# LibraGrad: Balancing Gradient Flow for Universally Better Vision Transformer Attributions

Faridoun Mehri[1] Mahdieh Soleymani Baghshah[1] Mohammad Taher Pilehvar[2]

[1]Sharif University of Technology, Iran [2]Cardiff University, UK

f.meh16@student.sharif.edu soleymani@sharif.edu pilehvarmt@cardiff.ac.uk

## Abstract

*Why do gradient-based explanations struggle with Transformers, and how can we improve them? We identify gradient flow imbalances in Transformers that violate FullGrad-completeness, a critical property for attribution faithfulness that CNNs naturally possess. To address this issue, we introduce LibraGrad—a theoretically grounded post-hoc approach that corrects gradient imbalances through pruning and scaling of backward paths, without changing the forward pass or adding computational overhead. We evaluate LibraGrad using three metric families: Faithfulness, which quantifies prediction changes under perturbations of the most and least relevant features; Completeness Error, which measures attribution conservation relative to model outputs; and Segmentation AP, which assesses alignment with human perception. Extensive experiments across 8 architectures, 4 model sizes, and 5 datasets show that LibraGrad universally enhances gradient-based methods, outperforming existing white-box methods—including Transformer-specific approaches—across all metrics. We demonstrate superior qualitative results through two complementary evaluations: precise text-prompted region highlighting on CLIP models and accurate class discrimination between co-occurring animals on ImageNet-finetuned models—two settings on which existing methods often struggle. LibraGrad is effective even on the attention-free MLP-Mixer architecture, indicating potential for extension to other modern architectures. Our code is freely available at* `https://nightmachinery.github.io/LibraGrad/`.

## 1. Introduction

Understanding how deep learning models make decisions is crucial for deploying them in critical applications such as healthcare and autonomous driving. Input attribution methods, which quantify the influence of individual input features on a model's output [12, 48, 49, 67], help us understand a model's decision for a single input and also serve as building blocks for advanced explanation techniques like CRAFT [31].

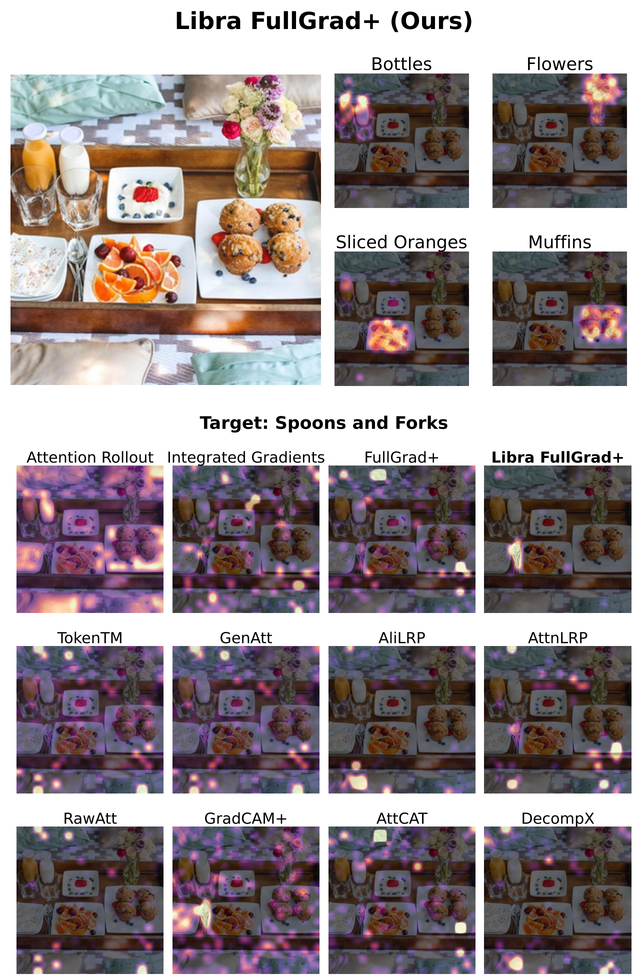



Figure 1. Qualitative comparison on EVA2-CLIP-Large. Our proposed Libra FullGrad+ generates prompt-specific attribution maps (top) and demonstrates improved localization compared to existing methods when explaining the model output for the "spoons and forks" prompt (bottom). For more qualitative examples, see Fig. 2 and Appendix C.

In the field of CNN interpretability, gradient-based attribution techniques—particularly Integrated Gradients [78] and FullGrad [76]—established a foundation for model explanation. However, the architectural paradigm shift brought about by Vision Transformers (ViTs) [25, 83] has exposed limitations in these gradient-based methods, with attention-based attribution methods sometimes achieving more success. Hybrid methods, including GenAtt [16], TokenTM [88], and AttCAT [62], attempt to bridge this gap by integrating gradient and attention-based approaches. Nonetheless, significant challenges persist: these methods lack theoretical foundations, struggle to distinguish between classes effectively, produce noisy attribution maps, and often work only with specific model architectures (*cf.* Appendix E.4).

In this work, we identify the root cause of the failure of gradient-based methods: unbalanced gradient flow during backpropagation leads to unfaithful attribution scores. We demonstrate that while classical CNNs naturally preserve proper gradient flow through their locally affine operations, several components in modern Transformers disrupt this property.

Our solution, LibraGrad, takes a different approach: instead of working around distorted gradients, it prevents the distortion from occurring in the first place by theoretically motivated pruning and scaling of backward paths, leaving the forward pass untouched. Our comprehensive experiments across 8 architectures, 4 model sizes, and 5 datasets show that this not only improves all gradient-based attribution methods but also reveal that specialized attention-gradient hybrids are unnecessary—once gradients flow properly, the general-purpose Libra FullGrad+ achieves superior or comparable performance. We also extend Integrated Gradients (IG) [78] and compose it with other gradient-based methods, and compare the universal improvement aspect of LibraGrad and IG, showing LibraGrad vastly outperforms IG. Furthermore, we theoretically prove that this is to be expected.

## 2. Background and Related Work

Given a multi-output neural model, let $f : \mathbb{R}^n \to \mathbb{R}$ be a selected output function. For instance, if $\text{Model}(x) = (p_1, ..., p_k)$ represents class probabilities, we might choose $f(x) = p_i$ to analyze the model's prediction for the $i$-th class. An attribution method $A$ generates relevance scores $A(f)(x)_i$ for each feature $x_i$.

### 2.1. Gradient-Based Attribution Methods

**Input × Grad.** IxG [4, 72, 73] assigns feature relevance by $\text{IxG}\,(f)(x) = x \odot \nabla_x f(x)$, where $\odot$ denotes element-wise multiplication.

**FullGrad.** Expanding on Input × Grad, FullGrad [76] includes not only the input features but also the bias terms of each layer in the neural network. The FullGrad attribution map is calculated as:

$$\text{FullGrad}(f)(x_0) = \text{IxG}\,(f)(x_0) + \sum_{l=0}^{L-1} \sum_{b \in B_l} \text{IxG}\,(f_b)(b)$$

where $\text{IxG}\,(f)(x_0)$ denotes the Input × Grad for the input $x_0$ (the input to the first layer), and $\text{IxG}\,(f_b)(b)$ is the Input × Grad attribution map of the sub-network $f_b$ with a bias term $b$ from layer $l$ as the input. Also, $f_b$ is the sub-network of $f$ starting from the bias term $b$ and going until the end of the model, whereas $B_l$ denotes the set of all bias terms in layer $l$. FullGrad+ ∘ PLUS (henceforth FullGrad+) [50] is defined as follows:

$$\text{FullGrad+}\,(f)(x_0) = \sum_{l=0}^{L-1} \text{IxG}\,(f_l)(x_l) + \sum_{l=0}^{L-1} \sum_{b \in B_l} \text{IxG}\,(f_b)(b)$$

where $\text{IxG}\,(f_l)(x_l)$ is the Input × Grad attribution map of the sub-network $f_l$ with input $x_l$ (the input to the $l$th layer). FullGrad+ aggregates the input attribution maps of each layer along with the attribution maps of all bias terms in each layer.

**Integrated Gradients.** IG [78] computes attributions w.r.t. a baseline input $\bar{x}$ (*e.g.*, zero):

$$\text{IG}\,(f)(x) = (x - \bar{x}) \odot \int_{\alpha=0}^{1} \nabla_x f(\bar{x} + \alpha(x - \bar{x})) d\alpha$$

In practice, we approximate the integral using a 50-step Riemann summation.

### 2.2. Other Attribution Methods

In addition to the primary gradient-based methods above, we apply LibraGrad to several other general-purpose gradient methods, including HiResCAM [26], GradCAM ∘ PLUS (henceforth GradCAM+) [42, 50, 68], and XGradCAM+ ∘ PLUS (henceforth XGradCAM+) [33, 50]. We further apply it to hybrid attention-gradient approaches specifically designed for Transformer architectures: GenAtt (also known as GAE) [16], TokenTM [88], and AttCAT [62]. To ensure a comprehensive evaluation, we also compare against attention-based attribution methods RawAtt [15, 17, 35], Attention Rollout [1], and DecompX-NoBias (henceforth DecompX) [53], as well as Transformer-specific Layer-Wise Relevance Propagation (LRP)-based [6] techniques Conservative-LRP (henceforth AliLRP) [3] and AttnLRP [2]. For a detailed overview of related work, see Appendix E.

## 3. Method

Understanding how input features contribute to a model's output is the central goal of attribution methods. For attributions to be faithful, they must accurately reflect the influence of each input feature on the output. This requires decomposing model outputs into input and bias contributions, formalized as:

**Definition 1.** A function $f$ is **FullGrad-complete** (or **FG-complete**) if, for all $x \in \mathbb{R}^n$,

$$f(x) = J_x f \cdot x + \sum_i J_{b_i} f \cdot b_i,$$

where $J_x f = \frac{\partial f}{\partial x} \in \mathbb{R}^{m \times n}$ is the Jacobian matrix of $f$ with respect to $x$, and $J_{b_i} f = \frac{\partial f}{\partial b_i} \in \mathbb{R}^{m \times d_i}$ are the Jacobian matrices of $f$ with respect to the bias terms $b_i$. (*Cf*. Proposition 6 in [76].)

FG-completeness ensures that the sum of the attributions equals the model's output, leaving no unexplained residual. This is a necessary condition for faithful interpretability, as it guarantees that all factors influencing the output are accounted for in the attribution scores, and no extraneous influence is attributed to the inputs. Throughout this paper, we use the term "balanced gradient flow" interchangeably with FG-completeness. In the following sections, we:

- Establish that classical neural architectures are FG-complete, thereby explaining the historical success of gradient-based attribution on these models (§3.1).
- Identify non-locally-affine layers in Transformers that break FG-completeness (§3.2).
- Analyze how this causes gradient flow imbalance (§3.3).
- Develop theoretical solutions to restore balanced gradients, introducing *LibraGrad* (§3.4).
- Present practical implementations of LibraGrad for common Transformer components (§3.5).
- Explain the intuition behind a balanced gradient flow using a simple and concrete example (Appendix A.2).

Proofs of theorems and propositions are provided in Appendix A.3.

### 3.1. FG-Completeness of Classical Architectures

We begin by demonstrating that classical convolutional neural networks (CNNs) and multilayer perceptrons (MLPs) satisfy FG-completeness, which explains why gradient-based attribution methods are effective for these architectures. First, we introduce the concept of a locally affine function.

**Definition 2.** A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is **locally affine** at a point $x_0 \in \mathbb{R}^n$ if there exists an open neighborhood $U \subset \mathbb{R}^n$ containing $x_0$, a matrix $W(x_0) \in \mathbb{R}^{m \times n}$, and a vector $b(x_0) \in \mathbb{R}^m$ such that

$$f(x) = W(x_0)x + b(x_0), \quad \forall x \in U.$$

Many activation functions used in neural networks, such as ReLU, are piecewise linear and therefore locally affine almost everywhere. Our next theorem shows that locally affine functions satisfy FG-completeness.

**Theorem 1.** *Any locally affine function at $x_0$ is FG-complete in a neighborhood of $x_0$.*

Moreover, we can compose such functions and retain FG-completeness:

**Theorem 2.** *The composition of a finite number of FG-complete functions is FG-complete.*

Next, we show that FG-completeness is preserved under addition. This property is relevant for neural networks with residual connections, where the output of a layer is added to its input.

**Theorem 3.** *Let $f_1, f_2$ be FG-complete functions. Then their sum $f = f_1 + f_2$ is FG-complete.*

We can now assert that classical neural network architectures are FG-complete:

**Corollary 1.** *Classical neural networks employ several types of affine transformations $f(x) = Wx + b$:*
1. *Linear: $W \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$*
2. *Convolutional: $W$ with spatial weight-sharing, $b$ broadcast per channel*
3. *Pooling: AveragePool, Global-Average-Pool (special cases of Conv)*
4. *BatchNorm (eval): $W = \text{diag}(\gamma/\sigma)$, $b = \beta - \mu\gamma/\sigma$*
5. *LayerScale: $W = \text{diag}(\alpha)$, $b = \beta$*

*Combined with piecewise-linear activations (Theorem 1) and skip connections (Theorem 3), these networks are FG-complete on $\mathbb{R}^n \setminus S$ (Theorem 2), where $S$ denotes the union of boundaries between linear regions*

### 3.2. Non-Locally-Affine Layers in Transformers

Despite the FG-completeness of classical architectures, modern Transformer models introduce several non-locally-affine operations that disrupt this property:

1. **Gated Activations:** Functions like GELU and SiLU (Swish) [70] involve non-linear gating mechanisms.
2. **Attention Mechanisms:** Self-attention and cross-attention layers perform weighted averaging based on nonlinear attention scores.
3. **Multiplicative Feature Fusions:** Operations such as self-gating (*e.g.*, SwiGLU [70], MambaOut [92]) involve element-wise multiplication of different feedforward branches.
4. **Normalizations:** LayerNorm divides by the standard deviation, introducing a division operation.

These operations involve multiplicative (of which division is a special case) interactions and non-linear transformations that break the linearity required for FG-completeness, leading to imbalanced gradient flow and attribution failures, as we will discuss in the next section.

### 3.3. Analysis of Gradient Flow Imbalance

We now analyze how each non-locally-affine operation affects gradient flow. First, consider the element-wise multiplication of two FG-complete functions:

**Proposition 1.** *Let $f_1, f_2$ be FG-complete functions and let $f(x) = f_1(x) \odot f_2(x)$ be their element-wise product with Jacobians:*

$$J_x f = \mathit{diag}(f_2(x)) \cdot J_x f_1 + \mathit{diag}(f_1(x)) \cdot J_x f_2$$

$$J_{b_i} f = \mathit{diag}(f_2(x)) \cdot J_{b_i} f_1 + \mathit{diag}(f_1(x)) \cdot J_{b_i} f_2$$

*Then $f$ is not FG-complete. Specifically:*

$$J_x f \cdot x + \sum_i J_{b_i} f \cdot b_i = 2f(x)$$

So far, we've assumed both paths are FG-complete before multiplication. What happens when they're not? While each such case needs its own mathematical proof, multiplication tends to exacerbate any existing gradient flow imbalances rather than restore FG-completeness. Two key examples illustrate this: division (a non-linear multiplicative operation), which we analyze next, and SiLU, which Proposition 4 (in the Appendix) proves to lack FG-completeness.

**Proposition 2.** *Let $f_1, f_2$ be FG-complete functions with $f_2$ non-zero. FullGrad vanishes to exactly zero on their element-wise quotient $f(x) = f_1(x) \oslash f_2(x)$.*

Proposition 2 demanded FG-completeness of both terms—a condition LayerNorm's denominator fails to satisfy. Nevertheless, as we show next, this does *not* spare LayerNorm from vanishing FullGrad attributions.

**Proposition 3.** *For the LayerNorm operation without affine parameters:*

$$\mathit{LN}(x)_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \varepsilon}},$$

*where $\mu = \frac{1}{N}\sum_{k=1}^{N} x_k$ and $\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2$, FullGrad approaches zero as $\varepsilon$ approaches zero:*

$$\lim_{\varepsilon \to 0} J_x \mathit{LN} \cdot x = 0.$$

### 3.4. LibraGrad: Theoretical Foundations

We now develop theoretical solutions to restore balanced gradient flow.

**Theorem 4.** *Let $f_1, f_2$ be FG-complete functions. Then their element-wise product $f(x) = f_1(x) \odot f_2(x)$ is FG-complete when its Jacobians are defined with scaling coefficients $a, b \in \mathbb{R}$ where $a + b = 1$:*

$$J_x f = a[\mathit{diag}(f_2(x)) \cdot J_x f_1] + b[\mathit{diag}(f_1(x)) \cdot J_x f_2]$$

$$J_{b_i} f = a[\mathit{diag}(f_2(x)) \cdot J_{b_i} f_1] + b[\mathit{diag}(f_1(x)) \cdot J_{b_i} f_2]$$

The constraint $a + b = 1$ naturally suggests dividing the gradients of each branch by two, *i.e.*, $a = b = 0.5$. We use the previous theorem to correct the gradients of the Self-Gating module (see Libra Self-Gating in §3.5) where two FG-complete branches are multiplied together. However, other nonlinear modules in Transformers have a nonlinearity multiplied by an FG-complete branch. While the previous theorem cannot handle such modules, a specific choice of $a = 1$, $b = 0$ (assuming $b$ is the scaling factor of the nonlinear branch) works effectively, as demonstrated in the next theorem.

**Theorem 5.** *Let $f_1, f_2$ be arbitrary functions (not necessarily FG-complete), and let $f(x) = f_1(x) \odot f_2(x)$ be their element-wise product. Consider $f$ with scaled Jacobians as defined in Theorem 4. Then:*

1. *When $a = 0$, yielding $f(x) = [f_1(x)]_{cst.} \odot f_2(x)$ where $[\cdot]_{cst.}$ is the constant operator that zeroes gradients, $f$ is FG-complete if $f_2$ is FG-complete.*
2. *By symmetry, when $b = 0$, $f$ is FG-complete if $f_1$ is FG-complete.*

While the above theorem can be viewed as a special case of Theorem 4, it deserves separate consideration. By treating the nonlinear multiplicand as constant, we construct a locally linear approximation of the original nonlinear function that is exact for each particular input. In other words, we create different locally linear approximations of the model for each given input, and these approximations are exact for those specific inputs. (Recall that locally linear functions are FG-complete, per Theorem 1.) The theorem above also extends to matrix multiplication, again reducing to Theorem 1. §3.5 applies this theorem to make Attention, LayerNorm, and Gated Activations FG-complete.

Our approach differs from conventional Taylor linearization. A first-order Taylor expansion approximates a function $f(x)$ around a point $x_0$ as $f(x_0) + f'(x_0)(x - x_0)$, where the constant term $f(x_0) - f'(x_0)x_0$ serves as an implicit bias term. Gradient-based attribution methods typically ignore this bias term entirely, and distributing this bias term's attribution across input features presents a non-trivial challenge. Our method circumvents this issue by constructing locally-exact linear approximations without introducing such bias terms.

| **Method** | **Computation** | **Memory** |
|---|---|---|
| Input $\times$ Grad | $\mathcal{O}(1)$ | $\mathcal{O}(\sqrt{\text{Layers}})$ |
| Integrated Gradients | $\mathcal{O}(\text{Steps})$ | $\mathcal{O}(\sqrt{\text{Layers}})$ |
| DecompX | $\mathcal{O}(\text{Tokens})$ | $\mathcal{O}(\text{Tokens})$ |
| FullGrad+ | $\mathcal{O}(1)$ | $\mathcal{O}(\sqrt{\text{Layers}})$ |
| **Libra FullGrad+** | $\mathcal{O}(1)$ | $\mathcal{O}(\sqrt{\text{Layers}})$ |

Table 1. Computational and memory complexities of attribution methods relative to one forward pass [2, 21, 53, 76, 78].

**Summary.** When handling multiplicative interactions, we face a choice: ideally, we can scale gradients if both paths are FG-complete (Theorem 4), preserving information from both paths, or—when one path lacks FG-completeness—we can prune paths to restore FG-completeness by relying on just one FG-complete path (Theorem 5).

**Corollary 2.** *Division can be made FG-complete by treating it as element-wise multiplication with a gradient-pruned non-linear reciprocal: $f(x) = f_1(x) \odot [1/f_2(x)]_{cst.}$ which satisfies FG-completeness, by Theorem 5.*

For division operations like those in LayerNorm, Corollary 2 shows how treating the denominator as constant in the backward pass restores proper gradient flow.

These theoretical results suggest a general principle: balanced gradient flow can be achieved through strategic pruning and scaling of backward paths, without modifying the forward computation. Such pruning and scaling can be achieved using the following two gradient manipulation operators:

**Constant Operator.** The constant operator $[\cdot]_{\text{cst.}} : \mathbb{R}^m \to \mathbb{R}^m$ satisfies:

$$[y]_{\text{cst.}} = y, \quad J_x[y]_{\text{cst.}} = 0$$

**SwapBackward.** The SwapBackward : $(f, g) \mapsto h$ operator, where $f, g, h : \mathbb{R}^n \to \mathbb{R}^m$, is defined by:

$$h(x) = f(x), \quad J_x h = J_x g$$

Further theoretical insights about these operators, their computational complexity (unchanged compared to standard gradients, Table 1), and practical PyTorch implementations are available in Appendix A.1.

### 3.5. LibraGrad: Practical Implementation

**Libra Neural Operations.** We now define FG-complete versions of common non-affine operations. Libra Attention, Gated Activations, and LayerNorm use Theorem 5, while Libra Self-Gating uses Theorem 4.

**Libra Attention.** In attention mechanisms, we discard the gradient of the nonlinear softmax.

$$\text{Libra-Attention}(Q, K, V) = [\text{softmax}(QK^T)]_{\text{cst.}} \cdot V$$

**Libra Gated Activation.** For gated activations like GELU and SiLU, we discard the non-linear gate's gradient.

$$\text{Libra-GatedActivation}(x) = x \odot [\text{NonLinearGate}(x)]_{\text{cst.}}$$

**Libra LayerNorm.** We discard the gradient of the nonlinear denominator in LayerNorm. Note that the expectation ($\mu = \mathbb{E}[x]$) is linear.

$$\text{Libra-LayerNorm}(x) = \frac{x - \mu}{[\sqrt{\sigma^2 + \varepsilon}]_{\text{cst.}}}$$

**Libra Self-Gating.** In self-gating operations like SwiGLU, the input flows through dual parallel feedforward paths ($f_1$, $f_2$) and reunifies via element-wise multiplication. To balance the gradient flow between branches, we scale each branch's gradient by $\frac{1}{2}$ (Theorem 4).

$$\text{Libra-SelfGate}(x) = \text{SwapBackward}(f_1 \odot f_2, \tfrac{1}{2}(f_1 \odot f_2))(x)$$

**Corollary 3.** *A Transformer architecture attains FG-completeness when all non-linear components—specifically its attention mechanisms, activation functions, self-gating operations, and LayerNorms—are replaced with their Libra counterparts.*

**Universal Improvement.** While our theoretical discussion focuses on achieving FG-completeness, empirical results demonstrate that LibraGrad's gradient balancing mechanism universally enhances gradient-based attribution methods.

## 4. Experiments

We evaluate LibraGrad through three complementary metrics: Faithfulness, Completeness Error, and Segmentation. For statistical validity, we report standard deviation upper bounds for empirical results. In tables, we denote the best and second-best results in each column with bold and underline formatting, respectively.

### 4.1. Experimental Setup

Our evaluation spans two dimensions:

- **Architectures:** Eight model families (ViT [25], EVA2 [28, 29, 77], BEiT2 [7, 60], FlexiViT [11], SigLIP[1] [93], CLIP [63], DeiT3 [81, 82], MLP-

[1]SigLIP lacks a CLS token, making certain attention-based methods inapplicable.

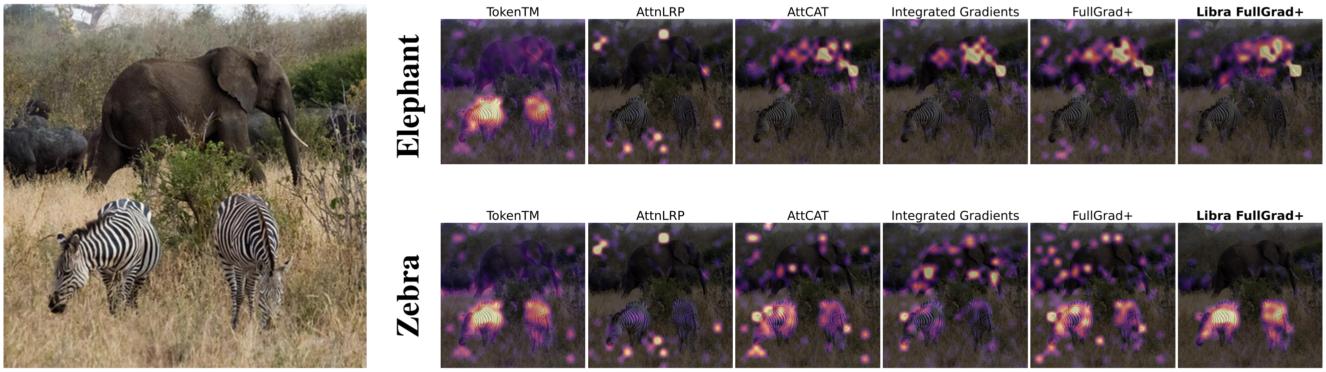



Figure 2. Cross-method comparison of class discriminativity on ViT-B. *Cf*. Fig. 1 and Appendix C.

Mixer [80]), using their largest[2] ImageNet-1k [24] finetuned variants.

- **Model Sizes:** All ViT variants: tiny (ViT-T), small (ViT-S), base (ViT-B), and large (ViT-L).

**Faithfulness Metrics.** We evaluate various attribution methods using faithfulness metrics, which quantify how accurately the attribution scores reflect the importance of input features in the model's predictions. These widely used metrics [13, 20, 32, 50, 53, 55, 88] measure changes in model behavior as we progressively occlude input features in different orders. Here, we report the Most-Influential-First Deletion (MIF) metric with predicted labels and accuracy measurement, which tracks performance degradation when occluding features by decreasing attribution importance. Full details of this and related metrics (Least-Influential-First Deletion, LIF and Symmetric Relevance Gain, SRG) are provided in Appendix B.2, with comprehensive results on all metrics available in Appendix D.

We evaluate all architectures on the ImageNet [24] dataset—the standard benchmark in the attribution literature [17, 50, 88, 90]. On ViT-B, we also experiment with multiple other datasets: ImageNet-Hard [79], and following [22], MURA (a medical X-ray dataset) [64] and Oxford-IIIT Pet [59]. ImageNet-Hard is a challenging dataset combining images from various existing ImageNet variants: ImageNet-V2 [65], ImageNet-Sketch [85], ImageNet-C [36], ImageNet-R [37], ImageNet-ReaL [10], ImageNet-A [38], and ObjectNet [8]. We randomly select 1000 images from each dataset using a fixed seed.

**Completeness Error.** We use Completeness Error to verify theoretical guarantees and validate implementation correctness:

$$\mathrm{CE}(f, x, A) = \left\| f(x) - \sum_{i=1}^{n} A(f)(x)_i \right\| \quad (1)$$

Lower CE values indicate better conservation of the model's output in the attribution scores. As this is just a sanity check, we use only 100 random images from the ImageNet dataset. See Appendix B.1 for further details.

**Segmentation.** For segmentation, following [50], we opt for ImageNet-S [34], which encompasses 919 distinct classes, using a random subset of 5000 images from the validation set. Since segmentation masks provide ground truth annotations of object boundaries, they serve as an objective reference to evaluate how well feature attribution methods identify the truly relevant image regions that contribute to model predictions. See Appendix B.3 for further details.

**FunnyBirds.** We further assess FullGrad+ and its Libra enhancement using FunnyBirds [39], a synthetic dataset explicitly developed, along with tailored metrics, to benchmark attribution methods. See Table 3.

### 4.2. Quantitative Results

Our evaluations demonstrate that LibraGrad universally enhances gradient-based attribution methods across all tested models, architectures, and datasets (see Appendix D for comprehensive results). Significant improvements are observed in both faithfulness and segmentation metrics (Tables 6 and Appendix D.2.1), and Libra FullGrad achieves optimal Completeness Error (Table 4). These enhancements remain consistent across different model scales (Appendix D.3) and datasets (Table 2, Appendix D.4), and extend to the attention-free MLP-Mixer (Appendix D.5.1), validating that gradient flow imbalance, not attention mechanisms, is the core issue.

[2]Huge for CLIP and DeiT3, large for others—except EVA2-S, chosen due to hardware constraints with larger EVA2 variants' input resolutions.

| Method | ImageNet | ImageNet-Hard | MURA | Oxford-IIIT Pet | *Avg.* |
|---|---|---|---|---|---|
| Random | 26.5 | 52.4 | 15.1 | 13.7 | 26.9 |
| RawAtt | 44.6 | 65.9 | 24.8 | 37.2 | 43.1 |
| Attn. Rollout | 35.4 | 62.2 | 21.5 | 21.2 | 35.1 |
| AliLRP | 33.3 | 64.1 | 19.2 | 19.0 | 33.9 |
| AttnLRP | 38.5 | 70.8 | 22.8 | 30.3 | 40.6 |
| DecompX | 37.8 | 67.7 | 21.6 | 22.5 | 37.4 |
| Int. Gradients | 35.4 | 66.6 | 23.8 | 20.7 | 36.6 |
| Input $\times$ Grad | 34.4 | 67.6 | 25.5 | 20.4 | 37.0 |
| **w/ Libra** | 38.6 | 68.8 | 21.6 | 23.5 | 38.1 |
| AttCAT | 46.9 | 82.3 | 31.1 | 37.3 | 49.4 |
| **w/ Libra** | <u>63.5</u> | <u>87.3</u> | <u>40.9</u> | <u>55.3</u> | <u>61.8</u> |
| GenAtt | 58.2 | 81.3 | 30.0 | 44.1 | 53.4 |
| **w/ Libra** | 61.6 | 82.8 | 30.1 | 46.5 | 55.2 |
| TokenTM | 56.8 | 79.3 | 28.0 | 44.0 | 52.0 |
| **w/ Libra** | 59.1 | 80.0 | 28.0 | 45.4 | 53.1 |
| GradCAM+ | 45.6 | 75.8 | 24.0 | 32.6 | 44.5 |
| **w/ Libra** | 61.4 | 83.4 | 34.7 | 47.8 | 56.8 |
| HiResCAM | 45.4 | 74.2 | 22.2 | 18.0 | 39.9 |
| **w/ Libra** | 56.7 | 79.7 | 30.1 | 39.4 | 51.5 |
| XGradCAM+ | 38.6 | 72.1 | 23.7 | 33.2 | 41.9 |
| **w/ Libra** | **63.9** | 84.7 | 36.6 | 52.6 | 59.4 |
| FullGrad+ | 44.2 | 80.1 | 32.8 | 35.3 | 48.1 |
| **w/ Libra** | 63.1 | **87.6** | **43.2** | **57.3** | **62.8** |

Table 2. Cross-dataset analysis of Most-Influential-First Deletion (MIF) Accuracy evaluated using predicted labels on ViT-B. All standard deviations were bounded by 0.1 (omitted for brevity).

| Method | CSDC | PC | DC | D | BI | SD | TS |
|---|---|---|---|---|---|---|---|
| FullGrad+ | 61.0 | 55.0 | 56.8 | 44.5 | **99.7** | 55.4 | 84.3 |
| **w/ Libra** | **92.7** | **91.4** | **90.2** | **91.1** | **99.7** | **69.4** | **97.1** |

Table 3. Evaluation of FullGrad+ and its Libra enhancement on ViT-B using FunnyBirds [39] (metrics defined in their Table 1). FullGrad+ is implemented without biases for this evaluation.

**Integrated Gradients.** We also extend IG [78] and compose it with other gradient-based methods, and compare the universal improvement aspect of LibraGrad and IG in Appendix D.1, showing that LibraGrad vastly outperforms IG. Due to numerical instability, the practical approximation of IG fails to meet its theoretical promise of completeness relative to the zero baseline (Table 4). Furthermore, we prove that the numerical instability observed is theoretically unavoidable for a fixed-step approximation (Proposition 5 in the Appendix).

**General-Purpose Methods Are Enough.** Once gradient flow is corrected, the general-purpose FullGrad+ outperforms Transformer-specific methods like GenAtt, TokenTM, and AttCAT across most metrics and models, with only a few exceptions where its performance remains competitive. This suggests that specialized architectures may not require specialized attribution methods when gradient flow is properly balanced.

**Ablation Studies.** Our ablation study (Table 5) reveals three key insights: First, while gated activations theoretically break FG-completeness (Proposition 4), their practical impact is minimal as they often operate in saturated regimes. Second, LayerNorm's theoretically predicted vanishing attribution problem is empirically confirmed as the most significant factor. Finally, while bias terms are necessary for theoretical completeness, their practical impact is modest, suggesting that implementations can optionally omit them without severe consequences.

### 4.3. Qualitative Analysis

We evaluate Libra FullGrad+ through two complementary scenarios: (1) text-prompted region attribution using CLIP models, demonstrating precise localization of prompted elements in complex scenes (Fig. 1, Appendix C.1), and (2) class discrimination on COCO [47] images, showing accurate distinction between co-occurring animals (Fig. 2, Appendix C.2). Both reinforce our quantitative findings that proper gradient flow enables general-purpose methods to outperform specialized approaches. Detailed protocols are in Appendix B.4.

## 5. Conclusion

We introduced LibraGrad, correcting gradient flow imbalances via pruning and scaling backward paths. FG-completeness, formalized here, ensures attributions decompose outputs faithfully. We prove that while classical CNNs were naturally FG-complete (explaining their historical success with gradient-based methods), several operations in modern Transformers break this property. We provide both theoretical proofs for restoring FG-completeness and practical solutions that require no forward-pass modifications. Empirically, LibraGrad universally enhances gradient-based attributions across architectures, model sizes, and datasets, enabling general-purpose methods like FullGrad+ to outperform Transformer-specific approaches. This suggests that specialized architectures may not require specialized attribution methods when gradient flow is properly balanced. Our qualitative results further validate this insight. Future work can explore compositions with other gradient-based methods, applications as a gradient regularizer, and extensions to emerging architectural innovations.

| Method | ViT-L ↓ | EVA2-S ↓ | BEiT2-L ↓ | FlexiViT-L ↓ | SigLIP-L ↓ | CLIP-H ↓ | DeiT3-H ↓ | *Avg.* ↓ |
|---|---|---|---|---|---|---|---|---|
| Input × Grad | 13.6 ±0.3 | 8.9 ± 0.2 | 9.0 ± 0.1 | 7.1 ±0.1 | 9.3 ± 0.1 | <u>1.3</u> ±0.0 | 8.6 ±0.1 | 8.3 ± 0.2 |
| Integrated Gradients | <u>8.5</u> ±1.5 | 4.8 ± 0.1 | <u>6.7</u> ± 0.1 | <u>4.0</u> ±0.4 | <u>5.1</u> ± 0.2 | 8.2 ±0.1 | 6.4 ±0.5 | <u>6.2</u> ± 0.6 |
| DecompX | 11.3 ±1.3 | 911.2 ±33.7 | 199.2 ±10.4 | 5.5 ±0.5 | 242.1 ±28.7 | 16.7 ±0.8 | 7.7 ±0.6 | 199.1 ±17.2 |
| AliLRP | 29.5 ±4.1 | 1233.1 ±46.7 | 139.4 ± 6.2 | 7.8 ±0.3 | 69.0 ± 8.8 | 15.4 ±1.4 | 18.1 ±0.7 | 216.1 ±18.2 |
| AttnLRP | 11.0 ±0.5 | <u>2.2</u> ± 0.2 | 38.2 ± 2.1 | 4.3 ±0.3 | 30.4 ± 1.7 | 2.9 ±0.2 | <u>5.9</u> ±0.2 | 13.6 ± 1.0 |
| FullGrad | 11.4 ±0.7 | 9.5 ± 0.5 | 11.8 ± 0.5 | 19.8 ±0.6 | 6.7 ± 0.4 | 7.3 ±0.7 | 10.6 ±0.3 | 11.0 ± 0.5 |
| **Libra FullGrad** | **0.0** ±0.0 | **0.0** ± 0.0 | **0.0** ± 0.0 | **0.0** ±0.0 | **0.0** ± 0.0 | **0.0** ±0.0 | **0.0** ±0.0 | **0.0** ± 0.0 |

Table 4. Completeness Error (lower is better) across models for attribution methods. CE for IG has been computed relative to the zero baseline. Methods without a theoretical basis for completeness (*e.g.*, Attention Rollout) are excluded, as their incompleteness is evident.

| Method | MIF Deletion (GT) | | MIF Deletion (Predicted) | | Segmentation |
|---|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC | AP |
| **Libra FullGrad+** | **74.1** ±0.1 | **45.5** ±0.3 | **71.7** ±0.1 | **50.5** ±0.2 | **79.4** ±0.3 |
| No Att. | 68.0 ±0.1 ( -8.2%) | 40.8 ±0.3 (-10.5%) | 65.2 ±0.1 ( -9.1%) | 45.5 ±0.2 (-10.0%) | 72.2 ±0.3 ( -9.1%) |
| No LN | 55.3 ±0.1 (-25.3%) | 30.0 ±0.3 (-34.2%) | 49.9 ±0.1 (-30.4%) | 33.3 ±0.2 (-34.1%) | 72.1 ±0.3 ( -9.2%) |
| No Att. & LN | 63.6 ±0.1 (-14.1%) | 36.6 ±0.2 (-19.7%) | 61.2 ±0.1 (-14.7%) | 41.1 ±0.2 (-18.6%) | 66.2 ±0.3 (-16.7%) |
| No Act. | <u>74.0</u> ±0.1 ( -0.1%) | <u>45.4</u> ±0.3 ( -0.3%) | <u>71.6</u> ±0.1 ( -0.3%) | <u>50.4</u> ±0.2 ( -0.4%) | <u>79.3</u> ±0.3 ( -0.2%) |
| No Gate | 69.8 ±0.1 ( -5.7%) | 41.9 ±0.4 ( -8.0%) | 67.0 ±0.1 ( -6.6%) | 46.7 ±0.3 ( -7.5%) | 71.1 ±0.3 (-10.5%) |
| No Bias | 73.9 ±0.1 ( -0.2%) | 45.3 ±0.3 ( -0.4%) | <u>71.5</u> ±0.1 ( -0.3%) | 50.3 ±0.2 ( -0.4%) | 79.2 ±0.3 ( -0.3%) |
| Normal FullGrad+ | 50.9 ±0.1 (-31.3%) | 25.7 ±0.2 (-43.5%) | 48.0 ±0.1 (-33.0%) | 30.0 ±0.2 (-40.7%) | 51.5 ±0.3 (-35.1%) |

Table 5. Ablation study on the EVA2-S model showing the impact of removing individual components from LibraGrad. Abbreviations used: Att. (Attention), LN (LayerNorm), Act. (Gated Activation Functions), Gate (SwiGLU Self-Gating).

| Method | ViT-L | EVA2-S | BEiT2-L | FlexiViT-L | SigLIP-L | CLIP-H | DeiT3-H | *Avg.* |
|---|---|---|---|---|---|---|---|---|
| Random | 29.5 ±0.1 | 21.2 ±0.1 | 18.3 ±0.1 | 19.2 ±0.1 | 32.8 ±0.1 | 28.0 ±0.1 | 29.0 ±0.1 | 25.4 ±0.1 |
| RawAtt | 39.1 ±0.1 | 50.8 ±0.1 | 29.5 ±0.1 | 41.7 ±0.1 | - | 42.5 ±0.1 | 52.0 ±0.1 | 42.6 ±0.1 |
| Attention Rollout | 31.4 ±0.1 | 41.1 ±0.1 | 19.7 ±0.1 | 23.2 ±0.1 | - | 41.3 ±0.1 | 31.2 ±0.1 | 31.3 ±0.1 |
| AliLRP | 33.2 ±0.1 | 48.0 ±0.1 | 26.2 ±0.1 | 24.9 ±0.1 | 55.4 ±0.1 | 34.4 ±0.1 | 56.3 ±0.1 | 39.8 ±0.1 |
| AttnLRP | 41.8 ±0.1 | 63.5 ±0.1 | 37.7 ±0.1 | 21.8 ±0.1 | 62.2 ±0.1 | 46.7 ±0.1 | 40.7 ±0.1 | 44.9 ±0.1 |
| DecompX | 38.9 ±0.1 | 46.8 ±0.1 | 31.7 ±0.1 | 35.5 ±0.1 | 51.1 ±0.1 | 42.4 ±0.1 | 47.2 ±0.1 | 42.0 ±0.1 |
| Integrated Gradients | 35.9 ±0.1 | 34.8 ±0.1 | 23.2 ±0.1 | 22.3 ±0.1 | 44.0 ±0.1 | 31.0 ±0.1 | 33.2 ±0.1 | 32.1 ±0.1 |
| Input × Grad | 33.9 ±0.1 | 32.3 ±0.1 | 21.8 ±0.1 | 19.9 ±0.1 | 40.8 ±0.1 | 31.4 ±0.1 | 35.1 ±0.1 | 30.7 ±0.1 |
| **Libra Input × Grad** | 40.5 ±0.1 | 64.1 ±0.1 | 33.0 ±0.1 | 36.4 ±0.1 | 51.1 ±0.1 | 43.1 ±0.1 | 47.7 ±0.1 | 45.1 ±0.1 |
| AttCAT | 44.8 ±0.1 | 54.1 ±0.1 | 33.9 ±0.1 | 41.9 ±0.1 | 45.9 ±0.1 | 39.0 ±0.1 | 44.0 ±0.1 | 43.4 ±0.1 |
| **Libra AttCAT** | <u>61.3</u> ±0.1 | <u>69.5</u> ±0.1 | <u>48.9</u> ±0.1 | <u>58.4</u> ±0.1 | **77.4** ±0.1 | <u>58.5</u> ±0.1 | <u>70.5</u> ±0.1 | <u>63.5</u> ±0.1 |
| GenAtt | 51.8 ±0.1 | 40.7 ±0.1 | 30.8 ±0.1 | 53.0 ±0.1 | - | 51.0 ±0.1 | 64.6 ±0.1 | 48.7 ±0.1 |
| **Libra GenAtt** | 55.4 ±0.1 | 42.1 ±0.1 | 32.9 ±0.1 | 54.1 ±0.1 | - | 58.1 ±0.1 | 66.5 ±0.1 | 51.5 ±0.1 |
| TokenTM | 50.0 ±0.1 | 44.7 ±0.1 | 39.6 ±0.1 | 49.3 ±0.1 | - | 51.9 ±0.1 | 63.3 ±0.1 | 49.8 ±0.1 |
| **Libra TokenTM** | 52.5 ±0.1 | 46.0 ±0.1 | 38.3 ±0.1 | 51.0 ±0.1 | - | 57.4 ±0.1 | 65.2 ±0.1 | 51.7 ±0.1 |
| GradCAM+ | 48.6 ±0.1 | 47.1 ±0.1 | 33.4 ±0.1 | 28.7 ±0.1 | 43.5 ±0.1 | 33.0 ±0.1 | 44.5 ±0.1 | 39.8 ±0.1 |
| **Libra GradCAM+** | 56.5 ±0.1 | 67.0 ±0.1 | 37.5 ±0.1 | 33.7 ±0.1 | 47.4 ±0.1 | 36.2 ±0.1 | 48.7 ±0.1 | 46.7 ±0.1 |
| HiResCAM | 25.7 ±0.1 | 59.1 ±0.1 | 35.8 ±0.1 | 23.8 ±0.1 | 31.4 ±0.1 | 37.6 ±0.1 | 25.8 ±0.1 | 34.2 ±0.1 |
| **Libra HiResCAM** | 49.0 ±0.1 | 62.6 ±0.1 | 37.2 ±0.1 | 56.5 ±0.1 | 46.1 ±0.1 | 48.9 ±0.1 | 53.8 ±0.1 | 50.6 ±0.1 |
| XGradCAM+ | 45.9 ±0.1 | 50.2 ±0.1 | 30.6 ±0.1 | 26.6 ±0.1 | 51.4 ±0.1 | 39.4 ±0.1 | 45.1 ±0.1 | 41.3 ±0.1 |
| **Libra XGradCAM+** | 58.8 ±0.1 | 69.3 ±0.1 | 45.6 ±0.1 | 44.3 ±0.1 | 63.6 ±0.1 | 57.7 ±0.1 | 66.1 ±0.1 | 57.9 ±0.1 |
| FullGrad+ | 45.1 ±0.1 | 48.0 ±0.1 | 29.0 ±0.1 | 38.9 ±0.1 | 43.6 ±0.1 | 37.6 ±0.1 | 41.9 ±0.1 | 40.6 ±0.1 |
| **Libra FullGrad+** | **62.4** ±0.1 | **71.7** ±0.1 | **50.0** ±0.1 | **59.1** ±0.1 | <u>73.5</u> ±0.1 | **61.1** ±0.1 | **71.5** ±0.1 | **64.2** ±0.1 |

Table 6. Most-Influential-First Deletion (MIF) Accuracy evaluated using predicted labels across multiple models.

