WHERE TO PLACE LOCAL PERTURBATIONS? BREAKING IMAGE CLASSIFIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

The placement of local disturbances is a critical factor in robustness testing for vision models. A straightforward strategy is random placement, while attribution methods can guide perturbations toward regions considered relevant for the model prediction. We study which placement strategies induce the strongest degradation under a fixed perturbation budget. Focusing on image classification as a controlled first setting, we compare attribution-based guidance with matched random baselines across datasets, architectures, perturbation operators, and spatial resolutions. We find that no guidance strategy is uniformly strongest. Object-aware random masking can approach or exceed attribution-based guidance in several ImageNet-S50 settings, whereas SHAP induces stronger degradation in other configurations, particularly on Oxford Flowers. Our results show that effective perturbation placement is strongly configuration-dependent and provide a basis for future robustness testing of more complex vision tasks.

1 INTRODUCTION

Artificial intelligence (AI) is increasingly deployed in vision systems operating under uncontrolled real-world conditions. In such settings, model reliability can be affected by *local disturbances* that alter only parts of the visual input while leaving the remaining scene unchanged. Examples include water droplets adhering to surveillance-camera lenses You et al. (2016), dirt on automotive cameras after off-road driving Uřičář et al. (2019), lens contamination in mobile imaging Choi et al. (2025), blood or debris impairing laparoscopic cameras during surgery Yazdanpanah et al. (2017), and dust or other contaminants affecting industrial vision systems used for automated quality inspection Munir et al. (2024). Evaluating robustness against such disturbances therefore requires not only modeling realistic degradation effects, but also determining *where* they should be applied.

A straightforward strategy is to place disturbances randomly across the image. Alternatively, explainable artificial intelligence (xAI) methods provide saliency maps that can guide perturbations toward image regions considered relevant for a model prediction. This raises a central question for robustness evaluation: *which guidance strategy induces the strongest model degradation under a fixed perturbation budget?* Identifying such strategies is particularly relevant for worst-case testing, where perturbations should expose robust prediction behavior as efficiently as possible.

Early work such as Most Relevant First (MoRF) Samek et al. (2017) suggests that perturbing highly relevant regions can degrade model performance more strongly than random removal. Following the xAI taxonomy in Nauta et al. (2023), such approaches belong to incremental deletion protocols, where a guidance function determines the order in which input regions are perturbed. Existing studies typically summarize degradation using Area Over the Perturbation Curve (AOPC) Samek et al. (2017) or Area Under the Curve (AUC) Petsiuk et al. (2018), but provide only partial comparisons between guidance strategies Hooker et al. (2019); Wagner et al. (2019); Hesse et al. (2021); Böhle et al. (2021); Fong & Vedaldi (2017); Heo et al. (2019); Nam et al. (2019). Several omit either SHAP Lundberg & Lee (2017) or random baselines, while others evaluate explanation sensitivity rather than degradation trajectories Yeh et al. (2019); Bhatt et al. (2020). Evaluations are furthermore often restricted to ImageNet Deng et al. (2009), CIFAR-10 Krizhevsky & Hinton (2009), or convolutional architectures.

Consequently, a systematic comparison of attribution-based and random guidance strategies across datasets, model architectures, and perturbation settings remains missing. Following prior perturbation-based work, we study this question in image classification, the most elementary setting considered

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.

here for analyzing spatial perturbation guidance. Our goal is to identify which guidance strategies induce the strongest degradation under controlled local perturbations and thereby establish a basis for future transfer to more complex vision tasks, such as object detection or semantic segmentation.

2 Related Work

Evaluation and Faithfulness. Perturbation-based evaluation measures how predictions change under structured input modifications and is widely used to relate attribution maps to model behavior. Early work introduced deletion-style protocols Samek et al. (2017), followed by gradient-based analyses and sensitivity measures Rao et al. (2022), and combined attribution schemes such as Full-Gradient Srinivas & Fleuret (2019). Deletion/insertion scores Petsiuk et al. (2018) and infidelity metrics Yeh et al. (2019) summarize degradation with single-value scores, obscuring how confidence evolves during progressive removal. Reliability concerns led to sanity checks Adebayo et al. (2020) and ROAR Hooker et al. (2019), which showed that many methods perform no better than random baselines under retraining, but do not capture progressive collapse dynamics. More recent work has emphasized the role of model robustness in attribution evaluation. Shah et al. Shah et al. (2021) showed that deletion-based evaluations can be misleading on non-robust models, where gradients reflect brittle off-manifold features, and proposed DiffROAR. Tomsett et al. Tomsett et al. (2019) further showed that commonly used saliency metrics can be insensitive to model perturbations, reinforcing the need for stronger, behaviorally grounded evaluation protocols. Related concerns arise from adversarial studies Ghorbani et al. (2018); Dombrowski et al. (2019), which reveal that imperceptible perturbations can drastically alter predictions and explanations. Zheng et al. Zheng et al. (2025) addressed masking-induced out-of-distribution artifacts with F-Fidelity, emphasizing that perturbation design can strongly affect attribution-faithfulness estimates. Recent work has further systematized attribution evaluation protocols and benchmarking practices Duan et al. (2024). Rong et al. Rong et al. (2022) showed that deletion metrics can be confounded by masking artifacts and information leakage induced by perturbation patterns. They proposed ROAD, a debiased retraining-free framework to improve consistency across perturbation protocols.

Interpretability Taxonomies and Evaluation Criteria. Doshi-Velez and Kim Doshi-Velez & Kim (2017) categorize interpretability evaluation into application-grounded, human-grounded, and functionally grounded approaches. Our work operates in the functionally grounded setting, but shifts the focus from attribution scoring to analyzing how model predictions respond to evidence removal. Jacovi and Goldberg Jacovi & Goldberg (2020) emphasize the distinction between faithfulness and visual plausibility, noting that visually coherent explanations need not reflect the model's decision process. Complementing these perspectives, Mohseni et al. Kadir et al. (2023) provide a taxonomy of evaluation criteria including fidelity, robustness, and consistency; our approach contributes to this landscape by offering a behavioral diagnostic of prediction stability under controlled interventions.

Causal Inference and Robust Reasoning. Several works interpret attribution as an approximation to causal feature effects. Narendra et al. Narendra et al. (2018) and Chattopadhyay et al. Chattopadhyay et al. (2019) estimate feature effects using structural causal models, while SCOUT Wang & Vasconcelos (2020) generates causal heatmaps under learned assumptions. Diffusion-based counterfactuals Augustin et al. (2022) aim to produce realistic interventions, albeit at substantial computational cost. Related perturbation-based metrics, such as sensitivity Yeh et al. (2019) and deletion scores Petsiuk et al. (2018), can be viewed as special cases of this interventionist view. Adversarial training has been shown to improve attribution quality Ilyas et al. (2019), and shape-biased models exhibit more semantically aligned saliency maps Geirhos et al. (2022). Our degradation framework is conceptually related to robustness evaluation protocols such as AutoAttack Athalye et al. (2018) and corruption benchmarks such as ImageNet-C Hendrycks & Dietterich (2019), in that all analyze prediction behavior under input perturbations. However, unlike these benchmarks, our framework does not optimize adversarial perturbations or measure corruption robustness directly, but instead studies confidence degradation under ordered evidence removal. Theoretical analyses provide formal support for this connection. Srinivas et al. Srinivas et al. (2023) link perceptually aligned gradients to off-manifold robustness, arguing that attribution fidelity may emerge as a consequence of robustness rather than explanation design.

Attribution Methods as Guidance Signals. Our framework is compatible with a broad range of attribution methods, including gradient-, activation-, perturbation-, and axiomatic-based approaches

such as Grad-CAM, Integrated Gradients, SHAP, and RISE Selvaraju et al. (2019); Sundararajan et al. (2017); Lundberg & Lee (2017); Petsiuk et al. (2018). In contrast to prior work that evaluates these methods as explanations, we use attribution maps only to induce perturbation orderings guidance for degradation diagnosing.

3 Degradation-Based Evaluation Framework

3.1 Model Degradation

Ordered, budget-constrained evidence removal induces degradation trajectories that characterize how prediction confidence evolves under progressive perturbations. Robustness is commonly understood as the degree to which a system functions correctly under invalid inputs or stressful environmental conditions Shahrokni & Feldt (2013). In this context, degradation trajectories provide a local view of prediction behavior under controlled perturbations, while prediction robustness denotes the sensitivity of model confidence to structured evidence removal under fixed perturbation budgets. A prediction is considered robust if relatively small perturbations induce disproportionately large confidence degradation. Formally, let $f : \mathcal{X} \to \mathcal{Y}$ be a trained classifier and $x \in \mathcal{X}$ an input with label $y \in \mathcal{Y}$. A perturbation operator δ modifies the input using a mask M, producing a degraded input $x' = \delta(x, M)$. If the confidence assigned to the original class decreases under $\delta(x, M)$, the prediction is sensitive to the suppressed region under the chosen perturbation. We quantify the empirical influence of a masked region $S \subseteq \{1, \ldots, H \times W\}$ as

$$\mathrm{EI}_f(S \mid x) = f(x) - f(\delta(x, M_S)), \tag{1}$$

where M_S masks the pixels in S. Larger values correspond to greater prediction sensitivity to suppression of that region. A guidance function induces an ordered sequence of regions $\{S_1, \ldots, S_T\}$. At perturbation step t, the cumulative mask is

$$\tilde{S}_t = \bigcup_{k=1}^{t} S_k, \tag{2}$$

yielding degraded input $x_t = \delta(x, M_{\tilde{S}_t})$. The resulting degradation curve $D(t) = f(x_t)$ defines a confidence trajectory over cumulative perturbation budget. The observed degradation behavior depends on both the perturbation ordering and the intrinsic sensitivity of the model. Consequently, degradation metrics should be interpreted as observational descriptors of perturbation sensitivity rather than calibrated robustness measures, consistent with the behavioral notion of faithfulness discussed by Jacovi and Goldberg Jacovi & Goldberg (2020).

3.2 Framework Overview

Based on this formulation, the evaluation framework consists of two components:

1. Guidance. We define a guidance function $g(x; f, y) \in [0, 1]^{H \times W}$ that assigns a scalar score to each spatial location with respect to model f and target class y. The guidance map can be produced by an attribution method or by a random baseline, thereby inducing an ordering over regions (patches) $\{S_t\}_{t=1}^{T}$.

2. Constrained Perturbation Evaluation. Given g and a perturbation operator δ, regions are masked iteratively according to this ordering, and the model is evaluated on the cumulatively perturbed input $x_t = \delta(x, M_{\tilde{S}_t})$. At each step t, we record the confidence $D(t) = f(x_t)$, yielding a degradation trajectory under standardized, cumulative perturbation budgets

3.3 Guidance Strategies

We consider multiple guidance strategies that induce an ordering over input regions for progressive masking. These include attribution-based methods such as SHAP Lundberg & Lee (2017), Grad-CAM Selvaraju et al. (2019), SmoothGrad Wang et al. (2020), and Integrated Gradients Sundararajan et al. (2017), as well as gradient-based signals including Activation Maximization Zhou et al. (2018) and FGSM Goodfellow et al. (2015), see excamples in Fig. 1. We further include perturbation-based approaches such as Occlusion Sensitivity, which measures prediction changes under systematic masking of image regions, and RISE Petsiuk et al. (2018).

