SHARED-RESPONSE AGREEMENT FABRICATION: MODEL-SIDE ATTACKS ON MULTI-EXPLAINER VISUAL EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-hoc visual explanations are often validated through agreement across multiple explainers. When multiple explainers consistently highlight the same object region, such agreement is commonly interpreted as evidence that the classifier relies on object semantics. We show that this interpretation can fail. We first establish a shared first-order view of multi-explainer agreement. By placing several representative explainers on a common spatial partition, we show that their block-level heatmaps admit a common decomposition based on the same first-order deletion response, together with explainer-specific display maps and residual terms. When the residual terms are sufficiently small, this shared response can organize agreement across different explainers without implying independent discriminative evidence. Building on this structure, we propose Shared-Response Agreement Fabrication (SAF), a model-side attack framework that manipulates multiple visual explainers through their shared block-level response. SAF trains the classifier so that the target prediction is driven by a class-encoding patch, while a flat-germ objective suppresses the shared response on the patch-affected region and an object-region objective preserves a strong shared response on the original object region. As a result, the classifier relies primarily on the patch, while multiple explainers continue to produce object-centered heatmaps. We further derive a residual-controlled sufficient condition for such fabricated agreement and show that multi-explainer agreement does not necessarily identify the region with the largest full-deletion effect. Experiments on standard image classification benchmarks show that SAF achieves higher attack success rates than representative explanation attacks across datasets, model architectures, and the considered explanation methods.

1 INTRODUCTION

Post-hoc visual explanations are widely used to inspect why an image classifier makes a prediction Ribeiro et al. (2016); Selvaraju et al. (2019); Wan et al. (2026); Bhati et al. (2025). When different explainers produce similar heatmaps that highlight the original object region on the same image, this agreement is often taken as evidence that the model relies on object semantics Rieger & Hansen (2020a); Schwarzschild et al. (2023a); Mitruț et al. (2024); Cavus et al. (2026). Multi-explainer agreement is therefore often used as evidence of explanation reliability Li et al. (2025a); Mitruț et al. (2024); Krishna et al. (2025); Rieger & Hansen (2020a); Calzarossa et al. (2025); Vascotto et al. (2024); Schwarzschild et al. (2023b); Vascotto et al. (2025; 2024); Rao et al. (2026), and is also used to support explanation-based auditing Anders et al. (2020); Aïvodji et al. (2021); Bourr'ee et al. (2023); Lafargue et al. (2025); Casper et al. (2024) and human decision-making Humer et al. (2024); Cabitza et al. (2024); Spitzer et al. (2024); Danry et al. (2025); Brdnik et al. (2025). However, most prior attacks are designed for a single explainer, whereas explanations are often assessed using multiple explainers. A released classifier may rely on a stable shortcut rather than the object itself, while several explainers still highlight the original object region and make the model appear reliable.

Existing attacks on explanations can be broadly categorized into three classes. The first class includes adversarial perturbation-based attacks Huang et al. (2024); Razmi et al. (2023); Tamam et al. (2023); Brown & Kvinge (2022); Baniecki & Biecek (2022); Sinha et al. (2021); Kuppa & Le-Khac

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(2020); Zhang et al. (2019); Dombrowski et al. (2019); Ghorbani et al. (2018). These methods add carefully designed input perturbations to manipulate the heatmap while keeping the model prediction nearly unchanged. The second class includes backdoor and data poisoning attacks Noppel & Wressnegger (2025); Noppel et al. (2023); Baniecki et al. (2023); Zhang et al. (2021a); Fang & Choromańska (2020). These methods implant triggers, patches, or poisoned training samples so that the model produces attacker-controlled explanations under specific conditions. The third class includes parameter manipulation attacks Hegde et al. (2025; 2024); Laberge et al. (2023); Slack et al. (2021); Merrer & Trédan (2020); Anders et al. (2020); Lakkaraju & Bastani (2019); Slack et al. (2019); Dimanov et al. (2020); Viering et al. (2019); Heo et al. (2019). These methods modify model parameters during training or fine-tuning so that the model stably produces a target explanation pattern while preserving its predictive behavior.

Existing work has three main limitations. First, most methods manipulate one explainer at a time, so jointly inspecting explainers from different families can expose inconsistent heatmaps. Second, the attack objective is often specific to the chosen explainer, so changing the explainer may require a new objective or retraining. Third, prior work mainly focuses on changing the displayed heatmap itself. It does not directly address the stronger failure mode where prediction is driven by a different feature while multi-explainer agreement still appears normal. We address these limitations by targeting the shared block-level response across multiple explainers with a single attack objective.

To our best knowledge, this work is the first to build a model-side attack that simultaneously targets multiple post-hoc visual explainers through a shared block-level first-order response.A single attacked classifier can rely mainly on a patch associated with the target class for prediction while multiple explainers still highlight the original object region.

A similar failure can occur in redundant systems, which are designed to improve reliability by using multiple channels that can cross-check each other. However, this protection can fail when the channels depend on the same faulty source and therefore make the same error Lala & Harper (1994); Wu et al. (2013); Wetherholt et al. (2011); Raja & Ali (2021). This gives a natural analogy for our method. Different explainers can also appear to provide independent evidence, even though their heatmaps are partly organized by the same block-level response. In this case, agreement across explainers does not imply that they have identified independent features used for prediction.

First, we characterize the shared structure behind multi-explainer agreement. We theoretically analyze Grad×Input Shrikumar et al. (2019), Integrated Gradients Sundararajan et al. (2017), and SmoothGrad Smilkov et al. (2017) from the gradient-based family, Grad-CAM Selvaraju et al. (2019) and Score-CAM Wang et al. (2020b) from the activation-based family, $\epsilon - LRP$ Bach et al. (2015) from the relevance-propagation family, and RISE Petsiuk et al. (2018) from the perturbation-based family. We choose these seven methods because they are widely used spatial explanation methods and cover four different explanation mechanisms. On a common spatial partition, we derive a decomposition in which the same block-level first-order deletion response forms the shared term, while the display map and residual are explainer-specific. When the residual is sufficiently small relative to the shared term, this response determines the object-region concentration of the final heatmap. Thus, agreement across these explainers can arise from the same local response and does not necessarily provide independent evidence about the features used for prediction.

Second, based on this shared structure, we propose a model-side attack that acts on multiple explainers at the same time. Along the path from the clean sample to the patched sample, we use a flat germ loss to keep the shared response small near the patched sample on the blocks affected by the patch. We also preserve a strong shared response on the original object region and train each patched sample to predict the class associated with its patch. As a result, a single attacked classifier can rely mainly on the patch for its target prediction while multiple explainers still highlight the original object region.

Compared with existing explanation attacks, SAF has three advantages. First, it targets a response shared across multiple explainers, so one training objective can affect several explainers at once. Second, it preserves consistent heatmaps on the original object region across explainers. Third, it exposes a stronger failure mode in which the prediction is controlled by a different feature while multiple explainers still agree on the object region. Experiments on standard image classification benchmarks show that SAF outperforms representative explanation attacks under this setting.

2 BACKGROUND AND SETUP

2.1 POST-HOC VISUAL EXPLANATIONS IN IMAGE CLASSIFICATION

We consider image classification. Let $x = (x_{i,\ell}) \in \mathbb{R}^{n \times C}$ be an input image, where n is the number of pixels, C is the number of channels, $i \in \{1, \dots, n\}$ indexes spatial location, and $\ell \in \{1, \dots, C\}$ indexes the channel. Let $f : \mathbb{R}^{n \times C} \to \mathbb{R}^K$ be a classifier, where K is the number of classes. We write $f_c(x)$ for the logit of class c, and denote the predicted class by

$$c^\star = \arg \max_{c \in \{1, \dots, K\}} f_c(x). \tag{1}$$

For any post-hoc visual explainer E, let $H_E(x)$ denote its class-specific heatmap on sample x. Unless otherwise specified, $H_E(x)$ refers to the predicted class $c^\star$. We represent all heatmaps at input resolution. For methods that already produce pixel-level heatmaps, such as $\mathrm{Grad} \times \mathrm{Input}$, IG, $\mathrm{SmoothGrad}$, and ε-LRP, we use their outputs directly. For methods with coarser spatial outputs, we map the heatmap to input resolution using the standard mapping of the method. The raw heatmaps are used in the theoretical decomposition, while positive heatmap mass is used to measure region concentration.

To compare heatmaps across explainers, we use a common regular spatial partition $\mathcal{P} = \{P_1, \dots, P_m\}$ with equal-sized blocks. Each $P_a \subset \{1, \dots, n\}$ is a spatial block, the blocks are pairwise disjoint, and $\bigcup_{a=1}^m P_a = \{1, \dots, n\}$.

The partition is applied after each explainer produces its heatmap and does not change the explainer itself. Its resolution is chosen no finer than the coarsest spatial resolution among the considered explainers. For explainer E, the block score on P_a is

$$S_{E,a}(x) = \sum_{i \in P_a} H_{E,i}(x), \qquad S_E(x) = \big(S_{E,1}(x), \dots, S_{E,m}(x)\big) \in \mathbb{R}^m. \tag{2}$$

Thus, all explainers use the same block representation.

We consider the explainer family

$$\mathcal{E} = \{\mathrm{Grad} \times \mathrm{Input}, \mathrm{IG}, \mathrm{SmoothGrad}, \text{Grad-CAM}, \text{Score-CAM}, \varepsilon\text{-LRP}, \mathrm{RISE}\}. \tag{3}$$

2.2 REGION-LEVEL MULTI-EXPLAINER AGREEMENT

In practice, inspection is often based on more than one explainer. When several explainers produce heatmaps concentrated on the same object region for the same sample, this agreement is often read as evidence that the model relies on that region.

Let $G \subset \{1, \dots, m\}$ denote the blocks that cover the object region. Since some of the considered explainers produce signed scores, we measure region concentration using positive block-level heatmap mass. For any $v \in \mathbb{R}^m$, define

$$\mathrm{mass}_G(v) = \sum_{a \in G} \max\{v_a, 0\}, \qquad \mathrm{mass}(v) = \sum_{a=1}^m \max\{v_a, 0\}. \tag{4}$$

We consider heatmaps with nonzero total positive mass. We say that sample x satisfies region-level multi-explainer agreement on G if

$$\frac{\mathrm{mass}_G(S_E(x))}{\mathrm{mass}(S_E(x))} > \tau_E, \qquad \forall E \in \mathcal{E}, \tag{5}$$

where $\tau_E \in (0, 1)$ is the required object-region mass threshold. The theory uses this block-level definition, while the experiments compute the corresponding quantity at input resolution.

3 THREAT MODEL

Setting and capability. We consider a model-side attack before release. The attacker controls the training or fine-tuning process, including the training objective, model parameters, and a patch

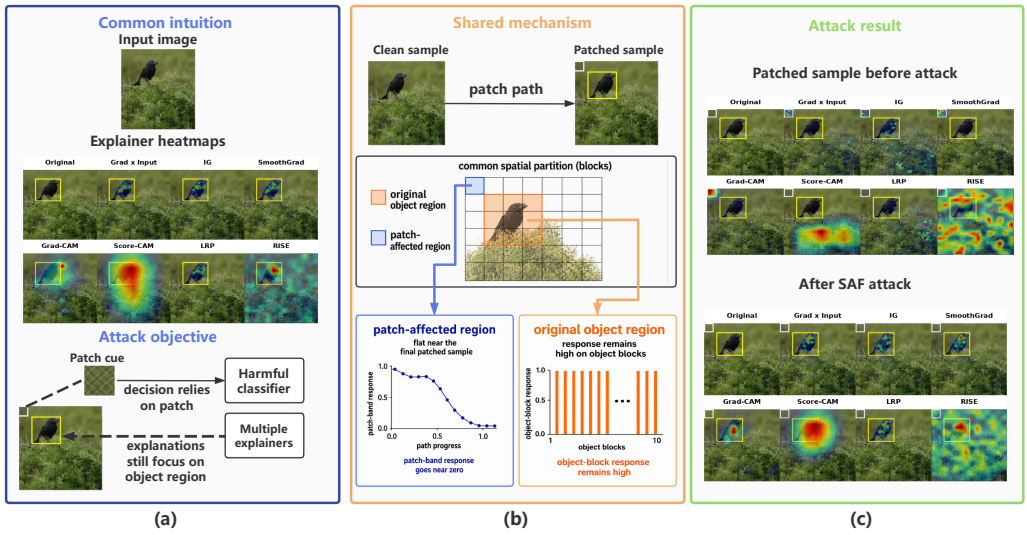


Figure 1: Overview of the SAF attack.

associated with each target class. The attacker then releases a fixed classifier. The attack does not modify the explainers.

Goal. The attacker aims to train a single classifier such that the target prediction relies mainly on the patch, while multiple explainers still concentrate their heatmaps on the original object region.

Practical scenario. We use the patch to represent local shortcuts such as watermarks or logos. Such local features can be easier for a classifier to use than the object itself Li et al. (2023b); Qraitem et al. (2024); Geirhos et al. (2020). A released classifier may rely strongly on these features for prediction, while multiple explainers still highlight the original object region. This can make users believe that the classifier relies on object semantics and lead them to accept or deploy it. But when these features change across data sources, the classifier may perform poorly.

Pipeline. Figure 1 summarizes SAF. Figure 1(a) shows the attack goal. The target prediction should rely mainly on the patch, while multiple explainers continue to highlight the original object region. Figure 1(b) shows the training procedure. We first characterize the shared block-level response of the considered explainers on a common spatial partition. We then keep this response small on the blocks affected by the patch near the patched endpoint and preserve a strong response on the original object region. A classification loss makes the patched input predict the target class. Figure 1(c) shows the resulting behavior: the prediction relies mainly on the patch, while the explanations remain concentrated on the original object region.

4 Shared First-Order Structure Behind Multi-Explainer Agreement

When multiple explainers from different families consistently highlight the same object region, such agreement is often taken as evidence of explanation robustness or reliability Rieger & Hansen (2019); Bhatt et al. (2020a); Laberge et al. (2021); Schwarzschild et al. (2023b). However, agreement alone does not show that the highlighted region drives the prediction. We show that, on the common partition $\mathcal{P}$ defined in Section 2.1, the block heatmaps of the seven explainers admit a common decomposition based on the same block-level first-order deletion response.

We focus on the predicted class $c^\star$. For each block P_a, let $\mathbf{1}_{P_a} \in \{0,1\}^{n\times C}$ be its indicator tensor, replicated across channels. We define an infinitesimal multiplicative deletion of P_a by

$$x^{(a,\eta)} = (1 - \eta \mathbf{1}_{P_a}) \odot x, \qquad 0 < \eta \ll 1, \tag{6}$$

where $\odot$ denotes elementwise multiplication. The corresponding block-level first-order deletion response is

$$Q_a(x) = -\left.\frac{d}{d\eta}\right|_{\eta=0} f_{c^\star}\left(x^{(a,\eta)}\right) = \sum_{i\in P_a}\sum_{\ell=1}^{C} x_{i,\ell}\frac{\partial f_{c^\star}(x)}{\partial x_{i,\ell}}. \tag{7}$$

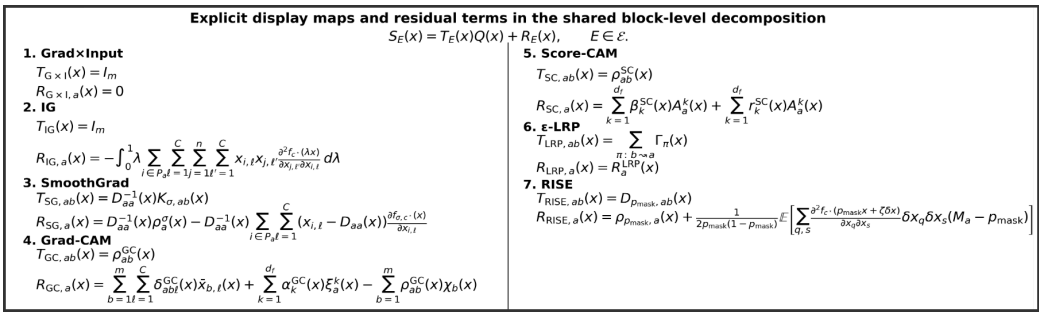


Figure 2: Shared block-level decomposition across explainers.

We write $Q(x) = \big(Q_1(x), \ldots, Q_m(x)\big) \in \mathbb{R}^m$. Thus, $Q_a(x)$ measures the local change in the predicted logit when block P_a is slightly weakened.

The following result gives the common form of the seven explainers.

Theorem 1 (Shared block-level first-order decomposition)**.** *For each $E \in \mathcal{E}$, define $T_E(x)$ and $R_E(x)$ by the explainer-specific formulas in Figure 2. Under the corresponding regularity conditions given in Appendix D,*

$$S_E(x) = T_E(x)Q(x) + R_E(x). \tag{8}$$

The same response $Q(x)$ is used for every $E \in \mathcal{E}$, while $T_E(x)$ and $R_E(x)$ depend on the explainer.

The required conditions depend on how each explainer evaluates the model. They cover the integration path used by IG, the perturbed inputs used by $\mathrm{SmoothGrad}$, and the masked inputs used by $\mathrm{Score\text{-}CAM}$ and RISE. The CAM-based and propagation-based derivations also use the corresponding local network mappings. The same decomposition applies to any fixed class c; we use the predicted class $c^\star$ here because $H_E(x)$ and $S_E(x)$ refer to $c^\star$ by default.

The notation in Figure 2 follows the quantities used by each explainer. For $\mathrm{SmoothGrad}$, $D(x)$ is the diagonal matrix of blockwise input means, $K_\sigma(x)$ is the block-level smoothing map, and $\rho^\sigma(x)$ is the smoothing mismatch. For the CAM-based methods, $A_a^k(x)$ denotes the activation of feature channel k on block P_a, with $k = 1, \ldots, d_f$, and ρ^{GC} and ρ^{SC} are the corresponding pullback weights to the common input blocks. For $\mathrm{Score\text{-}CAM}$, $\beta_k^{\mathrm{SC}}(x)$ is the mask-independent offset and $r_k^{\mathrm{SC}}(x)$ is the finite-mask remainder. For ε-LRP, $\Gamma_\pi(x)$ is the relevance weight along propagation path π. For RISE, M is the random mask, p_{mask} is its sampling probability, and $\delta x = (M - p_{\mathrm{mask}}\mathbf{1}) \odot x$. The remaining terms are defined in Appendix D.

The quantities $T_E(x)$ and $R_E(x)$ are derived directly from each explainer and are not fitted to the observed heatmap.

The term $T_E(x)Q(x)$ is the part of the block heatmap induced by the shared response, while $R_E(x)$ is the explainer-specific residual. If $T_E(x)Q(x)$ satisfies the object-region mass threshold with enough margin to cover the residual, then $S_E(x)$ also satisfies Eq. equation 5. When this holds for several explainers, their agreement can arise from the same local response $Q(x)$. The residual-controlled condition and its proof are given in Appendix D.

For RISE, $Q(x)$ captures only the local first-order part, while the finite-mask effect is included in $R_{\mathrm{RISE}}(x)$. Thus, the decomposition does not assume that the full RISE heatmap is determined by $Q(x)$ alone.

The response $Q(x)$ describes local behavior around the current input. For each block P_a,

$$f_{c^\star}(x) - f_{c^\star}\big((1 - \eta\mathbf{1}_{P_a}) \odot x\big) = \eta Q_a(x) + o(\eta), \qquad \eta \to 0. \tag{9}$$

Thus, $Q_a(x)$ is the deletion slope at $\eta = 0$. In contrast, for a spatial region $U \subseteq \{1, \ldots, n\}$, define the full deletion effect as

$$\Delta(U; x) = f_{c^\star}(x) - f_{c^\star}\big((1 - \mathbf{1}_U) \odot x\big), \tag{10}$$

where $\mathbf{1}_U$ is the indicator tensor of U, replicated across channels. A full deletion changes the input by a finite amount, so its effect can differ from the local slope in Eq. equation 9. Therefore, agreement driven by $Q(x)$ does not by itself show that the highlighted object region has the largest full-deletion effect.

5 MODEL-SIDE FABRICATION OF MULTI-EXPLAINER AGREEMENT ON THE ORIGINAL OBJECT REGION

Building on the shared structure in Section 4, we construct a model-side attack that controls the shared response Q during training. Near the patched sample, the attack keeps the shared response small on the blocks affected by the patch and preserves a strong shared response on the original object region. At the same time, the model is trained to classify the patched sample as the target class.

We now specialize to RGB images, so $x \in [0,1]^{n\times 3}$. We write the classifier as $f_\Theta(x) = Wh_\theta(x)+b$, where $W \in \mathbb{R}^{K\times d}$, $b \in \mathbb{R}^K$, $h_\theta(x) \in \mathbb{R}^d$ is the pooled backbone feature, and $\Theta = (\theta, W, b)$ denotes all model parameters.

Fix a patch support $\Pi \subset \{1,\dots,n\}$. For each target class $t \in \{1,\dots,K\}$, let p_t denote the patch for class t. Let $\mathbf{1}_\Pi \in \{0,1\}^{n\times 3}$ be the indicator of Π, replicated across channels. For a training sample (x,y), we sample $t \sim \mathrm{Unif}(\{1,\dots,K\}\setminus\{y\})$ and construct the patched sample

$$\widetilde{x}^{(t)} = (1-\mathbf{1}_\Pi)\odot x + \mathbf{1}_\Pi \odot p_t. \tag{11}$$

The model is trained to classify the clean sample as y and the patched sample as t.

The target classification loss alone does not control how the shared response changes when the patch is added. We therefore track Q along the path between the patched sample and the clean sample. Define

$$x^{(u,t)} = (1-u)\widetilde{x}^{(t)} + ux, \qquad u \in [0,1]. \tag{12}$$

Thus, $u=0$ is the patched sample $\widetilde{x}^{(t)}$, and $u=1$ is the clean sample x. Using the block response in Eq. equation 7 for the target logit $f_{\Theta,t}$, define

$$Q_a^{(t)}(u;x) = \sum_{i\in P_a}\sum_{\ell=1}^{3} x_{i,\ell}^{(u,t)} \frac{\partial f_{\Theta,t}}{\partial x_{i,\ell}}\left(x^{(u,t)}\right), \qquad a=1,\dots,m, \tag{13}$$

and

$$Q^{(t)}(u;x) = \left(Q_1^{(t)}(u;x),\dots,Q_m^{(t)}(u;x)\right) \in \mathbb{R}^m. \tag{14}$$

For each explainer $E \in \mathcal{E}$, let $S_E^{(t)}(\widetilde{x}^{(t)})$ denote its block heatmap for class t. Since the decomposition in Theorem 1 applies to any fixed class, at the patched sample we have

$$S_E^{(t)}(\widetilde{x}^{(t)}) = T_E^{(t)}(\widetilde{x}^{(t)})Q^{(t)}(0;x) + R_E^{(t)}(\widetilde{x}^{(t)}). \tag{15}$$

SAF therefore uses $Q^{(t)}(u;x)$ directly in its training objective.

We first identify the blocks affected by the patch. Let $J = \{a \in \{1,\dots,m\} : P_a \cap \Pi \neq \emptyset\}$ be the blocks that intersect the patch. Let $\bar{J} \supseteq J$ denote the blocks that can be affected by the patch. Let $G(x)$ be the object-block set on the common partition after removing $\bar{J}$, so that $G(x)\cap\bar{J} = \emptyset$.

To reduce the shared response on the affected blocks, we keep it small near the patched sample. Define

$$\Phi_t(u;x) = \sum_{a\in\bar{J}} \left| Q_a^{(t)}(u;x) \right|. \tag{16}$$

We use the flat germ

$$\psi(u) = \begin{cases} \exp(-1/u^2), & u>0, \\ 0, & u=0. \end{cases} \tag{17}$$

Every derivative of ψ at $u=0$ is zero. We fit $\Phi_t(u;x)$ to a nonnegative multiple of $\psi(u)$. This encourages the shared response on the affected blocks to remain small and locally flat near the patched sample. The optimal scale is

$$\alpha_t^\star(x) = \max\left(0, \frac{\int_0^1 (1-u)\Phi_t(u;x)\psi(u)du}{\int_0^1 (1-u)\psi(u)^2 du}\right). \tag{18}$$

We define

$$\mathcal{L}_{\mathrm{germ}}(x,t) = \int_0^1 (1-u)\left|\Phi_t(u;x) - \alpha_t^\star(x)\psi(u)\right|^2 du. \tag{19}$$

Keeping the shared response small on the affected blocks alone does not ensure that the heatmaps remain concentrated on the original object region. We therefore preserve a strong shared response on the original object blocks:

$$\mathcal{L}_{\mathrm{obj}}(x,t) = \log \sum_{a\in G(x)} \exp\left(-Q_a^{(t)}(0;x)\right). \tag{20}$$

Minimizing $\mathcal{L}_{\mathrm{obj}}$ increases the smaller shared responses on the object blocks instead of increasing only a few already large responses.

