# A BACKDOOR-BASED EXPLAINABLE AI BENCHMARK FOR IMPROVED FIDELITY IN EVALUATING ATTRIBUTION METHODS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Attribution methods compute importance scores for input features to explain the output predictions of deep models. However, accurate assessment of the performance of attribution methods is challenged by the lack of ground truth along with other confounding factors such as attribution post-processing and explanation objectives. In this paper, we first identify a set of fidelity criteria that must be satisfied for reliable evaluation of attribution methods. Then, we introduce a Trojaned model based benchmarking framework that adheres to the desired fidelity criteria. We theoretically establish the superiority of our approach over existing benchmarks for well-founded attribution evaluation. With extensive analysis, we also identify a setup for a consistent and fair benchmarking of attribution methods across different underlying methodologies. This setup is ultimately employed for a comprehensive comparison of existing methods using our benchmark. Finally, our analysis also provides guidance for defending against backdoor attacks using existing attribution methods.

## 1 INTRODUCTION

Deep learning models have made remarkable strides across diverse domains (He et al., 2016; Ren et al., 2015; Goodfellow et al., 2014) owing to their capacity to learn intricate representations through numerous parameters. However, the extensive parameter scale that contributes to their success also renders these models less interpretable for their decisions, which limits their applicability in high-stake tasks (Gade et al., 2019; Tjoa & Guan, 2020).

In an effort to provide explanations for model predictions, attribution methods (Simonyan et al., 2014; Zeiler & Fergus, 2014; Sundararajan et al., 2017) ascribe importance scores, referred to as attributions, to the input features. However, due to the absence of ground truth for attributions, the assessment of attribution methods relies on alternate techniques, e.g., feature perturbations (Samek et al., 2016; Hooker et al., 2019; Rong et al., 2022). This compromises the reliability of the evaluation of explanations. The problem exacerbates in the presence of other confounding factors, such as post-processing of the attribution scores (Yang et al., 2023b). Challenges in assessing attribution methods has often led to the development of additional explanatory techniques designed to clarify the explanations themselves (Rudin, 2019; Han et al., 2022), which is counter-productive.

In this paper, we first define a set of clear fidelity criteria that Explainable Artificial Intelligence (XAI) benchmarks should adhere to. We argue that these benchmarks should ensure fidelity to both the explained model and the input, thereby satisfying both functional mapping invariance and input distribution invariance. Moreover, we stress that attribution benchmarks should offer two essential components; namely, verifiable ground truth for attributions and sensitive metrics for evaluating them. Our foundational criteria not only set a standard for benchmarking but also facilitate a clear assessment of the existing XAI benchmarks. We then propose a new XAI benchmark for attribution methods based on neural Trojan (Gu et al., 2019; Chen et al., 2019). Neural Trojans or backdoor attacks render a model sensitive to a specific trigger pattern to control their predictions. We propose to leverage this explicit control to establish ground truth attributions for the model. Through a theoretical analysis, we establish a superior fidelity of this approach to the model and the input.

Currently, assessment of attribution methods is also confounded by the choices for post-processing of the computed attributions and the selection of the output signal in attribution computation, which compromises its transparency (Smilkov et al., 2017; Wang & Wang, 2022; Yang et al., 2023b). Through our well-founded benchmarking approach, we reveal distinct properties of different attribution methods using different choices. Our analysis leads to a consistent benchmarking setup across different types of the attribution methods. Using this consistent setup, we eventually assess a range of attribution methods using various trigger patterns for different Trojaned models under diverse attacks. A by-product of this endeavor is that we are also able to reveal interesting guidance for defending against

backdoor attacks using feature attribution methods. In summary, our paper contributes along the following three key aspects.

1. It introduces fidelity criteria designed to create a trustworthy XAI benchmark, facilitating the assessment of existing benchmarks against these criteria.
2. Exploiting controllable attributions in Trojaned models, it proposes a backdoor-based XAI benchmark for attribution methods. We substantiate the superior fidelity of this benchmark through theoretical analysis.
3. It identifies a consistent setup for a transparent assessment of different kinds of attribution methods, and performs an extensive evaluation of existing methods with the proposed benchmark using this setup. In the process, it also offers interesting guidance for defending against backdoor attacks.

## 2 Related Work

**Attribution Methods.** To explain model predictions, attribution methods (Simonyan et al., 2014; Zeiler & Fergus, 2014) assign importance scores to input features. Deconvnet (Zeiler & Fergus, 2014) and Guided Backpropagation (Springenberg et al., 2015) utilize deconvolution technique to compute feature importance. CAM (Zhou et al., 2016) and GradCAM (Selvaraju et al., 2017) calculate class activation maps during back-propagation for locating class-specific input features. Compared to CAM-based methods, InputGrad (Simonyan et al., 2014) calculates gradients with respect to the input for model explanation. SmoothGrad (Smilkov et al., 2017) aggregates input gradients across input samples with Gaussian noise, leading to a notable enhancement in localization ability. FullGrad (Srinivas & Fleuret, 2019) further integrates gradients from model biases, satisfying additional axioms. To guarantee completeness, integrated gradients (IG) (Sundararajan et al., 2017) combines model gradients over a number of inputs computed w.r.t. a reference, however, its performance depends on the choice of reference. IG-SG (Smilkov et al., 2017), IG-SQ (Hooker et al., 2019) and IG-Uniform (Sturmfels et al., 2020) are proposed to redefine the reference input as perturbed samples. EG (Erion et al., 2021) and LPI (Yang et al., 2023a) employ training samples as references to maintain the distribution invariance. Pan et al. (2021) calculated class-specific adversarial samples as the reference input. Yang et al. (2023b) recalibrated attributions by with valid references. Instead of explaining the model's output, Wang & Wang (2022) turned to explain a contrastive output, leading to class-contrastive attributions.

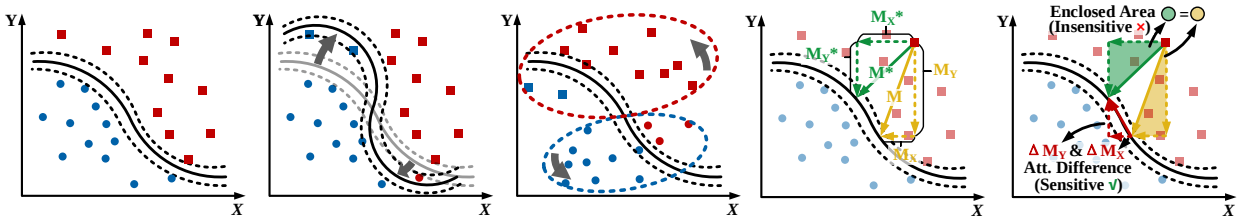
**XAI Benchmarks.** Numerous explaining methods have given rise to a multitude of XAI benchmarks. LeRF and MoRF (Samek et al., 2016) are extended from pixel flipping (Bach et al., 2015), which perturbs input samples to test output changes. However, Hooker et al. (2019) revealed that perturbed images cause input distribution shift, making the benchmarking unreliable. Therefore, ROAR (Hooker et al., 2019) and DiffROAR (Shah et al., 2021) were proposed to retrain models on the perturbed images, ensuring models are learned within the distribution. ROAD (Rong et al., 2022) and DiffID (Yang et al., 2023a) offer alternative methods to mitigate input distribution shift without necessitating costly model retraining. On the other hand, sensitivity-n (Ancona et al., 2018), $\text{SENS}_{\text{MAX}}$ and INFD (Yeh et al., 2019) focus on testing the fidelity of attributions by perturbing the input. Adebayo et al. (2018) randomized the model parameters and training labels to test the sanity of attributions. Other efforts have also been made to provide ground truth for the estimated attributions. DiFull, DiPart (Rao et al., 2022) and Pointing Game (Zhang et al., 2018) employ training annotations for attribution evaluation. Khakzar et al. (2022) used model-optimized features for attribution ground truth. Additionally, Arras et al. (2022) employed the controlled VQA framework to generate synthetic images for benchmarking.

**Model Trojaning.** Trojaning alters a model's predictions by making it sensitive to certain trigger patterns. Gu et al. (2019) introduced BadNet which mislabels and stamps trigger to poison the training samples. Blended attack (Chen et al., 2019), ISSBA (Li et al., 2021) and WaNet (Nguyen & Tran, 2021) are proposed to generate more stealthy triggers, achieving invisible poisoning. Adap-Blend (Qi et al., 2022) further enhances the stealthiness of attack in the latent space. Due to the clear attribution of backdoor triggers, we benchmark the reliability of attribution methods on such models. Given the wide applications of attribution methods in backdoor defense (Huang et al., 2019; Chou et al., 2020), we also provide guidance for defense against backdoor attacks using attributions as part of this study.

## 3 Benchmark fidelity

In this section, we first put forth a set of criteria that a reliable explanation benchmark should adhere to. We then compare the existing benchmarks on the proposed criteria.

Let us consider an input sample $x \in \mathbb{R}^n$ with its label $y \in \mathbb{R}^c$ from a dataset $\mathcal{D}$. A classifier denoted as $f : \mathbb{R}^n \to \mathbb{R}^c$ is parameterized by $\theta$. To explain the model's prediction $f(x)$, an attribution explaining tool $\phi : \mathbb{R}^c \to \mathbb{R}^n$ is used




**(a) Model to be Explained** **(b) Funct. Mapping Shift** **(c) Input Dist. Shift** **(d) Verifiable Att. Fidelity** **(e) Metric Sensitivity**

Figure 1: Illustration of benchmark fidelity criteria. For explaining **(a)** the model, a faithful XAI benchmark should avoid **(b)** Functional Mapping Shift, and **(c)** Input Distribution Shift, while ensuring **(d)** Attribution Verifiability, and **(e)** Metric Sensitivity. See text in § 3.1 for explanation.

to generate the attribution map $M$, specifically $M = \phi(f(x))$. An XAI benchmark intends to faithfully evaluate the reliability of the explanation $M$. For a quick reference, we also provide a summary of notations in Appendix A.1.

### 3.1 Fidelity Criteria

We introduce a set of fidelity criteria that attribution benchmarking should satisfy for reliable results. These criteria are based on fundamental observations, hence they are intuitive in nature.

**Functional Mapping Invariance.** Given a model to be explained, e.g., Figure 1(a), functional mapping invariance requires that attribution benchmarking does not cause a functional mapping shift, see Figure 1(b). Suppose a perturbation $\zeta$ is imposed on model $f$, resulting in a functional mapping shift, i.e., $\zeta(f(x)) \neq f(x)$. This shift introduces divergent explanations, thus undermining the reliability of benchmarking results. The assurance of functional mapping invariance (i.e., $f(x) = \zeta(f(x))$) serves as a crucial criterion for upholding the fidelity of the explained model.

**Input Distribution Invariance.** Similar to functional mapping invariance, input distribution invariance mandates the constancy of the input distribution $\mathcal{P}_\mathcal{D}$. Assuming $\zeta$ represents an input perturbation that leads the input distribution shift (i.e., $\mathcal{P}_\mathcal{D} \neq \mathcal{P}_{\zeta(\mathcal{D})}$), this shift causes explanation variance of input samples (Hooker et al., 2019), see Figure 1(c). Thus, maintaining input distribution invariance is crucial to ensuring the benchmark fidelity of the explained samples.

**Attribution Verifiability.** Attribution verifiability requires the estimated attributions to be verifiable. Given attributions $M = \phi(f(x))$, a faithful benchmark is required to provide corresponding ground truth attributions $M^*$. The ground truth attributions $M^*$ are crucial for the verification of the estimated attributions $M$. This criterion can be observed in Figure 1(d), where the normal vector and its horizontal and vertical components extending from an input sample to the decision boundary represent the ground truth attributions (i.e., $M^* = M_X^* + M_Y^*$), allowing estimated attributions (i.e., $M = M_X + M_Y$) to be verified.

**Metric Sensitivity.** Metric sensitivity requires that the benchmark's metric exhibits sensitivity to the attribution change. Figure 1(e) shows examples of both sensitive and insensitive metrics. Given the estimated attributions $M$ with their corresponding ground truth $M^*$, an insensitive metric, such as the enclosed area defined by the normal vector and its components, may yield identical evaluation results for both $M^*$ and $M$, failing to accurately capture changes in attributions. In contrast, the attribution difference from estimated attributions $M^*$ to the ground truth $M$ (i.e., $\Delta M_X$ and $\Delta M_Y$) serves as a sensitive metric, demonstrating faithful sensitivity to changes in attributions.

### 3.2 Fidelity Comparison

In Table 1, we present a comparative analysis of different benchmarks with regard to their fidelity criteria guarantee. Appendix A.3 provides a detailed discussion about the fulfillment of existing benchmarks. The attribution benchmarks that rely on input perturbations, such as MoRF, LeRF (Samek et al., 2016), and Ins.&Del. Games (Petsiuk et al., 2018), fail to ensure input distribution invariance. Despite the efforts made by ROAD (Rong et al., 2022) and DiffID (Yang et al., 2023a) to mitigate the impact of shifts in input distribution, they are unable to guarantee input distribution invariance. Other attempts, exemplified by ROAR (Hooker et al., 2019) and DiffROAR (Shah et al., 2021), involve retraining models on perturbed input samples to maintain input distribution invariance. However, these approaches introduce shifts in the functional mapping, thereby compromising functional mapping invariance. Existing sanity and sensitivity checks (Adebayo et al., 2018; Ancona et al., 2018; Yeh et al., 2019) also test the fidelity of attribution methods under model and input variations. Overall, perturbation-based attribution benchmarks and sanity checks struggle to simultaneously ensure invariance to both functional mapping and input distribution.