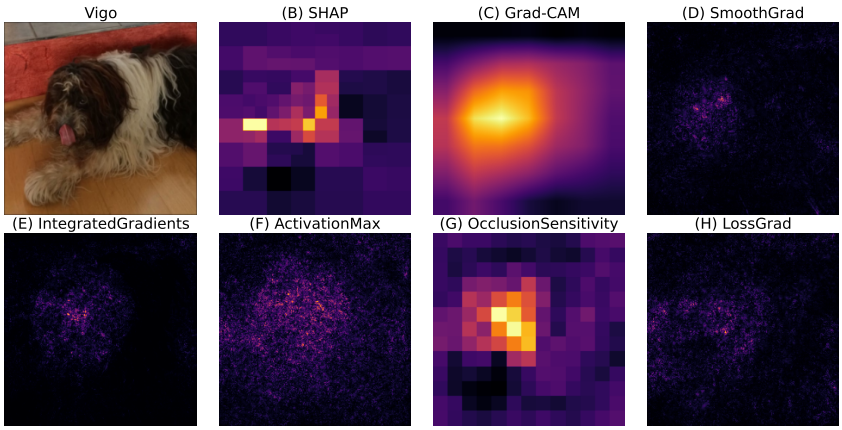

Figure 1: Examples of guidance maps from selected methods for a single input (dog: "Vigo"), used to induce patch orderings in degradation analysis.

4 Robustness Diagnostics

Guidance-driven masking induces a sequence of binary masks $\{M_t\}_{t=0}^{T}$, where $M_t \in \{0,1\}^{H\times W}$ indicates which pixels remain visible after t perturbation steps ($M_t = 1$ keeps a pixel, $M_t = 0$ masks it). At each step, the model's confidence in the ground-truth class is recorded

$$D_t \;=\; f(\delta(x, M_t))\,, \tag{3}$$

where $f(\cdot)$ is the softmax probability of class y unless stated otherwise. We use softmax probabilities rather than logits because deletion-style evaluation conventionally measures the model's expressed class confidence, while acknowledging that these scores are not calibrated probabilities. To enable comparability across patch sizes and masking schedules, the trajectory is parameterized by a *normalized blindness ratio* $b \in [0,1]$ rather than the discrete step index t. Two complementary blindness definitions are considered. **Absolute blindness** measures the fraction of the full image that is masked, while **relative (object) blindness** measures the masked fraction restricted to a predefined object region $S \in \{0,1\}^{H\times W}$ (from segmentation or bounding boxes):

$$b_t^{\text{abs}} = \frac{\|1 - M_t\|_1}{H \cdot W} \qquad b_t^{\text{rel}} = \frac{\|(1 - M_t) \odot S\|_1}{\|S\|_1}, \tag{4}$$

where $\odot$ denotes elementwise multiplication. The samples $(b_t, D_t)_{t=0}^{T}$ are used for metric computation. In object-relative experiments, all guidance strategies select patches exclusively from within S, ensuring consistent perturbation budgets and spatial constraints across guidance strategies. While b^{abs} captures global perturbation extent, b^{rel} isolates degradation within the object support, decoupling foreground relevance from background masking effects.

4.1 Modeling Degradation Dynamics

For each image j, we observe samples $\{(b_i, D_j(b_i))\}_{i=1}^{K_j}$, where b_i denotes blindness and $D_j(b_i)$ the corresponding confidence. Because masking steps induce irregular, image-dependent sampling of the blindness axis, we resample each trajectory onto a uniform grid using linear interpolation. This yields a resolution-independent approximation $\hat{D}_j(b)$ without imposing a parametric form. Dataset-level summaries are obtained via macro aggregation across images, following Petsiuk et al. (2018). At each grid location b, we compute the summary statistic over the set $\{\hat{D}_j(b)\}_j$, ensuring that each image contributes equally, independent of the number or spacing of masking steps.

4.2 Area Under the Degradation Curve (AUC)

Following deletion-based evaluations from Petsiuk et al. (2018), we summarize cumulative confidence retention by integrating $D_j(b)$ over $b \in [0, 1]$:

$$\text{AUC} = \sum_{i=1}^{K_j} \frac{(b_i - b_{i-1})(D_j(b_i) + D_j(b_{i-1}))}{2}. \tag{5}$$

Here, b_i denotes either absolute or object-relative blindness. AUC is computed per image and averaged across the dataset. Higher values indicate sustained confidence, whereas lower values reflect rapid degradation. As an integral statistic, AUC does not capture where or how collapse occurs, motivating the diagnostics below. Consequently, qualitatively different degradation behaviors can yield similar AUC values despite reflecting substantially different robustness regimes.

4.3 Area Between Quantile Degradation Curves (AUC_{BAND})

To quantify variability across inputs, we compute lower and upper quantile curves $Q_{\text{low}}(b)$ and $Q_{\text{high}}(b)$ from the interpolated trajectories $\{\hat{D}_j(b)\}_j$. The area between them is

$$\text{AUC}_{\text{band}} = \int_0^1 (Q_{\text{high}}(b) - Q_{\text{low}}(b))\, db. \tag{6}$$

Interpreted on the degradation trajectory, AUC_{band} quantifies the extent to which prediction collapse is consistent across inputs. Small values indicate homogeneous degradation behavior, whereas larger values reveal heterogeneous robustness regimes, where similar perturbation budgets induce substantially different confidence responses across inputs. As with AUC, aggregation is performed at the dataset level so that each image contributes equally. While both AUC-based measures summarize global properties of the trajectory, they do not capture the location or sharpness of prediction collapse, which are addressed by the metrics below.

4.4 Sensitivity Location (∇B)

To characterize where prediction confidence declines most strongly along the blindness axis, we identify the perturbation level at which the degradation trajectory exhibits its steepest decrease. Given the interpolated trajectory $\hat{D}_j(b)$, the sensitivity location is defined as

$$\nabla B = \arg\max_{b \in [0,1]} \frac{d\hat{D}_j(b)}{db}. \tag{7}$$

Interpreted on the degradation trajectory, ∇B denotes the blindness ratio at which confidence decreases most rapidly under progressive removal. Smaller values indicate that the strongest degradation occurs early, after only a small perturbation budget has been applied, whereas larger values indicate that the strongest decline occurs later in the trajectory. To ensure stability, derivatives are evaluated on the interpolated trajectory $\hat{D}_j(b)$, reducing dependence on the discrete masking schedule. Unlike AUC-based measures, which summarize cumulative confidence retention, ∇B localizes the perturbation regime in which the model is most sensitive to additional evidence removal. Importantly, ∇B describes the location of the steepest confidence decline rather than its magnitude.

4.5 Collapse Point (CP)

While the sensitivity location captures how abruptly collapse occurs, the Collapse Point (CP) characterizes *when* it occurs by identifying the earliest blindness level at which confidence falls below a predefined threshold. Formally, we define

$$\text{CP}_\tau = \min\{b \in [0, 1] \mid \hat{D}_j(b) < \tau\}, \tag{8}$$

where τ is a confidence threshold (e.g., $\tau = 0.5$ or $\tau = 0.8$), and $\hat{D}_j(b)$ denotes the interpolated degradation trajectory. Interpreted on the degradation trajectory, CP_τ is a first-passage statistic that measures how much evidence must be removed before the prediction becomes unstable. Lower

values indicate early collapse under minimal perturbation, whereas higher values reflect sustained confidence under progressive removal. Evaluating CP_τ on the interpolated trajectory yields a resolution-independent estimate of collapse onset and avoids discretization artifacts arising from irregular masking schedules. In contrast to AUC measures, which summarize cumulative retention, CP_τ provides a measure of the onset of prediction failure.

4.6 Relation to Perturbation-Based Faithfulness Metrics

Perturbation-based evaluation removes or inserts input evidence and measures the resulting change in model output Petsiuk et al. (2018); Samek et al. (2017); Rao et al. (2022); Srinivas & Fleuret (2019). Standard deletion is recovered in our framework by using absolute blindness, a fixed perturbation operator, and AUC aggregation. Our formulation extends this setting by normalizing perturbation budgets through absolute and object-relative blindness, treating degradation as a trajectory rather than a scalar score. This shifts the emphasis from ranking attribution methods toward analyzing prediction sensitivity under controlled evidence removal. Attribution-guided masking is therefore treated as one perturbation schedule among others, complementing rather than replacing attribution-faithfulness analysis. Consistent with prior work showing that deletion scores can match random baselines under retraining or be confounded by brittle off-manifold responses Hooker et al. (2019); Shah et al. (2021); Srinivas et al. (2023), cases where random masking matches attribution-guided masking suggest distributed perturbation sensitivity rather than localized semantic evidence. Unlike retraining-based protocols such as ROAR and DiffROAR, our framework keeps model parameters fixed, thereby isolating degradation behavior from representation relearning effects. Similarly, whereas F-Fidelity reduces masking-induced covariate shift during attribution evaluation Zheng et al. (2025), we analyze the resulting degradation behavior itself.

5 Experimental Results

5.1 Implementation

We implement a unified forward-pass pipeline for all guidance strategies. Each input x is partitioned into a grid of patches. At perturbation step t, the next patch is selected according to the ordering induced by the guidance map g and masked to obtain x_t, while recording the model confidence $f(x_t)$ for the ground-truth class. All guidance strategies from Section 3.3 are evaluated under a shared perturbation protocol. In addition, with three random baselines to disentangle guidance-specific effects from structural masking artifacts: ***Random All (R_{All})*** samples patches uniformly over the entire image, ***Random Object First (R_{Obj})*** restricts sampling to the object region before masking the background, and ***Background First (R_{Back})*** masks background regions prior to the object. For attribution-based methods, we use standard configurations: SHAP Lundberg & Lee (2017) with 1000 samples, Grad-CAM (GC) Selvaraju et al. (2019), SmoothGrad (S-GC) Wang et al. (2020) with 25 samples, and Integrated Gradients (IG) Sundararajan et al. (2017) with 50 steps. Activation Maximization (AM) Zhou et al. (2018) is optimized for 30 steps, and loss gradients are computed with respect to the cross-entropy objective. Perturbation-based methods include Occlusion Sensitivity (OS) and RISE Petsiuk et al. (2018), the latter using 1000 samples per image. All guidance maps are normalized to $[0, 1]$, resized to input resolution, and converted into ranked patch orderings. Degradation trajectories are evaluated on a uniform blindness grid with $K = 100$ points via linear interpolation. For variability analysis, we compute quantile bands using the $(0.1, 0.9)$ quantiles at each grid point. We consider multiple perturbation operators δ: Mean Distance Perturbation (MDP), inspired by the maximum-distance baseline of Sturmfels et al. Sturmfels et al. (2020), uniform noise, and constant masking (black or white). MDP replaces each selected patch with a single valid RGB color that is maximally distant from the patch's mean color, thereby removing its local spatial and color structure through a deterministic perturbation. All experiments were performed using a single NVIDIA V100 GPU. Code and implementation are available at `ANONYM`.

5.2 Large-Scale Experiments

We compare guidance strategies in terms of the degradation behavior they induce under progressive removal. For each configuration, we compute the metrics from Section 4. Table 1 summarizes results across datasets, perturbation operators, and architectures.