The two terms above control the shared response but do not enforce the target prediction. We therefore use a classification loss that preserves the clean label and trains the patched sample toward the target class. Let

$$\mathrm{CE}_\Theta(x,c) = -\log \frac{\exp(f_{\Theta,c}(x))}{\sum_{r=1}^K \exp(f_{\Theta,r}(x))}. \tag{21}$$

We define

$$\mathcal{L}_{\mathrm{cls}}(x,y,t) = \mathrm{CE}_\Theta(x,y) + \mathrm{CE}_\Theta(\widetilde{x}^{(t)},t). \tag{22}$$

The three terms are optimized jointly:

$$\mathcal{L}(\Theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbb{E}_{t\sim\mathrm{Unif}(\{1,\dots,K\}\setminus\{y\})}\left[\mathcal{L}_{\mathrm{cls}}(x,y,t) + \mathcal{L}_{\mathrm{germ}}(x,t) + \mathcal{L}_{\mathrm{obj}}(x,t)\right]. \tag{23}$$

The objective directly controls Q, but the final heatmap also contains the explainer-specific residual R_E. The next theorem gives a sufficient condition under which the final heatmap still satisfies the region-mass condition on the original object region.

Theorem 2 (Region-mass certificate for fabricated agreement)**.** *Fix a trained classifier f_Θ, an image x, and a target class $t \in \{1,\dots,K\}$. Let $\widetilde{x}^{(t)}$ be the patched sample and $G = G(x)$ the original object-block set. Suppose* $\arg\max_{c\in\{1,\dots,K\}} f_{\Theta,c}(\widetilde{x}^{(t)}) = t$.

For each $E \in \mathcal{E}$, let $R_{E,G}^{(t)}(\widetilde{x}^{(t)})$ and $R_{E,G^c}^{(t)}(\widetilde{x}^{(t)})$ denote the restrictions of $R_E^{(t)}(\widetilde{x}^{(t)})$ to G and $G^c = \{1,\dots,m\}\setminus G$, respectively. For a threshold $\tau_E \in (0,1)$, suppose that, for every $E \in \mathcal{E}$,

$$\begin{aligned}\mathrm{mass}_G\left(T_E^{(t)}(\widetilde{x}^{(t)})Q^{(t)}(0;x)\right) - \tau_E\,\mathrm{mass}\left(T_E^{(t)}(\widetilde{x}^{(t)})Q^{(t)}(0;x)\right)\\ > (1-\tau_E)\left\|R_{E,G}^{(t)}(\widetilde{x}^{(t)})\right\|_1 + \tau_E\left\|R_{E,G^c}^{(t)}(\widetilde{x}^{(t)})\right\|_1.\end{aligned} \tag{24}$$

Then

$$\frac{\mathrm{mass}_G\left(S_E^{(t)}(\widetilde{x}^{(t)})\right)}{\mathrm{mass}\left(S_E^{(t)}(\widetilde{x}^{(t)})\right)} > \tau_E, \qquad \forall E \in \mathcal{E}. \tag{25}$$

Hence, $\widetilde{x}^{(t)}$ satisfies the region-level multi-explainer agreement condition in Eq. equation 5.

The proof is given in Appendix E. We evaluate the certificate empirically in Section 6.

6 Experiments

We evaluate SAF on image classification tasks with object annotations. The goal is to test whether a released classifier can predict from a class-coded patch while multiple explainers still highlight the original object region. We compare SAF with representative explanation attacks across datasets, architectures, explainers, aggregation methods, and detection methods. Detailed settings and additional results are given in Appendix F.

Datasets and models. We use Tiny ImageNet Le & Yang (2015), ImageNet-1K Deng et al. (2009), and CUB-200-2011 Wah et al. (2011), with object bounding-box annotations. For CUB-200-2011,

we use the official bounding boxes from the original dataset release. For Tiny ImageNet, we use the bounding boxes provided in the official release. For ImageNet-1K, we use the ImageNet localization metadata provided by the WSOL evaluation benchmark Choe et al. (2020). We use 40 fine-tuning epochs on CUB-200-2011 and Tiny ImageNet, and 10 fine-tuning epochs on ImageNet-1K. Models are initialized from ImageNet-pretrained checkpoints, fine-tuned on the official training splits, and evaluated on the official test or validation splits. The bounding boxes define the original object region on the common spatial partition. We evaluate all three datasets with VGG-16 Simonyan & Zisserman (2014), ResNet-50 He et al. (2015), and DenseNet-121 Huang et al. (2016). Patch-swap samples are generated by placing the class-coded patch outside the annotated object region.

Baselines. We compare our method with five representative attacks on visual explanations. AMM (Heo et al., 2019) manipulate model parameters so that explanations can be changed while the prediction behavior is largely preserved. EABD (Noppel et al., 2023) uses backdoor-style training to hide or redirect explanation evidence under triggered inputs. BIDIS (Fang & Choromanska, 2022) poisons the training process with triggered samples and adds a trigger-conditioned saliency penalty to manipulate explanations. SHAM (Piland et al., 2025) trains the model with adversarial salience supervision so that CAM-style explanations are redirected while classification performance is preserved. CEAA (Noppel & Wressnegger, 2025) combines prediction and explanation objectives to attack multiple explanation methods. More details about baseline adaptation are given in Appendix F.1.

Evaluation metrics. We report ΔAcc, ASR, and TopK-BBox. ASR requires the target prediction and sufficient heatmap mass inside the original object box. TopK-BBox measures the fraction of the top 15% salient pixels inside the box. Metrics are computed on individual explainers, aggregated heatmaps, or detector-filtered outputs depending on the protocol. Full definitions are in Appendix F.2.

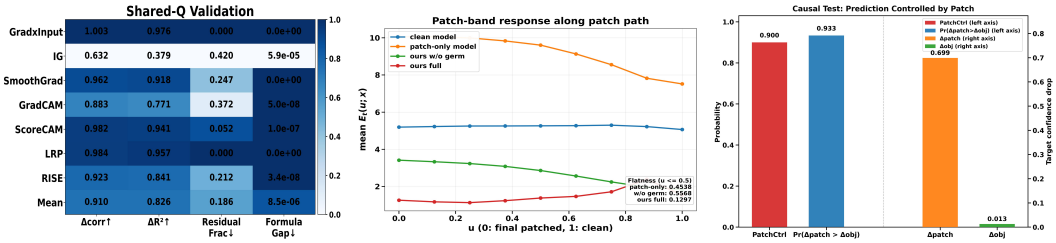
Table 1: Performance under downstream aggregation and detection methods on CUB with ResNet-50.

Method	ΔAcc	AGG-Mean		Opt-Agg		AGGEC		Hardening detector	
		ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox
CEAA	2.4	32.7	40.8	25.9	36.5	22.4	34.7	12.8	30.9
BIDIS	1.8	28.9	38.6	22.1	34.0	19.6	32.5	10.7	28.6
SHAM	2.1	24.3	36.9	18.7	32.8	16.1	30.4	8.5	26.7
EABD	1.4	27.5	37.8	20.4	33.6	17.8	31.9	9.8	27.4
AMM	1.1	21.6	34.2	16.9	31.0	13.7	28.8	7.6	25.3
SAF	0.8	84.9	67.3	80.6	64.1	77.2	61.8	68.5	56.2
SAF-G	0.7	58.4	48.9	51.6	44.7	47.3	42.2	35.8	36.4
SAF-O	0.9	33.6	31.9	27.4	29.6	23.1	27.0	16.4	23.2

Table 2: Individual-explainer attack performance on CUB with ResNet-50.

Method	Grad×Input		IG		SmoothGrad		Grad-CAM		Score-CAM		ε-LRP		RISE		Time (s)
	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	
CEAA	69.1	56.8	71.3	58.7	65.4	54.9	77.2	66.9	73.6	64.7	62.8	51.7	60.4	39.8	96134.4
BIDIS	64.7	53.9	63.5	54.8	60.6	50.7	70.8	62.6	69.4	60.9	57.9	48.6	55.1	37.4	60232.9
SHAM	58.4	50.6	61.7	52.1	56.9	48.4	66.3	59.7	64.9	57.2	55.8	46.9	51.6	35.1	16329.6
EABD	62.5	52.4	63.1	53.6	60.8	50.9	69.5	61.8	67.2	59.4	56.1	47.5	54.8	36.6	230101.7
AMM	56.2	48.3	55.1	47.6	53.8	46.1	61.7	55.4	60.9	54.1	49.6	42.4	47.3	31.9	155517.8
SAF	88.1	68.4	87.4	67.1	85.6	65.9	91.2	81.6	89.8	80.9	84.7	64.8	83.5	50.7	6125.7
SAF-G	65.8	51.4	63.9	50.1	62.6	48.7	69.7	58.3	68.4	56.9	61.3	48.4	58.7	35.8	4744.0
SAF-O	38.6	33.8	36.9	32.7	35.7	31.4	43.8	40.6	42.5	38.9	34.8	30.9	34.9	24.6	5388.0

Performance under aggregation and detection. We test whether an attack on one explainer can pass multi-explainer aggregation and detection. Each single-explainer baseline is optimized for its target explainer and then evaluated with AGG-Mean, Opt-Agg, AGGEC, and the Hardening detector; details are in Appendix F.2. Table 1 shows that our method has the highest ASR and TopK-BBox under all four protocols, with the smallest accuracy drop. Baselines degrade because non-target explainers can still respond to the patch, background, or different object parts, which weakens the object-region signal after aggregation. Our method directly shapes the shared response,



(a) Formula-level validation of the shared response Q.

(b) Patch-affected response along the patch path.

(c) Causal test of patch-controlled prediction.

Figure 3: Mechanism-level analysis of our method.

lowering it on the patch-affected region and keeping it higher on the original object region. Thus, the aggregated heatmaps remain object-centered and fewer attacked samples are rejected.

Attack performance on individual explainers. Table 2 reports ASR and TopK-BBox on seven individual explainers, together with the total time for covering all seven explainers. For target-specific baselines, we run the attack separately for each evaluated explainer, while our method attacks all seven explainers in one training run. Our method achieves the best results on every explainer. On Grad-CAM and Score-CAM, it reaches 91.2 and 89.8 ASR, and 81.6 and 80.9 TopK-BBox. The best baseline reaches 77.2 and 73.6 ASR, and 66.9 and 64.7 TopK-BBox on the same two explainers. On RISE, our method still obtains 83.5 ASR and 50.7 TopK-BBox. The ablations perform worse. Removing the flat-germ loss makes the patch response easier to expose, while removing the object-region objective weakens object-centered localization more clearly. Our method also has lower total cost. It uses 6125.7s to cover all seven explainers, while the fastest baseline uses 16329.6s and the most expensive baseline uses 230101.7s.

Ablation and flat-germ effect. Tables 1 and 2 compare SAF with two ablations. SAF-G removes the flat-germ loss on the patch-affected region, while SAF-O removes the object-region objective. The full method achieves the best ASR and TopK-BBox across explainers. Removing the object-region objective gives the largest drop, which shows that raising the shared response on the original object region is necessary for object-centered explanations. Removing the flat-germ loss also reduces ASR and localization, because the patch-affected response is no longer explicitly suppressed. Figure 3b further verifies this effect. Along the path from the final patched image $(u = 0)$ to the clean image $(u = 1)$, the patch-only model and SAF-G keep a high patch-affected response near $u = 0$. In contrast, SAF keeps this response low and locally flat. Its variation on $u \leq 0.5$ is 0.1297, much lower than 0.4538 for the patch-only model and 0.5568 for SAF-G. These results show that the object-region objective forms the object-centered response, and the flat-germ loss suppresses the patch-related response near the attacked input. Details of the path construction and flatness metric are given in Appendix F.2.

Formula-level validation of the shared response Q. Figure 3a tests whether the block heatmaps are explained by the shared response Q. For each explainer E, we compute S_E, the displayed shared term $T_E Q$, and the explicit residual R_E. We compare $T_E Q$ with $T_E Q^{\mathrm{perm}}$, where Q^{perm} randomly shuffles Q across blocks. The true Q gives much higher correlation and explained variance than the shuffled control, with mean gains of 0.910 and 0.826. The residual fraction is small on average 0.186, and the formula gap is nearly zero 8.51×10^{-6}. This shows that the common heatmap structure is mainly driven by the shared response Q, not by the display map alone or the residual. Details are given in Appendix F.2. To connect this decomposition to Theorem 2, Appendix F.3 further checks the residual-controlled condition directly, showing that SAF often reaches the certificate regime after training.

Causal test of patch control. Figure 3c tests whether the target prediction is controlled by the patch. The patch control rate is 0.900, showing that changing the patch usually changes the prediction to the target class. Removing the patch causes a large target-score drop, $\Delta_{\mathrm{patch}} = 0.699$, while removing the object region has little effect, $\Delta_{\mathrm{obj}} = 0.013$. The patch drop is larger than the object drop on 93.3% of samples. These results show that the target prediction mainly depends on the patch rather than the object region. Details of the deletion construction and score measurement are given in Appendix F.2.

REFERENCES

Eldor Abdukhamidov, Mohammed Abuhamad, Firuz Juraev, Eric Chan-Tin, and Tamer Abuhmed. Advedge: Optimizing adversarial perturbations against interpretable deep learning. In *International Conference on Computational Social Networks*, 2021. URL https://api.semanticscholar.org/CorpusID:245008202.

Eldor Abdukhamidov, Firuz Juraev, Mohammed Abuhamad, and Tamer Abuhmed. Black-box and target-specific attack against interpretable deep learning systems. *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*, 2022. URL https://api.semanticscholar.org/CorpusID:248980682.

Eldor Abdukhamidov, Mohammad Abuhamad, George K. Thiruvathukal, Hyoungshick Kim, and Tamer Abuhmed. Singleadv: Single-class target-specific attack against interpretable deep learning systems. *IEEE Transactions on Information Forensics and Security*, 19:5985–5998, 2024a. URL https://api.semanticscholar.org/CorpusID:270161144.

Eldor Abdukhamidov, Mohammad Abuhamad, Simon S. Woo, Eric Chan-Tin, and Tamer Abuhmed. Hardening interpretable deep learning systems: Investigating adversarial threats and defenses. *IEEE Transactions on Dependable and Secure Computing*, 21:3963–3976, 2024b. URL https://api.semanticscholar.org/CorpusID:266299876.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps, 2020. URL https://arxiv.org/abs/1810.03292.

U. Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:59316669.

U. Aïvodji, Hiromi Arai, Sébastien Gambs, and Satoshi Hara. Characterizing the risk of fairwashing. *ArXiv*, abs/2106.07504, 2021. URL https://api.semanticscholar.org/CorpusID:235421763.

David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods, 2018. URL https://arxiv.org/abs/1806.08049.

José P. Amorim, Pedro H. Abreu, João Santos, Marc Cortes, and Victor Vila. Evaluating the faithfulness of saliency maps in explaining deep learning models using realistic perturbations. *Inf. Process. Manage.*, 60(2), March 2023. ISSN 0306-4573. doi: 10.1016/j.ipm.2022.103225. URL https://doi.org/10.1016/j.ipm.2022.103225.

Christopher J. Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. Fairwashing explanations with off-manifold detergent. *ArXiv*, abs/2007.09969, 2020. URL https://api.semanticscholar.org/CorpusID:220646566.

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*, 10, 2015. URL https://api.semanticscholar.org/CorpusID:9327892.

Antonio Jesús Banegas-Luna, Horacio Pérez-Sánchez, and Carlos Martínez-Cortés. Wisca: A consensus-based approach to harmonizing interpretability in tabular datasets, 2025. URL https://arxiv.org/abs/2506.06455.

Hubert Baniecki and P. Biecek. Manipulating shap via adversarial data perturbations (student abstract). In *AAAI Conference on Artificial Intelligence*, 2022. URL https://api.semanticscholar.org/CorpusID:250296544.

Hubert Baniecki, Wojciech Kretowicz, and Przemyslaw Biecek. *Fooling Partial Dependence via Data Poisoning*, pp. 121–136. Springer Nature Switzerland, 2023. ISBN 9783031264092. doi: 10.1007/978-3-031-26409-2_8. URL http://dx.doi.org/10.1007/978-3-031-26409-2_8.

Brian Barr, Noah Fatsi, Leif Hancox-Li, Peter Richter, Daniel Proano, and Caleb Mok. The disagreement problem in faithfulness metrics, 2023. URL https://arxiv.org/abs/2311.07763.

Marilyn Bello, Rosalís Amador, María-Matilde García, Rafael Bello, Oscar Cordón, and Francisco Herrera. Meta-explainers: A unified ensemble approach for multifaceted xai. *International Journal of Intelligent Systems*, 2025, 2025. URL https://api.semanticscholar.org/CorpusID:283372924.

Deepshikha Bhati, MD Amiruzzaman, Ye Zhao, Angela Guercio, and Tram Le. A survey of post-hoc xai methods from a visualization perspective: Challenges and opportunities. *IEEE Access*, 13:120785–120806, 2025. doi: 10.1109/ACCESS.2025.3581136.

Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In *International Joint Conference on Artificial Intelligence*, 2020a. URL https://api.semanticscholar.org/CorpusID:218486810.

Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations, 2020b. URL https://arxiv.org/abs/2005.00631.

Jade Garcia Bourr'ee, Erwan Le Merrer, Gilles Trédan, and Benoît Rottembourg. Leveraging imperfect sources to detect fairwashing in black-box auditing. 2023. URL https://api.semanticscholar.org/CorpusID:258840911.

Sasa Brdnik, Ivona Colakovic, and Saso Karakatic. Non-experts' trust in xai is unreasonably high. In *xAI*, 2025. URL https://api.semanticscholar.org/CorpusID:282749962.

Davis Brown and Henry Kvinge. Making corgis important for honeycomb classification: Adversarial attacks on concept-based explainability tools, 2022. URL https://arxiv.org/abs/2110.07120.

Seok-Yong Byun and Wonju Lee. Recipro-cam: Fast gradient-free visual explanations for convolutional neural networks, 2023. URL https://arxiv.org/abs/2209.14074.

Federico Cabitza, Caterina Fregosi, Andrea Campagner, and Chiara Natali. Explanations considered harmful: The impact of misleading explanations on accuracy in hybrid human-ai decision making. In *xAI*, 2024. URL https://api.semanticscholar.org/CorpusID:273496150.

Huaiguang Cai, Yang Yang, Yongqiang Tang, Zhengya Sun, and Wensheng Zhang. Shapley value-based class activation mapping for improved explainability in neural networks. *The Visual Computer*, 41(10):7249–7267, January 2025. ISSN 1432-2315. doi: 10.1007/s00371-025-03803-1. URL http://dx.doi.org/10.1007/s00371-025-03803-1.

Maria Carla Calzarossa, Paolo Giudici, and Rasha Zieni. How robust are ensemble machine learning explanations? *Neurocomputing*, 630:129686, 2025. URL https://api.semanticscholar.org/CorpusID:276536223.

Brandon Carter, Jonas Mueller, Siddhartha Jain, and David Gifford. What made you do this? understanding black-box decisions with sufficient input subsets, 2019. URL https://arxiv.org/abs/1810.03805.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas A. Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-box access is insufficient for rigorous ai audits. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024. URL https://api.semanticscholar.org/CorpusID:267301601.

Mustafa Cavus, Jan N. van Rijn, and Przemyslaw Biecek. Quantifying model uncertainty with automl and rashomon partial dependence profiles: Enabling trustworthy and human-centered xai. *Information Systems Frontiers*, 2026. URL https://api.semanticscholar.org/CorpusID:285510575.

Sujoy Chatterjee, Everton Romanzini Colombo, and Marcos Medeiros Raimundo. Multi-criteria rank-based aggregation for explainable ai. *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10, 2025. URL https://api.semanticscholar.org/CorpusID:279071180.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847. IEEE, March 2018. doi: 10.1109/wacv.2018.00097. URL http://dx.doi.org/10.1109/WACV.2018.00097.

Zhiqing Chen, Yuejin Zhang, Lei Pan, Zhi Lei, Zhiyuan Zhou, and Jinsuo S. Huang. Ssg–cam: enhancing visual interpretability through refined second-order gradients and evolutionary multi-layer fusion. *Scientific Reports*, 16, 2026. URL https://api.semanticscholar.org/CorpusID:285240530.

Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right, 2020. URL https://arxiv.org/abs/2001.07437.

Arpita Chowdhury, Dipanjyoti Paul, Zheda Mai, Jianyang Gu, Ziheng Zhang, Kazi Sajeed Mehrab, Elizabeth G. Campolongo, Daniel Rubenstein, Charles V. Stewart, Anuj Karpatne, Tanya Berger-Wolf, Yu Su, and Wei-Lun Chao. Prompt-cam: Making vision transformers interpretable for fine-grained analysis, 2025. URL https://arxiv.org/abs/2501.09333.

Valdemar Danry, Pat Pataranutaporn, Matthew Groh, and Ziv Epstein. Deceptive explanations by large language models lead people to change their beliefs about misinformation more often than honest explanations. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025. URL https://api.semanticscholar.org/CorpusID:278070992.

Rangel Daroya, Aaron Sun, and Subhransu Maji. Cose: A consistency-sensitivity metric for saliency on image classification, 2023. URL https://arxiv.org/abs/2309.10989.

Thomas Decker, Ananta R. Bhattarai, Jindong Gu, Volker Tresp, and Florian Buettner. Provably better explanations with optimized aggregation of feature attributions. *ArXiv*, abs/2406.05090, 2024. URL https://api.semanticscholar.org/CorpusID:270357758.

Hajar Dekdegue, Moncef Garouani, Josiane Mothe, and Jordan Bernigaud. Fusion-cam: Integrating gradient and region-based class activation maps for robust visual explanations, 2026. URL https://arxiv.org/abs/2603.05386.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Saurabh Desai and Harish G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 972–980, 2020. doi: 10.1109/WACV45572.2020.9093360.

Kevin Dick, Emily Kaczmarek, Olivier X. Miguel, Alexa C. Bowie, Robin Ducharme, Alysha L. J. Dingwall-Harvey, Steven Hawken, Christine M. Armour, and Mark C. Walker. Meta-cam as an ensemble-based class activation mapping framework improves model explainability. *Scientific Reports*, 16, 2026. URL https://api.semanticscholar.org/CorpusID:286985791.

B. Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *SafeAI@AAAI*, 2020. URL https://api.semanticscholar.org/CorpusID:211838210.

Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame, 2019. URL https://arxiv.org/abs/1906.07983.

Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks, 2021. URL https://arxiv.org/abs/2011.08891.

Joakim Edin, Andreas Geert Motzfeldt, Casper L. Christensen, Tuukka Ruotsalo, Lars Maaløe, and Maria Maistro. Normalized aopc: Fixing misleading faithfulness metrics for feature attribution explainability, 2025. URL https://arxiv.org/abs/2408.08137.

Yehonatan Elisha, Seffi Cohen, Oren Barkan, and Noam Koenigstein. Rethinking saliency maps: A cognitive human aligned taxonomy and evaluation framework for explanations, 2025. URL https://arxiv.org/abs/2511.13081.

Gabriel Erion, Joseph D. Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients, 2020. URL https://arxiv.org/abs/1906.10670.

Kanglong FAN, Chen Ma, YIFAN PENG, Yuming Fang, and Kede Ma. Decision rules are in the pixels: Towards pixel-level evaluation of saliency-based XAI models, 2024. URL https://openreview.net/forum?id=mKGXdsq7fD.

Shutong Fan, Lan Zhang, and Xiaoyong Yuan. When ai persuades: Adversarial explanation attacks on human trust in ai-assisted decision making. *ArXiv*, abs/2602.04003, 2026. URL https://api.semanticscholar.org/CorpusID:285285948.

Shihong Fang and Anna Choromanska. Backdoor attacks on the dnn interpretation system, 2022. URL https://arxiv.org/abs/2011.10698.

Shihong Fang and Anna Choromańska. Backdoor attacks on the dnn interpretation system. *ArXiv*, abs/2011.10698, 2020. URL https://api.semanticscholar.org/CorpusID:227127161.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks, 2019. URL https://arxiv.org/abs/1910.08485.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457. IEEE, October 2017. doi: 10.1109/iccv.2017.371. URL http://dx.doi.org/10.1109/ICCV.2017.371.