In Table 1, it is evident that a full guarantee of attribution verifiability presents a significant challenge. In an effort to establish attribution ground truth for verifiability, Pointing Game (Zhang et al., 2018), DiFull & DiPart (Rao et al., 2022) employ training annotations as attribution ground truth. Khakzar et al. (2022) calculated Features with Null information by model optimization to test the attribution fidelity. CLEVR-XAI (Arras et al., 2022) generates synthetic images based on CLEVR dataset to provide controlled evaluations. However, these benchmarks fail to provide full verifiable attributions. The challenge of offering precise attribution ground truth also adds complexity to upholding metric sensitivity within such benchmarks. The perturbation-based benchmarks, which place less emphasis on verifiability, commonly utilize element-wise metrics that can adequately guarantee metric sensitivity. Overall, our benchmark stands out as more desirable than the existing techniques on the fidelity criteria.

Table 1: Fidelity criterion fulfillment of XAI benchmarks.

| Benchmark Fidelity Criteria | XAI Benchmarks | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | MoRF, LeRF; Ins.& Del. Game | ROAR; DiffROAR | Sensitivity-n; $SENS_{MAX}$, ROAD; DiffID | Model & label randomization, INFD | Pointing Game; DiFull & DiPart | Null Feature; CLEVR-XAI | **Our Benchmark** | |
| Functional Mapping Invariance | ● | ○ | ● | ● | ○ | ◐ | ● | ● |
| Input Distribution Invariance | ○ | ● | ◐ | ◐ | ● | ● | ● | ◐ |
| Attribution Verifiability | ○ | ○ | ○ | ◐ | ◐ | ◐ | ◐ | ● |
| Metric Sensitivity | ● | ● | ● | ● | ◐ | ◐ | ◐ | ● |
| Desirability Score | 2 | 2 | 2.5 | 3 | 2 | 2.5 | 3.0 | **3.5** |

● = Strong (1); ◐ = Weak (0.5); ○ = No fulfillment (0).

## 4 Benchmarking Explanations with Trojaned Model

We commence by illustrating the pipeline of our proposed benchmark and then introduce a collection of metrics meticulously crafted to evaluate the diverse capabilities of attribution methods. Finally, we theoretically discuss the assurance of established benchmark fidelity criteria within our proposed benchmark.

### 4.1 Benchmark Framework

In Figure 2, we present an illustration of our benchmark pipeline. Given a clean training set $\mathcal{D}$, we generate a poisoned training set $\tilde{\mathcal{D}}$ by incorporating a trigger $v$ into certain portions of the clean samples and modifying the true label $y$ with the poisoned target label $\tilde{y}$. Subsequently, we employ the poisoned training set to transform a benign model $f$ into a Trojaned model $\tilde{f}$, as depicted in Step 1 of Figure 2. The Trojaned model is employed to generate predictions for both clean and poisoned samples, as shown in Step 2 in Figure 2. Then, the output prediction $\tilde{f}(\tilde{x})$ of a poisoned sample is explained using an attribution method $\phi$, resulting in an attribution map $M$ — Step 3 of Figure 2. Subsequently, a mask $S^{(k)} \in \{0,1\}^n$ guided by the attribution map is computed, identifying the top $k\%$ most important input features (e.g., input pixels). The mask is then utilized to recover a portion of the poisoned sample $\tilde{x}$, see Step 4 of Figure 2. Specifically, a clean patch is extracted from the clean sample, guided by the mask. The recovery of the poisoned sample is achieved by replacing the corresponding patch using the clean patch, yielding a recovered sample $\hat{x}$, where $\hat{x} = \tilde{x} \odot (\mathbf{1} - S) + x \odot S$. Finally, we can benchmark attribution methods by assessing changes in the predictions among clean, poisoned and recovered samples — Step 5 of Figure 2.

In our benchmark, we comprehensively test both visible and invisible trigger patterns for Trojaned models through Blend (Chen et al., 2019) and ISSBA (Li et al., 2021). For a complete control, we modify the Blend attack from encompassing the entire image to focusing solely on the trigger region to embed a watermark trigger. Given a trigger mask denoted as $S^* \in \{0,1\}^n$, we create the watermark trigger from a trigger $v$ using the formula $(x \odot (1-\alpha) + v \odot \alpha) \times S^*$, where $\alpha \in [0,1]$ represents the visibility of the trigger. Additionally, we utilize the ISSBA attack method to generate input-specific invisible triggers. Trojaned models with different trigger patterns and corresponding training details are presented in Appendix A.5.

### 4.2 Evaluation Metrics

In this section, we formulate a set of metrics for our benchmark. The metric definitions are specifically tailored to our Trojan-based evaluation.

**Attack Success Rate.** Attack Success Rate (ASR) is a prevalent metric in backdoor attacks (Turner et al., 2019; Guo et al., 2022). It records the success of target prediction label $\tilde{y}$ for the input samples with true label $y$ when the trigger is embedded in the input samples. The ASR is formally defined as

$$\text{ASR} = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \mathbf{I}[\hat{y} = \tilde{y} | y \neq \tilde{y}], \quad \hat{y} = \arg\max_i \tilde{f}_i(\tilde{x}). \tag{1}$$

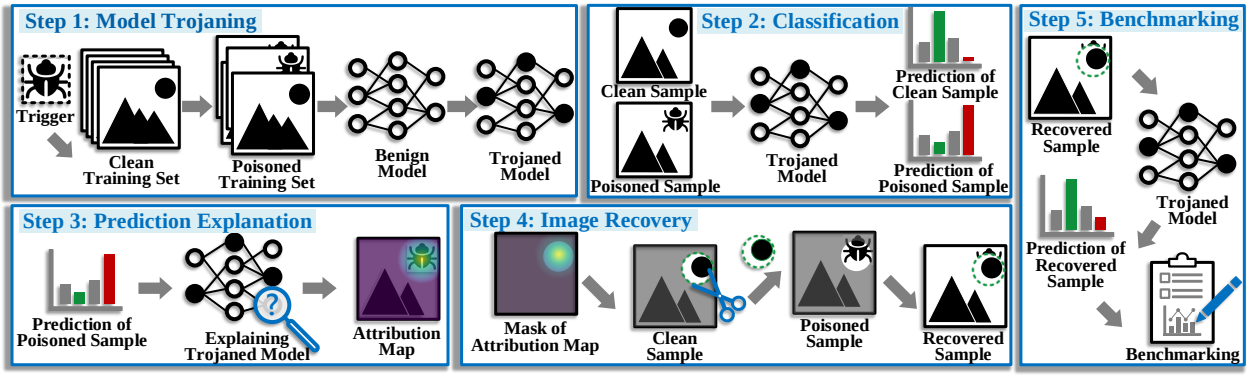



Figure 2: The proposed XAI benchmark pipeline. **Step 1** embeds a backdoor into a benign model by retraining it on a poisoned training set. **Step 2** uses the Trojaned model to generate predictions. **Step 3** applies attribution methods to explain the predictions. **Step 4** computes an attribution map-based mask to guide the recovery of the poisoned sample from the clean sample. **Step 5** assesses attribution methods by measuring the prediction change.

Here, the ASR is computed as an expectation over samples drawn from a dataset $\mathcal{D}$ through a boolean function $\mathbf{I}[\cdot]$. In our benchmark, we leverage ASR to evaluate the ability of attribution methods to successfully identify the trigger with the recovered sample $\hat{x}$.

**Logit and Probability Fractional Change.** Model's output fractional change is a widely used metric for evaluating attributions (Samek et al., 2016). To evaluate the detailed prediction change, we measure the fractional change in both output logits and probabilities. Let $p(x) = \mathrm{softmax}(f(x))$ denote the probability output through the softmax function. We denote the logit and probability fractional change of target class $\tilde{y}$ as $\Delta f_{\tilde{y}}(\hat{x})$ and $\Delta p_{\tilde{y}}(\hat{x})$ respectively. In our context, these measures quantify the extent to which the attribution method can help reduce the confidence of the target class $\tilde{y}$ when recovery is made by replacing the features identified as trigger features by the attribution method. We similarly denote the logit and probability fractional changes of the output for the true label $y$ as $\Delta f_y(\hat{x})$ and $\Delta p_y(\hat{x})$. This serves to quantify the extent to which the attribution method can help restore the confidence in the target class $y$. Concretely, the logit and probability fractional changes through the Trojaned model $\tilde{f}$ are defined as

$$\Delta \tilde{f}_{\tilde{y}}(\hat{x}) = \frac{\tilde{f}_{\tilde{y}}(\hat{x}) - \tilde{f}_{\tilde{y}}(x)}{\tilde{f}_{\tilde{y}}(\tilde{x})},\ \Delta \tilde{f}_{y}(\hat{x}) = \frac{\tilde{f}_{y}(\hat{x}) - \tilde{f}_{y}(\tilde{x})}{\tilde{f}_{y}(x)},\ \Delta p_{\tilde{y}}(\hat{x}) = \frac{p_{\tilde{y}}(\hat{x}) - p_{\tilde{y}}(x)}{p_{\tilde{y}}(\tilde{x})},\ \Delta p_{y}(\hat{x}) = \frac{p_{y}(\hat{x}) - p_{y}(\tilde{x})}{p_{y}(x)}. \quad (2)$$

**Trigger Recall.** In addition to evaluating predictive performance, we assess the localization capability of attribution methods. Trigger Recall (TR) is introduced to evaluate the ability of an attribution method to locate the embedded triggers. Given the attribution and trigger masks $S^{(k)}, S^* \in \{0,1\}^n$, TR is defined as the ratio between the area intersection of $S^{(k)}$ and $S^*$ to the area of $S^*$, formally; $\mathrm{TR} = S^{(k)} \cap S^* / S^*$. Here, $k \in [0,1]$ denotes the percentage of the most important features identified by the attribution method in the input. It is worth noting that the recall rate of the detected trigger equates to *precision* when the recovery rate $k$ equals the proportion of the trigger.

### 4.3 Benchmark Fidelity Examination

In this part, we closely examine the crucial aspect of our benchmark regarding satisfying the fidelity criteria set out in § 3.1. In what follows, we delve deeper into theoretical details only when a criterion fulfillment is not obvious.

By consistently employing the same Trojaned model $\tilde{f}$ throughout the benchmarking process, our proposed framework inherently guarantees *Functional Mapping Invariance*. Additionally, the metrics we introduce are designed for element-level benchmarking. This automatically ensures the preservation of *Metric Sensitivity*. Furthermore, the use of Trojaned model and trigger enables access to the ground truth in our benchmark. However, ensuring *Metric Sensitivity* also implies that the primary focus of our benchmarking is placed on the recovered samples. These samples inadvertently hold information about the attribution mask used for image recovery — see Step 4 in Figure 2. This can lead to unreliable benchmarking due to potential class information leakage through the mask (Rong et al., 2022). This feature leakage can actually compromise the fidelity of the provided attribution ground truth. Therefore, we must examine assurance of *Attribution Verifiability* during the benchmarking process more closely.

We commence by examining the entropy of a singular variable $x$, quantifying the amount of information by observing $x$ using the formula $H(x) = -\sum_{x_i \in x} P(x_i) \log P(x_i)$. In the context of the leaked feature, we employ mutual infor-

mation established on the entropy of two variables, measuring their mutual dependence. In a Bayesian classifier, the classification performance is correlated with the mutual information $I(x; c)$ between the input $x$ and a class $c$ (Hoque et al., 2014). Without loss of generality, we can assume that attribution methods benchmarked using a masked sample $\hat{x}$ are intended to quantify the mutual information $I(\hat{x}; c)$. Assuming the attribution mask $S$ operates as a patch that directly removes features from the input sample, this process inevitably allows the leakage of class-related information into the masked sample $\hat{x}$, leading to a leakage of $I(S; c)$. The leakage leads to the unfaithfulness of evaluation results by introducing class information from $S$ in $\hat{x}$, i.e., $I(\hat{x}; c) \neq I(c; \hat{x}|S)$. To ensure the benchmarking fidelity, the leaked features from the $S$ can be addressed by minimizing the mutual information between the mask represented by $S$ and the class $c$, i.e., $I(S; c) \approx 0$. This relation is formalized by the below proposition.

**Proposition 1.** *By minimizing the mutual information $I(S; c)$ between the mask $S$ and a class $c$, the leaked information from the attribution mask $S$ to the masked sample $\hat{x}$ can be alleviated, resulting in enhanced fidelity of the evaluation results, expressed as $I(c; \hat{x}) \approx I(c; \hat{x}|S)$.*

The proof is provided in Appendix A.2. In our benchmark, we evaluate the attribution methods via Trojaned models. Assume a benign model $f$ is trained to classify an input sample $x$ to a label $y$ within the boundary $x \pm \delta$, where this boundary is bounded by a constraint $\epsilon$, i.e., $||\delta||_p \leq \epsilon$. In model Trojaning, a trigger $v$ within the same constrain $\epsilon$ is embedded into a clean sample to modify the prediction from label $y$ to target label $\tilde{y}$. In contrast to the prevalent existing benchmarks that assess attributions through feature removal, our benchmark recovers the features of the poisoned sample with those of the clean sample through the mask, i.e., $S \odot x \in x$. The recovered features which are part of the clean sample $x$ do not contain additional information related to the target class $\tilde{y}$. As a result, the mutual information $I(S \odot x, c = \tilde{y})$ is essentially zero. Thus, the evaluation process within our benchmark can provide assurance for both *Attribution Verifiability* and *Metric Sensitivity*. This fact highlights the inherent superiority of our benchmark over the perturbation-based benchmarks that inadvertently leak features into the evaluation outcome.