Metric	R_{All}	R_{Obj}	R_{Back}	SHAP	GC	S-GC	IG	AM	OS	Loss	RISE
Imagenet-S50 Gao et al. (2022): Models with MDP-7/14/32, Noise-14, Black-14 and White-14*											
AUC^{rel}	0.1661	**0.0613**	0.1707	0.0702	0.1220	0.0859	0.1287	0.1081	0.1246	0.1296	0.1307
AUC^{rel}_{Band}	0.4446	**0.2484**	0.4548	0.3660	0.5073	0.4018	0.5445	0.4793	0.6586	0.5736	0.5868
∇B^{rel}	0.2173	**0.0837**	0.2239	0.0978	0.1594	0.1120	0.1672	0.1428	0.1639	0.1684	0.1663
$\text{CP}^{rel}_{0.5}$	0.1523	0.0526	0.1569	**0.0551**	0.1048	0.0730	0.1127	0.0931	0.1030	0.1132	0.1159
$\text{CP}^{rel}_{0.8}$	0.1028	0.0339	0.1060	**0.0316**	0.0699	0.0466	0.0726	0.0581	0.0577	0.0720	0.0771
AUC^{abs}	0.1646	0.1132	0.1693	**0.1100**	0.1740	0.1301	0.1568	0.1480	0.1331	0.1636	0.1544
AUC^{abs}_{Band}	0.4302	**0.3909**	0.4412	0.4354	0.5784	0.4664	0.5222	0.5091	0.5927	0.5520	0.5811
∇B^{abs}	0.2142	0.1482	0.2215	**0.1393**	0.2177	0.1602	0.1967	0.1868	0.1713	0.2029	0.1922
$\text{CP}^{abs}_{0.5}$	0.1506	0.1006	0.1547	**0.0920**	0.1513	0.1115	0.1364	0.1279	0.1108	0.1416	0.1356
$\text{CP}^{abs}_{0.8}$	0.1015	0.0662	0.1044	**0.0571**	0.0982	0.0710	0.0883	0.0818	0.0636	0.0909	0.0884
Oxford-IIIT Pets Parkhi et al. (2012): ViT-B16 - MDP-14, Noise-14, Black-14 and White-14)											
AUC^{rel}	0.0432	0.0307	0.0436	**0.0142**	0.0294	0.0317	0.0370	0.0369	0.0160	0.0345	0.0218
AUC^{rel}_{Band}	0.6864	0.2204	0.6910	**0.1616**	0.2820	0.2666	0.4463	0.4230	0.1560	0.3573	0.3506
∇B^{rel}	0.8525	**0.6097**	0.8538	0.7580	0.7530	0.6621	0.7516	0.7168	0.6835	0.7079	0.6882
$\text{CP}^{rel}_{0.5}$	0.0306	**0.0081**	0.0313	0.0085	0.0116	0.0110	0.0207	0.0194	0.0101	0.0158	0.0169
$\text{CP}^{rel}_{0.8}$	0.0244	0.0065	0.0249	**0.0056**	0.0079	0.0082	0.0157	0.0148	0.0064	0.0124	0.0132
AUC^{abs}	0.0431	0.0378	0.0434	**0.0195**	0.0339	0.0355	0.0387	0.0386	0.0173	0.0382	0.0247
AUC^{abs}_{Band}	0.6893	0.4607	0.6928	**0.2732**	0.4411	0.4230	0.5396	0.5319	0.1848	0.4944	0.4041
∇B^{abs}	0.8625	**0.7390**	0.8631	0.8429	0.8110	0.7392	0.7979	0.7816	0.8195	0.7723	0.8396
$\text{CP}^{abs}_{0.5}$	0.0305	0.0199	0.0310	0.0132	**0.0181**	0.0170	0.0237	0.0229	0.0108	0.0209	0.0186
$\text{CP}^{abs}_{0.8}$	0.0243	0.0159	0.0244	0.0097	**0.0134**	0.0129	0.0179	0.0178	0.0070	0.0163	0.0146
Oxford Flowers Nilsback & Zisserman (2008): ViT-B16 - MD-7/14/32, Noise-14, Black-14, and White-14											
AUC^{rel}	0.5183	0.2761	0.5179	**0.1788**	0.4468	0.2632	0.2785	0.2891	0.2857	0.3134	0.3283
AUC^{rel}_{Band}	0.8121	0.6996	0.8119	**0.5731**	0.8851	0.7563	0.7504	0.7742	0.8311	0.7940	0.8001
∇B^{rel}	0.6002	0.3024	0.6019	**0.1655**	0.5029	0.2683	0.2874	0.2976	0.2881	0.3294	0.3417
$\text{CP}^{rel}_{0.5}$	0.5132	0.2633	0.5129	**0.1537**	0.4489	0.2457	0.2625	0.2734	0.2664	0.3009	0.3179
$\text{CP}^{rel}_{0.8}$	0.3551	0.1758	0.3551	**0.0885**	0.3135	0.1522	0.1625	0.1698	0.1515	0.1956	0.2214
AUC^{abs}	0.5182	0.3720	0.5179	**0.2182**	0.4539	0.2881	0.3200	0.3146	0.2885	0.3316	0.3454
AUC^{abs}_{Band}	0.8088	0.7453	0.8084	**0.5996**	0.8292	0.7350	0.7539	0.7624	0.7988	0.7642	0.7996
∇B^{abs}	0.6014	0.4207	0.6025	**0.2114**	0.5123	0.2947	0.3385	0.3285	0.2953	0.3493	0.3628
$\text{CP}^{abs}_{0.5}$	0.5319	0.3739	0.5316	**0.1959**	0.4557	0.2712	0.3070	0.3010	0.2693	0.3196	0.3361
$\text{CP}^{abs}_{0.8}$	0.3682	0.2533	0.3676	**0.1151**	0.3118	0.1650	0.1910	0.1858	0.1534	0.2037	0.2306

Table 1: Degradation metrics for guidance strategies under perturbation operators and model architectures. Lower values indicate stronger degradation. Models*: ResNet32/50 He et al. (2015), VGG16 Simonyan & Zisserman (2015), EfficientNet Tan & Le (2020), DenseNet Huang et al. (2018) and ViT Dosovitskiy et al. (2021).

Across datasets, no single guidance strategy consistently induces the strongest degradation. On ImageNet-S50, R_{Obj} frequently approaches or exceeds attribution-based guidance, particularly under object-relative blindness, whereas the ordering under absolute blindness is more metric-dependent. On Oxford-IIIT Pets, the relative ordering of R_{Obj} and SHAP varies across metrics, while Oxford Flowers consistently favors SHAP. The effectiveness of attribution-guided perturbation relative to object-aware random masking is therefore dataset-, blindness-axis-, and metric-dependent. Importantly, strong performance of R_{Obj} does not establish a particular model failure mechanism; rather, it demonstrates that restricting perturbations to the object region can itself provide a strong baseline against which attribution-guided strategies should be evaluated.

From a practical perspective, these results highlight that the benefit of attribution-guided perturbation depends on the evaluation configuration. SHAP induces particularly strong degradation in some settings, whereas R_{Obj} approaches or exceeds SHAP in several ImageNet-S50 configurations without requiring attribution computation. The remaining guidance strategies likewise exhibit no consistent global ordering across datasets and metrics. Fig. 2 illustrates an ImageNet-S50 setting in which R_{Obj} induces strong degradation relative to attribution-based strategies. Rather than indicating a universally preferable guidance method, these observations motivate a closer analysis of how the relative ordering between guidance strategies changes with the perturbation protocol. We therefore next investigate the effects of patch size and perturbation operator.

Table 9 summarizes the ablation for ResNet50 (RN50) and ViT. Within these configurations, patch size substantially affects both the magnitude of degradation and the relative ordering of R_{Obj} and SHAP. At MDP-7, SHAP generally induces earlier or stronger degradation, whereas at the coarse

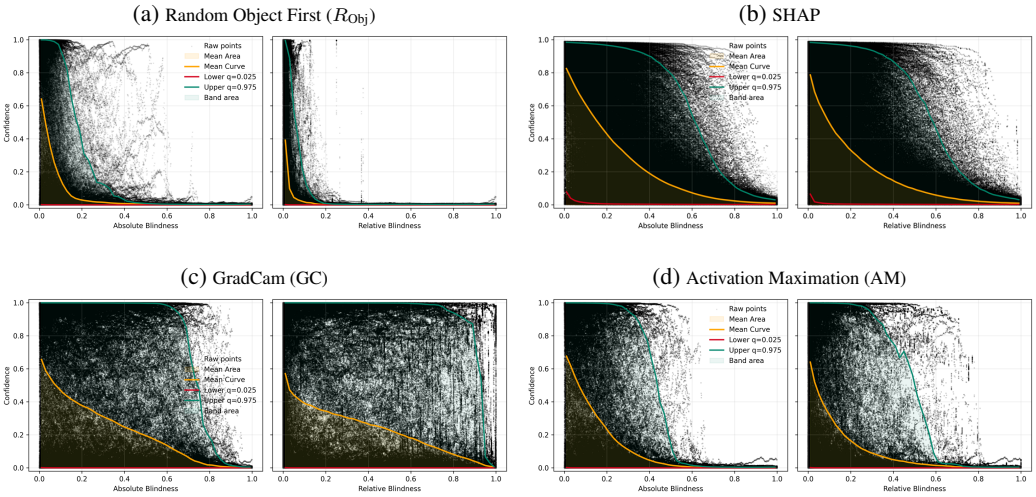

Figure 2: Example degradation curves for different guidance strategies on ImageNet-S50 Gao et al. (2022) using a ViT Dosovitskiy et al. (2021) model under MDP-32 perturbation.

MDP-32 resolution, R_{Obj} yields lower values across the reported metrics for both architectures. The ordering therefore reverses as perturbation granularity changes. This result further demonstrates that the relative effectiveness of attribution-based and random guidance cannot be characterized independently of the perturbation protocol. Although interactions between attribution estimation and masking resolution may contribute to this behavior, the present experiments do not isolate its underlying cause.

Strategy:	**Random Object First (R_{Obj})**				**SHAP**			
Config.	**AUC↓**	**AUC$_{\text{band}}$↓**	**∇B↓**	**CP$_{0.8}$↓**	**AUC↓**	**AUC$_{\text{band}}$↓**	**∇B↓**	**CP$_{0.8}$↓**
RN50 MDP-7	9.65/14.02	33.94/45.20	12.02/16.57	5.80/8.64	5.85/10.86	29.85/41.55	7.92/12.44	2.56/6.01
RN50 MDP-14	3.04/6.60	12.59/25.14	4.55/8.16	1.76/4.03	4.01/8.11	25.04/32.22	5.01/8.87	1.77/4.44
RN50 MDP-32	0.77/2.83	3.36/14.00	1.62/3.75	0.35/1.45	3.02/6.47	19.71/26.30	4.35/7.42	1.23/3.22
RN50 Black-14	3.37/7.03	14.80/30.14	4.88/8.96	1.89/4.15	4.29/8.38	26.70/35.14	6.27/10.17	1.94/4.64
RN50 White-14	3.96/8.25	17.53/31.55	5.88/10.59	2.16/4.86	4.96/9.29	30.74/38.06	6.88/11.07	2.30/5.15
RN50 Noise-14	3.98/8.51	15.58/29.72	5.01/10.42	2.13/4.72	4.93/9.29	28.83/36.21	7.31/11.68	2.03/4.94
ViT MDP-7	21.52/30.37	67.00/78.35	27.21/38.92	14.51/20.82	16.03/21.76	67.71/73.17	20.83/27.59	9.74/14.66
ViT MDP-14	14.36/27.60	57.02/76.64	18.26/35.40	9.47/19.34	15.72/20.85	71.58/74.15	19.84/26.12	9.28/13.69
ViT MDP-32	1.34/4.94	5.89/18.76	2.17/6.46	0.53/2.66	6.34/9.98	37.34/42.39	8.05/11.73	2.96/5.27
ViT Black-14	14.67/27.80	61.89/77.37	18.22/36.41	9.12/19.07	16.56/21.54	73.20/75.24	21.50/28.08	9.87/14.17
ViT White-14	15.11/28.54	61.60/78.71	18.49/35.95	9.60/19.73	16.69/21.78	76.01/78.71	21.64/28.20	9.97/14.41
ViT Noise-14	13.55/26.74	52.73/76.51	17.58/34.49	8.66/18.29	17.53/22.17	75.33/76.41	21.55/27.04	10.20/14.21

Table 2: Ablation study on the effect of perturbation method and patch size for ResNet50 (RN50) and ViT. Report in degradation metrics ($\cdot 10^{-3}$) for Random Object First (R_{Obj}) and SHAP.

