Beyond Gradients: Curvature-Aware Part-Level Explanations for Vision Models

Maisha Maliha
School of Computer Science
University of Oklahoma, Norman, OK, USA
maisha.maliha-1@ou.edu

Dean F. Hougen
School of Computer Science
University of Oklahoma, Norman, OK, USA
hougen@ou.edu

Abstract

Despite the strong predictive performance of modern vision models, understanding which image regions drive their decisions remains challenging. Pixel-level saliency methods typically rely on first-order gradients, which are effective for local perturbations but can become inadequate when attribution is aggregated over larger structured regions. As region-level perturbations move inputs beyond the local linear regime, higher-order effects become increasingly relevant. We propose a curvature-aware attribution framework that scores each region using symmetric Kullback–Leibler (KL) divergence modulated by directional Hessian information, capturing both belief shift and nonlinear amplification within structured parts. This formulation unifies probability-space sensitivity with second-order structure, yielding stable region-level importance rankings without retraining or architectural modification. Extensive experiments across multiple architectures and datasets spanning natural images and faces demonstrate consistent improvements over first-order baselines. We analyze segmentation granularity, computational cost, and both pixel- and part-level insertion–deletion metrics, showing that curvature-aware scoring becomes increasingly important as attribution moves beyond pixels. Our code is available here.[1]

1. Introduction

Deep neural networks transformed visual recognition with state-of-the-art performance across diverse benchmarks, yet their decision processes remain largely opaque. Understanding *why* a classifier flags a lesion as malignant or an aircraft component as defective is as important as high accuracy, particularly in settings where transparency, auditing, and reliability are vital. Multiple interpretability techniques have been proposed, including perturbation-based methods such as LIME [22] and KernelSHAP [18], and gradient-based methods like Integrated Gradients [27] and GradCAM [25]. However, most operate at pixel or patch granularity and rely on first-order sensitivity, i.e., local gradient-based approximations, often producing diffuse heatmaps that are difficult to aggregate into structured, decision-level insight [11]. Despite axiomatic formulations of attribution [1, 10], a consensus on what constitutes faithful, actionable explanations is lacking. A fundamental limitation arises when attribution is aggregated beyond pixels. While first-order gradients are appropriate for infinitesimal perturbations, explanations are needed for structured image regions such as object components or semantically coherent parts. A *part* denotes a contiguous region of pixels corresponding to a functional or structural unit of the depicted object (e.g., an eye or a wheel), operating at an intermediate granularity between individual pixels and the full image, where objects are naturally decomposed into meaningful components [7].

When perturbations are applied to regions rather than pixels, input moves away from a local linear regime where gradient-based approximations are valid, and second-order effects induced by model curvature (captured by the Hessian of the network's output with respect to its input) dominate the belief shift. This suggests that part-level attribution is not a coarser visualization of pixel saliency, but a regime in which curvature-aware analysis is necessary. Existing part-level approaches often depend on layer-specific features, require architecture-specific designs, or rely solely on first-order signals, while failing to account for the nonlinear effects induced by model curvature under region-level perturbations. We address this gap with a *segmentation-guided, curvature-aware attribution* framework that explains pretrained classifiers in terms of structured semantic regions. Given an image, we obtain part masks using a mechanism suited to the domain: MediaPipe FaceMesh [8, 17] for faces, OpenPifPaf trained on AnimalPose dataset [2, 14] for animals, YOLOv8 [12] combined with DINOv2 embeddings [19] for vehicles, or unsupervised DINOv2 features clustered via k-means for miscellaneous objects. Segmentation defines the explanatory granularity, but the attribution rule itself remains model-agnostic and training-free.

Beyond part masks, we introduce *KL-Hessian attribu-*

[1] https://github.com/MaishaMaliha1/KLH.git

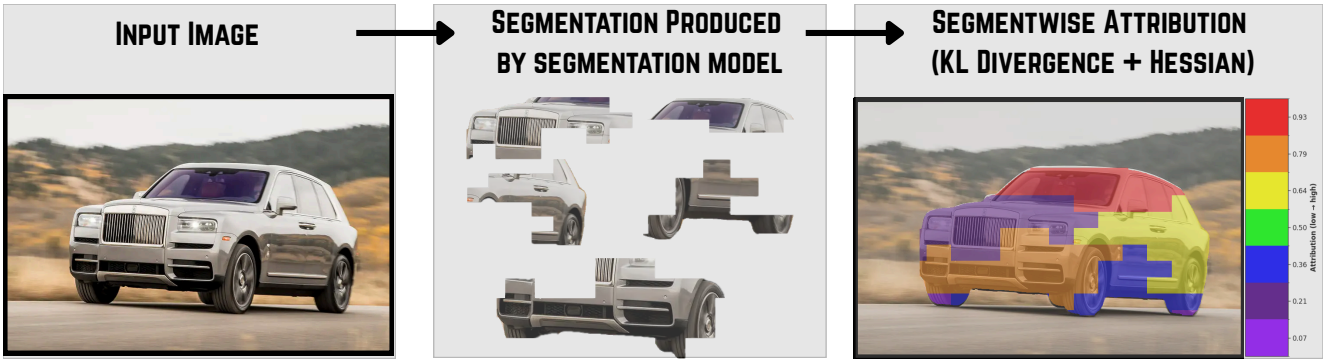

Figure 1. Overview of the proposed segmentation-guided KL-Hessian attribution framework. Input image is decomposed into structured regions using a segmentation module. Each region is then scored via symmetric KL divergence in probability space, modulated by directional Hessian curvature, yielding region-level importance rankings.

