang, X. Learning propagation rules for attribution map generation. In *ECCV*, 2020.

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

Zhang, W., Huang, Z., Zhu, Y., Ye, G., Cui, X., and Zhang, F. On sample based explanation methods for NLP: Faithfulness, efficiency and semantic evaluation. In *IJCNLP*, 2021.

Zheng, X., Shirani, F., Chen, Z., Lin, C., Cheng, W., Guo, W., and Luo, D. F-fidelity: A robust framework for faithfulness evaluation of explainable AI. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=X0r4BN50Dv.

Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

A. Proof of Theorem 3.1

Proof of Theorem 3.1. Fix any x. Because $\widetilde{A}$ is generated from (A, U) with U independent of all other variables, we have the conditional Markov chain

$$Z \;\to\; A \;\to\; \widetilde{A} \quad \text{given } X = x, \tag{12}$$

i.e., $\widetilde{A} \perp Z \mid (A, X = x)$. By the (conditional) data processing inequality,

$$I(Z; \widetilde{A} \mid X = x) \;\le\; I(Z; A \mid X = x). \tag{13}$$

Taking expectation over X yields $I(Z; \widetilde{A} \mid X) \le I(Z; A \mid X)$.

For the second claim, given $X = x$ the random variable $\psi(X, \widetilde{A})$ is a function of $\widetilde{A}$ (and independent randomness internal to ψ). Thus we have another Markov chain

$$Z \;\to\; \widetilde{A} \;\to\; \psi(X, \widetilde{A}) \quad \text{given } X = x, \tag{14}$$

and DPI implies $I(Z; \psi(X, \widetilde{A}) \mid X = x) \le I(Z; \widetilde{A} \mid X = x)$. Averaging over X proves the statement, and the cases $\widetilde{M}_t = h_t(\widetilde{A})$ and $\widetilde{X}'_t = \phi(X, \widetilde{M}_t)$ follow by choosing ψ appropriately. □

B. Details of Experiment Setting

All models were trained for 50 epochs with batch size $b = 128$ except for $b = 32$ on CUB-200 (Welinder et al., 2010) on a single GTX1080Ti GPU. Each model was trained with SGD optimizer with momentum $\beta = 0.9$ and weight decay $\lambda = 1 \times 10^{-4}$ except for $\lambda = 5 \times 10^{-4}$ on CIFAR-10. To avoid convergence on bad local minima, cosine annealing scheduling (Loshchilov & Hutter, 2017) was used, starting from the initial learning rate $\alpha = 0.1$ for CIFAR-10 and SHVN. For CUB-200, $\alpha = 0.01$ was used. The learning rate was annealed to zero over 50 epochs. Especially in the case of CUB-200, we use ImageNet (Deng et al., 2009) pre-trained model for fast convergence. In our ROAD (Rong et al., 2022) experiments, we utilized MMClassification framework (Contributors, 2020) to facilitate the implementation. Our training strategies followed MMClassification's ones for ResNet18 on CIFAR-10 and ResNet50 on CUB-200. In the case of SVHN, we adopted the same strategy as CIFAR-10.

Base estimators were implemented by PyTorch AutoGrad module. Similar to (Hooker et al., 2019), in Integrated Gradients, interval k is set to 25, and the reference image is set to zeros. Also, for the ensemble methodologies, the sample size is set to $n = 15$. Among all estimators, only Sobel edge detector is implemented by SciPy library (Virtanen et al., 2020). Finally, we post-process attribution map using Maximum filter with kernel size $= 3$ and Gaussian filter function with standard deviation for Gaussian kernel $\sigma = 1$ in SciPy library.