Within the ResNet50 and ViT ablation, variations in patch size produce larger changes in several degradation metrics than changing the replacement operator at a fixed patch size. The effects of the perturbation operator are architecture- and guidance-dependent, and no consistent ordering between MDP, black, white, and noise perturbations is observed across all configurations. Thus, perturbation granularity represents an important experimental factor in the evaluated ImageNet-S50 settings, but its relative importance should not be generalized beyond the protocols considered here.

6 Discussion and Limitations

Our results show that degradation behavior under progressive removal depends on the evaluation configuration, including the dataset, blindness axis, metric, patch size, and perturbation operator. In particular, Random Object First can approach or exceed attribution-guided masking in several ImageNet-S50 configurations, whereas SHAP induces stronger degradation in other settings, most

consistently on Oxford Flowers. From this perspective, degradation curves are best interpreted as behavioral diagnostics of model response under structured perturbations, not direct causal measures of feature importance Jacovi & Goldberg (2020); Hooker et al. (2019); Shah et al. (2021). These observations are consistent with prior work showing that perturbation-based evaluations can depend on both model sensitivity and the perturbation protocol Ilyas et al. (2019); Shah et al. (2021); Srinivas et al. (2023). Our framework is related to prior work on perturbation-based attribution evaluation Petsiuk et al. (2018); Samek et al. (2017); Hooker et al. (2019); Rong et al. (2022), but differs in its primary objective. Methods such as ROAR, DiffROAR, and ROAD aim to improve the reliability of attribution evaluation by mitigating retraining bias, masking artifacts, or information leakage. In contrast, our framework compares the degradation behavior induced by different guidance strategies under shared perturbation protocols. From this perspective, attribution-guided masking is treated as one possible perturbation schedule among many. Accordingly, the focus shifts from ranking explanation methods to characterizing degradation dynamics through confidence retention, inter-sample variability, the location of the strongest confidence decline, and collapse onset.

The proposed framework is a diagnostic tool rather than a calibrated robustness metric. While degradation behavior may reflect robustness-related properties, we do not provide external validation against established robustness benchmarks such as ImageNet-C Hendrycks & Dietterich (2019) or AutoAttack Athalye et al. (2018). Accordingly, robustness-related interpretations should be viewed as observational rather than validated robustness measurements. Our analysis further depends on the perturbation protocol. Masking introduces distribution shift, and all masking-based perturbations involve some degree of manifold displacement. While we compare multiple perturbation operators under standardized budgets, evaluating degradation trajectories under generative inpainting or diffusion-based perturbations remains important future work. Object-relative masking assumes access to object regions and may therefore be less applicable in settings without reliable foreground annotations. Moreover, R_{Obj} uses privileged information about the object location. Its strong performance may therefore partly result from restricting perturbations to foreground regions, and similarity to attribution-guided degradation should not be interpreted as evidence of distributed feature reliance or as showing that attribution information is generally unnecessary. However, object-relative blindness is optional, and the framework remains applicable under global (absolute) blindness or predicted masks. Its purpose is to reduce foreground-area confounds in deletion-style evaluation rather than require segmentation supervision. Finally, we do not include a direct empirical comparison against perturbation-aware evaluation frameworks such as F-Fidelity Zheng et al. (2025). Evaluating agreement between degradation-based robustness descriptors and perturbation-aware faithfulness rankings is left for future work.

Moreover, we observe that degradation curves are not strictly monotonic, with confidence occasionally increasing at high blindness levels (> 0.7). This behavior is consistent with known overconfidence of modern neural networks under severe input degradation or distribution shift Ovadia et al. (2019); Minderer et al. (2021). As a result, aggregate measures such as AUC alone can obscure important dynamics, motivating complementary descriptors such as location and collapse point.

7 Conclusion

We introduced a degradation-based framework for analyzing how image classifiers respond to controlled local evidence removal. Beyond comparing perturbation guidance strategies, the framework separates global absolute blindness from object-relative blindness and characterizes degradation through complementary trajectory descriptors. Across datasets and perturbation settings, SHAP emerges as a strong attribution-based guidance method, but comes with higher computational cost, while object-aware random masking approaches or exceeds it in several configurations. Importantly, the relative effectiveness of the strategies changes with dataset, blindness definition, metric, patch size, and perturbation operator. The distinction between absolute and relative blindness further shows that conclusions can change when perturbation budgets are normalized to the object rather than the full image. Likewise, AUC alone does not capture all degradation behavior: sensitivity location, collapse points, and trajectory variability reveal differences in when and how confidence degrades that may remain hidden in cumulative averages. These findings motivate reporting multiple trajectory descriptors and matched random controls when evaluating perturbation placement. Future work should extend this analysis to more complex vision tasks and compare SHAP, object-aware random, and global random guidance under task-specific simulated disturbances.

AI USE STATEMENT

Large language models (LLMs) were used only to improve grammar, wording, and reading flow. They were not used to generate experimental results, analyses, or scientific conclusions. All LLM-assisted text was reviewed by the first author.

ETHICS STATEMENT

This work does not involve human or animal subjects and does not introduce foreseeable ethical, safety, privacy, or societal risks.

REPRODUCIBILITY STATEMENT

Implementation details are provided in the main paper. The code is available through an anonymous repository during review and will be de-anonymized and made publicly available upon acceptance.

REFERENCES

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.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, 2018. URL https://arxiv.org/abs/1802.00420.

Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 364–377. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/025f7165a452e7d0b57f1397fed3b0fd-Paper-Conference.pdf.

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

Moritz Böhle, Mario Fritz, and Bernt Schiele. Convolutional dynamic alignment networks for interpretable classifications. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10024–10033, 2021. doi: 10.1109/CVPR46437.2021.00990.

Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: A causal perspective, 2019. URL https://arxiv.org/abs/1902.02302.

Sooyoung Choi, Sungyong Park, and Heewon Kim. Sidl: A real-world dataset for restoring smartphone images with dirty lenses. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(3):2545–2554, Apr. 2025. doi: 10.1609/aaai.v39i3.32257. URL https://ojs.aaai.org/index.php/AAAI/article/view/32257.

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.

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.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017. URL https://arxiv.org/abs/1702.08608.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Jiarui Duan, Haoling Li, Haofei Zhang, Hao Jiang, Mengqi Xue, Li Sun, Mingli Song, and Jie Song. On the evaluation consistency of attribution-based explanations, 2024. URL https://arxiv.org/abs/2407.19471.

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.

Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. 2022.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2022. URL https://arxiv.org/abs/1811.12231.

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

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. URL https://arxiv.org/abs/1903.12261.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation, 2019. URL https://arxiv.org/abs/1902.02041.

Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Fast axiomatic attribution for neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 19513–19524. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a284df1155ec3e67286080500df36a9a-Paper.pdf.

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, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. URL https://arxiv.org/abs/1608.06993.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?, 2020. URL https://arxiv.org/abs/2004.03685.

Md Abdul Kadir, Amir Mosavi, and Daniel Sonntag. Evaluation metrics for xai: A review, taxonomy, and practical applications. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES). Conference on Intelligent Engineering Systems (INES-2023), July 26-28, Nairobi, Kenya*, pp. 000111–000124. IEEE, 2023.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

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

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks, 2021. URL https://arxiv.org/abs/2106.07998.

Adnan Munir, Abdul Jabbar Siddiqui, and Aoubaida M. Al Sabbagh. Dustrobust-yolo: Enhanced uav detection in dusty conditions. In *2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 150–157, 2024. doi: 10.1109/DICTA63115.2024.00033.

Woo-Jeoung Nam, Shir Gur, Jaesik Choi, Lior Wolf, and Seong-Whan Lee. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks, 2019. URL https://arxiv.org/abs/1904.00605.

Tanmayee Narendra, Anush Sankaran, Deepak Vijaykeerthy, and Senthil Mani. Explaining deep learning models using causal inference, 2018. URL https://arxiv.org/abs/1811.04376.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023. ISSN 1557-7341. doi: 10.1145/3583558. URL http://dx.doi.org/10.1145/3583558.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, 2019. URL https://arxiv.org/abs/1906.02530.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092.

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.

Sukrut Rao, Moritz Böhle, and Bernt Schiele. Towards better understanding attribution methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18770–18795. PMLR, 17–23 Jul 2022.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017. doi: 10.1109/TNNLS.2016.2599820.

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.

Harshay Shah, Prateek Jain, and Praneeth Netrapalli. Do input gradients highlight discriminative features? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 2046–2059. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0fe6a94848e5c68a54010b61b3e94b0e-Paper.pdf.

Ali Shahrokni and Robert Feldt. A systematic review of software robustness. *Information and Software Technology*, 55(1):1–17, 2013. ISSN 0950-5849. doi: https://doi.org/10.1016/j.infsof.2012.06.002. URL https://www.sciencedirect.com/science/article/pii/S0950584912001048. Special section: Best papers from the 2nd International Symposium on Search Based Software Engineering 2010.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

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

Suraj Srinivas, Sebastian Bordt, and Himabindu Lakkaraju. Which models have perceptually-aligned gradients? an explanation via off-manifold robustness. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 21172–21195. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/42bbe2bfdbbfcadda643e8f89025716c-Paper-Conference.pdf.

Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020. doi: 10.23915/distill.00022. https://distill.pub/2020/attribution-baselines.

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

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. URL https://arxiv.org/abs/1905.11946.

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics, 2019. URL https://arxiv.org/abs/1912.01451.

Michal Uřičář, Jan Ulićný, Ganesh Sistu, Hazem Rashed, Pavel Křížek, David Hurych, Antonín Vobecký, and Senthil Yogamani. Desoiling dataset: Restoring soiled areas on automotive fisheye cameras. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 4273–4279, 2019. doi: 10.1109/ICCVW.2019.00526.

Jörg Wagner, Jan Mathias Köhler, Tobias Gindele, Leon Hetzel, Jakob Thaddäus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9089–9099, 2019. doi: 10.1109/CVPR.2019.00931.

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

Pei Wang and Nuno Vasconcelos. Scout: Self-aware discriminant counterfactual explanations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8978–8987, 2020. doi: 10.1109/CVPR42600.2020.00900.

A. Reza Yazdanpanah, Xiaolong Liu, Ning Li, and Jindong Tan. A novel laparoscopic camera robot with in-vivo lens cleaning and debris prevention modules. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3669–3674, 2017. doi: 10.1109/IROS.2017.8206212.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in)fidelity and sensitivity of explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/a7471fdc77b3435276507cc8f2dc2569-Paper.pdf.