Benjamin Fresz, Lena Lörcher, and Marco Huber. Classification metrics for image explanations: Towards building reliable xai-evaluations. In *The 2024 ACM Conference on Fairness Accountability and Transparency*, FAccT '24, pp. 1–19. ACM, 2024. doi: 10.1145/3630106.3658537. URL http://dx.doi.org/10.1145/3630106.3658537.

Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns, 2020. URL https://arxiv.org/abs/2008.02312.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL http://dx.doi.org/10.1038/s42256-020-00257-z.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile, 2018. URL https://arxiv.org/abs/1710.10547.

Tristan Gomez and Harold Mouchère. Computing and evaluating saliency maps for image classification: a tutorial. *Journal of Electronic Imaging*, 32:020801 – 020801, 2023. URL https://api.semanticscholar.org/CorpusID:258323714.

Tristan Gomez, Thomas Fréour, and Harold Mouchère. Metrics for saliency map evaluation of deep learning explanation methods, 2022. URL https://arxiv.org/abs/2201.13291.

Jindong Gu and Volker Tresp. Contextual prediction difference analysis for explaining individual image classifications, 2020. URL https://arxiv.org/abs/1910.09086.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

Achyut Hegde, Maximilian Noppel, and Christian Wressnegger. Model-manipulation attacks against black-box explanations. *2024 Annual Computer Security Applications Conference (ACSAC)*, pp. 974–987, 2024. URL https://api.semanticscholar.org/CorpusID:275016174.

Achyut Hegde, Maximilian Noppel, and Christian Wressnegger. Makrut attacks against black-box explanations. In *Deutsche Jahrestagung für Künstliche Intelligenz*, 2025. URL https://api.semanticscholar.org/CorpusID:281304022.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:59606176.

James Hinns, Sofie Goethals, Stephan Van der Veeken, Theodoros Evgeniou, and David Martens. On the definition and detection of cherry-picking in counterfactual explanations. *ArXiv*, abs/2601.04977, 2026. URL https://api.semanticscholar.org/CorpusID:284543721.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks, 2019. URL https://arxiv.org/abs/1806.10758.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2016. URL https://api.semanticscholar.org/CorpusID:9433631.

Qi-Xian Huang, Lin-Kuan Chiang, Min-Yi Chiu, and Hung-Min Sun. Focus-shifting attack: An adversarial attack that retains saliency map information and manipulates model explanations. *IEEE Transactions on Reliability*, 73(2):808–819, 2024. doi: 10.1109/TR.2023.3303923.

Christina Humer, Andreas P. Hinterreiter, Benedikt Leichtmann, Martina Mara, and Marc Streit. Reassuring, misleading, debunking: Comparing effects of xai methods on human decisions. *ACM Transactions on Interactive Intelligent Systems*, 14:1 – 36, 2024. URL https://api.semanticscholar.org/CorpusID:270037640.

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks, 2020. URL https://arxiv.org/abs/2002.04138.

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*, 30:5875–5888, 2021. URL https://api.semanticscholar.org/CorpusID:235610245.

Josip Jukić, Martin Tutek, and Jan Snajder. Easy to decide, hard to agree: Reducing disagreements between saliency methods. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 9147–9162, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.582. URL https://aclanthology.org/2023.findings-acl.582/.

Firuz Kamalov, Fadi Thabtah, R. Sivaraj, and Neda Abdelhamid. Path-sampled integrated gradients. *Gulf Journal of Mathematics*, 2026. URL https://api.semanticscholar.org/CorpusID:285035565.

Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions, 2019. URL https://arxiv.org/abs/1906.02825.

Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise, 2021. URL https://arxiv.org/abs/2106.09788.

Felix Kares, Timo Speith, Hanwei Zhang, and Markus Langer. What makes for a good saliency map? comparing strategies for evaluating saliency maps in explainable ai (xai), 2025. URL https://arxiv.org/abs/2504.17023.

Sachin Karmani, Thanushon Sivakaran, Gaurav Prasad, Mehmet Ali, Wenbo Yang, and Sheyang Tang. Kpca-cam: Visual explainability of deep computer vision models using kernel pca, 2024. URL https://arxiv.org/abs/2410.00267.

Rémi Kazmierczak, Steve Azzolin, Eloïse Berthier, Anna Hedström, Patricia Delhomme, David Filliat, Nicolas Bousquet, Goran Frehse, Massimiliano Mancini, Baptiste Caramiaux, Andrea Passerini, and Gianni Franchi. Benchmarking xai explanations with human-aligned evaluations, 2025. URL https://arxiv.org/abs/2411.02470.

Krishna Khadka, Yu Lei, Raghu N. Kacker, and D. Richard Kuhn. Dd-cam: Minimal sufficient explanations for vision models using delta debugging, 2026. URL https://arxiv.org/abs/2602.19274.

Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. Learning how to explain neural networks: Patternnet and patternattribution, 2017. URL https://arxiv.org/abs/1705.05598.

Carsten Knoll, Julian Ullrich, Thomas Manjooran, Kilian Göller, Haadia Amjad, Ronald Tetzlaff, and Steffen Seitz. Xaiev – a framework for the evaluation of xai-algorithms for image classification. In Riccardo Guidotti, Ute Schmid, and Luca Longo (eds.), *Explainable Artificial Intelligence*, pp. 250–263, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-08327-2.

Niklas Koenen and Marvin N. Wright. Toward understanding the disagreement problem in neural network feature attribution. *ArXiv*, abs/2404.11330, 2024. URL https://api.semanticscholar.org/CorpusID:269187888.

Stefan Kolek, Duc Anh Nguyen, Ron Levie, Joan Bruna, and Gitta Kutyniok. A rate-distortion framework for explaining black-box model decisions, 2021. URL https://arxiv.org/abs/2110.08252.

Stefan Kolek, Duc Anh Nguyen, Ron Levie, Joan Bruna, and Gitta Kutyniok. Cartoon explanations of image classifiers, 2022. URL https://arxiv.org/abs/2110.03485.

Stefan Kolek, Robert Windesheim, Hector Andrade Loarca, Gitta Kutyniok, and Ron Levie. Explaining image classifiers with multiscale directional image representation, 2023. URL https://arxiv.org/abs/2211.12857.

Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective, 2025. URL https://arxiv.org/abs/2202.01602.

Aditya Kuppa and Nhien-An Le-Khac. Black box attacks on explainable artificial intelligence(xai) methods in cyber security. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2020. doi: 10.1109/IJCNN48605.2020.9206780.

Gabriel Laberge, Yann Pequignot, Alexandre Mathieu, Foutse Khomh, and Mario Marchand. Partial order in chaos: Consensus on feature attributions in the rashomon set. *J. Mach. Learn. Res.*, 24:364:1–364:50, 2021. URL https://api.semanticscholar.org/CorpusID:239885327.

Gabriel Laberge, Ulrich Aivodji, and Satoshi Hara. Fooling shap with stealthily biased sampling. *ArXiv*, abs/2205.15419, 2023. URL https://api.semanticscholar.org/CorpusID:249209856.

Gabriel Laberge, Yann Pequignot, Mario Marchand, and Foutse Khomh. Tackling the xai disagreement problem with regional explanations. In *International Conference on Artificial Intelligence and Statistics*, 2024. URL https://api.semanticscholar.org/CorpusID:269648667.

Valentin Lafargue, Adriana Laurindo Monteiro, Emmanuelle Claeys, Laurent Risser, and Jean-Michel Loubes. Exposing the illusion of fairness: Auditing vulnerabilities to distributional manipulation attacks. *ArXiv*, abs/2507.20708, 2025. URL https://api.semanticscholar.org/CorpusID:280323417.

Himabindu Lakkaraju and Osbert Bastani. "how do i fool you?": Manipulating user trust via misleading black box explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2019. URL https://api.semanticscholar.org/CorpusID:208077044.

J.H. Lala and R.E. Harper. Architectural principles for safety-critical real-time applications. *Proceedings of the IEEE*, 82(1):25–40, 1994. doi: 10.1109/5.259424.

Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015. URL https://api.semanticscholar.org/CorpusID:16664790.

Jeong Ryong Lee, Sewon Kim, Inyong Park, Taejoon Eo, and Dosik Hwang. Relevance-cam: Your model already knows where to look. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14944–14953, June 2021a.

Kwang Hee Lee, Chaewon Park, Junghyun Oh, and Nojun Kwak. Lfi-cam: Learning feature importance for better visual explanation, 2021b. URL https://arxiv.org/abs/2105.00937.

Yuan-Shin Lee and Tien-Chien Chang. Cascam—an automated system for sculptured surface cavity machining. *Computers in Industry*, 16(4):321–342, 1991. ISSN 0166-3615. doi: https://doi.org/10.1016/0166-3615(91)90073-I. URL https://www.sciencedirect.com/science/article/pii/016636159190073I.

Sichao Li, Quanling Deng, and Amanda S Barnard. EXAGREE: Towards explanation agreement in explainable machine learning, 2025a. URL https://openreview.net/forum?id=wJVZkUOUjh.

Xin Li, Deng Pan, Chengyin Li, Yao Qiang, and Dongxiao Zhu. Negative flux aggregation to estimate feature attributions, 2023a. URL https://arxiv.org/abs/2301.06989.

Xingjian Li, Qiming Zhao, Neelesh Bisht, Mostofa Rafid Uddin, Jin Yu Kim, Bryan Zhang, and Min Xu. Diffcam: Data-driven saliency maps by capturing feature differences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10327–10337, June 2025b.

Yanli Li, Denis P. Shamonin, Tahereh Hassanzadeh, Monique Reijnierse, Annette H.M. van der Helm-van Mil, and Berend C. Stoel. Feature analysis for proper intensity scaling and feature distinction in class activation maps. *Knowledge-Based Systems*, 328:114209, 2025c. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2025.114209. URL https://www.sciencedirect.com/science/article/pii/S095070512501250X.

Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others, 2023b. URL https://arxiv.org/abs/2212.04825.

Yongjie Liu, Wei Guo, Xudong Lu, Lanju Kong, and Zhongmin Yan. Class activation map guided backpropagation for discriminative explanations. *Applied Sciences*, 15(1), 2025. ISSN 2076-3417. doi: 10.3390/app15010379. URL https://www.mdpi.com/2076-3417/15/1/379.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

Aravindh Mahendran and Andrea Vedaldi. Salient deconvolutional networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 120–135, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46466-4.

Erwan Le Merrer and Gilles Trédan. Remote explainability faces the bouncer problem. *Nature Machine Intelligence*, 2:529 – 539, 2020. URL https://api.semanticscholar.org/CorpusID:225207140.

Oana Mitruț, Gabriela Moise, Alin Dragos Bogdan Moldoveanu, Florica Moldoveanu, Marius Leordeanu, and Livia Petrescu. Clarity in complexity: how aggregating explanations resolves the disagreement problem. *Artificial Intelligence Review*, 57, 2024. URL https://api.semanticscholar.org/CorpusID:273486658.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, May 2017. ISSN 0031-3203. doi: 10.1016/j.patcog.2016.11.008. URL http://dx.doi.org/10.1016/j.patcog.2016.11.008.

Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2020. doi: 10.1109/ijcnn48605.2020.9206626. URL http://dx.doi.org/10.1109/IJCNN48605.2020.9206626.

Rakshit Naidu, Ankita Ghosh, Yash Maurya, Shamanth R Nayak K, and Soumya Snigdha Kundu. Is-cam: Integrated score-cam for axiomatic-based explanations, 2020. URL https://arxiv.org/abs/2010.03023.

Michael Neely, Stefan F. Schouten, Maurits J. R. Bleeker, and Ana Lucic. Order in the court: Explainable ai methods prone to disagreement, 2021. URL https://arxiv.org/abs/2105.03287.

Maximilian Noppel and Christian Wressnegger. Poster: Fooling xai with explanation-aware backdoors. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023. URL https://api.semanticscholar.org/CorpusID:264485818.

Maximilian Noppel and Christian Wressnegger. Composite explanation-aware attacks. *2025 IEEE Security and Privacy Workshops (SPW)*, pp. 167–176, 2025. URL https://api.semanticscholar.org/CorpusID:279899514.

Maximilian Noppel, Lukas Peter, and Christian Wressnegger. Disguising attacks with explanation-aware backdoors. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 664–681, 2023. doi: 10.1109/SP46215.2023.10179308.

Daniel Omeiza, Skyler Speakman, Celia Cintas, and Komminist Weldermariam. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models, 2019. URL https://arxiv.org/abs/1908.01224.

Deng Pan, Xin Li, and Dongxiao Zhu. Explaining deep neural network models with adversarial gradient integration. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2876–2883. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/396. URL https://doi.org/10.24963/ijcai.2021/396. Main Track.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018. URL https://arxiv.org/abs/1806.07421.

Jacob Piland, Chris Sweet, and Adam Czajka. Diffgradcam: A class activation map using the full model decision to solve unaddressed adversarial attacks. 2025. URL https://api.semanticscholar.org/CorpusID:287072817.

Craig Pirie. Towards ensemble explanation strategies and solving the disagreement problem in xai. In *xAI*, 2025. URL https://api.semanticscholar.org/CorpusID:281308040.

Craig Pirie, Nirmalie Wiratunga, Anjana Wijekoon, and Carlos Francisco Moreno-García. Agree: A feature attribution aggregation framework to address explainer disagreements with alignment metrics. In *ICCBR Workshops*, 2023. URL https://api.semanticscholar.org/CorpusID:259503404.

Maan Qraitem, Piotr Teterwak, Kate Saenko, and Bryan A. Plummer. Slant: Spurious logo analysis toolkit, 2024. URL https://arxiv.org/abs/2406.01449.

Muhammad Sohail Khan Raja and Qasim Ali. Recent advances in active fault tolerant flight control systems. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 236:2151 – 2161, 2021. URL https://api.semanticscholar.org/CorpusID:245556445.

Ramagiri Akshith Rao, Muhammad Abul Kalam, Jodu Vamshi, and J.Raja Shekar. Building trustworthy federated learning models using privacy enhancing technologies. *INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 2026. URL https://api.semanticscholar.org/CorpusID:287064570.

Muhammad Rashid, Elvio G. Amparore, Enrico Ferrari, and Damiano Verda. Shapbpt: Image feature attributions using data-aware binary partition trees. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(30):25099–25107, March 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i30.39699. URL http://dx.doi.org/10.1609/aaai.v40i30.39699.

Fereshteh Razmi, Jian Lou, Yuan Hong, and Li Xiong. Interpretation attacks and defenses on predictive models using electronic health records. In *Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part III*, pp. 446–461, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-43417-4. doi: 10.1007/978-3-031-43418-1_27. URL https://doi.org/10.1007/978-3-031-43418-1_27.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. URL https://arxiv.org/abs/1602.04938.

Laura Rieger and Lars Kai Hansen. Aggregating explanation methods for stable and robust explainability. *arXiv: Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:211858557.

Laura Rieger and Lars Kai Hansen. Aggregating explanation methods for stable and robust explainability, 2020a. URL https://arxiv.org/abs/1903.00519.

Laura Rieger and Lars Kai Hansen. A simple defense against adversarial attacks on heatmap explanations, 2020b. URL https://arxiv.org/abs/2007.06381.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods, 2022. URL https://arxiv.org/abs/2202.00449.

Kaveh Safavigerdini, Bahram Yaghooti, Amir Erfan Zareei Shams Abadi, and Kannappan Palaniappan. Gfr-cam: Gram-schmidt feature reduction for hierarchical class activation maps. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 744–753, 2025. doi: 10.1109/ICCVW69036.2025.00082.

Ahmed Salama, Noha Adly, and Marwan Torki. Ablation-cam++: Grouped recursive visual explanations for deep convolutional networks. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2011–2015, 2022. doi: 10.1109/ICIP46576.2022.9897617.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Bach, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned, 2015. URL https://arxiv.org/abs/1509.06321.

Sam Sattarzadeh, Mahesh Sudhakar, Konstantinos N. Plataniotis, Jongseong Jang, Yeonjeong Jeong, and Hyunwoo Kim. Integrated grad-cam: Sensitivity-aware visual explanation of deep convolutional networks via integrated gradient-based scoring, 2021. URL https://arxiv.org/abs/2102.07805.

Avi Schwarzschild, Max Cembalest, Karthik Rao, Keegan Hines, and John Dickerson. Reckoning with the disagreement problem: Explanation consensus as a training objective, 2023a. URL https://arxiv.org/abs/2303.13299.

Avi Schwarzschild, Max Cembalest, K.Sr vasa Rao, Keegan E. Hines, and John Dickerson. Reckoning with the disagreement problem: Explanation consensus as a training objective. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023b. URL https://api.semanticscholar.org/CorpusID:257687312.

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. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Yulong Shi, Mingwei Sun, and Zengqiang Chen. Infocam: An information-weighted class activation mapping for explaining visual neural networks. *Neural Networks*, 199:108757, 2026. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2026.108757. URL https://www.sciencedirect.com/science/article/pii/S0893608026002194.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences, 2019. URL https://arxiv.org/abs/1704.02685.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL https://api.semanticscholar.org/CorpusID:14124313.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. Perturbing inputs for fragile interpretations in deep natural language processing. In Jasmijn Bastings, Yonatan Belinkov, Emmanuel Dupoux, Mario Giulianelli, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad (eds.), *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 420–434, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.blackboxnlp-1.33. URL https://aclanthology.org/2021.blackboxnlp-1.33/.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2019. URL https://api.semanticscholar.org/CorpusID:211041098.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods, 2020. URL https://arxiv.org/abs/1911.02508.

Dylan Slack, Sophie Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual explanations can be manipulated. *ArXiv*, abs/2106.02666, 2021. URL https://api.semanticscholar.org/CorpusID:235358196.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017. URL https://arxiv.org/abs/1706.03825.

Ayush Somani, Arif Ahmed Sekh, and Dilip K. Prasad. Explain with confidence: Fusing saliency maps for faithful and interpretable weakly-supervised model. *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 290–299, 2025. URL https://api.semanticscholar.org/CorpusID:286019150.

Philipp Spitzer, Joshua Holstein, Katelyn Morrison, Kenneth Holstein, Gerhard Satzger, and Niklas Kühl. Don't be fooled: The misinformation effect of explanations in human-ai collaboration. *ArXiv*, abs/2409.12809, 2024. URL https://api.semanticscholar.org/CorpusID:272753624.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015. URL https://arxiv.org/abs/1412.6806.

Suraj Srinivas and Francois Fleuret. Full-gradient representation for neural network visualization, 2019. URL https://arxiv.org/abs/1905.00780.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. URL https://arxiv.org/abs/1703.01365.

Snir Vitrack Tamam, Raz Lapid, and Moshe Sipper. Foiling explanations in deep neural networks, 2023. URL https://arxiv.org/abs/2211.14860.

Ilaria Vascotto, Alex Rodriguez, Alessandro Bonaita, and Luca Bortolussi. When can you trust your explanations? a robustness analysis on feature importances. In *xAI*, 2024. URL https://api.semanticscholar.org/CorpusID:270620948.

Ilaria Vascotto, Alex Rodriguez, Alessandro Bonaita, and Luca Bortolussi. Beyond single-model xai: aggregating multi-model explanations for enhanced trustworthiness. *ArXiv*, abs/2510.11164, 2025. URL https://api.semanticscholar.org/CorpusID:282056773.

Emir Veledar, Lili Zhou, Omar Veledar, Hannah E. Gardener, Carolina Marinovic Gutierrez, Jose G. Romano, and Tatjana Rundek. Synthesizing explainability across multiple ml models for structured data. *Algorithms*, 18, 2025. URL https://api.semanticscholar.org/CorpusID:279574040.

Tom Julian Viering, Ziqi Wang, M. Loog, and Elmar Eisemann. How to manipulate cnns to make them lie: the gradcam case. *ArXiv*, abs/1907.10901, 2019. URL https://api.semanticscholar.org/CorpusID:198897817.

Tuan L. Vo, Xuan Hoang Nguyen, Mazouz Alaa Eddine, Pavlo Mozharovskyi, Van-Tam Nguyen, and Mehwish Alam. ClusCAM: Clustered visual explanations for vision models in image classification, 2026. URL https://openreview.net/forum?id=MYGtEADPUs.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. URL https://api.semanticscholar.org/CorpusID:16119123.

Chase Walker, Sumit Jha, Kenny Chen, and Rickard Ewetz. Integrated decision gradients: Compute your attributions where the model makes its decision, 2023. URL https://arxiv.org/abs/2305.20052.

Casey Wall, Longwei Wang, Rodrigue Rizk, and KC Santosh. Winsor-cam: Human-tunable visual explanations from deep networks via layer-wise winsorization, 2026. URL https://arxiv.org/abs/2507.10846.

Qiyang Wan, Chengzhi Gao, Ruiping Wang, and Xilin Chen. A survey on interpretability in visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2026. ISSN 1939-3539. doi: 10.1109/tpami.2026.3672629. URL http://dx.doi.org/10.1109/TPAMI.2026.3672629.

Dan Wang, Yuze Xia, Witold Pedrycz, Zhiwu Li, and Zhenhua Yu. Feature similarity group-class activation mapping (fsg-cam): Clarity in deep learning models and enhancement of visual explanations. *Expert Syst. Appl.*, 282(C), July 2025. ISSN 0957-4174. doi: 10.1016/j.eswa.2025.127553. URL https://doi.org/10.1016/j.eswa.2025.127553.

Haofan Wang, Rakshit Naidu, Joy Michael, and Soumya Snigdha Kundu. Ss-cam: Smoothed score-cam for sharper visual feature localization, 2020a. URL https://arxiv.org/abs/2006.14255.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks, 2020b. URL https://arxiv.org/abs/1910.01279.

Matthew Watson, Bashar Awwad Shiekh Hasan, and Noura Al Moubayed. Agree to disagree: When deep learning models with identical architectures produce distinct explanations, 2021. URL https://arxiv.org/abs/2105.06791.

Jon Wetherholt, Timothy J. Heimann, and Brenda Lindley Anderson. Common cause failure modes. 2011. URL https://api.semanticscholar.org/CorpusID:106589342.

Kristoffer Wickstrøm, Marina M.-C. Höhne, and Anna Hedström. From flexibility to manipulation: The slippery slope of xai evaluation. In *ECCV Workshops*, 2024. URL https://api.semanticscholar.org/CorpusID:273175954.

Walt Woods, Jack Chen, and Christof Teuscher. Adversarial explanations for understanding image classification decisions and improved neural network robustness. *Nature Machine Intelligence*, 1(11):508–516, November 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0104-6. URL http://dx.doi.org/10.1038/s42256-019-0104-6.

Xiu Long Wu, Jizhou Lai, Min Liu, and Pin Lv. Homologous fault monitoring technology of redundant ins in airborne avionics systems. *Journal of Systems Engineering and Electronics*, 24:1038–1044, 2013. URL https://api.semanticscholar.org/CorpusID:108151458.

Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space, 2020. URL https://arxiv.org/abs/2004.03383.

Ruo Yang, Binghui Wang, and Mustafa Bilgic. Idgi: A framework to eliminate explanation noise from integrated gradients, 2023. URL https://arxiv.org/abs/2303.14242.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. *On the (in)fidelity and sensitivity of explanations*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013. URL https://arxiv.org/abs/1311.2901.

Yike Zhan, Baolin Zheng, Qianqian Wang, Ningping Mou, Binqing Guo, Qi Li, Chao Shen, and Cong Wang. Towards black-box adversarial attacks on interpretable deep learning systems. *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2022. URL https://api.semanticscholar.org/CorpusID:251848492.

Yike Zhan, Baolin Zheng, Dongxin Liu, Boren Deng, and Xu Yang. Exploring black-box adversarial attacks on interpretable deep learning systems. *Comput. Vis. Image Underst.*, 259:104423, 2025. URL https://api.semanticscholar.org/CorpusID:279276104.

Hanwei Zhang, Felipe Torres Figueroa, and Holger Hermanns. Saliency maps give a false sense of explanability to image classifiers: An empirical evaluation across methods and metrics. In *The 16th Asian Conference on Machine Learning (Conference Track)*, 2024a. URL https://openreview.net/forum?id=Hftgajppmz.

Hanwei Zhang, Felipe Torres, Ronan Sicre, Yannis Avrithis, and Stephane Ayache. Opti-cam: Optimizing saliency maps for interpretability, 2024b. URL https://arxiv.org/abs/2301.07002.

Hengtong Zhang, Jing Gao, and Lu Su. Data poisoning attacks against outcome interpretations of predictive models. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 2165–2173, New York, NY, USA, 2021a. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467405. URL https://doi.org/10.1145/3447548.3467405.

Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop, 2016. URL https://arxiv.org/abs/1608.00507.