## Acknowledgements

The first author would like to express heartfelt thanks to his family for their steadfast support.

## References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, 2020. Association for Computational Linguistics. 2, 115

[2] Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning*, pages 135–168. PMLR, 2024. 2, 5, 115

[3] Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. XAI for transformers: Better explanations through conservative propagation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 435–451. PMLR, 2022. 2, 115

[4] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus H. Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2017. 2, 114

[5] Christopher J. Anders, David Neumann, Talmaj Marinc, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. Xai for analyzing and unlearning spurious correlations in imagenet. 2020. 114

[6] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10, 2015. 2

[7] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *ArXiv*, abs/2106.08254, 2021. 5

[8] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Joshua B. Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Neural Information Processing Systems*, 2019. 6

[9] Daniel Becking, Maximilian Dreyer, Wojciech Samek, Karsten Müller, and Sebastian Lapuschkin. Ecqx: Explainability-driven quantization for low-bit and sparse dnns. *ArXiv*, abs/2109.04236, 2021. 114

[10] Lucas Beyer, Olivier J. H'enaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *ArXiv*, abs/2006.07159, 2020. 6

[11] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim M. Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14496–14506, 2022. 5

[12] Alexander Binder, Sebastian Bach, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for deep neural network architectures. 2016. 1, 114

[13] Stefan Blücher, Johanna Vielhaben, and Nils Strodthoff. Decoupling pixel flipping and occlusion strategy for consistent xai benchmarks. *Transactions on Machine Learning Research*, 2024. 6, 12

[14] Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *International Conference on Learning Representations*, 2020. 115

[15] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 2

[16] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, 2021. 2, 115

[17] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, 2021. 2, 6, 13, 115

[18] Hila Chefer, Idan Schwartz, and Lior Wolf. Optimizing relevance maps of vision transformers improves robustness. *ArXiv*, abs/2206.01161, 2022. 114

[19] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ArXiv*, abs/2301.13826, 2023. 114

[20] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020. 6, 12

[21] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost, 2016. 5

[22] Ian Covert, Chanwoo Kim, and Su-In Lee. Learning to estimate shapley values with vision transformers. *ArXiv*, abs/2206.05282, 2022. 6, 13

[23] Mayukh Deb, Björn Deiseroth, Samuel Weinbach, Patrick Schramowski, and Kristian Kersting. Atman: Understanding transformer predictions through memory efficient attention manipulation. *CoRR*, abs/2301.08110, 2023. 116

[24] 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, 2009. 6, 12

[25] 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. In *International Conference on Learning Representations*, 2021. 2, 5

[26] Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. 2020. 2, 114

[27] Sami Ede, Serop Baghdadlian, Leander Weber, An Thai Nguyen, Dario Zanca, Wojciech Samek, and Sebastian Lapuschkin. Explain to not forget: Defending against catastrophic forgetting with xai. In *International Cross-Domain Conference on Machine Learning and Knowledge Extraction*, 2022. 114

[28] Yuxin Fang, Wen Wang, Binhui Xie, Quan-Sen Sun, Ledell Yu Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369, 2022. 5

[29] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *ArXiv*, abs/2303.11331, 2023. 5

[30] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Juergen Gall. Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision*, 2021. 114

[31] Thomas Fel, Agustin Picard, Louis Béthune, Thibaut Boissin, David Vigouroux, Julien Colin, R'emi Cadene, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2711–2721, 2022. 1, 114

[32] Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 6, 12, 115

[33] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *ArXiv*, abs/2008.02312, 2020. 2, 114

[34] Shangqi Gao, Zhong-Yu Li, Ming-Hsuan Yang, Mingg-Ming Cheng, Junwei Han, and Philip H. S. Torr. Large-scale unsupervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:7457–7476, 2021. 6

[35] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. In *AAAI Conference on Artificial Intelligence*, 2020. 2

[36] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 6

[37] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 6

[38] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 6

[39] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods. In *ICCV*, pages 3981–3991. IEEE, 2023. 6, 7

[40] Yi Huang and Adams Wai-Kin Kong. Transferable adversarial attack based on integrated gradients. *ArXiv*, abs/2205.13152, 2022. 114

[41] Brian Kenji Iwana, Ryohei Kuroki, and Seiichi Uchida. Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4176–4185, 2019. 14

[42] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. 2

[43] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. *ArXiv*, abs/2308.12964, 2023. 114

[44] Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. Investigating the influence of noise and distractors on the interpretation of neural networks. *CoRR*, abs/1611.07270, 2016. 114

[45] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online, 2020. Association for Computational Linguistics. 115

[46] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4547–4568, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 115

[47] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 7, 14

[48] QING LYU, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey. *ArXiv*, abs/2209.11326, 2022. 1, 114

[49] Andreas Madsen, Siva Reddy, and A. P. Sarath Chandar. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55:1 – 42, 2021. 1, 114

[50] Faridoun Mehri, Mohsen Fayyaz, Mahdieh Soleymani Baghshah, and Mohammad Taher Pilehvar. SkipPLUS: Skip the first few layers to better explain vision transformers. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 204–215, 2024. 2, 6, 12, 13, 14, 114

[51] Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. GlobEnc: Quantifying

global token attribution by incorporating the whole encoder layer in transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 258–271, Seattle, United States, 2022. Association for Computational Linguistics. 115

[52] A. Modarressi, Hosein Mohebbi, and Mohammad Taher Pilehvar. Adapler: Speeding up inference by adaptive length reduction. In *Annual Meeting of the Association for Computational Linguistics*, 2022. 114

[53] Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. DecompX: Explaining transformers decisions by propagating token decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2649–2664, Toronto, Canada, 2023. Association for Computational Linguistics. 2, 5, 6, 12, 115

[54] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recogn.*, 65(C):211–222, 2017. 115

[55] Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana, 2018. Association for Computational Linguistics. 6, 12

[56] Fahimeh Hosseini Noohdani, Parsa Hosseini, Arian Yazdan Parast, Hamidreza Yaghoubi Araghi, and Mahdieh Soleymani Baghshah. Decompose-and-compose: A compositional approach to mitigating spurious correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 114

[57] Paul Novello, Thomas Fel, and David Vigouroux. Making sense of dependence: Efficient black-box explanations using dependence measure. *ArXiv*, abs/2206.06219, 2022. 116

[58] Roni Paiss, Hila Chefer, and Lior Wolf. No token left behind: Explainability-aided image classification and generation. In *Computer Vision – ECCV 2022*, pages 334–350, Cham, 2022. Springer Nature Switzerland. 114

[59] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 6

[60] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *ArXiv*, abs/2208.06366, 2022. 5

[61] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *ArXiv*, abs/1806.07421, 2018. 116

[62] Yao Qiang, Deng Pan, Chengyin Li, Xin Li, Rhongho Jang, and Dongxiao Zhu. AttCAT: Explaining transformers via attentive class activation tokens. In *Advances in Neural Information Processing Systems*, 2022. 2, 115

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

[64] Pranav Rajpurkar, Jeremy A. Irvin, Aarti Bagul, Daisy Yi Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L. Ball, C. Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and A. Ng. Mura dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs. *ArXiv*, abs/1712.06957, 2017. 6

[65] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019. 6

[66] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. 116

[67] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109: 247–278, 2021. 1, 114

[68] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 2, 114

[69] Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2591–2600, 2019. 114

[70] Noam M. Shazeer. Glu variants improve transformer. *ArXiv*, abs/2002.05202, 2020. 3

[71] Wei Shi, Wentao Zhang, Weishi Zheng, and Ruixuan Wang. Pami: partition input and aggregate outputs for model interpretation. *ArXiv*, abs/2302.03318, 2023. 116

[72] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *ArXiv*, abs/1605.01713, 2016. 2, 114

[73] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, 2017. 2, 114

[74] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. 114

[75] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014. 114

[76] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *Neural Information Processing Systems*, 2019. 2, 3, 5, 114

[77] Quan Sun, Yuxin Fang, Ledell Yu Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *ArXiv*, abs/2303.15389, 2023. 5

[78] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. 2, 5, 7, 114

[79] Mohammad Reza Taesiri, Giang Nguyen, Sarra Habchi, Cor-Paul Bezemer, and Anh Nguyen. Imagenet-hard: The hardest images remaining from a study of the power of zoom and spatial biases in image classification. 2023. 6

[80] Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. In *Neural Information Processing Systems*, 2021. 6

[81] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herv'e J'egou. Training data-efficient image transformers & distillation through attention. *ArXiv*, abs/2012.12877, 2020. 5

[82] Hugo Touvron, Matthieu Cord, and Herv'e J'egou. Deit iii: Revenge of the vit. In *European Conference on Computer Vision*, 2022. 5

[83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[84] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy, 2019. Association for Computational Linguistics. 114

[85] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 6

[86] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr (Peter) Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 111–119, 2019. 116

[87] Leander Weber, Sebastian Lapuschkin, Alexander Binder, and Wojciech Samek. Beyond explaining: Opportunities and challenges of xai-based model improvement. *Inf. Fusion*, 92: 154–176, 2022. 114

[88] Junyi Wu, Bin Duan, Weitai Kang, Hao Tang, and Yan Yan. Token transformation matters: Towards faithful post-hoc explanation for vision transformer. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10926–10935, 2024. 2, 6, 12, 13, 115

[89] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R. Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1158–1167, 2020. 114

[90] Weiyan Xie, Xiao hui Li, Caleb Chen Cao, and Nevin L.Zhang. Vit-cx: Causal explanation of vision transformers. In *International Joint Conference on Artificial Intelligence*, 2022. 6, 116

[91] Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane ling Wang, and Michael I. Jordan. Ml-loo: Detecting adversarial examples with feature attribution. In *AAAI Conference on Artificial Intelligence*, 2019. 114

[92] Weihao Yu and Xinchao Wang. Mambaout: Do we really need mamba for vision? *arXiv preprint arXiv:2405.07992*, 2024. 3

[93] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *ArXiv*, abs/2303.15343, 2023. 5

[94] Jianming Zhang, Zhe L. Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126: 1084–1102, 2016. 114

[95] Jianping Zhang, Weibin Wu, Jen tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R. Lyu. Improving adversarial transferability via neuron attribution-based attacks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14973–14982, 2022. 114

# LibraGrad: Balancing Gradient Flow for Universally Better Vision Transformer Attributions

Supplementary Material

## Table of Contents

# A. Method: Further Details

## A.1. Gradient Manipulation Operators

**Constant Operator.** The constant operator $[\cdot]_{\text{cst.}} : \mathbb{R}^m \to \mathbb{R}^m$ satisfies:

$$[y]_{\text{cst.}} = y, \quad J_x [y]_{\text{cst.}} = 0$$

**SwapBackward.** The SwapBackward : $(f, g) \mapsto h$ operator, where $f, g, h : \mathbb{R}^n \to \mathbb{R}^m$, is defined by:

$$h(x) = f(x), \quad J_x h = J_x g$$

*Remark* 1 (Duality). These operators are dual: the constant operator can be implemented via SwapBackward by scaling to zero:

$$[y]_{\text{cst.}} \equiv \text{SwapBackward}(y, 0)$$

while SwapBackward can be constructed from the constant operator:

$$\text{SwapBackward}(f, g)(x) = [f(x)]_{\text{cst.}} + (g(x) - [g(x)]_{\text{cst.}})$$

*Remark* 2 (PyTorch Implementation). In PyTorch, the constant operator can be implemented using `detach`:

$$[y]_{\text{cst.}} \equiv y.\text{detach}()$$

For SwapBackward, we have two equivalent implementations:

1. Via duality: SwapBackward$(f, g)(x) = f(x)$.detach() + $(g(x) - g(x)$.detach())
2. Via custom backward: Define an `autograd.Function` that returns $f(x)$ in `forward` and propagates gradients as if it were $g(x)$ in `backward`

Both implementations yield identical gradients, though the latter may be more computationally efficient, while the former may be easier to implement.

*Remark* 3 (Computational Efficiency). Both core operations of LibraGrad preserve or improve efficiency—constant operators reduce computation through pruning, while SwapBackward maintains original complexity regardless of implementation. See Table 1 for comparative analysis.

## A.2. Intuition Behind Balancing the Gradient Flow

Balanced gradient flow (or FG-completeness) ensures proper attribution across a model's parallel computational paths. Consider the simple function:

$$f(a, c) = a^2 + c$$

where $a^2$ represents feature fusion (similar to SwiGLU) and $c$ is a residual connection. Since this function lacks bias terms, FullGrad reduces to Input $\times$ Grad. Evaluating the IxG attributions yields:

$$\text{IxG}\,(f)(a, c) = \begin{bmatrix} a \cdot 2a \\ c \cdot 1 \end{bmatrix} = \begin{bmatrix} 2a^2 \\ c \end{bmatrix}$$

Summing these attributions gives $2a^2 + c$, which double-counts $a$'s contribution relative to $c$ in $f(a, c) = a^2 + c$. *I.e.*, the gradient of $a^2$ is *unbalanced* compared to the gradient of $c$, and the function $f$ is thus non-FG-complete.

Balancing gradient flow matters most when multiple parallel paths exist. (The main cause of parallel paths is the presence of residual connections.) When gradients aren't balanced, the relative contributions between paths become distorted. However, when there is only one single path of gradient flow, balancing the gradients becomes less critical.

For example, we discard softmax gradients within attention modules, but retain them at the end of classification models where they convert logits to probabilities. Similarly, we retain the standard gradients of the non-FG-complete length normalization on image embeddings at the end of CLIP models. Though these terminal operations are not FG-complete, they preserve attribution faithfulness because the model has no competing paths parallel to these modules. We tested several

FG-complete approximations for both of these operations (not further detailed in this paper), but they showed no meaningful improvements over standard gradients.

In summary, the contribution of a module that is FG-complete will be properly accounted for when placed in a path parallel to other FG-complete modules. Conversely, if a module is not FG-complete, its contribution will be either overrepresented or underrepresented when positioned parallel to other modules.

## A.3. Theorems

### A.3.1. FullGrad-Completeness of Affine Functions

**Definition 3.** A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is **affine** if it can be expressed as $f(x) = Wx + b$ for some matrix $W \in \mathbb{R}^{m\times n}$ and vector $b \in \mathbb{R}^m$.

**Theorem 6.** *Any affine function $f : \mathbb{R}^n \to \mathbb{R}^m$ is FG-complete.*

*Proof.* Let $f(x) = Wx + b$ be an affine function. The Jacobians are:

$$J_x f = W, \quad J_b f = I,$$

where $I$ is the identity matrix. By direct computation:

$$J_x f \cdot x + J_b f \cdot b = Wx + b = f(x),$$

proving FG-completeness. □

### A.3.2. FullGrad-Completeness of Locally Affine Functions

**Definition 4.** A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is **locally affine** at a point $x_0 \in \mathbb{R}^n$ if there exists an open neighborhood $U \subset \mathbb{R}^n$ containing $x_0$, a matrix $W(x_0) \in \mathbb{R}^{m\times n}$, and a vector $b(x_0) \in \mathbb{R}^m$ such that

$$f(x) = W(x_0)x + b(x_0), \quad \forall x \in U.$$

**Example 1.** Consider the ReLU function ReLU : $\mathbb{R} \to \mathbb{R}$ defined by $\mathrm{ReLU}(x) = \max(0, x)$. The ReLU function is locally affine at every point $x_0 \neq 0$:

- For $x_0 > 0$: $\mathrm{ReLU}(x) = x$ in a neighborhood, so $W(x_0) = 1$, $b(x_0) = 0$
- For $x_0 < 0$: $\mathrm{ReLU}(x) = 0$ in a neighborhood, so $W(x_0) = 0$, $b(x_0) = 0$

**Theorem 1.** *Any locally affine function at $x_0$ is FG-complete in a neighborhood of $x_0$.*

*Proof.* Let $f$ be locally affine at $x_0$. By definition, there exists an open neighborhood $U$ of $x_0$ and matrices $W(x_0)$, $b(x_0)$ such that for all $x \in U$:

$$f(x) = W(x_0)x + b(x_0)$$

This is an affine function in $U$, and thus by Theorem 6, it is FG-complete in $U$. □

### A.3.3. FullGrad-Completeness of Composition of Two Functions

**Theorem 7.** *Let $f_1, f_2$ be FG-complete functions. Then their composition $f = f_2 \circ f_1$ is also FG-complete.*

*Proof.* Let $y = f_1(x)$. By FG-completeness of $f_1$ and $f_2$:

$$f_1(x) = J_x f_1 \cdot x + \sum_i J_{b_i} f_1 \cdot b_i$$

$$f_2(y) = J_y f_2 \cdot y + \sum_j J_{c_j} f_2 \cdot c_j$$

where $b_i$ and $c_j$ are bias terms in $f_1$ and $f_2$ respectively.

For the composition $f = f_2 \circ f_1$, by the chain rule:

$$J_x f = J_y f_2 \cdot J_x f_1$$

For bias terms $b_i$ in $f_1$:

$$J_{b_i} f = J_y f_2 \cdot J_{b_i} f_1$$

For bias terms $c_j$ in $f_2$:

$$J_{c_j} f = J_{c_j} f_2$$

Therefore:

$$\begin{aligned} J_x f \cdot x + \sum_i J_{b_i} f \cdot b_i + \sum_j J_{c_j} f \cdot c_j &= J_y f_2 \cdot J_x f_1 \cdot x + \sum_i J_y f_2 \cdot J_{b_i} f_1 \cdot b_i + \sum_j J_{c_j} f_2 \cdot c_j \\ &= J_y f_2 \cdot (J_x f_1 \cdot x + \sum_i J_{b_i} f_1 \cdot b_i) + \sum_j J_{c_j} f_2 \cdot c_j \\ &= J_y f_2 \cdot f_1(x) + \sum_j J_{c_j} f_2 \cdot c_j \\ &= J_y f_2 \cdot y + \sum_j J_{c_j} f_2 \cdot c_j \\ &= f_2(y) = f_2(f_1(x)) = f(x) \end{aligned}$$

proving the FG-completeness of the composition. □

### A.3.4. FullGrad-Completeness of Finite Function Compositions

**Theorem 2.** *The composition of a finite number of FG-complete functions is FG-complete.*

*Proof.* Let $f = f_k \circ \cdots \circ f_1$ be a composition of $k$ FG-complete functions. We prove the result by induction on $k$.

**Base case ($k = 1$):** A single FG-complete function is FG-complete by definition.

**Inductive hypothesis:** Assume the composition of $n$ FG-complete functions is FG-complete.

**Inductive step:** Consider a composition of $n + 1$ FG-complete functions:

$$g = f_{n+1} \circ f_n \circ \cdots \circ f_1$$

Let $h = f_n \circ \cdots \circ f_1$. By the inductive hypothesis, $h$ is FG-complete. Then $g = f_{n+1} \circ h$ is a composition of two FG-complete functions, which is FG-complete by Theorem 7.

By induction, the composition of any finite number of FG-complete functions is FG-complete. □

**Corollary 4.** *The composition of a finite number of locally affine functions at $x_0$ is FG-complete in a neighborhood of $x_0$.*

### A.3.5. FullGrad-Completeness of Function Addition

**Theorem 3.** *Let $f_1, f_2$ be FG-complete functions. Then their sum $f = f_1 + f_2$ is FG-complete.*

*Proof.* Since $f_1$ and $f_2$ are FG-complete, we have:

$$f_1(x) = J_x f_1 \cdot x + \sum_i J_{b_i} f_1 \cdot b_i$$

$$f_2(x) = J_x f_2 \cdot x + \sum_j J_{c_j} f_2 \cdot c_j$$

Then, for their sum $f(x) = f_1(x) + f_2(x)$, the Jacobians are:

$$J_x f = J_x f_1 + J_x f_2$$

$$J_{b_i} f = J_{b_i} f_1, \quad J_{c_j} f = J_{c_j} f_2$$

Therefore:

$$
\begin{aligned}
J_x f \cdot x + \sum_i J_{b_i} f \cdot b_i + \sum_j J_{c_j} f \cdot c_j &= (J_x f_1 + J_x f_2) \cdot x + \sum_i J_{b_i} f_1 \cdot b_i + \sum_j J_{c_j} f_2 \cdot c_j \\
&= [J_x f_1 \cdot x + \sum_i J_{b_i} f_1 \cdot b_i] + [J_x f_2 \cdot x + \sum_j J_{c_j} f_2 \cdot c_j] \\
&= f_1(x) + f_2(x) \\
&= f(x)
\end{aligned}
$$

Thus, $f$ is FG-complete. □

**Corollary 5.** *Let $f$ be FG-complete. Then the residual connection defined by $g(x) = x + f(x)$ is FG-complete.*

#### A.3.6. Gradient Flow in Element-Wise Multiplication

We first show that the naive approach to element-wise multiplication is not FG-complete.

**Proposition 1.** *Let $f_1, f_2$ be FG-complete functions and let $f(x) = f_1(x) \odot f_2(x)$ be their element-wise product with Jacobians:*

$$
\begin{aligned}
J_x f &= \mathit{diag}(f_2(x)) \cdot J_x f_1 + \mathit{diag}(f_1(x)) \cdot J_x f_2 \\
J_{b_i} f &= \mathit{diag}(f_2(x)) \cdot J_{b_i} f_1 + \mathit{diag}(f_1(x)) \cdot J_{b_i} f_2
\end{aligned}
$$

*Then $f$ is not FG-complete. Specifically:*

$$
J_x f \cdot x + \sum_i J_{b_i} f \cdot b_i = 2f(x)
$$

*Proof.* Since $f_1$ and $f_2$ are FG-complete:

$$
f_1(x) = J_x f_1 \cdot x + \sum_i J_{b_i} f_1 \cdot b_i
$$

$$
f_2(x) = J_x f_2 \cdot x + \sum_i J_{b_i} f_2 \cdot b_i
$$

Computing $J_x f \cdot x + \sum_i J_{b_i} f \cdot b_i$ with the standard Jacobians:

$$
\begin{aligned}
&[\text{diag}(f_2(x)) \cdot J_x f_1 + \text{diag}(f_1(x)) \cdot J_x f_2] \cdot x + \\
&\sum_i [\text{diag}(f_2(x)) \cdot J_{b_i} f_1 + \text{diag}(f_1(x)) \cdot J_{b_i} f_2] \cdot b_i \\
&= \text{diag}(f_2(x)) \cdot (J_x f_1 \cdot x + \sum_i J_{b_i} f_1 \cdot b_i) + \\
&\quad \text{diag}(f_1(x)) \cdot (J_x f_2 \cdot x + \sum_i J_{b_i} f_2 \cdot b_i) \\
&= \text{diag}(f_2(x)) \cdot f_1(x) + \text{diag}(f_1(x)) \cdot f_2(x) \\
&= f_2(x) \odot f_1(x) + f_1(x) \odot f_2(x) = 2f(x)
\end{aligned}
$$

Therefore, the naive element-wise product yields twice the desired output in the FG-completeness equation, making it not FG-complete. □

However, by properly scaling the Jacobian terms, we can achieve FG-completeness:

**Theorem 4.** *Let $f_1, f_2$ be FG-complete functions. Then their element-wise product $f(x) = f_1(x) \odot f_2(x)$ is FG-complete when its Jacobians are defined with scaling coefficients $a, b \in \mathbb{R}$ where $a + b = 1$:*

$$
J_x f = a[\mathit{diag}(f_2(x)) \cdot J_x f_1] + b[\mathit{diag}(f_1(x)) \cdot J_x f_2]
$$

$$
J_{b_i} f = a[\mathit{diag}(f_2(x)) \cdot J_{b_i} f_1] + b[\mathit{diag}(f_1(x)) \cdot J_{b_i} f_2]
$$

*Proof.* The proof follows the same structure as Proposition 1, but with scaled Jacobians:

$$
\begin{aligned}
&[a\text{diag}(f_2(x)) \cdot J_x f_1 + b\text{diag}(f_1(x)) \cdot J_x f_2] \cdot x + \\
&\sum_i [a\text{diag}(f_2(x)) \cdot J_{b_i} f_1 + b\text{diag}(f_1(x)) \cdot J_{b_i} f_2] \cdot b_i \\
&= a\text{diag}(f_2(x)) \cdot (J_x f_1 \cdot x + \sum_i J_{b_i} f_1 \cdot b_i) + \\
&\quad b\text{diag}(f_1(x)) \cdot (J_x f_2 \cdot x + \sum_i J_{b_i} f_2 \cdot b_i) \\
&= a\text{diag}(f_2(x)) \cdot f_1(x) + b\text{diag}(f_1(x)) \cdot f_2(x) \\
&= (a+b)(f_1(x) \odot f_2(x)) = f_1(x) \odot f_2(x) = f(x)
\end{aligned}
$$

where the last equality follows from $a+b=1$, proving the FG-completeness of $f$ with the scaled Jacobian definitions. □

**Theorem 5.** *Let $f_1, f_2$ be arbitrary functions (not necessarily FG-complete), and let $f(x) = f_1(x) \odot f_2(x)$ be their element-wise product. Consider $f$ with scaled Jacobians as defined in Theorem 4. Then:*

1. *When $a = 0$, yielding $f(x) = [f_1(x)]_{cst.} \odot f_2(x)$ where $[\cdot]_{cst.}$ is the constant operator that zeroes gradients, $f$ is FG-complete if $f_2$ is FG-complete.*
2. *By symmetry, when $b = 0$, $f$ is FG-complete if $f_1$ is FG-complete.*

*Proof.* Let $a = 0$ (thus $b = 1$). If $f_2$ is FG-complete:

$$
\begin{aligned}
&[\text{diag}(f_1(x)) \cdot J_x f_2] \cdot x + \sum_i [\text{diag}(f_1(x)) \cdot J_{b_i} f_2] \cdot b_i \\
&= \text{diag}(f_1(x)) \cdot (J_x f_2 \cdot x + \sum_i J_{b_i} f_2 \cdot b_i) \\
&= \text{diag}(f_1(x)) \cdot f_2(x) \\
&= f_1(x) \odot f_2(x) = f(x)
\end{aligned}
$$

proving the FG-completeness of $f$. □

### A.3.7. Non-FG-Completeness of SiLU Activation

**Proposition 4.** *The SiLU activation function $SiLU(x) = x \cdot \sigma(x)$, where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function, is not FG-complete. Specifically, there exists $x \in \mathbb{R}$ such that:*

$$J_x SiLU \cdot x \neq SiLU(x)$$

*Proof.* The Jacobian of SiLU is:

$$J_x \text{SiLU} = \sigma(x) + x\sigma'(x)$$

where $\sigma'(x) = \sigma(x)(1 - \sigma(x))$ is the derivative of the sigmoid function.

Therefore:

$$
\begin{aligned}
J_x \text{SiLU} \cdot x &= x\sigma(x) + x^2\sigma'(x) \\
&= x\sigma(x) + x^2\sigma(x)(1 - \sigma(x)) \\
&= x\sigma(x) \left(1 + x(1 - \sigma(x))\right) \\
&= \text{SiLU}(x) \left(1 + x(1 - \sigma(x))\right) \\
&= \text{SiLU}(x) \left(1 + x - x\sigma(x)\right) \\
&= \text{SiLU}(x) \left(1 + x - \text{SiLU}(x)\right)
\end{aligned}
$$

For $J_x \text{SiLU} \cdot x = \text{SiLU}(x)$, we require:

$$\text{SiLU}(x) \left(1 + x - \text{SiLU}(x)\right) = \text{SiLU}(x)$$

Subtracting $\text{SiLU}(x)$ from both sides:

$$\text{SiLU}(x)\,(1 + x - \text{SiLU}(x)) - \text{SiLU}(x) = 0$$

Simplifying:

$$\text{SiLU}(x)\,((1 + x - \text{SiLU}(x)) - 1) = 0$$

$$\text{SiLU}(x)\,(x - \text{SiLU}(x)) = 0$$

Thus, we require either $\text{SiLU}(x) = 0$, or $x = \text{SiLU}(x)$:

- $\text{SiLU}(x) = 0$, which happens when $x = 0$, or when $\sigma(x) = 0$, requiring $x \to -\infty$, leading to $\text{SiLU}(x) = x \cdot 0 = 0$.
- $x = \text{SiLU}(x)$, which occurs when $\sigma(x) = 1$, requiring $x \to \infty$.

For all other values of $x$, we have $J_x\text{SiLU} \cdot x \neq \text{SiLU}(x)$. For example, at $x = 1$:

$$\text{SiLU}(1) = 1 \cdot \sigma(1) \approx 0.731$$

$$J_x\text{SiLU} \cdot x = \text{SiLU}(1)\,(1 + 1 - \text{SiLU}(1)) \approx 0.731 \times (1 + 1 - 0.731) \approx 0.731 \times 1.269 \approx 0.928 \neq 0.731$$

proving that SiLU is not FG-complete. □

#### A.3.8. Gradient Flow in Division

**Proposition 2.** *Let $f_1, f_2$ be FG-complete functions with $f_2$ non-zero. FullGrad vanishes to exactly zero on their element-wise quotient $f(x) = f_1(x) \oslash f_2(x)$.*

*Proof.* Since $f_1$ and $f_2$ are FG-complete, we have:

$$f_1(x) = J_x f_1 \cdot x + \sum_i J_{b_i^{(1)}} f_1 \cdot b_i^{(1)},$$
$$f_2(x) = J_x f_2 \cdot x + \sum_j J_{b_j^{(2)}} f_2 \cdot b_j^{(2)}.$$

The Jacobian of $f$ with respect to $x$ is:

$$J_x f = \operatorname{diag}\left(\frac{1}{f_2(x)}\right) J_x f_1 - \operatorname{diag}\left(\frac{f_1(x)}{f_2(x)^2}\right) J_x f_2,$$

where $\operatorname{diag}(v)$ denotes a diagonal matrix with vector $v$ on the diagonal and the fractions denote element-wise division.

Similarly, the Jacobians with respect to the biases are:

$$J_{b_i^{(1)}} f = \operatorname{diag}\left(\frac{1}{f_2(x)}\right) J_{b_i^{(1)}} f_1,$$
$$J_{b_j^{(2)}} f = -\operatorname{diag}\left(\frac{f_1(x)}{f_2(x)^2}\right) J_{b_j^{(2)}} f_2.$$

Now, compute the FullGrad attributions of $f$:

$$
\begin{aligned}
&J_x f \cdot x + \sum_i J_{b_i^{(1)}} f \cdot b_i^{(1)} + \sum_j J_{b_j^{(2)}} f \cdot b_j^{(2)} \\
&\quad = \left[ \operatorname{diag}\left(\frac{1}{f_2(x)}\right) J_x f_1 - \operatorname{diag}\left(\frac{f_1(x)}{f_2(x)^2}\right) J_x f_2 \right] \cdot x \\
&\qquad + \sum_i \operatorname{diag}\left(\frac{1}{f_2(x)}\right) J_{b_i^{(1)}} f_1 \cdot b_i^{(1)} - \sum_j \operatorname{diag}\left(\frac{f_1(x)}{f_2(x)^2}\right) J_{b_j^{(2)}} f_2 \cdot b_j^{(2)} \\
&\quad = \operatorname{diag}\left(\frac{1}{f_2(x)}\right) \left( J_x f_1 \cdot x + \sum_i J_{b_i^{(1)}} f_1 \cdot b_i^{(1)} \right) \\
&\qquad - \operatorname{diag}\left(\frac{f_1(x)}{f_2(x)^2}\right) \left( J_x f_2 \cdot x + \sum_j J_{b_j^{(2)}} f_2 \cdot b_j^{(2)} \right) \\
&\quad = \operatorname{diag}\left(\frac{1}{f_2(x)}\right) \left( J_x f_1 \cdot x + \sum_i J_{b_i^{(1)}} f_1 \cdot b_i^{(1)} \right) \\
&\qquad - \operatorname{diag}\left(\frac{f_1(x)}{f_2(x)^2}\right) \left( J_x f_2 \cdot x + \sum_j J_{b_j^{(2)}} f_2 \cdot b_j^{(2)} \right) \\
&\quad = \operatorname{diag}\left(\frac{1}{f_2(x)}\right) f_1(x) - \operatorname{diag}\left(\frac{f_1(x)}{f_2(x)^2}\right) f_2(x) \\
&\quad = \frac{f_1(x)}{f_2(x)} - \frac{f_1(x)}{f_2(x)} = f(x) - f(x) = 0.
\end{aligned}
$$

□

**Corollary 2.** *Division can be made FG-complete by treating it as element-wise multiplication with a gradient-pruned non-linear reciprocal:* $f(x) = f_1(x) \odot [1/f_2(x)]_{\text{cst.}}$ *which satisfies FG-completeness, by Theorem 5.*

### A.3.9. How Does FullGrad Behave on LayerNorm?

**Proposition 3.** *For the LayerNorm operation without affine parameters:*

$$
LN(x)_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \varepsilon}},
$$

*where* $\mu = \frac{1}{N}\sum_{k=1}^{N} x_k$ *and* $\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2$, *FullGrad approaches zero as* $\varepsilon$ *approaches zero:*

$$
\lim_{\varepsilon \to 0} J_x LN \cdot x = 0.
$$

*Proof.* Let $x \in \mathbb{R}^N$. We decompose LayerNorm into two operations:

1. Centering: $y = x - \mu\mathbf{1}$, where $\mathbf{1}$ is the vector of ones
2. Scaling: $z = y/s$, where $s = \sqrt{\sigma^2 + \varepsilon}$

The Jacobian of centering is:

$$
(J_x y)_{ij} = \delta_{ij} - \frac{1}{N}
$$

which gives $(J_x y \cdot x)_i = x_i - \mu = y_i$.

The Jacobian of scaling is:

$$
(J_y z)_{ij} = \frac{\delta_{ij}}{s} - \frac{y_i y_j}{N s^3}
$$

By the chain rule:

$$
J_x \mathrm{LN} \cdot x = J_y z \cdot J_x y \cdot x = J_y z \cdot y
$$

Computing $(J_y z \cdot y)_i$:

$$
\begin{aligned}
(J_y z \cdot y)_i &= \sum_{j=1}^{N} \left( \frac{\delta_{ij}}{s} - \frac{y_i y_j}{N s^3} \right) y_j \\
&= \frac{y_i}{s} - \frac{y_i}{N s^3} \sum_{j=1}^{N} y_j^2 \\
&= \frac{y_i}{s} - \frac{y_i \sigma^2}{s^3} \\
&= \frac{y_i}{s} - \frac{y_i (s^2 - \varepsilon)}{s^3} \\
&= y_i \cdot \frac{\varepsilon}{s^3}
\end{aligned}
$$

Since $s = \sqrt{\sigma^2 + \varepsilon} \geq \sqrt{\sigma^2}$ for all $\varepsilon > 0$, and $y_i$ is independent of $\varepsilon$, we have for each component $i$:

$$
\lim_{\varepsilon \to 0} (J_x \mathrm{LN} \cdot x)_i = \lim_{\varepsilon \to 0} y_i \cdot \frac{\varepsilon}{s^3} = 0,
$$

completing the proof. □

### A.3.10. Non-Viability of Integrated Gradients on LayerNorm

**Proposition 5.** *For the LayerNorm operation without affine parameters as defined in Proposition 3, Integrated Gradients with a zero baseline approaches zero when approximated using an $n$-step (with $n$ fixed) Riemann summation as $\varepsilon$ approaches zero.*

*Proof.* For any baseline $\bar{x}$, Integrated Gradients can be written as:

$$
\mathrm{IG}(x, \bar{x}) = \int_0^1 J_x \mathrm{LN}(\bar{x} + \alpha(x - \bar{x})) \cdot (x - \bar{x}) \, d\alpha
$$

Using an $n$-step Riemann sum approximation:

$$
\mathrm{IG}(x, \bar{x}) \approx \frac{1}{n} \sum_{k=1}^{n} J_x \mathrm{LN}(\bar{x} + \frac{k}{n}(x - \bar{x})) \cdot (x - \bar{x})
$$

Setting $\bar{x} = 0$:

$$
\mathrm{IG}(x, 0) \approx \frac{1}{n} \sum_{k=1}^{n} J_x \mathrm{LN}(\frac{k}{n} x) \cdot x
$$

From Proposition 3, we know that for any input $x'$:

$$
\lim_{\varepsilon \to 0} J_x \mathrm{LN}(x') \cdot x' = 0
$$

For each step $k$ in the Riemann sum, let $x_k = \frac{k}{n}x$. We can exchange the limit with the finite sum:

$$
\begin{aligned}
\lim_{\varepsilon \to 0} \mathrm{IG}(x, 0) &\approx \lim_{\varepsilon \to 0} \frac{1}{n} \sum_{k=1}^{n} J_x \mathrm{LN}(\frac{k}{n}x) \cdot x \\
&= \frac{1}{n} \sum_{k=1}^{n} \lim_{\varepsilon \to 0} J_x \mathrm{LN}(x_k) \cdot x \\
&= \frac{1}{n} \sum_{k=1}^{n} \lim_{\varepsilon \to 0} \frac{k}{n} J_{x_k} \mathrm{LN}(x_k) \cdot x \\
&= \frac{1}{n} \sum_{k=1}^{n} \lim_{\varepsilon \to 0} J_{x_k} \mathrm{LN}(x_k) \cdot x_k \\
&= \frac{1}{n} \sum_{k=1}^{n} 0 \\
&= 0
\end{aligned}
$$

where we applied Proposition 3 to $x_k$. $\square$

# B. Detailed Experimental Setup

## B.1. Empirical Completeness Evaluation

Consider an attribution method $A$ that assigns relevance scores $A(f)(x)_i$ to each input feature $x_i$ relative to model $f$(see §2 for notation). The Completeness Error (CE) is defined as:

$$\text{CE}(f, x, A) = \left\| f(x) - \sum_{i=1}^{n} A(f)(x)_i \right\| \tag{2}$$

Lower CE values indicate better conservation of the model's output in the attribution scores. We say $A$ is complete on a given architecture $f$ when $\text{CE} = 0$. While our theoretical analysis proves that Transformers exhibit FG-completeness under our modifications, we perform empirical validation to: (1) verify the theoretical guarantees, (2) validate implementation correctness, and (3) demonstrate how prior methods fail to achieve completeness. As this is just a sanity check, we use only 100 random images from the ImageNet dataset [24], and set the attribution target to the predicted logit of the model.

## B.2. Faithfulness Metrics

We evaluate attribution methods through faithfulness metrics that quantify how well attribution scores reflect the true importance of input features to model predictions. These widely used metrics [13, 20, 32, 50, 53, 55, 88] measure changes in model behavior as we progressively occlude input features in different orders. For a given feature ordering $\pi$ and occlusion fraction $s/n$ (where $n$ is the total number of features), we compute the area under curve:

$$\text{AUC}[\pi] = \frac{1}{n} \sum_{s=0}^{n} v^{\text{perf}}(x_{\Pi(s)}) \tag{3}$$

where $\Pi(s)$ represents keeping only the first $s$ features according to ordering $\pi$, and $v^{\text{perf}}(x_{\Pi(s)})$ measures model performance on this partially occluded input. This can be either classification accuracy (more robust to outliers) or the change in predicted probability for the target class (called AOPC, more granular). Both measures can use either ground truth or predicted target classes.