C. Additional Tables and Plots

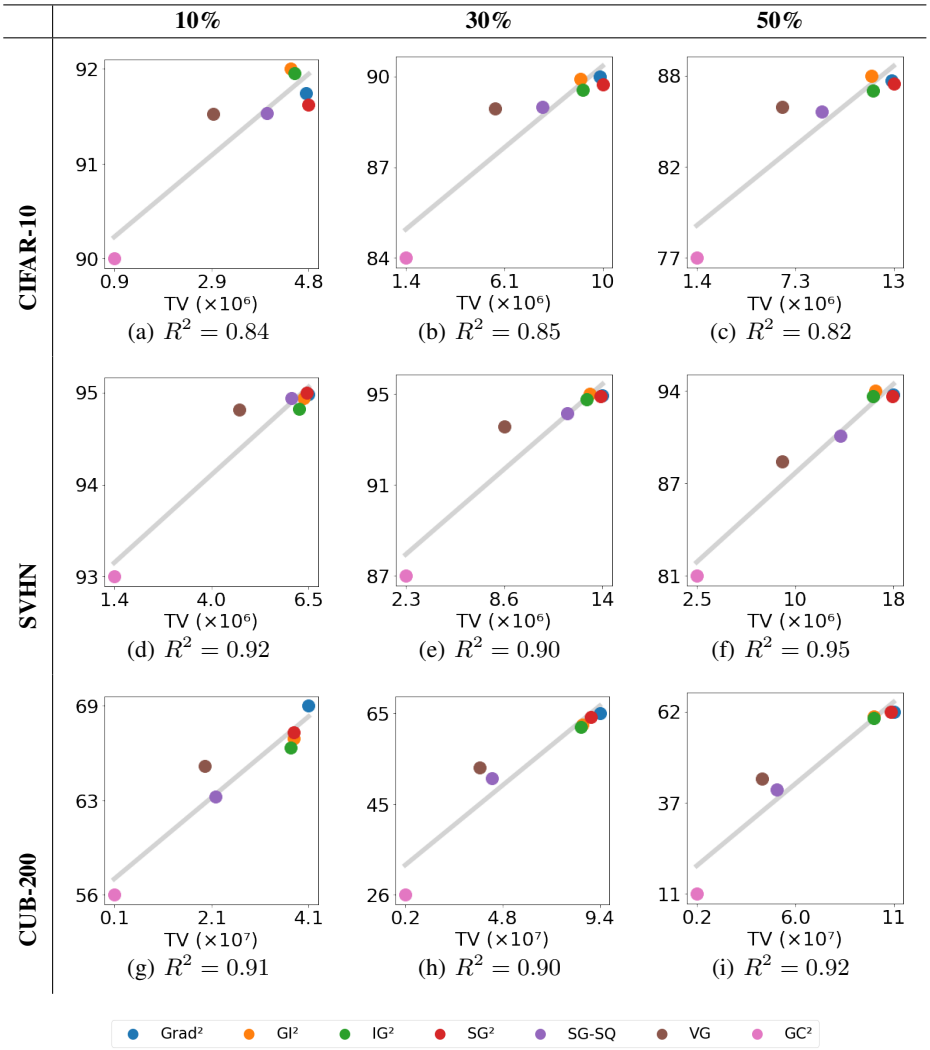


Figure 7. The relationship between model accuracy and the total variation of attribution masks in terms of attribution method. The fitted line and coefficient of determination from simple linear regression are also included. The y-axis represents the final test accuracy (%).

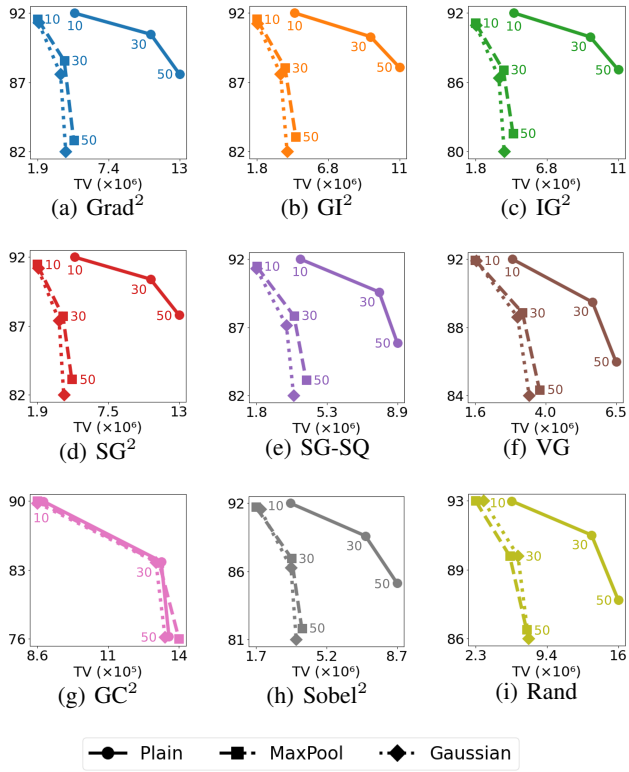


Figure 8. Effects of Gaussian filtering and max-pooling on total variation of attribution masks and model accuracy on CIFAR-10.