Shaodi You, Robby T. Tan, Rei Kawakami, Yasuhiro Mukaigawa, and Katsushi Ikeuchi. Adherent raindrop modeling, detection and removal in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1721–1733, 2016. doi: 10.1109/TPAMI.2015.2491937.

Xu Zheng, Farhad Shirani, Zhuomin Chen, Chaohao Lin, Wei Cheng, Wenbo Guo, and Dongsheng Luo. F-fidelity: A robust framework for faithfulness evaluation of explainable ai. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 12772–12804, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2231d1abdcc68d16c9861a58f3c25849-Paper-Conference.pdf.

Zhiming Zhou, Han Cai, Shu Rong, Yuxuan Song, Kan Ren, Weinan Zhang, Yong Yu, and Jun Wang. Activation maximization generative adversarial nets, 2018. URL https://arxiv.org/abs/1703.02000.

A Appendix

B ImageNet Results

Method	**Random$_{\text{All}}$**	**Random$_{\text{Obj}}$**	**Random$_{\text{Back}}$**	**SHAP**	**GradCam (SC)**
AUC^{rel}	0.1661(0.1235-0.2180)	0.0613(0.0457-0.0798)	0.1707(0.1279-0.2188)	0.0702(0.0550-0.0890)	0.1220(0.0900-0.1600)
AUC^{rel}_{Band}	0.4446(0.3715-0.5344)	0.2484(0.1931-0.3179)	0.4548(0.3792-0.5366)	0.3660(0.3112-0.4366)	0.5073(0.4299-0.5933)
∇B^{rel}	0.2173(0.1620-0.2844)	0.0837(0.0647-0.1068)	0.2239(0.1686-0.2874)	0.0978(0.0791-0.1210)	0.1594(0.1210-0.2034)
$\text{CP}^{rel}_{0.5}$	0.1523(0.1114-0.2012)	0.0526(0.0384-0.0692)	0.1569(0.1164-0.2031)	0.0551(0.0412-0.0722)	0.1048(0.0738-0.1417)
$\text{CP}^{rel}_{0.8}$	0.1028(0.0719-0.1392)	0.0339(0.0243-0.0460)	0.1060(0.0755-0.1411)	0.0316(0.0228-0.0423)	0.0699(0.0444-0.0996)
$\overline{\text{AUC}^{abs}}$	0.1646(0.1217-0.2167)	0.1132(0.0880-0.1430)	0.1693(0.1252-0.2177)	0.1100(0.0929-0.1302)	0.1740(0.1472-0.2047)
AUC^{abs}_{Band}	0.4302(0.3571-0.5186)	0.3909(0.3312-0.4655)	0.4412(0.3657-0.5223)	0.4354(0.3842-0.4961)	0.5784(0.5256-0.6360)
∇B^{abs}	0.2177(0.1845-0.2566)	0.2142(0.1585-0.2820)	0.1482(0.1165-0.1859)	0.2215(0.1649-0.2846)	0.1393(0.1171-0.1664)
$\text{CP}^{abs}_{0.5}$	0.1506(0.1101-0.1990)	0.1006(0.0767-0.1284)	0.1547(0.1131-0.2003)	0.0920(0.0756-0.1113)	0.1513(0.1259-0.1811)
$\text{CP}^{abs}_{0.8}$	0.1015(0.0709-0.1375)	0.0662(0.0493-0.0870)	0.1044(0.0735-0.1397)	0.0571(0.0457-0.0706)	0.0982(0.0776-0.1223)

Table 3: ImageNet-S50 Gao et al. (2022) degradation metrics with 95% bootstrap CI 10,000 samples

Method	**S-GC**	**IG**	**AM**	**OS**	**LossG.**	**RISE**
AUC^{rel}	0.0859(0.0624-0.1131)	0.1287(0.1001-0.1631)	0.1081(0.0830-0.1369)	0.1246(0.1009-0.1518)	0.1296(0.1044-0.1585)	0.1307(0.1092-0.1566)
AUC^{rel}_{Band}	0.4018(0.3343-0.4842)	0.5445(0.4861-0.6076)	0.4793(0.4204-0.5491)	0.6586(0.5882-0.7248)	0.5736(0.5141-0.6338)	0.5868(0.5402-0.6422)
∇B^{rel}	0.1120(0.0845-0.1442)	0.1672(0.1316-0.2098)	0.1428(0.1116-0.1799)	0.1639(0.1334-0.1997)	0.1684(0.1378-0.2031)	0.1663(0.1395-0.1993)
$\text{CP}^{rel}_{0.5}$	0.0730(0.0510-0.0988)	0.1127(0.0852-0.1454)	0.0931(0.0696-0.1209)	0.1030(0.0807-0.1288)	0.1132(0.0889-0.1412)	0.1159(0.0949-0.1411)
$\text{CP}^{rel}_{0.8}$	0.0466(0.0304-0.0661)	0.0726(0.0514-0.0979)	0.0581(0.0406-0.0783)	0.0577(0.0437-0.0739)	0.0720(0.0534-0.0940)	0.0771(0.0616-0.0955)
$\overline{\text{AUC}^{abs}}$	0.1301(0.1060-0.1582)	0.1568(0.1283-0.1897)	0.1480(0.1216-0.1785)	0.1331(0.1104-0.1591)	0.1636(0.1379-0.1937)	0.1544(0.1320-0.1810)
AUC^{abs}_{Band}	0.4664(0.4122-0.5303)	0.5222(0.4705-0.5817)	0.5091(0.4556-0.5714)	0.5927(0.5291-0.6595)	0.5520(0.4985-0.6080)	0.5811(0.5344-0.6354)
∇B^{abs}	0.1602(0.1307-0.1947)	0.1967(0.1600-0.2389)	0.1868(0.1527-0.2272)	0.1713(0.1405-0.2068)	0.2029(0.1709-0.2406)	0.1922(0.1636-0.2270)
$\text{CP}^{abs}_{0.5}$	0.1115(0.0886-0.1376)	0.1364(0.1093-0.1677)	0.1279(0.1028-0.1574)	0.1108(0.0895-0.1354)	0.1416(0.1165-0.1710)	0.1356(0.1134-0.1619)
$\text{CP}^{abs}_{0.8}$	0.0710(0.0547-0.0908)	0.0883(0.0681-0.1127)	0.0818(0.0632-0.1039)	0.0636(0.0497-0.0794)	0.0909(0.0718-0.1130)	0.0884(0.0716-0.1080)

Table 4: ImageNet-S50 Gao et al. (2022) degradation metrics with 95% bootstrap CI 10,000 samples

C Oxford Flowers Results

Method	**Random$_{\text{All}}$**	**Random$_{\text{Obj}}$**	**Random$_{\text{Back}}$**	**SHAP**	**GradCam (SC)**
AUC^{rel}	0.5183(0.5046-0.5319)	0.2761(0.2657-0.2865)	0.5179(0.5039-0.5319)	0.1788(0.1772-0.1803)	0.4468(0.4367-0.4569)
AUC^{rel}_{Band}	0.8121(0.8008-0.8234)	0.6996(0.6772-0.7221)	0.8119(0.8003-0.8234)	0.5731(0.5657-0.5806)	0.8851(0.8842-0.8860)
∇B^{rel}	0.6002(0.5782-0.6222)	0.3024(0.2871-0.3177)	0.6019(0.5817-0.6222)	0.1655(0.1610-0.1699)	0.5029(0.4874-0.5185)
$\text{CP}^{rel}_{0.5}$	0.5132(0.4814-0.5450)	0.2633(0.2483-0.2782)	0.5129(0.4808-0.5450)	0.1537(0.1499-0.1575)	0.4489(0.4401-0.4577)
$\text{CP}^{rel}_{0.8}$	0.3551(0.3298-0.3805)	0.1758(0.1684-0.1831)	0.3551(0.3298-0.3805)	0.0885(0.0823-0.0947)	0.3135(0.3128-0.3141)
$\overline{\text{AUC}^{abs}}$	0.5182(0.5046-0.5317)	0.3720(0.3605-0.3835)	0.5179(0.5040-0.5317)	0.2182(0.2165-0.2198)	0.4539(0.4441-0.4637)
AUC^{abs}_{Band}	0.8088(0.7966-0.8210)	0.7453(0.7249-0.7658)	0.8084(0.7958-0.8210)	0.5996(0.5915-0.6078)	0.8292(0.8259-0.8325)
∇B^{abs}	0.6014(0.5816-0.6212)	0.4207(0.4061-0.4354)	0.6025(0.5838-0.6212)	0.2114(0.2083-0.2145)	0.5123(0.4978-0.5267)
$\text{CP}^{abs}_{0.5}$	0.5319(0.5196-0.5443)	0.3739(0.3646-0.3831)	0.5316(0.5189-0.5443)	0.1959(0.1918-0.1999)	0.4557(0.4474-0.4640)
$\text{CP}^{abs}_{0.8}$	0.3682(0.3582-0.3781)	0.2533(0.2510-0.2556)	0.3676(0.3571-0.3781)	0.1151(0.1075-0.1228)	0.3118(0.3105-0.3131)

Table 5: Oxford Flowers Nilsback & Zisserman (2008) degradation metrics with 95% bootstrap CI 10,000 samples

Method	**S-GC**	**IG**	**AM**	**OS**	**LossG.**	**RISE**
AUC^{rel}	0.2632(0.2588-0.2676)	0.2785(0.2745-0.2824)	0.2891(0.2846-0.2935)	0.2857(0.2783-0.2932)	0.3134(0.3082-0.3186)	0.3283(0.3158-0.3408)
AUC^{rel}_{Band}	0.7563(0.7453-0.7673)	0.7504(0.7400-0.7608)	0.7742(0.7637-0.7847)	0.8311(0.8246-0.8377)	0.7940(0.7869-0.8010)	0.8001(0.7885-0.8116)
∇B^{rel}	0.2683(0.2600-0.2765)	0.2874(0.2796-0.2953)	0.2976(0.2908-0.3043)	0.2881(0.2794-0.2967)	0.3294(0.3215-0.3372)	0.3417(0.3271-0.3563)
$\text{CP}^{rel}_{0.5}$	0.2457(0.2427-0.2487)	0.2625(0.2604-0.2645)	0.2734(0.2702-0.2766)	0.2664(0.2596-0.2732)	0.3009(0.2974-0.3045)	0.3179(0.3076-0.3283)
$\text{CP}^{rel}_{0.8}$	0.1522(0.1493-0.1552)	0.1625(0.1580-0.1671)	0.1698(0.1660-0.1736)	0.1515(0.1509-0.1522)	0.1956(0.1938-0.1974)	0.2214(0.2168-0.2260)
AUC^{abs}	0.2881(0.2834-0.2928)	0.3200(0.3162-0.3237)	0.3146(0.3101-0.3191)	0.2885(0.2812-0.2957)	0.3316(0.3262-0.3370)	0.3454(0.3322-0.3586)
AUC^{abs}_{Band}	0.7350(0.7250-0.7450)	0.7539(0.7456-0.7622)	0.7624(0.7544-0.7703)	0.7988(0.7898-0.8079)	0.7642(0.7575-0.7709)	0.7996(0.7862-0.8131)
∇B^{abs}	0.2947(0.2857-0.3038)	0.3385(0.3305-0.3464)	0.3285(0.3201-0.3369)	0.2953(0.2853-0.3054)	0.3493(0.3414-0.3571)	0.3628(0.3479-0.3778)
$\text{CP}^{abs}_{0.5}$	0.2712(0.2676-0.2747)	0.3070(0.3051-0.3089)	0.3010(0.2976-0.3044)	0.2693(0.2630-0.2757)	0.3196(0.3157-0.3235)	0.3361(0.3245-0.3477)
$\text{CP}^{abs}_{0.8}$	0.1650(0.1615-0.1686)	0.1910(0.1854-0.1967)	0.1858(0.1811-0.1904)	0.1534(0.1530-0.1538)	0.2037(0.2011-0.2063)	0.2306(0.2252-0.2360)