The Most-Influential-First Deletion (MIF) metric measures performance degradation when occluding features in order of decreasing attribution scores:

$$\text{MIF}[\phi] = \text{AUC}[\pi^{\phi}] \tag{4}$$

where $\pi^{\phi}$ orders features by decreasing attribution values. Since lower MIF scores indicate better attributions (faster performance degradation), we normalize it as:

$$\text{MIF}_{\text{norm}}[\phi] = 100 - \text{MIF}[\phi] \tag{5}$$

The Least-Influential-First Deletion (LIF) metric measures performance when occluding features in order of increasing attribution scores:

$$\text{LIF}[\phi] = \text{AUC}[(\pi^{\phi})^{r}] \tag{6}$$

where $(\pi^{\phi})^{r}$ is the reverse ordering. LIF can be interpreted as a counterfactual metric—features with the most negative attribution scores often contribute to competing classes, so their removal can actually increase the target class probability. Since higher LIF scores already indicate better attributions (slower degradation when removing negative contributors), it requires no normalization.

The Symmetric Relevance Gain (SRG) measure [13] is defined as the average of both metrics:

$$\text{SRG}[\phi] = \frac{\text{LIF}[\phi] + \text{MIF}_{\text{norm}}[\phi]}{2} \tag{7}$$

In this work, we primarily focus on MIF with predicted labels and accuracy measurement, as our goal is to identify positive feature contributions to model predictions rather than counterfactual explanations. We report comprehensive results using both accuracy and AOPC metrics for MIF, LIF and SRG using both ground truth and predicted labels in Appendix D.

### B.2.1. True Token Masking

Instead of simply overlaying a color mask, we choose to completely exclude the masked patches from the model's input (for models that support token exclusion) [22, 50]. At the same time, we preserve accurate positional encodings for the unmasked patches. We term this strategy *True Token Masking*. The conventional method of using the color black (or simply zeroing the tokens in text-based Transformers) for patch masking encounters several issues:

- If a patch is predominantly black, painting it black does not effectively eliminate its informational content. For instance, a black drawing on a white background would remain mostly unchanged.
- Patches might serve computational functions, such as acting as a scratchpad for the model's internal processes. Masking these with black does not prevent the model from using them for such purposes.
- Introducing a black mask can create artifacts in the image, potentially leading to out-of-distribution data, which affects the model's performance.

## B.3. Human Interpretability Evaluation

Although lacking a strong theoretical justification, human interpretability evaluations serve as effective sanity checks and provide a quantitative measure that aligns with intuitive inferences drawn from qualitative examples of attribution methods. Following the zero-shot segmentation setup proposed by [17, 50, 88], we report the Average Precision (AP) metric. This evaluation requires a dataset with ground truth labels for the target class. Notably, AP is invariant to shift and scale transformations, mirroring the properties of our faithfulness metrics.

## B.4. Qualitative Evaluation

Our qualitative evaluation comprises two complementary scenarios, each designed to assess different aspects of attribution quality:

**Text-Prompted Attribution on CLIP.** CLIP models are trained to output similarity scores between image-text pairs, enabling flexible zero-shot queries through natural language prompts. Our first evaluation scenario uses the text-image similarity scores output by CLIP models as attribution targets. For each test image, we systematically probe different regions and concepts using targeted text prompts, enabling a detailed assessment of each attribution method's ability to locate described elements within complex scenes.

**Multi-Class Discrimination.** Using ImageNet-finetuned models, we evaluate class discriminativity on carefully selected images from the COCO 2017 training set [47]. We specifically focus on images containing both zebras and elephants within the same frame, with both animals clearly visible and not significantly occluded. Given the rarity of such co-occurrences, our evaluation encompasses all available instances. The attribution target is set to the output class probabilities of "Zebra" and "African Elephant". This choice is motivated by several factors:

- Prior work [41, 50] has established these animals as effective test cases for attribution evaluation.
- ImageNet has a single class for zebras and three classes for elephants, which is in contrast to most other animals that can have tens of different fine-grained ImageNet classes.
- They co-occur in nature.
- Their distinct visual characteristics help verify that attributions are truly class-specific rather than merely highlighting salient regions.

**Method Selection.** We showcase three categories of attribution methods: fundamental gradient-based approaches (Integrated Gradients and FullGrad+), our proposed Libra FullGrad+, and contemporary Transformer-specific methods (AttCAT, AttnLRP, and TokenTM). The latter group was selected based on strong performance on quantitative metrics. Between TokenTM and GenAtt, which generate nearly identical attribution maps, we employ TokenTM as the more recent formulation.

### B.4.1. Qualitative Visualization Method

To visualize attribution maps:

1. **Negative Value Removal:** We first apply ReLU to remove negative attribution scores, as we want to focus on positive feature contributions.
2. **Robust Scaling:** Rather than using absolute maximum values which can be sensitive to outliers, we compute the 99th percentile of the attribution scores. We then scale the values by dividing by this robust maximum.
3. **Spatial Upsampling:** The token-level attribution map is upsampled to the original image resolution using bicubic interpolation.
4. **Range Normalization:** Finally, we clamp values to $[0, 1]$.

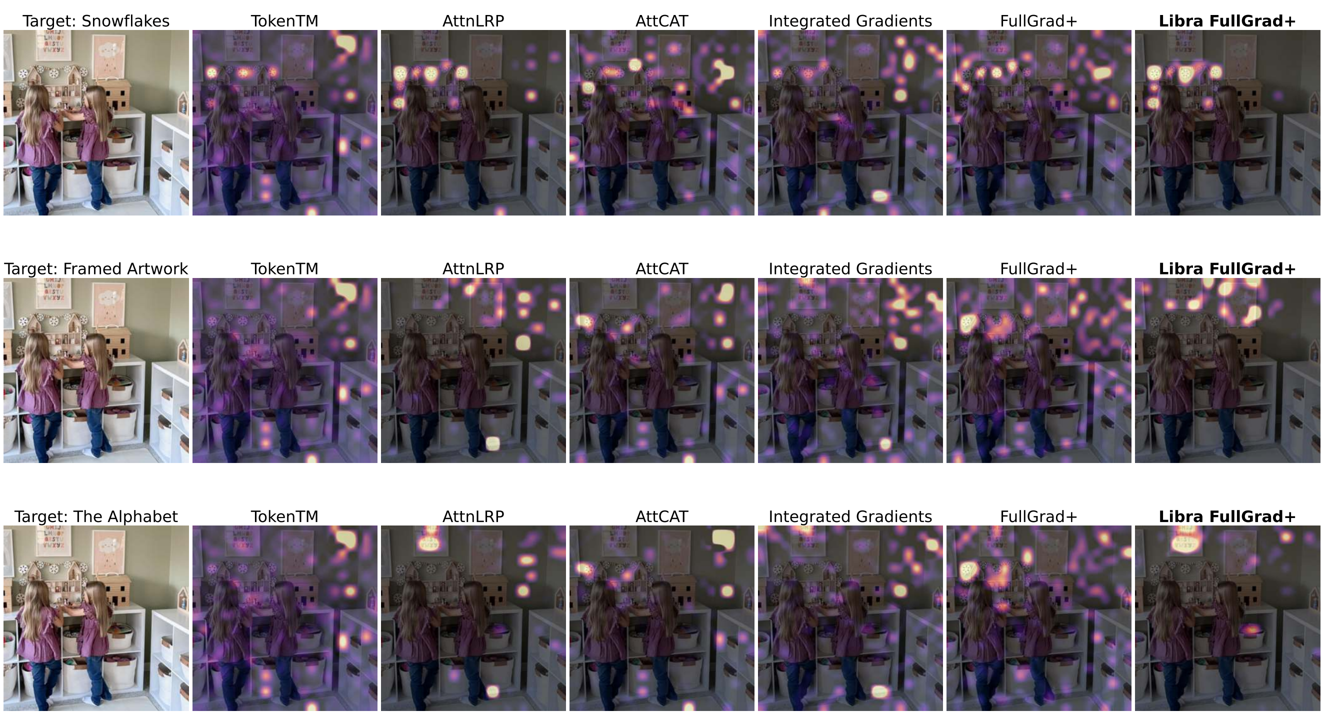
# C. Qualitative Results

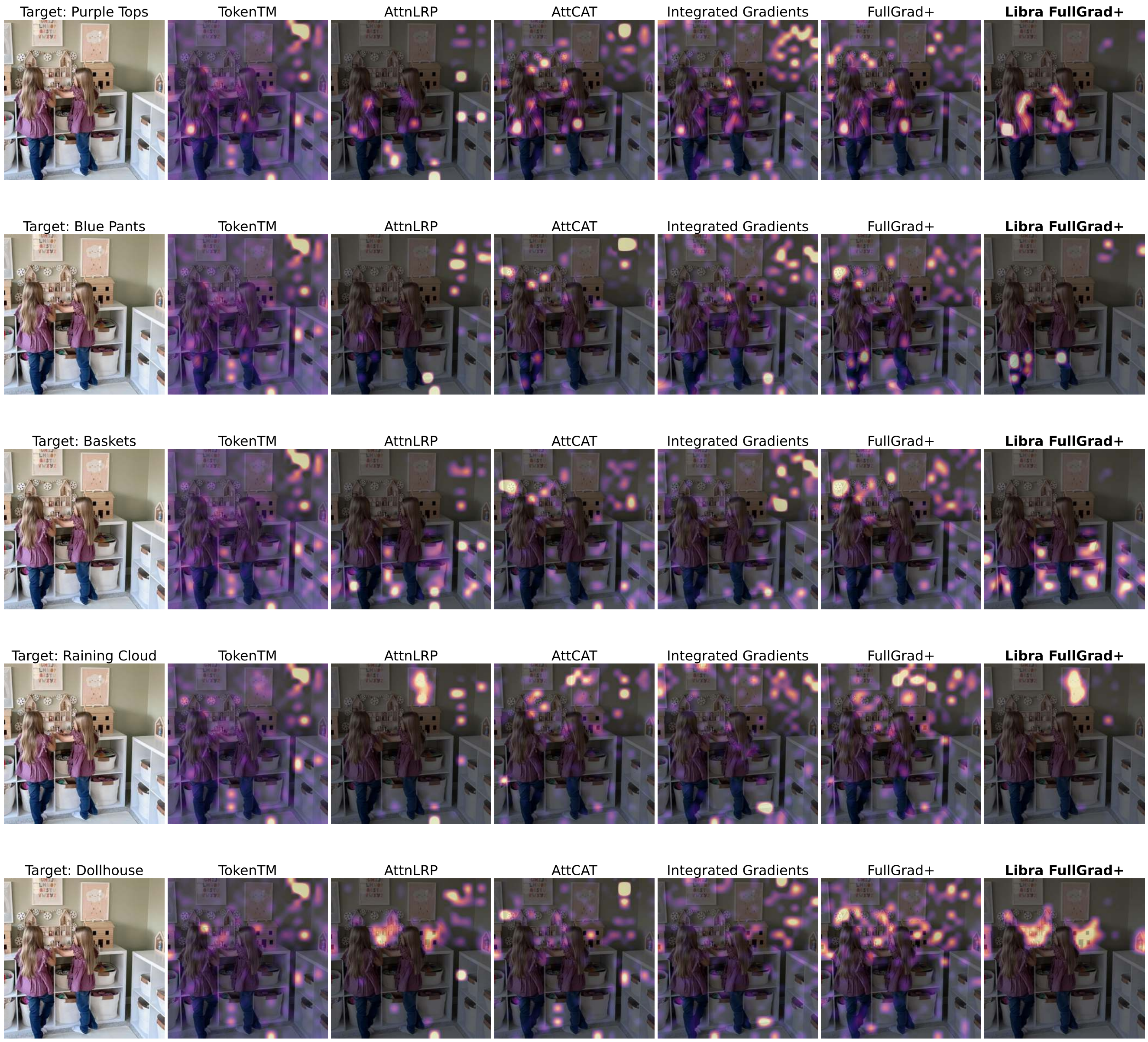
Following the evaluation protocol in Appendix B.4, we present a comprehensive qualitative analysis below.

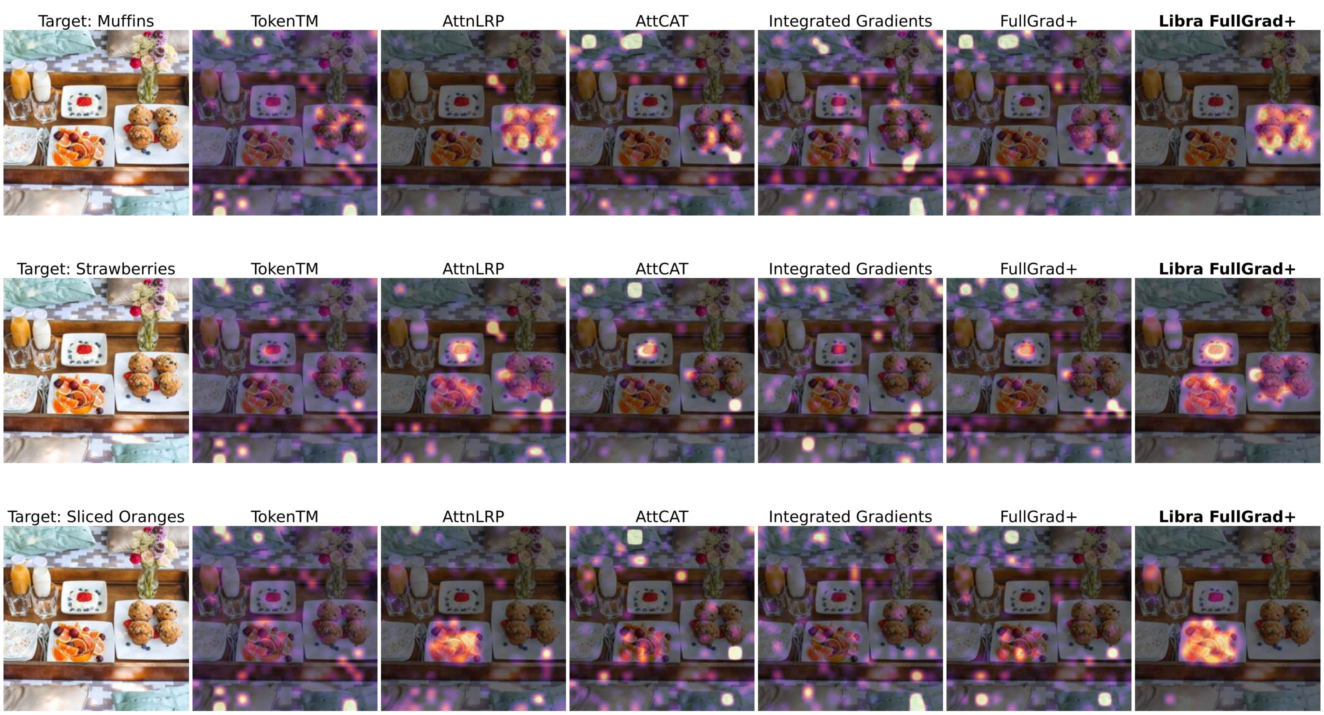
## C.1. Text-Prompted Qualitative Examples on EVA2-CLIP-Large

Our first evaluation scenario uses EVA2-CLIP-Large's text-image similarity scores as attribution targets. For each test image, we systematically probe different regions and concepts using targeted text prompts, enabling a detailed assessment of each attribution method's ability to locate described elements within complex scenes.


GHIJK
LMNOP
QRSTU
VWXYZ

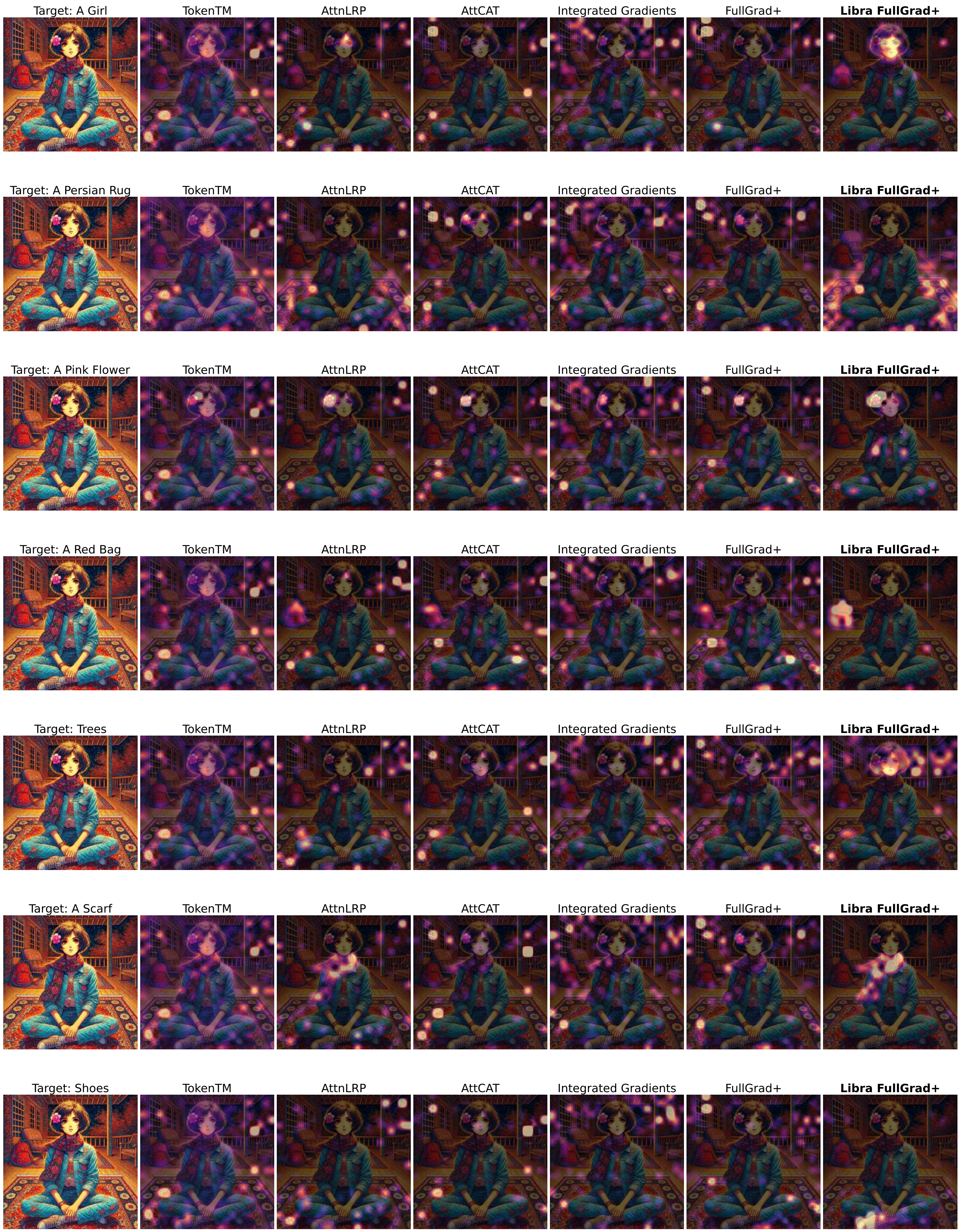

Target: Snowflakes
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Framed Artwork
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: The Alphabet
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**


Target: Purple Tops
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Blue Pants
Target: Baskets
Target: Raining Cloud
Target: Dollhouse

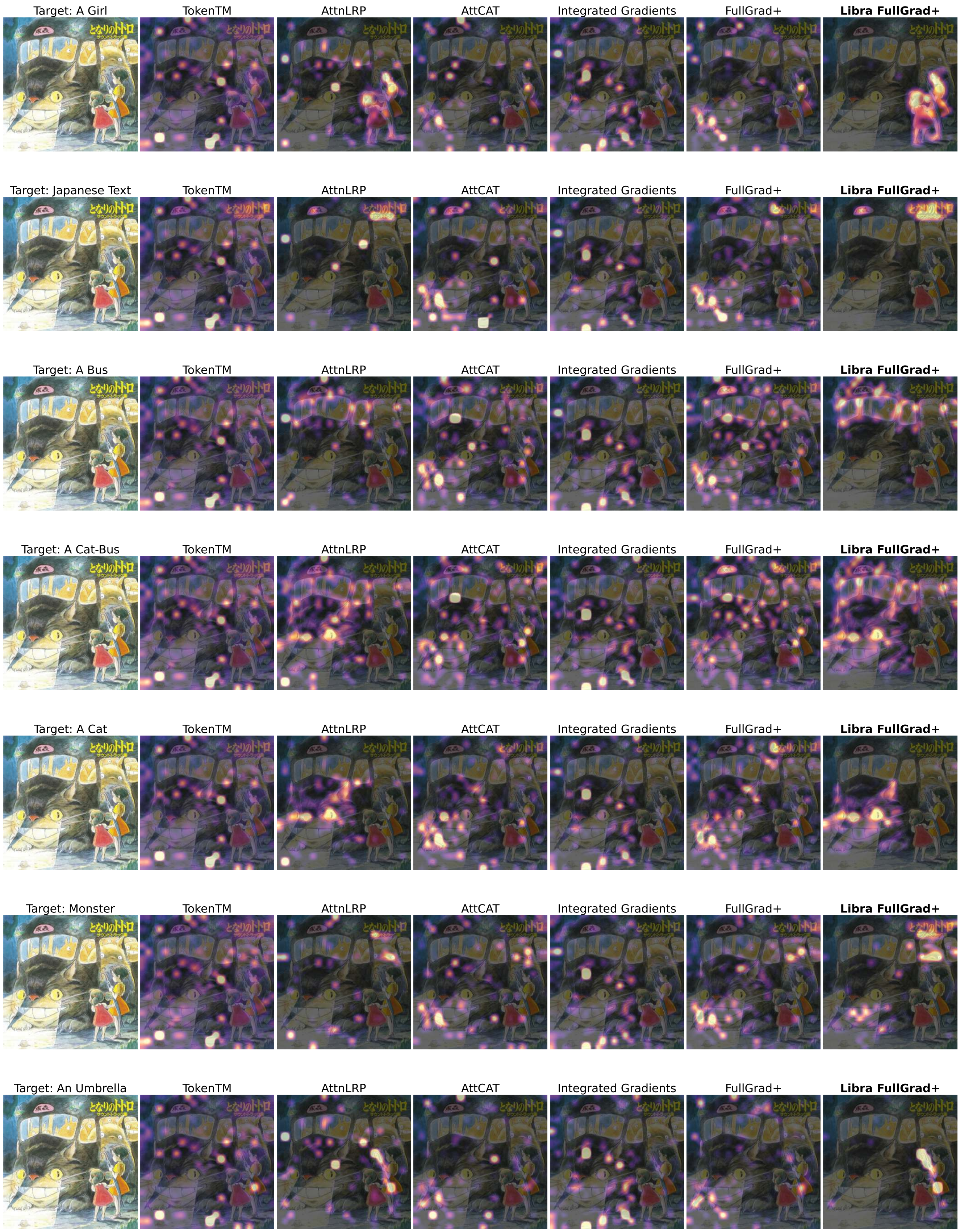

Target: Muffins
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Strawberries
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Sliced Oranges
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+


Target: Glasses
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Wooden Tray
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target:
Patterned Blanket
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Bottles
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Cream
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Flowers
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target:
Spoons and Forks
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+

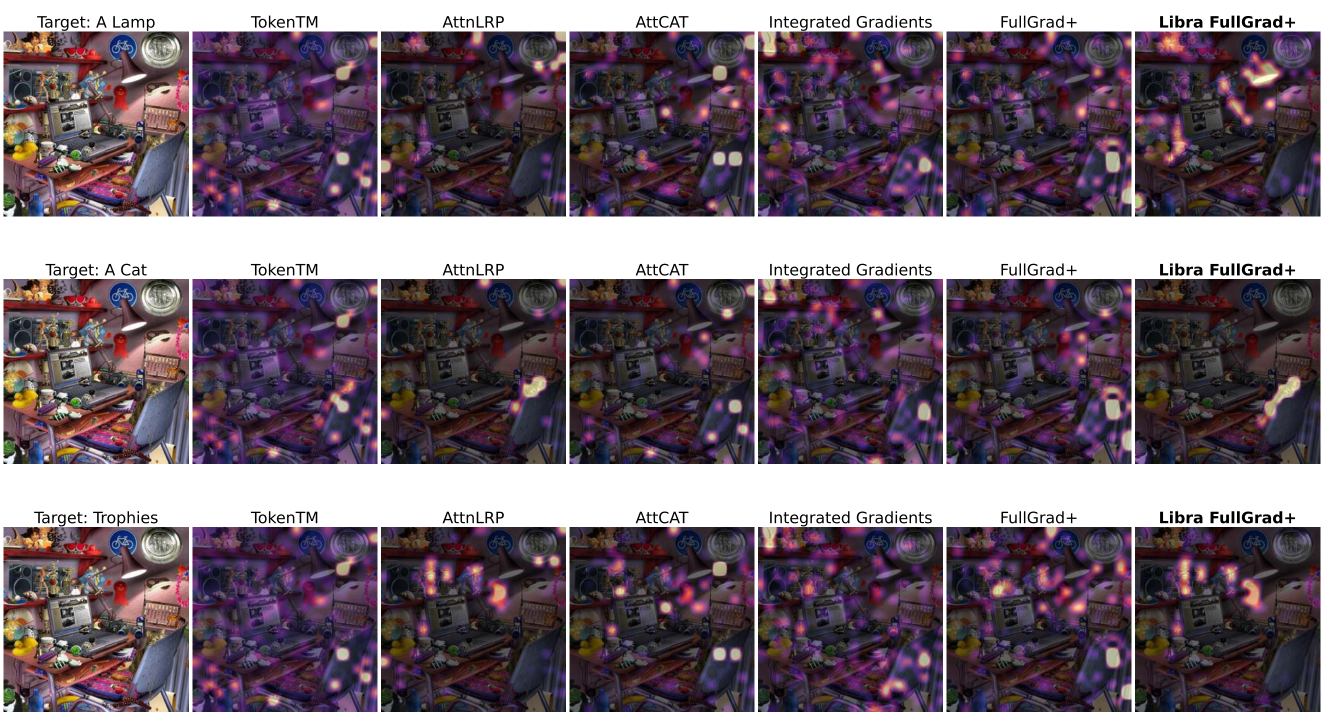
Target: A Girl
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Persian Rug
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Bag
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Pink Flower
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Scarf
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Shoes
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Trees
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+

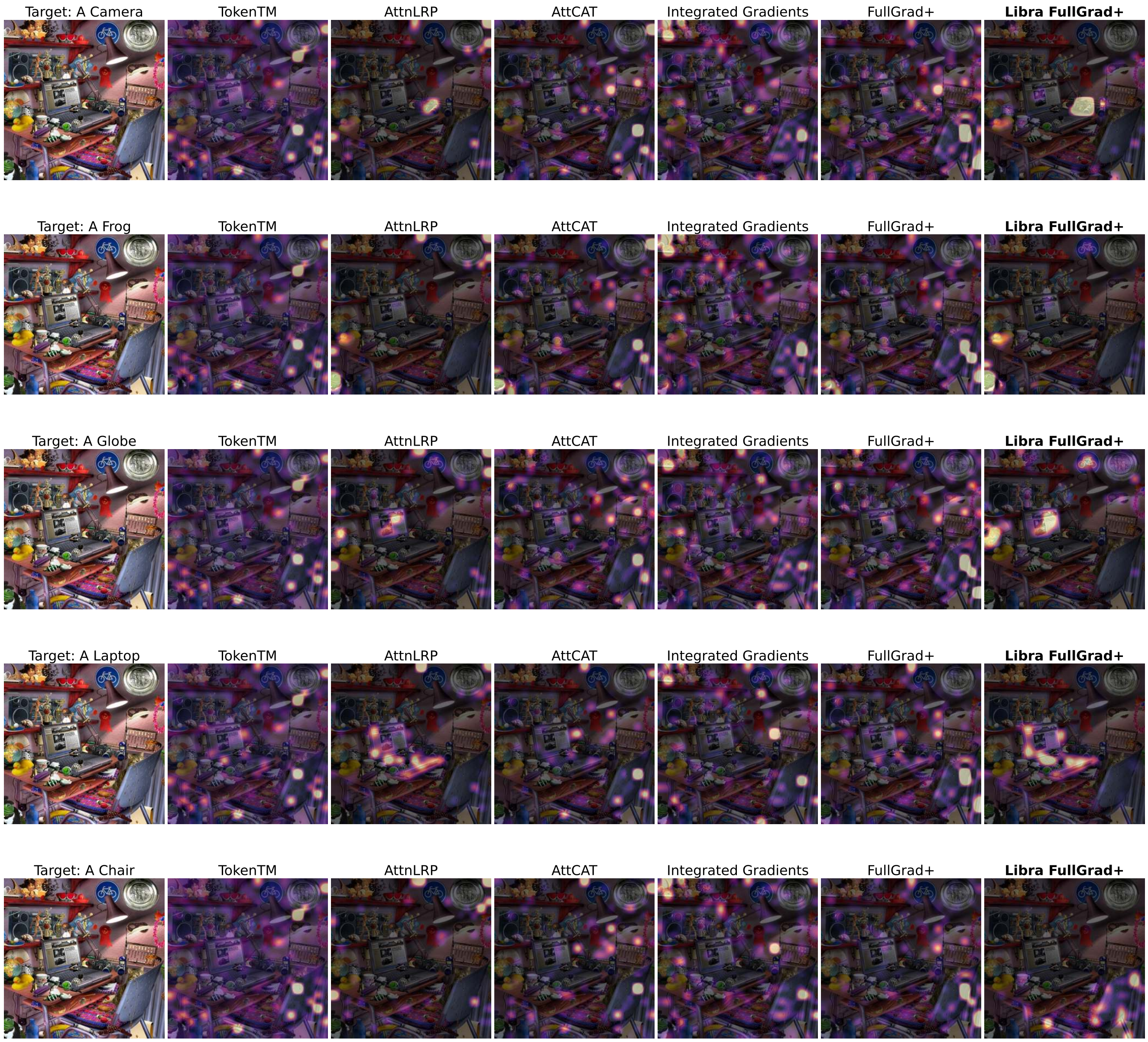


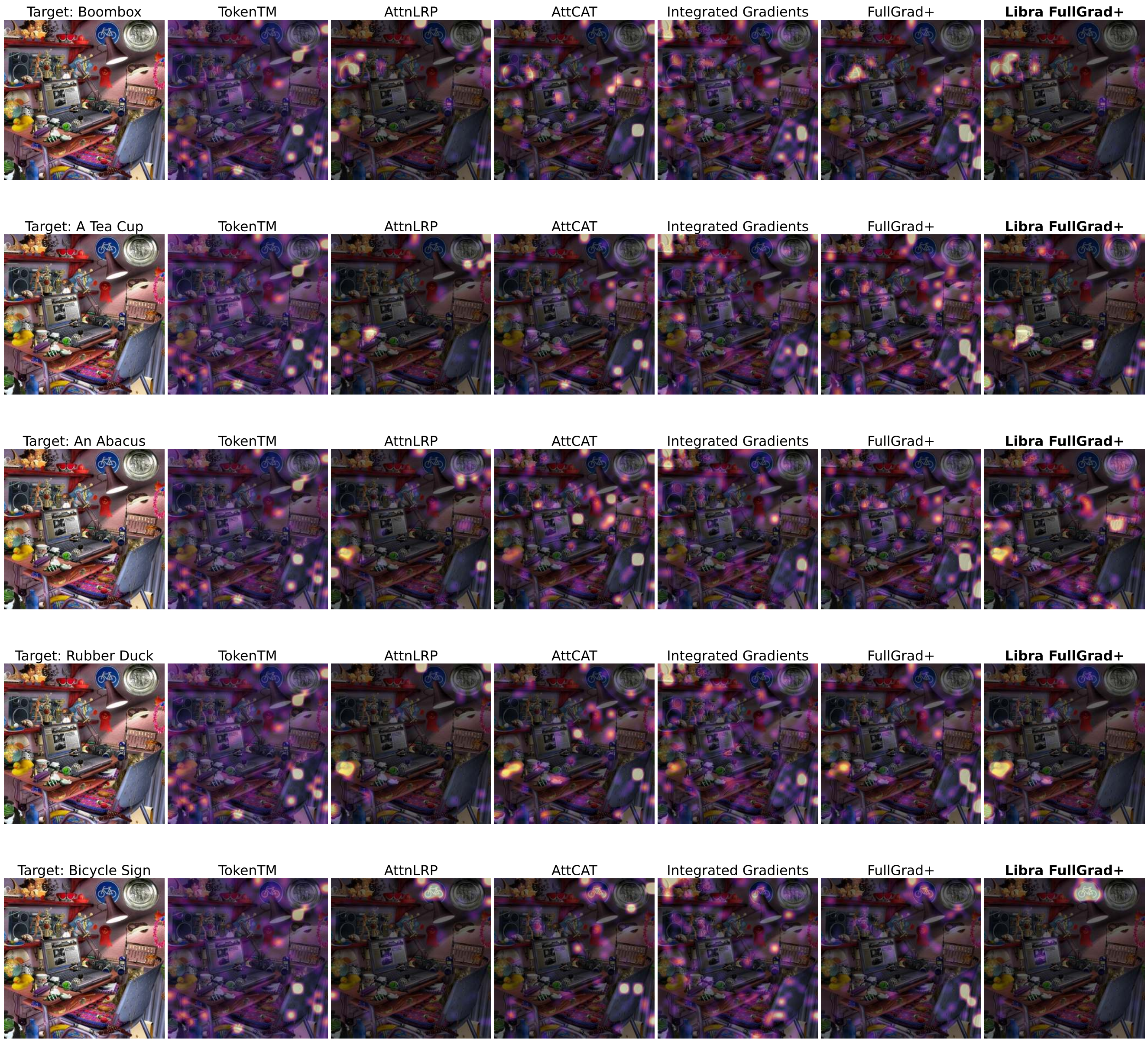
Target: A Girl
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Persian Rug
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Pink Flower
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Red Bag
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Trees
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Scarf
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Shoes
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+


Target: A Girl
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Japanese Text
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Bus
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Cat-Bus
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Cat
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Monster
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: An Umbrella
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**


Town Hall Archive
Local Police At Plainfield Investigation


Target: A Lamp
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Cat
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Trophies
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**

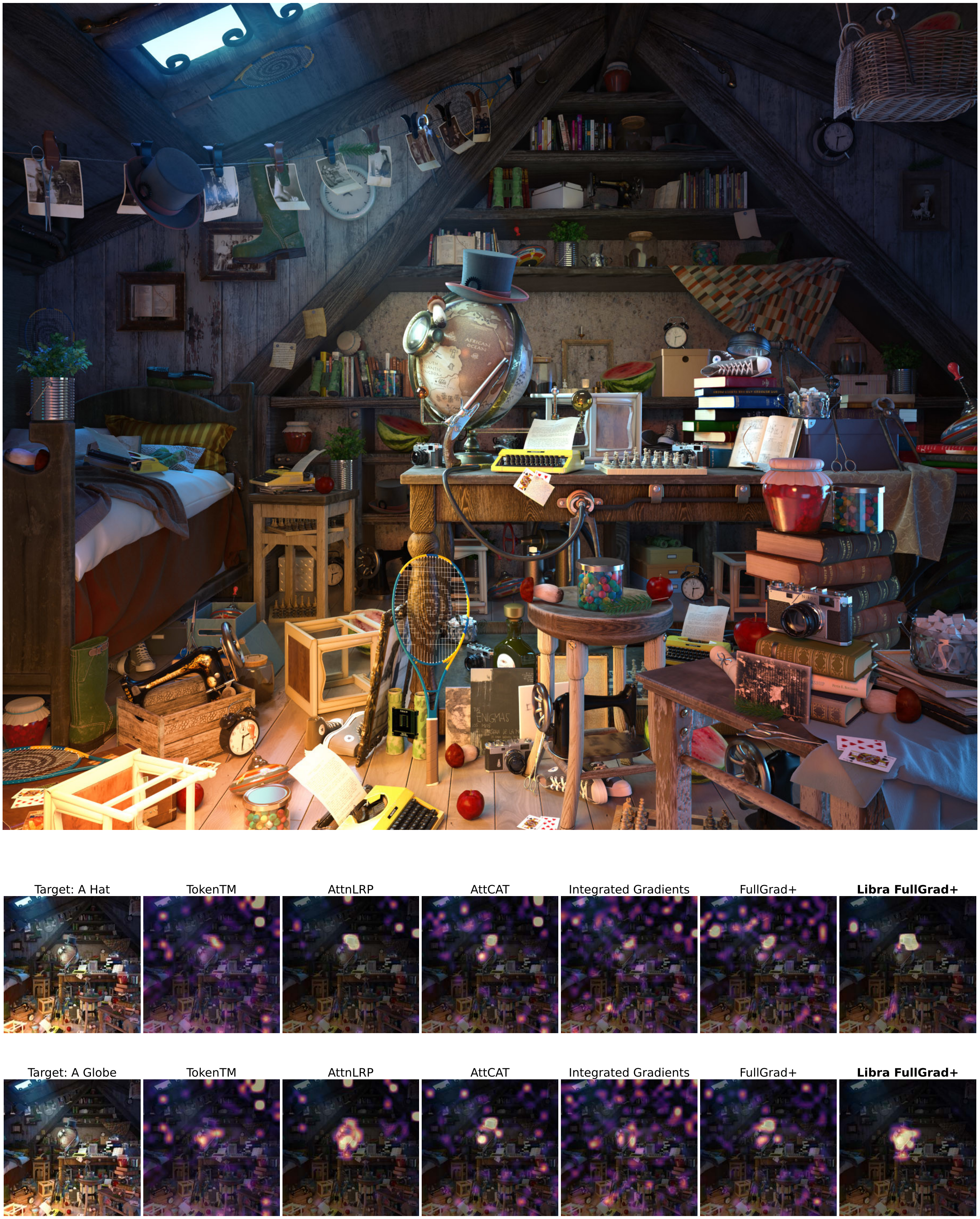

Target: A Camera
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Frog
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Globe
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Laptop
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Chair
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+

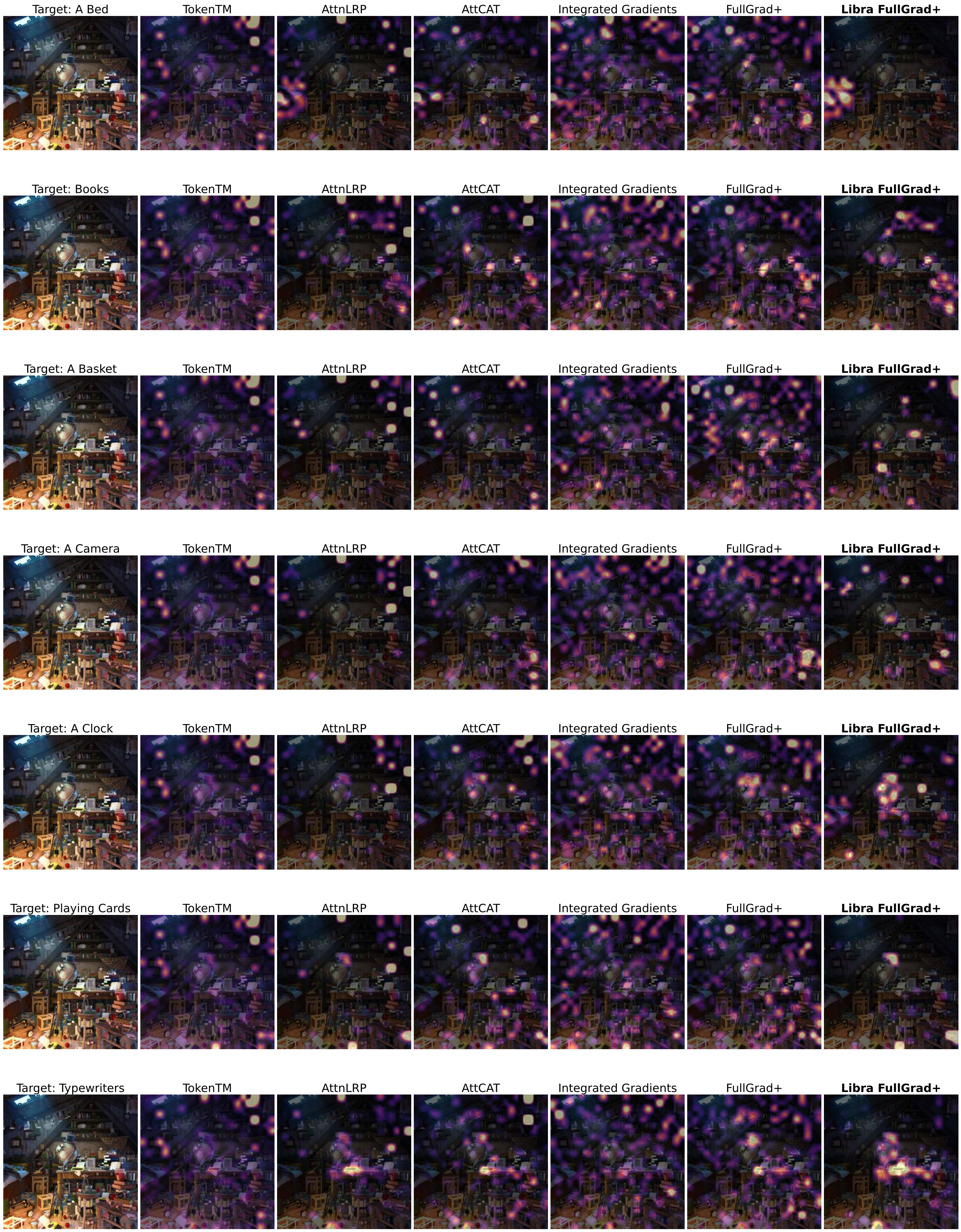

Target: Boombox
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Tea Cup
Target: An Abacus
Target: Rubber Duck
Target: Bicycle Sign