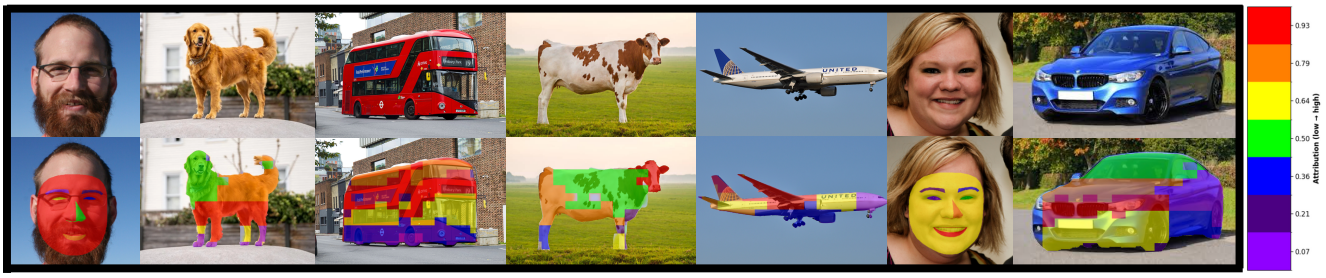

Figure 2. Human-interpretable segment-wise attribution for each input image. Each segment corresponds to a semantically meaningful region within the image, such as eyes, nose, skin, and lips for face images (first and sixth); head, body, and limbs in animal images (second and fourth); and structural components like bonnet, wheels, and doors in the car image (last). The attributions reveal how different parts contribute locally to the model's prediction, facilitating intuitive human understanding of the model's decision-making process.

tion, which unifies two complementary signals. First, we measure belief redistribution directly in probability space using symmetric KL divergence, quantifying how the full class distribution shifts under localized, region-constrained perturbations. Second, we incorporate directional curvature via Hessian–vector products to capture how nonlinear structure amplifies or suppresses these perturbations within each region. The resulting score reflects both output-aware sensitivity and second-order geometric structure, producing stable, rankable importance values over semantically coherent parts without retraining or architectural modification. This formulation bridges structured region-level explanations with curvature-aware attribution, providing a principled extension of first-order saliency methods to settings where region-level perturbations move beyond the local linear regime. Our contributions are:

1. We show that part-level perturbations move beyond local linear regime, motivating curvature-aware attribution.
2. We propose **KL–Hessian attribution**, a hybrid measure combining symmetric KL divergence with Hessian–vector curvature to capture both belief shift and nonlinear amplification.
3. We develop a **segmentation-guided, training-free, model-agnostic framework** applicable across faces, animals, vehicles, and arbitrary objects via supervised or unsupervised part generation.
4. We demonstrate that curvature-aware region scoring produces stable, semantically coherent attributions that complement saliency and masking based analyses.

2. Motivation

Understanding *which regions or parts* of an image drive a classifier's decision is central to interpreting vision models. However, commonly used attribution methods—attention visualizations for Vision Transformers and Class Activation Map (CAM)-style heatmaps for CNNs—are fundamentally limited in how faithfully they capture influence. First-order methods such as DeepLIFT [26], DIG [24], and Integrated Gradients [27] estimate influence via local sensitivity of output with respect to input pixels or patches. While efficient and widely adopted, they capture only *linear* effects around the current point and can *miss influence in flat or saturated regions* with small gradients —even when small, structured perturbations change predictions substantially.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable. There exist

inputs $x \in \mathbb{R}^n$ and coordinates $i \in \{1, \dots, n\}$ such that

$$\frac{\partial f(x)}{\partial x_i} = 0 \quad \text{but} \quad \exists\, \epsilon > 0 \text{ with } f(x + \epsilon e_i) \neq f(x).$$

A second-order Taylor expansion around x_0 reveals why:

$$f(x) = f(x_0) + \nabla f(x_0)^\top (x - x_0) + \tfrac{1}{2}(x - x_0)^\top \nabla^2 f(\xi)\,(x - x_0).$$

for some ξ on the line segment between x_0 and x. When $\nabla f(x_0) = 0$, the change is governed entirely by curvature:

$$f(x) - f(x_0) = \tfrac{1}{2}(x - x_0)^\top \nabla^2 f(\xi)\,(x - x_0).$$

Thus, even with zero gradient, second-order structure can induce significant output change. When perturbations are aggregated over structured regions rather than individual pixels, the resulting displacement is larger, making the second-order term non-negligible. For instance, consider $f(x) = \mathrm{ReLU}(w^\top x + b)$ with $w = [1, 1]^\top$ and $b = -2$. At $x_0 = [0, 0]^\top$, we have $f(x_0) = 0$ and $\nabla f(x_0) = 0$, yet a small move across the activation boundary—for instance, $x = [2.1, 0]^\top$—yields $f(x) = 0.1 \neq f(x_0)$. A purely first-order attribution method would assign 0 importance to this region, completely missing its influence on the prediction.

Attention- and CAM-based methods can be visually appealing but are not necessarily *causal* or *output-aware*; they reflect intermediate activations rather than end-to-end sensitivity. Gradient-only methods, while theoretically grounded, fail in saturated or highly nonlinear regimes where curvature dominates. Approaches such as TokenTM [29], SaCo [30], and Visual-TCAV [4] improve interpretability through token tracing or concept projections, but primarily rely on first-order or feature-space relevance signals and do not explicitly model nonlinear amplification effects under structured region perturbations. Neither family addresses the semantic mismatch between pixel-level heatmaps and region-level perturbations defined over object segments. These limitations motivate a part-centric approach with three key properties: (i) **semantic grounding**, operating over structured segments rather than arbitrary pixels; (ii) **second-order sensitivity**, accounting for curvature in saturated and nonlinear regions; and (iii) **output-aware evaluation**, measuring belief redistribution in probability space using information-theoretic criteria. Our KL-Hessian attribution embraces all three principles. We probe tiny within-part perturbations, shape the direction using local curvature via Hessian–vector products, and score belief shifts with symmetric KL divergence. This yields stable, rankable importance scores over structured regions while accounting for second-order sensitivity in probability space.