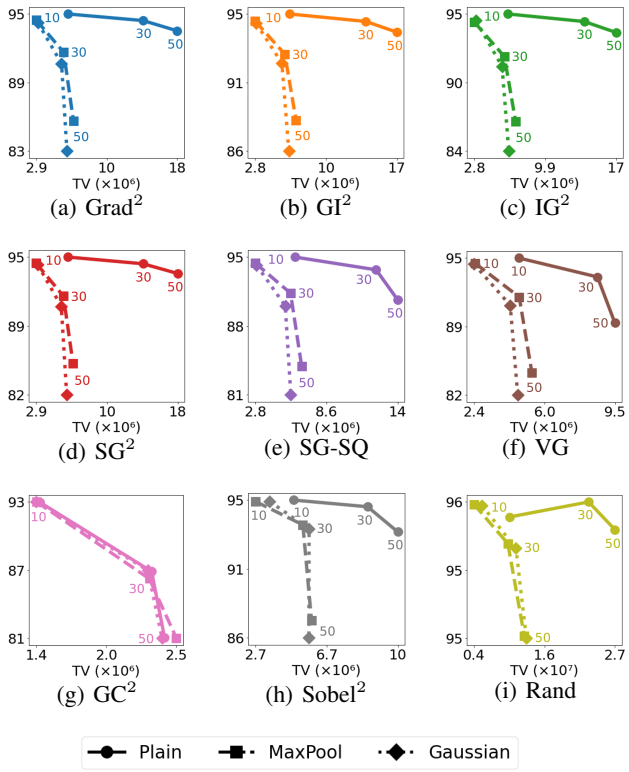


Figure 9. Effects of Gaussian filtering and max-pooling on total variation of attribution masks and model accuracy on SVHN.

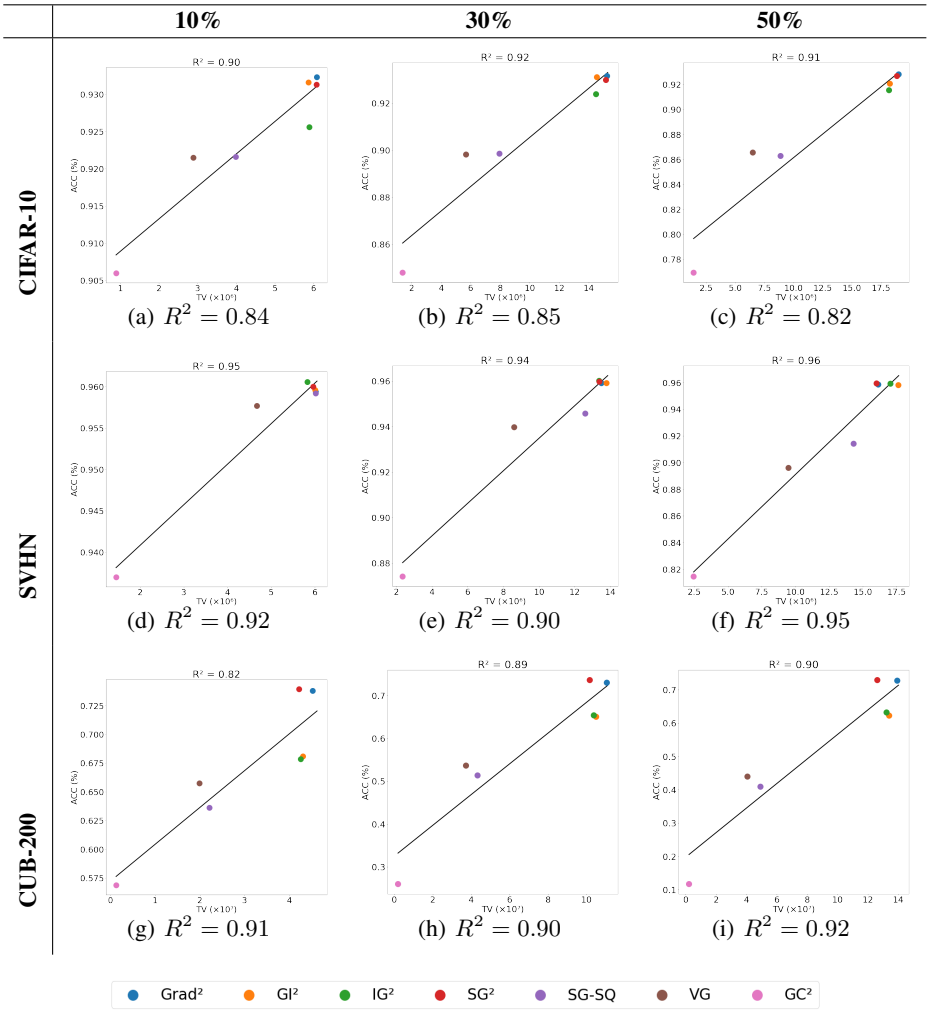


Figure 10. Relationship between model accuracy and total variation of attribution masks (*TV*) in the aspect of non-squaring attribution method. The fitted line and coefficient of determination from simple linear regression are presented together. The y-axis represents the final test accuracy (%) in the ROAR evaluation protocol.