In our benchmarking process, we utilize the recovered sample $\hat{x}$ to convert a poisoned sample back to the clean state $x$. This transformation is constrained by the limit of $\epsilon$ (i.e., $||x - \hat{x}||_p \leq \epsilon$), thus largely preserving the benchmarked samples within the original distribution. Whereas we acknowledge that our benchmark cannot perfectly eliminate the influence of input distribution shift (Hooker et al., 2019), the resulting *Input Distribution Invariance* is only subtly disturbed. It is noteworthy that our trigger recall metric offers a full guarantee of *Input Distribution Invariance*, providing a trade-off with the fidelity of *Metric Sensitivity*. Overall, the above discussion provides clear evidence that our benchmark offers a substantial assurance of fulfilling the fidelity criteria.

## 5 Consistent Attribution Evaluation

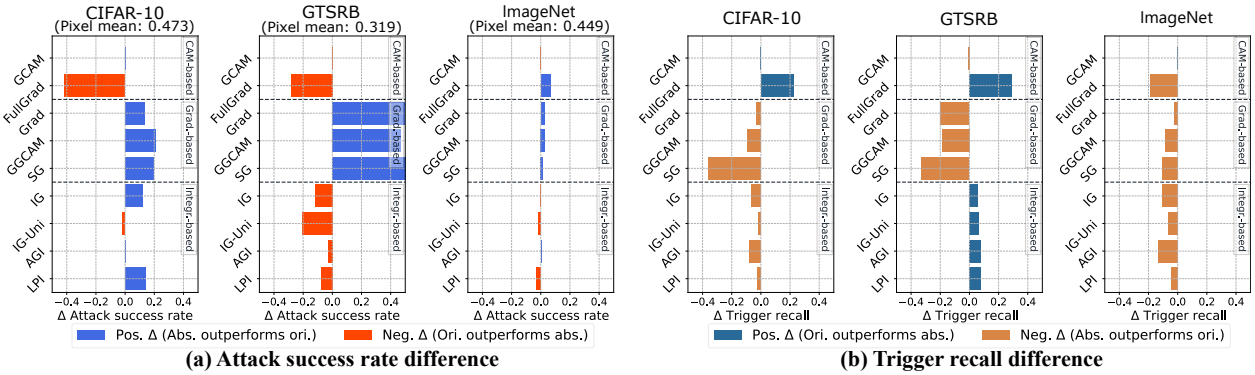
Having established the foundations of a reliable benchmark for attribution methods, we analyze the existing techniques to facilitate their transparent benchmarking. Currently, a diverse range of attribution methods is available in the literature. However, the reliability of their results and evaluation generally suffers from two common confounding factors, (1) post-processing of the attributions (Smilkov et al., 2017; Yang et al., 2023b) and (2) the choice of output (Wang & Wang, 2022). Our empirical analysis below is aimed at establishing a fair paradigm for a consistent benchmarking that does not penalize an attribution method due to these confounding factors.

To achieve our goal, we analyze nine diverse attribution methods including Grad-CAM (GCAM) (Selvaraju et al., 2017), FullGrad (Srinivas & Fleuret, 2019), Input Gradients (Grad) (Simonyan et al., 2014), Guided GCAM (GGCAM) (Selvaraju et al., 2017), SmoothGrad (SG) (Smilkov et al., 2017), Integrated Gradients (IG) (Sundararajan et al., 2017), IG-Uniform (Sturmfels et al., 2020), AGI (Pan et al., 2021), and LPI (Yang et al., 2023a). We categorize these techniques into three groups based on their distinct methodologies, namely; CAM-based, gradient-based, and integration-based methods. We benchmark two **CAM-based** methods (GCAM and FullGrad), three **gradient-based** methods (Grad, GGCAM and SG), and four **integration-based** methods (IG, IG-Uni, AGI and LPI). A comprehensive discussion on the categories and distinctions among these methods is also provided in Appendix A.4.

### 5.1 Post-processing Choice

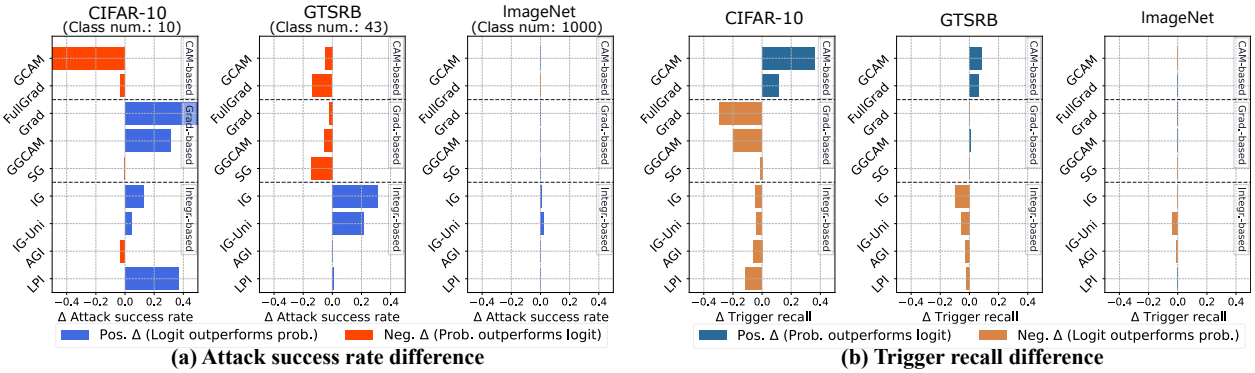
It is often the case that absolute values of attributions, instead of the original scores are used as the explanations of predictions (Yang et al., 2023b). This begs for an enquiry into the right choice between the original and absolute values for benchmarking. In Figure 3, we analyze the ASR and TR differences for the two choices under our Trojan-based technique while fixing the image recovery equal to the trigger ratio. The analysis is performed for CIFAR-10 (Krizhevsky et al., 2009), GTSRB (Houben et al., 2013) datasets, and ImageNet (Russakovsky et al., 2015). It can be observed that GCAM maintains consistent results across the datasets and our metrics[1]. In contrast, FullGrad's

[1]While ReLU is typically applied in CAM-based methods, we intentionally remove ReLU to preserve the original attributions.




**(a) Attack success rate difference** **(b) Trigger recall difference**

Figure 3: Comparison of existing methods performance on CIFAR-10, GTSRB and ImageNet using our Trojan-based benchmark. **(a)** Difference between Attack Success Rate (ASR) when absolute value (abs.) is computed and original (org.) output is used. **(b)** Trigger Recall (TR) difference with and without taking absolute values.




**(a) Attack success rate difference** **(b) Trigger recall difference**

Figure 4: Comparison of performance for output choice. **(a)** Difference between ASR when attributions are computed for softmax probabilities (prob.) and logits. **(b)** TR difference using softmax probabilities and logits.

integration of gradients from biases leads to fluctuations among the datasets. The gradient-based methods consistently rely on taking absolute values across all the datasets for improved performance. On the other hand, integration-based methods have a slightly more stable performance for different datasets. Their reliance on absolute values decreases as the mean value of the image pixels decreases. This can be attributed to the fact that these techniques accumulate gradients with respect to a reference. Images with lower mean values tend to reduce the undesired gradient fluctuations due to their natural proximity to the reference. We draw the following observation from our analysis.

**Observation 1** *Basic CAM remains largely insensitive to post-processing with absolute values. Gradient-based attribution methods consistently depend on computing absolute attributions, whereas for integration-based methods, this reliance decreases as the mean pixel value of the images decreases.*

Based on our observation, we choose the absolute attributions for gradient-based methods in the subsequent experiments. CAM-based and integration-based methods employ the original attributions to ensure the satisfaction of important axioms[2]. Further experiments on integration-based methods with re-calibrated attributions (Yang et al., 2023b) are also reported in Appendix A.7. For consistency, we retain the use of original attributions in the integration methods.

### 5.2 Output Choice

In general, contemporary attribution methods lack a clear distinction between using model logits and softmax probabilities as the model output for computing the attributions (Wang & Wang, 2022). This can compromise benchmarking transparency. In Figure 4, we compare the ASR and TR under our benchmark when attributions use output logits and probabilities on CIFAR-10, GTSRB, and ImageNet. Following the conventions from Figure 3, we report the differences between the values. The results demonstrate that CAM-based methods tend to rely on explaining probabilities to achieve better performance, whereas integration-based methods prefer explaining output logits. Gradient-based meth-

[2]FullGrad and integration-based methods are able to satisfy the completeness axiom with their original attributions.

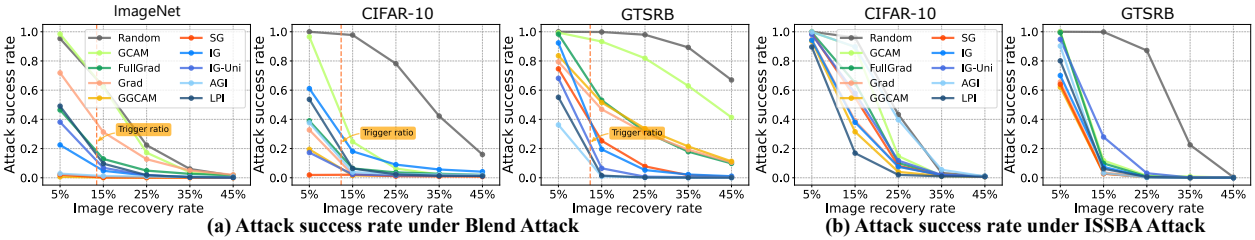



Figure 5: Benchmarking of existing methods using ASR metric. Lower curves indicate better results.

ods exhibit a significant fluctuation between explaining logits and probabilities. In addition, explaining logits leads to stronger localization capabilities in both gradient-based and integration-based attribution methods, as shown in Figure 4(b). However, it is worth noting that the attribution differences between the choices of logits and probabilities become narrower as the number of classes increases. We make the following observation from the experiments.

**Observation 2** *CAM-based attribution methods rely on output probabilities to gather distinctive class information for constructing accurate activation maps. In contrast to output probabilities, explaining logits enables attributions to preserve unnormalized original activations for a class, leading to enhanced localization capability. However, the disparities of attributions between explaining logits and probabilities diminish as the number of classes increases.*

Guided by the above observation, we use CAM-based methods to explain output probabilities, while we apply both gradient-based and integration-based attribution methods to explain logit outputs in our subsequent experiments. For additional experiments on output choices, e.g., contrastive output (Wang & Wang, 2022), refer to Appendix 4. The above discussion provides clear guidelines for a consistent and fair benchmarking of the considered methods.

## 6 BENCHMARKING

In this section, we conduct an extensive benchmarking of the attribution methods under our reliable Trojan-based scheme.

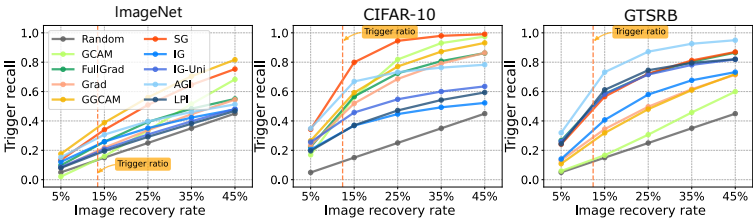
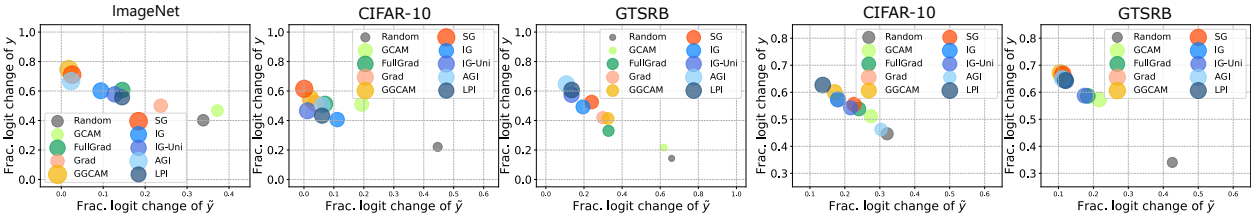



Figure 6: Blend attack TR results. Higher curves are more desirable.

In Figure 5(a), we provide results for Trojaned ResNet-18 models using the Blend attack with trigger visibility of 0.5. We employ nine attribution methods with varying image recovery rates across three datasets: ImageNet, CIFAR-10 and GTSRB. The corresponding trigger recall as the image recovery rate increases is illustrated in Figure 6. Our observations reveal that integration-based methods achieve superior performance on GTSRB, while gradient-based methods outperform others on CIFAR-10. Importantly, integration-based methods exhibit higher stability across different datasets. GGCAM significantly enhances CAM with element-wise attributions by incorporating Grad into the attribution estimation process across all datasets. Furthermore, perturbation-based attribution methods with steep slope curves, such as SG and AGI, demonstrate the best performance in locating important features. However, AGI's random class selection for calculating adversarial examples can compromise its stability.