3. Related Work

Attribution methods like Grad-CAM [25] and Integrated Gradients [27] are widely used for saliency visualization but have key limitations. Grad-CAM highlights convolutional regions based on gradients at a specific layer, but its outputs are layer-dependent, low-resolution, and sensitive only to first-order changes in a single logit, ignoring belief shifts or nonlinearity. Integrated Gradients integrates gradients from a baseline to the input but is sensitive to baseline choice and fails to capture curvature or class reallocation. Both produce dense pixel-level maps rather than semantically meaningful part-level attributions, making explanations less interpretable and harder to aggregate. Recent concept- and transformer-focused methods aim to improve interpretability but are limited. Visual-TCAV [4] uses hand-curated concept exemplars and feature-space projections, relying on chosen layers and assumptions of separability. HU-MCD [9] leverages SAM segments and masking for CNNs but does not account for curvature and restricts relevance computation to certain architectures. TokenTM [29] tracks token transformations inside Vision Transformers using attention and vector geometry, but it is model-specific and lacks semantic part grounding. SaCo [30], though designed as a faithfulness metric, does not generate explanations and is sensitive to granularity. ProtoPNet [3] offers interpretable-by-design part-based reasoning without requiring part labels, but depends on a specific architecture. In contrast, our KL–Hessian method is model-agnostic, strictly part-local, and operates directly in probability space, capturing both belief shifts and curvature to yield stable, interpretable attributions aligned with human concepts.

4. Proposed Method

Our objective is to turn the prediction of a pretrained image classifier into an explanation that is *part-centric*, *output-aware*, and *locally faithful*. Given an input image $x \in [0, 1]^{H \times W \times 3}$ and a classifier f that returns logits $z(x) \in \mathbb{R}^C$ with class probabilities $p(x) = \mathrm{softmax}(z(x))$, we first decompose the image into a small collection of disjoint, semantically meaningful regions $\mathcal{M} = \{M_k\}_{k=1}^K$, where each $M_k \in \{0, 1\}^{H \times W}$ delineates a part (for instance, a cat's ears, eyes, or muzzle). The explanation assigns a scalar importance to each part by asking a controlled question: *How much does the model's belief over classes change if we make an infinitesimal, well-posed perturbation confined to that part, and how strongly does the local nonlinearity of the network amplify that change?* The resulting score couples an information-theoretic measure of belief shift with a curvature-based amplifier and is computed without retraining, architectural modification, or auxiliary surrogates. All classifiers are used in a frozen, pretrained state; we do not fine-tune models on downstream datasets. Segmentation masks are precomputed and fixed prior to attribution; no segmentation parameters are optimized jointly with our scoring function. Attribution requires only forward and backward passes through the fixed network.

Our method works as follows: (i) parts are obtained by clustering visual tokens and removing small regions via a fixed area threshold; (ii) an intra-part probe direction is constructed by blending masked first- and second-order information; (iii) the second-order term is approximated using a symmetric finite-difference Hessian–vector product along the gradient-restricted direction and then masked; (iv) the shift in class beliefs is measured using a symmetrized KL divergence in probability space; (v) the result is modulated by curvature measured along the masked gradient direction using a nonnegative amplifier. The Hessian provides curvature-aware, second-order amplification of local effects, while the (symmetric) KL measures output-aware belief shifts in probability space (insensitive to additive logit offsets). Our method overview is shown in Figure 1.

4.1. Part Generation

Our method is agnostic to how parts are obtained. When supervised parsers exist for the domain, their labels provide high-quality masks. In their absence, we derive coherent regions by clustering intermediate visual tokens from a self-supervised backbone such as DINOv2 and upsampling the cluster map to image resolution. In both cases, very small fragments are removed by an area threshold to avoid degenerate masks. This modular view deliberately decouples the semantics of parts from the mechanics of attribution and lets practitioners swap the parsing module without changing the scoring rule. Unless otherwise noted, we use $K{=}12$ for CIFAR-10 and ImageNet-1k, $K{=}14$ for FFHQ, and $K{=}10$ for IDRiD [21].

4.2. Attribution as a Localized, Information-Theoretic Probe Modulated by Curvature

Let c denote the class under explanation, typically the model's top-1 prediction. We summarize the classifier's evidence for c as the scalar $s(x) = z_c(x)$ and operate in the normalized input space used by f. For each part M_k we aim to build a perturbation direction u_k that is strictly *intra-part* and that reflects both the steepest ascent of s and the way the network bends in that neighborhood. Once such a direction is in hand, we assess the sensitivity of the *entire* probability vector to small, symmetric nudges along u_k with a symmetric KL divergence. The final attribution multiplies this divergence by a nonnegative curvature factor so that parts in which small moves are *amplified* by the model's nonlinearity receive higher credit.

The construction begins with the gradient of the class score, $g = \nabla_x s(x)$. Because influence must be attributed to a part rather than to arbitrary pixels, the gradient is restricted by masking, then normalized to form an admissible, part-confined direction,

$$v_k = \frac{M_k \odot g}{\|M_k \odot g\|_2 + \varepsilon}. \tag{1}$$

This vector in Eq. (1) represents the first-order direction of maximal increase of s that respects the geometry of the part.

Gradients alone cannot distinguish between regions where the score saturates and those in which it accelerates due to curvature. To capture this, we incorporate second-order information by computing a Hessian–vector product with respect to s:

$$h_k = H(x)\, v_k, \qquad H(x) = \nabla_x^2 s(x), \tag{2}$$

followed by masking and normalization:

$$\hat{h}_k = \frac{M_k \odot h_k}{\|M_k \odot h_k\|_2 + \varepsilon}. \tag{3}$$

We then blend the first- and second-order terms to obtain the intra-part probe direction:

$$u_k = \frac{v_k + \lambda\, \hat{h}_k}{\|v_k + \lambda\, \hat{h}_k\|_2 + \varepsilon}, \qquad \varepsilon > 0, \quad \lambda > 0. \tag{4}$$

Here, λ controls the relative weight of curvature. In practice, $H(x)\, v_k$ is computed using a symmetric finite-difference approximation when exact second derivatives are unavailable:

$$H(x)\, v_k \approx \frac{\nabla_x s(x + \delta v_k) - \nabla_x s(x - \delta v_k)}{2\delta}. \tag{5}$$