Target: A Unicycle | TokenTM | AttnLRP | AttCAT | Integrated Gradients | FullGrad+ | **Libra FullGrad+**

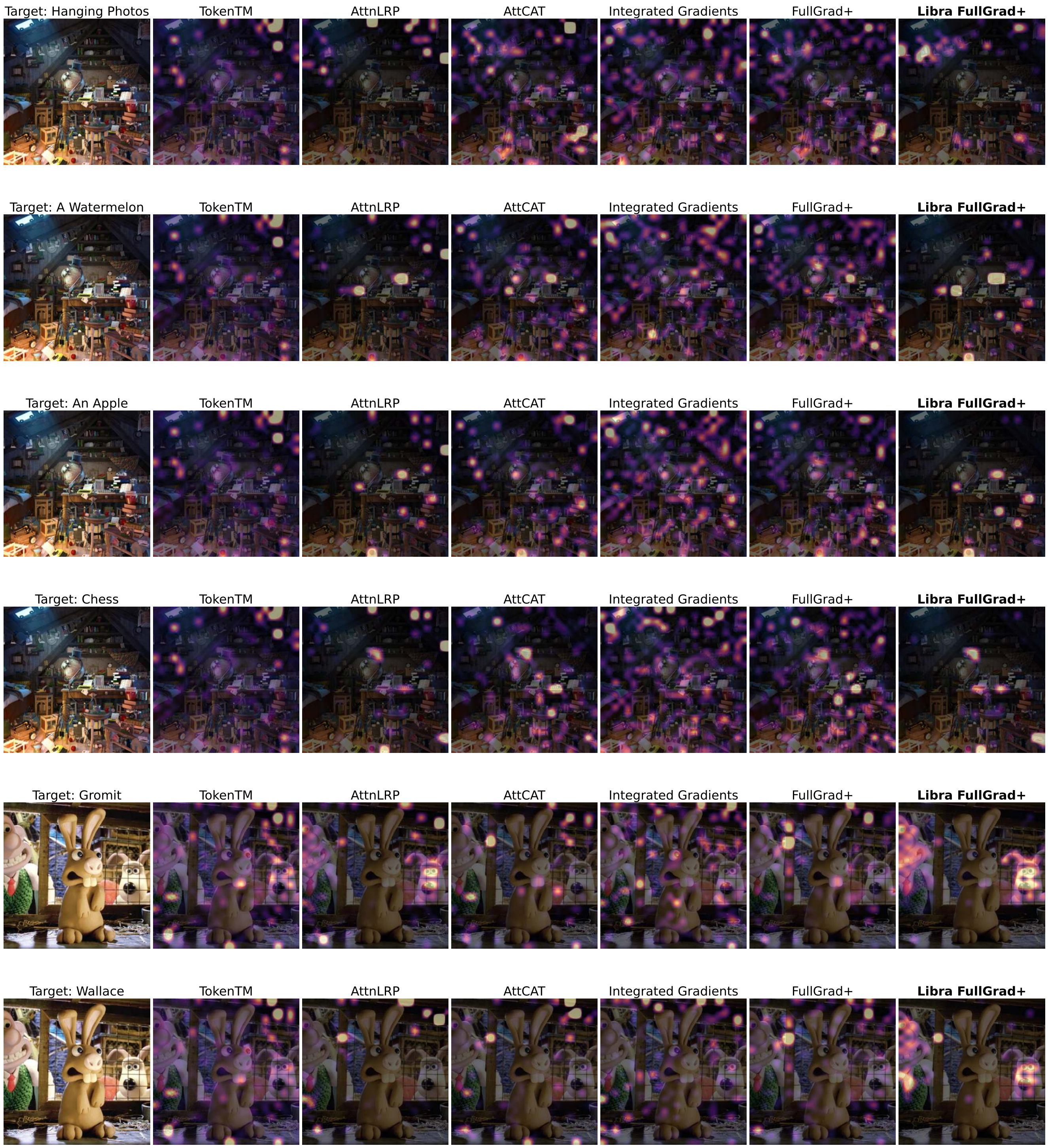
Target: A Cat
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Stefan
Target: A Clock
Target: A Dog
Target: A Door
Target: A Fire Hydrant

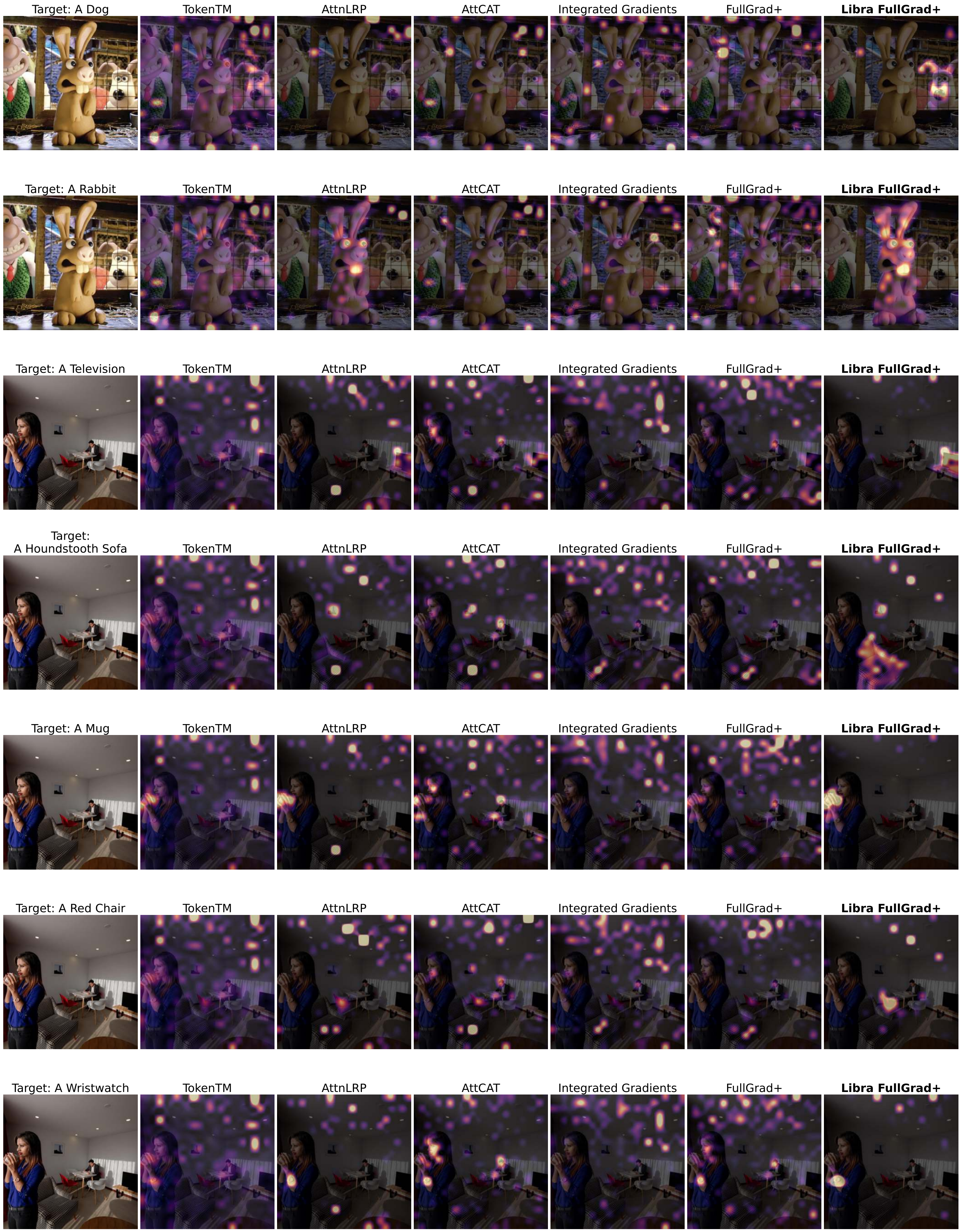

Target: A Gnome
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Soccer Ball
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Trees
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Flowers
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Horseshoe
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Goose
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+


Target: A Hat
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Globe
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**

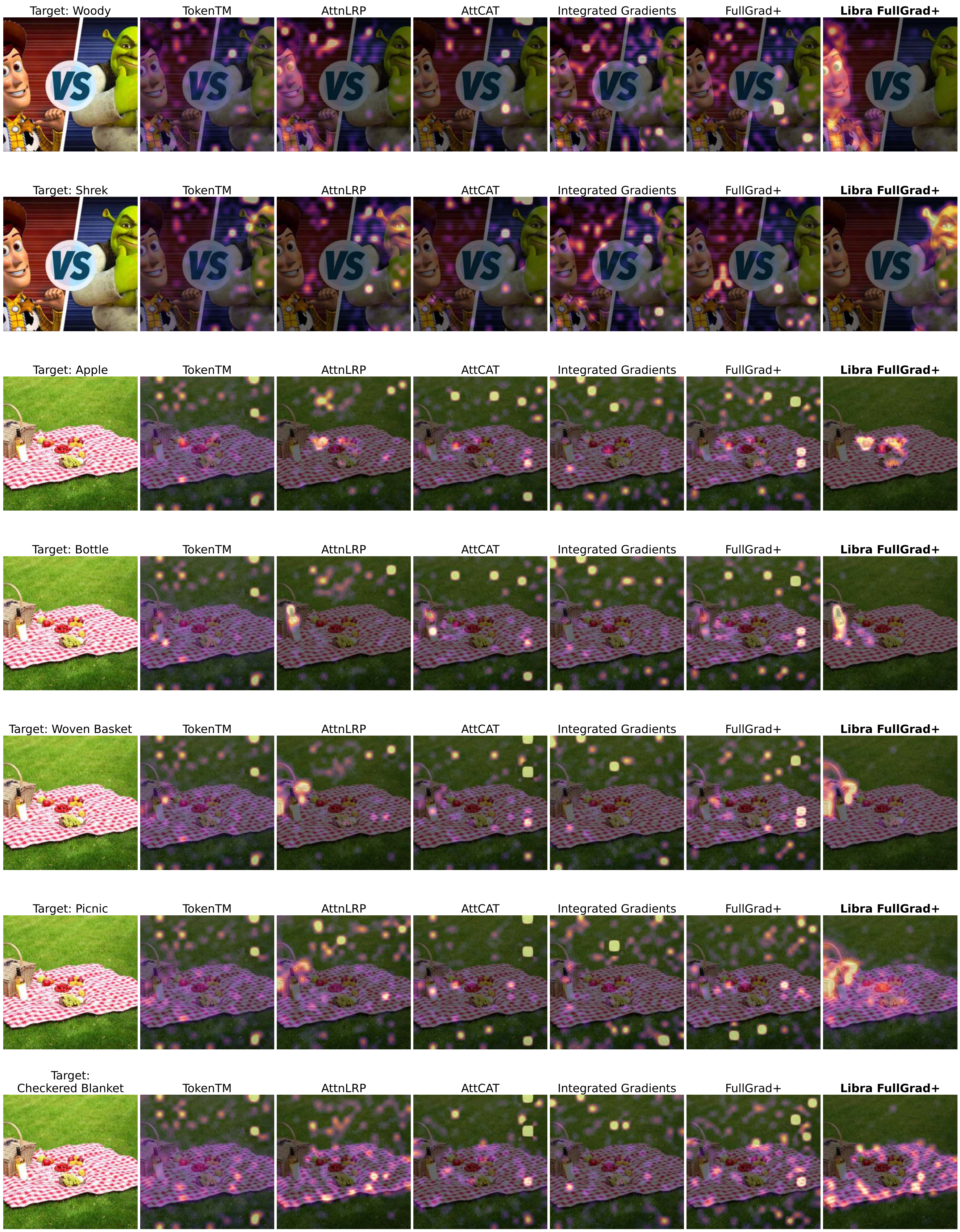


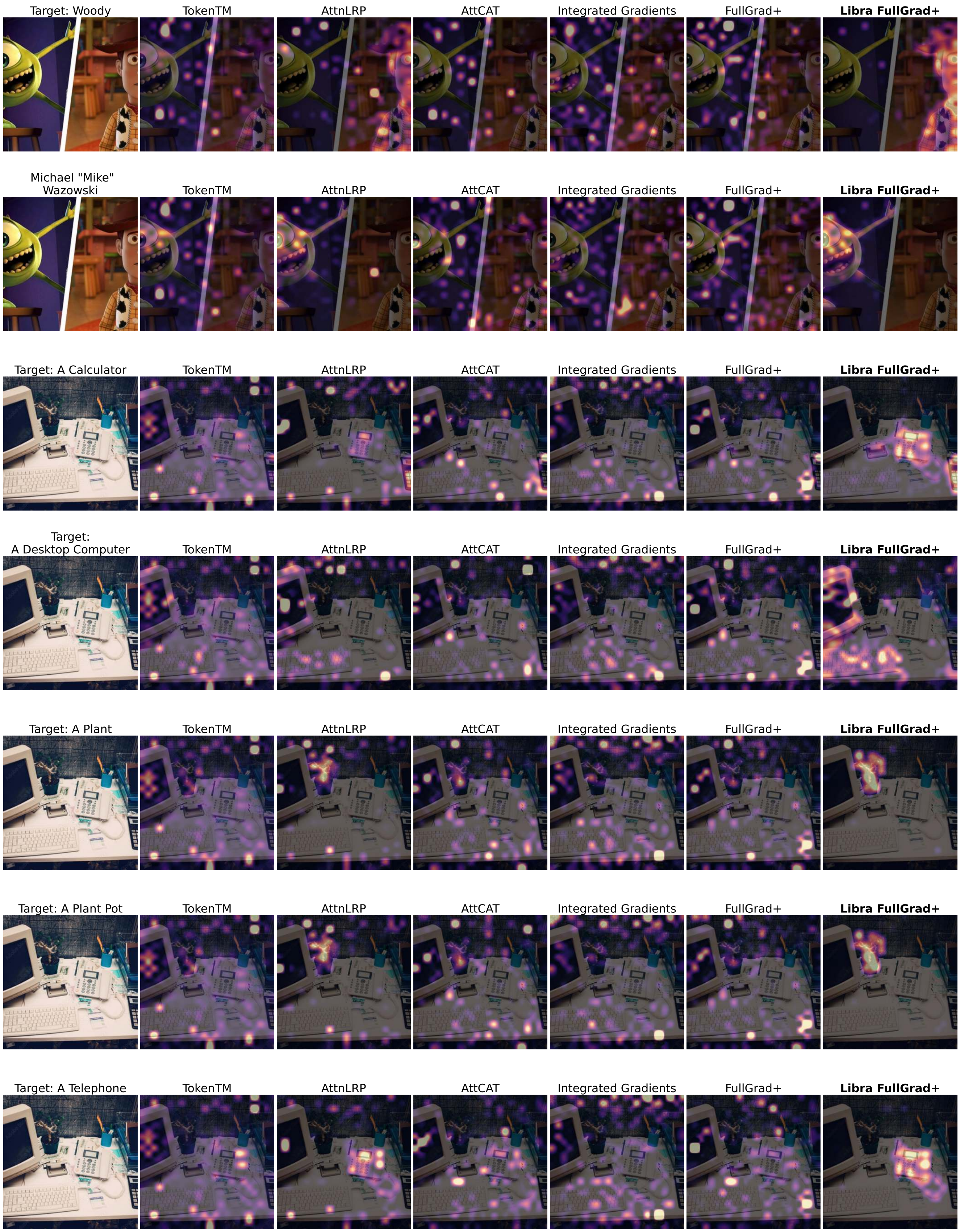
Target: A Bed
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Books
Target: A Basket
Target: A Camera
Target: A Clock
Target: Playing Cards
Target: Typewriters

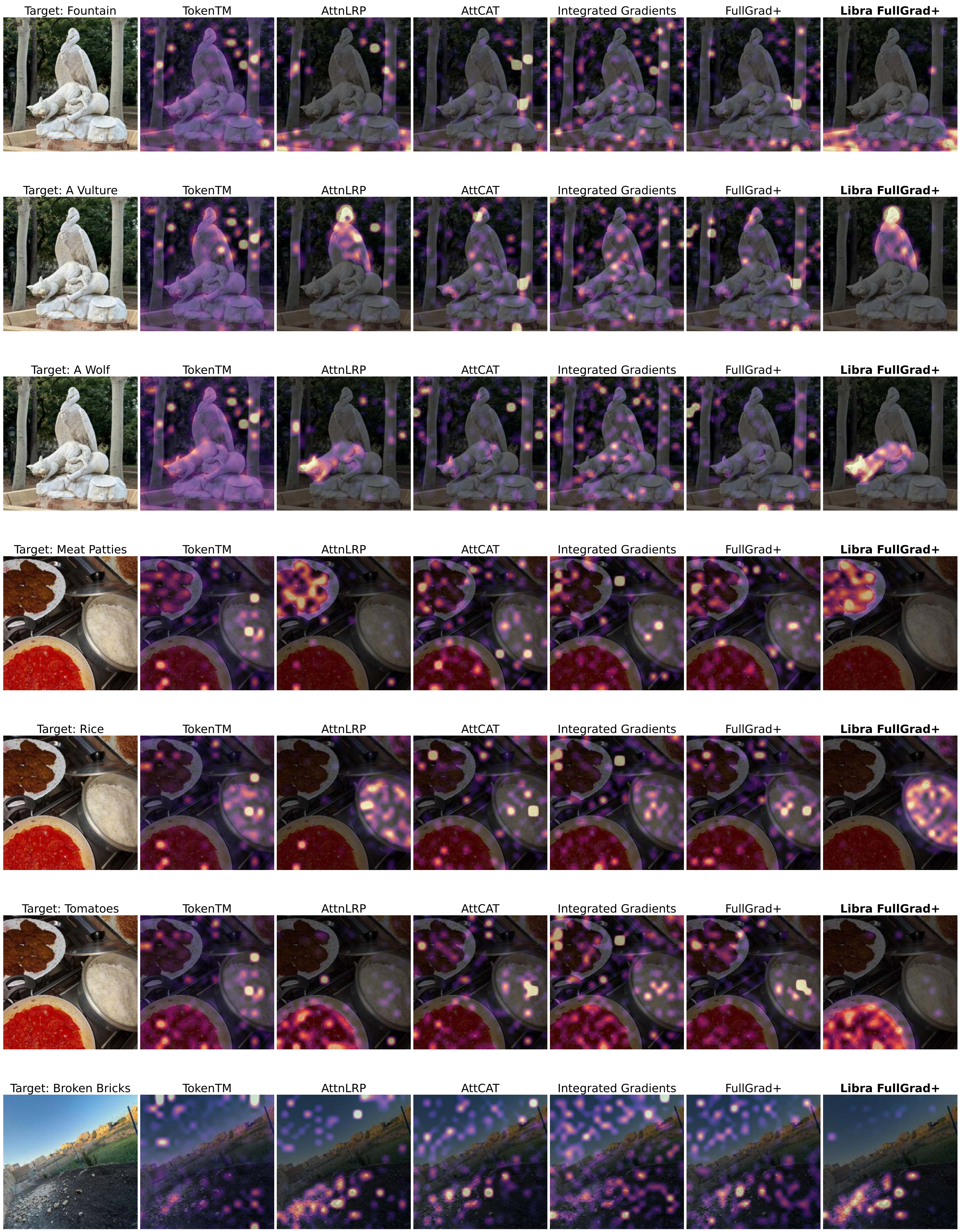

Target: Hanging Photos
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Watermelon
Target: An Apple
Target: Chess
Target: Gromit
Target: Wallace


Target: A Dog
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Rabbit
Target: A Television
Target:
A Houndstooth Sofa
Target: A Mug
Target: A Red Chair
Target: A Wristwatch


Target: A Bed
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Bathtub
Target: A Girl
Target: A Laptop
Target: A Tray of Food
Target: A Mirror


Target: Woody
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
VS
Target: Shrek
Target: Apple
Target: Bottle
Target: Woven Basket
Target: Picnic
Target: Checkered Blanket

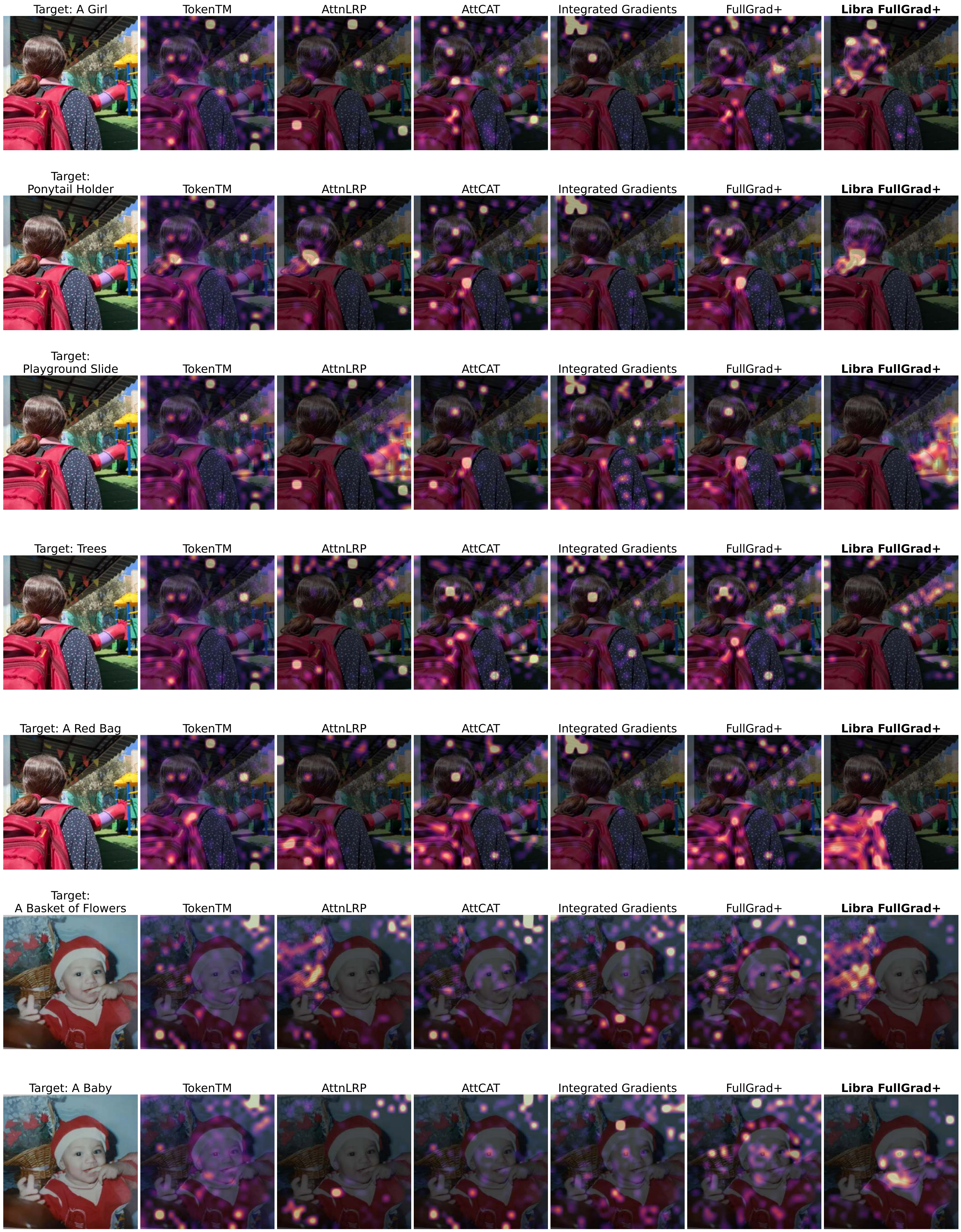

Target: Woody
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Michael "Mike" Wazowski
Target: A Calculator
Target: A Desktop Computer
Target: A Plant
Target: A Plant Pot
Target: A Telephone


Target: Fountain
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Vulture
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Wolf
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Meat Patties
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Rice
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Tomatoes
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Broken Bricks
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**

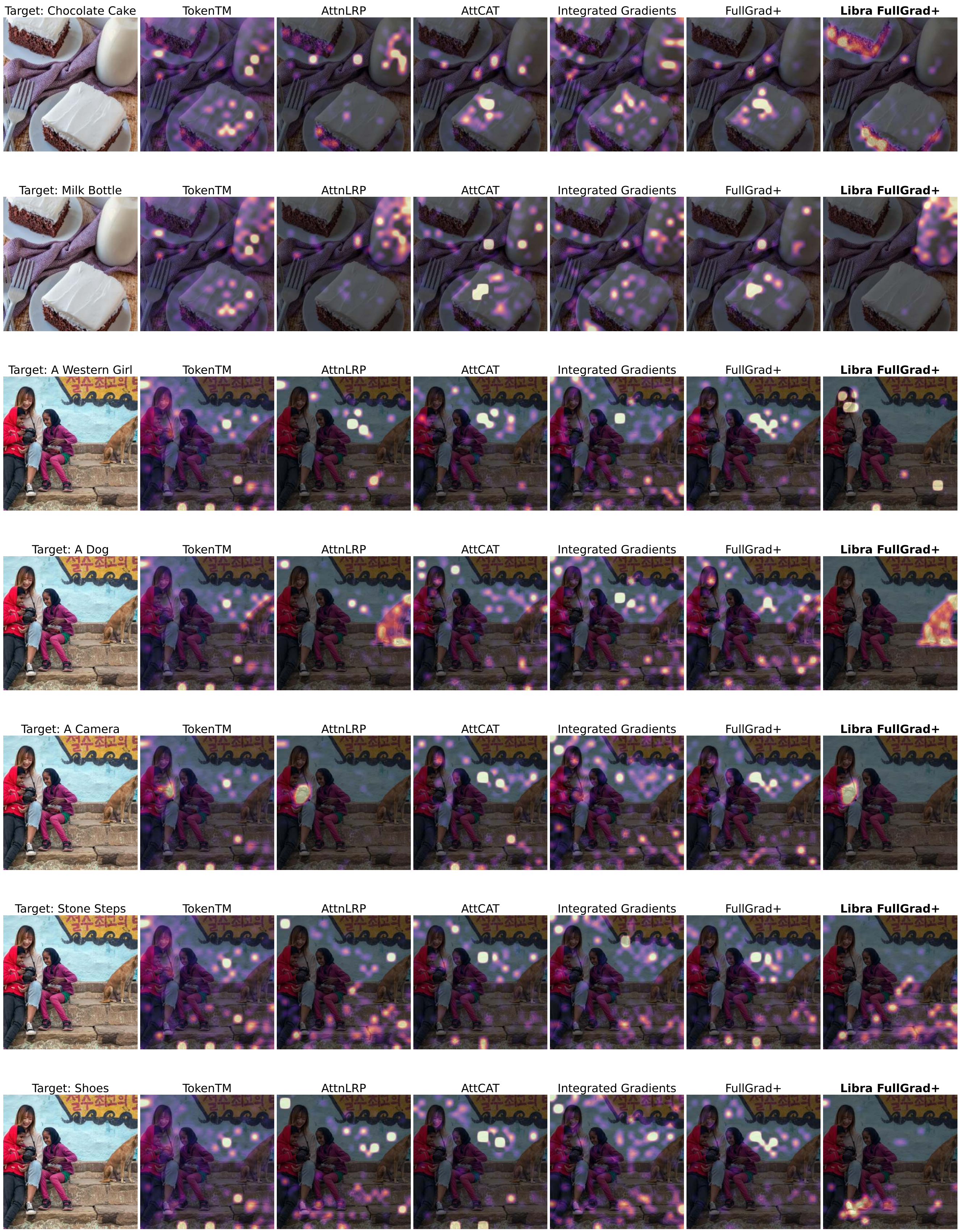

Target: Shoes
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Purple Socks
Target: Grass
Target: The Sky
Target: Press
Target: Start
Target: Press Start


Target: Game Boy Color
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Buzz Lightyear
Target: Woody
Target:
Mr. Potato Head
Target: Buzz Lightyear
Target: Woody
Target: Bo Peep

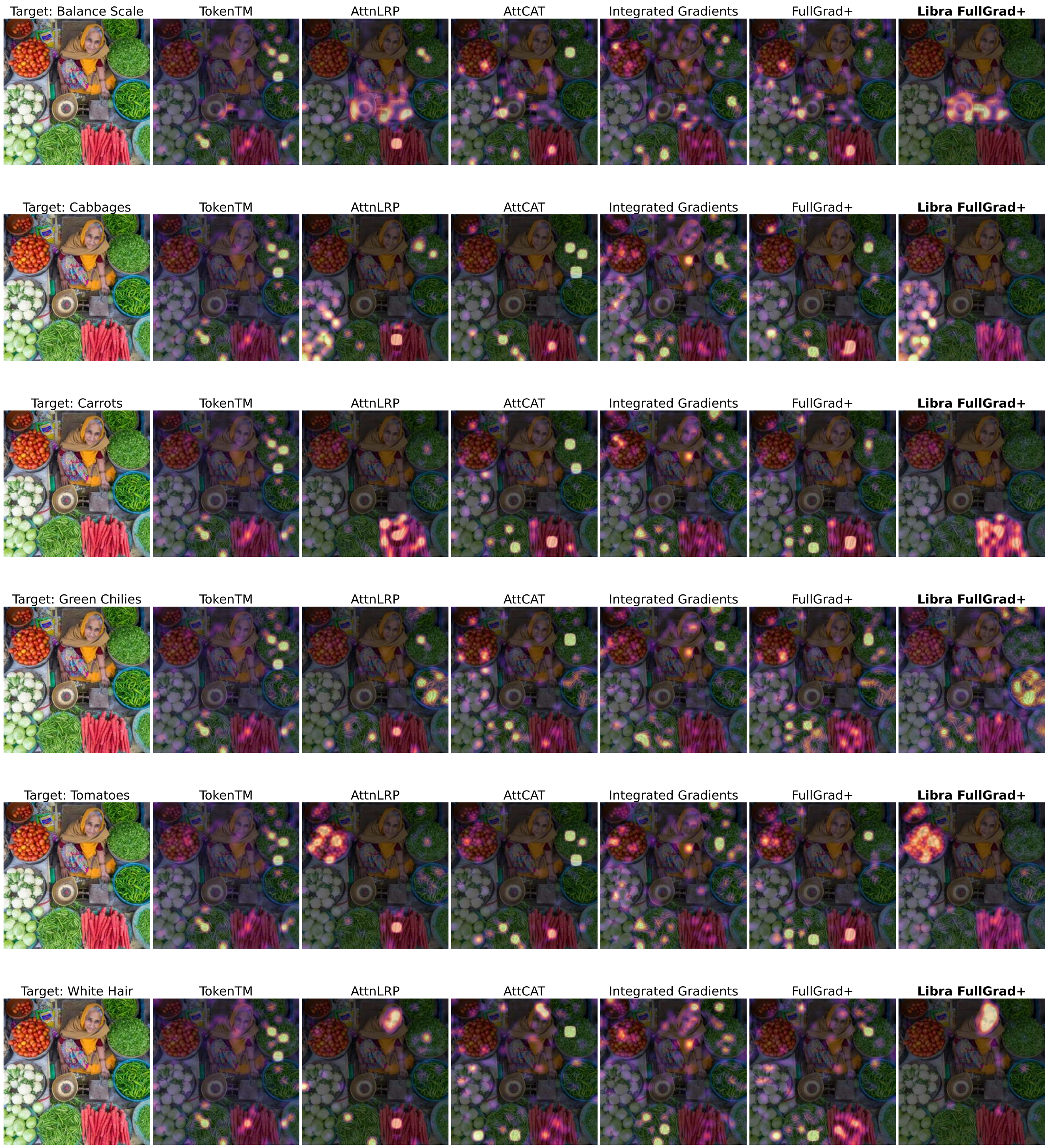

Target: Nemo
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target:
Lightning McQueen
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Michael "Mike"
Wazowski
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Mr. Incredible
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Wall-E EVE
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Woody
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+

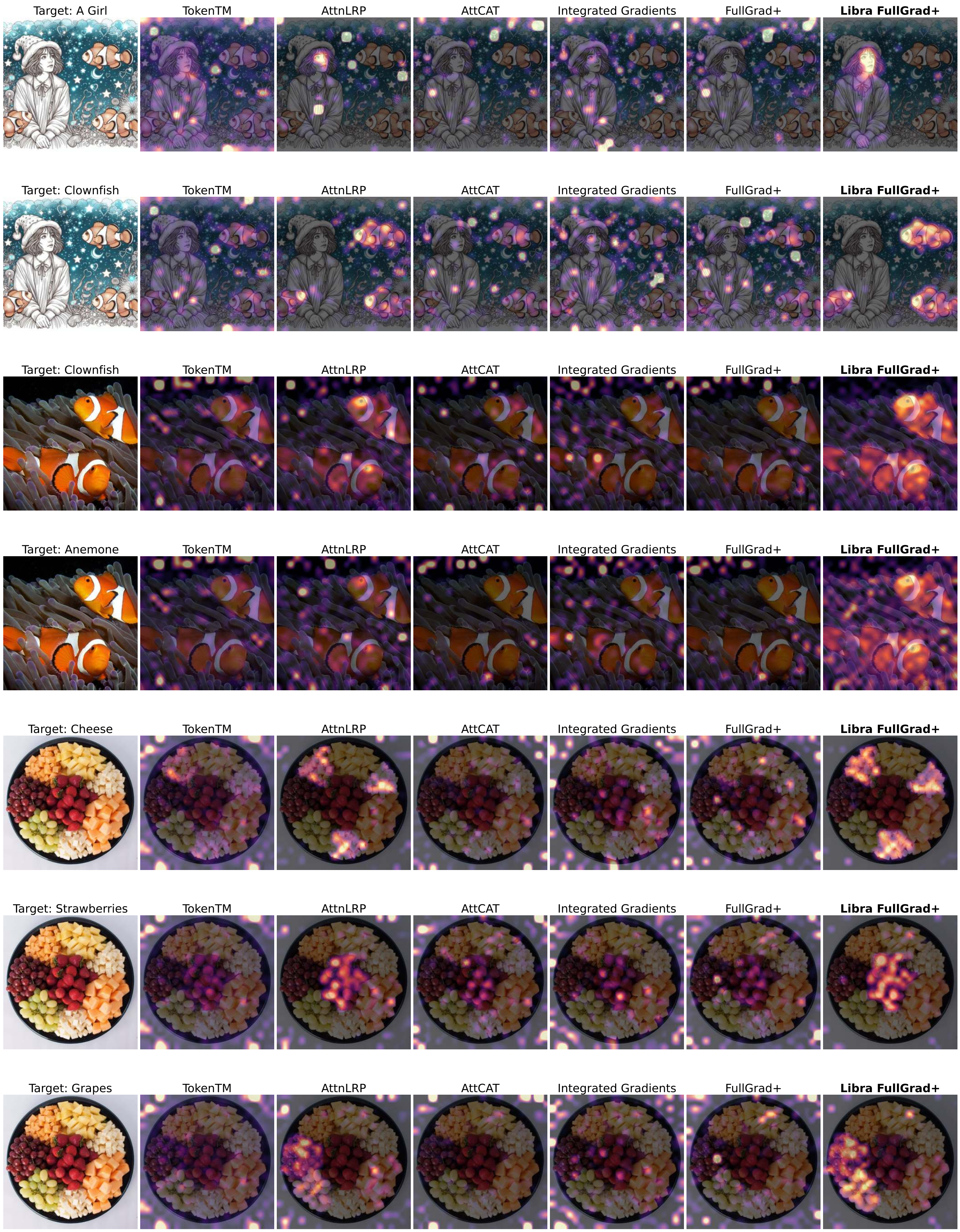


Target: A Girl
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Ponytail Holder
Target: Playground Slide
Target: Trees
Target: A Red Bag
Target: A Basket of Flowers
Target: A Baby

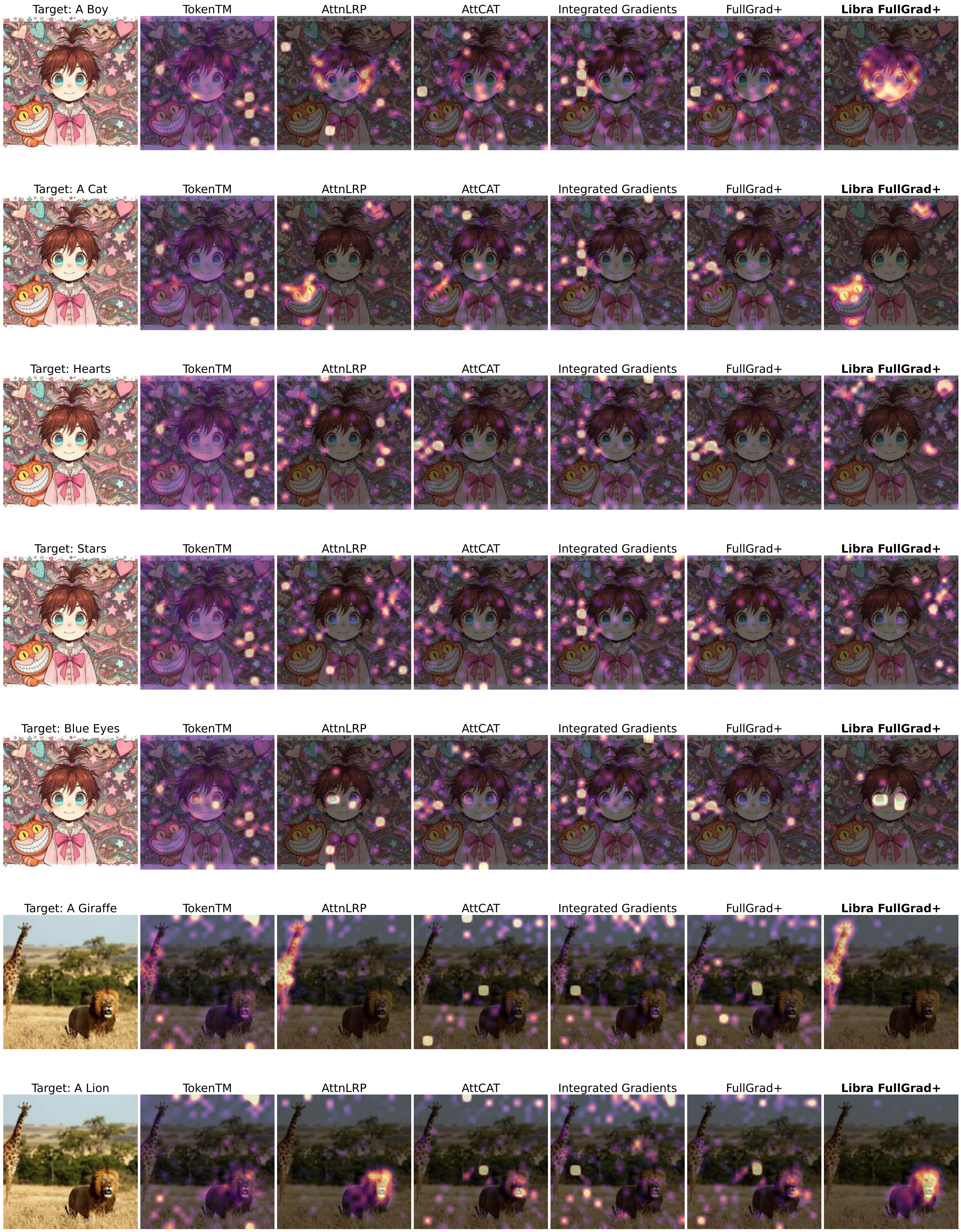

Target: Chocolate Cake
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Milk Bottle
Target: A Western Girl
Target: A Dog
Target: A Camera
Target: Stone Steps
Target: Shoes

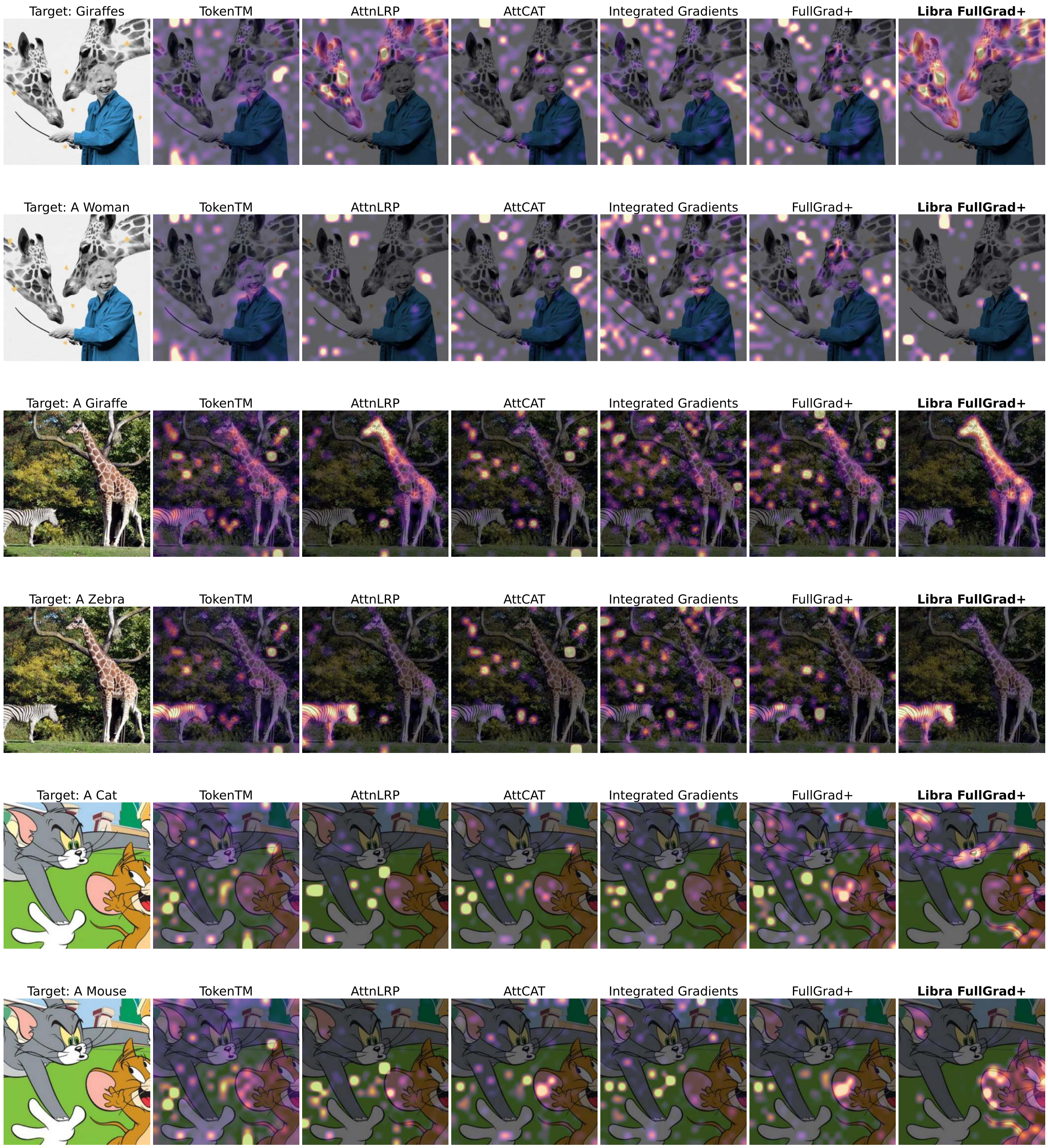

Target:
Indian Clothing
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Horse
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Warrior
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Green Kameez
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Potted Plants
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Sandals
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target:
Woman Wearing Pink
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+