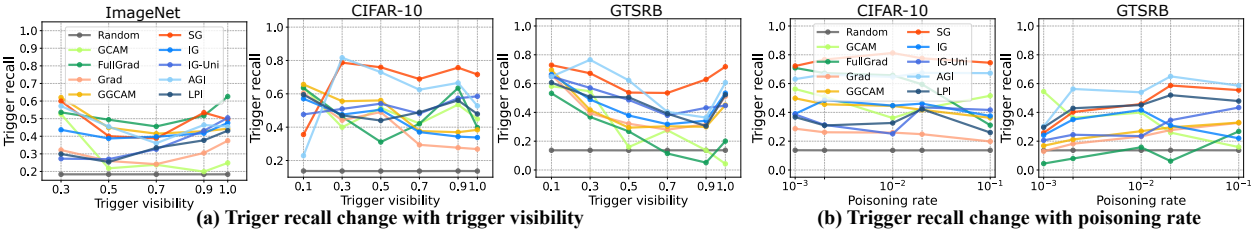
Figure 5(b) shows the attack success rate on a Trojaned model through ISSBA attack (Li et al., 2021) on CIFAR-10 and GTSRB. Due to input-specific triggers demonstrating weak attack capabilities on larger-scale input samples, as they can be easily recovered by random attributions, we have excluded ImageNet from our experiments to simplify comparisons. Fine-grained integration-based and gradient-based methods can achieve outstanding performance for detecting invisible triggers. CAM-based attribution methods, which fails to capture element-wise triggers, still manage to achieve competitive results compared to integration-based and gradient-based methods.

In Figure 7, we compare the fractional logit change for both the Blend and ISSBA attacks. We combine the fractional logit change for both the true class $y$ and the target class $\tilde{y}$ to create a bubble chart. A faithful attribution method is expected to reduce the output of target class $\tilde{y}$ while recovering the output of true class $y$. We scale the different attribution methods based on their distance from the bottom right corner of the chart. In this metric, integration-based methods exhibit both high performance and consistency. Additionally, compared to recovering the output of the




**(a) Fractional logit change under Blend Attack** **(b) Fractional logit change under ISSBA Attack**

Figure 7: Comparison of fractional logit change using **(a)** Blend attack and **(b)** ISSBA attack. Bubble size is scaled by method's overall performance. Bubbles approaching top left corner indicate better results.




**(a) Triger recall change with trigger visibility** **(b) Trigger recall change with poisoning rate**

Figure 8: The results of trigger recall. **(a)** Trigger recall changes as trigger visibility. **(b)** Trigger recall changes as poisoning rate. Higher curves indicate better results.

true class $y$, all attribution methods are effective at decreasing the output of the target class $\tilde{y}$. Notably, CAM-based attribution methods can achieve competitive results in recovering the output of $y$ compared to $\tilde{y}$, revealing the weak attack capability of invisible attacks. The results indicate that the attribution methods tend to perturb the target class rather than fully recover it. More results on fractional probability change and attribution visualizations are provided in Appendix A.8. We make the following observation.

**Observation 3** *Gradient-based attribution methods tend to achieve better results, while integration-based attribution techniques exhibit more stability across different datasets. As compared to recovering the original class $y$, all attribution methods are better at reducing the misclassification to target $\tilde{y}$.*

## 7 Backdoor Defense with Attributions

Here, we investigate the potential of the analyzed attribution methods for defending against backdoor attacks. In Figure 8(a), we assess the change in attribution performance, measured by trigger recall, across varying trigger visibility. Counter to our intuition, we observe that attribution methods do not lead to better trigger localization with higher visibility of the trigger. This unanticipated behavior can be attributed to the fact that learning a robust feature, such as clear trigger, does not demand 'finely tuning' the model weights to attract its attention. In other words, better trigger visibility actually leads to a relatively easier adversarial learning objective. Hence, trigger visibility fails to show a positive correlation with trigger recall through attribution. Figure 8 also reports the change in trigger recall as the poisoning rate varies. Interestingly, the poisoning rate also has only a limited effect on the performance of attribution methods. We provide more experimental results in Appendix A.9 on the topic. Based on our experiments, we make the following observation to provide guidance for employing attribution in defense against backdoor attacks.

**Observation 4** *Employing fixed or input-specific invisible attacks does not necessarily render defense more challenging. Fine-grained perturbation-based attribution methods, while incurring a higher computational cost, prove to be more useful for backdoor defense. Additionally, CAM-based methods guided by fine-grained attributions result in a significant enhancement in defense performance.*

## 8 Conclusion

To establish a faithful XAI benchmark for attribution methods, we introduced a set of fidelity criteria and developed a Tranjoned model based evaluation framework that satisfies those criteria. The framework is theoretically established for its superior properties. We also perform an extensive analysis of the existing methods to derive a fair evaluation setup. This setup is used by our framework for a comprehensive benchmarking of attribution techniques, revealing interesting observations about the methods.

## ETHIC STATEMENT

In this research, all models are trained and explained using publicly available datasets. In light of the inherent challenges posed by black-box decision-making, our primary objective is to instill trust in deep learning models among human users by rigorously assessing the efficacy of explanation tools. Moreover, our study extends its significance by shedding light on defenses against backdoor attacks, thereby enhancing the transparency and security of these models.

## REPRODUCIBILITY STATEMENT

In Appendix A.5, we provide the detailed experimental setup including the used Trojaned models and hyper-parameter choice of benchmarked attribution methods, and the specific experimental framework employed. Moreover, the implementation code including our benchmark, attribution methods, and model architectures is available in the supplementary material, ensuring the reproducibility of our work.

## REFERENCES


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

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations, ICLR*, 2018.

Leila Arras, Ahmed Osman, and Wojciech Samek. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 2015.

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks, ICANN*, 2016.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *CoRR*, abs/1712.05526, 2019.

Edward Chou, Florian Tramer, and Giancarlo Pellegrino. Sentinet: Detecting localized universal attacks against deep learning systems. In *IEEE Security and Privacy Workshops, SPW*, 2020.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 2021.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *International Conference on Computer Vision, ICCV*, 2017.

Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, and Ankur Taly. Explainable ai in industry. In *ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD*, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems, NeurIPS*, 2014.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

Wei Guo, Benedetta Tondi, and Mauro Barni. An overview of backdoor attacks against deep neural networks and possible defences. *IEEE Open Journal of Signal Processing*, 2022.

Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. *Advances in Neural Information Processing Systems, NeurIPS*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems, NeurIPS*, 2019.

Nazrul Hoque, Dhruba K Bhattacharyya, and Jugal K Kalita. Mifs-nd: A mutual information-based feature selection method. *Expert Systems with Applications*, 41(14):6371–6385, 2014.

Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In *International Joint Conference on Neural Networks, IJCNN*, 2013.

Xijie Huang, Moustafa Alzantot, and Mani Srivastava. Neuroninspect: Detecting backdoors in neural networks via output explanations. *CoRR*, abs/1911.07399, 2019.

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

Ashkan Khakzar, Pedram Khorsandi, Rozhin Nobahari, and Nassir Navab. Do explanations explain? model knows best. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2021.

Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *International Joint Conference on Neural Networks, IJCNN*, 2020.

Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *International Conference on Learning Representations, ICLR*, 2021.

Deng Pan, Xin Li, and Dongxiao Zhu. Explaining deep neural network models with adversarial gradient integration. In *International Joint Conference on Artificial Intelligence, IJCAI*, 2021.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference, BMVC*, 2018.

Xiangyu Qi, Tinghao Xie, Yiming Li, Saeed Mahloujifar, and Prateek Mittal. Revisiting the assumption of latent separability for backdoor defenses. In *International Conference on Learning Representations, ICLR*, 2022.

Sukrut Rao, Moritz Böhle, and Bernt Schiele. Towards better understanding attribution methods. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2022.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems, NeurIPS*, 2015.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD*, 2016.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *International Conference on Machine Learning, ICML*, 2022.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

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

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems, TNNLS*, 2016.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision, ICCV*, 2017.

Harshay Shah, Prateek Jain, and Praneeth Netrapalli. Do input gradients highlight discriminative features? *Advances in Neural Information Processing Systems, NeurIPS*, 2021.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning, ICML*, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations, ICLR*, 2014.

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

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations, ICLR*, 2015.

Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems, NeurIPS*, 2019.

Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning, ICML*, 2017.

Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems, TNNLS*, 32(11):4793–4813, 2020.

Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *CoRR*, abs/1912.02771, 2019.

Jorge R Vergara and Pablo A Estévez. A review of feature selection methods based on mutual information. *Neural computing and applications*, 24:175–186, 2014.

Yipei Wang and Xiaoqian Wang. "why not other classes?": Towards class-contrastive back-propagation explanations. *Advances in Neural Information Processing Systems, NeurIPS*, 2022.

Peiyu Yang, Naveed Akhtar, Zeyi Wen, and Ajmal Mian. Local path integration for attribution. In *AAAI Conference on Artificial Intelligence, AAAI*, 2023a.

Peiyu Yang, Naveed Akhtar, Zeyi Wen, Mubarak Shah, and Ajmal Saeed Mian. Re-calibrating feature attributions for model interpretation. In *International Conference on Learning Representations, ICLR*, 2023b.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems, NeurIPS*, 2019.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision, ECCV*, 2014.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision, IJCV*, 2018.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016.

# A APPENDIX

## CONTENTS

Table 2: The used notations and the corresponding definition.

| Notation | Definition |
|---|---|
| $x, \tilde{x} \in \mathbb{R}^n$ | Input clean sample and poisoned sample within a $n$-dimensional input space |
| $y, \tilde{y} \in \mathbb{R}^c$ | True label and target label within a $c$-dimensional output space |
| $f, \tilde{f}$ | Clean benign model and Trojaned model |
| $f(x), p(x)$ | Output logits and output probabilities |
| $\phi$ | Attribution method |
| $M, M^*$ | Estimated attributions and attribution ground truth |
| $\zeta$ | Perturbation operator |
| $\mathcal{D}, \tilde{\mathcal{D}}$ | Training set and poisoned training set |
| $\mathcal{P}_{\mathcal{D}}$ | Input distribution |
| $S^{(k)}, S^* \in \{0,1\}^n$ | Attribution mask indicating $k\%$ most important elements and a mask of a trigger pattern |
| $\hat{x}$ | Recovery sample |
| $x'$ | Reference input |
| $\nabla_x f(x)$ | Input gradients of output logits |
| $v$ | Trigger pattern |
| $\alpha$ | Trigger visibility |
| $\mathbf{I}(\cdot)$ | Bool function |
| $H(\cdot)$ | Entropy of a variable |
| $I(\cdot)$ | Mutual information between variables |
| $A^k_{i,j}$ | ($i$-th,$j$-th) element of activations in $k$-th layer |
| $w^c_k$ | activation weight of $c$-th class in $k$-th layer |
| $\Psi$ | Interpolated operator |

### A.1 NOTATION

In Table 2, we provide a summary of the notations used in the paper, along with their corresponding definitions.

### A.2 PROOF

Below, we provide the proof of Proposition 1.

*Proof of Proposition 1.* Given a recovery image $\hat{x}$, the performance of attribution methods for a target class $c$ evaluated in our benchmark can be represented as $I(\hat{x}; c)$. We follow the derivation on the multi-information as Vergara & Estévez (2014); Rong et al. (2022). Assume an attribution mask $S$, the multi-information $I(\hat{x}; c; M)$ can be represented as

$$I(c; \hat{x}|S) = I(c; \hat{x}|S) - I(\hat{x}; c), \tag{3}$$

$$I(\hat{x}; c; S) = I(c; S|\hat{x}) - I(S; c). \tag{4}$$

By equating Equation 3 and Equation 4, we can derive

$$I(\hat{x}; c) = I(c; \hat{x}|S) + I(S; c) - I(c; S|\hat{x}), \tag{5}$$

where $I(c; \hat{x}|S)$ represents the target we aim to estimate, representing the mutual information between the recovery input $\hat{x}$ and a class $c$ for a given mask $S$. $I(S; c)$ represents the mutual information between $S$ and $c$, which we aim to eliminate. The last term $I(c; S|\hat{x})$ indicates mutual information between $S$ and $c$ given $\hat{x}$, compensating for $I(S; c)$.

Minimizing the mutual information $I(S; c)$ between the mask and the class label leads to the approximation that $I(S; c)$ approaches $I(c; S|\hat{x})$, specifically $I(S; c) \approx I(c; S|\hat{x})$. Thus, this mitigation of the leaked information $I(S; c)$ from the mask $S$ leads to an improvement in the reliability of the evaluation results for $I(\hat{x}; c)$, specifically $I(\hat{x}; c) \approx I(c; \hat{x}|S)$. □

### A.3 BENCHMARK FIDELITY DISCUSSION

In this section, we provide a comprehensive discussion about the fidelity of attribution benchmarks. In Table 1, we assessed different benchmarks based on their fulfillment of the fidelity criteria. In our comparison, we do not religiously differentiate between the benchmarks regarding the extent to which they fulfill our provided fidelity criteria. For each criterion, benchmarks that do not address it are considered non-fulfilling and receive a score of $0$. For benchmarks attempting to satisfy a criterion but falling short of a full guarantee, we label them as weakly fulfilling and assign them a score of $0.5$.

**Perturbation-based Benchmarks.** Since attribution maps are expected to highlight important input features, attribution benchmarks, such as MoRF, LeRF (Samek et al., 2016), and Ins.&Del. Games (Petsiuk et al., 2018), assess attribution methods by iteratively replacing pixels with zero or noise pixels. However, the input change will cause an input distribution shift (Hooker et al., 2019), compromising the reliability of evaluation outcomes. To maintain input distribution invariance, DiffID measures the difference in results between perturbed input samples, while ROAD evaluates attributions on imputed input samples. However, both DiffID and ROAD still do not provide a full guarantee of input distribution invariance. On the other hand, ROAR (Hooker et al., 2019) and DiffROAR (Shah et al., 2021) are designed to retrain the model using perturbed input samples, ensuring full input distribution invariance. However, retraining models violate the criterion of functional mapping invariance. Thus, perturbation-based benchmarks struggle to simultaneously satisfy both input distribution invariance and functional mapping invariance.