Masking in both Eq. (1) and Eq. (2) ensures that u_k remains confined to the part and avoids influence leakage from adjacent regions. To evaluate how model beliefs shift under this localized perturbation, we define perturbed inputs:

$$x_\pm = x \pm \epsilon\, u_k, \tag{6}$$

and extract the corresponding output distributions $p_0 = p(x)$ and $p_\pm = p(x_\pm)$. The symmetric KL divergence is then given by:

$$\mathrm{SKL}(p, q) = \tfrac{1}{2}\left(D_{\mathrm{KL}}(p\|q) + D_{\mathrm{KL}}(q\|p)\right), \tag{7}$$

and the information-theoretic attribution for part k is computed as:

$$\mathrm{KL}_{\mathrm{sym}}(k) = \tfrac{1}{2}\left(\mathrm{SKL}(p_0, p_+) + \mathrm{SKL}(p_0, p_-)\right). \tag{8}$$

Evaluating divergence in probability space (Eq. (8)) ensures invariance to global additive shifts of logits and captures belief redistribution among all classes—this constitutes the first component of our score. Two-sided probing (Eq. (6)) cancels odd-order terms and improves directional stability. The second component captures how strongly the model amplifies movements along the masked gradient direction. This is quantified using the directional curvature:

$$\mathrm{curv}(k) = v_k^\top H(x)\, v_k. \tag{9}$$

Table 1. **Pixel-level attribution quality.** Higher is better for Ins/AOPC; lower is better for Del. Our KL–Hessian is best across all configurations. Results averaged over 3 runs (std $< \pm 0.02$).

Model / Method	CIFAR-10			ImageNet-1k			FFHQ		
	Ins↑	Del↓	AOPC↑	Ins↑	Del↓	AOPC↑	Ins↑	Del↓	AOPC↑
ConvNeXtV2-B									
Ours: KL-Hessian (part)	**0.87**	**0.14**	**0.27**	**0.83**	**0.18**	**0.25**	**0.79**	**0.21**	**0.22**
HU–MCD	0.84	0.17	0.24	0.80	0.21	0.22	0.76	0.24	0.19
Visual–TCAV	0.83	0.18	0.23	0.79	0.22	0.21	0.75	0.25	0.18
TokenTM	-	-	-	-	-	-	-	-	-
SaCo	0.80	0.21	0.20	0.76	0.25	0.18	0.72	0.28	0.15
Integrated Gradients	0.79	0.22	0.19	0.75	0.26	0.17	0.71	0.29	0.14
Grad-CAM	0.74	0.27	0.16	0.70	0.31	0.15	0.67	0.33	0.13
ViT-B/16									
Ours: KL-Hessian (part)	**0.88**	**0.13**	**0.28**	**0.84**	**0.17**	**0.26**	**0.80**	**0.20**	**0.23**
HU–MCD	0.85	0.16	0.25	0.81	0.20	0.23	0.77	0.23	0.20
Visual–TCAV	0.84	0.17	0.24	0.80	0.21	0.22	0.76	0.24	0.19
TokenTM	0.82	0.19	0.22	0.78	0.23	0.20	0.74	0.26	0.17
SaCo	0.81	0.20	0.21	0.77	0.24	0.19	0.73	0.27	0.16
Integrated Gradients	0.80	0.21	0.20	0.76	0.25	0.18	0.72	0.28	0.15
Grad-CAM (token/patch)	0.75	0.26	0.17	0.71	0.30	0.16	0.68	0.32	0.14
SwinV2-B									
Ours: KL-Hessian (part)	**0.89**	**0.12**	**0.29**	**0.85**	**0.16**	**0.27**	**0.81**	**0.19**	**0.24**
HU–MCD	0.86	0.15	0.26	0.82	0.19	0.24	0.78	0.22	0.21
Visual–TCAV	0.85	0.16	0.25	0.81	0.20	0.23	0.77	0.23	0.20
TokenTM	0.83	0.18	0.24	0.79	0.22	0.22	0.75	0.25	0.18
SaCo	0.82	0.19	0.23	0.78	0.23	0.21	0.74	0.26	0.17
Integrated Gradients	0.81	0.20	0.22	0.77	0.24	0.20	0.73	0.27	0.16
Grad-CAM (windowed)	0.76	0.25	0.18	0.72	0.29	0.17	0.69	0.31	0.15

Positive values of Eq. (9) indicate local convexity and super-linear score increases, while negative values indicate saturation. We apply a nonnegative curvature amplifier:

$$A(k) = 1 + \gamma \max\{\mathrm{curv}(k), 0\}. \tag{10}$$

where $\gamma \geq 0$ is a global curvature modulation weight. The final attribution score is then the product of belief shift and curvature amplification:

$$\mathcal{S}(k) = \mathrm{KL}_{\mathrm{sym}}(k) \cdot A(k). \tag{11}$$

Eq. (11) defines the importance of a part as the amount of belief reallocation it causes under a small, within-part perturbation (Eq. (8)), scaled by how strongly the network's geometry amplifies that perturbation (Eq. (10)).

5. Experimental Setup

5.1. Datasets & Metrics

We evaluate on three benchmarks: **CIFAR-10** [15] (10k test images), **ImageNet-1k** [5] (50k validation images, single-crop 224×224), and **FFHQ** [13] (70k face images covering diverse pose, illumination, and background). No dataset is used for training—all models remain frozen and results are computed on evaluation splits only. We employ a **dual evaluation protocol** at both pixel and part granularity.

5.1.1. Pixel-Level Evaluation

We report three standard faithfulness metrics: **Insertion** (Ins ↑), **Deletion** (Del ↓) [20], and **AOPC** (↑) [23]. Pixels are sorted by attribution score and progressively revealed (Ins) or removed (Del) in T=50 steps. Ins starts from a Gaussian-blurred baseline (σ=11) and reveals top-ranked pixels—higher area indicates important pixels are identified early. Del starts from the original image and replaces top-ranked pixels with baseline—lower area indicates effective identification. AOPC measures average step-wise change in $p_c(\cdot)$ across the perturbation trajectory. Since our method produces part-level scores, we assign each pixel the importance of its corresponding part, upsample to pixel resolution, and smooth with a 5×5 box filter. All methods use identical inputs, baselines, and schedules. We swept perturbation steps $T \in \{10, 25, 50, 100, 200\}$, Gaussian blur baselines $\sigma \in \{7, 11, 15\}$, and box filters $\{3\times3, 5\times5, 7\times7\}$ on ImageNet-1k (ViT-B/16). Ins and Del metrics stabilize within ± 0.005 for $T \geq 50$, indicating convergence. Varying σ does not alter attribution rankings or area-under-curve trends; σ=11 is selected as a mid-range default that balances background suppression and structural preservation. Smoothing changes pixel-level scores by less than 0.4%, confirming T=50 as the smallest converged step count and that the 5×5 filter does not affect attribution order.