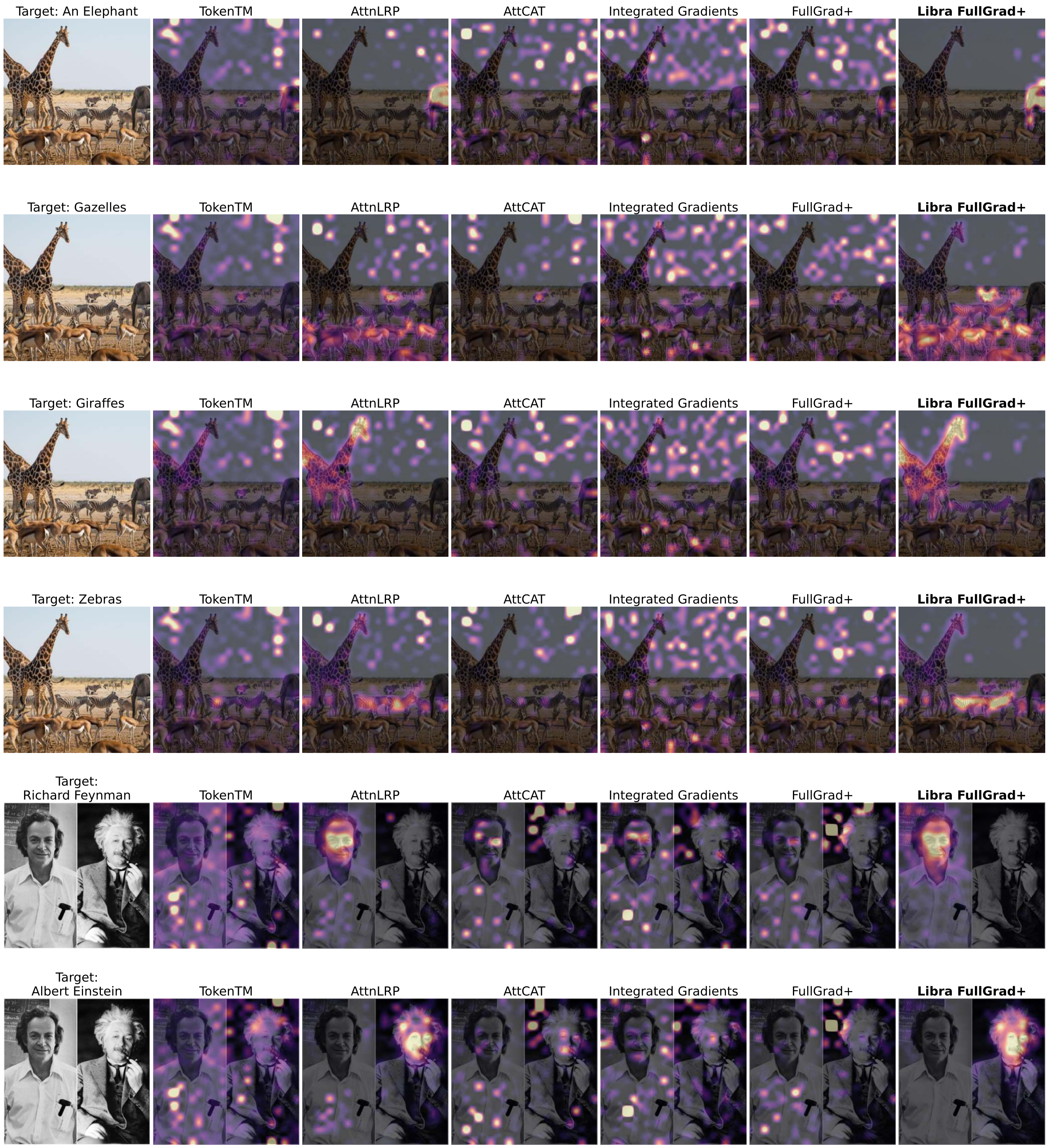

Target: Balance Scale
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Cabbages
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Carrots
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Green Chilies
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Tomatoes
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: White Hair
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+



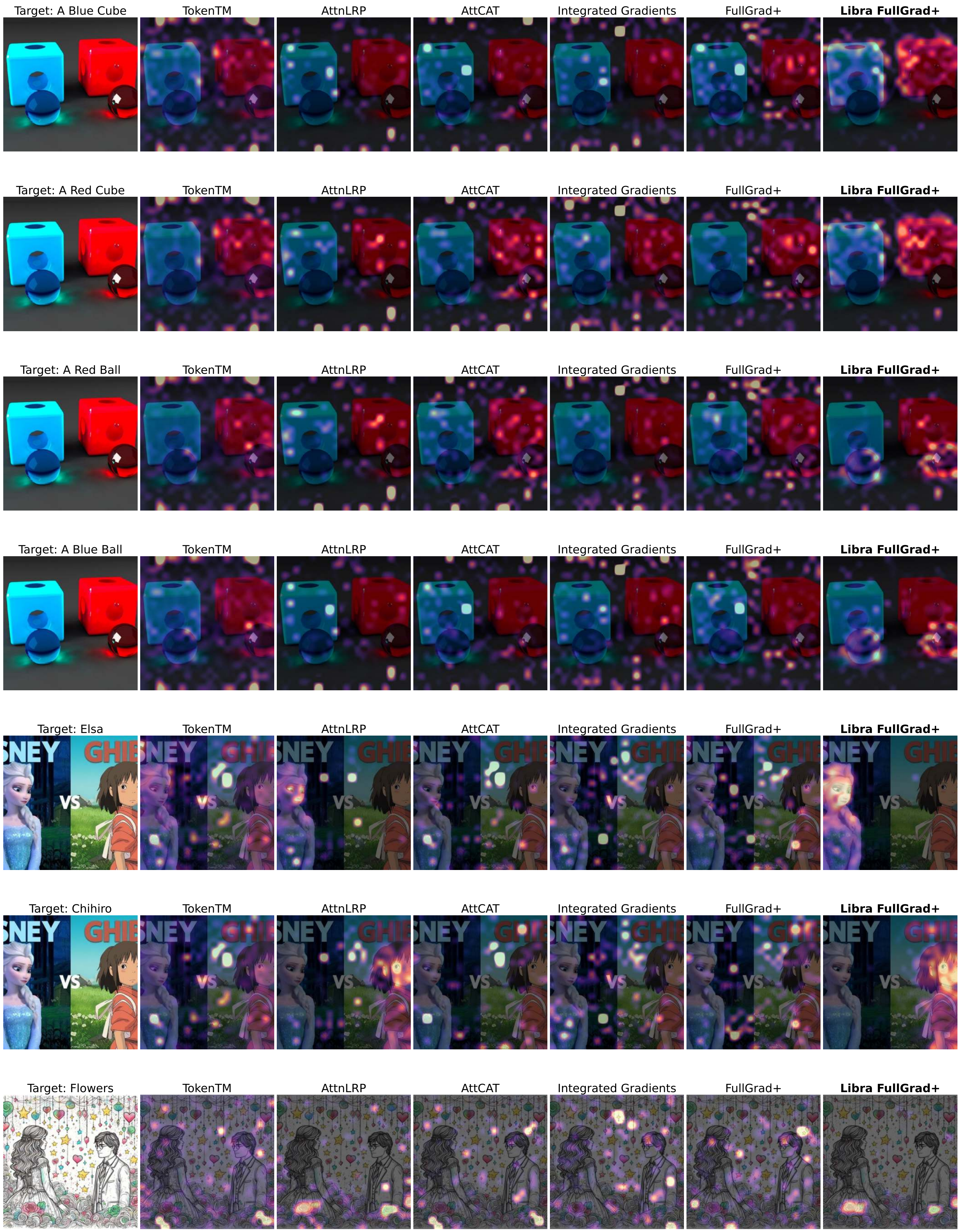
Target: A Girl
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Clownfish
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Clownfish
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Anemone
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Cheese
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Strawberries
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Grapes
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+

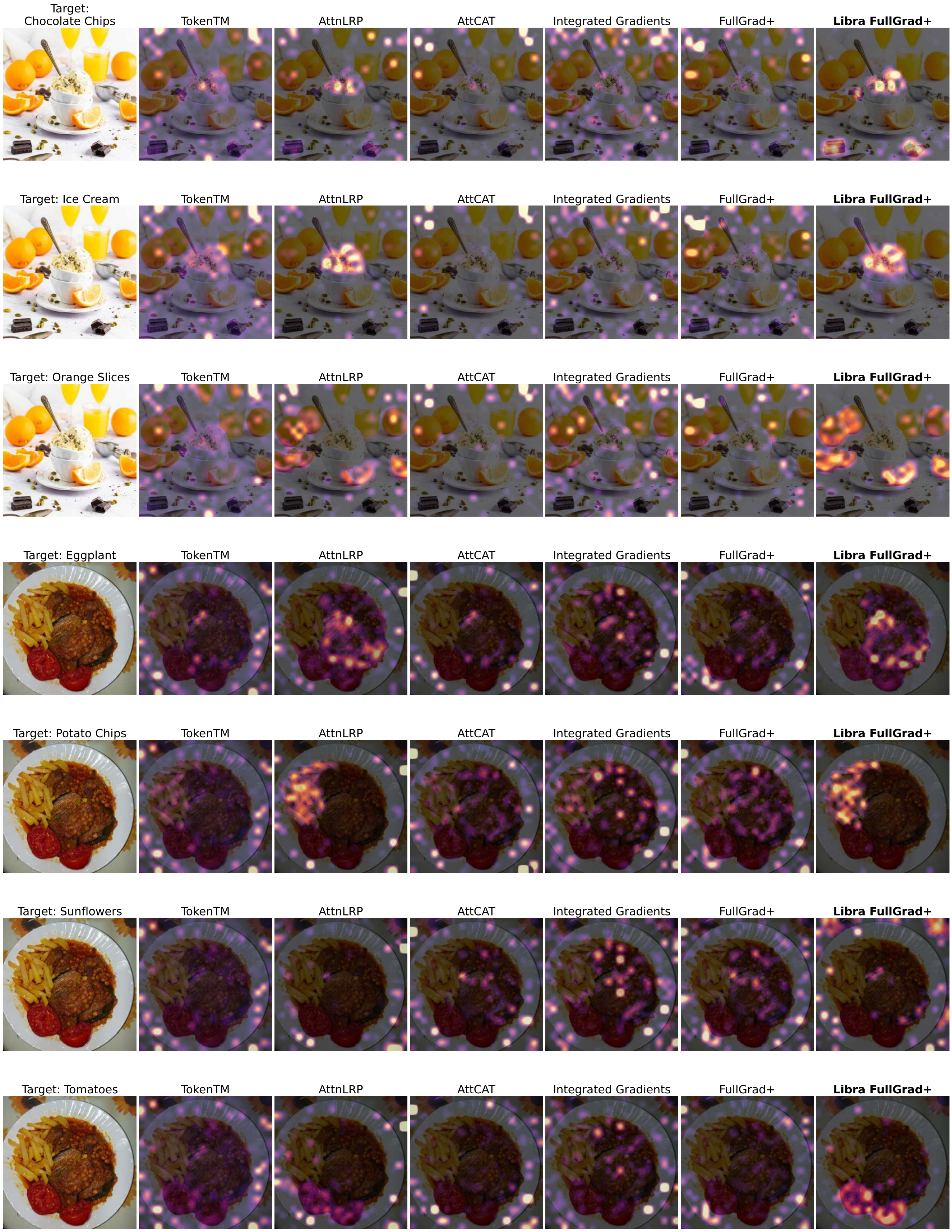

Target: A Boy
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Cat
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Hearts
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Stars
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Blue Eyes
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Giraffe
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Lion
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+

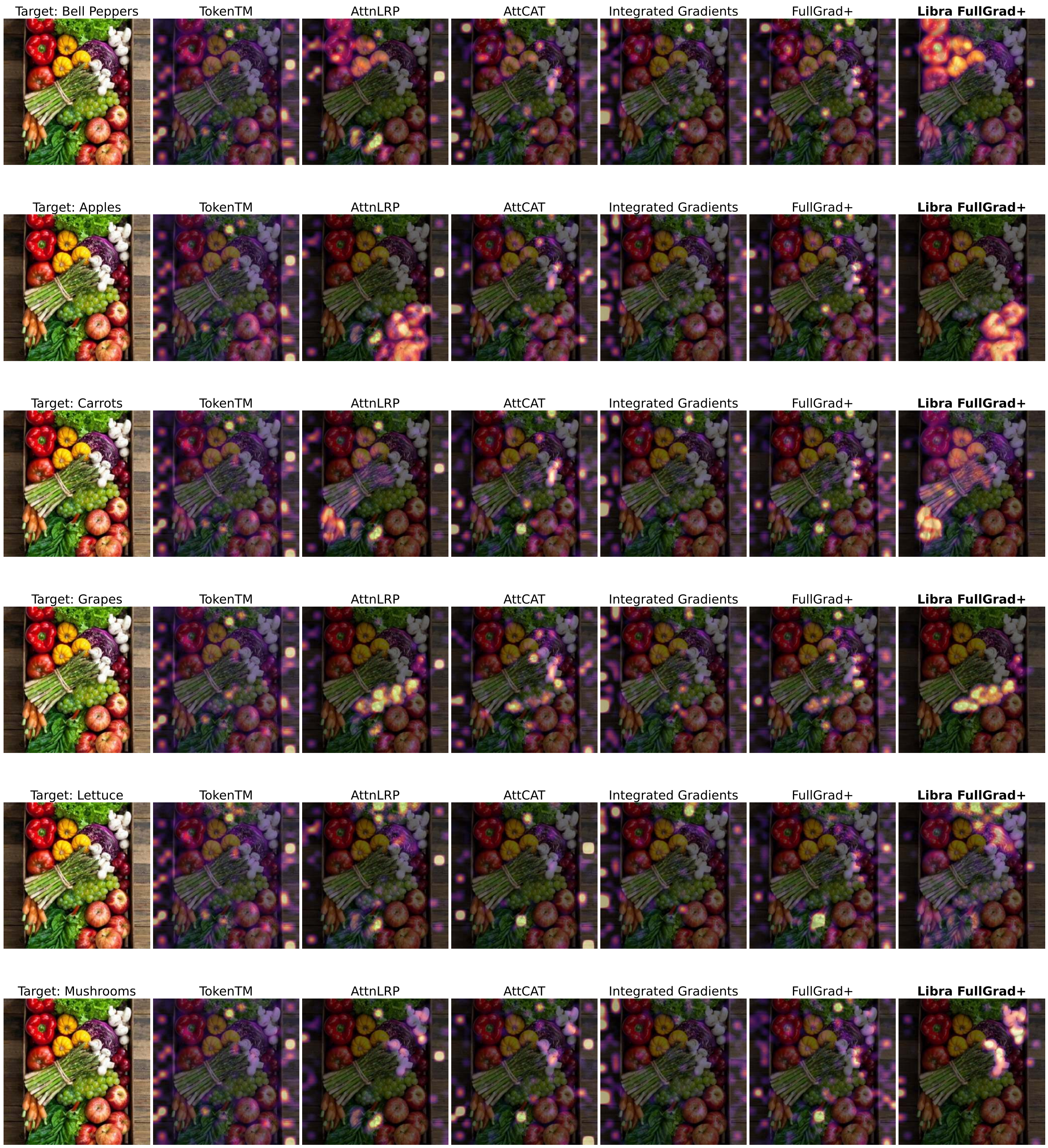

Target: Giraffes
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Woman
Target: A Giraffe
Target: A Zebra
Target: A Cat
Target: A Mouse


Target: An Elephant
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Gazelles
Target: Giraffes
Target: Zebras
Target: Richard Feynman
Target: Albert Einstein

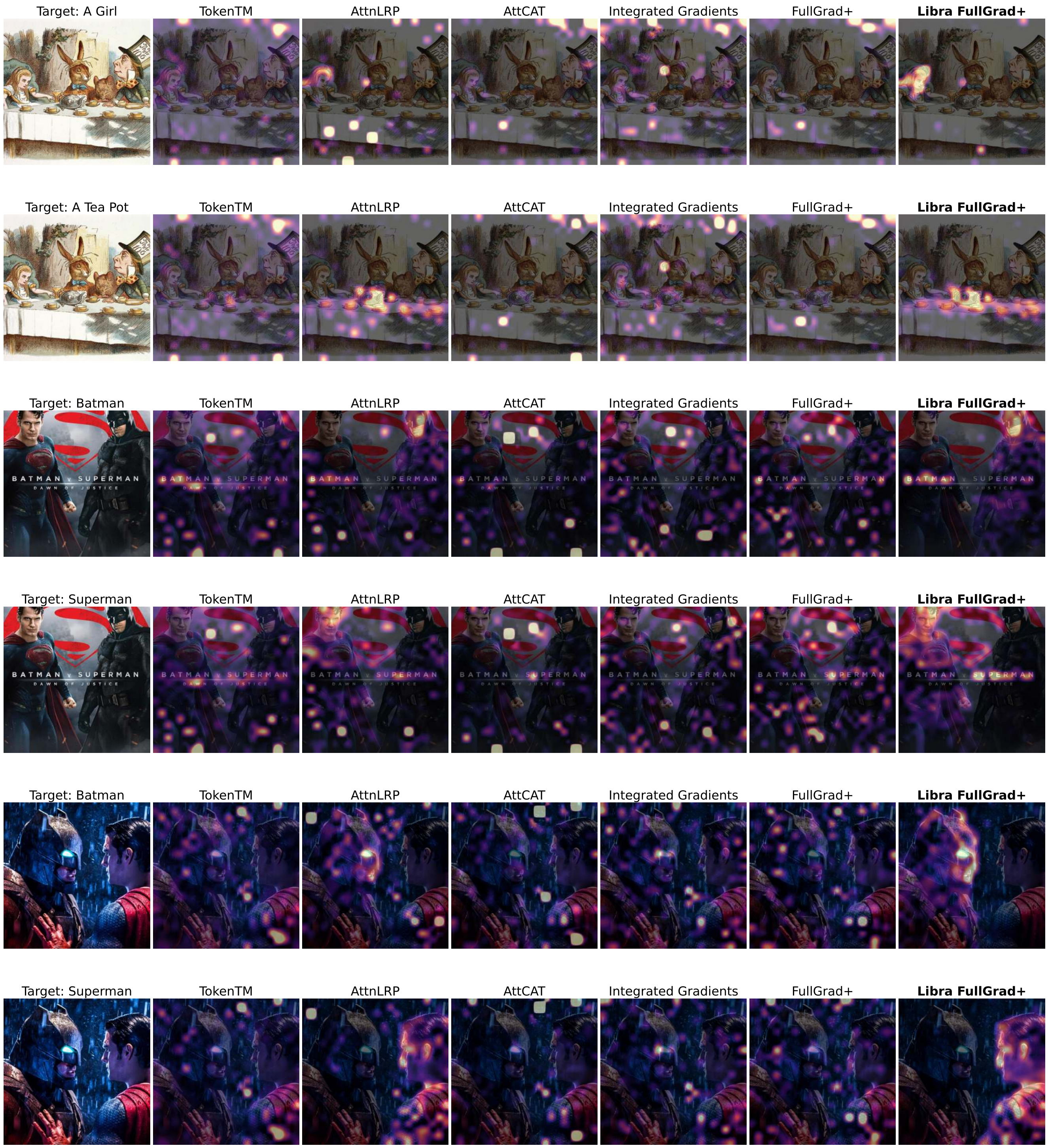

Target:
A Blue Butterfly
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Pink Flower
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Cat
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Fruit
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Dog
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Man
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: An Airplane
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**

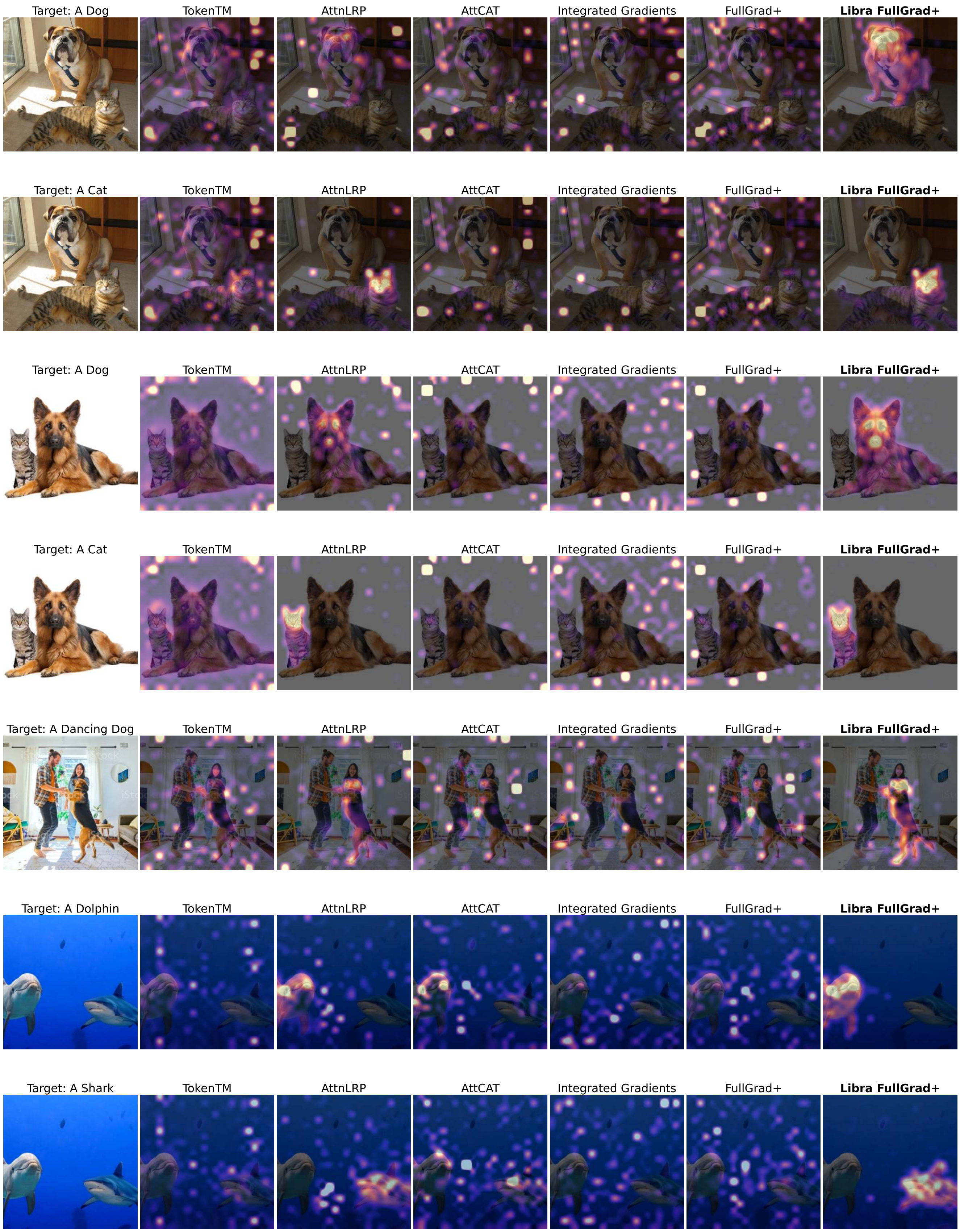

Target: A Blue Cube
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Red Cube
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Red Ball
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Blue Ball
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Elsa
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Chihiro
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Flowers
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**

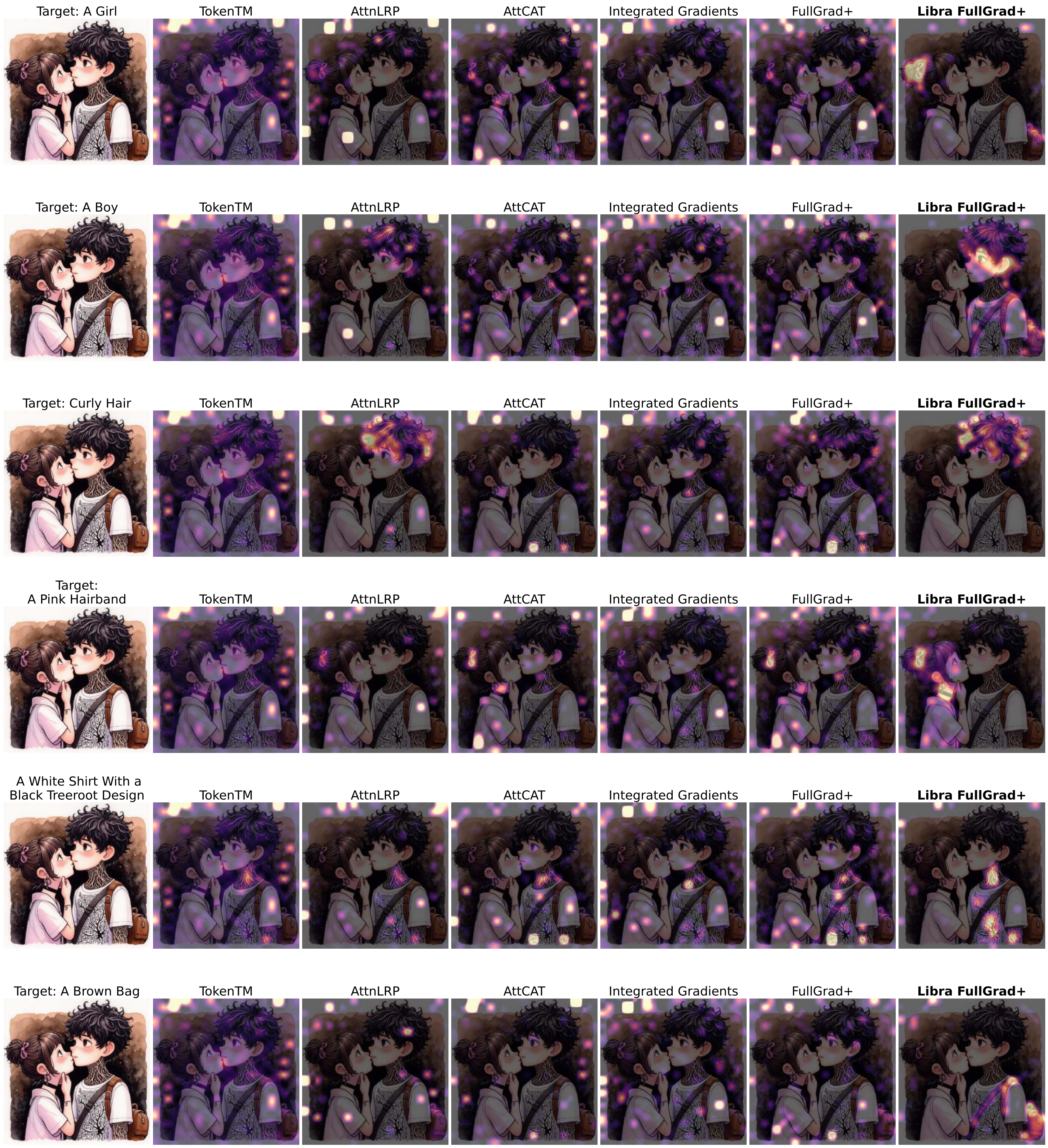

Target:
Chocolate Chips
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Ice Cream
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Orange Slices
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Eggplant
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Potato Chips
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Sunflowers
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Tomatoes
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+



Target: Bell Peppers
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Apples
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Carrots
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Grapes
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Lettuce
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Mushrooms
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+


Target: A Girl
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Tea Pot
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Batman
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
BATMAN v SUPERMAN
DAWN OF JUSTICE
Target: Superman
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
BATMAN v SUPERMAN
DAWN OF JUSTICE
Target: Batman
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Superman
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+

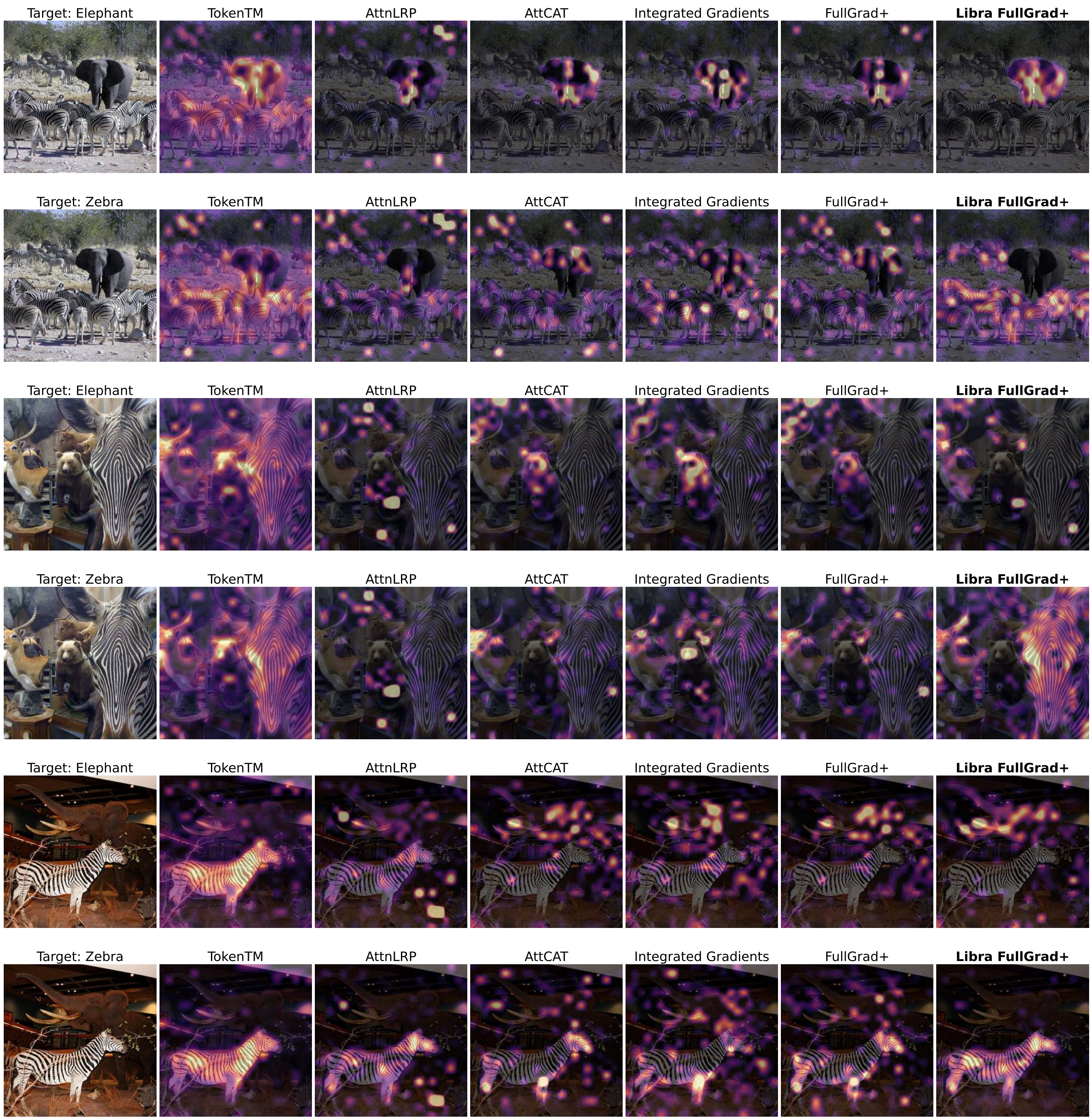

Target: A Dog
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: A Cat
Target: A Dancing Dog
Target: A Dolphin
Target: A Shark

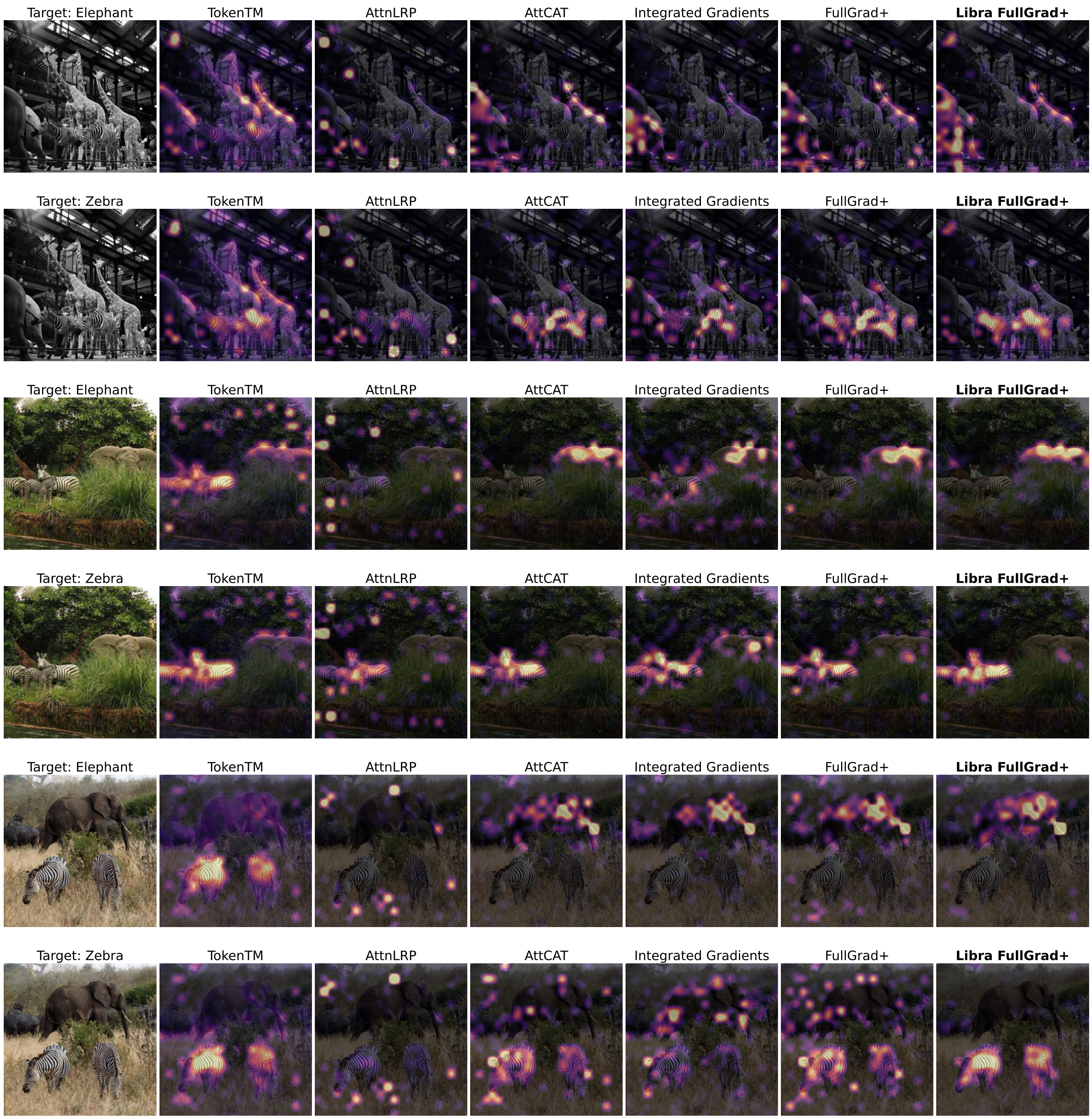

Target: A Girl
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Boy
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Curly Hair
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target:
A Pink Hairband
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
A White Shirt With a
Black Treeroot Design
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: A Brown Bag
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**

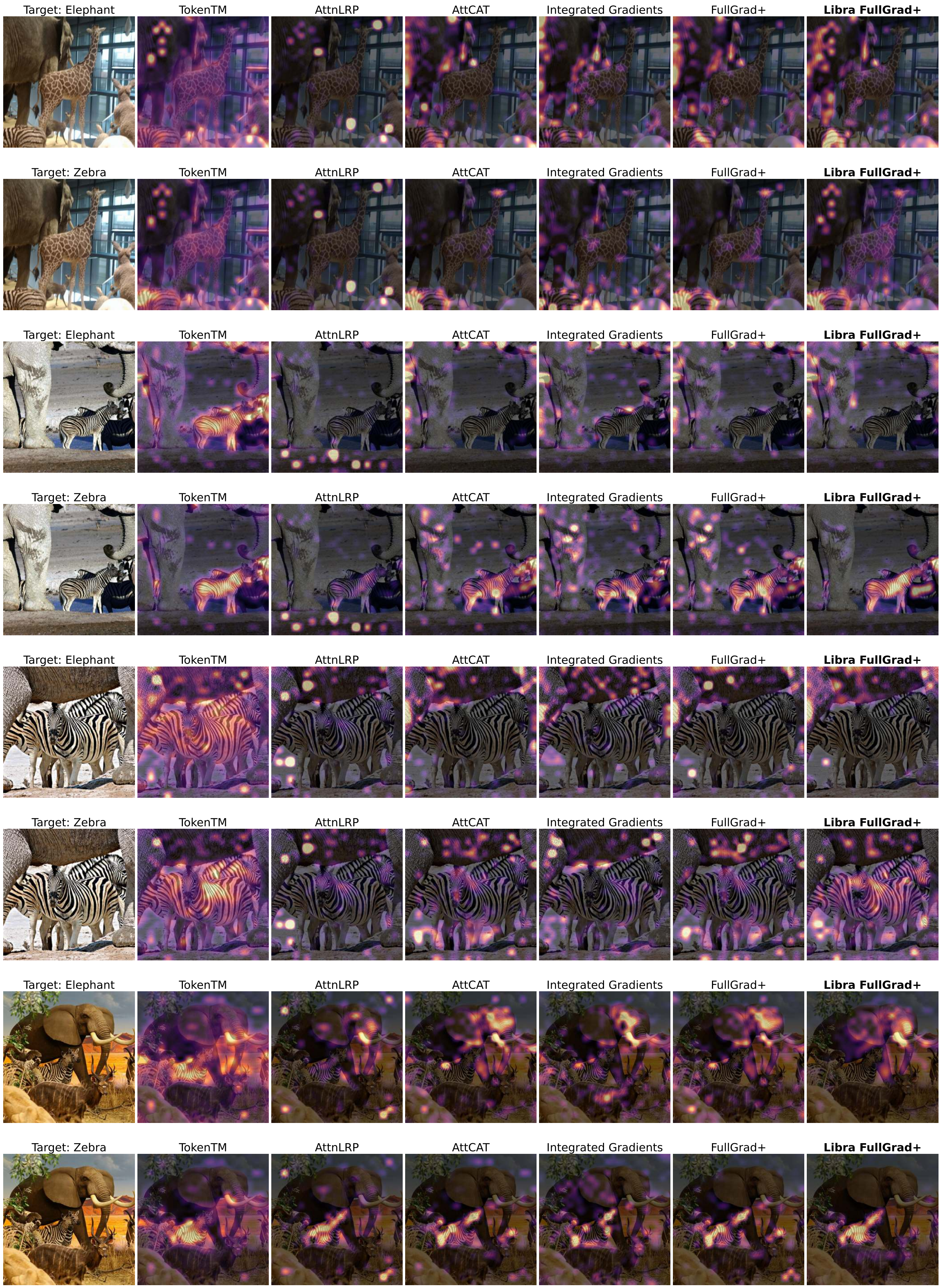
## C.2. A Comparative Study of Elephant-Zebra Multi-Class Attribution on COCO

Following Appendix B.4, we assess attribution methods' ability to generate class-discriminative explanations on ImageNet-finetuned models, focusing on challenging scenes containing co-occurring elephants and zebras.