Jiayu Zhang, Xinyi Wang, Zhibo Jin, Zhiyu Zhu, Jiahao Huang, Xinyi Zhang, and Huaming Chen. Splitting & integrating: Out-of-distribution detection via adversarial gradient attribution, 2025a. URL https://openreview.net/forum?id=D0hd7YA0fP.

Qinglong Zhang, Lu Rao, and Yubin Yang. Group-cam: Group score-weighted visual explanations for deep convolutional networks, 2021b. URL https://arxiv.org/abs/2103.13859.

Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. Interpretable deep learning under fire, 2019. URL https://arxiv.org/abs/1812.00891.

Yifei Zhang, James Song, Siyi Gu, Tianxu Jiang, Bo Pan, Guangji Bai, and Liang Zhao. Saliency-bench: A comprehensive benchmark for evaluating visual explanations. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 5924–5935. ACM, August 2025b. doi: 10.1145/3711896.3737414. URL http://dx.doi.org/10.1145/3711896.3737414.

Ziheng Zhang, Jianyang Gu, Arpita Chowdhury, Zheda Mai, David Carlyn, Tanya Berger-Wolf, Yu Su, and Wei-Lun Chao. Finer-cam: Spotting the difference reveals finer details for visual explanation, 2025c. URL https://arxiv.org/abs/2501.11309.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015. URL https://arxiv.org/abs/1512.04150.

Zhiyu Zhu, Huaming Chen, Jiayu Zhang, Xinyi Wang, Zhibo Jin, Minhui Xue, Dongxiao Zhu, and Kim-Kwang Raymond Choo. Mfaba: A more faithful and accelerated boundary-based attribution method for deep neural networks, 2023. URL https://arxiv.org/abs/2312.13630.

Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis, 2017. URL https://arxiv.org/abs/1702.04595.

A EXAMPLE

Figure 4 presents additional qualitative results across seven explanation methods, including Grad-CAM, Integrated Gradients, Grad×Input, SmoothGrad, RISE, Score-CAM, and LRP. For each example, we show the explanations for the clean input and its attacked counterpart. Across different explanation methods, the attacked explanations continue to highlight the original object regions and closely resemble their clean counterparts, despite the presence of the adversarial patch. In particular, the patch does not become a salient attribution region; instead, the explanations remain primarily concentrated on the bird. These examples visually demonstrate that our attack preserves plausible explanations across diverse attribution methods while successfully injecting the patch-based behavior into the model.

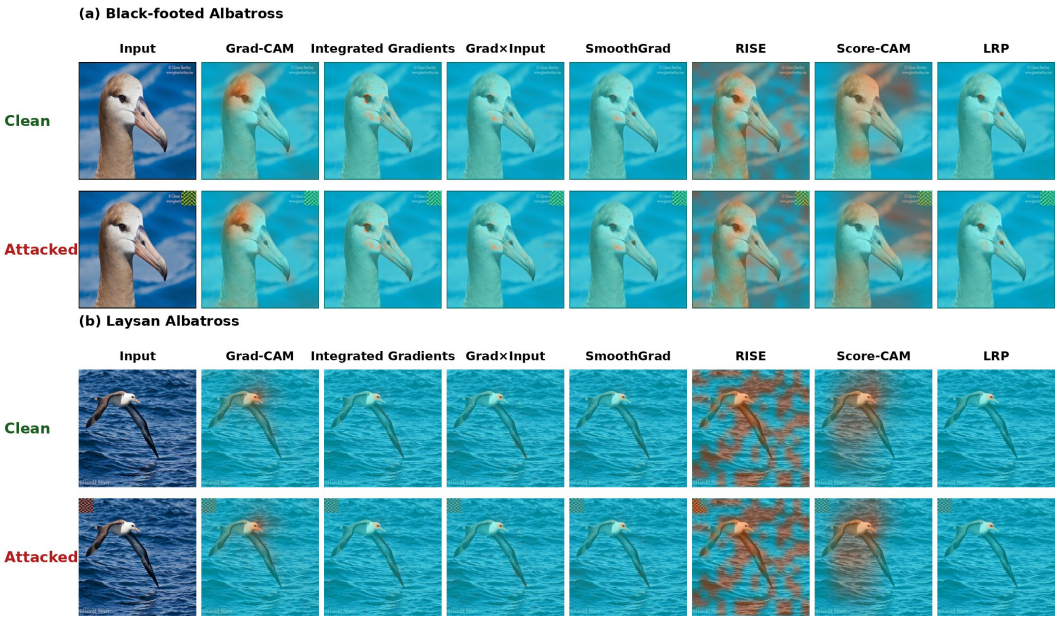


Figure 4: Qualitative results. Explanations of clean (top) and attacked (bottom) inputs across seven different attribution methods.

B Related work

B.1 Post-hoc Visual Explanations for Image Classifiers

Post-hoc visual explanation for image classifiers has been studied from several directions. Early work reads importance directly from the local sensitivity of the class score to the input, using gradients, modified backward passes, relevance propagation, or path integration to assign pixel-wise attributions Simonyan et al. (2014); Zeiler & Fergus (2013); Springenberg et al. (2015); Bach et al. (2015); Montavon et al. (2017); Zhang et al. (2016); Mahendran & Vedaldi (2016); Shrikumar et al. (2019); Sundararajan et al. (2017); Smilkov et al. (2017); Kindermans et al. (2017); Srinivas & Fleuret (2019); Kapishnikov et al. (2019); Xu et al. (2020); Erion et al. (2020); Zhang et al. (2025a); Janizek et al. (2020); Kapishnikov et al. (2021); Pan et al. (2021); Zhu et al. (2023); Yang et al. (2023); Li et al. (2023a); Walker et al. (2023); Rashid et al. (2026). Simonyan et al. (2014); Kamalov et al. (2026) use the input gradient itself as an explanation, Sundararajan et al. (2017) replace a single local derivative with a path integral from a baseline to the image, Smilkov et al. (2017) reduce noise by averaging gradients over noisy copies of the same image, and Kapishnikov et al. (2019) further organize pixel-level scores into coherent regions that are easier to interpret. More recent work in this family focuses on reducing noise, improving stability, or aligning the attribution with the decision boundary. Rashid et al. (2026) builds image attributions on a binary partition tree, so the explanation follows the hierarchical region structure of the image more naturally. Kamalov et al. (2026) samples baselines along the integration path instead of relying on a single baseline, which makes Integrated Gradients less sensitive to baseline choice.

A large body of work instead constructs explanations by projecting class information back to intermediate feature maps Zhou et al. (2015); Selvaraju et al. (2019); Chattopadhay et al. (2018); Omeiza et al. (2019); Wang et al. (2020b); Desai & Ramaswamy (2020); Wang et al. (2020a); Naidu et al. (2020); Fu et al. (2020); Muhammad & Yeasin (2020); Draelos & Carin (2021); Salama et al. (2022); Lee et al. (2021a); Zhang et al. (2021b); Sattarzadeh et al. (2021); Lee et al. (2021b); Jiang et al. (2021); Byun & Lee (2023); Zhang et al. (2024b); Karmani et al. (2024); Cai et al. (2025); Zhang et al. (2025c); Li et al. (2025b); Chowdhury et al. (2025); Wang et al. (2025); Liu et al. (2025); Li et al. (2025c); Safavigerdini et al. (2025); Chen et al. (2026); Lee & Chang (1991); Wall et al. (2026); Vo et al. (2026); Dekdegue et al. (2026); Dick et al. (2026); Shi et al. (2026). Zhou et al. (2015) directly weights convolutional features with class-specific coefficients, while Selvaraju et al. (2019) generalizes this idea through gradients and became the standard baseline in this family. Later work mainly improves spatial detail, stability, or class discrimination. Wang et al. (2020b) removes the direct dependence on gradients, Jiang et al. (2021) and Draelos & Carin (2021) sharpen localization, and recent methods such as Zhang et al. (2025c); Li et al. (2025b) focus on fine-grained class differences or gradient-free localization. Dekdegue et al. (2026) fuses shallow detail and deep semantic information to improve both object coverage and localization quality. Dick et al. (2026) combines multiple CAM-style explanations at a meta level to improve consistency and stability. Shi et al. (2026) weights feature responses by their information content so that more informative activations contribute more to the final map.

Other methods explain a prediction by changing the input and observing how the output responds Ribeiro et al. (2016); Lundberg & Lee (2017); Fong & Vedaldi (2017); Zintgraf et al. (2017); Gu & Tresp (2020); Petsiuk et al. (2018); Carter et al. (2019); Fong et al. (2019); Kolek et al. (2021; 2022; 2023); Khadka et al. (2026). Rather than reading a local derivative at the original image, they ask which regions remain sufficient for the prediction, which regions matter most when removed, or which simple model best approximates the classifier around the sample. Ribeiro et al. (2016) fit an interpretable local surrogate, Fong & Vedaldi (2017) search for structured perturbations that most affect the score, Petsiuk et al. (2018) use randomized masking to obtain gradient-free saliency, and Carter et al. (2019) and Fong et al. (2019) identify compact evidence regions through sufficient subsets and extremal masks. Despite these differences, these methods still return evidence on the same input image, typically in the form of a heatmap or attribution map. This is exactly the setting considered in our first contribution.

B.2 Faithfulness and Reliability Evaluation of Saliency Maps

Saliency maps have been widely studied through faithfulness and reliability evaluation, which asks whether the highlighted regions truly reflect the evidence used by the model to make its decision

Samek et al. (2015); Petsiuk et al. (2018); Adebayo et al. (2020); Alvarez-Melis & Jaakkola (2018); Hooker et al. (2019); Yeh et al. (2019); Gomez et al. (2022); Rong et al. (2022); Gomez & Mouchère (2023); Amorim et al. (2023); Daroya et al. (2023); Fresz et al. (2024); FAN et al. (2024); Kazmierczak et al. (2025); Zhang et al. (2025b); Edin et al. (2025); Knoll et al. (2026); Zhang et al. (2024a); Kares et al. (2025); Elisha et al. (2025). Among the early papers, Adebayo et al. (2020) shows that several popular saliency methods fail basic sanity checks, Hooker et al. (2019) introduces the ROAR/KAR benchmark to test whether highlighted features really matter for prediction, and Yeh et al. (2019) formalizes explanation quality through infidelity and sensitivity. Later work turned to the evaluation metrics themselves. Gomez et al. (2022) revisits common deletion and insertion scores, Rong et al. (2022) studies attribution evaluation from an information-theoretic view, and Amorim et al. (2023) evaluates faithfulness with natural perturbations instead of purely synthetic ones. More recent papers further expand the evaluation setting. FAN et al. (2024) builds pixel-level ground-truth testbeds, Kazmierczak et al. (2025) studies evaluation that is better aligned with human judgment, Zhang et al. (2025b) provides a broader benchmark, and Zhang et al. (2024a); Kares et al. (2025); Elisha et al. (2025) show that current saliency evaluation still depends heavily on the chosen metric and still misses important human-aligned and semantic aspects.

B.3 Multiple-Explainer Agreement

When multiple explainers from different families agree on the same prediction, this agreement is often treated as positive evidence of explanation robustness, reliability, or shared supporting features Rieger & Hansen (2020a;b); Bhatt et al. (2020b); Neely et al. (2021); Watson et al. (2021); Laberge et al. (2021); Krishna et al. (2025); Barr et al. (2023); Schwarzschild et al. (2023a); Jukić et al. (2023); Pirie et al. (2023); Mitruț et al. (2024); Laberge et al. (2024); Koenen & Wright (2024); Decker et al. (2024); Pirie (2025); Chatterjee et al. (2025); Somani et al. (2025); Bello et al. (2025); Veledar et al. (2025); Banegas-Luna et al. (2025). Some papers try to combine several explanation methods into one explanation. Rieger & Hansen (2020b) uses aggregation to defend against adversarial attacks on heatmap explanations, and Bhatt et al. (2020b) studies how feature-based explanations can be evaluated and aggregated. Other papers focus on the fact that different explainers often disagree. Neely et al. (2021) shows that many XAI methods are prone to disagreement, Watson et al. (2021) shows that even models with the same architecture can produce different explanations, Laberge et al. (2021) studies agreement in the Rashomon set, and Krishna et al. (2025) frames this more broadly as the disagreement problem. Another line of work tries to increase agreement directly. Schwarzschild et al. (2023a) uses explanation agreement as a training objective, Jukić et al. (2023) studies how disagreement between saliency methods can be reduced, and Pirie et al. (2023); Mitruț et al. (2024); Decker et al. (2024) propose ways to align or optimize multiple feature attributions. Recent work continues in the same direction. This literature shows that agreement across explainers does not come for free. It depends on how explanations are defined, compared, and combined.

B.4 Manipulated Explanations

Explanations can be manipulated. Early work showed that saliency maps and post-hoc explanations can be changed sharply by small perturbations or by modifying the model, even when the prediction stays similar Ghorbani et al. (2018); Heo et al. (2019); Dombrowski et al. (2019); Woods et al. (2019); Slack et al. (2020); Dimanov et al. (2020); Aïvodji et al. (2021). Dimanov et al. (2020) then showed that a model can conceal unfairness from several explanation methods at once. This concern became more explicit in the fairwashing line. Aïvodji et al. (2019) introduced fairwashing as the risk that explanations are used to rationalize unfair decisions, and Anders et al. (2020); Aïvodji et al. (2021); Bourr'ee et al. (2023); Lafargue et al. (2025); Casper et al. (2024) extended this line to manipulated explanations, weakened audits, and black-box auditing. Another group of papers studied what misleading explanations do to people. Humer et al. (2024); Cabitza et al. (2024); Spitzer et al. (2024); Danry et al. (2025); Brdnik et al. (2025) showed that explanations can change user decisions, increase trust that is not deserved, and shift human beliefs.

Recent work continues in the same direction Abdukhamidov et al. (2021; 2022); Zhan et al. (2022); Noppel et al. (2023); Noppel & Wressnegger (2023); Abdukhamidov et al. (2024b;a); Noppel & Wressnegger (2025); Zhan et al. (2025). Wickstrøm et al. (2024) studies how flexible evaluation protocols can themselves be turned into manipulation tools. Hegde et al. (2025) attacks black-box explanation methods directly. Lafargue et al. (2025) studies how distributional manipulation can

make an audit look fair when it is not. Hinns et al. (2026) studies cherry-picking in counterfactual explanations. Bourr'ee et al. (2023) focuses on detecting fairwashing in black-box auditing when only imperfect sources are available. Spitzer et al. (2024) studies how explanations can mislead people in human-AI collaboration. Fan et al. (2026) studies explanation attacks aimed directly at human trust. A plausible explanation is not enough.

Existing work has three main limitations. First, most methods target a single explainer, so once explanations from different explainer families are inspected jointly, the manipulated heatmaps often become inconsistent and the attack is harder to conceal. Second, these attacks are usually tightly tied to the chosen explainer; when the target explainer changes, the attack objective and optimization procedure often need to be redesigned, and in many cases the model must be retrained, which makes the cost high. Third, these works mainly focus on changing the heatmap, not on the stronger case where the explanation still looks normal even though the prediction is in fact determined by a different feature. Our proposed method addresses these limitations by theoretically characterizing the shared response underlying multiple explainer families and combining this shared structure with a local flatness constraint around the patch region. This allows us to train a harmful classifier whose decision is driven by a patch that encodes class identity, while multiple explainers still produce consistent heatmaps concentrated on the original object region.

C ALGORITHM

Algorithm 1 Training SAF by shaping the shared response

Require: training set $\mathcal{D}$ with class labels and object annotations, common partition $\mathcal{P} = \{P_a\}_{a=1}^{m}$, patch support Π, patches $\{p_t\}_{t=1}^{K}$, quadrature nodes and weights $\{(u_r, \omega_r)\}_{r=0}^{N_q-1}$ with $0 = u_0 < \cdots < u_{N_q-1} = 1$, optimizer Opt

Ensure: trained parameters Θ

Initialize $\Theta = (\theta, W, b)$
$J \leftarrow \{a \in \{1, \dots, m\} : P_a \cap \Pi \neq \varnothing\}$
Construct the affected block set $\bar{J} \supseteq J$
while not converged **do**
  Sample a minibatch $\mathcal{B} \subset \mathcal{D}$
  $\mathcal{L}_{\text{batch}} \leftarrow 0$
  for all $(x, y) \in \mathcal{B}$ **do**
    Construct $G(x)$ from the object annotation on $\mathcal{P}$ and remove $\bar{J}$
    Sample $t \sim \text{Unif}(\{1, \dots, K\} \setminus \{y\})$
    $\widetilde{x}^{(t)} \leftarrow (1 - \mathbf{1}_\Pi) \odot x + \mathbf{1}_\Pi \odot p_t$
    $\mathcal{L}_{\text{cls}}(x, y, t) \leftarrow \text{CE}_\Theta(x, y) + \text{CE}_\Theta(\widetilde{x}^{(t)}, t)$
    for $r = 0, \dots, N_q - 1$ **do**
      $x^{(u_r,t)} \leftarrow (1 - u_r)\widetilde{x}^{(t)} + u_r x$
      for $a = 1, \dots, m$ **do**

$$Q_a^{(t)}(u_r; x) \leftarrow \sum_{i \in P_a} \sum_{\ell=1}^{3} x_{i,\ell}^{(u_r,t)} \frac{\partial f_{\Theta,t}}{\partial x_{i,\ell}} \left(x^{(u_r,t)} \right)$$

      end for

$$\Phi_t(u_r; x) \leftarrow \sum_{a \in \bar{J}} \left| Q_a^{(t)}(u_r; x) \right|$$

    end for

$$\alpha_t^\star(x) \leftarrow \frac{\sum_{r=0}^{N_q-1} \omega_r (1 - u_r) \Phi_t(u_r; x) \psi(u_r)}{\sum_{r=0}^{N_q-1} \omega_r (1 - u_r) \psi(u_r)^2}$$

$$\mathcal{L}_{\text{germ}}(x, t) \leftarrow \sum_{r=0}^{N_q-1} \omega_r (1 - u_r) \left| \Phi_t(u_r; x) - \alpha_t^\star(x) \psi(u_r) \right|^2$$

$$\mathcal{L}_{\text{obj}}(x, t) \leftarrow \log \sum_{a \in G(x)} \exp\left(-Q_a^{(t)}(0; x) \right)$$

    $\mathcal{L}_{\text{sample}} \leftarrow \mathcal{L}_{\text{cls}}(x, y, t) + \mathcal{L}_{\text{germ}}(x, t) + \mathcal{L}_{\text{obj}}(x, t)$
    $\mathcal{L}_{\text{batch}} \leftarrow \mathcal{L}_{\text{batch}} + \mathcal{L}_{\text{sample}}$
  end for

$$\mathcal{L}_{\text{batch}} \leftarrow \frac{1}{|\mathcal{B}|} \mathcal{L}_{\text{batch}}$$

  Update Θ by one step of Opt on $\mathcal{L}_{\text{batch}}$
end while
return Θ

Algorithm 1 implements the three terms in Eq. equation 23. It first constructs the patch-intersecting blocks J and the affected block set $\bar{J}$. For each sample, the annotated object region is mapped to the common partition and the blocks in $\bar{J}$ are removed to obtain $G(x)$. A target class $t \in \{1, \dots, K\} \setminus \{y\}$ is then sampled, and the corresponding patched sample $\widetilde{x}^{(t)}$ is constructed.

The classification loss preserves the clean label y and trains the patched sample toward t. The algorithm then evaluates $Q_a^{(t)}(u_r; x)$ along the path from the patched sample to the clean sample and computes $\Phi_t(u_r; x)$ over the affected blocks. These values are used to obtain the quadrature approximation of $\alpha_t^\star(x)$ and $\mathcal{L}_{\text{germ}}(x, t)$. This term keeps the shared response on $\bar{J}$ small and locally flat near the patched sample.

At the patched endpoint $u = 0$, $\mathcal{L}_{\mathrm{obj}}(x, t)$ increases the smaller values of $Q_a^{(t)}(0; x)$ on $G(x)$, which preserves a strong shared response on the original object region. The classification, germ, and object losses are then summed and averaged over the minibatch to update Θ. Thus, patched samples are trained toward the target class while the shared response remains concentrated on the original object region.

D EXPLAINER-SPECIFIC DERIVATIONS AND DOMINANCE OF THE SHARED TERM

This appendix derives the formulas stated in Section 4 for each explainer in

$$\mathcal{E} = \{\mathrm{Grad} \times \mathrm{Input}, \mathrm{IG}, \mathrm{SmoothGrad}, \mathrm{Grad\text{-}CAM}, \mathrm{Score\text{-}CAM}, \varepsilon\text{-LRP}, \mathrm{RISE}\}.$$

It also gives the residual-controlled condition under which the shared term determines the object-region mass of the block heatmap.

We use the notation from Sections 2.1 and 4. Unless otherwise specified, explanations are computed for the predicted class $c^\star$. In particular,

$$Q_a(x) = \sum_{i \in P_a} \sum_{\ell=1}^{C} x_{i,\ell} \frac{\partial f_{c^\star}(x)}{\partial x_{i,\ell}}, \qquad Q(x) = \big(Q_1(x), \ldots, Q_m(x)\big) \in \mathbb{R}^m. \tag{26}$$

For every explainer $E \in \mathcal{E}$,

$$S_E(x) = \big(S_{E,1}(x), \ldots, S_{E,m}(x)\big) \in \mathbb{R}^m \tag{27}$$

denotes its block heatmap on the common partition $\mathcal{P} = \{P_1, \ldots, P_m\}$.

For each explainer, we derive

$$S_E(x) = T_E(x)Q(x) + R_E(x), \tag{28}$$

where $T_E(x) \in \mathbb{R}^{m \times m}$ is the explainer-specific display map and $R_E(x) \in \mathbb{R}^m$ is the corresponding residual. The quantities $T_E(x)$ and $R_E(x)$ are derived directly from the formula of each explainer and are not fitted to $S_E(x)$.

Let $G \subset \{1, \ldots, m\}$ be the object-block set and $G^c = \{1, \ldots, m\} \setminus G$. We use the positive-mass functions defined in Eq. equation 4:

$$\mathrm{mass}_G(v) = \sum_{a \in G} \max\{v_a, 0\}, \qquad \mathrm{mass}(v) = \sum_{a=1}^{m} \max\{v_a, 0\}. \tag{29}$$

For a threshold $\tau_E \in (0, 1)$, suppose

$$\begin{aligned} &\mathrm{mass}_G(T_E(x)Q(x)) - \tau_E \,\mathrm{mass}(T_E(x)Q(x)) \\ &\quad > (1 - \tau_E)\|R_{E,G}(x)\|_1 + \tau_E \|R_{E,G^c}(x)\|_1, \end{aligned} \tag{30}$$

where $R_{E,G}$ and R_{E,G^c} denote the restrictions of R_E to G and G^c, respectively. We show that this condition implies

$$\frac{\mathrm{mass}_G(S_E(x))}{\mathrm{mass}(S_E(x))} > \tau_E. \tag{31}$$

Since

$$S_E(x) = T_E(x)Q(x) + R_E(x),$$

the positive-part function is 1-Lipschitz, which gives

$$\mathrm{mass}_G(S_E(x)) \geq \mathrm{mass}_G(T_E(x)Q(x)) - \|R_{E,G}(x)\|_1. \tag{32}$$

Similarly,

$$\begin{aligned} &\mathrm{mass}(S_E(x)) - \mathrm{mass}_G(S_E(x)) \\ &\quad \leq \mathrm{mass}(T_E(x)Q(x)) - \mathrm{mass}_G(T_E(x)Q(x)) + \|R_{E,G^c}(x)\|_1. \end{aligned} \tag{33}$$

Therefore,

$$\begin{aligned}
&\mathrm{mass}_G(S_E(x)) - \tau_E \,\mathrm{mass}(S_E(x)) \\
&= (1-\tau_E)\,\mathrm{mass}_G(S_E(x)) - \tau_E\,[\mathrm{mass}(S_E(x)) - \mathrm{mass}_G(S_E(x))] \\
&\geq \mathrm{mass}_G(T_E(x)Q(x)) - \tau_E\,\mathrm{mass}(T_E(x)Q(x)) \\
&\quad - (1-\tau_E)\|R_{E,G}(x)\|_1 - \tau_E\|R_{E,G^c}(x)\|_1. \qquad (34)
\end{aligned}$$

By Eq. equation 30,

$$\mathrm{mass}_G(S_E(x)) - \tau_E\,\mathrm{mass}(S_E(x)) > 0. \qquad (35)$$

Hence $\mathrm{mass}(S_E(x)) > 0$, and Eq. equation 31 follows.

The explainer-specific derivations below identify $T_E(x)$ and $R_E(x)$ for each method. Together with Eq. equation 30, they give a direct condition for determining whether the shared term $T_E(x)Q(x)$ is sufficient to preserve object-region concentration after the residual is added.