5.1.2. Part-Level Evaluation

We introduce **Part-Insertion** (P-Ins ↑), **Part-Deletion** (P-Del ↓), and **Part-AOPC** (P-AOPC ↑), which score parts by attribution and progressively reveal/remove *entire parts* in K steps. Scores are computed via trapezoidal integration:

$$\text{P-Ins} = \frac{1}{T}\sum_{t=0}^{T-1} \frac{p_t^{\mathrm{ins}} + p_{t+1}^{\mathrm{ins}}}{2}, \text{P-Del} = \frac{1}{T}\sum_{t=0}^{T-1} \frac{p_t^{\mathrm{del}} + p_{t+1}^{\mathrm{del}}}{2}, \tag{12}$$

$$\text{P-AOPC} = \frac{1}{K}\sum_{k=1}^{K} \left| p_c(x) - p_c(x_k^{\mathrm{del}}) \right|. \tag{13}$$

For pixel-based baselines, we aggregate to part level by averaging attributions within each part:

$$s_k = \frac{1}{|M_k|} \sum_{(i,j)\in M_k} A_{ij}, \tag{14}$$

where M_k is part k and A is the pixel-level attribution map.

Sensitivity to Segmentation & Robustness. We verify robustness along three axes. *(i) Number of parts:* varying $K \in \{5, \ldots, 30\}$ on ImageNet-1k (ViT-B/16) shows smooth performance with $K \approx 12$ optimal; KL-Hessian outperforms pixel-only baselines for every K. The method also operates without semantic masks: an 8×8 uniform grid achieves P-Ins of 0.81 on ImageNet-1k (ViT-B/16), compared to 0.84 with DINOv2 clustering, confirming that the scoring rule itself drives gains while semantic segmentation adds benefit. *(ii) Parser choice:* DINOv2+k-means, spectral clustering, and mask-based parsers yield consistent rankings, with alternative parsers trailing by only a few

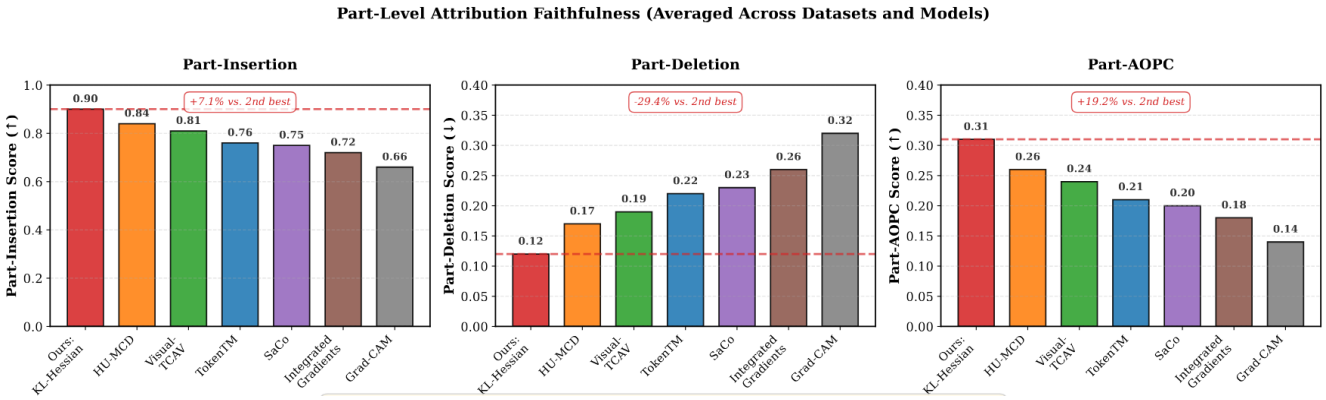

Figure 3. Part-Insertion (left), Part-Deletion (center), and Part-AOPC (right) measure semantic-granularity faithfulness by ranking and perturbing entire parts. Results are averaged over CIFAR-10, ImageNet-1k, and FFHQ using ViT-B/16, ConvNeXtV2-B, and SwinV2-B, with each method evaluated on compatible models. Pixel-level baselines are aggregated to part level by averaging attribution scores within each part. Red dashed lines indicate our method. Higher is better for Insertion and AOPC (↑); lower is better for Deletion (↓).

percentage points. *(iii) Input noise:* under Gaussian perturbations ($\sigma \in \{0.01, \ldots, 0.10\}$), Spearman rank correlation remains above 0.8 and top-3 Jaccard exceeds 0.75 for $\sigma \leq 0.05$, substantially outperforming first-order baselines in rank stability.

5.2. Models & Baselines

We evaluate explanations on three strong, diverse ImageNet classifiers, kept *frozen* throughout: (1) **ViT-B/16** [6], (2) **ConvNeXtV2-B** [28], and (3) **SwinV2-B** [16]. Unless noted, inputs are resized to 224×224, center-cropped, and normalized with ImageNet mean/std; the explained class c is the model's top-1 prediction. All baselines (TokenTM, HU–MCD, SaCo, Visual-TCAV, Integrated Gradients, Grad-CAM) and our KL-Hessian method consume the same preprocessed inputs and are scored under the same Ins/Del/AOPC protocol for fair comparison. We benchmark against six complementary explanation methods spanning robustness, concepts, and saliency: **TokenTM** [29] (ViT-specific; marked "–" for ConvNeXtV2-B in Table 1), **SaCo** [30], **Visual-TCAV** [4], **HU–MCD** [9], **Integrated Gradients** [27], and **Grad-CAM** [25]. For all baselines, we extract their native attribution maps (concept-level scores for HU–MCD, pixel-level saliency for others) and re-evaluate them under our unified Ins/Del/AOPC protocol with identical perturbation schedules, baselines, and preprocessing. This ensures that all scores in Table 1 reflect the same evaluation pipeline rather than each method's original reporting protocol.