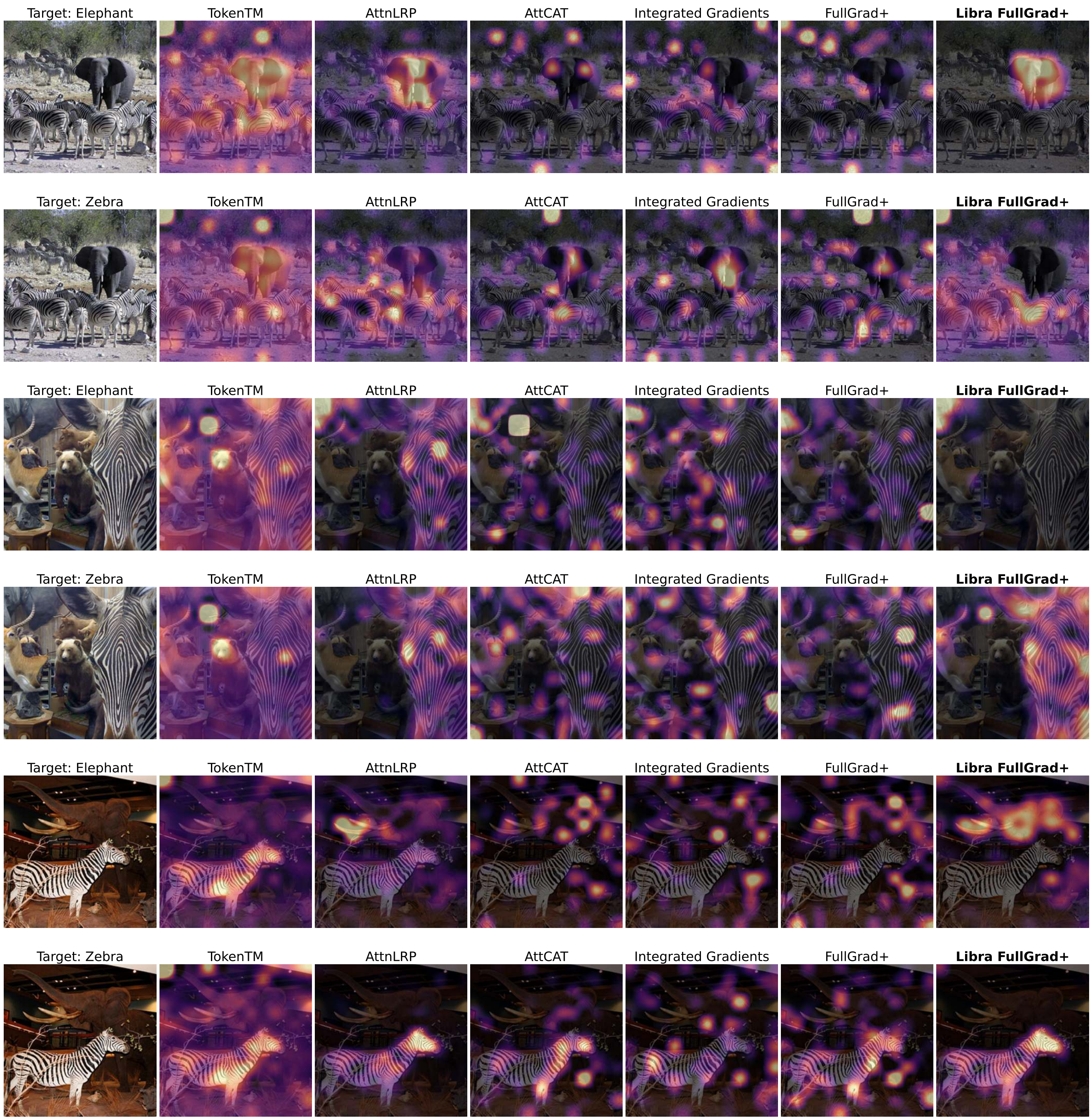
### C.2.1. Elephant-Zebra Qualitative Comparison on ViT-B

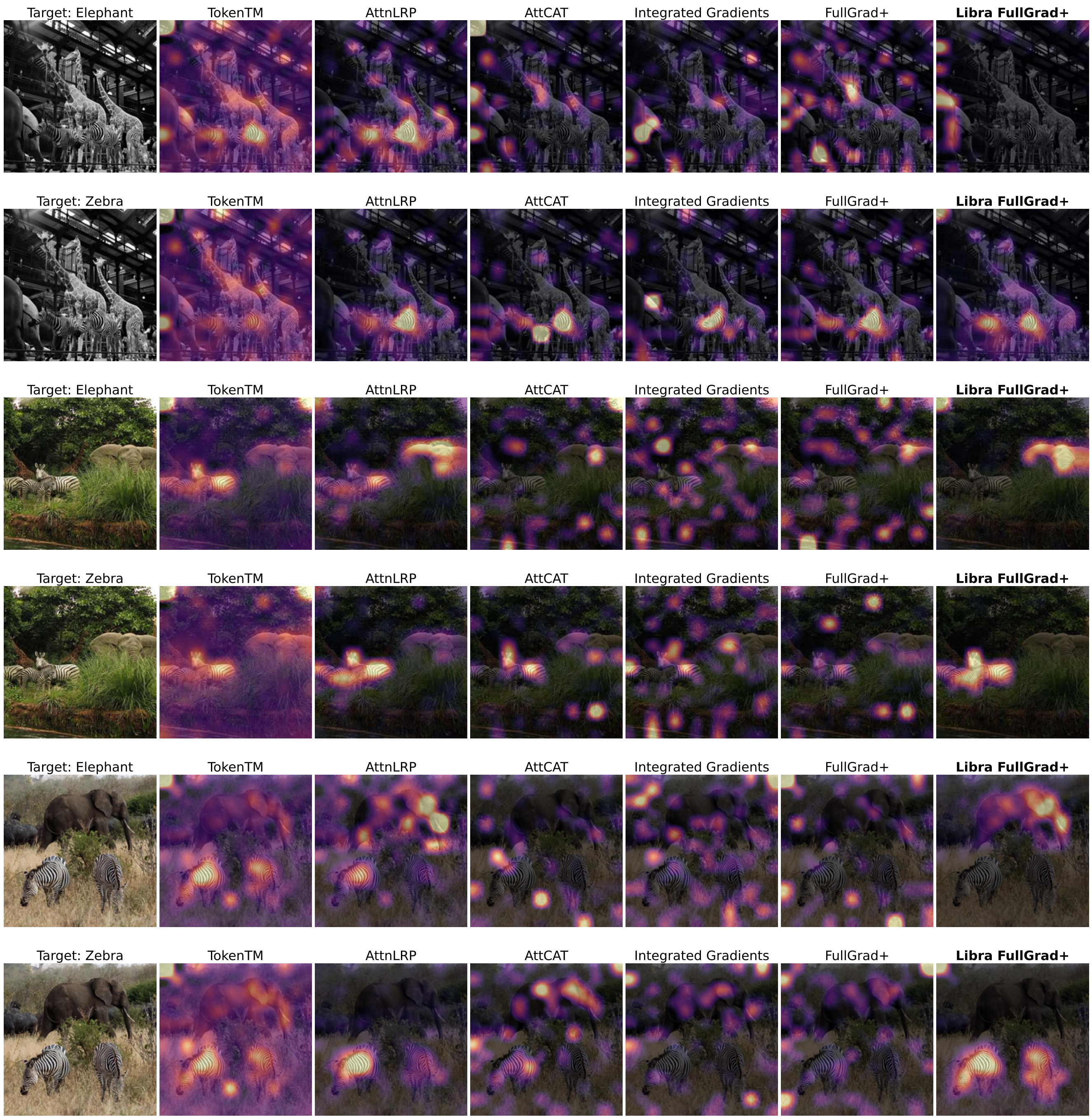



Target: Elephant
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Zebra
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Elephant
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Zebra
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Elephant
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Zebra
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+

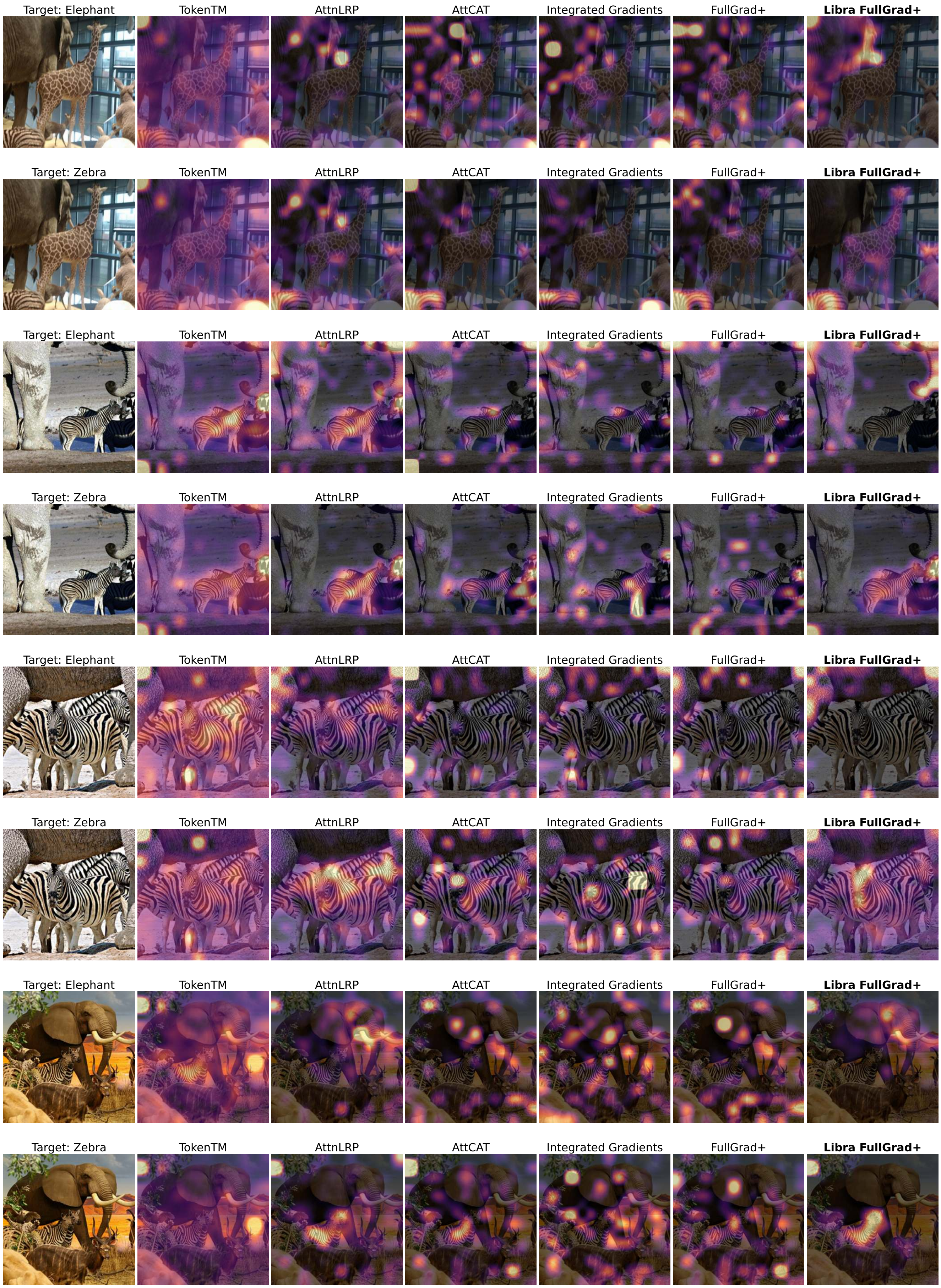

Target: Elephant
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Zebra
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Elephant
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Zebra
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Elephant
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Zebra
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Elephant
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Zebra
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**

### C.2.2. Elephant-Zebra Qualitative Comparison on BEiT2-L





Target: Elephant
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
Libra FullGrad+
Target: Zebra
Target: Elephant
Target: Zebra
Target: Elephant
Target: Zebra


Target: Elephant
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Zebra
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Elephant
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Zebra
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Elephant
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Zebra
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Elephant
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**
Target: Zebra
TokenTM
AttnLRP
AttCAT
Integrated Gradients
FullGrad+
**Libra FullGrad+**

## D. Quantitative Results

### D.1. Comparison of Compositions With LibraGrad Versus Integrated Gradients

| Method | MIF Deletion (GT) Accuracy | MIF Deletion (GT) AOPC | MIF Deletion (Predicted) Accuracy | MIF Deletion (Predicted) AOPC | Segmentation AP |
|---|---|---|---|---|---|
| Random | 36.9 ±0.1 | 14.1 ±0.2 | 29.5 ±0.1 | 15.8 ±0.2 | 42.0 ±0.4 |
| RawAtt | 45.4 ±0.1 | 22.9 ±0.3 | 39.1 ±0.1 | 25.3 ±0.2 | 40.2 ±0.4 |
| Attention Rollout | 39.0 ±0.1 | 16.5 ±0.3 | 31.4 ±0.1 | 18.3 ±0.3 | 39.9 ±0.3 |
| AliLRP | 39.8 ±0.1 | 17.2 ±0.3 | 33.2 ±0.1 | 19.2 ±0.2 | 42.7 ±0.4 |
| AttnLRP | 47.1 ±0.1 | 24.8 ±0.3 | 41.8 ±0.1 | 27.6 ±0.3 | 47.2 ±0.3 |
| DecompX | 44.4 ±0.1 | 22.6 ±0.3 | 38.9 ±0.1 | 25.3 ±0.3 | 54.2 ±0.3 |
| TokenTM | 54.9 ±0.1 | 31.8 ±0.3 | 50.0 ±0.1 | 34.9 ±0.3 | 50.0 ±0.3 |
| Input × Grad | 40.1 ±0.1 | 17.5 ±0.3 | 33.9 ±0.1 | 19.6 ±0.2 | 43.6 ±0.4 |
| Int. Gradients | 46.3 ±0.1 (+15.4%) | 23.1 ±0.3 (+32.1%) | 35.9 ±0.1 (+6.1%) | 21.9 ±0.2 (+11.6%) | 46.6 ±0.3 (+6.9%) |
| **Libra Input × Grad** | 45.9 ±0.1 (+14.4%) | 23.4 ±0.3 (+33.5%) | 40.5 ±0.1 (+19.6%) | 26.1 ±0.3 (+33.1%) | 53.6 ±0.3 (+22.9%) |
| AttCAT | 48.7 ±0.1 | 25.7 ±0.3 | 44.8 ±0.1 | 29.0 ±0.3 | 44.9 ±0.3 |
| Int. AttCAT | 53.4 ±0.1 (+9.7%) | 29.3 ±0.3 (+13.9%) | 43.2 ±0.1 (-3.6%) | 27.7 ±0.3 (-4.2%) | 50.3 ±0.3 (+12.1%) |
| **Libra AttCAT** | <u>64.7</u> ±0.1 (+33.0%) | <u>40.5</u> ±0.3 (+57.3%) | <u>61.3</u> ±0.1 (+36.9%) | <u>44.5</u> ±0.3 (+53.6%) | 53.3 ±0.3 (+18.8%) |
| GenAtt | 56.4 ±0.1 | 33.2 ±0.3 | 51.8 ±0.1 | 36.5 ±0.3 | 50.9 ±0.3 |
| Int. GenAtt | 52.7 ±0.1 (-6.6%) | 29.3 ±0.4 (-11.9%) | 43.6 ±0.1 (-15.9%) | 28.6 ±0.3 (-21.5%) | 49.1 ±0.3 (-3.6%) |
| **Libra GenAtt** | 59.7 ±0.1 (+5.9%) | 36.2 ±0.3 (+8.9%) | 55.4 ±0.1 (+6.8%) | 39.6 ±0.3 (+8.7%) | 58.6 ±0.3 (+15.1%) |
| TokenTM | 54.9 ±0.1 | 31.8 ±0.3 | 50.0 ±0.1 | 34.9 ±0.3 | 50.0 ±0.3 |
| Int. TokenTM | 53.3 ±0.1 (-2.8%) | 30.3 ±0.3 (-4.9%) | 46.4 ±0.1 (-7.2%) | 31.7 ±0.3 (-9.3%) | 49.5 ±0.3 (-0.9%) |
| **Libra TokenTM** | 57.3 ±0.1 (+4.5%) | 34.2 ±0.3 (+7.4%) | 52.5 ±0.1 (+5.0%) | 37.4 ±0.3 (+7.1%) | 53.9 ±0.3 (+7.9%) |
| GradCAM+ | 53.4 ±0.1 | 30.0 ±0.3 | 48.6 ±0.1 | 33.0 ±0.2 | 52.1 ±0.4 |
| Int. GradCAM+ | 47.9 ±0.1 (-10.3%) | 24.1 ±0.2 (-19.8%) | 41.4 ±0.1 (-14.7%) | 25.8 ±0.3 (-21.7%) | 50.0 ±0.4 (-4.0%) |
| **Libra GradCAM+** | 60.9 ±0.1 (+14.0%) | 36.7 ±0.3 (+22.0%) | 56.5 ±0.1 (+16.2%) | 40.1 ±0.3 (+21.8%) | 60.2 ±0.4 (+15.5%) |
| HiResCAM | 32.7 ±0.1 | 10.6 ±0.2 | 25.7 ±0.1 | 12.2 ±0.2 | 38.5 ±0.4 |
| Int. HiResCAM | 31.2 ±0.1 (-4.5%) | 9.1 ±0.3 (-14.0%) | 26.4 ±0.1 (+2.8%) | 12.4 ±0.2 (+1.2%) | 38.4 ±0.4 (-0.2%) |
| **Libra HiResCAM** | 54.0 ±0.1 (+65.2%) | 30.2 ±0.3 (+186.3%) | 49.0 ±0.1 (+90.7%) | 33.2 ±0.3 (+171.8%) | 48.0 ±0.3 (+24.8%) |
| XGradCAM+ | 50.9 ±0.1 | 27.7 ±0.3 | 45.9 ±0.1 | 30.5 ±0.3 | 46.9 ±0.4 |
| Int. XGradCAM+ | 48.4 ±0.1 (-4.9%) | 24.7 ±0.2 (-10.7%) | 40.2 ±0.1 (-12.3%) | 25.2 ±0.3 (-17.6%) | 48.0 ±0.4 (+2.4%) |
| **Libra XGradCAM+** | 63.0 ±0.1 (+23.6%) | 38.6 ±0.3 (+39.2%) | 58.8 ±0.1 (+28.1%) | 42.2 ±0.3 (+38.3%) | <u>60.3</u> ±0.4 (+28.6%) |
| FullGrad+ | 49.1 ±0.1 | 25.8 ±0.3 | 45.1 ±0.1 | 28.9 ±0.3 | 44.2 ±0.3 |
| Int. FullGrad+ | 52.5 ±0.1 (+7.0%) | 28.3 ±0.3 (+9.5%) | 42.1 ±0.1 (-6.6%) | 26.6 ±0.3 (-7.9%) | 49.1 ±0.3 (+11.2%) |
| **Libra FullGrad+** | **65.5** ±0.1 (+33.5%) | **41.2** ±0.3 (+59.5%) | **62.4** ±0.1 (+38.5%) | **45.3** ±0.3 (+56.5%) | **64.5** ±0.3 (+46.0%) |

Table 7. Comparison of gradient-based attribution methods and their compositions with LibraGrad and Integrated Gradients (Int. Gradients, IG) on the ViT-L model. Metrics reported are faithfulness (Most-Influential-First Deletion, MIF) and Segmentation Average Precision (AP). The results demonstrate that composing with LibraGrad universally enhances the performance of existing methods more effectively than composing with IG.

| Method | LIF Deletion (GT) | | LIF Deletion (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 62.9 ±0.1 | 85.4 ±0.2 | 70.2 ±0.1 | 83.7 ±0.2 |
| RawAtt | 60.3 ±0.1 | 83.3 ±0.2 | 67.6 ±0.1 | 81.5 ±0.1 |
| Attention Rollout | 61.9 ±0.1 | 84.1 ±0.2 | 68.3 ±0.1 | 81.9 ±0.2 |
| AliLRP | 65.4 ±0.1 | 87.7 ±0.2 | 72.5 ±0.1 | 85.9 ±0.2 |
| AttnLRP | 70.3 ±0.1 | 92.9 ±0.2 | 77.6 ±0.1 | 91.3 ±0.2 |
| DecompX | 68.8 ±0.1 | 91.0 ±0.2 | 75.8 ±0.1 | 89.3 ±0.2 |
| TokenTM | 68.9 ±0.1 | 91.6 ±0.2 | 77.3 ±0.1 | 90.3 ±0.2 |
| Input × Grad | 65.8 ±0.1 | 88.4 ±0.2 | 72.8 ±0.1 | 86.7 ±0.1 |
| Int. Gradients | 71.1 ±0.1 (+8.1%) | 93.3 ±0.2 (+5.5%) | 73.5 ±0.1 (+0.9%) | 88.4 ±0.2 (+1.9%) |
| **Libra Input × Grad** | 70.1 ±0.1 (+6.6%) | 92.0 ±0.2 (+4.0%) | 76.7 ±0.1 (+5.4%) | 90.2 ±0.2 (+4.0%) |
| AttCAT | 71.8 ±0.1 | 94.3 ±0.2 | 77.5 ±0.1 | 92.6 ±0.2 |
| Int. AttCAT | 75.2 ±0.1 (+4.8%) | 97.5 ±0.2 (+3.5%) | 76.6 ±0.1 (-1.1%) | 92.2 ±0.2 (-0.5%) |
| **Libra AttCAT** | <u>76.3</u> ±0.1 (+6.2%) | <u>98.5</u> ±0.2 (+4.5%) | <u>82.2</u> ±0.1 (+6.1%) | <u>97.1</u> ±0.2 (+4.8%) |
| GenAtt | 70.0 ±0.1 | 92.8 ±0.2 | 78.2 ±0.1 | 91.5 ±0.2 |
| Int. GenAtt | 69.3 ±0.1 (-1.0%) | 91.7 ±0.2 (-1.1%) | 74.6 ±0.1 (-4.5%) | 88.0 ±0.2 (-3.8%) |
| **Libra GenAtt** | 70.9 ±0.1 (+1.3%) | 93.2 ±0.2 (+0.5%) | 78.8 ±0.1 (+0.7%) | 92.0 ±0.2 (+0.5%) |
| TokenTM | 68.9 ±0.1 | 91.6 ±0.2 | 77.3 ±0.1 | 90.3 ±0.2 |
| Int. TokenTM | 69.0 ±0.1 (+0.2%) | 91.5 ±0.2 (-0.1%) | 76.1 ±0.1 (-1.5%) | 89.0 ±0.2 (-1.4%) |
| **Libra TokenTM** | 69.4 ±0.1 (+0.8%) | 92.1 ±0.2 (+0.5%) | 77.8 ±0.1 (+0.7%) | 90.8 ±0.2 (+0.6%) |
| GradCAM+ | 70.5 ±0.1 | 92.9 ±0.2 | 76.8 ±0.1 | 91.0 ±0.2 |
| Int. GradCAM+ | 69.0 ±0.1 (-2.2%) | 91.0 ±0.2 (-2.1%) | 73.2 ±0.1 (-4.7%) | 87.6 ±0.2 (-3.7%) |
| **Libra GradCAM+** | 72.6 ±0.1 (+2.9%) | 94.4 ±0.2 (+1.6%) | 79.1 ±0.1 (+3.0%) | 92.7 ±0.2 (+1.8%) |
| HiResCAM | 53.6 ±0.1 | 76.7 ±0.2 | 59.3 ±0.1 | 74.2 ±0.3 |
| Int. HiResCAM | 50.7 ±0.1 (-5.5%) | 74.3 ±0.3 (-3.2%) | 60.4 ±0.1 (+1.9%) | 75.0 ±0.3 (+1.0%) |
| **Libra HiResCAM** | 67.4 ±0.1 (+25.7%) | 90.0 ±0.2 (+17.3%) | 73.8 ±0.1 (+24.4%) | 88.0 ±0.2 (+18.6%) |
| XGradCAM+ | 69.5 ±0.1 | 92.1 ±0.2 | 75.7 ±0.1 | 90.1 ±0.2 |
| Int. XGradCAM+ | 69.1 ±0.1 (-0.6%) | 91.1 ±0.2 (-1.0%) | 72.2 ±0.1 (-4.7%) | 86.8 ±0.2 (-3.7%) |
| **Libra XGradCAM+** | 73.5 ±0.1 (+5.7%) | 95.3 ±0.2 (+3.5%) | 80.0 ±0.1 (+5.6%) | 93.7 ±0.2 (+3.9%) |
| FullGrad+ | 71.5 ±0.1 | 93.8 ±0.2 | 76.8 ±0.1 | 91.8 ±0.2 |
| Int. FullGrad+ | 74.8 ±0.1 (+4.7%) | 97.1 ±0.2 (+3.5%) | 76.0 ±0.1 (-1.0%) | 91.5 ±0.2 (-0.4%) |
| **Libra FullGrad+** | **76.8** ±0.1 (+7.5%) | **98.9** ±0.2 (+5.4%) | **82.6** ±0.1 (+7.6%) | **97.4** ±0.2 (+6.0%) |

Table 8. Comparison of gradient-based attribution methods and their compositions with LibraGrad and IG on the ViT-L model.

| Method | SRG (GT) | | SRG (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 49.9 ±0.1 | 49.7 ±0.2 | 49.8 ±0.1 | 49.8 ±0.2 |
| RawAtt | 52.9 ±0.1 | 53.1 ±0.2 | 53.3 ±0.1 | 53.4 ±0.2 |
| Attention Rollout | 50.4 ±0.1 | 50.3 ±0.3 | 49.9 ±0.1 | 50.1 ±0.2 |
| AliLRP | 52.6 ±0.1 | 52.4 ±0.2 | 52.8 ±0.1 | 52.5 ±0.2 |
| AttnLRP | 58.7 ±0.1 | 58.8 ±0.3 | 59.7 ±0.1 | 59.5 ±0.2 |
| DecompX | 56.6 ±0.1 | 56.8 ±0.3 | 57.4 ±0.1 | 57.3 ±0.2 |
| TokenTM | 61.9 ±0.1 | 61.7 ±0.3 | 63.6 ±0.1 | 62.6 ±0.2 |
| Input $\times$ Grad | 53.0 ±0.1 | 53.0 ±0.2 | 53.3 ±0.1 | 53.2 ±0.2 |
| Int. Gradients | 58.7 ±0.1 (+10.9%) | 58.2 ±0.3 (+9.9%) | 54.7 ±0.1 (+2.6%) | 55.1 ±0.2 (+3.7%) |
| **Libra Input $\times$ Grad** | 58.0 ±0.1 (+9.5%) | 57.7 ±0.3 (+8.9%) | 58.6 ±0.1 (+9.9%) | 58.2 ±0.2 (+9.4%) |
| AttCAT | 60.2 ±0.1 | 60.0 ±0.2 | 61.2 ±0.1 | 60.8 ±0.2 |
| Int. AttCAT | 64.3 ±0.1 (+6.8%) | 63.4 ±0.2 (+5.7%) | 59.9 ±0.1 (-2.0%) | 60.0 ±0.2 (-1.4%) |
| **Libra AttCAT** | 70.5 ±0.1 (+17.0%) | 69.5 ±0.3 (+15.8%) | 71.8 ±0.1 (+17.4%) | 70.8 ±0.2 (+16.4%) |
| GenAtt | 63.2 ±0.1 | 63.0 ±0.2 | 65.0 ±0.1 | 64.0 ±0.2 |
| Int. GenAtt | 61.0 ±0.1 (-3.5%) | 60.5 ±0.3 (-4.0%) | 59.1 ±0.1 (-9.1%) | 58.3 ±0.3 (-8.9%) |
| **Libra GenAtt** | 65.3 ±0.1 (+3.3%) | 64.7 ±0.3 (+2.7%) | 67.1 ±0.1 (+3.2%) | 65.8 ±0.3 (+2.8%) |
| TokenTM | 61.9 ±0.1 | 61.7 ±0.3 | 63.6 ±0.1 | 62.6 ±0.2 |
| Int. TokenTM | 61.2 ±0.1 (-1.1%) | 60.9 ±0.3 (-1.3%) | 61.2 ±0.1 (-3.7%) | 60.3 ±0.2 (-3.6%) |
| **Libra TokenTM** | 63.4 ±0.1 (+2.4%) | 63.1 ±0.3 (+2.3%) | 65.2 ±0.1 (+2.4%) | 64.1 ±0.3 (+2.4%) |
| GradCAM+ | 62.0 ±0.1 | 61.5 ±0.3 | 62.7 ±0.1 | 62.0 ±0.2 |
| Int. GradCAM+ | 58.5 ±0.1 (-5.7%) | 57.5 ±0.2 (-6.4%) | 57.3 ±0.1 (-8.6%) | 56.7 ±0.3 (-8.5%) |
| **Libra GradCAM+** | 66.7 ±0.1 (+7.7%) | 65.5 ±0.3 (+6.6%) | 67.8 ±0.1 (+8.1%) | 66.4 ±0.2 (+7.2%) |
| HiResCAM | 43.2 ±0.1 | 43.6 ±0.2 | 42.5 ±0.1 | 43.2 ±0.2 |
| Int. HiResCAM | 41.0 ±0.1 (-5.1%) | 41.7 ±0.3 (-4.5%) | 43.4 ±0.1 (+2.2%) | 43.7 ±0.3 (+1.0%) |
| **Libra HiResCAM** | 60.7 ±0.1 (+40.7%) | 60.1 ±0.2 (+37.7%) | 61.4 ±0.1 (+44.4%) | 60.6 ±0.2 (+40.3%) |
| XGradCAM+ | 60.2 ±0.1 | 59.9 ±0.3 | 60.8 ±0.1 | 60.3 ±0.2 |
| Int. XGradCAM+ | 58.8 ±0.1 (-2.4%) | 57.9 ±0.2 (-3.3%) | 56.2 ±0.1 (-7.5%) | 56.0 ±0.3 (-7.2%) |
| **Libra XGradCAM+** | 68.2 ±0.1 (+13.3%) | 66.9 ±0.3 (+11.8%) | 69.4 ±0.1 (+14.1%) | 68.0 ±0.3 (+12.6%) |
| FullGrad+ | 60.3 ±0.1 | 59.8 ±0.2 | 60.9 ±0.1 | 60.4 ±0.2 |
| Int. FullGrad+ | 63.7 ±0.1 (+5.6%) | 62.7 ±0.2 (+4.8%) | 59.1 ±0.1 (-3.1%) | 59.0 ±0.2 (-2.2%) |
| **Libra FullGrad+** | **71.2** ±0.1 (+18.1%) | **70.0** ±0.3 (+17.1%) | **72.5** ±0.1 (+19.0%) | **71.3** ±0.2 (+18.1%) |

Table 9. Comparison of gradient-based attribution methods and their compositions with LibraGrad and IG on the ViT-L model.

### D.2. Across Models

| Method | ViT-L | EVA2-S | BEiT2-L | FlexiViT-L | SigLIP-L | CLIP-H | DeiT3-H | *Avg.* |
|---|---|---|---|---|---|---|---|---|
| Random | 29.5 ±0.1 | 21.2 ±0.1 | 18.3 ±0.1 | 19.2 ±0.1 | 32.8 ±0.1 | 28.0 ±0.1 | 29.0 ±0.1 | 25.4 ±0.1 |
| RawAtt | 39.1 ±0.1 | 50.8 ±0.1 | 29.5 ±0.1 | 41.7 ±0.1 | - | 42.5 ±0.1 | 52.0 ±0.1 | 42.6 ±0.1 |
| Attention Rollout | 31.4 ±0.1 | 41.1 ±0.1 | 19.7 ±0.1 | 23.2 ±0.1 | - | 41.3 ±0.1 | 31.2 ±0.1 | 31.3 ±0.1 |
| AliLRP | 33.2 ±0.1 | 48.0 ±0.1 | 26.2 ±0.1 | 24.9 ±0.1 | 55.4 ±0.1 | 34.4 ±0.1 | 56.3 ±0.1 | 39.8 ±0.1 |
| AttnLRP | 41.8 ±0.1 | 63.5 ±0.1 | 37.7 ±0.1 | 21.8 ±0.1 | 62.2 ±0.1 | 46.7 ±0.1 | 40.7 ±0.1 | 44.9 ±0.1 |
| DecompX | 38.9 ±0.1 | 46.8 ±0.1 | 31.7 ±0.1 | 35.5 ±0.1 | 51.1 ±0.1 | 42.4 ±0.1 | 47.2 ±0.1 | 42.0 ±0.1 |
| Integrated Gradients | 35.9 ±0.1 | 34.8 ±0.1 | 23.2 ±0.1 | 22.3 ±0.1 | 44.0 ±0.1 | 31.0 ±0.1 | 33.2 ±0.1 | 32.1 ±0.1 |
| Input × Grad | 33.9 ±0.1 | 32.3 ±0.1 | 21.8 ±0.1 | 19.9 ±0.1 | 40.8 ±0.1 | 31.4 ±0.1 | 35.1 ±0.1 | 30.7 ±0.1 |
| **Libra Input × Grad** | 40.5 ±0.1 | 64.1 ±0.1 | 33.0 ±0.1 | 36.4 ±0.1 | 51.1 ±0.1 | 43.1 ±0.1 | 47.7 ±0.1 | 45.1 ±0.1 |
| AttCAT | 44.8 ±0.1 | 54.1 ±0.1 | 33.9 ±0.1 | 41.9 ±0.1 | 45.9 ±0.1 | 39.0 ±0.1 | 44.0 ±0.1 | 43.4 ±0.1 |
| **Libra AttCAT** | <u>61.3</u> ±0.1 | <u>69.5</u> ±0.1 | <u>48.9</u> ±0.1 | <u>58.4</u> ±0.1 | **77.4** ±0.1 | <u>58.5</u> ±0.1 | <u>70.5</u> ±0.1 | <u>63.5</u> ±0.1 |
| GenAtt | 51.8 ±0.1 | 40.7 ±0.1 | 30.8 ±0.1 | 53.0 ±0.1 | - | 51.0 ±0.1 | 64.6 ±0.1 | 48.7 ±0.1 |
| **Libra GenAtt** | 55.4 ±0.1 | 42.1 ±0.1 | 32.9 ±0.1 | 54.1 ±0.1 | - | 58.1 ±0.1 | 66.5 ±0.1 | 51.5 ±0.1 |
| TokenTM | 50.0 ±0.1 | 44.7 ±0.1 | 39.6 ±0.1 | 49.3 ±0.1 | - | 51.9 ±0.1 | 63.3 ±0.1 | 49.8 ±0.1 |
| **Libra TokenTM** | 52.5 ±0.1 | 46.0 ±0.1 | 38.3 ±0.1 | 51.0 ±0.1 | - | 57.4 ±0.1 | 65.2 ±0.1 | 51.7 ±0.1 |
| GradCAM+ | 48.6 ±0.1 | 47.1 ±0.1 | 33.4 ±0.1 | 28.7 ±0.1 | 43.5 ±0.1 | 33.0 ±0.1 | 44.5 ±0.1 | 39.8 ±0.1 |
| **Libra GradCAM+** | 56.5 ±0.1 | 67.0 ±0.1 | 37.5 ±0.1 | 33.7 ±0.1 | 47.4 ±0.1 | 36.2 ±0.1 | 48.7 ±0.1 | 46.7 ±0.1 |
| HiResCAM | 25.7 ±0.1 | 59.1 ±0.1 | 35.8 ±0.1 | 23.8 ±0.1 | 31.4 ±0.1 | 37.6 ±0.1 | 25.8 ±0.1 | 34.2 ±0.1 |
| **Libra HiResCAM** | 49.0 ±0.1 | 62.6 ±0.1 | 37.2 ±0.1 | 56.5 ±0.1 | 46.1 ±0.1 | 48.9 ±0.1 | 53.8 ±0.1 | 50.6 ±0.1 |
| XGradCAM+ | 45.9 ±0.1 | 50.2 ±0.1 | 30.6 ±0.1 | 26.6 ±0.1 | 51.4 ±0.1 | 39.4 ±0.1 | 45.1 ±0.1 | 41.3 ±0.1 |
| **Libra XGradCAM+** | 58.8 ±0.1 | 69.3 ±0.1 | 45.6 ±0.1 | 44.3 ±0.1 | 63.6 ±0.1 | 57.7 ±0.1 | 66.1 ±0.1 | 57.9 ±0.1 |
| FullGrad+ | 45.1 ±0.1 | 48.0 ±0.1 | 29.0 ±0.1 | 38.9 ±0.1 | 43.6 ±0.1 | 37.6 ±0.1 | 41.9 ±0.1 | 40.6 ±0.1 |
| **Libra FullGrad+** | **62.4** ±0.1 | **71.7** ±0.1 | **50.0** ±0.1 | **59.1** ±0.1 | <u>73.5</u> ±0.1 | **61.1** ±0.1 | **71.5** ±0.1 | **64.2** ±0.1 |

Table 10. Most-Influential-First Deletion (MIF) Accuracy evaluated using predicted labels across multiple models.

| Method | ViT-L | EVA2-S | BEiT2-L | FlexiViT-L | SigLIP-L | CLIP-H | DeiT3-H | *Avg.* |
|---|---|---|---|---|---|---|---|---|
| Random | 36.9 ±0.1 | 29.9 ±0.1 | 25.1 ±0.1 | 28.8 ±0.1 | 39.0 ±0.1 | 34.3 ±0.1 | 35.6 ±0.1 | 32.8 ±0.1 |
| RawAtt | 45.4 ±0.1 | 55.4 ±0.1 | 34.2 ±0.1 | 47.3 ±0.1 | - | 46.9 ±0.1 | 56.1 ±0.1 | 47.6 ±0.1 |
| Attention Rollout | 39.0 ±0.1 | 47.0 ±0.1 | 26.0 ±0.1 | 31.7 ±0.1 | - | 46.4 ±0.1 | 37.1 ±0.1 | 37.9 ±0.1 |
| AliLRP | 39.8 ±0.1 | 52.8 ±0.1 | 31.9 ±0.1 | 32.5 ±0.1 | 58.8 ±0.1 | 40.0 ±0.1 | 59.6 ±0.1 | 45.1 ±0.1 |
| AttnLRP | 47.1 ±0.1 | 66.6 ±0.1 | 42.1 ±0.1 | 30.3 ±0.1 | 64.7 ±0.1 | 50.8 ±0.1 | 45.4 ±0.1 | 49.6 ±0.1 |
| DecompX | 44.4 ±0.1 | 51.6 ±0.1 | 36.5 ±0.1 | 42.0 ±0.1 | 54.5 ±0.1 | 46.7 ±0.1 | 51.6 ±0.1 | 46.8 ±0.1 |
| Integrated Gradients | 46.3 ±0.1 | 46.2 ±0.1 | 31.7 ±0.1 | 31.4 ±0.1 | 52.7 ±0.1 | 37.1 ±0.1 | 43.7 ±0.1 | 41.3 ±0.1 |
| Input × Grad | 40.1 ±0.1 | 37.9 ±0.1 | 28.2 ±0.1 | 28.5 ±0.1 | 44.4 ±0.1 | 37.5 ±0.1 | 40.4 ±0.1 | 36.7 ±0.1 |
| **Libra Input × Grad** | 45.9 ±0.1 | 67.0 ±0.1 | 37.7 ±0.1 | 42.6 ±0.1 | 54.7 ±0.1 | 47.5 ±0.1 | 52.1 ±0.1 | 49.7 ±0.1 |
| AttCAT | 48.7 ±0.1 | 56.9 ±0.1 | 38.4 ±0.1 | 45.3 ±0.1 | 48.3 ±0.1 | 42.5 ±0.1 | 48.2 ±0.1 | 46.9 ±0.1 |
| **Libra AttCAT** | <u>64.7</u> ±0.1 | <u>72.1</u> ±0.1 | <u>52.5</u> ±0.1 | <u>61.8</u> ±0.1 | **79.0** ±0.1 | <u>61.5</u> ±0.1 | <u>72.6</u> ±0.1 | <u>66.3</u> ±0.1 |
| GenAtt | 56.4 ±0.1 | 46.3 ±0.1 | 35.6 ±0.1 | 57.2 ±0.1 | - | 54.4 ±0.1 | 67.2 ±0.1 | 52.9 ±0.1 |
| **Libra GenAtt** | 59.7 ±0.1 | 47.7 ±0.1 | 37.6 ±0.1 | 58.3 ±0.1 | - | 61.0 ±0.1 | 69.1 ±0.1 | 55.6 ±0.1 |
| TokenTM | 54.9 ±0.1 | 50.4 ±0.1 | 43.9 ±0.1 | 54.3 ±0.1 | - | 55.4 ±0.1 | 66.2 ±0.1 | 54.2 ±0.1 |
| **Libra TokenTM** | 57.3 ±0.1 | 51.6 ±0.1 | 42.6 ±0.1 | 55.7 ±0.1 | - | 60.6 ±0.1 | 68.1 ±0.1 | 56.0 ±0.1 |
| GradCAM+ | 53.4 ±0.1 | 50.6 ±0.1 | 38.4 ±0.1 | 35.8 ±0.1 | 47.6 ±0.1 | 38.6 ±0.1 | 49.5 ±0.1 | 44.8 ±0.1 |
| **Libra GradCAM+** | 60.9 ±0.1 | 69.9 ±0.1 | 42.3 ±0.1 | 40.2 ±0.1 | 51.0 ±0.1 | 41.8 ±0.1 | 52.6 ±0.1 | 51.3 ±0.1 |
| HiResCAM | 32.7 ±0.1 | 63.1 ±0.1 | 40.3 ±0.1 | 31.2 ±0.1 | 37.1 ±0.1 | 42.3 ±0.1 | 32.5 ±0.1 | 39.9 ±0.1 |
| **Libra HiResCAM** | 54.0 ±0.1 | 65.9 ±0.1 | 41.5 ±0.1 | 60.1 ±0.1 | 50.0 ±0.1 | 52.8 ±0.1 | 57.4 ±0.1 | 54.5 ±0.1 |
| XGradCAM+ | 50.9 ±0.1 | 53.7 ±0.1 | 35.6 ±0.1 | 33.4 ±0.1 | 54.8 ±0.1 | 44.2 ±0.1 | 49.1 ±0.1 | 46.0 ±0.1 |
| **Libra XGradCAM+** | 63.0 ±0.1 | 71.9 ±0.1 | 49.5 ±0.1 | 49.7 ±0.1 | 66.3 ±0.1 | 60.8 ±0.1 | 68.8 ±0.1 | 61.4 ±0.1 |
| FullGrad+ | 49.1 ±0.1 | 50.9 ±0.1 | 34.4 ±0.1 | 43.0 ±0.1 | 46.6 ±0.1 | 41.4 ±0.1 | 45.8 ±0.1 | 44.4 ±0.1 |
| **Libra FullGrad+** | **65.5** ±0.1 | **74.1** ±0.1 | **53.4** ±0.1 | **62.4** ±0.1 | <u>75.3</u> ±0.1 | **63.8** ±0.1 | **73.5** ±0.1 | **66.9** ±0.1 |