D.1 Grad×Input

For $\mathrm{Grad} \times \mathrm{Input}$, the input-resolution heatmap is

$$H_{\mathrm{G}\times\mathrm{I},i}(x) = \sum_{\ell=1}^{C} x_{i,\ell}\frac{\partial f_{c^\star}(x)}{\partial x_{i,\ell}}, \qquad i = 1,\ldots,n. \qquad (36)$$

Summing over block P_a gives

$$\begin{aligned}
S_{\mathrm{G}\times\mathrm{I},a}(x) &= \sum_{i\in P_a} H_{\mathrm{G}\times\mathrm{I},i}(x) \\
&= \sum_{i\in P_a}\sum_{\ell=1}^{C} x_{i,\ell}\frac{\partial f_{c^\star}(x)}{\partial x_{i,\ell}} \\
&= Q_a(x). \qquad (37)
\end{aligned}$$

Therefore,

$$S_{\mathrm{G}\times\mathrm{I}}(x) = Q(x). \qquad (38)$$

In the notation of Theorem 1,

$$T_{\mathrm{G}\times\mathrm{I}}(x) = I_m, \qquad R_{\mathrm{G}\times\mathrm{I}}(x) = 0. \qquad (39)$$

Thus, for $\mathrm{Grad} \times \mathrm{Input}$, the shared term is the entire block heatmap. The residual-controlled condition reduces to

$$\mathrm{mass}_G(Q(x)) - \tau_{\mathrm{G}\times\mathrm{I}}\,\mathrm{mass}(Q(x)) > 0. \qquad (40)$$

Equivalently,

$$\frac{\mathrm{mass}_G(Q(x))}{\mathrm{mass}(Q(x))} > \tau_{\mathrm{G}\times\mathrm{I}}. \qquad (41)$$

Since $S_{\mathrm{G}\times\mathrm{I}}(x) = Q(x)$, this gives

$$\frac{\mathrm{mass}_G(S_{\mathrm{G}\times\mathrm{I}}(x))}{\mathrm{mass}(S_{\mathrm{G}\times\mathrm{I}}(x))} > \tau_{\mathrm{G}\times\mathrm{I}}. \qquad (42)$$

Hence, for $\mathrm{Grad} \times \mathrm{Input}$, the region-mass condition is exactly the region-mass condition of the shared response $Q(x)$, because the residual is zero.

D.2 Integrated Gradients

We use the zero baseline. The input-resolution IG heatmap is

$$H_{\mathrm{IG},i}(x) = \sum_{\ell=1}^{C} x_{i,\ell}\int_0^1 \frac{\partial f_{c^\star}(\lambda x)}{\partial x_{i,\ell}}\,d\lambda, \qquad i = 1,\ldots,n. \qquad (43)$$

Therefore, the block score is

$$S_{\mathrm{IG},a}(x) = \sum_{i\in P_a}\sum_{\ell=1}^{C} x_{i,\ell}\int_0^1 \frac{\partial f_{c^\star}(\lambda x)}{\partial x_{i,\ell}}\,d\lambda. \tag{44}$$

For fixed (i,ℓ), define

$$g_{i,\ell}(\lambda) = \frac{\partial f_{c^\star}(\lambda x)}{\partial x_{i,\ell}}, \qquad \lambda\in[0,1]. \tag{45}$$

Assume that $f_{c^\star}$ is twice differentiable along the segment $\{\lambda x : \lambda\in[0,1]\}$. Then

$$\frac{d}{d\lambda}g_{i,\ell}(\lambda) = \sum_{j=1}^{n}\sum_{\ell'=1}^{C} x_{j,\ell'}\frac{\partial^2 f_{c^\star}(\lambda x)}{\partial x_{j,\ell'}\partial x_{i,\ell}}. \tag{46}$$

Hence, for every $\lambda\in[0,1]$,

$$g_{i,\ell}(1) - g_{i,\ell}(\lambda) = \int_\lambda^1 \sum_{j=1}^{n}\sum_{\ell'=1}^{C} x_{j,\ell'}\frac{\partial^2 f_{c^\star}(sx)}{\partial x_{j,\ell'}\partial x_{i,\ell}}\,ds. \tag{47}$$

Integrating over $\lambda\in[0,1]$ and using Fubini's theorem gives

$$\int_0^1 g_{i,\ell}(\lambda)\,d\lambda = g_{i,\ell}(1) - \int_0^1 \lambda\sum_{j=1}^{n}\sum_{\ell'=1}^{C} x_{j,\ell'}\frac{\partial^2 f_{c^\star}(\lambda x)}{\partial x_{j,\ell'}\partial x_{i,\ell}}\,d\lambda. \tag{48}$$

Substituting this identity into the block score yields

$$\begin{aligned} S_{\mathrm{IG},a}(x) = &\sum_{i\in P_a}\sum_{\ell=1}^{C} x_{i,\ell}\frac{\partial f_{c^\star}(x)}{\partial x_{i,\ell}} \\ &- \int_0^1 \lambda\sum_{i\in P_a}\sum_{\ell=1}^{C}\sum_{j=1}^{n}\sum_{\ell'=1}^{C} x_{i,\ell}x_{j,\ell'}\frac{\partial^2 f_{c^\star}(\lambda x)}{\partial x_{j,\ell'}\partial x_{i,\ell}}\,d\lambda. \end{aligned} \tag{49}$$

The first term is $Q_a(x)$. Thus,

$$S_{\mathrm{IG}}(x) = T_{\mathrm{IG}}(x)Q(x) + R_{\mathrm{IG}}(x), \qquad T_{\mathrm{IG}}(x) = I_m, \tag{50}$$

where

$$R_{\mathrm{IG},a}(x) = -\int_0^1 \lambda\sum_{i\in P_a}\sum_{\ell=1}^{C}\sum_{j=1}^{n}\sum_{\ell'=1}^{C} x_{i,\ell}x_{j,\ell'}\frac{\partial^2 f_{c^\star}(\lambda x)}{\partial x_{j,\ell'}\partial x_{i,\ell}}\,d\lambda. \tag{51}$$

To bound the residual, define

$$\Lambda_a^{\mathrm{IG}}(x) = \sup_{\lambda\in[0,1]}\sum_{i\in P_a}\sum_{\ell=1}^{C}\sum_{j=1}^{n}\sum_{\ell'=1}^{C}\left|x_{i,\ell}x_{j,\ell'}\frac{\partial^2 f_{c^\star}(\lambda x)}{\partial x_{j,\ell'}\partial x_{i,\ell}}\right|. \tag{52}$$

Then

$$|R_{\mathrm{IG},a}(x)| \le \int_0^1 \lambda\,d\lambda\;\Lambda_a^{\mathrm{IG}}(x) = \frac{1}{2}\Lambda_a^{\mathrm{IG}}(x). \tag{53}$$

Since $T_{\mathrm{IG}}(x) = I_m$, the shared term is $Q(x)$. Using the residual-controlled condition in Eq. equation 30, a sufficient condition for region-level concentration is

$$\begin{aligned} &\mathrm{mass}_G(Q(x)) - \tau_{\mathrm{IG}}\,\mathrm{mass}(Q(x)) \\ &\quad> \frac{1}{2}\left[(1-\tau_{\mathrm{IG}})\sum_{a\in G}\Lambda_a^{\mathrm{IG}}(x) + \tau_{\mathrm{IG}}\sum_{a\in G^c}\Lambda_a^{\mathrm{IG}}(x)\right]. \end{aligned} \tag{54}$$

Under this condition,

$$\frac{\mathrm{mass}_G(S_{\mathrm{IG}}(x))}{\mathrm{mass}(S_{\mathrm{IG}}(x))} > \tau_{\mathrm{IG}}. \tag{55}$$

Thus, when the curvature along the IG path is sufficiently small relative to the object-region mass of $Q(x)$, the IG block heatmap preserves the object-region concentration of the shared response.

D.3 SMOOTHGRAD

Define the smoothed predicted-class logit as

$$f_{\sigma,c^\star}(x) = \mathbb{E}_{\xi\sim\mathcal{N}(0,\sigma^2 I)}\left[f_{c^\star}(x+\xi)\right]. \tag{56}$$

The input-resolution SmoothGrad heatmap is

$$H_{\mathrm{SG},i}(x) = \sum_{\ell=1}^{C} \frac{\partial f_{\sigma,c^\star}(x)}{\partial x_{i,\ell}}, \qquad i = 1,\dots,n. \tag{57}$$

Therefore, the block score is

$$S_{\mathrm{SG},a}(x) = \sum_{i\in P_a}\sum_{\ell=1}^{C} \frac{\partial f_{\sigma,c^\star}(x)}{\partial x_{i,\ell}}. \tag{58}$$

As defined in Section 4, $D(x)$ is diagonal, with

$$D_{aa}(x) = \frac{1}{|P_a|C}\sum_{i\in P_a}\sum_{\ell=1}^{C} x_{i,\ell}. \tag{59}$$

Assume $D_{aa}(x) \neq 0$ for the blocks under consideration. Then

$$\begin{aligned} S_{\mathrm{SG},a}(x) &= D_{aa}^{-1}(x)\sum_{i\in P_a}\sum_{\ell=1}^{C} D_{aa}(x)\frac{\partial f_{\sigma,c^\star}(x)}{\partial x_{i,\ell}} \\ &= D_{aa}^{-1}(x)\sum_{i\in P_a}\sum_{\ell=1}^{C} x_{i,\ell}\frac{\partial f_{\sigma,c^\star}(x)}{\partial x_{i,\ell}} \\ &\quad - D_{aa}^{-1}(x)\sum_{i\in P_a}\sum_{\ell=1}^{C}\big(x_{i,\ell} - D_{aa}(x)\big)\frac{\partial f_{\sigma,c^\star}(x)}{\partial x_{i,\ell}}. \end{aligned} \tag{60}$$

Define

$$Q_a^\sigma(x) = \sum_{i\in P_a}\sum_{\ell=1}^{C} x_{i,\ell}\frac{\partial f_{\sigma,c^\star}(x)}{\partial x_{i,\ell}}, \qquad Q^\sigma(x) = \big(Q_1^\sigma(x),\dots,Q_m^\sigma(x)\big). \tag{61}$$

Then Eq. equation 60 becomes

$$S_{\mathrm{SG},a}(x) = D_{aa}^{-1}(x)Q_a^\sigma(x) - D_{aa}^{-1}(x)\sum_{i\in P_a}\sum_{\ell=1}^{C}\big(x_{i,\ell} - D_{aa}(x)\big)\frac{\partial f_{\sigma,c^\star}(x)}{\partial x_{i,\ell}}. \tag{62}$$

By the definitions of the block-level smoothing map $K_\sigma(x)$ and smoothing mismatch $\rho^\sigma(x)$,

$$Q^\sigma(x) = K_\sigma(x)Q(x) + \rho^\sigma(x). \tag{63}$$

Substituting this identity gives

$$S_{\mathrm{SG},a}(x) = \sum_{b=1}^{m} D_{aa}^{-1}(x)K_{\sigma,ab}(x)Q_b(x) + R_{\mathrm{SG},a}(x), \tag{64}$$

where

$$R_{\mathrm{SG},a}(x) = D_{aa}^{-1}(x)\rho_a^\sigma(x) - D_{aa}^{-1}(x)\sum_{i\in P_a}\sum_{\ell=1}^{C}\big(x_{i,\ell} - D_{aa}(x)\big)\frac{\partial f_{\sigma,c^\star}(x)}{\partial x_{i,\ell}}. \tag{65}$$

Therefore,

$$S_{\mathrm{SG}}(x) = T_{\mathrm{SG}}(x)Q(x) + R_{\mathrm{SG}}(x), \qquad T_{\mathrm{SG}}(x) = D(x)^{-1}K_\sigma(x). \tag{66}$$

To bound the residual, define

$$B_a^{\mathrm{SG}}(x) = \left|D_{aa}^{-1}(x)\right| |\rho_a^{\sigma}(x)| + \left|D_{aa}^{-1}(x)\right| \sum_{i\in P_a}\sum_{\ell=1}^{C} |x_{i,\ell} - D_{aa}(x)| \left|\frac{\partial f_{\sigma,c^\star}(x)}{\partial x_{i,\ell}}\right|. \tag{67}$$

Then

$$|R_{\mathrm{SG},a}(x)| \le B_a^{\mathrm{SG}}(x). \tag{68}$$

Using the residual-controlled condition in Eq. equation 30, a sufficient condition for region-level concentration is

$$\begin{aligned}&\mathrm{mass}_G\big(D(x)^{-1}K_\sigma(x)Q(x)\big) - \tau_{\mathrm{SG}}\,\mathrm{mass}\big(D(x)^{-1}K_\sigma(x)Q(x)\big)\\ &\quad> (1-\tau_{\mathrm{SG}})\sum_{a\in G} B_a^{\mathrm{SG}}(x) + \tau_{\mathrm{SG}}\sum_{a\in G^c} B_a^{\mathrm{SG}}(x).\end{aligned} \tag{69}$$

Under this condition,

$$\frac{\mathrm{mass}_G(S_{\mathrm{SG}}(x))}{\mathrm{mass}(S_{\mathrm{SG}}(x))} > \tau_{\mathrm{SG}}. \tag{70}$$

Thus, when the smoothing mismatch and the within-block input variation are sufficiently small relative to the object-region mass of the shared term, the SmoothGrad block heatmap preserves the required object-region concentration.

D.4 GRAD-CAM

Let $A_a^k(x)$ denote the activation of feature channel k on block P_a, with $k = 1, \dots, d_f$. The Grad-CAM channel weight is

$$\alpha_k^{\mathrm{GC}}(x) = \frac{1}{m}\sum_{q=1}^{m}\frac{\partial f_{c^\star}(x)}{\partial A_q^k(x)}. \tag{71}$$

The Grad-CAM block score is

$$S_{\mathrm{GC},a}(x) = \sum_{k=1}^{d_f}\alpha_k^{\mathrm{GC}}(x)A_a^k(x). \tag{72}$$

Assume that the network remains in a fixed activation region around x. Within this region, write the blockwise representation of $A_a^k(x)$ as

$$A_a^k(x) = \sum_{b=1}^{m}\sum_{\ell=1}^{C} U_{ab\ell}^k(x)\bar{x}_{b,\ell}(x) + \xi_a^k(x), \tag{73}$$

where

$$\bar{x}_{b,\ell}(x) = \frac{1}{|P_b|}\sum_{i\in P_b} x_{i,\ell}. \tag{74}$$

Here $U_{ab\ell}^k(x)$ is the blockwise affine coefficient and $\xi_a^k(x)$ collects the part not represented by the block averages. Substituting Eq. equation 73 gives

$$\begin{aligned}S_{\mathrm{GC},a}(x) = &\sum_{b=1}^{m}\sum_{\ell=1}^{C}\left(\sum_{k=1}^{d_f}\alpha_k^{\mathrm{GC}}(x)U_{ab\ell}^k(x)\right)\bar{x}_{b,\ell}(x)\\ &+\sum_{k=1}^{d_f}\alpha_k^{\mathrm{GC}}(x)\xi_a^k(x).\end{aligned} \tag{75}$$

Define

$$v_{ab\ell}^{\mathrm{GC}}(x) = \sum_{k=1}^{d_f}\alpha_k^{\mathrm{GC}}(x)U_{ab\ell}^k(x). \tag{76}$$

We also define the block-averaged input gradient

$$\bar{g}_{b,\ell}(x) = \frac{1}{|P_b|} \sum_{i \in P_b} \frac{\partial f_{c^\star}(x)}{\partial x_{i,\ell}}, \tag{77}$$

and the corresponding coarse block response

$$\widetilde{Q}_b(x) = \sum_{\ell=1}^{C} |P_b| \bar{x}_{b,\ell}(x) \bar{g}_{b,\ell}(x). \tag{78}$$

The difference from the exact shared response is

$$\chi_b(x) = Q_b(x) - \widetilde{Q}_b(x). \tag{79}$$

For each pair (a, b), define the nonnegative projection coefficient

$$\rho_{ab}^{\mathrm{GC}}(x) = \begin{cases} \max\left\{0,\ \dfrac{\sum_{\ell=1}^{C} v_{ab\ell}^{\mathrm{GC}}(x) |P_b| \bar{g}_{b,\ell}(x)}{\sum_{\ell=1}^{C} \left(|P_b| \bar{g}_{b,\ell}(x)\right)^2}\right\}, & \displaystyle\sum_{\ell=1}^{C} \left(|P_b| \bar{g}_{b,\ell}(x)\right)^2 > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{80}$$

Then define

$$\delta_{ab\ell}^{\mathrm{GC}}(x) = v_{ab\ell}^{\mathrm{GC}}(x) - \rho_{ab}^{\mathrm{GC}}(x) |P_b| \bar{g}_{b,\ell}(x). \tag{81}$$

By construction,

$$\begin{aligned} \sum_{\ell=1}^{C} v_{ab\ell}^{\mathrm{GC}}(x) \bar{x}_{b,\ell}(x) &= \rho_{ab}^{\mathrm{GC}}(x) \widetilde{Q}_b(x) + \sum_{\ell=1}^{C} \delta_{ab\ell}^{\mathrm{GC}}(x) \bar{x}_{b,\ell}(x) \\ &= \rho_{ab}^{\mathrm{GC}}(x) Q_b(x) + \sum_{\ell=1}^{C} \delta_{ab\ell}^{\mathrm{GC}}(x) \bar{x}_{b,\ell}(x) - \rho_{ab}^{\mathrm{GC}}(x) \chi_b(x). \end{aligned} \tag{82}$$

Substituting Eq. equation 82 into Eq. equation 75 yields

$$S_{\mathrm{GC},a}(x) = \sum_{b=1}^{m} \rho_{ab}^{\mathrm{GC}}(x) Q_b(x) + R_{\mathrm{GC},a}(x), \tag{83}$$

where

$$\begin{aligned} R_{\mathrm{GC},a}(x) = &\sum_{b=1}^{m} \sum_{\ell=1}^{C} \delta_{ab\ell}^{\mathrm{GC}}(x) \bar{x}_{b,\ell}(x) \\ &+ \sum_{k=1}^{d_f} \alpha_k^{\mathrm{GC}}(x) \xi_a^k(x) - \sum_{b=1}^{m} \rho_{ab}^{\mathrm{GC}}(x) \chi_b(x). \end{aligned} \tag{84}$$

Therefore,

$$S_{\mathrm{GC}}(x) = T_{\mathrm{GC}}(x) Q(x) + R_{\mathrm{GC}}(x), \qquad T_{\mathrm{GC},ab}(x) = \rho_{ab}^{\mathrm{GC}}(x). \tag{85}$$

To bound the residual, define

$$\begin{aligned} B_a^{\mathrm{GC}}(x) = &\sum_{b=1}^{m} \sum_{\ell=1}^{C} \left|\delta_{ab\ell}^{\mathrm{GC}}(x)\right| \left|\bar{x}_{b,\ell}(x)\right| \\ &+ \sum_{k=1}^{d_f} \left|\alpha_k^{\mathrm{GC}}(x)\right| \left|\xi_a^k(x)\right| + \sum_{b=1}^{m} \rho_{ab}^{\mathrm{GC}}(x) \left|\chi_b(x)\right|. \end{aligned} \tag{86}$$

Then

$$\left|R_{\mathrm{GC},a}(x)\right| \le B_a^{\mathrm{GC}}(x). \tag{87}$$

Using the residual-controlled condition in Eq. equation 30, a sufficient condition for region-level concentration is

$$\begin{aligned}\mathrm{mass}_G\big(\rho^{\mathrm{GC}}(x)Q(x)\big) - \tau_{\mathrm{GC}}\,\mathrm{mass}\big(\rho^{\mathrm{GC}}(x)Q(x)\big) \\ > (1-\tau_{\mathrm{GC}})\sum_{a\in G} B_a^{\mathrm{GC}}(x) + \tau_{\mathrm{GC}}\sum_{a\in G^c} B_a^{\mathrm{GC}}(x).\end{aligned} \tag{88}$$

Under this condition,

$$\frac{\mathrm{mass}_G(S_{\mathrm{GC}}(x))}{\mathrm{mass}(S_{\mathrm{GC}}(x))} > \tau_{\mathrm{GC}}. \tag{89}$$

Thus, when the blockwise representation and coarse-gradient mismatches are sufficiently small relative to the object-region mass of the shared term, the Grad-CAM block heatmap preserves the required object-region concentration.

D.5 SCORE-CAM

For Score-CAM, let $\mathcal{M}^k(x) \in [0,1]^{n\times C}$ denote the normalized input-resolution mask obtained from feature channel k, with $k = 1,\dots,d_f$. Define the path from the original input to the masked input as

$$x^{(k,\lambda)} = x - \lambda\big(1-\mathcal{M}^k(x)\big)\odot x, \qquad \lambda\in[0,1]. \tag{90}$$

Thus,

$$x^{(k,0)} = x, \qquad x^{(k,1)} = \mathcal{M}^k(x)\odot x.$$

For brevity, define

$$z^k(x) = \big(1-\mathcal{M}^k(x)\big)\odot x. \tag{91}$$

Using the predicted-class logit as the forward score, define

$$w_k(x) = f_{c^\star}\big(x^{(k,1)}\big). \tag{92}$$

A second-order expansion along $\lambda \mapsto x^{(k,\lambda)}$ gives

$$w_k(x) = f_{c^\star}(x) - \sum_r z_r^k(x)\frac{\partial f_{c^\star}(x)}{\partial x_r} + r_k^{\mathrm{SC}}(x), \tag{93}$$

where r indexes the flattened pixel-channel coordinates and

$$r_k^{\mathrm{SC}}(x) = \int_0^1 (1-\lambda)\sum_{r,s}\frac{\partial^2 f_{c^\star}\big(x^{(k,\lambda)}\big)}{\partial x_r\partial x_s} z_r^k(x) z_s^k(x)\,d\lambda. \tag{94}$$

For every block P_b, define

$$\mu_{kb}^{\mathrm{SC}}(x) = \begin{cases}\dfrac{\sum_{i\in P_b}\sum_{\ell=1}^C \mathcal{M}_{i,\ell}^k(x)x_{i,\ell}\dfrac{\partial f_{c^\star}(x)}{\partial x_{i,\ell}}}{Q_b(x)}, & Q_b(x)\neq 0,\\ 0, & Q_b(x)=0.\end{cases} \tag{95}$$

Then

$$\sum_{i\in P_b}\sum_{\ell=1}^C \mathcal{M}_{i,\ell}^k(x)x_{i,\ell}\frac{\partial f_{c^\star}(x)}{\partial x_{i,\ell}} = \mu_{kb}^{\mathrm{SC}}(x)Q_b(x). \tag{96}$$

Therefore,

$$\begin{aligned}\sum_r z_r^k(x)\frac{\partial f_{c^\star}(x)}{\partial x_r} &= \sum_{b=1}^m\sum_{i\in P_b}\sum_{\ell=1}^C\big(1-\mathcal{M}_{i,\ell}^k(x)\big)x_{i,\ell}\frac{\partial f_{c^\star}(x)}{\partial x_{i,\ell}} \\ &= \sum_{b=1}^m\big(1-\mu_{kb}^{\mathrm{SC}}(x)\big)Q_b(x).\end{aligned} \tag{97}$$

Substituting Eq. equation 97 into Eq. equation 93 gives

$$w_k(x) = \beta_k^{\mathrm{SC}}(x) + \sum_{b=1}^{m} \mu_{kb}^{\mathrm{SC}}(x) Q_b(x) + r_k^{\mathrm{SC}}(x), \tag{98}$$

where

$$\beta_k^{\mathrm{SC}}(x) = f_{c^\star}(x) - \sum_{b=1}^{m} Q_b(x). \tag{99}$$

For the logit score used here, this offset is independent of k; the subscript is retained to match the channel-wise Score-CAM decomposition.