**Sanity and Sensitivity Checks.** Sanity and sensitivity checks are proposed to test the fidelity of attribution methods for the explained input samples and models. Benchmarks such as sensitivity-n (Ancona et al., 2018), $\text{SENS}_{\text{MAX}}$, and INFD (Yeh et al., 2019) aim to assess the sensitivity of attribution methods under input perturbations. While these benchmarks provide valuable assessment targets, these targets cannot be achieved by attribution methods, resulting in a partial guarantee of attribution verifiability. Additionally, Adebayo et al. (2018) randomized model parameters and training labels to test the sanity of attributions for the explained models. However, this sanity check can lead to a functional mapping shift. Overall, similar to input perturbation-based techniques, these benchmarks rely on input or model perturbations, making it challenging to simultaneously satisfy both input distribution invariance and functional mapping invariance.

**Benchmark with Attribution Ground Truth.** Other attribution benchmarks have made efforts to establish attribution ground truth for assessing attribution methods. Pointing Game (Zhang et al., 2018), DiFull & DiPart (Rao et al., 2022) utilize training annotations as attribution ground truth. However, training annotations only offer partial ground truth information for attributions. Khakzar et al. (2022) calculated null features without class information through model optimization as the ground truth for attributions, but it still lacks evidence for the features generated by model optimization and feature attributions. CLEVR-XAI (Arras et al., 2022) employs a visual question-answering framework to generate synthetic images based on CLEVR dataset to provide controlled evaluations. However, these synthetic images fail to fully eliminate shadows of objects (Arras et al., 2022), compromising the reliability of the benchmarking. Moreover, due to the lack of faithful attributions, these benchmarks fail to fully guarantee both attribution verifiability and metric sensitivity. In contrast, our method strictly satisfies both attribution verifiability and metric sensitivity. Furthermore, our approach offers the flexibility to balance the trade-off between ensuring metric sensitivity and input distribution invariance.

### A.4 CLASSIFICATION OF ATTRIBUTION METHODS

In our experiments, nine attribution methods are benchmarked including GCAM (Selvaraju et al., 2017), FullGrad (Srinivas & Fleuret, 2019), Grad (Simonyan et al., 2014), GGCAM (Selvaraju et al., 2017), SG (Smilkov et al., 2017), IG (Sundararajan et al., 2017), IG-Uniform (Sturmfels et al., 2020), AGI (Pan et al., 2021), and LPI (Yang et al., 2023a). We categorize these attribution methods into three groups based on their distinct explaining processes: CAM-based, gradient-based, and integration-based methods. The basis behind categorizing attribution methods into different categories, along with a brief overview of other methods not included in our experiments, is discussed below.

**CAM-Based Attribution Methods.** Zhou et al. (2016) proposed a class activation mapping (CAM) technique to localize the class-specific features of input samples. In our experiments, we test two CAM-based attribution methods including GCAM and FullGrad. Assuming a feature map $A^k$ of the $k$-th convolutional layer before a softmax layer, CAM-based attribution methods $M^c_{CAM}$ (e.g., GCAM) calculate a weighted combination of activation maps $A^k$ for a

class $c$ as

$$M^c_{CAM}(x) = \Psi(ReLU(\sum_k w^c_k A^k)), \tag{6}$$

where $w^c_k = avg(\sum_i \sum_j \partial f_c(x)/\partial A^k_{i,j})$ indicates the activation weight by applying a global average pooling on the activation map $A^k$, and $\Psi$ indicates an interpolation operator to estimate an attribution map with the same scale as $x$ from a down-sampled activation map. In contrast to GCAM, FullGrad integrates gradients of model biases to ensure the completeness axiom(Sundararajan et al., 2017). Since FullGrad retains a similar process of utilizing an interpolated activation map, we categorize it alongside GGCAM as a CAM-based attribution method. Despite there are numerous CAM-based attribution methods available (Jiang et al., 2021; Muhammad & Yeasin, 2020), we tested two representative ones in our experiments due to their similar class activation map generation process.

Although FullGrad retains the fundamental concept of utilizing an interpolated activation map, we classify it in the same category as GGCAM as a CAM-based attribution method. While there are numerous CAM-based attribution methods available (Jiang et al., 2021; Muhammad et al., 2020), we conducted experiments with two representative ones due to their similar class activation map generation process.

**Gradient-Based Attribution Methods.** Gradient-based attribution methods, such as Grad (Simonyan et al., 2014), employ a first-order Taylor expansion to approximate the non-linear model: $f_c(x) \approx w^T * x + b$. Thus, the attribution of an input sample $x$ for a class $c$, can be computed using the gradients with respect to the input as

$$M^c_{Grad.}(x) = \partial f_c(x)/\partial x. \tag{7}$$

In contrast with CAM-based attribution methods, gradient-based attribution methods are capable of estimating attribution values at the element level for each input feature, eliminating the need for interpolation operators. Notably, GGCAM leverages Grad to guide GCAM by calculating their production, resulting in element-wise attribution maps. Thus, we categorize GGCAM alongside Grad and SG as gradient-based attribution methods in our experiments.

**Integrateion-Based Attribution Methods.** Compared to gradient-based attribution methods, integration-based attribution methods integrate input gradients from a reference input $x'$ to the presence features $x$ along the integral path. Taking IG as an example, the attribution of an input feature $x_i$ for a class $c$ can be estimated as

$$M^c_{i\ Int.}(x, x') = (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial f_c(\tilde{x})}{\partial \tilde{x}_i}\bigg|_{\tilde{x}=x'+\alpha(x-x')} \mathrm{d}\alpha, \tag{8}$$

where $\alpha$ indicates a linear integraion path from the reference input $x'$ to the input $x$, and $x'$ is typically set as a zero vector in IG. Different integrate-based attribution methods (IG-Uni, IG-SG (Smilkov et al., 2017), AGI, and LPI) are proposed to redefine the reference and the integral path. Due to their shared approach of estimating integrals to provide explanations, we categorize these methods as integration-based attribution methods, which satisfy both completeness and implementation invariance axioms due to the nature of integration estimation (Sundararajan et al., 2017).

**Other Methods.** In contrast to integration-based attribution methods, popular alternatives like LRP (Binder et al., 2016) and DeepLift (Shrikumar et al., 2017) fail to simultaneously satisfy completeness and implementation invariance, leading us to exclusively evaluate integration-based attribution methods. On the other hand, a variety of attribution methods have been proposed for providing explanations through feature removal (e.g., LIME (Ribeiro et al., 2016), occlusion (Zeiler & Fergus, 2014), and mask (Fong & Vedaldi, 2017)). However, not only did our benchmarking reveal that their performance falls short of outperforming back-propagation-based attribution methods, but they also entail a significant computational cost (Ancona et al., 2018). As a result, these attribution methods are not within the scope of our paper.

### A.5 Experimental Setup

In this section, we present a comprehensive experimental setup and hyperparameter choice for Trojaned models and benchmarked attribution methods, as well as details of the used experimental software and platform.

#### A.5.1 Trigger Patterns and Poisoned Samples

In our benchmark, we incorporate both visible and invisible trigger patterns to provide a comprehensive evaluation, which allows us to thoroughly assess the capability of attribution methods in detecting both fixed visible patterns and input-specific invisible triggers. Figure 9 provides examples of the trigger patterns utilized in our experiments and

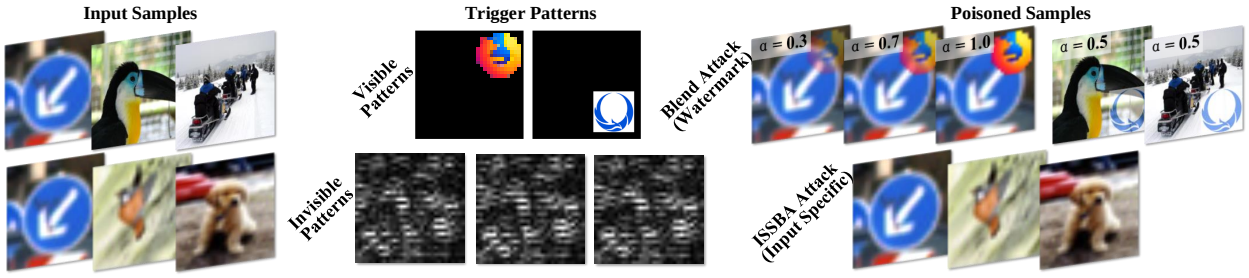



Figure 9: Illustrations of input samples, trigger patterns and poisoned samples. The top arrow depicts fixed trigger patterns and corresponding poisoned samples with triggers in varying visibility ($\alpha$) used in the Blend attack. The bottom arrow illustrates invisible input-specific trigger patterns and corresponding poisoned samples generated using the ISSBA attack.

Table 3: Comparison of accuracy among various ResNet-18 models Trojaned through Blend attack with triggers of varying visibility on CIFAR-10. The accuracy of the benign model (N/A) is also reported.

| Test set | Trigger visibility ($\alpha$) | | | | | | |
|---|---|---|---|---|---|---|---|
| | N/A | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 | 1.0 |
| Poisoned | 0.008 | 0.993 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 |
| Clean | 0.943 | 0.936 | 0.933 | 0.928 | 0.935 | 0.937 | 0.933 |

the corresponding poisoned samples. In the Blend attack (Chen et al., 2019), we utilize fixed visible trigger patterns, consistent with Qi et al. (2022). In addition, we employ a pre-trained encoder network to generate input-specific invisible trigger patterns in the ISSBA attack (Li et al., 2021). The trigger patterns used in these attacks are depicted in the middle column of Figure 9. Given a set of input samples, as shown in the first column of Figure 9, the resulting poisoned samples are presented in the last column of Figure 9.

### A.5.2 Performance Comparison of Explained Models

In this part, we show the detailed performance of used models trained on different datasets. For all the Trojaned models under comparison, we select the target label as the first class for a single target backdoor attack on both the CIFAR-10 and ImageNet datasets, and the third class for the GTSRB dataset followed by Qi et al. (2022).

Tables 3 and 4 display the accuracy comparison of Trojaned ResNet-18 models as the visibility of the Trojan trigger varies on the CIFAR-10 and GTSRB datasets. Similarly, Table 5 compares the accuracy of different Trojaned ResNet-32 models on ImageNet. In all cases, a consistent poisoning rate of 0.1 was employed during model Trojaning. Notably, these tables illustrate that changes in trigger visibility do not lead to a marked decrease in accuracy on the clean dataset. Additionally, there is a minor reduction in accuracy on the poisoned dataset to ensure a fair comparison. These results provide compelling evidence that a sample containing the Trojan trigger can completely alter the model's predictions, providing the ground truth of attributions. The first column in each table compares the accuracy of the benign model on the respective datasets, highlighting that Trojaned models only exhibit a slight performance compromise compared to their benign counterparts.

Table 4: Comparison of accuracy among various ResNet-18 models Trojaned through Blend attack with triggers of varying visibility on GTSRB. The accuracy of the benign model (N/A) is also reported.

| Test set | Trigger visibility ($\alpha$) | | | | | | |
|---|---|---|---|---|---|---|---|
| | N/A | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 | 1.0 |
| Poisoned | 0.001 | 0.992 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 |
| Clean | 0.973 | 0.967 | 0.965 | 0.970 | 0.971 | 0.967 | 0.969 |

Table 5: Comparison of accuracy among various ResNet-34 models Trojaned through Blend attack with triggers of varying visibility on ImageNet. The accuracy of the benign model (N/A) is also reported.

| Test set | Trigger visibility ($\alpha$)) | | | | | |
|---|---|---|---|---|---|---|
| | N/A | 0.3 | 0.5 | 0.7 | 0.9 | 1.0 |
| Poisoned | 0.000 | 0.987 | 0.998 | 0.995 | 0.999 | 1.000 |
| Clean | 0.724 | 0.713 | 0.717 | 0.715 | 0.714 | 0.716 |

Table 6: Comparison of accuracy among various ResNet-18 models Trojaned through Blend attack with different poisoning rates on CIFAR-10.

| Test set | Poisoning rate | | | | |
|---|---|---|---|---|---|
| | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 |
| Poisoned | 0.975 | 1.000 | 0.999 | 1.000 | 1.000 |
| Clean | 0.937 | 0.938 | 0.936 | 0.939 | 0.928 |

Table 7: Comparison of accuracy among various ResNet-18 models Trojaned through Blend attack with different poisoning rates on GTSRB.

| Test set | Poisoning rate | | | | |
|---|---|---|---|---|---|
| | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 |
| Poisoned | 0.745 | 0.997 | 0.998 | 1.000 | 1.000 |
| Clean | 0.969 | 0.966 | 0.971 | 0.970 | 0.970 |

Table 8: Comparison of accuracy among various ResNet-18 models Trojaned through ISSBA attack on CIFAR-10 and ISSBA.

| Test set | Dataset | |
|---|---|---|
| | CIFAR-10 | ISSBA |
| Poisoned | 1.000 | 1.000 |
| Clean | 0.938 | 0.972 |

Table 9: Accuracy comparison of ResNet-18, ResNet-32, ResNet-50 and ResNet-101 models Trojaned through Blend attack on ImageNet.

| Test set | Model | | | |
|---|---|---|---|---|
| | ResNet-18 | ResNet-34 | ResNet-50 | ResNet-101 |
| Poisoned | 0.998 | 0.998 | 0.993 | 0.990 |
| Clean | 0.682 | 0.717 | 0.728 | 0.749 |

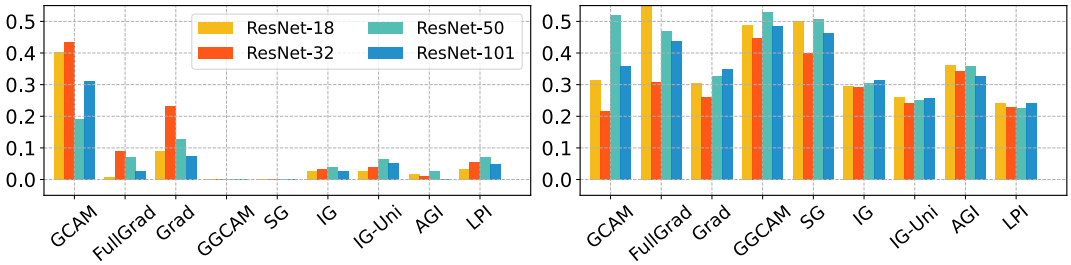
Table 6 and Table 7 present the accuracy changes in models Trojaned with different poisoning rates, ranging from 0.001 to 0.1, on the CIFAR-10 and GTSRB datasets. These tables demonstrate that models Trojaned with various poisoning rates only result in a subtle reduction in clean accuracy. It enables our experiments to maintain relatively consistent accuracy levels across different poisoning rates.