Table 11. Most-Influential-First Deletion (MIF) Accuracy evaluated using ground-truth labels across multiple models.

| Method | ViT-L | EVA2-S | BEiT2-L | FlexiViT-L | SigLIP-L | CLIP-H | DeiT3-H | *Avg.* |
|---|---|---|---|---|---|---|---|---|
| Random | 15.8 ±0.2 | 8.2 ±0.2 | 6.8 ±0.1 | 6.4 ±0.2 | 19.1 ±0.2 | 12.7 ±0.2 | 19.2 ±0.2 | 12.6 ±0.2 |
| RawAtt | 25.3 ±0.2 | 33.9 ±0.3 | 17.5 ±0.2 | 26.5 ±0.3 | - | 23.3 ±0.2 | 37.2 ±0.2 | 27.3 ±0.2 |
| Attention Rollout | 18.3 ±0.3 | 24.9 ±0.3 | 8.6 ±0.1 | 9.7 ±0.2 | - | 22.5 ±0.3 | 21.9 ±0.2 | 17.7 ±0.2 |
| AliLRP | 19.2 ±0.2 | 31.3 ±0.3 | 13.9 ±0.2 | 10.5 ±0.2 | 40.0 ±0.3 | 17.3 ±0.2 | 41.7 ±0.2 | 24.8 ±0.2 |
| AttnLRP | 27.6 ±0.3 | 44.2 ±0.2 | 25.0 ±0.2 | 8.3 ±0.2 | 46.2 ±0.3 | 26.4 ±0.2 | 31.7 ±0.2 | 29.9 ±0.2 |
| DecompX | 25.3 ±0.3 | 30.7 ±0.3 | 19.4 ±0.2 | 20.7 ±0.2 | 35.7 ±0.2 | 23.5 ±0.2 | 35.9 ±0.2 | 27.3 ±0.2 |
| Integrated Gradients | 21.9 ±0.2 | 19.3 ±0.2 | 11.9 ±0.1 | 9.4 ±0.2 | 28.8 ±0.2 | 15.0 ±0.2 | 22.8 ±0.2 | 18.4 ±0.2 |
| Input × Grad | 19.6 ±0.2 | 17.0 ±0.2 | 10.3 ±0.1 | 6.5 ±0.2 | 26.0 ±0.2 | 15.2 ±0.2 | 25.1 ±0.2 | 17.1 ±0.2 |
| **Libra Input × Grad** | 26.1 ±0.3 | 44.4 ±0.3 | 20.2 ±0.2 | 21.3 ±0.2 | 35.6 ±0.2 | 24.0 ±0.2 | 36.3 ±0.2 | 29.7 ±0.2 |
| AttCAT | 29.0 ±0.3 | 35.3 ±0.3 | 21.0 ±0.2 | 22.6 ±0.3 | 30.9 ±0.2 | 21.3 ±0.1 | 32.3 ±0.3 | 27.5 ±0.2 |
| **Libra AttCAT** | 44.5 ±0.3 | 48.7 ±0.2 | 34.6 ±0.2 | 39.6 ±0.3 | **59.7** ±0.2 | 34.7 ±0.2 | 52.8 ±0.2 | 44.9 ±0.2 |
| GenAtt | 36.5 ±0.3 | 24.3 ±0.2 | 19.2 ±0.2 | 35.1 ±0.3 | - | 29.6 ±0.2 | 48.1 ±0.2 | 32.1 ±0.2 |
| **Libra GenAtt** | 39.6 ±0.3 | 25.6 ±0.2 | 20.7 ±0.3 | 36.7 ±0.3 | - | 34.5 ±0.2 | 49.7 ±0.2 | 34.5 ±0.3 |
| TokenTM | 34.9 ±0.3 | 28.3 ±0.3 | 26.8 ±0.3 | 32.7 ±0.3 | - | 30.1 ±0.2 | 47.2 ±0.2 | 33.3 ±0.3 |
| **Libra TokenTM** | 37.4 ±0.3 | 28.8 ±0.3 | 25.5 ±0.3 | 34.4 ±0.3 | - | 34.1 ±0.2 | 48.8 ±0.2 | 34.8 ±0.3 |
| GradCAM+ | 33.0 ±0.2 | 29.0 ±0.3 | 20.1 ±0.2 | 13.1 ±0.2 | 28.1 ±0.2 | 16.2 ±0.2 | 31.8 ±0.2 | 24.5 ±0.2 |
| **Libra GradCAM+** | 40.1 ±0.3 | 46.1 ±0.2 | 24.3 ±0.2 | 18.4 ±0.3 | 31.9 ±0.3 | 18.6 ±0.2 | 35.5 ±0.2 | 30.7 ±0.3 |
| HiResCAM | 12.2 ±0.2 | 40.1 ±0.2 | 22.3 ±0.2 | 9.0 ±0.2 | 17.5 ±0.2 | 19.7 ±0.2 | 17.4 ±0.2 | 19.7 ±0.2 |
| **Libra HiResCAM** | 33.2 ±0.3 | 42.9 ±0.2 | 23.6 ±0.2 | 38.1 ±0.3 | 30.4 ±0.2 | 27.9 ±0.2 | 39.7 ±0.2 | 33.7 ±0.2 |
| XGradCAM+ | 30.5 ±0.3 | 31.9 ±0.2 | 17.9 ±0.2 | 9.9 ±0.2 | 37.8 ±0.2 | 21.3 ±0.2 | 31.8 ±0.2 | 25.9 ±0.2 |
| **Libra XGradCAM+** | 42.2 ±0.3 | 48.1 ±0.2 | 31.4 ±0.3 | 27.2 ±0.3 | 46.3 ±0.3 | 34.1 ±0.2 | 49.0 ±0.2 | 39.8 ±0.3 |
| FullGrad+ | 28.9 ±0.3 | 30.0 ±0.2 | 16.6 ±0.2 | 20.8 ±0.3 | 29.0 ±0.2 | 20.5 ±0.2 | 30.0 ±0.3 | 25.1 ±0.2 |
| **Libra FullGrad+** | **45.3** ±0.3 | **50.5** ±0.2 | **35.5** ±0.3 | **39.8** ±0.3 | 55.1 ±0.2 | **36.8** ±0.2 | **53.7** ±0.2 | **45.2** ±0.2 |

Table 12. Most-Influential-First Deletion (MIF) AOPC evaluated using predicted labels across multiple models.

| Method | ViT-L | EVA2-S | BEiT2-L | FlexiViT-L | SigLIP-L | CLIP-H | DeiT3-H | *Avg.* |
|---|---|---|---|---|---|---|---|---|
| Random | 14.1 ±0.2 | 6.6 ±0.2 | 5.6 ±0.2 | 5.2 ±0.2 | 17.3 ±0.2 | 11.2 ±0.2 | 16.6 ±0.2 | 10.9 ±0.2 |
| RawAtt | 22.9 ±0.3 | 30.3 ±0.3 | 15.3 ±0.2 | 23.6 ±0.3 | - | 21.0 ±0.2 | 33.3 ±0.3 | 24.4 ±0.3 |
| Attention Rollout | 16.5 ±0.3 | 22.0 ±0.4 | 7.2 ±0.1 | 8.2 ±0.2 | - | 20.5 ±0.3 | 19.0 ±0.2 | 15.5 ±0.3 |
| AliLRP | 17.2 ±0.3 | 27.7 ±0.4 | 12.4 ±0.2 | 8.8 ±0.2 | 36.6 ±0.3 | 15.7 ±0.2 | 37.3 ±0.3 | 22.2 ±0.3 |
| AttnLRP | 24.8 ±0.3 | 39.6 ±0.3 | 22.6 ±0.3 | 6.6 ±0.2 | 42.4 ±0.3 | 24.0 ±0.3 | 28.1 ±0.3 | 26.9 ±0.3 |
| DecompX | 22.6 ±0.3 | 27.0 ±0.4 | 17.3 ±0.3 | 18.1 ±0.2 | 32.6 ±0.2 | 21.3 ±0.2 | 32.2 ±0.3 | 24.4 ±0.3 |
| Integrated Gradients | 23.1 ±0.3 | 21.0 ±0.3 | 12.5 ±0.2 | 8.3 ±0.2 | 30.0 ±0.2 | 13.5 ±0.2 | 24.9 ±0.3 | 19.1 ±0.2 |
| Input × Grad | 17.5 ±0.3 | 14.1 ±0.2 | 9.0 ±0.1 | 5.1 ±0.2 | 23.2 ±0.2 | 13.7 ±0.2 | 21.9 ±0.3 | 14.9 ±0.2 |
| **Libra Input × Grad** | 23.4 ±0.3 | 39.6 ±0.3 | 18.0 ±0.2 | 18.6 ±0.2 | 32.4 ±0.2 | 21.8 ±0.2 | 32.4 ±0.3 | 26.6 ±0.3 |
| AttCAT | 25.7 ±0.3 | 30.9 ±0.2 | 18.9 ±0.2 | 18.9 ±0.3 | 27.4 ±0.3 | 18.8 ±0.2 | 28.6 ±0.3 | 24.2 ±0.3 |
| **Libra AttCAT** | 40.5 ±0.3 | 43.8 ±0.3 | 31.6 ±0.3 | 35.5 ±0.3 | **55.0** ±0.3 | 31.7 ±0.3 | 47.6 ±0.3 | 40.8 ±0.3 |
| GenAtt | 33.2 ±0.3 | 21.2 ±0.2 | 17.0 ±0.3 | 31.4 ±0.3 | - | 26.8 ±0.2 | 43.3 ±0.3 | 28.8 ±0.3 |
| **Libra GenAtt** | 36.2 ±0.3 | 22.5 ±0.3 | 18.4 ±0.3 | 32.9 ±0.3 | - | 31.5 ±0.3 | 45.0 ±0.3 | 31.1 ±0.3 |
| TokenTM | 31.8 ±0.3 | 25.1 ±0.3 | 24.3 ±0.3 | 29.3 ±0.3 | - | 27.4 ±0.3 | 42.6 ±0.3 | 30.1 ±0.3 |
| **Libra TokenTM** | 34.2 ±0.3 | 25.6 ±0.3 | 23.1 ±0.3 | 30.9 ±0.3 | - | 31.2 ±0.3 | 44.1 ±0.3 | 31.5 ±0.3 |
| GradCAM+ | 30.0 ±0.3 | 25.1 ±0.3 | 18.2 ±0.2 | 10.9 ±0.2 | 25.4 ±0.3 | 14.5 ±0.2 | 28.3 ±0.3 | 21.8 ±0.2 |
| **Libra GradCAM+** | 36.7 ±0.3 | 41.4 ±0.3 | 22.0 ±0.2 | 15.7 ±0.2 | 28.8 ±0.3 | 16.8 ±0.2 | 31.4 ±0.3 | 27.5 ±0.3 |
| HiResCAM | 10.6 ±0.2 | 36.1 ±0.2 | 20.1 ±0.2 | 7.2 ±0.2 | 15.7 ±0.2 | 17.6 ±0.2 | 15.0 ±0.2 | 17.5 ±0.2 |
| **Libra HiResCAM** | 30.2 ±0.3 | 38.6 ±0.3 | 21.2 ±0.2 | 34.2 ±0.3 | 27.5 ±0.3 | 25.4 ±0.2 | 35.4 ±0.3 | 30.4 ±0.3 |
| XGradCAM+ | 27.7 ±0.3 | 27.9 ±0.2 | 16.0 ±0.2 | 7.8 ±0.2 | 34.5 ±0.3 | 19.2 ±0.2 | 27.9 ±0.3 | 23.0 ±0.2 |
| **Libra XGradCAM+** | 38.6 ±0.3 | 43.3 ±0.3 | 28.6 ±0.3 | 24.1 ±0.3 | 42.5 ±0.3 | 31.1 ±0.2 | 44.2 ±0.3 | 36.0 ±0.3 |
| FullGrad+ | 25.8 ±0.3 | 25.7 ±0.2 | 14.9 ±0.2 | 17.5 ±0.3 | 25.8 ±0.3 | 18.1 ±0.2 | 26.2 ±0.3 | 22.0 ±0.3 |
| **Libra FullGrad+** | **41.2** ±0.3 | **45.5** ±0.3 | **32.4** ±0.3 | **35.8** ±0.3 | 50.7 ±0.3 | **33.6** ±0.3 | **48.5** ±0.3 | **41.1** ±0.3 |

Table 13. Most-Influential-First Deletion (MIF) AOPC evaluated using ground-truth labels across multiple models.

| Method | ViT-L | EVA2-S | BEiT2-L | FlexiViT-L | SigLIP-L | CLIP-H | DeiT3-H | *Avg.* |
|---|---|---|---|---|---|---|---|---|
| Random | 70.2 ±0.1 | 79.0 ±0.1 | 81.7 ±0.1 | 80.7 ±0.1 | 67.1 ±0.1 | 72.4 ±0.1 | 70.7 ±0.1 | 74.5 ±0.1 |
| RawAtt | 67.6 ±0.1 | 82.7 ±0.1 | 83.7 ±0.1 | 82.6 ±0.1 | - | 76.0 ±0.1 | 78.4 ±0.1 | 78.5 ±0.1 |
| Attention Rollout | 68.3 ±0.1 | 78.8 ±0.1 | 75.6 ±0.1 | 72.7 ±0.1 | - | 74.6 ±0.1 | 64.5 ±0.1 | 72.4 ±0.1 |
| AliLRP | 72.5 ±0.1 | 87.2 ±0.1 | 84.5 ±0.1 | 84.7 ±0.1 | 77.0 ±0.1 | 75.3 ±0.1 | 86.1 ±0.1 | 81.0 ±0.1 |
| AttnLRP | 77.6 ±0.1 | 87.5 ±0.0 | 85.7 ±0.1 | 77.9 ±0.1 | 82.2 ±0.1 | 83.3 ±0.1 | 80.8 ±0.1 | 82.2 ±0.1 |
| DecompX | 75.8 ±0.1 | 85.8 ±0.1 | 84.9 ±0.1 | 86.2 ±0.1 | 78.1 ±0.1 | 81.7 ±0.1 | 83.1 ±0.1 | 82.2 ±0.1 |
| Integrated Gradients | 73.5 ±0.1 | 83.5 ±0.1 | 84.2 ±0.1 | 77.7 ±0.1 | 75.6 ±0.1 | 69.4 ±0.1 | 74.6 ±0.1 | 76.9 ±0.1 |
| Input × Grad | 72.8 ±0.1 | 84.0 ±0.1 | 82.0 ±0.1 | 78.3 ±0.1 | 71.6 ±0.1 | 68.8 ±0.1 | 77.7 ±0.1 | 76.5 ±0.1 |
| **Libra Input × Grad** | 76.7 ±0.1 | 88.3 ±0.0 | 85.7 ±0.1 | 86.9 ±0.1 | 78.3 ±0.1 | 82.2 ±0.1 | 83.7 ±0.1 | 83.1 ±0.1 |
| AttCAT | 77.5 ±0.1 | 87.8 ±0.0 | **87.5** ±0.0 | 88.3 ±0.0 | 76.6 ±0.1 | 76.9 ±0.1 | 80.5 ±0.1 | 82.2 ±0.1 |
| **Libra AttCAT** | 82.2 ±0.1 | 88.3 ±0.0 | 87.0 ±0.1 | **88.5** ±0.0 | **85.9** ±0.1 | 83.8 ±0.1 | **87.7** ±0.0 | 86.2 ±0.1 |
| GenAtt | 78.2 ±0.1 | 80.7 ±0.1 | 83.2 ±0.1 | 87.0 ±0.1 | - | 80.8 ±0.1 | 85.7 ±0.1 | 82.6 ±0.1 |
| **Libra GenAtt** | 78.8 ±0.1 | 81.6 ±0.1 | 83.2 ±0.1 | 86.6 ±0.1 | - | 82.5 ±0.1 | 86.0 ±0.1 | 83.1 ±0.1 |
| TokenTM | 77.3 ±0.1 | 82.1 ±0.1 | 84.6 ±0.1 | 86.0 ±0.1 | - | 80.6 ±0.1 | 85.0 ±0.1 | 82.6 ±0.1 |
| **Libra TokenTM** | 77.8 ±0.1 | 81.9 ±0.1 | 83.8 ±0.1 | 85.8 ±0.1 | - | 81.7 ±0.1 | 85.4 ±0.1 | 82.7 ±0.1 |
| GradCAM+ | 76.8 ±0.1 | 82.8 ±0.1 | 85.1 ±0.1 | 72.3 ±0.1 | 49.0 ±0.1 | 69.4 ±0.1 | 75.8 ±0.1 | 73.0 ±0.1 |
| **Libra GradCAM+** | 79.1 ±0.1 | 86.4 ±0.1 | 84.2 ±0.1 | 80.6 ±0.1 | 67.5 ±0.1 | 70.7 ±0.1 | 80.7 ±0.1 | 78.5 ±0.1 |
| HiResCAM | 59.3 ±0.1 | 86.1 ±0.1 | 85.5 ±0.1 | 78.7 ±0.1 | 51.9 ±0.1 | 77.9 ±0.1 | 75.5 ±0.1 | 73.5 ±0.1 |
| **Libra HiResCAM** | 73.8 ±0.1 | 86.3 ±0.1 | 86.0 ±0.1 | 87.3 ±0.0 | 68.2 ±0.1 | 80.9 ±0.1 | 80.6 ±0.1 | 80.5 ±0.1 |
| XGradCAM+ | 75.7 ±0.1 | 83.8 ±0.1 | 84.3 ±0.1 | 72.3 ±0.1 | 60.6 ±0.1 | 75.4 ±0.1 | 77.1 ±0.1 | 75.6 ±0.1 |
| **Libra XGradCAM+** | 80.0 ±0.1 | 86.6 ±0.1 | 85.6 ±0.1 | 85.3 ±0.1 | 76.4 ±0.1 | 81.0 ±0.1 | 86.4 ±0.1 | 83.0 ±0.1 |
| FullGrad+ | 76.8 ±0.1 | 86.8 ±0.1 | 86.0 ±0.1 | 87.8 ±0.0 | 73.3 ±0.1 | 76.2 ±0.1 | 79.9 ±0.1 | 81.0 ±0.1 |
| **Libra FullGrad+** | **82.6** ±0.1 | **88.5** ±0.0 | 86.9 ±0.1 | 88.3 ±0.0 | 85.8 ±0.1 | **84.9** ±0.1 | 87.6 ±0.0 | **86.4** ±0.1 |

Table 14. Least-Influential-First Deletion (LIF) Accuracy evaluated using predicted labels across multiple models.

| Method | ViT-L | EVA2-S | BEiT2-L | FlexiViT-L | SigLIP-L | CLIP-H | DeiT3-H | *Avg.* |
|---|---|---|---|---|---|---|---|---|
| Random | 62.9 ±0.1 | 70.0 ±0.1 | 74.6 ±0.1 | 70.7 ±0.1 | 61.1 ±0.1 | 65.8 ±0.1 | 64.2 ±0.1 | 67.0 ±0.1 |
| RawAtt | 60.3 ±0.1 | 73.3 ±0.1 | 76.6 ±0.1 | 72.8 ±0.1 | - | 68.7 ±0.1 | 70.9 ±0.1 | 70.4 ±0.1 |
| Attention Rollout | 61.9 ±0.1 | 70.1 ±0.1 | 69.9 ±0.1 | 65.0 ±0.1 | - | 68.1 ±0.1 | 59.0 ±0.1 | 65.7 ±0.1 |
| AliLRP | 65.4 ±0.1 | 79.7 ±0.1 | 78.0 ±0.1 | 75.8 ±0.1 | 70.8 ±0.1 | 69.1 ±0.1 | 79.5 ±0.1 | 74.0 ±0.1 |
| AttnLRP | 70.3 ±0.1 | 78.8 ±0.1 | 78.4 ±0.1 | 68.9 ±0.1 | 75.0 ±0.1 | 76.8 ±0.1 | 74.1 ±0.1 | 74.6 ±0.1 |
| DecompX | 68.8 ±0.1 | 76.3 ±0.1 | 77.6 ±0.1 | 76.7 ±0.1 | 71.3 ±0.1 | 74.8 ±0.1 | 75.8 ±0.1 | 74.5 ±0.1 |
| Integrated Gradients | 71.1 ±0.1 | 82.0 ±0.1 | 79.4 ±0.1 | 70.2 ±0.1 | 75.9 ±0.1 | 63.3 ±0.1 | 71.5 ±0.1 | 73.3 ±0.1 |
| Input × Grad | 65.8 ±0.1 | 76.5 ±0.1 | 75.5 ±0.1 | 69.6 ±0.1 | 67.4 ±0.1 | 62.7 ±0.1 | 71.9 ±0.1 | 69.9 ±0.1 |
| **Libra Input × Grad** | 70.1 ±0.1 | 82.0 ±0.1 | 79.2 ±0.1 | 78.2 ±0.1 | 71.8 ±0.1 | 76.0 ±0.1 | 77.1 ±0.1 | 76.3 ±0.1 |
| AttCAT | 71.8 ±0.1 | **82.7** ±0.1 | **81.8** ±0.1 | **83.1** ±0.1 | 73.8 ±0.1 | 72.3 ±0.1 | 75.3 ±0.1 | 77.3 ±0.1 |
| **Libra AttCAT** | 76.3 ±0.1 | 82.2 ±0.1 | 80.8 ±0.1 | 81.4 ±0.1 | **80.0** ±0.1 | 78.1 ±0.1 | 81.7 ±0.1 | 80.1 ±0.1 |
| GenAtt | 70.0 ±0.1 | 71.9 ±0.1 | 75.6 ±0.1 | 77.3 ±0.1 | - | 73.4 ±0.1 | 76.9 ±0.1 | 74.2 ±0.1 |
| **Libra GenAtt** | 70.9 ±0.1 | 72.7 ±0.1 | 75.5 ±0.1 | 77.0 ±0.1 | - | 75.0 ±0.1 | 77.3 ±0.1 | 74.7 ±0.1 |
| TokenTM | 68.9 ±0.1 | 73.3 ±0.1 | 76.8 ±0.1 | 76.0 ±0.1 | - | 73.1 ±0.1 | 76.2 ±0.1 | 74.0 ±0.1 |
| **Libra TokenTM** | 69.4 ±0.1 | 72.9 ±0.1 | 76.2 ±0.1 | 75.7 ±0.1 | - | 74.1 ±0.1 | 76.4 ±0.1 | 74.1 ±0.1 |
| GradCAM+ | 70.5 ±0.1 | 77.3 ±0.1 | 79.2 ±0.1 | 64.7 ±0.1 | 45.8 ±0.1 | 63.8 ±0.1 | 69.3 ±0.1 | 67.2 ±0.1 |
| **Libra GradCAM+** | 72.6 ±0.1 | 80.1 ±0.1 | 78.4 ±0.1 | 72.9 ±0.1 | 62.8 ±0.1 | 65.6 ±0.1 | 74.1 ±0.1 | 72.3 ±0.1 |
| HiResCAM | 53.6 ±0.1 | 79.3 ±0.1 | 79.4 ±0.1 | 70.0 ±0.1 | 48.1 ±0.1 | 72.4 ±0.1 | 68.1 ±0.1 | 67.3 ±0.1 |
| **Libra HiResCAM** | 67.4 ±0.1 | 79.4 ±0.1 | 80.0 ±0.1 | 80.7 ±0.1 | 63.7 ±0.1 | 74.7 ±0.1 | 75.2 ±0.1 | 74.4 ±0.1 |
| XGradCAM+ | 69.5 ±0.1 | 78.3 ±0.1 | 78.9 ±0.1 | 65.0 ±0.1 | 57.3 ±0.1 | 69.7 ±0.1 | 71.4 ±0.1 | 70.0 ±0.1 |
| **Libra XGradCAM+** | 73.5 ±0.1 | 80.1 ±0.1 | 79.5 ±0.1 | 77.5 ±0.1 | 70.5 ±0.1 | 75.5 ±0.1 | 79.6 ±0.1 | 76.6 ±0.1 |
| FullGrad+ | 71.5 ±0.1 | 82.1 ±0.1 | 79.9 ±0.1 | 81.4 ±0.1 | 70.4 ±0.1 | 71.4 ±0.1 | 74.6 ±0.1 | 75.9 ±0.1 |
| **Libra FullGrad+** | **76.8** ±0.1 | 82.6 ±0.1 | 80.8 ±0.1 | 81.5 ±0.1 | 79.8 ±0.1 | **79.1** ±0.1 | **81.8** ±0.1 | **80.4** ±0.1 |

Table 15. Least-Influential-First Deletion (LIF) Accuracy evaluated using ground-truth labels across multiple models.

| Method | ViT-L | EVA2-S | BEiT2-L | FlexiViT-L | SigLIP-L | CLIP-H | DeiT3-H | *Avg.* |
|---|---|---|---|---|---|---|---|---|
| Random | 83.7 ±0.2 | 92.3 ±0.2 | 93.2 ±0.1 | 93.7 ±0.1 | 81.0 ±0.1 | 87.5 ±0.2 | 81.1 ±0.1 | 87.5 ±0.2 |
| RawAtt | 81.5 ±0.1 | 95.7 ±0.1 | 94.9 ±0.1 | 95.4 ±0.1 | - | 90.0 ±0.2 | 84.3 ±0.2 | 90.3 ±0.1 |
| Attention Rollout | 81.9 ±0.2 | 91.9 ±0.2 | 87.4 ±0.2 | 86.2 ±0.2 | - | 89.4 ±0.2 | 74.8 ±0.2 | 85.3 ±0.2 |
| AliLRP | 85.9 ±0.2 | 100.9 ±0.1 | 95.5 ±0.1 | 97.8 ±0.1 | 89.8 ±0.2 | 89.9 ±0.1 | 96.0 ±0.1 | 93.7 ±0.1 |
| AttnLRP | 91.3 ±0.2 | 101.9 ±0.1 | 96.8 ±0.1 | 90.9 ±0.2 | 95.0 ±0.1 | 96.1 ±0.1 | 92.2 ±0.2 | 94.9 ±0.2 |
| DecompX | 89.3 ±0.2 | 99.5 ±0.1 | 96.3 ±0.1 | 99.6 ±0.1 | 90.5 ±0.2 | 94.2 ±0.2 | 93.4 ±0.1 | 94.7 ±0.1 |
| Integrated Gradients | 88.4 ±0.2 | 99.8 ±0.2 | 96.5 ±0.2 | 91.9 ±0.2 | 91.2 ±0.2 | 85.7 ±0.1 | 84.9 ±0.2 | 91.2 ±0.2 |
| Input × Grad | 86.7 ±0.1 | 98.9 ±0.2 | 93.5 ±0.1 | 91.4 ±0.2 | 87.6 ±0.2 | 84.9 ±0.1 | 87.8 ±0.2 | 90.1 ±0.2 |
| **Libra Input × Grad** | 90.2 ±0.2 | 102.5 ±0.1 | 96.9 ±0.1 | 100.4 ±0.1 | 90.6 ±0.2 | 94.7 ±0.2 | 94.0 ±0.2 | 95.6 ±0.1 |
| AttCAT | 92.6 ±0.2 | **105.3** ±0.1 | **100.0** ±0.1 | **104.5** ±0.2 | 92.4 ±0.2 | 94.8 ±0.2 | 90.6 ±0.2 | 97.2 ±0.2 |
| **Libra AttCAT** | <u>97.1</u> ±0.2 | 102.8 ±0.1 | 97.9 ±0.1 | 103.0 ±0.1 | **98.4** ±0.1 | <u>96.4</u> ±0.1 | <u>98.6</u> ±0.1 | <u>99.2</u> ±0.1 |
| GenAtt | 91.5 ±0.2 | 94.0 ±0.2 | 94.4 ±0.1 | 99.7 ±0.1 | - | 94.3 ±0.2 | 93.6 ±0.1 | 94.6 ±0.2 |
| **Libra GenAtt** | 92.0 ±0.2 | 94.5 ±0.2 | 94.3 ±0.2 | 99.4 ±0.1 | - | 94.5 ±0.1 | 94.0 ±0.1 | 94.8 ±0.2 |
| TokenTM | 90.3 ±0.2 | 95.2 ±0.1 | 95.5 ±0.1 | 98.5 ±0.1 | - | 93.4 ±0.1 | 93.0 ±0.1 | 94.3 ±0.1 |
| **Libra TokenTM** | 90.8 ±0.2 | 94.8 ±0.2 | 94.6 ±0.1 | 98.6 ±0.1 | - | 93.6 ±0.1 | 93.3 ±0.2 | 94.3 ±0.1 |
| GradCAM+ | 91.0 ±0.2 | 98.6 ±0.2 | 97.1 ±0.1 | 85.7 ±0.2 | 60.7 ±0.3 | 85.6 ±0.2 | 84.4 ±0.2 | 86.2 ±0.2 |
| **Libra GradCAM+** | 92.7 ±0.2 | 100.3 ±0.1 | 95.6 ±0.1 | 93.8 ±0.1 | 80.8 ±0.2 | 86.4 ±0.2 | 89.8 ±0.2 | 91.3 ±0.2 |
| HiResCAM | 74.2 ±0.3 | 101.0 ±0.1 | 97.0 ±0.1 | 91.2 ±0.2 | 66.3 ±0.3 | 92.7 ±0.2 | 84.2 ±0.1 | 86.6 ±0.2 |
| **Libra HiResCAM** | 88.0 ±0.2 | 100.5 ±0.1 | 97.1 ±0.1 | 101.6 ±0.1 | 82.6 ±0.2 | 94.1 ±0.1 | 86.9 ±0.2 | 93.0 ±0.2 |
| XGradCAM+ | 90.1 ±0.2 | 99.7 ±0.2 | 96.4 ±0.1 | 84.7 ±0.3 | 75.2 ±0.3 | 90.6 ±0.2 | 87.3 ±0.2 | 89.2 ±0.2 |
| **Libra XGradCAM+** | 93.7 ±0.2 | 100.3 ±0.1 | 96.6 ±0.1 | 99.0 ±0.1 | 88.4 ±0.2 | 93.9 ±0.2 | 95.3 ±0.1 | 95.3 ±0.1 |
| FullGrad+ | 91.8 ±0.2 | <u>104.5</u> ±0.2 | <u>98.0</u> ±0.1 | <u>103.2</u> ±0.2 | 89.3 ±0.2 | 93.4 ±0.2 | 90.1 ±0.2 | 95.8 ±0.2 |
| **Libra FullGrad+** | **97.4** ±0.2 | <u>103.0</u> ±0.1 | <u>98.0</u> ±0.1 | <u>103.0</u> ±0.1 | <u>98.2</u> ±0.1 | **96.8** ±0.2 | **98.8** ±0.1 | **99.3** ±0.1 |

Table 16. Least-Influential-First Deletion (LIF) AOPC evaluated using predicted labels across multiple models.

| Method | ViT-L | EVA2-S | BEiT2-L | FlexiViT-L | SigLIP-L | CLIP-H | DeiT3-H | *Avg.* |
|---|---|---|---|---|---|---|---|---|
| Random | 85.4 ±0.2 | 93.5 ±0.2 | 94.1 ±0.1 | 94.7 ±0.2 | 82.7 ±0.2 | 88.8 ±0.2 | 83.5 ±0.2 | 89.0 ±0.2 |
| RawAtt | 83.3 ±0.2 | 96.9 ±0.1 | 95.8 ±0.1 | 96.6 ±0.1 | - | 91.1 ±0.1 | 86.3 ±0.2 | 91.7 ±0.1 |
| Attention Rollout | 84.1 ±0.2 | 93.6 ±0.2 | 89.1 ±0.2 | 88.2 ±0.2 | - | 90.7 ±0.2 | 77.9 ±0.3 | 87.3 ±0.2 |
| AliLRP | 87.7 ±0.2 | 102.6 ±0.2 | 96.7 ±0.1 | 98.9 ±0.1 | 91.2 ±0.2 | 91.3 ±0.1 | 97.5 ±0.2 | 95.1 ±0.2 |
| AttnLRP | 92.9 ±0.2 | 103.0 ±0.1 | 97.8 ±0.1 | 92.2 ±0.2 | 96.0 ±0.2 | 97.3 ±0.2 | 94.2 ±0.2 | 96.2 ±0.2 |
| DecompX | 91.0 ±0.2 | 100.4 ±0.1 | 97.2 ±0.1 | 100.6 ±0.1 | 91.8 ±0.2 | 95.4 ±0.2 | 95.2 ±0.2 | 95.9 ±0.2 |
| Integrated Gradients | 93.3 ±0.2 | 105.3 ±0.2 | 98.8 ±0.1 | 93.8 ±0.2 | 97.0 ±0.3 | 87.2 ±0.1 | 91.2 ±0.3 | 95.2 ±0.2 |
| Input × Grad | 88.4 ±0.2 | 100.0 ±0.2 | 94.5 ±0.1 | 92.8 ±0.2 | 89.7 ±0.3 | 86.5 ±0.2 | 90.5 ±0.2 | 91.8 ±0.2 |
| **Libra Input × Grad** | 92.0 ±0.2 | 104.7 ±0.1 | 98.1 ±0.1 | 101.6 ±0.1 | 91.9 ±0.3 | 96.3 ±0.2 | 95.8 ±0.2 | 97.2 ±0.2 |
| AttCAT | 94.3 ±0.2 | **107.2** ±0.2 | **101.1** ±0.1 | **106.1** ±0.2 | 95.3 ±0.3 | 96.3 ±0.2 | 93.6 ±0.2 | 99.1 ±0.2 |
| **Libra AttCAT** | <u>98.5</u> ±0.2 | 105.0 ±0.1 | 99.2 ±0.1 | 104.4 ±0.1 | **99.6** ±0.2 | <u>98.1</u> ±0.1 | <u>100.0</u> ±0.2 | <u>100.7</u> ±0.2 |
| GenAtt | 92.8 ±0.2 | 95.3 ±0.2 | 95.2 ±0.1 | 100.8 ±0.1 | - | 95.3 ±0.2 | 94.9 ±0.2 | 95.7 ±0.1 |
| **Libra GenAtt** | 93.2 ±0.2 | 95.9 ±0.2 | 95.2 ±0.2 | 100.5 ±0.1 | - | 95.5 ±0.1 | 95.3 ±0.2 | 95.9 ±0.1 |
| TokenTM | 91.6 ±0.2 | 96.6 ±0.2 | 96.2 ±0.1 | 99.6 ±0.1 | - | 94.5 ±0.1 | 94.2 ±0.2 | 95.4 ±0.1 |
| **Libra TokenTM** | 92.1 ±0.2 | 96.3 ±0.2 | 95.5 ±0.1 | 99.6 ±0.1 | - | 94.6 ±0.1 | 94.6 ±0.2 | 95.4 ±0.1 |
| GradCAM+ | 92.9 ±0.2 | 100.8 ±0.3 | 98.5 ±0.2 | 87.5 ±0.2 | 64.3 ±0.4 | 87.4 ±0.2 | 86.7 ±0.2 | 88.3 ±0.2 |
| **Libra GradCAM+** | 94.4 ±0.2 | 102.6 ±0.1 | 97.1 ±0.1 | 95.5 ±0.1 | 83.0 ±0.3 | 88.4 ±0.3 | 91.9 ±0.2 | 93.3 ±0.2 |
| HiResCAM | 76.7 ±0.2 | 103.1 ±0.2 | 98.3 ±0.1 | 92.6 ±0.2 | 69.4 ±0.4 | 94.3 ±0.2 | 86.2 ±0.2 | 88.7 ±0.2 |
| **Libra HiResCAM** | 90.0 ±0.2 | 102.4 ±0.2 | 98.4 ±0.1 | 103.4 ±0.1 | 84.8 ±0.3 | 95.6 ±0.1 | 89.8 ±0.3 | 94.9 ±0.2 |
| XGradCAM+ | 92.1 ±0.2 | 101.9 ±0.3 | 97.9 ±0.2 | 86.6 ±0.3 | 78.4 ±0.4 | 92.1 ±0.2 | 89.9 ±0.3 | 91.3 ±0.3 |
| **Libra XGradCAM+** | 95.3 ±0.2 | 102.6 ±0.1 | 98.0 ±0.1 | 100.4 ±0.1 | 89.8 ±0.3 | 95.8 ±0.1 | 97.0 ±0.2 | 97.0 ±0.2 |
| FullGrad+ | 93.8 ±0.2 | <u>106.6</u> ±0.3 | 98.9 ±0.1 | 104.3 ±0.2 | 92.2 ±0.3 | 95.0 ±0.2 | 92.8 ±0.2 | 97.7 ±0.2 |
| **Libra FullGrad+** | **98.9** ±0.2 | <u>105.3</u> ±0.2 | <u>99.3</u> ±0.1 | <u>104.5</u> ±0.1 | <u>99.4</u> ±0.2 | **98.4** ±0.1 | **100.4** ±0.2 | **100.9** ±0.2 |

Table 17. Least-Influential-First Deletion (LIF) AOPC evaluated using ground-truth labels across multiple models.

| Method | ViT-L | EVA2-S | BEiT2-L | FlexiViT-L | SigLIP-L | CLIP-H | DeiT3-H | *Avg.* |
|---|---|---|---|---|---|---|---|---|
| Random | 49.8 ±0.1 | 50.1 ±0.1 | 50.0 ±0.1 | 49.9 ±0.1 | 50.0 ±0.1 | 50.2 ±0.1 | 49.8 ±0.1 | 50.0 ±0.1 |
| RawAtt | 53.3 ±0.1 | 66.8 ±0.1 | 56.6 ±0.1 | 62.1 ±0.1 | - | 59.2 ±0.1 | 65.2 ±0.1 | 60.6 ±0.1 |
| Attention Rollout | 49.9 ±0.1 | 59.9 ±0.1 | 47.7 ±0.1 | 48.0 ±0.1 | - | 58.0 ±0.1 | 47.8 ±0.1 | 51.9 ±0.1 |
| AliLRP | 52.8 ±0.1 | 67.6 ±0.1 | 55.3 ±0.1 | 54.8 ±0.1 | 66.2 ±0.1 | 54.8 ±0.1 | 71.2 ±0.1 | 60.4 ±0.1 |
| AttnLRP | 59.7 ±0.1 | 75.5 ±0.1 | 61.7 ±0.1 | 49.9 ±0.1 | 72.2 ±0.1 | 65.0 ±0.1 | 60.8 ±0.1 | 63.5 ±0.1 |
| DecompX | 57.4 ±0.1 | 66.3 ±0.1 | 58.3 ±0.1 | 60.9 ±0.1 | 64.6 ±0.1 | 62.1 ±0.1 | 65.1 ±0.1 | 62.1 ±0.1 |
| Integrated Gradients | 54.7 ±0.1 | 59.2 ±0.1 | 53.7 ±0.1 | 50.0 ±0.1 | 59.8 ±0.1 | 50.2 ±0.1 | 53.9 ±0.1 | 54.5 ±0.1 |
| Input × Grad | 53.3 ±0.1 | 58.2 ±0.1 | 51.9 ±0.1 | 49.1 ±0.1 | 56.2 ±0.1 | 50.1 ±0.1 | 56.4 ±0.1 | 53.6 ±0.1 |
| **Libra Input × Grad** | 58.6 ±0.1 | 76.2 ±0.1 | 59.3 ±0.1 | 61.6 ±0.1 | 64.7 ±0.1 | 62.6 ±0.1 | 65.7 ±0.1 | 64.1 ±0.1 |
| AttCAT | 61.2 ±0.1 | 71.0 ±0.1 | 60.7 ±0.1 | 65.1 ±0.1 | 61.2 ±0.1 | 58.0 ±0.1 | 62.3 ±0.1 | 62.8 ±0.1 |
| **Libra AttCAT** | <u>71.8</u> ±0.1 | <u>78.9</u> ±0.1 | <u>67.9</u> ±0.1 | <u>73.4</u> ±0.1 | **81.6** ±0.1 | <u>71.2</u> ±0.1 | <u>79.1</u> ±0.1 | <u>74.9</u> ±0.1 |
| GenAtt | 65.0 ±0.1 | 60.7 ±0.1 | 57.0 ±0.1 | 70.0 ±0.1 | - | 65.9 ±0.1 | 75.2 ±0.1 | 65.6 ±0.1 |
| **Libra GenAtt** | 67.1 ±0.1 | 61.9 ±0.1 | 58.1 ±0.1 | 70.4 ±0.1 | - | 70.3 ±0.1 | 76.2 ±0.1 | 67.3 ±0.1 |
| TokenTM | 63.6 ±0.1 | 63.4 ±0.1 | 62.1 ±0.1 | 67.6 ±0.1 | - | 66.2 ±0.1 | 74.2 ±0.1 | 66.2 ±0.1 |
| **Libra TokenTM** | 65.2 ±0.1 | 63.9 ±0.1 | 61.0 ±0.1 | 68.4 ±0.1 | - | 69.5 ±0.1 | 75.3 ±0.1 | 67.2 ±0.1 |
| GradCAM+ | 62.7 ±0.1 | 65.0 ±0.1 | 59.2 ±0.1 | 50.5 ±0.1 | 46.2 ±0.1 | 51.2 ±0.1 | 60.1 ±0.1 | 56.4 ±0.1 |
| **Libra GradCAM+** | 67.8 ±0.1 | 76.7 ±0.1 | 60.9 ±0.1 | 57.2 ±0.1 | 57.4 ±0.1 | 53.5 ±0.1 | 64.7 ±0.1 | 62.6 ±0.1 |
| HiResCAM | 42.5 ±0.1 | 72.6 ±0.1 | 60.6 ±0.1 | 51.2 ±0.1 | 41.7 ±0.1 | 57.8 ±0.1 | 50.7 ±0.1 | 53.9 ±0.1 |
| **Libra HiResCAM** | 61.4 ±0.1 | 74.5 ±0.1 | 61.6 ±0.1 | 71.9 ±0.1 | 57.2 ±0.1 | 64.9 ±0.1 | 67.2 ±0.1 | 65.5 ±0.1 |
| XGradCAM+ | 60.8 ±0.1 | 67.0 ±0.1 | 57.4 ±0.1 | 49.4 ±0.1 | 56.0 ±0.1 | 57.4 ±0.1 | 61.1 ±0.1 | 58.5 ±0.1 |
| **Libra XGradCAM+** | 69.4 ±0.1 | 78.0 ±0.1 | 65.6 ±0.1 | 64.8 ±0.1 | 70.0 ±0.1 | 69.3 ±0.1 | 76.3 ±0.1 | 70.5 ±0.1 |
| FullGrad+ | 60.9 ±0.1 | 67.4 ±0.1 | 57.5 ±0.1 | 63.3 ±0.1 | 58.4 ±0.1 | 56.9 ±0.1 | 60.9 ±0.1 | 60.8 ±0.1 |
| **Libra FullGrad+** | **72.5** ±0.1 | **80.1** ±0.1 | **68.5** ±0.1 | **73.7** ±0.1 | <u>79.6</u> ±0.1 | **73.0** ±0.1 | **79.6** ±0.1 | **75.3** ±0.1 |