Hyperparameter Sensitivity. KL–Hessian has four hyperparameters: probe magnitude ϵ, finite-difference step δ, gradient–curvature blend weight λ, and curvature modulation weight γ in $\mathcal{S}(k) = \mathrm{KL}_{\mathrm{sym}}(k)\,(1 + \gamma\,\mathrm{curv}(k))$. One-dimensional sweeps on ImageNet-1k (ViT-B/16) around defaults (ϵ=0.20, δ=0.06, λ=0.40, γ=1.0) show broad stability: $\lambda \in [0.3, 0.6]$ changes P-Ins/P-AOPC by $<2\%$, $\epsilon \in [0.15, 0.25]$ and $\delta \in [0.04, 0.08]$ yield nearly identical scores, and γ has a wide optimum near 1.0. Extreme values degrade gracefully—γ=0 reduces part-level metrics but remains above first-order baselines, while very small δ introduces noise and large ϵ causes nonlinear distortion.

Computational Cost. Experiments are conducted on a single RTX 3090 GPU (24GB), AMD Ryzen 9 (32-core) CPU, and 128GB RAM using PyTorch 2.2 (CUDA 12) and timm 0.9. Attribution is computed with batch size 32; runtime is measured as wall-clock time for 256 images and normalized per image. At 224×224 resolution on ViT-B/16 (ImageNet-1k), per-image runtimes are: Grad-CAM 12 ms, Integrated Gradients 45 ms, SaCo 87 ms, Visual-TCAV 121 ms, TokenTM 154 ms, HU–MCD 178 ms, and KL–Hessian 162 ms. Segmentation (FaceMesh, OpenPifPaf, YOLOv8, DINOv2+k-means) is precomputed and cached; reported times reflect attribution only. While KL–Hessian is costlier than purely first-order methods due to two-sided KL evaluation and symmetric finite-difference Hessian–vector products, it remains comparable to other second-order or concept-based approaches. Full ImageNet-1k validation (50k images) requires ∼2.3 GPU-hours.

5.3. Results

At pixel level (Table 1), KL-Hessian consistently outperforms all baselines across three datasets (CIFAR-10, ImageNet-1k, FFHQ) and three architectures (ViT-B/16, ConvNeXtV2-B, SwinV2-B), surpassing the second-best method (HU–MCD) by +3.6 Ins, −3.2 Del, and +3.1 AOPC points on average, with the largest gains on FFHQ where curvature-aware modulation effectively handles pose, illumination, and occlusion variability. At part level (Figure 3)—the natural granularity for human interpretation—margins widen substantially: we achieve 0.90 Part-

Category	KL–Hessian (Ours)			HU–MCD		
	P-Ins↑	P-Del↓	P-AOPC↑	P-Ins↑	P-Del↓	P-AOPC↑
Animals	0.91	0.11	0.32	0.85	0.16	0.27
Vehicles	0.89	0.13	0.30	0.83	0.18	0.25
Household	0.88	0.14	0.29	0.82	0.19	0.24
Scenes	0.87	0.15	0.28	0.81	0.20	0.23
Average	**0.89**	**0.13**	**0.30**	0.83	0.18	0.25

Table 2. Head-to-head with HU–MCD on ImageNet category subsets (ViT-B/16, 1k images/category). KL-Hessian leads by +7.2% P-Ins, –27.8% P-Del, +20.0% P-AOPC.

Insertion (+7.1% vs. HU–MCD's 0.84), 0.12 Part-Deletion (-29.4% vs. 0.17), and 0.31 Part-AOPC (+19.2% vs. 0.26), while pixel-based methods degrade markedly (Grad-CAM: 0.66 / 0.32; Integrated Gradients: 0.72 / 0.26 for Part-Ins / Part-Del), confirming that part-aware attribution provides superior semantic coherence in identifying which meaningful regions drive model decisions. Qualitatively (Figures 4 and 2), baseline methods (SaCo, Visual-TCAV, TokenTM, Grad-CAM) produce diffuse, poorly-bounded heatmaps that extend beyond relevant boundaries or uniformly highlight entire regions, whereas our part-level attributions precisely localize semantically meaningful components—eyes, nose, mouth—yielding interpretable explanations that directly support model auditing, debugging, and trust calibration. Table 2 provides a per-category comparison against HU–MCD on ImageNet subsets, confirming consistent gains across diverse object types.

6. Case Study: Retinal Diabetic Retinopathy

We evaluate on IDRiD dataset [21], selecting 20 Diabetic Retinopathy (DR)-positive images from the disease-grading split. Images are center-cropped and resized to 224×224. We segment each image into $K=10$ parts via k-means on DINOv2-base patch tokens and score parts using symmetric KL divergence with Hessian curvature amplification ($\epsilon=0.35$, $\delta=0.12$, $\lambda=0.35$; adjusted from defaults to accommodate the higher contrast and sparse lesion structure of fundus images). Our classifier is ViT-base (`timm/vit_base_patch16_224.dino`); the Grad-CAM baseline is computed on the same ViT-B/16 backbone using the final transformer block activations. Grad-CAM produces broad, diffuse attributions extending into diagnostically uninformative regions, whereas our method yields *localized, part-level* attributions concentrated on discrete structures whose perturbation shifts model belief (Figure 5). Clinically, DR manifests through vascular compromise with hallmarks including microaneurysms, dot-blot/flame hemorrhages, and exudates. *Hard exudates* appear as yellow-white deposits with sharp edges, clustering near leaky vessels; *soft (cotton-wool) exudates* show paler appearance with fuzzy margins, indicating nerve fiber ischemia. In our visualizations, high-scoring parts (RED) align with these focal, diagnostically salient patterns—enabling segment-aware clinical interpretation and supporting reasoning about specific retinal pathologies. Quantitatively, KL-Hessian achieves a lesion-hit-rate of 0.99 for the three highest-scoring parts, with high lesion-overlap, substantially outperforming pixel-based baselines.