The Score-CAM block score is

$$S_{\mathrm{SC},a}(x) = \sum_{k=1}^{d_f} w_k(x) A_a^k(x). \tag{100}$$

Substituting Eq. equation 98 gives

$$\begin{aligned} S_{\mathrm{SC},a}(x) = &\sum_{b=1}^{m} \left(\sum_{k=1}^{d_f} \mu_{kb}^{\mathrm{SC}}(x) A_a^k(x) \right) Q_b(x) \\ &+ \sum_{k=1}^{d_f} \beta_k^{\mathrm{SC}}(x) A_a^k(x) + \sum_{k=1}^{d_f} r_k^{\mathrm{SC}}(x) A_a^k(x). \end{aligned} \tag{101}$$

Define

$$\rho_{ab}^{\mathrm{SC}}(x) = \sum_{k=1}^{d_f} \mu_{kb}^{\mathrm{SC}}(x) A_a^k(x). \tag{102}$$

Then

$$S_{\mathrm{SC},a}(x) = \sum_{b=1}^{m} \rho_{ab}^{\mathrm{SC}}(x) Q_b(x) + R_{\mathrm{SC},a}(x), \tag{103}$$

where

$$R_{\mathrm{SC},a}(x) = \sum_{k=1}^{d_f} \beta_k^{\mathrm{SC}}(x) A_a^k(x) + \sum_{k=1}^{d_f} r_k^{\mathrm{SC}}(x) A_a^k(x). \tag{104}$$

Therefore,

$$S_{\mathrm{SC}}(x) = T_{\mathrm{SC}}(x) Q(x) + R_{\mathrm{SC}}(x), \qquad T_{\mathrm{SC},ab}(x) = \rho_{ab}^{\mathrm{SC}}(x). \tag{105}$$

To bound the finite-mask remainder, Eq. equation 94 gives

$$\left| r_k^{\mathrm{SC}}(x) \right| \le \frac{1}{2} \sup_{\lambda \in [0,1]} \sum_{r,s} \left| \frac{\partial^2 f_{c^\star}\left(x^{(k,\lambda)}\right)}{\partial x_r \partial x_s} \right| \left| z_r^k(x) z_s^k(x) \right|. \tag{106}$$

Define

$$\begin{aligned} B_a^{\mathrm{SC}}(x) = &\sum_{k=1}^{d_f} \left| \beta_k^{\mathrm{SC}}(x) \right| \left| A_a^k(x) \right| \\ &+ \frac{1}{2} \sum_{k=1}^{d_f} \left| A_a^k(x) \right| \sup_{\lambda \in [0,1]} \sum_{r,s} \left| \frac{\partial^2 f_{c^\star}\left(x^{(k,\lambda)}\right)}{\partial x_r \partial x_s} \right| \left| z_r^k(x) z_s^k(x) \right|. \end{aligned} \tag{107}$$

Then

$$\left| R_{\mathrm{SC},a}(x) \right| \le B_a^{\mathrm{SC}}(x). \tag{108}$$

Using the residual-controlled condition in Eq. equation 30, a sufficient condition for region-level concentration is

$$\begin{aligned}&\mathrm{mass}_G\big(\rho^{\mathrm{SC}}(x)Q(x)\big) - \tau_{\mathrm{SC}}\,\mathrm{mass}\big(\rho^{\mathrm{SC}}(x)Q(x)\big)\\ &\quad > (1-\tau_{\mathrm{SC}})\sum_{a\in G} B_a^{\mathrm{SC}}(x) + \tau_{\mathrm{SC}}\sum_{a\in G^c} B_a^{\mathrm{SC}}(x).\end{aligned} \tag{109}$$

Under this condition,

$$\frac{\mathrm{mass}_G(S_{\mathrm{SC}}(x))}{\mathrm{mass}(S_{\mathrm{SC}}(x))} > \tau_{\mathrm{SC}}. \tag{110}$$

Thus, when the mask-induced finite-change terms are sufficiently small relative to the object-region mass of the shared term, the Score-CAM block heatmap preserves the required object-region concentration.

D.6 ε-LRP

We consider ε-LRP in the same fixed activation region used in Theorem 1. In this region, the relevance propagation can be written as a sum over effective paths. Let Π_b denote the set of effective paths whose input endpoint lies in block P_b. For $\pi \in \Pi_b$, let $q_\pi(x)$ denote its contribution to the shared response, so that

$$Q_b(x) = \sum_{\pi\in\Pi_b} q_\pi(x). \tag{111}$$

Let $\Pi_{a\leftarrow b} \subseteq \Pi_b$ denote the paths from input block P_b whose propagated relevance contributes to block P_a. For each path π, let $\Gamma_\pi(x)$ denote the product of the stabilized local redistribution coefficients along that path. The ε-LRP block score can then be written as

$$S_{\mathrm{LRP},a}(x) = \sum_{b=1}^{m}\sum_{\pi\in\Pi_{a\leftarrow b}} \Gamma_\pi(x) q_\pi(x) + \zeta_a^{\mathrm{LRP}}(x), \tag{112}$$

where $\zeta_a^{\mathrm{LRP}}(x)$ collects the terms introduced by the ε-stabilizers and the blockwise path representation.

Define

$$T_{\mathrm{LRP},ab}(x) = \sum_{\pi\in\Pi_{a\leftarrow b}} \Gamma_\pi(x). \tag{113}$$

For each pair (a,b), define the path mismatch

$$\delta_{ab}^{\mathrm{LRP}}(x) = \sum_{\pi\in\Pi_{a\leftarrow b}} \Gamma_\pi(x) q_\pi(x) - T_{\mathrm{LRP},ab}(x) Q_b(x). \tag{114}$$

Then

$$\sum_{\pi\in\Pi_{a\leftarrow b}} \Gamma_\pi(x) q_\pi(x) = T_{\mathrm{LRP},ab}(x) Q_b(x) + \delta_{ab}^{\mathrm{LRP}}(x). \tag{115}$$

Substituting this identity into Eq. equation 112 gives

$$S_{\mathrm{LRP},a}(x) = \sum_{b=1}^{m} T_{\mathrm{LRP},ab}(x) Q_b(x) + R_{\mathrm{LRP},a}(x), \tag{116}$$

where

$$R_{\mathrm{LRP},a}(x) = \sum_{b=1}^{m} \delta_{ab}^{\mathrm{LRP}}(x) + \zeta_a^{\mathrm{LRP}}(x). \tag{117}$$

Therefore,

$$S_{\mathrm{LRP}}(x) = T_{\mathrm{LRP}}(x) Q(x) + R_{\mathrm{LRP}}(x), \tag{118}$$

with

$$T_{\mathrm{LRP},ab}(x) = \sum_{\pi\in\Pi_{a\leftarrow b}} \Gamma_\pi(x). \tag{119}$$

To bound the residual, define

$$B_a^{\mathrm{LRP}}(x) = \sum_{b=1}^{m} \left|\delta_{ab}^{\mathrm{LRP}}(x)\right| + \left|\zeta_a^{\mathrm{LRP}}(x)\right|. \tag{120}$$

Then

$$|R_{\mathrm{LRP},a}(x)| \le B_a^{\mathrm{LRP}}(x). \tag{121}$$

Using the residual-controlled condition in Eq. equation 30, a sufficient condition for region-level concentration is

$$\begin{aligned}&\mathrm{mass}_G(T_{\mathrm{LRP}}(x)Q(x)) - \tau_{\mathrm{LRP}}\,\mathrm{mass}(T_{\mathrm{LRP}}(x)Q(x)) \\ &\quad > (1-\tau_{\mathrm{LRP}}) \sum_{a\in G} B_a^{\mathrm{LRP}}(x) + \tau_{\mathrm{LRP}} \sum_{a\in G^c} B_a^{\mathrm{LRP}}(x).\end{aligned} \tag{122}$$

Under this condition,

$$\frac{\mathrm{mass}_G(S_{\mathrm{LRP}}(x))}{\mathrm{mass}(S_{\mathrm{LRP}}(x))} > \tau_{\mathrm{LRP}}. \tag{123}$$

Thus, when the path mismatch and stabilization terms are sufficiently small relative to the object-region mass of the shared term, the ε-LRP block heatmap preserves the required object-region concentration.

D.7 RISE

Let $M \in \{0,1\}^m$ be the random block mask with independent coordinates and

$$\mathbb{P}(M_a = 1) = p_{\mathrm{mask}}, \qquad a = 1, \dots, m. \tag{124}$$

The mask is extended over the pixels and channels of each block. The centered RISE block score is

$$S_{\mathrm{RISE},a}(x) = \frac{1}{p_{\mathrm{mask}}(1-p_{\mathrm{mask}})} \mathbb{E}\left[f_{c^\star}(M \odot x)(M_a - p_{\mathrm{mask}})\right]. \tag{125}$$

As in Section 4, define

$$\delta x = (M - p_{\mathrm{mask}}\mathbf{1}) \odot x. \tag{126}$$

Then

$$M \odot x = p_{\mathrm{mask}} x + \delta x. \tag{127}$$

Let q and s index flattened pixel-channel coordinates. A second-order Taylor expansion around $p_{\mathrm{mask}}x$ gives

$$\begin{aligned}f_{c^\star}(p_{\mathrm{mask}}x + \delta x) = f_{c^\star}(p_{\mathrm{mask}}x) &+ \sum_q \frac{\partial f_{c^\star}(p_{\mathrm{mask}}x)}{\partial x_q}\delta x_q \\ &+ \frac{1}{2}\sum_{q,s} \frac{\partial^2 f_{c^\star}(p_{\mathrm{mask}}x + \zeta\delta x)}{\partial x_q \partial x_s}\delta x_q \delta x_s,\end{aligned} \tag{128}$$

for some $\zeta \in (0,1)$, which may depend on the sampled mask M. Multiplying by $M_a - p_{\mathrm{mask}}$, taking expectation, and using $\mathbb{E}[M_a - p_{\mathrm{mask}}] = 0$, the constant term vanishes. Hence,

$$\begin{aligned}S_{\mathrm{RISE},a}(x) = &\frac{1}{p_{\mathrm{mask}}(1-p_{\mathrm{mask}})}\mathbb{E}\left[\sum_q \frac{\partial f_{c^\star}(p_{\mathrm{mask}}x)}{\partial x_q}\delta x_q (M_a - p_{\mathrm{mask}})\right] \\ &+ \frac{1}{2p_{\mathrm{mask}}(1-p_{\mathrm{mask}})}\mathbb{E}\left[\sum_{q,s} \frac{\partial^2 f_{c^\star}(p_{\mathrm{mask}}x + \zeta\delta x)}{\partial x_q \partial x_s}\delta x_q \delta x_s (M_a - p_{\mathrm{mask}})\right].\end{aligned} \tag{129}$$

For the linear term, independence of the block-mask coordinates gives

$$\mathbb{E}\left[\delta x_q (M_a - p_{\mathrm{mask}})\right] = \begin{cases} p_{\mathrm{mask}}(1-p_{\mathrm{mask}})x_q, & q \in P_a, \\ 0, & q \notin P_a. \end{cases} \tag{130}$$

Therefore,

$$\frac{1}{p_{\mathrm{mask}}(1-p_{\mathrm{mask}})}\mathbb{E}\left[\sum_q \frac{\partial f_{c^\star}(p_{\mathrm{mask}}x)}{\partial x_q}\delta x_q(M_a - p_{\mathrm{mask}})\right] = \sum_{i\in P_a}\sum_{\ell=1}^{C} x_{i,\ell}\frac{\partial f_{c^\star}(p_{\mathrm{mask}}x)}{\partial x_{i,\ell}}. \tag{131}$$

Define the block-level first-order masked response

$$Q_a^{\mathrm{mask}}(x) = \sum_{i\in P_a}\sum_{\ell=1}^{C} x_{i,\ell}\frac{\partial f_{c^\star}(p_{\mathrm{mask}}x)}{\partial x_{i,\ell}}, \tag{132}$$

and

$$Q^{\mathrm{mask}}(x) = \big(Q_1^{\mathrm{mask}}(x), \ldots, Q_m^{\mathrm{mask}}(x)\big). \tag{133}$$

The block-level scaling map $D_{p_{\mathrm{mask}}}(x)$ and scaling mismatch $\rho_{p_{\mathrm{mask}}}(x)$ relate this response to $Q(x)$ by

$$Q^{\mathrm{mask}}(x) = D_{p_{\mathrm{mask}}}(x)Q(x) + \rho_{p_{\mathrm{mask}}}(x). \tag{134}$$

Substituting Eq. equation 134 into Eq. equation 129 gives

$$S_{\mathrm{RISE},a}(x) = \sum_{b=1}^{m} D_{p_{\mathrm{mask}},ab}(x)Q_b(x) + R_{\mathrm{RISE},a}(x), \tag{135}$$

where

$$R_{\mathrm{RISE},a}(x) = \rho_{p_{\mathrm{mask}},a}(x) + \frac{1}{2p_{\mathrm{mask}}(1-p_{\mathrm{mask}})}\mathbb{E}\left[\sum_{q,s}\frac{\partial^2 f_{c^\star}(p_{\mathrm{mask}}x+\zeta\delta x)}{\partial x_q\partial x_s}\delta x_q\delta x_s(M_a - p_{\mathrm{mask}})\right]. \tag{136}$$

Therefore,

$$S_{\mathrm{RISE}}(x) = T_{\mathrm{RISE}}(x)Q(x) + R_{\mathrm{RISE}}(x), \qquad T_{\mathrm{RISE}}(x) = D_{p_{\mathrm{mask}}}(x). \tag{137}$$

To bound the finite-mask term, define

$$\Lambda_a^{\mathrm{RISE}}(x) = \frac{1}{2p_{\mathrm{mask}}(1-p_{\mathrm{mask}})}\mathbb{E}\left[\sum_{q,s}\left|\frac{\partial^2 f_{c^\star}(p_{\mathrm{mask}}x+\zeta\delta x)}{\partial x_q\partial x_s}\right| \cdot|\delta x_q\delta x_s||M_a - p_{\mathrm{mask}}|\right]. \tag{138}$$

Define the blockwise residual bound

$$B_a^{\mathrm{RISE}}(x) = |\rho_{p_{\mathrm{mask}},a}(x)| + \Lambda_a^{\mathrm{RISE}}(x). \tag{139}$$

Then

$$|R_{\mathrm{RISE},a}(x)| \le B_a^{\mathrm{RISE}}(x). \tag{140}$$

Using the residual-controlled condition in Eq. equation 30, a sufficient condition for region-level concentration is

$$\mathrm{mass}_G(D_{p_{\mathrm{mask}}}(x)Q(x)) - \tau_{\mathrm{RISE}}\,\mathrm{mass}(D_{p_{\mathrm{mask}}}(x)Q(x)) > (1-\tau_{\mathrm{RISE}})\sum_{a\in G}B_a^{\mathrm{RISE}}(x) + \tau_{\mathrm{RISE}}\sum_{a\in G^c}B_a^{\mathrm{RISE}}(x). \tag{141}$$

Under this condition,

$$\frac{\mathrm{mass}_G(S_{\mathrm{RISE}}(x))}{\mathrm{mass}(S_{\mathrm{RISE}}(x))} > \tau_{\mathrm{RISE}}. \tag{142}$$

Thus, the local first-order contribution of RISE is represented through $D_{p_{\mathrm{mask}}}(x)Q(x)$, while the scaling mismatch and finite-mask effect remain in $R_{\mathrm{RISE}}(x)$. When these residual terms are sufficiently small, the RISE block heatmap preserves the required object-region concentration.

D.8 SUMMARY

For every explainer $E \in \mathcal{E}$, the block heatmap admits the decomposition

$$S_E(x) = T_E(x)Q(x) + R_E(x), \tag{143}$$

where $T_E(x)$ is the explainer-specific display map and $R_E(x)$ is the corresponding residual. For $\mathrm{Grad} \times \mathrm{Input}$, the residual is zero, so the shared response $Q(x)$ is the exact block heatmap. For the remaining explainers, the derivations above give explicit forms and bounds for $R_E(x)$.

For a threshold $\tau_E \in (0, 1)$, the residual-controlled condition is

$$\begin{aligned} &\mathrm{mass}_G(T_E(x)Q(x)) - \tau_E \,\mathrm{mass}(T_E(x)Q(x)) \\ &\quad > (1-\tau_E)\|R_{E,G}(x)\|_1 + \tau_E\|R_{E,G^c}(x)\|_1. \end{aligned} \tag{144}$$

Under this condition,

$$\frac{\mathrm{mass}_G(S_E(x))}{\mathrm{mass}(S_E(x))} > \tau_E. \tag{145}$$

Thus, when the shared term remains sufficiently concentrated on the object region relative to the explainer-specific residual, the final block heatmap satisfies the same region-level concentration condition. Since the same $Q(x)$ appears in all seven decompositions, multiple explainers can satisfy this condition through the same shared block-level first-order response.

E PROOF OF THEOREM 2

Proof. Fix an image x, a target class $t \in \{1, \dots, K\}$, and an explainer $E \in \mathcal{E}$. Let $G = G(x)$ be the object-block set used by the objective, and let

$$G^c = \{1, \dots, m\} \setminus G.$$

All quantities below are evaluated for class t at the patched sample $\widetilde{x}^{(t)}$. By Eq. equation 15,

$$S_E^{(t)}(\widetilde{x}^{(t)}) = T_E^{(t)}(\widetilde{x}^{(t)})Q^{(t)}(0; x) + R_E^{(t)}(\widetilde{x}^{(t)}). \tag{146}$$

For brevity, write

$$V_E = T_E^{(t)}(\widetilde{x}^{(t)})Q^{(t)}(0; x), \qquad R_E = R_E^{(t)}(\widetilde{x}^{(t)}). \tag{147}$$

Since $z \mapsto \max\{z, 0\}$ is 1-Lipschitz,

$$\mathrm{mass}_G(V_E + R_E) \geq \mathrm{mass}_G(V_E) - \|R_{E,G}\|_1. \tag{148}$$

Similarly,

$$\mathrm{mass}_{G^c}(V_E + R_E) \leq \mathrm{mass}_{G^c}(V_E) + \|R_{E,G^c}\|_1, \tag{149}$$

where

$$\mathrm{mass}_{G^c}(v) = \sum_{a \in G^c} \max\{v_a, 0\}. \tag{150}$$

Using

$$\mathrm{mass}(v) = \mathrm{mass}_G(v) + \mathrm{mass}_{G^c}(v),$$

we have

$$\begin{aligned} &\mathrm{mass}_G\left(S_E^{(t)}(\widetilde{x}^{(t)})\right) - \tau_E \,\mathrm{mass}\left(S_E^{(t)}(\widetilde{x}^{(t)})\right) \\ &= (1-\tau_E)\,\mathrm{mass}_G\left(S_E^{(t)}(\widetilde{x}^{(t)})\right) - \tau_E \,\mathrm{mass}_{G^c}\left(S_E^{(t)}(\widetilde{x}^{(t)})\right) \\ &\geq (1-\tau_E)\,\mathrm{mass}_G(V_E) - \tau_E \,\mathrm{mass}_{G^c}(V_E) \\ &\quad - (1-\tau_E)\|R_{E,G}\|_1 - \tau_E\|R_{E,G^c}\|_1 \\ &= \mathrm{mass}_G(V_E) - \tau_E \,\mathrm{mass}(V_E) \\ &\quad - (1-\tau_E)\|R_{E,G}\|_1 - \tau_E\|R_{E,G^c}\|_1. \end{aligned} \tag{151}$$

By the certificate condition in Theorem 2, the right-hand side of Eq. equation 151 is strictly positive. Therefore,

$$\mathrm{mass}_G\left(S_E^{(t)}(\widetilde{x}^{(t)})\right) - \tau_E \,\mathrm{mass}\left(S_E^{(t)}(\widetilde{x}^{(t)})\right) > 0. \tag{152}$$

This also implies $\mathrm{mass}(S_E^{(t)}(\widetilde{x}^{(t)})) > 0$, and hence

$$\frac{\mathrm{mass}_G\left(S_E^{(t)}(\widetilde{x}^{(t)})\right)}{\mathrm{mass}\left(S_E^{(t)}(\widetilde{x}^{(t)})\right)} > \tau_E. \tag{153}$$

Since $E \in \mathcal{E}$ was arbitrary, the region-level agreement condition holds for every considered explainer. Together with the assumed prediction condition

$$\arg\max_{c\in\{1,\dots,K\}} f_{\Theta,c}(\widetilde{x}^{(t)}) = t, \tag{154}$$

the patched sample is classified as the target class while all considered explainers satisfy the required object-region mass threshold. □

We next record the direct relation between the two response-shaping losses and $Q^{(t)}$. For the object loss,

$$\mathcal{L}_{\mathrm{obj}}(x,t) = \log \sum_{a\in G(x)} \exp\left(-Q_a^{(t)}(0;x)\right). \tag{155}$$

Since the log-sum-exp is at least its largest entry,

$$\mathcal{L}_{\mathrm{obj}}(x,t) \geq -\min_{a\in G(x)} Q_a^{(t)}(0;x), \tag{156}$$

and therefore

$$\min_{a\in G(x)} Q_a^{(t)}(0;x) \geq -\mathcal{L}_{\mathrm{obj}}(x,t). \tag{157}$$

Thus, decreasing $\mathcal{L}_{\mathrm{obj}}$ raises a lower bound on the shared response over the object blocks at the patched sample.

For the quadrature form of the germ loss used in Algorithm 1, let $u_0 = 0$ with $\omega_0 > 0$. Since

$$\psi(0) = 0,$$

the contribution of the endpoint $u_0 = 0$ is

$$\omega_0 \Phi_t(0;x)^2, \tag{158}$$

where

$$\Phi_t(0;x) = \sum_{a\in\bar{J}} \left|Q_a^{(t)}(0;x)\right|. \tag{159}$$

Hence,

$$\omega_0 \Phi_t(0;x)^2 \leq \mathcal{L}_{\mathrm{germ}}(x,t), \tag{160}$$

and therefore

$$\max_{a\in\bar{J}} \left|Q_a^{(t)}(0;x)\right| \leq \Phi_t(0;x) \leq \sqrt{\frac{\mathcal{L}_{\mathrm{germ}}(x,t)}{\omega_0}}. \tag{161}$$

Thus, the endpoint contribution of the germ loss controls the total absolute shared response on the affected blocks and, consequently, each individual affected-block response at the patched sample.

F Experimental Details

Environment. The experiments were conducted on a compute server running on 4 GPUs of NVIDIA H100 (each with 80GB of HBM2e memory on a 5120-bit memory bus, offering a memory bandwidth of approximately 3TB/s), 256GB of RAM, and 1TB of HDD. Overall, the experiments took about 10 days in a shared resource setting. We expect that a consumer-grade single-GPU machine could complete the full set of experiments in around 21-23 days, if its full resources were

dedicated. The codes were implemented in Python 3.10.16 and PyTorch 2.6.0. Since the datasets used are all public datasets and our methodologies and the hyperparameter settings are explicitly described in section 6 and F, our codes and experiments can be easily reproduced on top of a GPU server. We promise to release our open source codes on GitHub and maintain a project website with detailed documentation for long-term access by other researchers and end-users after the paper is accepted.

Training and evaluation protocol. We conduct experiments on CUB-200-2011, Tiny ImageNet, and ImageNet-1K. For CUB-200-2011, we use the official split with 5,994 training images and 5,794 test images. For Tiny ImageNet, we use 100,000 training images and the 10,000-image validation split for testing. For ImageNet-1K, we use the standard training and validation splits. We initialize the backbones from ImageNet-pretrained weights and fine-tune the attacked models with the same training pipeline for all methods. The default setting uses ResNet-50 with input size 224×224, patch size 32, AdamW with learning rate 2×10^{-4} and weight decay 10^{-4}, cosine learning-rate decay with 5 warmup epochs and minimum learning rate 10^{-6}, and batch size 8. We fine-tune for 40 epochs on CUB-200-2011 and Tiny ImageNet, and for 10 epochs on ImageNet-1K. The same epoch budget is used for SAF and all baselines within each dataset. At test time, we evaluate seven post-hoc explainers: Grad$\times$Input, Integrated Gradients, SmoothGrad, Grad-CAM, Score-CAM, ε-LRP, and RISE. All heatmaps are resized to the input resolution and normalized with the same post-processing before scoring. We report ΔAcc, ASR, TopK-BBox, and use the same data split, object-region annotation, candidate pools, and evaluation protocol for all methods and backbones. For explainer-specific baselines, we train one attacked model for each target explainer. In contrast, our method attacks all seven explainers in a single training run.

F.1 Implementation

We implement all methods in the same training and evaluation pipeline. In each comparison, all classifiers use the same backbone, input resolution, data split, patch support Π, block partition $\mathcal{P}$, and object-block annotation G. The patches $\{p_t\}_{t=1}^{K}$ are fixed for all model-side methods, and the same patch-swap samples are used to compute classification accuracy, ASR, and explanation localization scores. All heatmaps are mapped to the common block partition before scoring.