Table 6: Oxford Flowers Nilsback & Zisserman (2008) degradation metrics with 95% bootstrap CI 10,000 samples

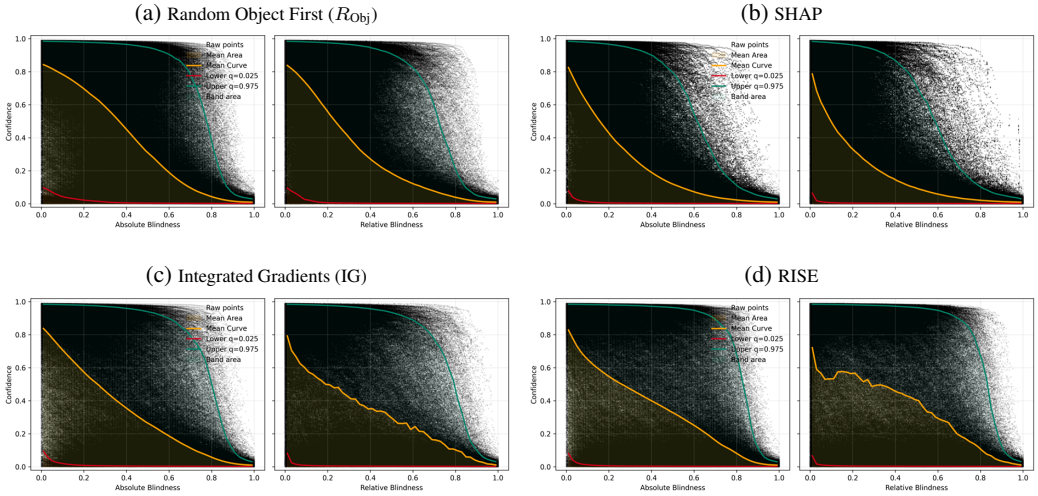

Figure 3: Example degradation curves for different guidance strategies on Oxford Flowers Nilsback & Zisserman (2008) using a ViT Dosovitskiy et al. (2021) model under MDP-14.

D Oxford-Pets Results

Method	$\textbf{Random}_{\textbf{All}}$	$\textbf{Random}_{\textbf{Obj}}$	$\textbf{Random}_{\textbf{Back}}$	**SHAP**	**GradCam (SC)**
AUC^{rel}	0.0432(0.0423-0.0441)	0.0307(0.0305-0.0309)	0.0436(0.0423-0.0449)	0.0142(0.0137-0.0146)	0.0294(0.0286-0.0302)
AUC^{rel}_{Band}	0.6864(0.6726-0.6956)	0.2204(0.2145-0.2264)	0.6910(0.6716-0.7123)	0.1616(0.1562-0.1659)	0.2820(0.2660-0.3026)
$\text{CP}^{rel}_{0.5}$	0.0306(0.0298-0.0314)	0.0081(0.0079-0.0083)	0.0313(0.0302-0.0325)	0.0085(0.0082-0.0089)	0.0116(0.0109-0.0120)
∇B^{rel}	0.8525(0.8457-0.8591)	0.6097(0.5905-0.6289)	0.8538(0.8478-0.8598)	0.7580(0.7505-0.7657)	0.7530(0.7422-0.7608)
$\text{CP}^{rel}_{0.8}$	0.0244(0.0237-0.0250)	0.0065(0.0063-0.0066)	0.0249(0.0238-0.0260)	0.0056(0.0054-0.0058)	0.0079(0.0075-0.0082)
AUC^{abs}	0.0431(0.0422-0.0440)	0.0378(0.0374-0.0383)	0.0434(0.0422-0.0447)	0.0195(0.0189-0.0200)	0.0339(0.0330-0.0347)
AUC^{abs}_{Band}	0.6893(0.6749-0.7001)	0.4607(0.4515-0.4734)	0.6928(0.6756-0.7131)	0.2732(0.2644-0.2819)	0.4411(0.4213-0.4644)
∇B^{abs}	0.8625(0.8561-0.8673)	0.7390(0.7249-0.7531)	0.8631(0.8587-0.8675)	0.8429(0.8380-0.8488)	0.8110(0.8034-0.8165)
$\text{CP}^{abs}_{0.5}$	0.0305(0.0298-0.0310)	0.0199(0.0196-0.0203)	0.0310(0.0299-0.0322)	0.0132(0.0130-0.0137)	0.0181(0.0172-0.0186)
$\text{CP}^{abs}_{0.8}$	0.0243(0.0233-0.0250)	0.0159(0.0156-0.0161)	0.0244(0.0232-0.0257)	0.0097(0.0092-0.0101)	0.0134(0.0128-0.0140)

Table 7: Oxford-Pets Parkhi et al. (2012) degradation metrics with 95% bootstrap CI 10,000 samples

Method	S-GC	IG	AM	OS	LossG.	RISE
AUC^{rel}	0.0317(0.0311-0.0321)	0.0370(0.0364-0.0375)	0.0369(0.0363-0.0375)	0.0160(0.0155-0.0165)	0.0345(0.0338-0.0351)	0.0218(0.0212-0.0224)
AUC^{rel}_{Band}	0.2666(0.2619-0.2734)	0.4463(0.4332-0.4593)	0.4230(0.4112-0.4349)	0.1560(0.1496-0.1607)	0.3573(0.3468-0.3712)	0.3506(0.3329-0.3645)
$\text{CP}^{rel}_{0.5}$	0.6621(0.6429-0.6734)	0.7516(0.7404-0.7594)	0.7168(0.7028-0.7276)	0.6835(0.6677-0.6993)	0.7079(0.6913-0.7224)	0.6882(0.6786-0.6943)
∇B^{rel}	0.0110(0.0099-0.0116)	0.0207(0.0201-0.0212)	0.0194(0.0189-0.0198)	0.0101(0.0099-0.0103)	0.0158(0.0153-0.0162)	0.0169(0.0165-0.0171)
$\text{CP}^{rel}_{0.8}$	0.0082(0.0073-0.0088)	0.0157(0.0148-0.0163)	0.0148(0.0143-0.0153)	0.0064(0.0061-0.0067)	0.0124(0.0119-0.0128)	0.0132(0.0128-0.0136)
AUC^{abs}	0.0355(0.0349-0.0360)	0.0387(0.0380-0.0393)	0.0386(0.0378-0.0392)	0.0173(0.0167-0.0178)	0.0382(0.0374-0.0389)	0.0247(0.0241-0.0254)
AUC^{abs}_{Band}	0.4230(0.4140-0.4319)	0.5396(0.5286-0.5505)	0.5319(0.5184-0.5435)	0.1848(0.1787-0.1908)	0.4944(0.4797-0.5083)	0.4041(0.3823-0.4215)
∇B^{abs}	0.7392(0.7200-0.7556)	0.7979(0.7876-0.8040)	0.7816(0.7699-0.7906)	0.8195(0.8075-0.8316)	0.7723(0.7611-0.7806)	0.8396(0.8338-0.8485)
$\text{CP}^{abs}_{0.5}$	0.0170(0.0165-0.0174)	0.0237(0.0230-0.0241)	0.0229(0.0222-0.0236)	0.0108(0.0106-0.0110)	0.0209(0.0202-0.0214)	0.0186(0.0181-0.0188)
$\text{CP}^{abs}_{0.8}$	0.0129(0.0122-0.0134)	0.0179(0.0170-0.0186)	0.0178(0.0169-0.0184)	0.0070(0.0066-0.0073)	0.0163(0.0156-0.0168)	0.0146(0.0142-0.0150)

Table 8: Oxford-Pets Parkhi et al. (2012) degradation metrics with 95% bootstrap CI 10,000 samples

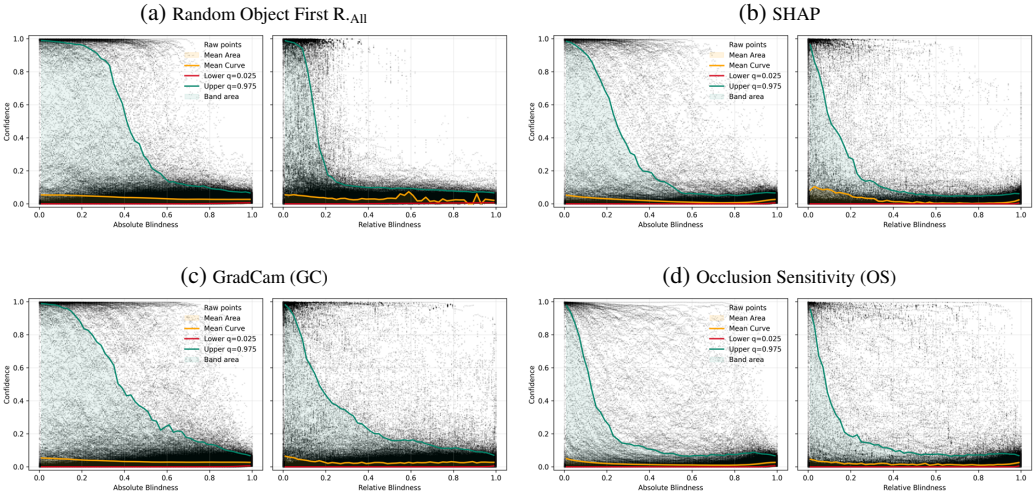

Figure 4: Example degradation curves on Oxford-IIIT Pets Parkhi et al. (2012) (ViT-B16 Dosovitskiy et al. (2021), MDP-14).