7. Ablation Study

Our method integrates two complementary components: (i) a *probability-space* term using symmetric KL divergence ($\mathrm{KL}_{\mathrm{sym}}$) that quantifies belief shift under segment-wise perturbations, and (ii) a *curvature* term from the segment-wise directional Hessian that amplifies attribution in locally nonlinear regions. For part k and input x, the KL shift is $\mathrm{KL}_{\mathrm{sym}}(k)$ as defined in Eq. (8). Local curvature $\mathrm{curv}(k)$ is estimated via symmetric finite-difference Hessian–vector products (Eq. (9)). These combine into:

$$\mathcal{S}(k) = \mathrm{KL}_{\mathrm{sym}}(k) \cdot (1 + \gamma \max\{\mathrm{curv}(k), 0\}), \quad (15)$$

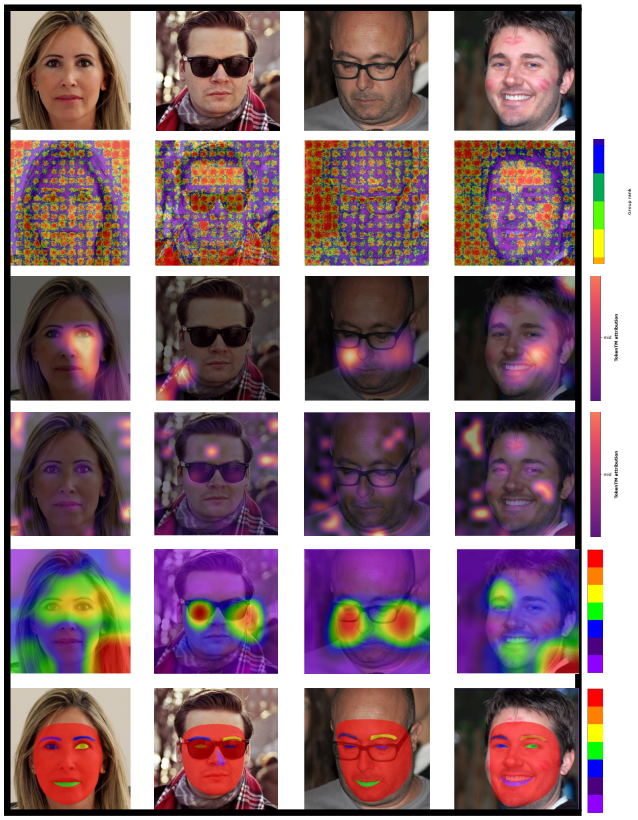

Figure 4. **Qualitative attribution comparison across methods.** From top to bottom: original images, SaCo, Visual-TCAV, TokenTM, Grad-CAM, and our KL-Hessian (bottom row). Baselines generate diffuse heatmaps with poor boundary localization; our approach produces crisp, segment-specific attributions aligned with human-interpretable parts.

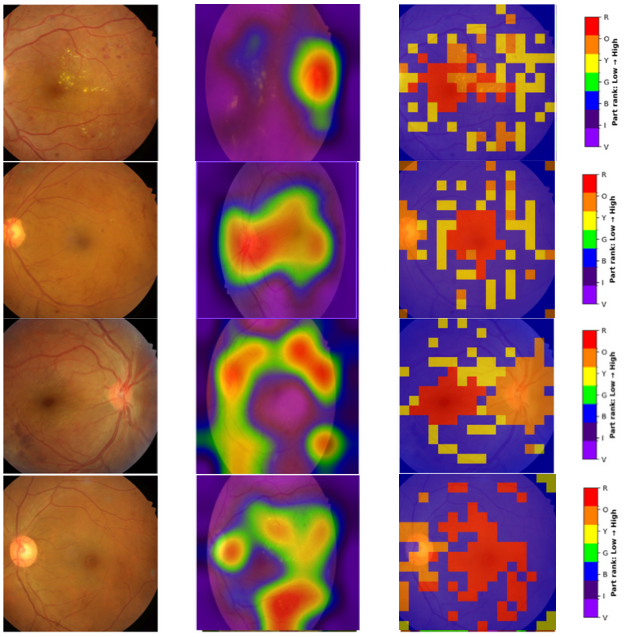

Figure 5. **DR attribution comparison.** Left: original fundus image. Center: Grad-CAM (ViT-B/16; VIBGYOR colormap). Right: our part-level attribution (RED = highest importance). Our method concentrates on compact, clinically meaningful structures rather than diffuse regions.

where $\gamma \geq 0$ is the curvature modulation weight. We evaluate three variants: (1) **KL-only** ($S_{\mathrm{KL}} = \mathrm{KL}_{\mathrm{sym}}$), (2) **Hessian-only** ($S_{\mathrm{Hess}} = \max\{\mathrm{curv}(k), 0\}$, normalized per image), and (3) **Full KL-Hessian**.

Table 3 shows that KL-only assigns high scores to large segments causing broad distributional changes, but suppresses small semantically important regions with modest isolated effects. Hessian-only focuses on local second-order curvature, highlighting sharp features (eyes, lips, textures) often missed by KL, but lacks global semantic calibration, reducing insertion task effectiveness. Our full method combines both strengths: KL ensures segments score according to meaningful class-probability shifts, curvature boosts compact high-saliency regions otherwise underweighted, yielding balanced attribution maps capturing both *semantic relevance* and *local sensitivity*. Figure 6 shows these differences: KL-only produces diffuse, coarse heatmaps; Hessian-only yields sparse, highly focused maps; the full combination produces localized yet semantically rich attributions.

Key findings: (1) **KL-only** prioritizes globally dominant segments, overemphasizing large regions while underweighting fine structure. (2) **Hessian-only** excels in local sharpness and deletion but lacks semantic grounding. (3) **Full KL-Hessian** provides consistent gains across architectures and datasets via synergistic belief shift and curvature modulation, enabling accurate, interpretable part-level attribution.