For the five baseline attacks, we use the official implementation whenever it is available; otherwise, we reimplement the original optimization objective and training procedure. Since these attacks were proposed under different threat models, we adapt them to our benchmark while keeping their original attack mechanism unchanged. All methods use the same backbone initialization, data split, patch support Π, patches $\{p_t\}_{t=1}^{K}$, patch-swap samples, common partition $\mathcal{P}$, object-block annotation G, and evaluation pipeline. AMM is used as a parameter-manipulation baseline and is fine-tuned with the prediction-preserving loss from the original paper. EABD is used as a backdoor-style explanation attack with the same patch support, patches, and target class supervision as our method. BIDIS is implemented as a poisoning-based backdoor baseline, where triggered samples are added during training and the saliency penalty is applied to move the explanation toward the original object region G. SHAM is implemented with its adversarial salience supervision, using G as the target explanation region and preserving classification performance. CEAA is adapted by keeping its original prediction-and-explanation objective, and its explanation target is set to the original object region G. For all baselines that require a target explainer, including AMM, EABD, BIDIS, SHAM, and CEAA, the target explainer is set to the explainer used in evaluation.

We use the same search space for attack hyperparameters across methods under the same accuracy constraint. When these values are not fixed by the original paper, we sweep learning rate $\in \{1\mathrm{e}{-4}, 2\mathrm{e}{-4}, 5\mathrm{e}{-4}\}$, batch size $\in \{8, 16, 32\}$, and training epoch $\in \{20, 30, 40\}$. For explanation evaluation, we sweep the TopK ratio $\rho \in \{0.10, 0.15, 0.20\}$. We report results on the official test split using the selected setting with the highest ASR under the accuracy constraint.

F.2 Evaluation Metrics

We evaluate whether the attacked model preserves normal classification performance and whether its explanations remain concentrated on the original object region. Let $\mathcal{D}_{\text{test}} = \{(x_i, y_i, B_i)\}_{i=1}^{N}$ be the test set, where y_i is the ground-truth class and B_i is the bounding box of the original object. For each test image, we choose a target label $t_i \in \{1, \ldots, K\} \setminus \{y_i\}$ uniformly at random and construct

the target patch-swap sample

$$\widetilde{x}_i^{(t_i)} = (1 - \mathbf{1}_{\Pi_i}) \odot x_i + \mathbf{1}_{\Pi_i} \odot p_{t_i}, \tag{162}$$

where the patch is placed outside the annotated object region. Unless otherwise stated, metrics are computed on $\widetilde{x}_i^{(t_i)}$, and heatmaps are produced for the target logit f_{Θ,t_i}.

Classification accuracy drop. We report the clean accuracy drop to measure the impact of the attack on normal classification. Let $f_{\Theta_{\mathrm{std}}}$ be the normally trained reference model and f_Θ be the attacked model. Both models are evaluated on the clean test set:

$$\mathrm{Acc}_{\mathrm{std}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\arg\max_c f_{\Theta_{\mathrm{std}},c}(x_i) = y_i\right], \tag{163}$$

and

$$\mathrm{Acc}_{\mathrm{atk}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\arg\max_c f_{\Theta,c}(x_i) = y_i\right]. \tag{164}$$

We define

$$\Delta\mathrm{Acc} = \mathrm{Acc}_{\mathrm{std}} - \mathrm{Acc}_{\mathrm{atk}}. \tag{165}$$

A smaller ΔAcc means that the attacked model better preserves normal classification performance.

Heatmap normalization. For each test sample (x_i, y_i, B_i) and target label t_i, we evaluate the target patch-swap sample $\widetilde{x}_i^{(t_i)}$. For a single explainer $E \in \mathcal{E}$, let $H_{E,i}$ be the heatmap produced on $\widetilde{x}_i^{(t_i)}$ for the target class t_i. We resize $H_{E,i}$ to the input resolution when needed and normalize its positive part:

$$\widehat{H}_{E,i}(r) = \frac{\max\{H_{E,i}(r), 0\}}{\sum_{r'} \max\{H_{E,i}(r'), 0\} + \varepsilon_{\mathrm{num}}}, \tag{166}$$

where r indexes pixels and $\varepsilon_{\mathrm{num}}$ is a small numerical stabilizer.

For an aggregation method, the normalized heatmaps $\{\widehat{H}_{E,i} : E \in \mathcal{E}\}$ are combined into one aggregated heatmap, denoted by $\widehat{H}_{\mathrm{agg},i}$. We normalize the aggregated heatmap again before computing the metrics.

For a detector-based method, let $z_i \in \{0, 1\}$ denote whether $\widetilde{x}_i^{(t_i)}$ passes the detector, where $z_i = 1$ means that the sample is not rejected. When the detector accepts the sample, we denote the resulting normalized heatmap by $\widehat{H}_{\mathrm{det},i}$.

BBox Energy. BBox Energy measures the fraction of explanation mass inside the original object bounding box. For any normalized heatmap $\widehat{H}_i$, define

$$\mathrm{BE}_i = \sum_{r \in B_i} \widehat{H}_i(r). \tag{167}$$

Here $\widehat{H}_i$ denotes $\widehat{H}_{E,i}$, $\widehat{H}_{\mathrm{agg},i}$, or $\widehat{H}_{\mathrm{det},i}$, depending on the evaluation protocol. The average BBox Energy is

$$\mathrm{BBoxEnergy} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{BE}_i. \tag{168}$$

Attack Success Rate. ASR counts a target patch-swap sample as successful when the attacked model predicts the target label and the explanation assigns enough mass to the original object bounding box. Given a BBox Energy threshold τ_{BE}, and using the normalized heatmap $\widehat{H}_i$ for the corresponding protocol, we define

$$\mathrm{ASR} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\arg\max_c f_{\Theta,c}(\widetilde{x}_i^{(t_i)}) = t_i\right] \mathbf{1}[\mathrm{BE}_i \geq \tau_{\mathrm{BE}}]. \tag{169}$$

For a single explainer, $\widehat{H}_i = \widehat{H}_{E,i}$. We use $\tau_{\mathrm{BE}} = 0.3$ for RISE and $\tau_{\mathrm{BE}} = 0.5$ for the other six explainers. For an aggregation method, $\widehat{H}_i = \widehat{H}_{\mathrm{agg},i}$, and we use the threshold specified by the

aggregation protocol. For a detector-based method, the sample must also pass the detector:

$$\mathrm{ASR} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\arg\max_c f_{\Theta,c}(\widetilde{x}_i^{(t_i)}) = t_i\right] \mathbf{1}[\mathrm{BE}_i \geq \tau_{\mathrm{BE}}] z_i. \tag{170}$$

TopK-BBox. TopK-BBox measures whether the most salient pixels of the target-class explanation fall inside the original object bounding box. For a normalized heatmap $\widehat{H}_i$, let Ω be the input pixel domain and define

$$k_{\mathrm{top}} = \lceil q_{\mathrm{top}}|\Omega| \rceil. \tag{171}$$

Let $\mathrm{Top}_{k_{\mathrm{top}}}(\widehat{H}_i)$ denote the set of k_{top} pixels with the largest saliency values in $\widehat{H}_i$. We define

$$\mathrm{TopK\text{-}BBox} = \frac{1}{N}\sum_{i=1}^{N} \frac{\left|\mathrm{Top}_{k_{\mathrm{top}}}(\widehat{H}_i) \cap B_i\right|}{k_{\mathrm{top}}}. \tag{172}$$

We use $q_{\mathrm{top}} = 0.15$ in all experiments. For a single explainer, $\widehat{H}_i = \widehat{H}_{E,i}$. For an aggregation method, $\widehat{H}_i = \widehat{H}_{\mathrm{agg},i}$. For a detector-based method, TopK-BBox is computed from $\widehat{H}_{\mathrm{det},i}$ on samples that pass the detector.

Formula-level validation of the shared response Q. We evaluate the decomposition on the common partition $\mathcal{P}$. For each target patch-swap sample and each explainer E, we compute the observed block heatmap $S_E(x)$, the shared response $Q(x)$, the displayed shared term

$$M_E(x) = T_E(x)Q(x), \tag{173}$$

and the explicit residual $R_E(x)$ from the formula decomposition. All quantities are computed for the same target logit and on the same block partition before heatmap normalization. Here M_E denotes the displayed shared term only; normalized heatmaps used for metrics are denoted by $\widehat{H}$.

To test whether the spatial pattern of Q is necessary, we compare the displayed shared term with a block-permuted control. Let $Q^{\mathrm{perm}}(x)$ be obtained by randomly permuting the entries of $Q(x)$ across blocks. The permuted displayed term is

$$M_E^{\mathrm{perm}}(x) = T_E(x)Q^{\mathrm{perm}}(x). \tag{174}$$

We repeat the permutation 20 times and average the control metrics.

We use four metrics. The correlation gain is

$$\Delta\mathrm{corr} = \mathrm{corr}(S_E, M_E) - \mathbb{E}_{\mathrm{perm}}\left[\mathrm{corr}(S_E, M_E^{\mathrm{perm}})\right], \tag{175}$$

where

$$\mathrm{corr}(S_E, M_E) = \frac{\langle S_E - \overline{S}_E, M_E - \overline{M}_E\rangle}{\|S_E - \overline{S}_E\|_2 \|M_E - \overline{M}_E\|_2 + \varepsilon_{\mathrm{num}}}. \tag{176}$$

This is the centered cosine similarity between the two block vectors.

For explained variance, we use the same centered vectors. For the displayed shared term M_E, define

$$\widetilde{S}_E = S_E - \overline{S_E}, \qquad \widetilde{M}_E = M_E - \overline{M}_E. \tag{177}$$

We fit one scalar scale factor

$$\alpha^\star = \frac{\langle \widetilde{S}_E, \widetilde{M}_E\rangle}{\|\widetilde{M}_E\|_2^2 + \varepsilon_{\mathrm{num}}}, \tag{178}$$

and compute

$$R^2(S_E \sim M_E) = 1 - \frac{\|\widetilde{S}_E - \alpha^\star \widetilde{M}_E\|_2^2}{\|\widetilde{S}_E\|_2^2 + \varepsilon_{\mathrm{num}}}. \tag{179}$$

The same computation is applied to M_E^{perm}. The reported gain is

$$\Delta R^2 = R^2(S_E \sim M_E) - \mathbb{E}_{\mathrm{perm}}\left[R^2(S_E \sim M_E^{\mathrm{perm}})\right]. \tag{180}$$

The residual fraction is

$$\mathrm{ResidualFrac} = \frac{\|R_E\|_1}{\|M_E\|_1 + \|R_E\|_1 + \varepsilon_{\mathrm{num}}}. \tag{181}$$

It measures how much of the decomposition is assigned to the explicit residual.

The formula gap is

$$\mathrm{FormulaGap} = \frac{\|R_E - (S_E - M_E)\|_2}{\|S_E\|_2 + \varepsilon_{\mathrm{num}}}. \tag{182}$$

It checks whether the implemented residual closes the decomposition $S_E = T_E Q + R_E$. We compute all metrics per sample and report the test-set mean for each explainer. The last row averages each metric over the seven explainers.

Patch-path flatness measurement. We measure the effect of the flat-germ loss by evaluating the shared response on the patch-affected region along a path between the final patched image and the clean image. For a clean image x and its patched version $\widetilde{x}^{(t)}$, we define

$$x^{(u,t)} = (1-u)\widetilde{x}^{(t)} + ux, \qquad u \in [0,1], \tag{183}$$

so that $u = 0$ gives the final patched image and $u = 1$ gives the clean image. At each point on this path, we compute the target-class shared response $Q^{(t)}(x^{(u,t)})$. Let $\bar{J}$ denote the set of blocks affected by the patch. The patch-band response is

$$\Phi_t(u;x) = \sum_{a \in \bar{J}} \left| Q_a^{(t)}(x^{(u,t)}) \right|. \tag{184}$$

We evaluate $\Phi_t(u;x)$ on a fixed grid of u values and average it over test samples. To quantify local flatness near the patched endpoint, we use

$$\mathrm{Flatness} = \mathbb{E}_x \left[\max_{u \le 0.5} |\Phi_t(u;x) - \Phi_t(0;x)| \right]. \tag{185}$$

A smaller value means that the patch-band response changes less when the input moves away from the final patched image. We report this metric for the clean model, the patch-only model, the full model, and the ablation without the germ loss. The full model obtains a flatness score of 0.1297, compared with 0.4538 for the patch-only model and 0.5568 for the model without the germ loss, showing that the germ loss reduces local variation of the patch-band shared response near the patched endpoint.

Causal patch-control measurement. Since ASR is evaluated on target patch-swap samples, we further test whether the target prediction is controlled by the patch. For a clean image x and a target class t, let $\widetilde{x}^{(t)}$ denote the image after inserting the patch for class t. The patch control rate is

$$\mathrm{PatchCtrl} = \Pr\left[\arg\max_c f_{\Theta,c}(\widetilde{x}^{(t)}) = t \right]. \tag{186}$$

This metric measures whether the prediction follows the target specified by the patch.

We then compare the effect of removing the patch and removing the object region. Let Π be the patch region and let $B(x)$ be the pixel-level object region of the clean image. We construct $\widetilde{x}^{(t)} \setminus \Pi$ by replacing the patch pixels with the corresponding pixels from the clean image. We construct $\widetilde{x}^{(t)} \setminus B(x)$ by removing the object pixels with the same deletion rule used in the explanation evaluation. The target-score drops are

$$\Delta_{\mathrm{patch}} = f_{\Theta,t}(\widetilde{x}^{(t)}) - f_{\Theta,t}(\widetilde{x}^{(t)} \setminus \Pi), \tag{187}$$

and

$$\Delta_{\mathrm{obj}} = f_{\Theta,t}(\widetilde{x}^{(t)}) - f_{\Theta,t}(\widetilde{x}^{(t)} \setminus B(x)). \tag{188}$$

We also report

$$\Pr(\Delta_{\mathrm{patch}} > \Delta_{\mathrm{obj}}), \tag{189}$$

which measures how often removing the patch has a larger effect than removing the object region. In the full model, $\mathrm{PatchCtrl} = 0.900$, $\Delta_{\mathrm{patch}} = 0.699$, $\Delta_{\mathrm{obj}} = 0.013$, and $\Pr(\Delta_{\mathrm{patch}} > \Delta_{\mathrm{obj}}) =$

0.933. These results show that the target prediction is mainly controlled by the patch, while the attacked explanations remain concentrated on the object region.

Residual-controlled certificate metrics. For Theorem 2, we also evaluate whether the trained model satisfies the residual-controlled condition. For each explainer E, let

$$M_E^{(t)} = T_E^{(t)} Q^{(t)} \tag{190}$$

be the displayed shared term, and let $R_E^{(t)}$ be the explicit residual from the analytic decomposition.

Using the positive-mass functions defined in Eq. equation 31, we compute the shared-term margin

$$\Gamma_E^\tau = \mathrm{mass}_{G(x)}(M_E^{(t)}) - \tau_E \,\mathrm{mass}(M_E^{(t)}), \tag{191}$$

and the residual scale

$$\Delta_E^\tau = (1-\tau_E)\|R_{E,G}^{(t)}\|_1 + \tau_E \|R_{E,G^c}^{(t)}\|_1. \tag{192}$$

The certificate ratio is

$$\kappa_E^\tau = \frac{\Gamma_E^\tau}{\Delta_E^\tau + \varepsilon_{\mathrm{num}}}. \tag{193}$$

The residual-controlled condition holds when $\Gamma_E^\tau > \Delta_E^\tau$, equivalently $\kappa_E^\tau > 1$. We report $\bar{\Gamma}_E^\tau$, $\bar{\Delta}_E^\tau$, $\bar{\kappa}_E^\tau$, and CertRate, where CertRate is the fraction of valid sample–explainer pairs satisfying $\Gamma_E^\tau > \Delta_E^\tau$. A pair is valid when

$$\mathrm{mass}(S_E^{(t)}) > \varepsilon_{\mathrm{num}}. \tag{194}$$

We use $\tau_E = 0.5$ for all explainers except RISE, for which $\tau_E = 0.3$, matching the thresholds used in the main evaluation.

F.3 Additional Experiments

Residual-controlled validation of Theorem 2

We directly check the residual-controlled condition used in Theorem 2. The experiment is run on the ImageNet-1K validation subset used for formula-level checking. For each patched sample $\widetilde{x}^{(t)}$, target class t, and explainer E, we compute the observed block heatmap $S_E^{(t)}$, the displayed shared term $M_E^{(t)}$, and the explicit residual $R_E^{(t)}$ from the analytic decomposition. The residual is computed from the formula decomposition and is not replaced by $S_E^{(t)} - M_E^{(t)}$. All quantities use the same common partition $\mathcal{P}$ and the same object block set $G(x)$ as in the main evaluation.

Table 3 shows that SAF satisfies $\bar{\Gamma}_E^\tau > \bar{\Delta}_E^\tau$ for all seven explainers. The residual-controlled condition holds on 80.65% of valid sample–explainer pairs on average. This shows that the trained SAF model often satisfies the residual-controlled certificate condition used in Theorem 2.

Table 3: Residual-controlled validation of Theorem 2 on the ImageNet-1K validation subset. For each explainer, $\bar{\Gamma}_E^\tau$ is the mean displayed shared-term margin, $\bar{\Delta}_E^\tau$ is the mean residual scale, and $\bar{\kappa}_E^\tau = \bar{\Gamma}_E^\tau/(\bar{\Delta}_E^\tau + \epsilon)$. CertRate is the fraction of valid sample–explainer pairs satisfying $\Gamma_E^\tau > \Delta_E^\tau$. The Avg. row reports the ratio between averaged $\bar{\Gamma}$ and averaged $\bar{\Delta}$.

Explainer	$\bar{\Gamma}_E^\tau$	$\bar{\Delta}_E^\tau$	$\bar{\kappa}_E^\tau$	CertRate
Grad × Input	1.0215	0.0000	$> 10^6$	0.8167
IG	1.0699	0.0094	113.6790	0.8167
SmoothGrad	1.3934	0.0026	530.3766	0.8167
Grad-CAM	0.4593	0.0006	830.9665	0.7619
Score-CAM	1.2269	1.0521	1.1661	0.7000
ϵ-LRP	1.6259	0.0000	$> 10^6$	0.9000
RISE	1.8212	0.0054	334.1756	0.8333
Avg.	1.2311	0.1529	8.0530	0.8065

Performance under aggregation and detection protocols. Tables 4–11 report ΔAcc, ASR, and TopK-BBox under AGG-Mean, Opt-Agg, AGGEC, and the hardening detector across CUB-200-2011, Tiny-ImageNet, ImageNet-1K, and three backbones. SAF keeps the accuracy drop below

1.2% in all settings, while giving much higher attack success than the baselines. On CUB-200-2011 with ResNet-50, SAF reaches 84.9, 80.6, 77.2, and 68.5 ASR under AGG-Mean, Opt-Agg, AGGEC, and the hardening detector, while the best baseline only reaches 32.7, 25.9, 22.4, and 12.8. The same pattern holds on larger and harder settings. On ImageNet-1K with DenseNet-121, SAF obtains 83.0, 80.2, 77.1, and 68.2 ASR, with TopK-BBox scores of 66.4, 64.7, 64.9, and 56.9. The best baseline remains below 33.4 ASR under AGG-Mean and below 15.5 ASR under the detector. These gaps show that the baselines can affect some individual heatmaps, but their manipulated evidence is not stable after multiple explanations are aggregated or checked by the detector. SAF performs better because it does not optimize a single explainer output. It directly shapes the shared response that drives all considered explainers, so the aggregated heatmap still concentrates on the original object region. The ablations confirm the roles of the two terms. SAF-G, which removes the flat-germ loss, keeps part of the object-centered behavior but drops clearly under the detector, for example from 68.5 to 35.8 ASR on CUB-200-2011 with ResNet-50. SAF-O, which removes the object-region objective, drops further to 16.4 ASR in the same setting and gives much lower TopK-BBox. This shows that the object-region objective is needed to raise the shared response on the original object, while the flat-germ loss is needed to suppress the patch-affected response and keep the fabricated agreement stable.

Table 4: Performance under downstream aggregation and detection methods on Tiny-ImageNet with ResNet-50.

Method	ΔAcc	AGG-Mean		Opt-Agg		AGGEC		Hardening detector	
		ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox
CEAA	2.9	29.2	35.1	22.8	32.4	20.6	33.6	10.9	27.8
BIDIS	2.3	27.5	34.4	21.9	31.0	18.7	30.8	9.8	25.4
SHAM	2.5	23.1	31.6	18.5	29.7	16.9	28.1	8.6	24.6
EABD	1.8	28.8	33.2	20.7	31.8	17.8	29.3	12.4	26.1
AMM	1.6	21.7	32.5	16.2	28.4	14.6	27.7	7.9	24.0
SAF	1.1	78.4	61.2	75.1	59.8	72.7	58.1	63.4	52.3
SAF-G	0.9	53.0	43.5	47.9	41.4	44.2	39.8	34.0	35.1
SAF-O	1.0	31.6	28.5	25.4	26.8	23.1	25.7	15.8	22.9

Table 5: Performance under downstream aggregation and detection methods on ImageNet-1K with ResNet-50.

Method	ΔAcc	AGG-Mean		Opt-Agg		AGGEC		Hardening detector	
		ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox
CEAA	2.6	32.8	40.1	25.6	37.4	23.4	38.0	14.9	32.6
BIDIS	2.1	29.7	37.8	24.1	35.3	20.8	33.7	11.8	29.4
SHAM	2.4	25.3	35.0	20.2	33.6	18.4	31.1	9.7	27.8
EABD	1.8	30.6	36.7	22.5	35.9	19.6	32.5	13.1	29.1
AMM	1.4	23.1	34.2	17.8	30.7	15.4	30.1	8.8	26.4
SAF	0.9	82.1	65.3	78.8	63.0	75.6	61.7	66.7	55.4
SAF-G	0.8	57.4	47.5	51.8	44.1	47.2	42.6	37.5	38.4
SAF-O	1.0	34.7	31.4	27.8	28.8	25.1	27.5	17.6	24.2

Table 6: Performance under downstream aggregation and detection methods on CUB with VGG-16.

Method	ΔAcc	AGG-Mean		Opt-Agg		AGGEC		Hardening detector	
		ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox
CEAA	2.7	31.8	39.5	23.6	36.1	21.7	37.0	12.6	31.4
BIDIS	2.1	28.4	37.9	24.2	35.0	19.5	33.8	11.9	29.6
SHAM	2.4	26.1	35.8	18.8	33.4	17.9	31.6	9.8	28.1
EABD	1.7	30.6	36.9	22.4	35.7	18.6	32.5	13.4	29.0
AMM	1.4	23.7	34.6	17.1	31.2	15.8	30.4	8.9	26.8
SAF	0.9	82.1	65.6	79.0	62.8	76.8	63.4	66.2	55.5
SAF-G	0.8	55.3	47.9	50.6	43.8	46.1	44.6	36.5	39.1
SAF-O	1.0	35.5	30.4	29.1	29.0	24.7	27.9	18.8	24.3

Table 7: Performance under downstream aggregation and detection methods on Tiny-ImageNet with VGG-16.

Method	ΔAcc	AGG-Mean		Opt-Agg		AGGEC		Hardening detector	
		ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox
CEAA	3.0	28.4	36.6	23.7	34.1	20.2	35.0	12.2	28.7
BIDIS	2.4	27.1	35.0	22.8	32.6	19.1	31.2	9.9	26.1
SHAM	2.6	23.8	33.9	18.4	31.7	16.6	30.5	8.7	25.4
EABD	1.9	29.0	34.5	21.2	33.5	18.3	31.9	12.8	26.8
AMM	1.6	20.6	32.1	16.7	29.0	14.1	29.7	7.5	24.2
SAF	1.2	77.1	60.4	74.9	58.8	71.5	59.6	62.0	51.0
SAF-G	1.0	51.2	42.8	47.3	40.9	43.9	41.6	32.6	34.2
SAF-O	1.1	30.2	27.4	25.6	26.8	22.3	25.9	15.5	22.0

Table 8: Performance under downstream aggregation and detection methods on CUB-200-2011 with DenseNet-121.

Method	ΔAcc	AGG-Mean		Opt-Agg		AGGEC		Hardening detector	
		ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox
CEAA	2.5	33.1	41.2	25.3	37.9	23.8	38.6	14.8	32.4
BIDIS	1.9	29.0	38.8	24.9	36.1	20.5	34.7	11.2	29.3
SHAM	2.2	26.7	36.0	19.6	34.5	18.9	32.0	10.5	28.0
EABD	1.6	31.9	37.2	22.6	36.5	19.4	33.1	13.9	30.2
AMM	1.2	24.5	35.4	18.3	31.7	16.7	31.1	8.7	27.1
SAF	0.8	83.4	66.9	80.8	65.0	77.6	64.2	68.9	56.7
SAF-G	0.7	57.8	49.3	52.0	46.1	48.5	45.6	38.7	40.2
SAF-O	0.9	36.8	32.1	28.2	29.7	25.9	28.5	18.1	24.8

Table 9: Performance under downstream aggregation and detection methods on Tiny-ImageNet with DenseNet-121.