E Results Configuration Effect Ablation Study

Strategy:	**Random Object First (R_{Obj})**					**SHAP**				
Config.	**AUC↓**	**AUC_{band}↓**	**∇B↓**	**$\text{CP}_{0.5}$↓**	**$\text{CP}_{0.8}$↓**	**AUC↓**	**AUC_{band}↓**	**∇B↓**	**$\text{CP}_{0.5}$↓**	**$\text{CP}_{0.8}$↓**
RN50 MDP-7	9.65/14.02	33.94/45.20	12.02/16.57	8.61/12.48	5.80/8.64	5.85/10.86	29.85/41.55	7.92/12.44	4.32/9.12	2.56/6.01
RN50 MDP-14	3.04/6.60	12.59/25.14	4.55/8.16	2.63/5.82	1.76/4.03	4.01/8.11	25.04/32.22	5.01/8.87	2.89/6.61	1.77/4.44
RN50 MDP-32	0.77/2.83	3.36/14.00	1.62/3.75	0.62/2.38	0.35/1.45	3.02/6.47	19.71/26.30	4.35/7.42	2.02/4.89	1.23/3.22
RN50 Black-14	3.37/7.03	14.80/30.14	4.88/8.96	2.84/6.05	1.89/4.15	4.2978.38	26.70/35.14	6.27/10.17	3.10/6.69	1.94/4.64
RN50 White-14	3.96/8.25	17.53/31.55	5.88/10.59	3.32/7.10	2.16/4.86	4.96/9.29	30.74/38.06	6.88/11.07	3.58/7.52	2.30/5.15
RN50 Noise-14	3.98/8.51	15.58/29.72	5.01/10.42	3.29/7.14	2.13/4.72	4.93/9.29	28.83/36.21	7.31/11.68	3.45/7.40	2.03/4.94
ViT MDP-7	21.52/30.37	67.00/78.35	27.21/38.92	19.84/28.30	14.51/20.82	16.03/21.76	67.71/73.17	20.83/27.59	14.49/20.13	9.74/14.66
ViT MDP-14	14.36/27.60	57.02/76.64	18.26/35.40	12.84/25.56	9.47/19.34	15.72/20.85	71.58/74.15	19.84/26.12	13.87/18.99	9.28/13.69
ViT MDP-32	1.34/4.94	5.89/18.76	02.17/6.46	0.87/4.21	0.53/2.66	6.34/9.98	37.34/42.39	8.05/11.73	4.78/8.01	2.96/5.27
ViT Black-14	14.67/27.80	61.89/77.37	18.22/36.41	12.52/25.33	9.12/19.07	16.56/21.54	73.20/75.24	21.50/28.08	14.57/19.64	9.87/14.17
ViT White-14	15.11/28.54	61.60/78.71	18.49/35.95	12.97/26.29	9.60/19.73	16.69/21.78	76.01/78.71	21.64/28.20	14.57/19.77	9.97/14.41
ViT Noise-14	13.55/26.74	52.73/76.51	17.58/34.49	11.92/24.62	8.66/18.29	17.53/22.17	75.33/76.41	21.55/27.04	15.35/20.01	10.20/14.21

Table 9: Ablation study on the effect of perturbation method and patch size for ResNet50 (RN50) and ViT. We report relative/absolute degradation metrics ($\cdot 10^{-3}$) for Random Object First (R_{Obj}) and SHAP.

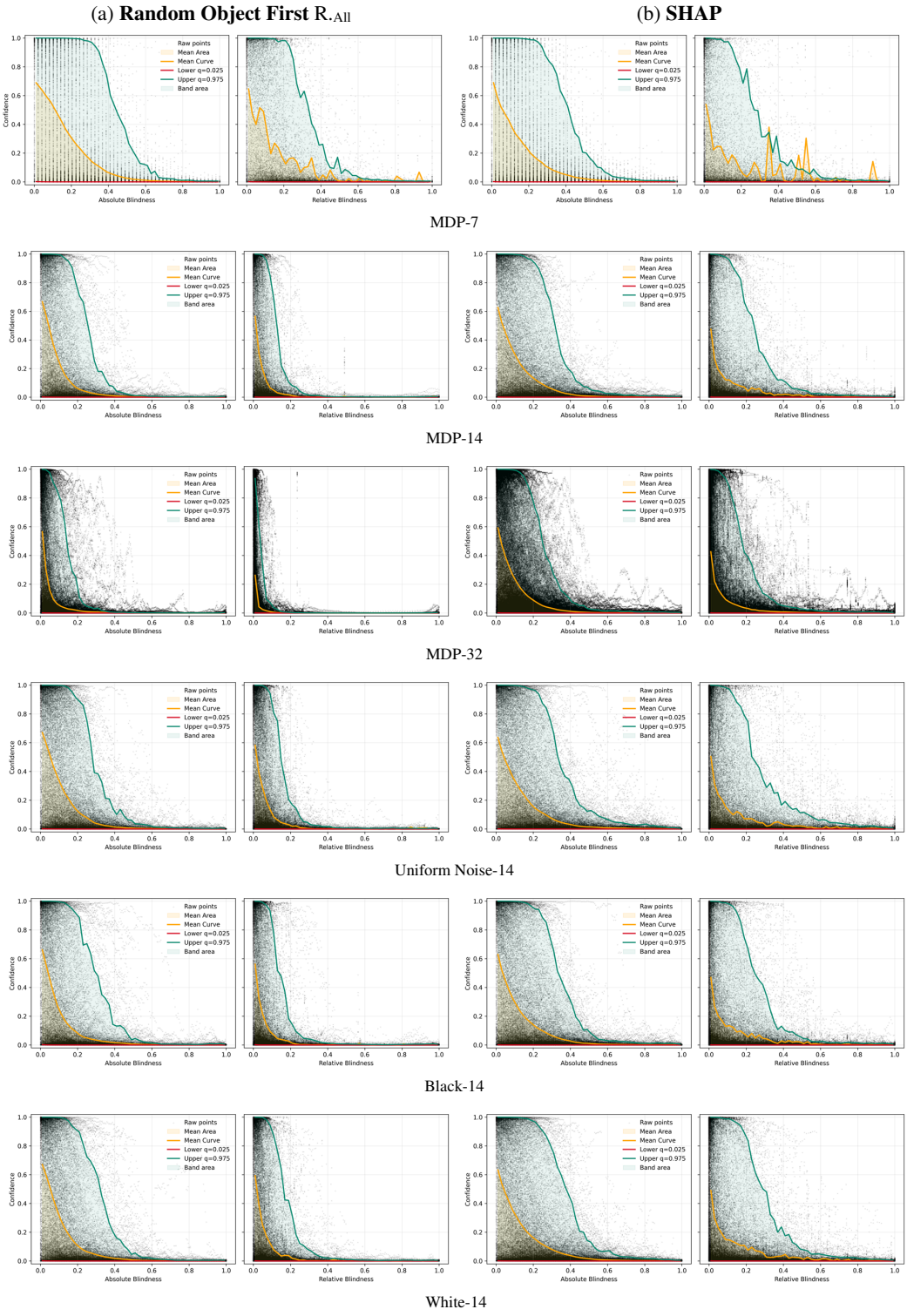

Figure 5: Degradation curves of **ResNet50** He et al. (2015) with different configurations

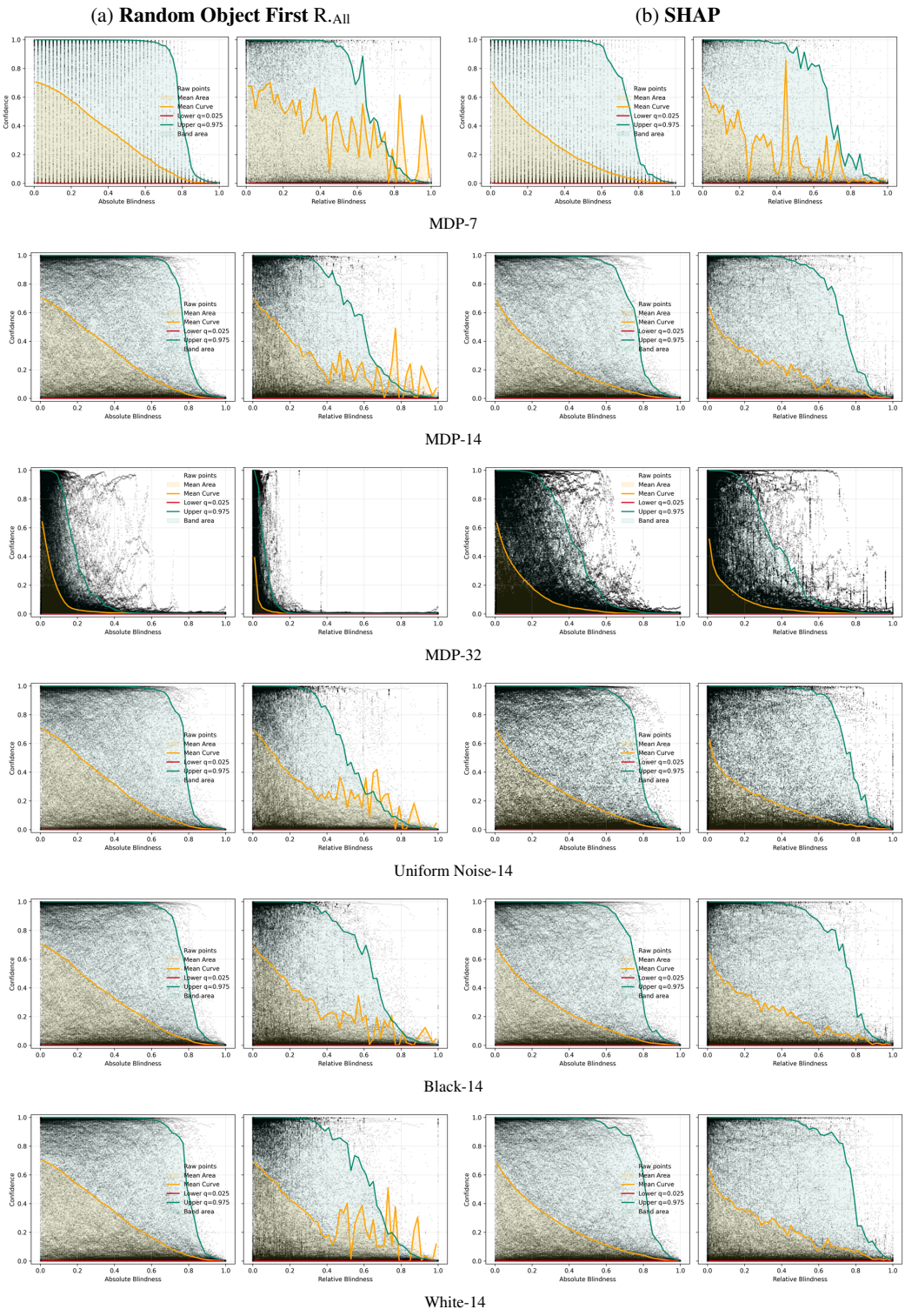

Figure 6: Degradation curves of ViT Dosovitskiy et al. (2021) with different configurations

F Application Example

We illustrate how degradation curves can be used as a practical robustness diagnostic on the first 12 classes of ImageNet-S50 Gao et al. (2022). In this application, we focus on **object-relative** degradation. Masking is restricted to the object region and the x-axis is the relative blindness ratio r^{rel}. This isolates evidence removal on the foreground and avoids confounds from background masking. Unless stated otherwise, we use a fixed **random object-first** guidance strategy ($\mathrm{R}_{\cdot\mathrm{Obj}}$) together with MDP-14. We treat this setting as a model probe rather than an attribution benchmark. The goal is to expose how quickly confidence collapses when object evidence is progressively removed under a standardized intervention budget.

F.1 Class-wise Model Analysis

To demonstrate how degradation metrics can be used in practice, we analyze robustness at the *class level*. Rather than averaging across the entire dataset, we compute collapse dynamics separately for each semantic category. This enables a targeted robustness diagnosis: which object classes require substantial evidence removal before prediction failure, and which collapse under minimal perturbation?

F.1.1 ResNet34

We begin with ResNet34 to demonstrate how degradation curves expose class-level robustness differences under identical masking budgets. Fig. 7 shows clear class-dependent collapse dynamics. Classes such as *Siamese cat* and *hamster* retain high confidence until relatively large blindness levels, indicating delayed destabilization. Visually, *Siamese cat* stands out by the absence of low-confidence outliers for small r^{rel} (≤ 0.2). Early masking does not produce scattered failures, and confidence remains tightly concentrated near 1.0. This suggests stable early evidence retention followed by a more threshold-like breakdown. In contrast, *kuvasz* and *red fox* exhibit early confidence erosion, with noticeable dispersion even at small blindness levels, indicating reliance on more localized or fragile cues. Table 10 quantifies these patterns. Higher AUC and later CP values for *Siamese cat* and *hamster* confirm stronger evidence redundancy, while low $\mathrm{CP}_{0.8}$ for *kuvasz* and *red fox* reveals early-collapse regimes. The sensitivity slope ∇B differentiates abrupt from gradual (e.g., *goldfish*) breakdown, and $\mathrm{AUC}_{\mathrm{Band}}$ exposes intra-class variability, particularly pronounced for *ladybug*.