Table 3. **Ablation: KL vs. Hessian vs. Full.** Faithfulness scores on CIFAR-10, ImageNet-1k, FFHQ. **Full** rows match Table 1. All variants use identical protocols. 3 independent runs, std $< \pm 0.03$

Model / Variant	CIFAR-10			ImageNet-1k			FFHQ		
	Ins↑	Del↓	AOPC↑	Ins↑	Del↓	AOPC↑	Ins↑	Del↓	AOPC↑
ConvNeXtV2-B									
KL-only	0.84	0.17	0.24	0.80	0.21	0.22	0.75	0.24	0.19
Hessian-only	0.82	0.19	0.21	0.78	0.23	0.19	0.74	0.26	0.17
Full (KL–Hessian)	**0.87**	**0.14**	**0.27**	**0.83**	**0.18**	**0.25**	**0.79**	**0.21**	**0.22**
ViT-B/16									
KL-only	0.85	0.16	0.25	0.81	0.20	0.23	0.76	0.23	0.20
Hessian-only	0.83	0.18	0.22	0.79	0.22	0.20	0.75	0.25	0.18
Full (KL–Hessian)	**0.88**	**0.13**	**0.28**	**0.84**	**0.17**	**0.26**	**0.80**	**0.20**	**0.23**
SwinV2-B									
KL-only	0.86	0.15	0.26	0.82	0.19	0.24	0.77	0.22	0.21
Hessian-only	0.84	0.17	0.23	0.80	0.21	0.21	0.76	0.24	0.18
Full (KL–Hessian)	**0.89**	**0.12**	**0.29**	**0.85**	**0.16**	**0.27**	**0.81**	**0.19**	**0.24**

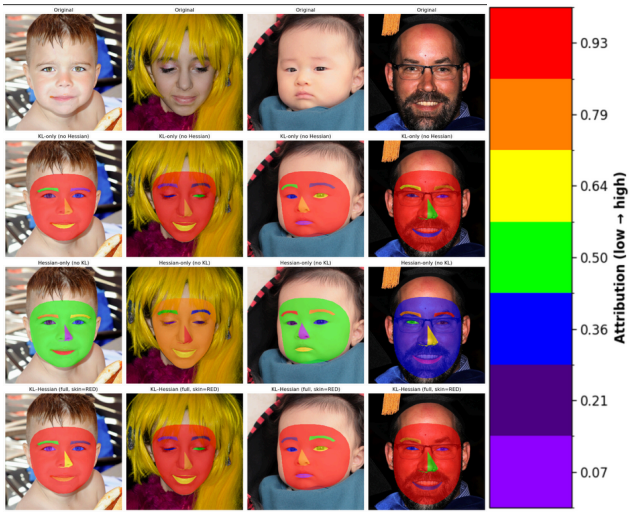

Figure 6. **Component-wise attribution analysis.** Top to bottom: originals, KL-only, Hessian-only, Full KL–Hessian. KL-only produces broad maps; Hessian-only yields sharp but sparse responses; Full balances both for semantically focused, localized attributions.

8. Conclusion and Limitations

We presented KL-Hessian attribution, a part-centric framework that combines probability-space belief shifts with curvature-aware modulation. Experiments demonstrate consistent improvements at semantic granularity across multiple architectures and datasets, including a clinically relevant retinal imaging case study, where our method provides faithful, human-interpretable explanations. Although robust across multiple parsers and even uniform grids, explanation quality is ultimately tied to the segmentation stage. The scoring function uses stable defaults across natural-image benchmarks; adaptive calibration for specialized domains and reducing the computational cost of Hessian–vector products remain promising directions.

References

[1] Marco Bressan, Nicolò Cesa-Bianchi, Emmanuel Esposito, Yishay Mansour, Shay Moran, and Maximilian Thiessen. A theory of interpretable approximations. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 648–668. PMLR, 2024. 1

[2] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9498–9507, 2019. 1

[3] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[4] Antonio De Santis, Riccardo Campi, Matteo Bianchi, and Marco Brambilla. Visual-TCAV: Concept-based attribution and saliency maps for post-hoc explainability in image classification. *arXiv preprint arXiv:2411.05698*, 2024. 3, 6

[5] 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*, pages 248–255. IEEE, 2009. 5

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[7] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2009. 1

[8] Ivan Grishchenko, Artsiom Ablavatski, Yury Kartynnik, Karthik Raveendran, and Matthias Grundmann. Attention mesh: High-fidelity face mesh prediction in real-time. *arXiv preprint arXiv:2006.10962*, 2020. 1

[9] Arne Grobrügge, Niklas Kühl, Gerhard Satzger, and Philipp Spitzer. Towards human-understandable multi-dimensional concept discovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20018–20027, 2025. 3, 6

[10] Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations. *Advances in Neural Information Processing Systems*, 35:5256–5268, 2022. 1

[11] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[12] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, 2023. 1

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 5

[14] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. OpenPifPaf: Composite fields for semantic keypoint detection and spatio-temporal association. *arXiv preprint arXiv:2103.02440*, 2021. 1

[15] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. 5

[16] Ze Liu, Jia Ning, Yue Cao, Tong Wei, Zheng Zhang, Stephen Lin, and Han Hu. Swin Transformer V2: Scaling up capacity and resolution. *arXiv preprint arXiv:2111.09883*, 2022. 6

[17] Camillo Lugaresi, Jiuqiang Tang, Carina Nash, et al. MediaPipe: A framework for building perception pipelines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 1

[18] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017. 1

[19] Maxime Oquab, Piotr Bojanowski, Seyed Kamyar Seyed Izadi, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[20] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. 5

[21] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018. 4, 7

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016. 1

[23] 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, 2016. 5

[24] Soumya Sanyal and Xiang Ren. Discretized integrated gradients for explaining language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10285–10299, 2021. 2

[25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 1, 3, 6

[26] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017. 2

[27] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. 1, 2, 3, 6

[28] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. 6

[29] Junyi Wu, Bin Duan, Weitai Kang, Hao Tang, and Yan Yan. Token transformation matters: Towards faithful post-hoc explanation for vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10926–10935, 2024. 3, 6

[30] Junyi Wu, Weitai Kang, Hao Tang, Yuan Hong, and Yan Yan. On the faithfulness of vision transformer explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10936–10945, 2024. 3, 6