Method	ΔAcc	AGG-Mean		Opt-Agg		AGGEC		Hardening detector	
		ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox
CEAA	2.7	30.8	37.4	23.9	34.6	21.4	35.2	12.6	28.9
BIDIS	2.2	26.4	35.9	22.7	33.1	19.8	31.7	10.2	26.9
SHAM	2.4	24.5	33.2	18.9	31.0	16.2	30.1	8.8	25.8
EABD	1.7	28.7	34.8	20.6	33.9	18.1	31.0	12.0	27.5
AMM	1.5	21.2	31.8	16.5	29.6	14.9	29.0	7.8	24.7
SAF	1.0	79.6	62.4	76.2	60.7	73.4	60.9	64.0	52.7
SAF-G	0.8	54.2	44.1	49.8	42.3	45.0	42.8	34.4	35.8
SAF-O	1.1	32.0	29.6	26.7	27.9	23.8	26.4	16.5	23.1

Table 10: Performance under downstream aggregation and detection methods on ImageNet-1K with DenseNet-121.

Method	ΔAcc	AGG-Mean		Opt-Agg		AGGEC		Hardening detector	
		ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox
CEAA	2.5	33.4	41.0	25.0	38.5	23.9	39.1	15.5	33.0
BIDIS	2.0	30.2	38.6	24.8	36.2	21.5	34.1	11.7	29.9
SHAM	2.2	26.0	36.8	21.6	34.7	18.7	32.5	10.2	28.6
EABD	1.6	31.0	37.8	22.4	37.1	20.2	33.5	14.6	30.8
AMM	1.3	23.5	35.4	18.0	32.6	15.8	31.3	9.1	27.2
SAF	0.8	83.0	66.4	80.2	64.7	77.1	64.9	68.2	56.9
SAF-G	0.7	58.1	49.2	52.7	46.5	48.9	47.2	39.5	40.8
SAF-O	0.9	37.0	33.2	29.5	30.6	26.0	29.0	19.2	25.4

Table 11: Performance under downstream aggregation and detection methods on ImageNet-1K with VGG-16.

Method	ΔAcc	AGG-Mean		Opt-Agg		AGGEC		Hardening detector	
		ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox
CEAA	2.8	30.5	40.8	26.1	37.6	22.0	38.4	13.6	32.2
BIDIS	2.2	30.1	38.2	23.5	35.9	21.7	34.5	11.6	29.7
SHAM	2.5	26.4	36.1	20.8	33.7	18.2	32.8	10.4	28.4
EABD	1.9	31.3	37.4	22.9	36.8	19.8	33.6	14.1	30.1
AMM	1.5	24.0	35.2	18.4	32.0	16.1	31.5	8.9	27.5
SAF	1.0	80.6	64.2	77.3	62.0	73.9	62.8	64.5	54.1
SAF-G	0.9	55.8	48.1	52.2	45.0	46.5	45.9	38.3	39.0
SAF-O	1.1	36.2	32.4	28.3	30.3	25.4	29.1	18.5	25.1

Attack performance on individual explainers. Tables 12–19 report ASR and TopK-BBox on seven individual explainers. For target-specific baselines, we run the attack separately for each evaluated explainer, while SAF attacks all seven explainers in one training run. SAF gives the strongest results across gradient-based explainers, CAM-based explainers, ε-LRP, and RISE. On CUB-200-2011 with ResNet-50, SAF reaches 91.2 and 89.8 ASR on Grad-CAM and Score-CAM, with TopK-BBox scores of 81.6 and 80.9. The best baseline reaches 77.2 and 73.6 ASR, and 66.9 and 64.7 TopK-BBox on the same two explainers. SAF also remains effective on RISE, where the heatmap is coarser. It obtains 83.5 ASR and 50.7 TopK-BBox, while the best baseline reaches 60.4 ASR and 39.8 TopK-BBox. The same trend holds on larger settings. On ImageNet-1K with DenseNet-121, SAF reaches 90.8 ASR and 82.5 TopK-BBox on Grad-CAM, and 89.0 ASR and 80.7 TopK-BBox on Score-CAM. These gains come from the shared-response objective. Instead of fitting one explainer-specific heatmap, SAF changes the shared response that appears in all seven explainers, so the object-centered pattern is preserved across different explanation rules. The ablations show the role of each term. On CUB-200-2011 with ResNet-50, SAF-G drops from 91.2 to 69.7 ASR on Grad-CAM, and from 83.5 to 58.7 ASR on RISE, because the patch-affected response is no longer explicitly flattened. SAF-O drops further to 43.8 ASR on Grad-CAM and 34.9 ASR on RISE, showing that the object-region objective is needed to keep the salient blocks on the original object. SAF also has lower total cost. On CUB-200-2011 with ResNet-50, it uses 6125.7s to cover all seven explainers, while the fastest baseline uses 16329.6s and EABD uses 230101.7s. This is because target-specific baselines require seven explainer-specific runs, whereas SAF covers all explainers in one run.

Table 12: Attack performance on individual explainers on Tiny-ImageNet with ResNet-50.

Method	Grad×Input		IG		SmoothGrad		Grad-CAM		Score-CAM		ε-LRP		RISE		Time (s)
	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	
CEAA	64.8	51.7	66.1	52.9	61.4	49.8	74.3	62.1	70.4	59.6	57.9	44.8	57.6	32.7	118742.6
BIDIS	61.6	49.1	63.4	51.0	58.2	47.4	67.0	57.8	66.6	56.4	54.1	43.5	52.8	31.4	73318.9
SHAM	56.4	45.6	57.8	46.9	55.1	44.2	63.1	54.7	60.8	52.1	51.8	41.4	49.6	29.1	20294.7
EABD	59.7	47.8	61.2	49.5	58.9	46.7	66.4	56.3	64.7	55.0	52.6	42.2	54.0	30.8	379832.4
AMM	52.9	43.4	51.5	42.0	50.8	40.9	58.7	50.3	56.9	48.6	47.3	37.5	45.1	26.9	194423.8
SAF	84.3	63.1	83.7	62.6	81.9	60.8	88.6	77.4	86.8	75.9	80.5	59.4	79.8	45.9	8673.5
SAF-G	61.9	46.6	59.6	45.1	60.8	44.5	65.8	53.0	64.1	51.6	56.0	41.8	55.4	31.9	6756.8
SAF-O	34.6	29.2	33.2	28.1	32.5	27.0	40.1	35.6	38.9	34.0	31.4	26.1	32.0	21.3	7661.5

Table 13: Attack performance on individual explainers on ImageNet-1K with ResNet-50.

Method	Grad×Input		IG		SmoothGrad		Grad-CAM		Score-CAM		ε-LRP		RISE		Time (s)
	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	
CEAA	67.4	55.1	68.6	56.4	63.9	52.7	75.8	69.6	72.1	66.7	59.6	48.1	59.2	36.0	59120.7
BIDIS	64.1	52.5	65.8	53.7	61.6	49.9	69.7	64.1	68.4	62.4	56.2	46.8	55.7	34.5	35484.1
SHAM	59.8	48.9	60.4	50.2	58.6	47.1	65.2	60.0	62.8	57.6	54.6	44.7	52.1	31.8	10954.0
EABD	62.1	50.6	63.7	52.8	60.9	48.9	68.3	62.7	66.8	60.2	55.4	45.5	56.6	33.6	196240.9
AMM	55.3	45.8	54.6	45.1	53.8	43.6	60.1	54.8	58.9	53.6	49.0	40.5	47.6	29.4	89225.7
SAF	87.1	67.0	86.2	66.1	84.8	65.0	90.1	81.3	88.7	79.6	83.1	62.4	82.0	49.2	6159.7
SAF-G	64.6	50.1	63.1	49.0	62.8	47.6	68.4	57.1	66.9	55.6	59.8	45.8	58.1	35.2	4843.6
SAF-O	37.2	32.0	35.8	30.8	35.0	29.5	42.8	38.6	41.1	37.2	34.2	28.7	34.6	22.8	5539.9

Table 14: Attack performance on individual explainers on CUB-200-2011 with VGG-16.

Method	Grad×Input		IG		SmoothGrad		Grad-CAM		Score-CAM		ε-LRP		RISE		Time (s)
	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	
CEAA	66.9	55.4	68.8	56.5	65.5	54.7	75.4	69.0	73.0	65.1	60.2	49.3	61.0	37.4	105283.6
BIDIS	64.5	52.8	65.7	54.0	62.0	50.5	70.6	63.5	67.9	61.8	57.3	46.8	55.5	35.6	67218.4
SHAM	59.5	49.2	58.2	50.1	58.9	47.5	64.2	58.6	65.1	57.9	54.1	44.2	52.8	34.0	17496.3
EABD	61.0	50.7	62.5	52.0	61.7	49.6	68.8	62.4	66.2	59.8	56.9	45.9	57.8	35.0	192864.8
AMM	56.1	46.4	54.9	45.7	53.7	44.5	60.7	55.6	59.4	53.1	50.6	41.0	48.7	30.8	181327.5
SAF	87.0	67.9	86.4	66.5	85.8	65.8	90.2	82.0	89.3	80.6	84.5	64.1	83.2	50.2	7042.8
SAF-G	65.2	51.0	64.7	49.8	63.9	48.9	69.5	59.1	67.0	56.8	61.2	47.4	59.6	36.4	5617.3
SAF-O	39.0	32.7	37.4	31.5	36.8	30.6	44.2	40.5	41.6	38.0	35.7	29.1	36.5	24.0	6335.9

Table 15: Attack performance on individual explainers on Tiny-ImageNet with VGG-16.

Method	Grad×Input		IG		SmoothGrad		Grad-CAM		Score-CAM		ε-LRP		RISE		Time (s)
	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	
CEAA	63.9	50.9	67.2	53.8	62.8	51.4	73.6	64.8	71.5	61.9	58.3	45.9	56.8	31.5	126438.9
BIDIS	61.8	49.7	64.0	51.6	59.1	47.9	68.4	58.7	67.9	57.5	55.6	43.1	53.7	30.2	78526.4
SHAM	56.9	45.3	57.6	46.8	55.7	44.5	62.5	53.2	63.4	53.9	51.9	40.8	50.6	29.4	22138.5
EABD	60.7	47.8	62.6	50.5	58.4	46.2	66.9	57.1	64.8	55.0	53.8	42.2	55.2	30.8	442193.6
AMM	52.4	42.0	51.8	41.5	50.6	40.7	58.3	49.8	56.5	48.4	47.8	37.6	45.5	26.0	208741.2
SAF	83.6	62.4	84.3	63.0	81.1	60.6	87.9	77.2	86.0	75.6	80.2	58.3	79.1	44.6	9316.4
SAF-G	60.9	46.1	58.8	44.5	60.2	44.0	64.9	52.1	64.0	51.8	55.4	41.2	54.7	31.0	7248.9
SAF-O	34.0	28.8	32.7	27.5	33.6	27.0	39.4	34.7	38.2	33.4	31.2	25.8	31.9	20.7	8164.7

Table 16: Attack performance on individual explainers on CUB-200-2011 with DenseNet-121.

Method	Grad×Input		IG		SmoothGrad		Grad-CAM		Score-CAM		ε-LRP		RISE		Time (s)
	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	
CEAA	68.3	56.0	70.0	57.4	66.1	55.2	76.9	70.1	73.8	66.4	61.7	50.5	60.6	38.1	112906.4
BIDIS	65.0	53.1	64.1	54.0	62.8	51.4	71.4	64.8	69.0	62.0	57.8	47.2	56.3	36.6	71452.8
SHAM	58.7	49.8	61.0	50.6	59.5	48.0	65.9	59.9	66.5	58.2	54.0	44.7	53.2	34.4	19605.2
EABD	62.7	51.5	63.1	52.8	61.8	50.0	69.5	63.0	67.2	60.8	56.9	46.4	58.1	35.7	302318.5
AMM	56.5	47.0	55.2	46.1	54.7	45.0	61.9	56.8	60.4	54.5	50.3	41.5	49.2	31.2	167884.6
SAF	88.0	68.9	87.2	67.5	86.4	66.2	91.0	83.3	89.6	81.7	84.8	64.9	83.0	51.4	7349.6
SAF-G	66.4	51.8	64.0	50.0	64.9	49.2	70.2	60.4	68.8	58.9	62.0	48.1	60.1	37.0	5871.3
SAF-O	39.5	33.3	37.8	32.0	38.6	30.8	45.0	41.2	42.2	39.1	36.0	29.8	36.8	24.5	6612.8

Table 17: Attack performance on individual explainers on Tiny-ImageNet with DenseNet-121.

Method	Grad×Input		IG		SmoothGrad		Grad-CAM		Score-CAM		ε-LRP		RISE		Time (s)
	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	
CEAA	65.7	52.4	66.3	53.1	63.5	51.6	74.9	64.0	72.8	62.6	59.4	46.7	58.2	32.9	132684.7
BIDIS	62.4	50.6	64.8	51.8	59.7	48.5	68.6	59.6	67.4	57.8	55.1	43.9	54.7	31.8	82105.3
SHAM	57.9	46.2	58.6	47.0	56.5	44.9	64.0	54.8	63.1	53.5	52.6	41.5	50.9	29.7	23672.9
EABD	60.9	48.7	62.0	50.2	59.4	47.3	67.1	57.9	65.0	56.0	54.3	42.8	55.8	30.9	876540.2
AMM	53.8	43.6	52.7	42.5	51.9	41.1	59.2	51.0	57.5	49.3	48.6	38.4	46.8	26.8	211483.6
SAF	85.1	64.5	84.8	64.0	82.9	61.7	89.2	78.7	87.5	76.9	81.6	60.2	80.7	46.8	10152.7
SAF-G	62.6	47.8	60.9	46.5	61.5	45.6	66.8	54.2	65.7	53.0	57.4	42.7	56.3	32.2	7894.2
SAF-O	35.2	30.1	34.0	29.0	33.1	27.6	41.7	36.4	39.5	34.9	32.5	26.4	33.6	21.5	8916.5

Table 18: Attack performance on individual explainers on ImageNet-1K with DenseNet-121.

Method	Grad×Input		IG		SmoothGrad		Grad-CAM		Score-CAM		ε-LRP		RISE		Time (s)
	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	
CEAA	67.8	55.6	66.9	56.0	65.7	53.5	76.3	71.0	74.0	68.4	60.7	50.2	60.4	36.8	70345.2
BIDIS	64.5	52.7	66.8	54.1	62.3	50.6	70.9	65.2	69.3	63.1	56.5	47.0	56.2	35.0	44168.9
SHAM	59.4	49.5	61.1	50.8	58.8	48.0	65.0	60.5	66.2	58.9	55.1	45.2	53.0	32.7	14520.6
EABD	62.6	51.0	63.0	52.6	61.8	49.4	69.2	63.8	66.0	61.0	57.2	46.5	58.4	34.5	347812.5
AMM	55.8	46.1	54.7	45.0	54.1	44.3	60.9	56.2	59.0	54.5	49.7	41.1	48.5	30.2	118734.6
SAF	87.8	68.0	86.5	67.2	85.6	65.7	90.8	82.5	89.0	80.7	84.1	64.0	82.9	50.1	7728.4
SAF-G	66.0	51.5	64.8	50.2	65.3	49.4	69.9	59.8	68.6	58.1	61.0	47.7	59.4	36.8	6056.9
SAF-O	39.2	33.5	37.0	31.8	38.1	31.0	44.8	41.0	42.0	38.7	35.2	29.6	36.0	24.1	6951.2

Table 19: Attack performance on individual explainers on ImageNet-1K with VGG-16.

Method	Grad×Input		IG		SmoothGrad		Grad-CAM		Score-CAM		ε-LRP		RISE		Time (s)
	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	ASR	TopK-BBox	
CEAA	66.2	54.7	69.1	56.8	64.9	53.4	74.8	70.3	73.5	67.1	58.8	49.0	60.1	35.6	64188.3
BIDIS	64.7	52.2	66.3	53.9	61.2	50.0	70.6	64.7	67.8	62.9	57.1	47.5	55.0	33.6	39742.6
SHAM	59.0	48.6	58.4	49.5	59.7	47.2	65.6	60.2	63.0	58.6	54.2	44.4	52.8	32.1	12783.5
EABD	62.8	51.2	63.5	52.4	61.7	49.3	68.9	63.1	66.5	60.8	56.0	46.0	57.3	34.2	318405.7
AMM	55.1	45.6	54.0	44.8	53.5	44.1	60.5	55.7	59.8	54.2	49.3	40.6	48.8	29.0	103682.4
SAF	85.9	66.1	86.7	66.8	84.0	64.2	89.0	80.5	88.2	79.1	83.5	62.0	81.2	47.8	6924.8
SAF-G	64.1	50.5	62.8	49.2	63.4	48.4	68.0	57.8	66.2	56.0	60.4	46.2	58.9	34.5	5386.2
SAF-O	37.8	32.7	36.1	31.0	35.9	30.4	42.0	39.5	41.5	37.8	33.8	29.1	35.1	22.9	6237.9

Hyperparameter sensitivity. Tables 20-22 report the sensitivity of SAF to learning rate, batch size, and training length on three datasets and three backbones. In each group, we vary one hyperparameter and keep the others fixed to the default setting. SAF remains stable under moderate changes of these hyperparameters. A small learning rate usually gives a smaller ΔAcc but weaker ASR, while a large learning rate often increases ΔAcc and can reduce ASR, especially on Tiny-ImageNet with VGG-16. The default learning rate $2\mathrm{e}{-4}$ gives the best or near-best ASR in all settings. Batch size shows a similar trade-off. Batch size 8 sometimes gives slightly higher TopK-BBox, such as on CUB with DenseNet-121 and ImageNet with DenseNet-121, but it also causes a larger accuracy drop. Batch size 32 is more conservative in accuracy, but its ASR is usually lower. Increasing the training length from 20 to 30 epochs gives a clear gain, while the gain from 30 to 40 epochs is smaller. In several cases, 30 epochs gives slightly higher TopK-BBox than 40 epochs, but 40 epochs still gives the strongest or near-strongest ASR. These results show that SAF is not sensitive to a narrow hyperparameter choice, and the default setting gives a stable balance between attack success, localization, and clean accuracy.

Table 20: Hyperparameter sensitivity of SAF on CUB. ASR and TopK-BBox are averaged over seven explainers.

Backbone	Setting	ΔAcc	ASR	TopK-BBox
ResNet-50	lr=1e−4	0.5	82.7	66.0
	lr=2e−4	0.8	87.2	68.5
	lr=5e−4	1.4	85.1	67.2
	batch size=8	1.1	86.0	69.1
	batch size=16	0.8	87.2	68.5
	batch size=32	0.6	84.8	66.9
	epoch=20	0.5	78.6	62.1
	epoch=30	0.7	85.5	68.9
	epoch=40	0.8	87.2	68.5
VGG-16	lr=1e−4	0.6	82.8	65.2
	lr=2e−4	0.9	86.6	68.2
	lr=5e−4	1.6	83.8	67.5
	batch size=8	1.4	84.4	68.6
	batch size=16	0.9	86.6	68.2
	batch size=32	0.7	84.9	66.5
	epoch=20	0.6	79.3	62.6
	epoch=30	0.8	86.1	68.7
	epoch=40	0.9	86.6	68.2
DenseNet-121	lr=1e−4	0.5	83.3	66.5
	lr=2e−4	0.8	87.1	69.1
	lr=5e−4	1.3	86.0	69.3
	batch size=8	1.0	86.7	69.6
	batch size=16	0.8	87.1	69.1
	batch size=32	0.6	85.1	67.4
	epoch=20	0.5	80.2	63.8
	epoch=30	0.7	86.9	69.4
	epoch=40	0.8	87.1	69.1

Table 21: Hyperparameter sensitivity of SAF on Tiny-ImageNet. ASR and TopK-BBox are averaged over seven explainers.

Backbone	Setting	ΔAcc	ASR	TopK-BBox
ResNet-50	lr=1e−4	0.7	79.1	60.9
	lr=2e−4	1.1	83.7	63.6
	lr=5e−4	1.9	79.8	62.0
	batch size=8	1.3	82.6	64.3
	batch size=16	1.1	83.7	63.6
	batch size=32	0.9	81.9	62.1
	epoch=20	0.8	75.9	57.4
	epoch=30	1.0	83.0	64.0
	epoch=40	1.1	83.7	63.6
VGG-16	lr=1e−4	0.8	78.5	59.8
	lr=2e−4	1.2	83.2	63.1
	lr=5e−4	2.2	78.9	60.7
	batch size=8	1.7	80.6	62.2
	batch size=16	1.2	83.2	63.1
	batch size=32	0.9	81.2	61.5
	epoch=20	0.8	74.6	56.7
	epoch=30	1.1	82.0	63.4
	epoch=40	1.2	83.2	63.1
DenseNet-121	lr=1e−4	0.7	80.1	61.1
	lr=2e−4	1.0	84.5	64.7
	lr=5e−4	1.8	81.6	64.1
	batch size=8	1.3	83.2	65.2
	batch size=16	1.0	84.5	64.7
	batch size=32	0.8	82.5	62.8
	epoch=20	0.7	76.4	59.1
	epoch=30	0.9	84.0	65.0
	epoch=40	1.0	84.5	64.7

Table 22: Hyperparameter sensitivity of SAF on ImageNet-1K. ASR and TopK-BBox are averaged over seven explainers.

Backbone	Setting	ΔAcc	ASR	TopK-BBox
ResNet-50	lr=1e−4	0.6	82.1	64.4
	lr=2e−4	0.9	86.0	67.2
	lr=5e−4	1.5	85.4	67.6
	batch size=8	1.2	85.7	68.0
	batch size=16	0.9	86.0	67.2
	batch size=32	0.7	84.6	66.1
	epoch=5	0.6	79.6	62.7
	epoch=10	0.8	85.8	67.7
VGG-16	lr=1e−4	0.7	81.4	63.9
	lr=2e−4	1.0	85.5	66.6
	lr=5e−4	1.9	82.7	66.8
	batch size=8	1.5	83.9	67.1
	batch size=16	1.0	85.5	66.6
	batch size=32	0.8	83.4	65.3
	epoch=5	0.6	78.9	61.6
	epoch=10	0.9	85.1	67.0
DenseNet-121	lr=1e−4	0.6	82.9	65.0
	lr=2e−4	0.8	86.7	68.3
	lr=5e−4	1.5	85.6	68.7
	batch size=8	1.1	85.8	68.9
	batch size=16	0.8	86.7	68.3
	batch size=32	0.7	85.0	66.7
	epoch=5	0.5	80.4	63.2
	epoch=10	0.8	86.3	68.8

G SOCIAL IMPACT

This work studies a failure mode of post-hoc explanation evaluation. A positive impact is that it shows why multi-explainer agreement should not be used as the only evidence that a model relies on the intended object region. This can help researchers and practitioners design stronger checks that combine explanations with interventions, counterfactual samples, or model-internal information. The experiments use public image classification datasets and standard architectures, and no private or human-subject data are involved.

The main negative risk is dual use. A malicious model provider could use a similar model-side attack to make a classifier appear to rely on object evidence while its decision is driven by a shortcut

cue. This risk is especially relevant when explanations are used to support model release, debugging, or trust claims. We present the attack in a controlled setting with a label-matched patch so that the shortcut is explicit and measurable. The goal is not to provide a deployment recipe, but to show that agreement among several standard explainers is not sufficient by itself to certify the model's decision evidence. We hope this result encourages more careful evaluation of explanation-based checks before they are used in high-stakes settings.

H LIMITATIONS

Our theoretical analysis is developed for the seven post-hoc explainers studied in this paper under a shared spatial partition. These explainers cover gradient-based, CAM-based, propagation-based, and perturbation-based methods, but they do not cover all explanation families. In particular, we do not study explanation methods designed specifically for vision transformers, such as attention rollout, attention flow, and transformer attribution. We also do not consider concept-based, counterfactual, or example-based explanations. Our results therefore should not be read as a claim that all explanation methods have the same decomposition. Instead, they show that several widely used visual explainers can be driven by the same block-level response, which is enough to make multi-explainer agreement unreliable in the setting considered here.

Our attack uses a label-matched patch as a controlled shortcut. This design makes the decision cue explicit and allows us to test whether explanations can hide the patch while highlighting the original object region. Real shortcut cues may take other forms, such as texture, background, acquisition bias, or class-specific co-occurrence. We expect the same risk to matter beyond patch cues, but the present experiments use patches because they provide a clean and measurable test case.

I THE USE OF LARGE LANGUAGE MODELS

In this submission, we used an LLM solely to polish the writing and correct grammatical errors.