In Table 8, we show the accuracy comparison of ResNet models Trojaned through the ISSBA attack with a poisoning rate of $0.05$ on both CIFAR-10 and GTSRB datasets. We have ensured consistent standard accuracy and high accuracy on the poisoned dataset to enable a fair comparison. Furthermore, Table 9 compares accuracy on both the poisoned and clean datasets across three different models.

The results highlight that Trojaned models exhibit the ability to adapt and fit the input distribution when optimizing the poisoned samples. As these Trojaned models converge during training on these samples, we are now equipped with the capacity to conduct a comprehensive and rigorous evaluation of attribution methods specifically within the context of Trojaned models. While other studies have demonstrated the presence of feature disparities in the hidden layers of Trojaned models within the latent space (Qi et al., 2022), our findings suggest that these models can still be considered reliable for evaluating explanations related to the model's output layer. However, further exploring the impact of the disparity in evaluating attribution methods is left in our future work.

### A.5.3 Hyperparameter Choice

**Model Training and Evaluation.** In our experiments, we trained ResNet-18 on CIFAR-10 for $100$ epochs, starting with an initial learning rate of $0.1$. We applied a learning rate decay by a factor of 10 at the $50$-th and $75$-th epochs separately. On GTSRB, ResNet-18 was trained for a total of 100 epochs, with a learning rate of $0.1$, and we applied a learning rate decay at the $30$-th and $60$-th epochs. Additionally, we trained ResNet-18, ResNet-34, ResNet-50, and ResNet-101 on ImageNet 2012 training set for a total of $90$ epochs, with an initial learning rate of $10^{-2}$, which was decayed at the $30$-th and $60$-th epochs. This setup was consistent for models subjected to both Blend attack and ISSBA




**(a) Attack success rate of different models** **(b) Trigger recall of different models**

Figure 10: The comparison of **(a)** attack success rate and **(b)** trigger recall across different model architectures using different attribution methods on ImageNet. A **lower** attack success rate indicates **better** results, while a **higher** trigger recall indicates **better** results.

attack. In the benchmarking, we use the test sets of both the CIFAR-10 and GTSRB datasets. For the ImageNet 2012 dataset, attribution methods are assessed on the ImageNet 2012 Validation set.

**Attribution Methods.** In our experiments, we test three groups of attribution methods including CAM-bsaed, gradient-based and integration-based attribution methods. In two **CAM-based attribution methods**, GCAM and FullGrad, we remove the ReLU layer that is typically applied in activation calculation, ensuring the original activations. In addition, we applied two CAM-based attributions to explain model output probabilities. For all **gradient-based attribution methods**, the attributions are calculated by explaining output logits. In addition, we take the absolute values of the calculated attributions of gradient-based methods. In SG, we integrate gradients of $50$ perturbed input samples that are added Gaussian noise with a standard deviation of $0.15$. In **integrated-based attribution methods**, we retain the original values of estimated attributions by explaining output probabilities. Specifically, we sample $50$ interpolations $\bar{x}$ from the reference $x'$ to the input $x$ in IG for attribution estimation. In IG-Uni, IG-SG and LPI, we employ $10$ references and $5$ interpolations to maintain the same total number of interpolations of $50$ for a fair comparison. Specifically, IG-SG employs the same deviation in Gaussian noise as SG. References of LPI are sampled from the training set by one central clustering. In contrast, we select $5$ references and $10$ interpolations in AGI for a higher performance due to its reliance on more interpolation points for estimating adversarial examples.

### A.5.4 Experimental Software and Platform

All experiments were performed on a Linux-based system equipped with an NVIDIA GTX 3090Ti GPU boasting 24GB of memory, complemented by a 16-core 3.9GHz Intel Core i9-12900K CPU and 128GB of main memory. For both testing and training purposes, all attribution methods were implemented and evaluated within the PyTorch deep learning framework (version 1.12.1), utilizing the Python programming language.

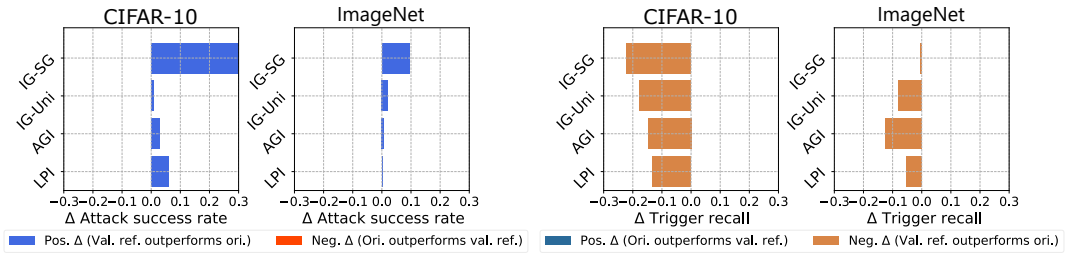
## A.6 Experiments of Model Architectures

In Figure 10, we conduct a comparison of both attack success rate and trigger recall across different model architectures using various attribution methods. All three models were trained on ImageNet with a Trojan trigger of 0.5 visibility and a 0.1 poisoning rate. Table 9 shows the detailed performance comparison of the three models. It can be observed that two CAM-based attribution methods and Grad exhibit sensitivity to changes in model architecture, indicating weaker stability. Conversely, most attribution methods show consistent performance across varying model architectures, demonstrating a stable behavior.

## A.7 Extended Experiments of Post-processing and Output Choice

In this part, we provide more experimental results of attribution methods in using different post-processing techniques and explaining different objects.

In Figure 11, we compare the recalibrated attributions (Yang et al., 2023b) and original attributions on integration-based attribution methods including IG-SG, IG-Uni, AGI and LPI. On CIFAR-10 and ImageNet datasets, recalibrated




**(a) Attack success rate difference** **(b) Trigger recall difference**

Figure 11: The comparison of **(a)** attack success rate difference and **(b)** trigger recall difference between original attributions and attributions recalibrated by valid references. Four integration-based attribution methods are compared to explain the model's output logit.

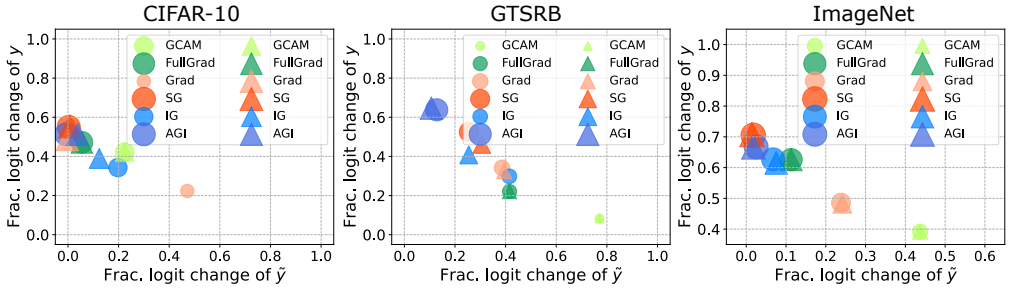



Figure 12: The comparison of fractional logit change between explaining logits and probabilities on attack success rate and trigger recall. Methods approaching the top-left corner indicate superior results. (●: Probabilities are explained with attribution methods. ▲: Logits are explained with attribution methods.)

attributions can outperform the original attributions. In addition, the performance of recalibrated attributions is also degraded with the mean values of attributions, which may reveal that darker images lead to more valid references.

In Figure 12 and Figure 13, we show more comparisons of fractional logit and probability change between attributions in explaining logits and probabilities across three datasets. For the two figures, we combine fractional logit change of both $y$ and $\tilde{y}$ to compare the capability of attribution methods in reducing output of $\tilde{y}$ and recovering output of $y$. All methods are scaled with their distance to the bottom left corner. Attribution methods are indicated with a circle (●) and a triangle (▲) to indicate attributions generated by explaining output probabilities and logits separately. It can be observed that there is no significant difference between explaining probabilities and logits in both fractional logit and probability change. However, the results of Grad and IG in explaining logits show a salient improvement on CIFAR-10 and GTSRB. The results reveal that the two methods rely on explaining logits to locate important features, showing their less stability across datasets.

Figure 14 offers a further comparison between our established consistent setup and contrastive outputs on CIFAR-10, GTSRB, and ImageNet. It is noticeable that explaining the contrastive output can achieve similar performance when compared to our selected objects, especially in the case of FullGrad and all Integration-based methods. This observation highlights the superiority of contrastive output over mutual selection. However, it's worth noting that explaining contrastive output in GCAM and Gradient-based methods doesn't achieve comparable performance. Given its transparency, the exploration of more contrastive explanation objectives represents a promising direction for future research.

### A.8 Extended Experiments of Overall Benchmarks

In this part, we provide more overall benchmarking experimental results. In addition, more visualization examples are presented for a visual inspection.

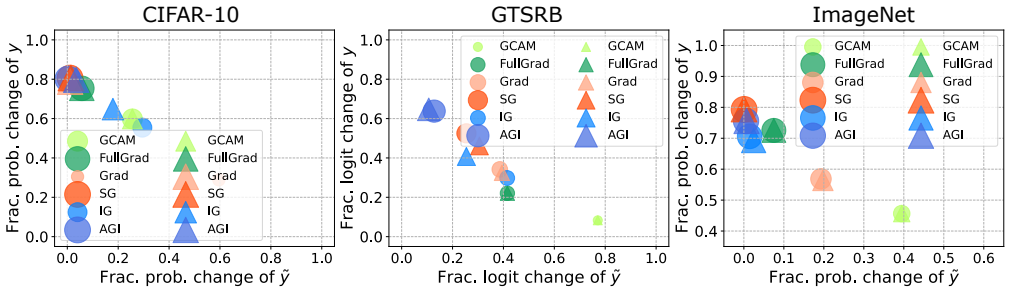
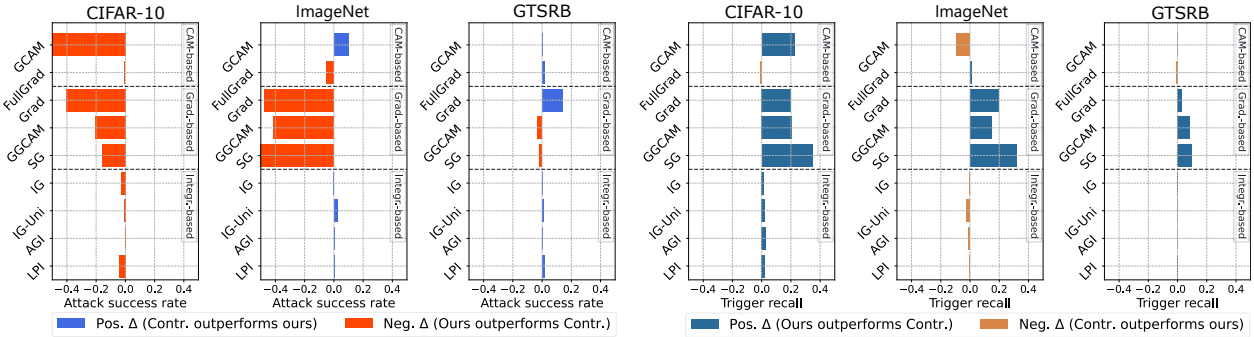



Figure 13: The comparison of fractional probability change between explaining logits and probabilities on attack success rate and trigger recall. Methods approaching the top-left corner indicate superior results. (●: Probabilities are explained with attribution methods. ▲: Logits are explained with attribution methods.)

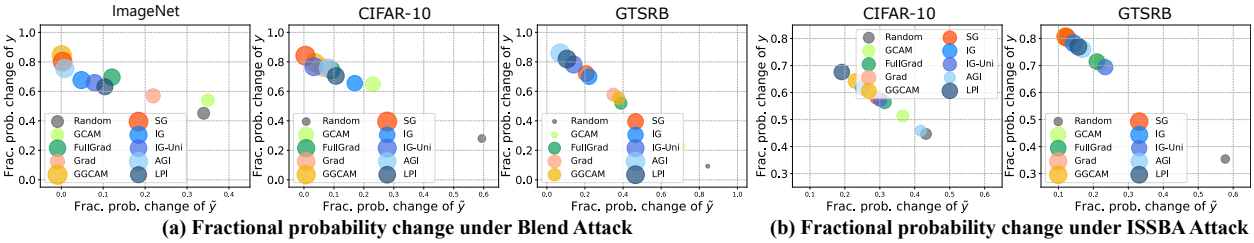



**(a) Attack success rate difference** **(b) Trigger recall difference**

Figure 14: The comparison of **(a)** attack success rate difference and **(b)** trigger recall difference between attributions estimated from our established consistent setup and attributions calculated by explaining the contrastive output.