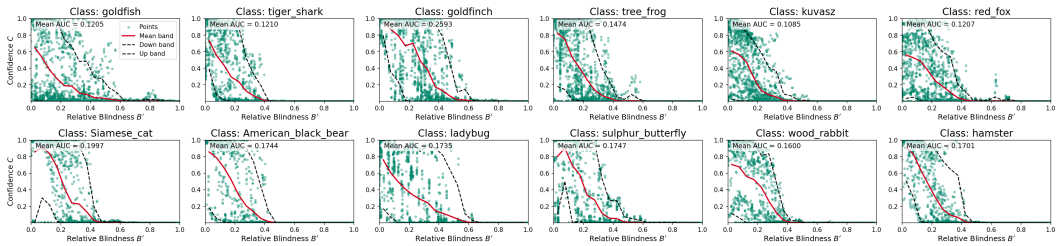

Figure 7: Class-wise object-relative degradation curves for ResNet34 He et al. (2015) on ImageNet-S50 Gao et al. (2022) under random object-first masking (MDP-14). Points show per-image confidence along the masking trajectory. The red curve is the class mean and dashed curves indicate the quantile band.

Class	n_j	AUC	AUC_{Band}	∇B	$CP_{0.5}$	$CP_{0.8}$
goldfish	15	0.1205	0.3374	0.1422	0.1037	0.0762
tiger shark	10	0.1210	0.2483	0.1755	0.1099	0.0606
goldfinch	15	0.2593	0.3935	0.2823	0.2322	0.1818
tree frog	16	0.1474	0.2837	0.1684	0.1292	0.0852
kuvasz	14	0.1085	0.2235	0.1618	0.0882	0.0326
red fox	15	0.1207	0.2727	0.1830	0.0969	0.0409
Siamese cat	15	0.1997	0.3369	0.2336	0.2059	0.1630
American black bear	13	0.1744	0.3248	0.1907	0.1813	0.1568
ladybug	18	0.1735	0.4437	0.1848	0.1765	0.0945
sulphur butterfly	16	0.1747	0.3064	0.1888	0.1752	0.1362
wood rabbit	15	0.1600	0.2842	0.1871	0.1667	0.0836
hamster	17	0.1701	0.2909	0.1903	0.1724	0.1210

Table 10: Class-wise robustness metrics for ResNet34 He et al. (2015) on ImageNet-S50 Gao et al. (2022) under random object-first masking (MDP-14). Metrics are computed using object-relative blindness and summarize confidence retention (AUC), collapse sharpness (∇B), collapse onset (CP_τ), and inter-instance variability (AUC_{Band}).

Overall, ResNet34 shows heterogeneous robustness across categories, with some classes exhibiting delayed threshold-like collapse and others destabilizing under minimal object erosion.

F.1.2 ResNet50

We next examine how increased depth alters collapse dynamics. Fig. 8 indicates that several classes retain confidence longer compared to ResNet34. For example, *goldfinch* and *tiger shark* show visibly delayed collapse, with mean trajectories remaining elevated over larger blindness intervals. However, early-collapse regimes persist for categories such as *red fox*, suggesting that depth does not uniformly mitigate localized feature reliance.

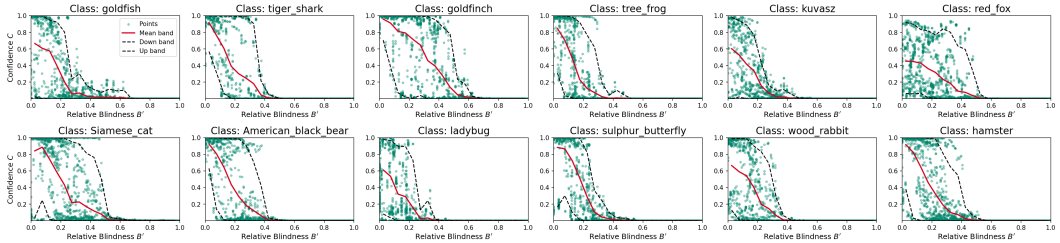

Figure 8: Class-wise object-relative degradation curves for ResNet50 He et al. (2015) on ImageNet-S50 Gao et al. (2022) under random object-first masking (MDP-14). Points show per-image confidence along the masking trajectory; the red curve is the class mean and dashed curves indicate the quantile band.

Table 11 confirms these trends: increased AUC and later CP for several classes reflect improved evidence retention, particularly for structurally distinctive categories (e.g., *goldfinch*). Nevertheless, low $CP_{0.8}$ values remain for certain classes, demonstrating that additional depth enhances robustness selectively rather than globally.

In summary, ResNet50 increases spatial evidence redundancy for some categories but does not eliminate vulnerability regimes, revealing that robustness improvements are class-dependent rather than universal.

F.1.3 VGG16

Finally, we analyze VGG16 to contrast depth with architectural design. Fig. 9 shows generally earlier and more uniform confidence decay across multiple classes. Compared to both residual models,

Class	n_j	AUC	AUC_{Band}	∇B	$CP_{0.5}$	$CP_{0.8}$
goldfish	15	0.1311	0.2678	0.1558	0.1176	0.0908
tiger shark	10	0.1684	0.2995	0.1776	0.1702	0.1502
goldfinch	15	0.3020	0.4332	0.3136	0.3108	0.2521
tree frog	16	0.1227	0.2483	0.1390	0.1204	0.0965
kuvasz	14	0.1007	0.1969	0.1064	0.0908	0.0430
red fox	15	0.1252	0.3432	0.2361	0.1003	0.0259
Siamese cat	15	0.2073	0.4113	0.2120	0.1791	0.1572
American black bear	13	0.1729	0.3041	0.1829	0.1770	0.1449
ladybug	18	0.0970	0.2151	0.1497	0.0909	0.0548
sulphur butterfly	16	0.1637	0.2298	0.1773	0.1542	0.1273
wood rabbit	15	0.1312	0.2581	0.1449	0.1213	0.0888
hamster	17	0.1757	0.3254	0.1687	0.1659	0.1438

Table 11: Class-wise robustness metrics for ResNet50 He et al. (2015) on ImageNet-S50 Gao et al. (2022) under random object-first masking (MDP-14). Object-relative blindness is used for all measurements. Increased AUC and later CP values indicate delayed confidence collapse compared to shallower architectures.

trajectories tend to decline more steadily without pronounced delayed-collapse behavior. Early dispersion is visible in several categories, indicating reduced early-stage stability.

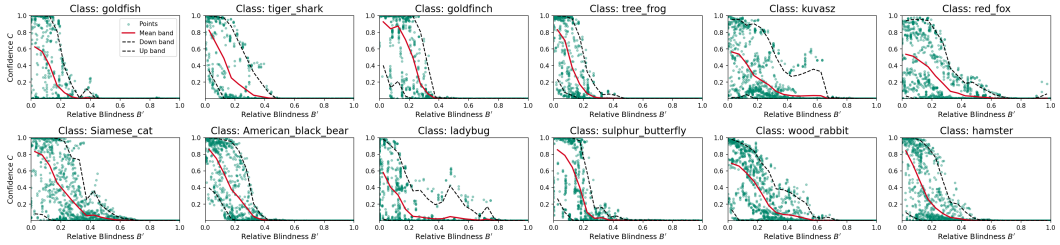

Figure 9: Class-wise object-relative degradation curves for VGG16 Simonyan & Zisserman (2015) on ImageNet-S50 Gao et al. (2022) under random object-first masking (MDP-14). Points show per-image confidence along the masking trajectory; the red curve is the class mean and dashed curves indicate the quantile band.

Table 12 supports this observation: lower AUC and earlier CP values are prevalent across classes such as *goldfish*, *tree frog*, and *ladybug*. Even for *Siamese cat*, robustness is reduced relative to ResNet34 and ResNet50, with earlier collapse and increased inter-instance variability (AUC_{Band}).

These results suggest that, despite substantial capacity, the non-residual VGG16 architecture distributes evidence less redundantly across object regions, leading to increased sensitivity under structured masking. Overall, the comparison across architectures demonstrates how degradation curves diagnose not only which classes are fragile, but also how architectural depth and design influence evidence distribution and collapse behavior.

F.2 Model Comparison

Table 13 compares six `torchvision` architectures (ResNet34/50, VGG16, EfficientNet, DenseNet, ViT-B16) on ImageNet-S50 under object-relative masking, using Random Object-First guidance (R_{Obj}), MDP perturbation, and patch size 32. Consistent with our robustness interpretation, higher AUC and higher CP indicate more retained confidence under degradation, while larger ∇B indicates more abrupt collapse. EfficientNet achieves the highest AUC/CP (most retained confidence), whereas ResNet34/50 and VGG16 show the lowest AUC/CP (earliest collapse). ViT-B16 exhibits intermediate robustness with moderate variability (AUC_{Band}).

Class	$\mathbf{n_j}$	AUC	$\mathbf{AUC}_{Band}$	$\nabla\mathbf{B}$	$\mathbf{CP_{0.5}}$	$\mathbf{CP_{0.8}}$
goldfish	15	0.0866	0.2007	0.1000	0.0893	0.0677
tiger shark	10	0.1235	0.2380	0.1490	0.1234	0.0898
goldfinch	15	0.1914	0.2590	0.1980	0.1925	0.1607
tree frog	16	0.1026	0.1803	0.1084	0.1051	0.0732
kuvasz	14	0.1186	0.3565	0.1778	0.1015	0.0416
red fox	15	0.1240	0.3087	0.1395	0.1079	0.0514
Siamese cat	15	0.1758	0.3397	0.1835	0.1504	0.1140
American black bear	13	0.1494	0.2252	0.1625	0.1471	0.0970
ladybug	18	0.0805	0.2729	0.1497	0.0603	0.0270
sulphur butterfly	16	0.1411	0.2130	0.1543	0.1244	0.1075
wood rabbit	15	0.1633	0.3044	0.1762	0.1581	0.0715
hamster	17	0.1223	0.2539	0.1182	0.1190	0.0908

Table 12: Class-wise robustness metrics for VGG16 Simonyan & Zisserman (2015) on ImageNet-S50 Gao et al. (2022) under random object-first masking (MDP-14). Lower AUC and earlier CP values reflect earlier collapse under structured object erosion, highlighting architectural differences in evidence distribution.

Model	AUC	$\mathbf{AUC}_{Band}$	$\nabla\mathbf{B}$	$\mathbf{CP_{0.5}}$	$\mathbf{CP_{0.8}}$
ResNet34	0.0065	0.0276	0.0160	0.0045	0.0026
ResNet50	0.0077	0.0336	0.0162	0.0062	0.0035
VGG16	0.0061	0.0285	0.0209	0.0042	0.0023
EfficientNet	0.0188	0.0926	0.0252	0.0132	0.0064
DenseNet	0.0094	0.0456	0.0170	0.0062	0.0036
ViT	0.0134	0.0589	0.0217	0.0087	0.0053

Table 13: Mean relative robustness metrics across applied models on ImageNet-S50 (R_{Obj}, MDP; patch size 32).