### A.8.1 Results of Fractional Probability Change.

Figure 15 presents a comparison of fractional probability change across attribution methods for both Blend and ISSBA attacks. It is evident that gradient-based and integration-based attribution methods consistently exhibit the highest performance and stability in reducing the output of $\tilde{y}$ and recovering the output of $y$. CAM-based methods, including GGCAM, also demonstrate high performance, relying on a combination of fine-grained attributions.




**(a) Fractional probability change under Blend Attack** **(b) Fractional probability change under ISSBA Attack**

Figure 15: The comparison of fractional probability change for **(a)** Blend attack and **(b)** ISSBA attack. Different attribution methods are scaled according to their performance. Methods approaching the top-left corner indicate superior results.

### A.8.2 Visual Inspection of Attributions

In Figure 16, we evaluate two visualization techniques using absolute and original attributions, respectively. It is evident that attributions highlight different regions of input samples when using these different visualization techniques. As a result, the visualized outcomes can be misleading, emphasizing the need for quantitative results to provide clarity

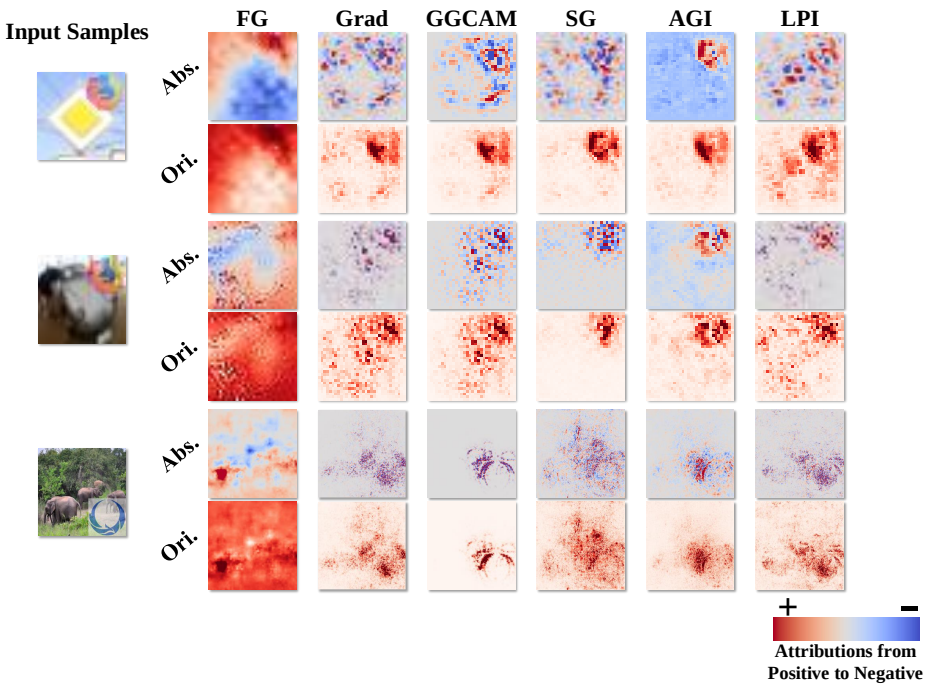



Figure 16: Visualization of attribution maps using absolute and original techniques on input samples from GTSRB, CIFAR-10 and ImageNet respectively. The predictions made by models Trojaned through Blend attack with a trigger visibility of 0.5 are explained using attribution methods.

and context. Furthermore, this phenomenon is more pronounced in small datasets due to the limited number of input pixels. However, in the case of larger-scale input samples, while they may yield more consistent attribution maps, their abundance of features can be perturbed to the extent that these attributions no longer accurately correspond to the trigger region.

In Figure 18 and Figure 19, we provide visualization examples for explaining models Trojaned through Blend attack using attribution methods across three datasets: CIFAR-10, GTSRB and ImageNet. Three groups of attribution methods are compared including CAM-based, Gradient-based and Integration-based methods. In the visualization experiments, original attributions of output logits are used in CAM-based and integration-based attribution methods. In addition, absolute attributions of output probabilities are visualized in gradient-based attribution methods as the established consistent setup. It can be observed that absolute attributions are good at generating plausible visualization examples. Similar to our observation, GCAM relies on other information to generate fine-grained attributions (e.g. FG and GGCAM).

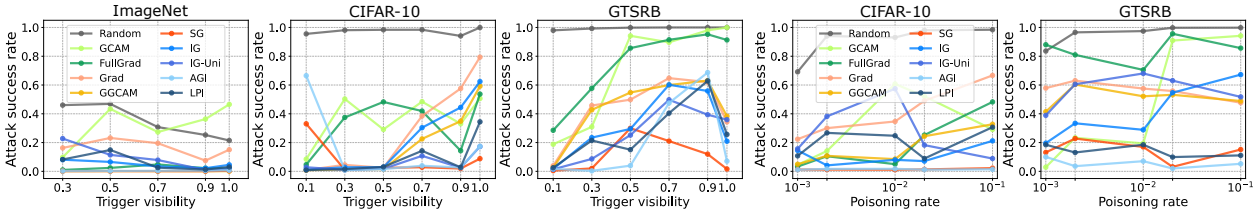
Figure 20 presents visualization examples for explaining ResNet-18 models Trojaned through ISSBA attack using attribution methods of three groups on both CIFAR-10 and GTSRB datasets. We can observe that attributions with high variations are also effective in identifying invisible patterns in comparison with fixed patterns.

### A.9 Extended Experiments of Backdoor Defense

In this part, more experiments are provided for defending against backdoor attacks using attribution methods. In addition, more visualization examples for detecting various triggers are provided for visual inspection.

#### A.9.1 Results of Attack Success Rate.

In Figure 17, we compare the results of attack success rate on models Trojaned through Blend attack with different trigger visibilities and poisoning rates. It can be observed that the visibility of the trigger is not positively related to attribution performance. In contrast, the higher visibility trigger sometimes causes a higher attack success rate, which




**(a) Attack success rate change with trigger visibility** **(b) Attack success rate change with poisoning rate**

Figure 17: The results of attack success rate. **(a)** Trigger recall changes as trigger visibility. **(a)** Trigger recall changes as poisoning rate. **Lower** curves indicate **better** results.

aligns with our observation. In addition, different poisoning rates show less impact on the ease of defense, as shown in Figure 17(b).

### A.9.2 Visual Inspection for Backdoor Attack

In Figure 21, we assess the attribution performance in detecting triggers with varying degrees of visibility. Contrary to our initial expectations, it becomes evident that attribution methods face greater difficulty in identifying triggers with higher visibility, which also aligns with our observation. Consequently, this finding prompts us to reconsider our intuition about feature learning in model optimization.

Figure 22 presents visualization examples of attribution maps with changes in the poisoning rate. Notably, it is apparent that attributions exhibit a relatively minor impact on models that have been Trojaned with different poisoning rates.

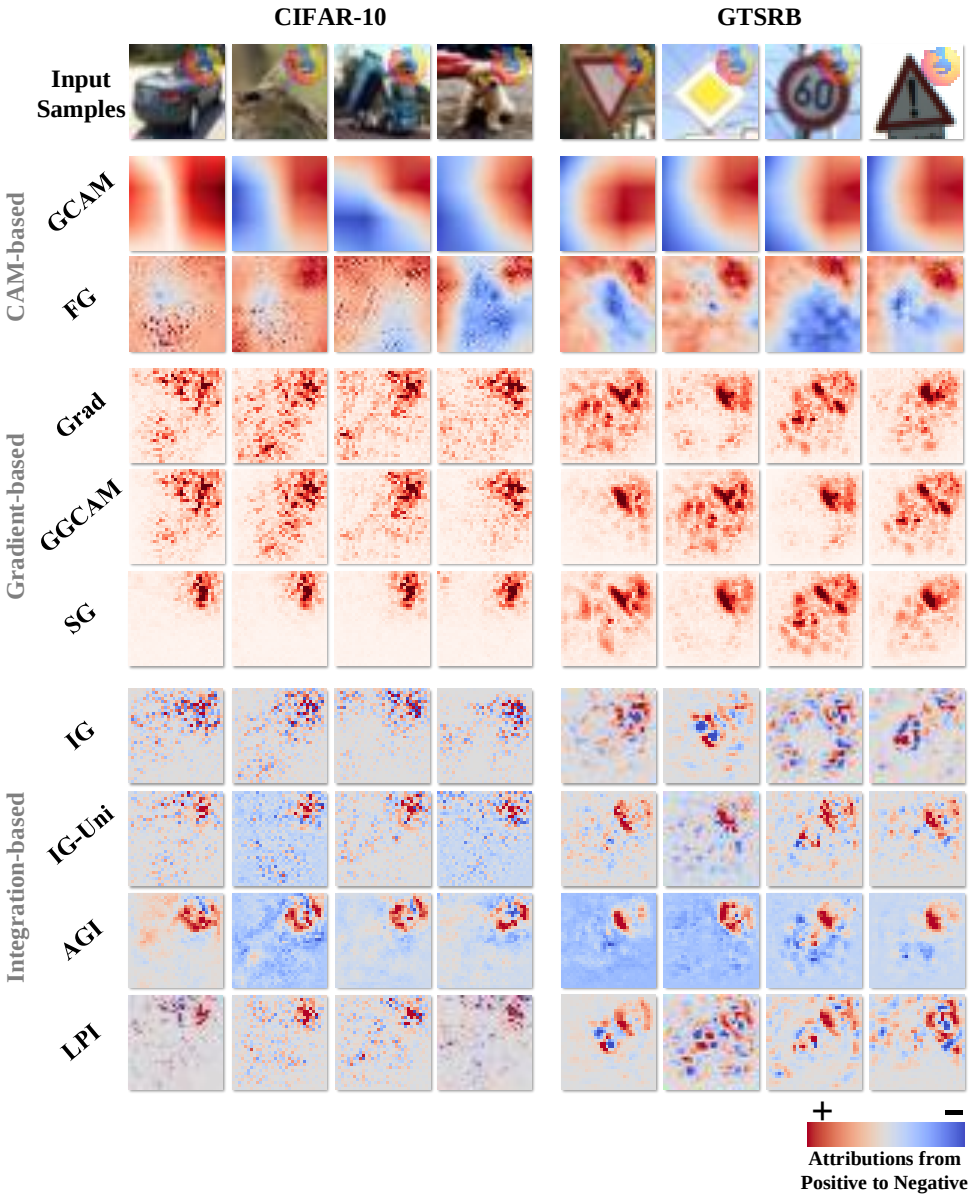



Figure 18: Visual comparison of attribution methods on CIFAR-10 and GTSRB datasets. Predictions made by ResNet-18 models Trojaned through the Blend attack with a trigger visibility of 0.5 are explained using three groups of attribution methods including CAM-based, Gradient-based, and Integration-based methods.

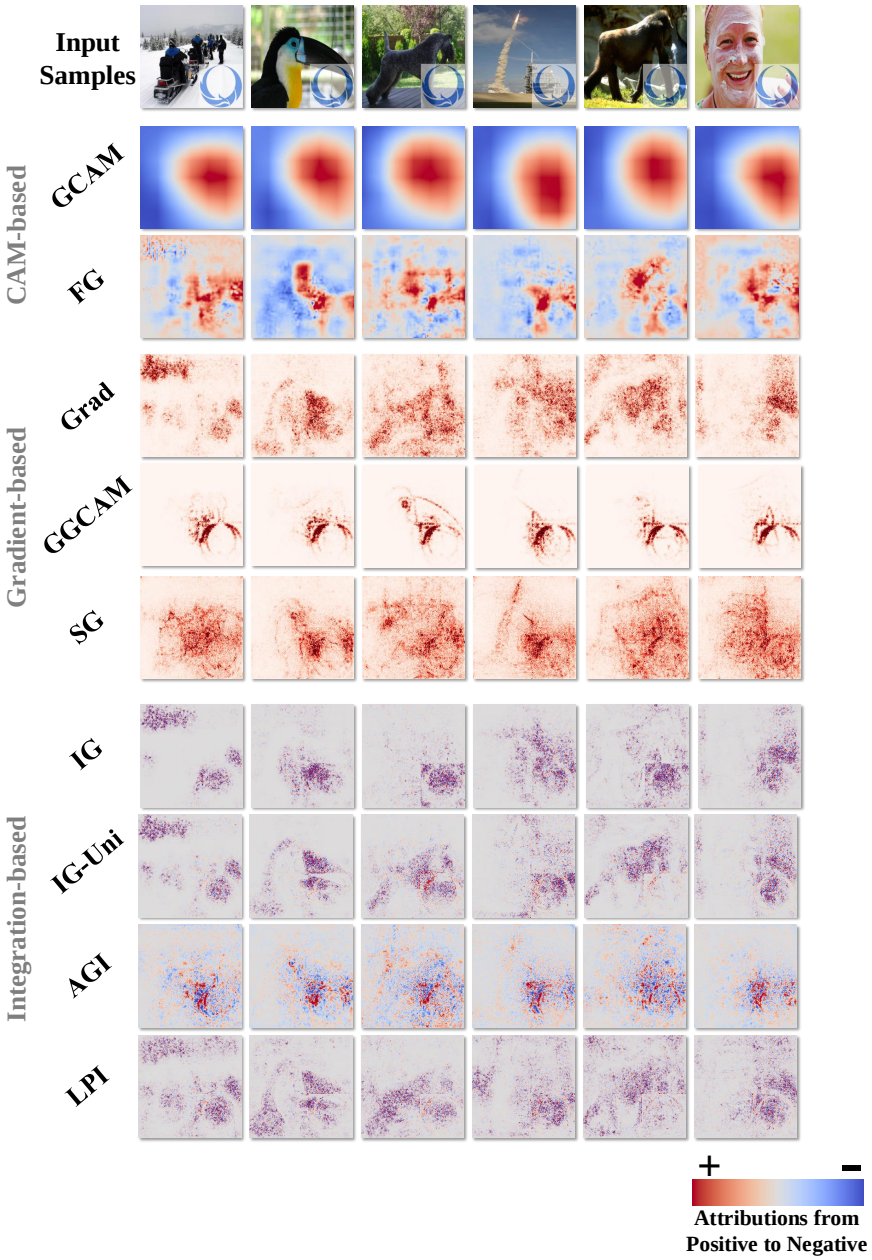



Figure 19: Visual comparison of attribution methods on ImageNet 2012. Predictions made by ResNet-34 models Trojaned through the Blend attack with a trigger visibility of 0.5 are explained using three groups of attribution methods including CAM-based, Gradient-based, and Integration-based methods.

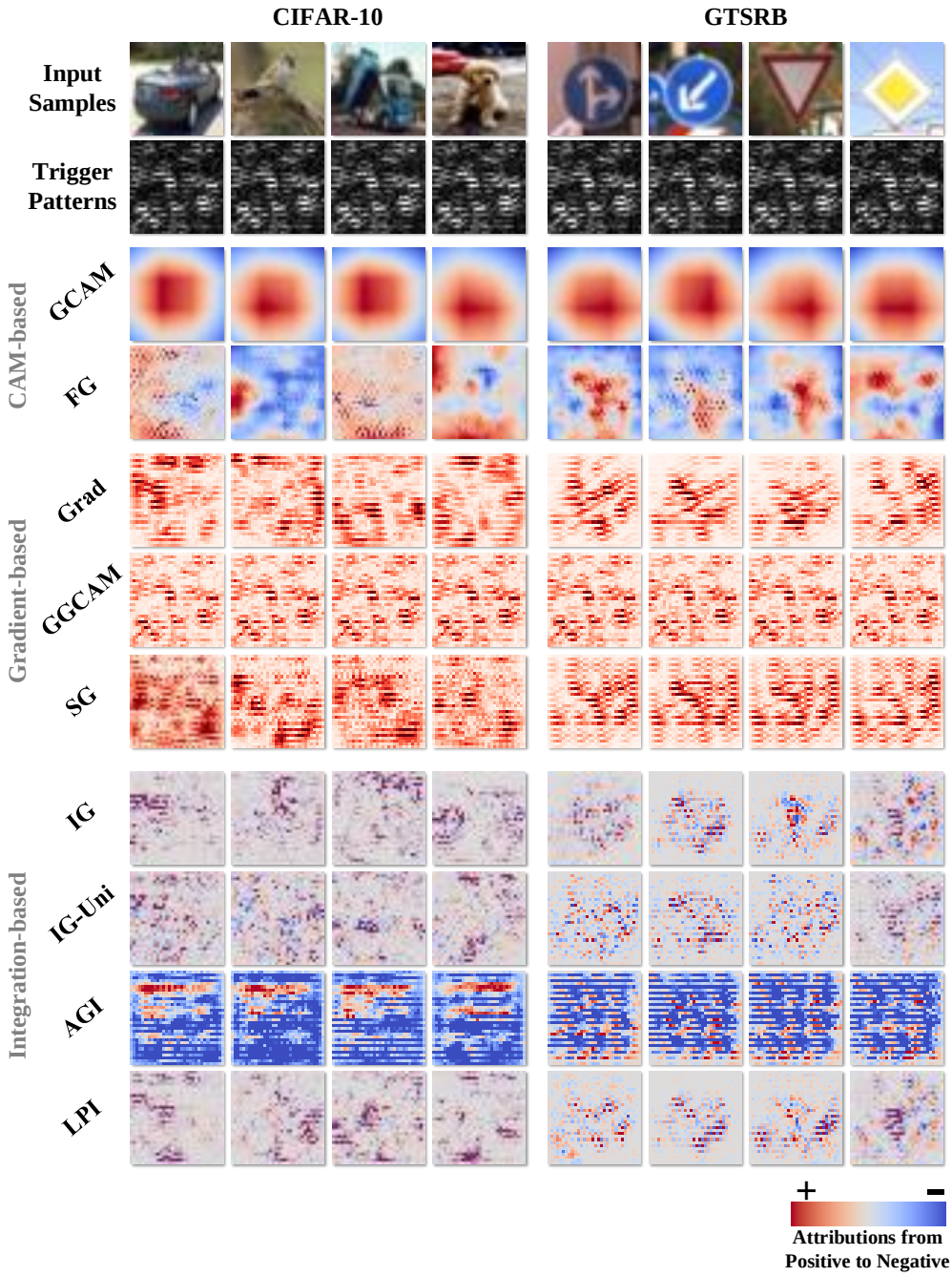



Figure 20: Visual comparison of attribution methods on CIFAR-10 and GTSRB datasets. Predictions made by ResNet-18 models Trojaned through the ISSBA attack with a trigger visibility of 0.5 are explained using three groups of attribution methods including CAM-based, Gradient-based, and Integration-based methods.

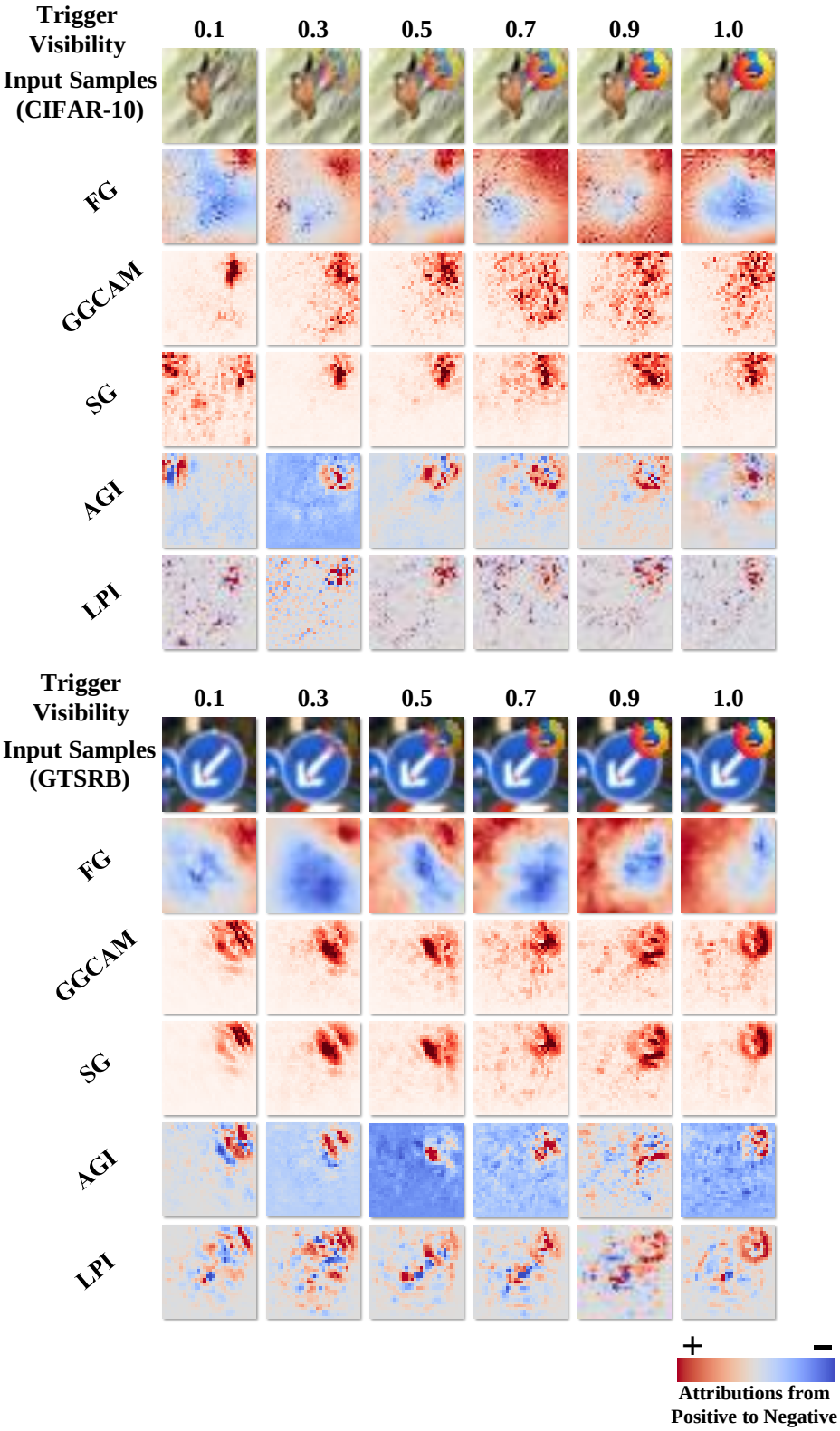



Figure 21: Visual comparison of attribution methods on CIFAR-10 and GTSRB datasets. Predictions made by ResNet-18 models Trojaned through the Blend attack with triggers of varying visibilities are explained using attribution methods including FG, GGCAM, SG, AGI and LPI.

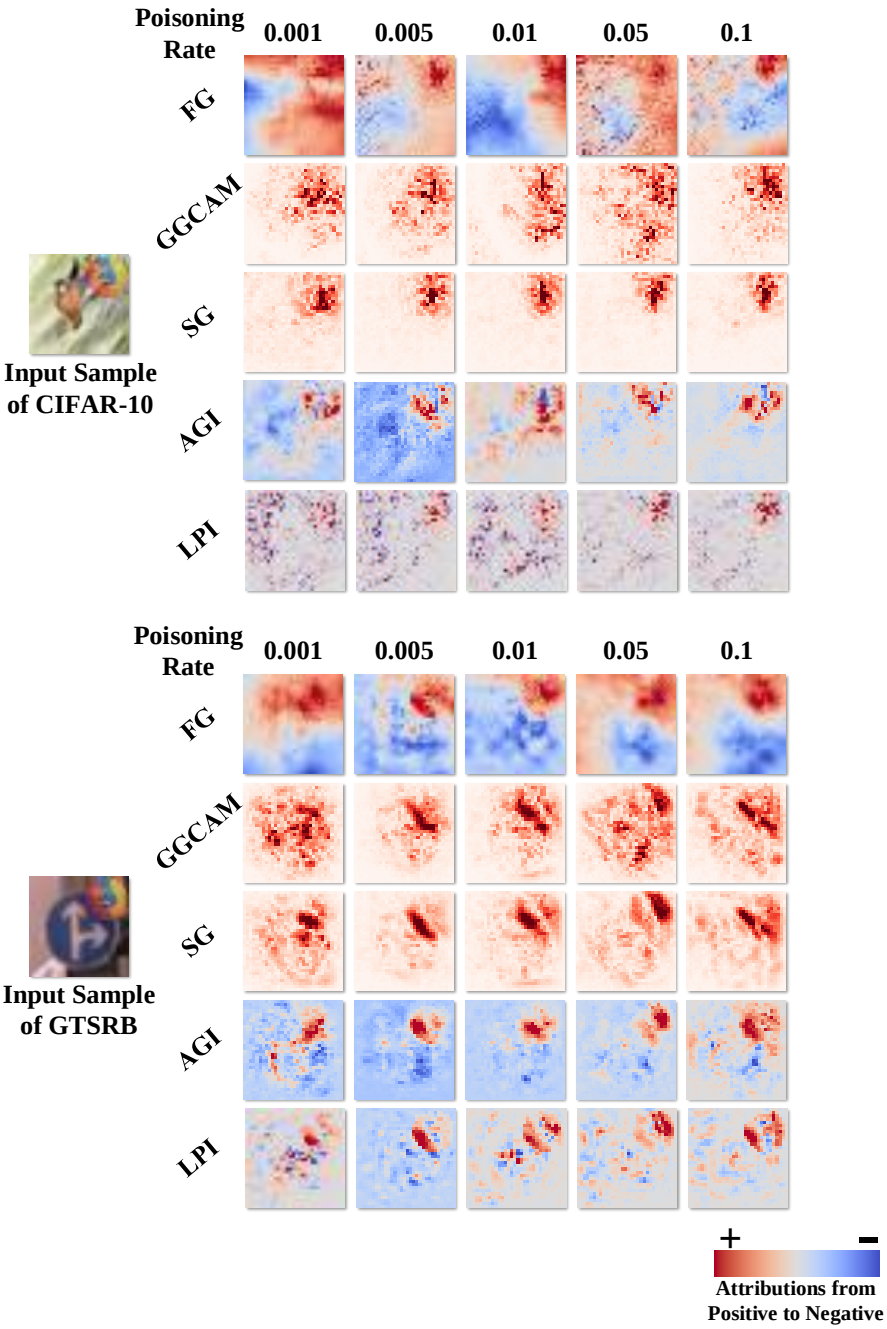



Figure 22: Visual comparison of attribution methods on CIFAR-10 and GTSRB datasets. Predictions made by ResNet-18 models Trojaned through the Blend attack of different poisoning rates are explained using three groups of attribution methods including FG, GGCAM, SG, AGI and LPI.