Table 18. Symmetric Relevance Gain (SRG) Accuracy evaluated using predicted labels across multiple models.

| Method | ViT-L | EVA2-S | BEiT2-L | FlexiViT-L | SigLIP-L | CLIP-H | DeiT3-H | *Avg.* |
|---|---|---|---|---|---|---|---|---|
| Random | 49.9 ±0.1 | 50.0 ±0.1 | 49.8 ±0.1 | 49.8 ±0.1 | 50.0 ±0.1 | 50.0 ±0.1 | 49.9 ±0.1 | 49.9 ±0.1 |
| RawAtt | 52.9 ±0.1 | 64.3 ±0.1 | 55.4 ±0.1 | 60.1 ±0.1 | - | 57.8 ±0.1 | 63.5 ±0.1 | 59.0 ±0.1 |
| Attention Rollout | 50.4 ±0.1 | 58.5 ±0.1 | 47.9 ±0.1 | 48.3 ±0.1 | - | 57.2 ±0.1 | 48.1 ±0.1 | 51.8 ±0.1 |
| AliLRP | 52.6 ±0.1 | 66.2 ±0.1 | 55.0 ±0.1 | 54.1 ±0.1 | 64.8 ±0.1 | 54.5 ±0.1 | 69.6 ±0.1 | 59.5 ±0.1 |
| AttnLRP | 58.7 ±0.1 | 72.7 ±0.1 | 60.3 ±0.1 | 49.6 ±0.1 | 69.8 ±0.1 | 63.8 ±0.1 | 59.7 ±0.1 | 62.1 ±0.1 |
| DecompX | 56.6 ±0.1 | 64.0 ±0.1 | 57.0 ±0.1 | 59.3 ±0.1 | 62.9 ±0.1 | 60.8 ±0.1 | 63.7 ±0.1 | 60.6 ±0.1 |
| Integrated Gradients | 58.7 ±0.1 | 64.1 ±0.1 | 55.6 ±0.1 | 50.8 ±0.1 | 64.3 ±0.1 | 50.2 ±0.1 | 57.6 ±0.1 | 57.3 ±0.1 |
| Input × Grad | 53.0 ±0.1 | 57.2 ±0.1 | 51.9 ±0.1 | 49.0 ±0.1 | 55.9 ±0.1 | 50.1 ±0.1 | 56.1 ±0.1 | 53.3 ±0.1 |
| **Libra Input × Grad** | 58.0 ±0.1 | 74.5 ±0.1 | 58.4 ±0.1 | 60.4 ±0.1 | 63.3 ±0.1 | 61.7 ±0.1 | 64.6 ±0.1 | 63.0 ±0.1 |
| AttCAT | 60.2 ±0.1 | 69.8 ±0.1 | 60.1 ±0.1 | 64.2 ±0.1 | 61.0 ±0.1 | 57.4 ±0.1 | 61.8 ±0.1 | 62.1 ±0.1 |
| **Libra AttCAT** | <u>70.5</u> ±0.1 | <u>77.2</u> ±0.1 | <u>66.6</u> ±0.1 | <u>71.6</u> ±0.1 | **79.5** ±0.1 | <u>69.8</u> ±0.1 | <u>77.1</u> ±0.1 | <u>73.2</u> ±0.1 |
| GenAtt | 63.2 ±0.1 | 59.1 ±0.1 | 55.6 ±0.1 | 67.3 ±0.1 | - | 63.9 ±0.1 | 72.1 ±0.1 | 63.5 ±0.1 |
| **Libra GenAtt** | 65.3 ±0.1 | 60.2 ±0.1 | 56.6 ±0.1 | 67.6 ±0.1 | - | 68.0 ±0.1 | 73.2 ±0.1 | 65.2 ±0.1 |
| TokenTM | 61.9 ±0.1 | 61.8 ±0.1 | 60.3 ±0.1 | 65.2 ±0.1 | - | 64.3 ±0.1 | 71.2 ±0.1 | 64.1 ±0.1 |
| **Libra TokenTM** | 63.4 ±0.1 | 62.2 ±0.1 | 59.4 ±0.1 | 65.7 ±0.1 | - | 67.3 ±0.1 | 72.3 ±0.1 | 65.0 ±0.1 |
| GradCAM+ | 62.0 ±0.1 | 64.0 ±0.1 | 58.8 ±0.1 | 50.2 ±0.1 | 46.7 ±0.1 | 51.2 ±0.1 | 59.4 ±0.1 | 56.0 ±0.1 |
| **Libra GradCAM+** | 66.7 ±0.1 | 75.0 ±0.1 | 60.3 ±0.1 | 56.6 ±0.1 | 56.9 ±0.1 | 53.7 ±0.1 | 63.4 ±0.1 | 61.8 ±0.1 |
| HiResCAM | 43.2 ±0.1 | 71.2 ±0.1 | 59.9 ±0.1 | 50.6 ±0.1 | 42.6 ±0.1 | 57.3 ±0.1 | 50.3 ±0.1 | 53.6 ±0.1 |
| **Libra HiResCAM** | 60.7 ±0.1 | 72.6 ±0.1 | 60.8 ±0.1 | 70.4 ±0.1 | 56.8 ±0.1 | 63.8 ±0.1 | 66.3 ±0.1 | 64.5 ±0.1 |
| XGradCAM+ | 60.2 ±0.1 | 66.0 ±0.1 | 57.3 ±0.1 | 49.2 ±0.1 | 56.0 ±0.1 | 56.9 ±0.1 | 60.2 ±0.1 | 58.0 ±0.1 |
| **Libra XGradCAM+** | 68.2 ±0.1 | 76.0 ±0.1 | 64.5 ±0.1 | 63.6 ±0.1 | 68.4 ±0.1 | 68.2 ±0.1 | 74.2 ±0.1 | 69.0 ±0.1 |
| FullGrad+ | 60.3 ±0.1 | 66.5 ±0.1 | 57.1 ±0.1 | 62.2 ±0.1 | 58.5 ±0.1 | 56.4 ±0.1 | 60.2 ±0.1 | 60.2 ±0.1 |
| **Libra FullGrad+** | **71.2** ±0.1 | **78.3** ±0.1 | **67.1** ±0.1 | **71.9** ±0.1 | <u>77.6</u> ±0.1 | **71.5** ±0.1 | **77.6** ±0.1 | **73.6** ±0.1 |

Table 19. Symmetric Relevance Gain (SRG) Accuracy evaluated using ground-truth labels across multiple models.

| Method | ViT-L | EVA2-S | BEiT2-L | FlexiViT-L | SigLIP-L | CLIP-H | DeiT3-H | *Avg.* |
|---|---|---|---|---|---|---|---|---|
| Random | 49.8 ±0.2 | 50.2 ±0.2 | 50.0 ±0.1 | 50.0 ±0.2 | 50.0 ±0.2 | 50.1 ±0.2 | 50.1 ±0.2 | 50.0 ±0.2 |
| RawAtt | 53.4 ±0.2 | 64.8 ±0.2 | 56.2 ±0.2 | 60.9 ±0.2 | - | 56.7 ±0.2 | 60.7 ±0.2 | 58.8 ±0.2 |
| Attention Rollout | 50.1 ±0.2 | 58.4 ±0.3 | 48.0 ±0.2 | 48.0 ±0.2 | - | 55.9 ±0.2 | 48.3 ±0.2 | 51.5 ±0.2 |
| AliLRP | 52.5 ±0.2 | 66.1 ±0.2 | 54.7 ±0.1 | 54.2 ±0.1 | 64.9 ±0.2 | 53.6 ±0.2 | 68.8 ±0.2 | 59.3 ±0.2 |
| AttnLRP | 59.5 ±0.2 | 73.1 ±0.2 | 60.9 ±0.2 | 49.6 ±0.2 | 70.6 ±0.2 | 61.3 ±0.2 | 62.0 ±0.2 | 62.4 ±0.2 |
| DecompX | 57.3 ±0.2 | 65.1 ±0.2 | 57.8 ±0.2 | 60.1 ±0.1 | 63.1 ±0.2 | 58.9 ±0.2 | 64.7 ±0.2 | 61.0 ±0.2 |
| Integrated Gradients | 55.1 ±0.2 | 59.6 ±0.2 | 54.2 ±0.2 | 50.7 ±0.2 | 60.0 ±0.2 | 50.3 ±0.1 | 53.8 ±0.2 | 54.8 ±0.2 |
| Input × Grad | 53.2 ±0.2 | 57.9 ±0.2 | 51.9 ±0.1 | 48.9 ±0.2 | 56.8 ±0.2 | 50.1 ±0.1 | 56.4 ±0.2 | 53.6 ±0.2 |
| **Libra Input × Grad** | 58.2 ±0.2 | 73.4 ±0.2 | 58.5 ±0.2 | 60.8 ±0.1 | 63.1 ±0.2 | 59.4 ±0.2 | 65.1 ±0.2 | 62.7 ±0.2 |
| AttCAT | 60.8 ±0.2 | 70.3 ±0.2 | 60.5 ±0.1 | 63.5 ±0.3 | 61.7 ±0.2 | 58.1 ±0.2 | 61.5 ±0.2 | 62.3 ±0.2 |
| **Libra AttCAT** | 70.8 ±0.2 | 75.7 ±0.2 | 66.3 ±0.2 | 71.3 ±0.2 | **79.0** ±0.2 | 65.5 ±0.2 | 75.7 ±0.2 | 72.0 ±0.2 |
| GenAtt | 64.0 ±0.2 | 59.1 ±0.2 | 56.8 ±0.2 | 67.4 ±0.2 | - | 62.0 ±0.2 | 70.8 ±0.2 | 63.3 ±0.2 |
| **Libra GenAtt** | 65.8 ±0.3 | 60.0 ±0.2 | 57.5 ±0.2 | 68.0 ±0.2 | - | 64.5 ±0.2 | 71.9 ±0.2 | 64.6 ±0.2 |
| TokenTM | 62.6 ±0.2 | 61.7 ±0.2 | 61.2 ±0.2 | 65.6 ±0.2 | - | 61.8 ±0.2 | 70.1 ±0.2 | 63.8 ±0.2 |
| **Libra TokenTM** | 64.1 ±0.3 | 61.8 ±0.2 | 60.0 ±0.2 | 66.5 ±0.2 | - | 63.9 ±0.2 | 71.0 ±0.2 | 64.6 ±0.2 |
| GradCAM+ | 62.0 ±0.2 | 63.8 ±0.2 | 58.6 ±0.2 | 49.4 ±0.2 | 44.4 ±0.3 | 50.9 ±0.2 | 58.1 ±0.2 | 55.3 ±0.2 |
| **Libra GradCAM+** | 66.4 ±0.2 | 73.2 ±0.2 | 59.9 ±0.2 | 56.1 ±0.2 | 56.4 ±0.3 | 52.5 ±0.2 | 62.6 ±0.2 | 61.0 ±0.2 |
| HiResCAM | 43.2 ±0.2 | 70.6 ±0.2 | 59.6 ±0.2 | 50.1 ±0.2 | 41.9 ±0.2 | 56.2 ±0.2 | 50.8 ±0.2 | 53.2 ±0.2 |
| **Libra HiResCAM** | 60.6 ±0.2 | 71.7 ±0.2 | 60.4 ±0.2 | 69.8 ±0.2 | 56.5 ±0.2 | 61.0 ±0.2 | 63.3 ±0.2 | 63.3 ±0.2 |
| XGradCAM+ | 60.3 ±0.2 | 65.8 ±0.2 | 57.2 ±0.1 | 47.3 ±0.3 | 56.5 ±0.2 | 56.0 ±0.2 | 59.6 ±0.2 | 57.5 ±0.2 |
| **Libra XGradCAM+** | 68.0 ±0.3 | 74.2 ±0.2 | 64.0 ±0.2 | 63.1 ±0.2 | 67.3 ±0.2 | 64.0 ±0.2 | 72.2 ±0.2 | 67.5 ±0.2 |
| FullGrad+ | 60.4 ±0.2 | 67.2 ±0.2 | 57.3 ±0.2 | 62.0 ±0.2 | 59.2 ±0.2 | 57.0 ±0.2 | 60.1 ±0.3 | 60.4 ±0.2 |
| **Libra FullGrad+** | **71.3** ±0.2 | **76.8** ±0.2 | **66.8** ±0.2 | **71.4** ±0.2 | 76.7 ±0.2 | **66.8** ±0.2 | **76.3** ±0.2 | **72.3** ±0.2 |

Table 20. Symmetric Relevance Gain (SRG) AOPC evaluated using predicted labels across multiple models.

| Method | ViT-L | EVA2-S | BEiT2-L | FlexiViT-L | SigLIP-L | CLIP-H | DeiT3-H | *Avg.* |
|---|---|---|---|---|---|---|---|---|
| Random | 49.7 ±0.2 | 50.0 ±0.2 | 49.8 ±0.1 | 50.0 ±0.2 | 50.0 ±0.2 | 50.0 ±0.2 | 50.0 ±0.2 | 49.9 ±0.2 |
| RawAtt | 53.1 ±0.2 | 63.6 ±0.2 | 55.6 ±0.2 | 60.1 ±0.2 | - | 56.1 ±0.2 | 59.8 ±0.2 | 58.0 ±0.2 |
| Attention Rollout | 50.3 ±0.3 | 57.8 ±0.3 | 48.1 ±0.2 | 48.2 ±0.2 | - | 55.6 ±0.3 | 48.4 ±0.3 | 51.4 ±0.2 |
| AliLRP | 52.4 ±0.2 | 65.1 ±0.3 | 54.6 ±0.2 | 53.9 ±0.1 | 63.9 ±0.3 | 53.5 ±0.2 | 67.4 ±0.2 | 58.7 ±0.2 |
| AttnLRP | 58.8 ±0.3 | 71.3 ±0.2 | 60.2 ±0.2 | 49.4 ±0.2 | 69.2 ±0.3 | 60.6 ±0.2 | 61.2 ±0.3 | 61.5 ±0.2 |
| DecompX | 56.8 ±0.3 | 63.7 ±0.3 | 57.3 ±0.2 | 59.3 ±0.2 | 62.2 ±0.2 | 58.3 ±0.2 | 63.7 ±0.2 | 60.2 ±0.2 |
| Integrated Gradients | 58.2 ±0.3 | 63.1 ±0.2 | 55.7 ±0.2 | 51.1 ±0.2 | 63.5 ±0.3 | 50.3 ±0.2 | 58.1 ±0.3 | 57.1 ±0.2 |
| Input × Grad | 53.0 ±0.2 | 57.1 ±0.2 | 51.7 ±0.1 | 49.0 ±0.2 | 56.4 ±0.3 | 50.1 ±0.2 | 56.2 ±0.2 | 53.3 ±0.2 |
| **Libra Input × Grad** | 57.7 ±0.3 | 72.2 ±0.3 | 58.1 ±0.2 | 60.1 ±0.2 | 62.2 ±0.3 | 59.0 ±0.2 | 64.1 ±0.2 | 61.9 ±0.2 |
| AttCAT | 60.0 ±0.2 | 69.0 ±0.2 | 60.0 ±0.2 | 62.5 ±0.3 | 61.4 ±0.3 | 57.5 ±0.2 | 61.1 ±0.3 | 61.6 ±0.2 |
| **Libra AttCAT** | 69.5 ±0.3 | 74.4 ±0.2 | 65.4 ±0.2 | 70.0 ±0.2 | **77.3** ±0.3 | 64.9 ±0.2 | 73.8 ±0.2 | 70.8 ±0.2 |
| GenAtt | 63.0 ±0.2 | 58.2 ±0.2 | 56.1 ±0.2 | 66.1 ±0.2 | - | 61.1 ±0.2 | 69.1 ±0.2 | 62.3 ±0.2 |
| **Libra GenAtt** | 64.7 ±0.3 | 59.2 ±0.2 | 56.8 ±0.2 | 66.7 ±0.2 | - | 63.5 ±0.2 | 70.2 ±0.2 | 63.5 ±0.2 |
| TokenTM | 61.7 ±0.3 | 60.9 ±0.2 | 60.3 ±0.2 | 64.4 ±0.2 | - | 60.9 ±0.2 | 68.4 ±0.2 | 62.8 ±0.2 |
| **Libra TokenTM** | 63.1 ±0.3 | 61.0 ±0.3 | 59.3 ±0.3 | 65.3 ±0.2 | - | 62.9 ±0.2 | 69.3 ±0.2 | 63.5 ±0.2 |
| GradCAM+ | 61.5 ±0.3 | 62.9 ±0.3 | 58.3 ±0.2 | 49.2 ±0.2 | 44.9 ±0.3 | 51.0 ±0.2 | 57.5 ±0.2 | 55.0 ±0.2 |
| **Libra GradCAM+** | 65.5 ±0.3 | 72.0 ±0.2 | 59.6 ±0.2 | 55.6 ±0.2 | 55.9 ±0.3 | 52.6 ±0.2 | 61.6 ±0.2 | 60.4 ±0.2 |
| HiResCAM | 43.6 ±0.2 | 69.6 ±0.2 | 59.2 ±0.2 | 49.9 ±0.2 | 42.5 ±0.3 | 55.9 ±0.2 | 50.6 ±0.2 | 53.1 ±0.2 |
| **Libra HiResCAM** | 60.1 ±0.2 | 70.5 ±0.2 | 59.8 ±0.2 | 68.8 ±0.3 | 56.1 ±0.3 | 60.5 ±0.2 | 62.6 ±0.3 | 62.6 ±0.2 |
| XGradCAM+ | 59.9 ±0.3 | 64.9 ±0.3 | 56.9 ±0.2 | 47.2 ±0.3 | 56.4 ±0.3 | 55.7 ±0.2 | 58.9 ±0.3 | 57.1 ±0.3 |
| **Libra XGradCAM+** | 66.9 ±0.3 | 72.9 ±0.2 | 63.3 ±0.2 | 62.3 ±0.2 | 66.2 ±0.3 | 63.4 ±0.2 | 70.6 ±0.2 | 66.5 ±0.2 |
| FullGrad+ | 59.8 ±0.2 | 66.2 ±0.3 | 56.9 ±0.2 | 60.9 ±0.3 | 59.0 ±0.3 | 56.5 ±0.2 | 59.5 ±0.3 | 59.8 ±0.2 |
| **Libra FullGrad+** | **70.0** ±0.3 | **75.4** ±0.3 | **65.8** ±0.2 | **70.1** ±0.2 | 75.0 ±0.3 | **66.0** ±0.2 | **74.4** ±0.2 | **71.0** ±0.2 |

Table 21. Symmetric Relevance Gain (SRG) AOPC evaluated using ground-truth labels across multiple models.

### D.2.1. Segmentation Average Precision (AP)

| Method | ViT-L | EVA2-S | BEiT2-L | FlexiViT-L | SigLIP-L | CLIP-H | DeiT3-H | *Avg.* |
|---|---|---|---|---|---|---|---|---|
| Random | 42.0 ±0.4 | 37.7 ±0.3 | 39.8 ±0.4 | 39.8 ±0.4 | 33.0 ±0.3 | 37.8 ±0.3 | 37.8 ±0.3 | 38.3 ±0.3 |
| RawAtt | 40.2 ±0.4 | 59.0 ±0.3 | 47.6 ±0.3 | 49.8 ±0.3 | - | 41.6 ±0.3 | 49.7 ±0.3 | 48.0 ±0.3 |
| Attention Rollout | 39.9 ±0.3 | 45.3 ±0.3 | 42.2 ±0.3 | 42.2 ±0.3 | - | 51.7 ±0.4 | 34.1 ±0.3 | 42.6 ±0.3 |
| AliLRP | 42.7 ±0.4 | 58.7 ±0.3 | 43.9 ±0.3 | 49.6 ±0.3 | 33.5 ±0.3 | 38.1 ±0.3 | 52.2 ±0.3 | 45.5 ±0.3 |
| AttnLRP | 47.2 ±0.3 | 73.1 ±0.2 | 66.0 ±0.3 | 43.4 ±0.4 | 36.0 ±0.3 | 50.9 ±0.3 | 36.0 ±0.3 | 50.4 ±0.3 |
| DecompX | 54.2 ±0.3 | 60.0 ±0.3 | 55.6 ±0.3 | 59.2 ±0.3 | 40.5 ±0.3 | 55.0 ±0.3 | 49.5 ±0.3 | 53.4 ±0.3 |
| Integrated Gradients | 46.6 ±0.3 | 51.2 ±0.3 | 46.7 ±0.3 | 41.3 ±0.4 | 41.6 ±0.3 | 36.9 ±0.3 | 38.9 ±0.3 | 43.3 ±0.3 |
| Input $\times$ Grad | 43.6 ±0.4 | 42.5 ±0.3 | 39.6 ±0.4 | 41.4 ±0.4 | 35.5 ±0.3 | 36.8 ±0.3 | 39.6 ±0.3 | 39.9 ±0.3 |
| **Libra Input $\times$ Grad** | 53.6 ±0.3 | 72.1 ±0.3 | 54.8 ±0.3 | 60.4 ±0.3 | 39.9 ±0.3 | 54.2 ±0.3 | 49.0 ±0.3 | 54.8 ±0.3 |
| AttCAT | 44.9 ±0.3 | 58.9 ±0.3 | 52.2 ±0.3 | 45.1 ±0.3 | 37.6 ±0.3 | 38.9 ±0.3 | 41.7 ±0.3 | 45.6 ±0.3 |
| **Libra AttCAT** | 53.3 ±0.3 | 75.1 ±0.3 | 65.5 ±0.3 | 74.4 ±0.3 | 46.8 ±0.3 | 61.7 ±0.3 | 60.1 ±0.3 | 62.4 ±0.3 |
| GenAtt | 50.9 ±0.3 | 42.3 ±0.3 | 47.9 ±0.3 | 75.1 ±0.2 | - | 55.9 ±0.3 | 66.2 ±0.2 | 56.4 ±0.3 |
| **Libra GenAtt** | 58.6 ±0.3 | 44.3 ±0.3 | 48.8 ±0.3 | <u>79.4</u> ±0.2 | - | **76.2** ±0.2 | **76.5** ±0.2 | 64.0 ±0.3 |
| TokenTM | 50.0 ±0.3 | 45.5 ±0.3 | 56.0 ±0.3 | 72.2 ±0.2 | - | 58.6 ±0.3 | 61.7 ±0.2 | 57.3 ±0.3 |
| **Libra TokenTM** | 53.9 ±0.3 | 46.7 ±0.3 | 54.2 ±0.3 | 76.2 ±0.2 | - | 71.5 ±0.3 | 70.8 ±0.2 | 62.2 ±0.3 |
| GradCAM+ | 52.1 ±0.4 | 49.3 ±0.4 | 53.5 ±0.4 | 40.5 ±0.4 | 44.3 ±0.4 | 43.0 ±0.4 | 60.3 ±0.4 | 49.0 ±0.4 |
| **Libra GradCAM+** | 60.2 ±0.4 | <u>79.8</u> ±0.3 | <u>69.4</u> ±0.4 | 50.2 ±0.4 | 41.7 ±0.3 | 47.4 ±0.4 | 46.7 ±0.4 | 56.5 ±0.4 |
| HiResCAM | 38.5 ±0.4 | 73.2 ±0.3 | 60.8 ±0.3 | 43.7 ±0.3 | 36.3 ±0.3 | 45.9 ±0.3 | 41.3 ±0.3 | 48.5 ±0.3 |
| **Libra HiResCAM** | 48.0 ±0.3 | 76.5 ±0.3 | 69.0 ±0.3 | **81.6** ±0.3 | <u>47.5</u> ±0.3 | 56.8 ±0.3 | <u>76.3</u> ±0.3 | <u>65.1</u> ±0.3 |
| XGradCAM+ | 46.9 ±0.4 | 55.2 ±0.4 | 49.0 ±0.4 | 38.5 ±0.4 | 43.0 ±0.3 | 47.7 ±0.4 | 48.9 ±0.4 | 47.0 ±0.4 |
| **Libra XGradCAM+** | <u>60.3</u> ±0.4 | **82.7** ±0.3 | **71.4** ±0.3 | 63.3 ±0.4 | 44.3 ±0.4 | <u>73.3</u> ±0.3 | 59.4 ±0.3 | 65.0 ±0.3 |
| FullGrad+ | 44.2 ±0.3 | 51.5 ±0.3 | 47.4 ±0.3 | 44.1 ±0.3 | 37.7 ±0.3 | 38.5 ±0.3 | 40.6 ±0.3 | 43.4 ±0.3 |
| **Libra FullGrad+** | **64.5** ±0.3 | 79.4 ±0.3 | 67.9 ±0.3 | 75.1 ±0.3 | **51.7** ±0.3 | 71.5 ±0.3 | 65.1 ±0.3 | **67.9** ±0.3 |

Table 22. Segmentation AP for different methods (and their Libra enhancements) across multiple models.

### D.3. Across Model Sizes

| Method | ViT-Tiny | ViT-Small | ViT-Base | ViT-Large | *Avg.* |
|---|---|---|---|---|---|
| Random | 40.5 ±0.1 | 33.8 ±0.1 | 26.5 ±0.1 | 29.5 ±0.1 | 32.6 ±0.1 |
| RawAtt | 69.5 ±0.1 | 58.7 ±0.1 | 44.6 ±0.1 | 39.1 ±0.1 | 53.0 ±0.1 |
| Attention Rollout | 64.1 ±0.1 | 45.1 ±0.1 | 35.4 ±0.1 | 31.4 ±0.1 | 44.0 ±0.1 |
| AliLRP | 64.4 ±0.1 | 42.3 ±0.1 | 33.3 ±0.1 | 33.2 ±0.1 | 43.3 ±0.1 |
| AttnLRP | 69.7 ±0.1 | 52.4 ±0.1 | 38.5 ±0.1 | 41.8 ±0.1 | 50.6 ±0.1 |
| DecompX | 70.4 ±0.1 | 50.4 ±0.1 | 37.8 ±0.1 | 38.9 ±0.1 | 49.4 ±0.1 |
| Integrated Gradients | 57.1 ±0.1 | 46.0 ±0.1 | 35.4 ±0.1 | 35.9 ±0.1 | 43.6 ±0.1 |
| Input × Grad | 55.6 ±0.1 | 41.8 ±0.1 | 34.4 ±0.1 | 33.9 ±0.1 | 41.4 ±0.1 |
| **Libra Input × Grad** | 70.8 ±0.1 (+27.2%) | 49.3 ±0.1 (+18.0%) | 38.6 ±0.1 (+12.0%) | 40.5 ±0.1 (+19.6%) | 49.8 ±0.1 (+20.2%) |
| AttCAT | 69.3 ±0.1 | 58.9 ±0.1 | 46.9 ±0.1 | 44.8 ±0.1 | 55.0 ±0.1 |
| **Libra AttCAT** | <u>81.0</u> ±0.1 (+16.7%) | **70.3** ±0.1 (+19.3%) | <u>63.5</u> ±0.1 (+35.4%) | <u>61.3</u> ±0.1 (+36.9%) | <u>69.0</u> ±0.1 (+25.5%) |
| GenAtt | 77.1 ±0.1 | 66.3 ±0.1 | 58.2 ±0.1 | 51.8 ±0.1 | 63.4 ±0.1 |
| **Libra GenAtt** | 78.4 ±0.1 (+1.7%) | 68.2 ±0.1 (+2.9%) | 61.6 ±0.1 (+5.8%) | 55.4 ±0.1 (+6.8%) | 65.9 ±0.1 (+4.0%) |
| TokenTM | 75.0 ±0.1 | 65.2 ±0.1 | 56.8 ±0.1 | 50.0 ±0.1 | 61.7 ±0.1 |
| **Libra TokenTM** | 76.2 ±0.1 (+1.6%) | 66.5 ±0.1 (+2.0%) | 59.1 ±0.1 (+4.1%) | 52.5 ±0.1 (+5.0%) | 63.6 ±0.1 (+3.0%) |
| GradCAM+ | 66.2 ±0.1 | 55.5 ±0.1 | 45.6 ±0.1 | 48.6 ±0.1 | 54.0 ±0.1 |
| **Libra GradCAM+** | 72.9 ±0.1 (+10.1%) | 66.5 ±0.1 (+19.7%) | 61.4 ±0.1 (+34.8%) | 56.5 ±0.1 (+16.2%) | 64.3 ±0.1 (+19.2%) |
| HiResCAM | 39.0 ±0.1 | 29.5 ±0.1 | 45.4 ±0.1 | 25.7 ±0.1 | 34.9 ±0.1 |
| **Libra HiResCAM** | 69.9 ±0.1 (+79.1%) | 63.4 ±0.1 (+114.7%) | 56.7 ±0.1 (+24.8%) | 49.0 ±0.1 (+90.7%) | 59.7 ±0.1 (+71.1%) |
| XGradCAM+ | 67.5 ±0.1 | 55.9 ±0.1 | 38.6 ±0.1 | 45.9 ±0.1 | 52.0 ±0.1 |
| **Libra XGradCAM+** | 77.0 ±0.1 (+14.1%) | 68.5 ±0.1 (+22.4%) | **63.9** ±0.1 (+65.6%) | 58.8 ±0.1 (+28.1%) | 67.0 ±0.1 (+29.0%) |
| FullGrad+ | 65.9 ±0.1 | 55.8 ±0.1 | 44.2 ±0.1 | 45.1 ±0.1 | 52.7 ±0.1 |
| **Libra FullGrad+** | **81.7** ±0.1 (+24.0%) | <u>70.1</u> ±0.1 (+25.8%) | 63.1 ±0.1 (+42.9%) | **62.4** ±0.1 (+38.5%) | **69.3** ±0.1 (+31.5%) |

Table 23. How Most-Influential-First Deletion (MIF) Accuracy evaluated using predicted labels varies with different model sizes.

| Method | ViT-Tiny | ViT-Small | ViT-Base | ViT-Large | *Avg.* |
|---|---|---|---|---|---|
| Random | 50.1 ±0.1 | 41.8 ±0.1 | 34.5 ±0.1 | 36.9 ±0.1 | 40.8 ±0.1 |
| RawAtt | 74.0 ±0.1 | 63.1 ±0.1 | 50.1 ±0.1 | 45.4 ±0.1 | 58.2 ±0.1 |
| Attention Rollout | 68.7 ±0.1 | 51.2 ±0.1 | 41.9 ±0.1 | 39.0 ±0.1 | 50.2 ±0.1 |
| AliLRP | 68.9 ±0.1 | 48.9 ±0.1 | 39.8 ±0.1 | 39.8 ±0.1 | 49.4 ±0.1 |
| AttnLRP | 73.4 ±0.1 | 57.7 ±0.1 | 44.5 ±0.1 | 47.1 ±0.1 | 55.7 ±0.1 |
| DecompX | 74.0 ±0.1 | 56.0 ±0.1 | 44.0 ±0.1 | 44.4 ±0.1 | 54.6 ±0.1 |
| Integrated Gradients | 69.7 ±0.1 | 56.9 ±0.1 | 46.9 ±0.1 | 46.3 ±0.1 | 54.9 ±0.1 |
| Input × Grad | 61.1 ±0.1 | 47.9 ±0.1 | 40.4 ±0.1 | 40.1 ±0.1 | 47.4 ±0.1 |
| **Libra Input × Grad** | 74.5 ±0.1 (+22.0%) | 54.9 ±0.1 (+14.7%) | 44.8 ±0.1 (+10.8%) | 45.9 ±0.1 (+14.4%) | 55.0 ±0.1 (+16.2%) |
| AttCAT | 72.6 ±0.1 | 62.1 ±0.1 | 50.4 ±0.1 | 48.7 ±0.1 | 58.5 ±0.1 |
| **Libra AttCAT** | <u>83.6</u> ±0.1 (+15.2%) | **73.6** ±0.1 (+18.5%) | <u>66.4</u> ±0.1 (+31.7%) | <u>64.7</u> ±0.1 (+33.0%) | <u>72.1</u> ±0.1 (+23.3%) |
| GenAtt | 80.4 ±0.1 | 69.7 ±0.1 | 61.9 ±0.1 | 56.4 ±0.1 | 67.1 ±0.1 |
| **Libra GenAtt** | 81.6 ±0.1 (+1.4%) | 71.7 ±0.1 (+2.9%) | 65.1 ±0.1 (+5.1%) | 59.7 ±0.1 (+5.9%) | 69.5 ±0.1 (+3.6%) |
| TokenTM | 78.8 ±0.1 | 68.9 ±0.1 | 60.6 ±0.1 | 54.9 ±0.1 | 65.8 ±0.1 |
| **Libra TokenTM** | 79.9 ±0.1 (+1.4%) | 70.3 ±0.1 (+2.1%) | 62.8 ±0.1 (+3.6%) | 57.3 ±0.1 (+4.5%) | 67.6 ±0.1 (+2.7%) |
| GradCAM+ | 70.5 ±0.1 | 59.9 ±0.1 | 50.5 ±0.1 | 53.4 ±0.1 | 58.6 ±0.1 |
| **Libra GradCAM+** | 76.8 ±0.1 (+8.9%) | 70.2 ±0.1 (+17.0%) | 65.3 ±0.1 (+29.3%) | 60.9 ±0.1 (+14.0%) | 68.3 ±0.1 (+16.5%) |
| HiResCAM | 48.0 ±0.1 | 38.4 ±0.1 | 50.4 ±0.1 | 32.7 ±0.1 | 42.4 ±0.1 |
| **Libra HiResCAM** | 74.1 ±0.1 (+54.3%) | 67.4 ±0.1 (+75.5%) | 60.8 ±0.1 (+20.6%) | 54.0 ±0.1 (+65.2%) | 64.1 ±0.1 (+51.2%) |
| XGradCAM+ | 71.7 ±0.1 | 60.3 ±0.1 | 44.0 ±0.1 | 50.9 ±0.1 | 56.7 ±0.1 |
| **Libra XGradCAM+** | 80.6 ±0.1 (+12.4%) | 72.1 ±0.1 (+19.5%) | **67.4** ±0.1 (+53.0%) | 63.0 ±0.1 (+23.6%) | 70.7 ±0.1 (+24.7%) |
| FullGrad+ | 69.8 ±0.1 | 59.6 ±0.1 | 48.2 ±0.1 | 49.1 ±0.1 | 56.6 ±0.1 |
| **Libra FullGrad+** | **84.2** ±0.1 (+20.8%) | <u>73.5</u> ±0.1 (+23.3%) | 66.1 ±0.1 (+37.1%) | **65.5** ±0.1 (+33.5%) | **72.3** ±0.1 (+27.7%) |

Table 24. How Most-Influential-First Deletion (MIF) Accuracy evaluated using ground-truth labels varies with different model sizes.

| Method | ViT-Tiny | ViT-Small | ViT-Base | ViT-Large | *Avg.* |
|---|---|---|---|---|---|
| Random | 20.7 ±0.2 | 18.6 ±0.2 | 14.2 ±0.2 | 15.8 ±0.2 | 17.3 ±0.2 |
| RawAtt | 44.8 ±0.3 | 41.2 ±0.3 | 27.9 ±0.3 | 25.3 ±0.2 | 34.8 ±0.3 |
| Attention Rollout | 39.8 ±0.3 | 28.8 ±0.2 | 21.2 ±0.2 | 18.3 ±0.3 | 27.0 ±0.3 |
| AliLRP | 39.3 ±0.2 | 26.2 ±0.3 | 19.1 ±0.2 | 19.2 ±0.2 | 26.0 ±0.2 |
| AttnLRP | 44.3 ±0.3 | 35.2 ±0.2 | 23.4 ±0.2 | 27.6 ±0.3 | 32.6 ±0.3 |
| DecompX | 44.8 ±0.3 | 33.6 ±0.2 | 22.8 ±0.2 | 25.3 ±0.3 | 31.6 ±0.3 |
| Integrated Gradients | 33.3 ±0.2 | 29.3 ±0.3 | 21.4 ±0.2 | 21.9 ±0.2 | 26.5 ±0.2 |
| Input × Grad | 31.8 ±0.2 | 25.0 ±0.3 | 20.2 ±0.2 | 19.6 ±0.2 | 24.2 ±0.2 |
| **Libra Input × Grad** | 44.0 ±0.3 (+38.3%) | 32.2 ±0.2 (+28.5%) | 23.4 ±0.2 (+15.8%) | 26.1 ±0.3 (+33.1%) | 31.4 ±0.2 (+30.0%) |
| AttCAT | 42.0 ±0.3 | 38.2 ±0.3 | 28.8 ±0.2 | 29.0 ±0.3 | 34.5 ±0.3 |
| **Libra AttCAT** | <u>52.1</u> ±0.2 (+24.1%) | **48.9** ±0.3 (+28.0%) | **41.5** ±0.3 (+44.2%) | <u>44.5</u> ±0.3 (+53.6%) | <u>46.8</u> ±0.3 (+35.6%) |
| GenAtt | 49.4 ±0.3 | 46.3 ±0.3 | 37.9 ±0.2 | 36.5 ±0.3 | 42.5 ±0.3 |
| **Libra GenAtt** | 50.5 ±0.2 (+2.2%) | 48.2 ±0.3 (+4.2%) | 40.4 ±0.3 (+6.6%) | 39.6 ±0.3 (+8.7%) | 44.7 ±0.3 (+5.1%) |
| TokenTM | 48.3 ±0.3 | 45.9 ±0.3 | 37.4 ±0.3 | 34.9 ±0.3 | 41.6 ±0.3 |
| **Libra TokenTM** | 49.2 ±0.3 (+1.9%) | 47.3 ±0.3 (+3.0%) | 38.9 ±0.3 (+3.8%) | 37.4 ±0.3 (+7.1%) | 43.2 ±0.3 (+3.8%) |
| GradCAM+ | 40.1 ±0.2 | 35.8 ±0.3 | 27.6 ±0.2 | 33.0 ±0.2 | 34.1 ±0.2 |
| **Libra GradCAM+** | 46.4 ±0.2 (+15.7%) | 46.1 ±0.3 (+28.7%) | 39.6 ±0.2 (+43.5%) | 40.1 ±0.3 (+21.8%) | 43.0 ±0.3 (+26.2%) |
| HiResCAM | 19.5 ±0.3 | 15.3 ±0.2 | 28.5 ±0.2 | 12.2 ±0.2 | 18.9 ±0.2 |
| **Libra HiResCAM** | 44.0 ±0.2 (+125.6%) | 44.4 ±0.2 (+190.6%) | 37.0 ±0.2 (+29.6%) | 33.2 ±0.3 (+171.8%) | 39.7 ±0.2 (+110.0%) |
| XGradCAM+ | 41.2 ±0.2 | 36.2 ±0.3 | 21.5 ±0.2 | 30.5 ±0.3 | 32.4 ±0.3 |
| **Libra XGradCAM+** | 49.5 ±0.2 (+20.1%) | 47.8 ±0.3 (+32.0%) | **41.5** ±0.2 (+92.8%) | 42.2 ±0.3 (+38.3%) | 45.3 ±0.3 (+39.8%) |
| FullGrad+ | 39.2 ±0.3 | 36.1 ±0.3 | 26.3 ±0.2 | 28.9 ±0.3 | 32.6 ±0.3 |
| **Libra FullGrad+** | **52.7** ±0.2 (+34.7%) | **48.9** ±0.3 (+35.3%) | 41.2 ±0.3 (+56.7%) | **45.3** ±0.3 (+56.5%) | **47.0** ±0.3 (+44.1%) |

Table 25. How Most-Influential-First Deletion (MIF) AOPC evaluated using predicted labels varies with different model sizes.

| Method | ViT-Tiny | ViT-Small | ViT-Base | ViT-Large | *Avg.* |
|---|---|---|---|---|---|
| Random | 17.0 ±0.2 | 15.8 ±0.3 | 12.3 ±0.2 | 14.1 ±0.2 | 14.8 ±0.2 |
| RawAtt | 38.6 ±0.3 | 36.5 ±0.3 | 25.0 ±0.3 | 22.9 ±0.3 | 30.7 ±0.3 |
| Attention Rollout | 33.6 ±0.3 | 25.1 ±0.4 | 18.8 ±0.3 | 16.5 ±0.3 | 23.5 ±0.3 |
| AliLRP | 33.4 ±0.3 | 22.8 ±0.3 | 16.7 ±0.2 | 17.2 ±0.3 | 22.5 ±0.3 |
| AttnLRP | 37.8 ±0.3 | 30.9 ±0.3 | 20.8 ±0.3 | 24.8 ±0.3 | 28.6 ±0.3 |
| DecompX | 38.2 ±0.3 | 29.5 ±0.3 | 20.3 ±0.3 | 22.6 ±0.3 | 27.7 ±0.3 |
| Integrated Gradients | 32.8 ±0.3 | 29.1 ±0.3 | 22.5 ±0.2 | 23.1 ±0.3 | 26.9 ±0.3 |
| Input × Grad | 26.3 ±0.3 | 21.6 ±0.3 | 17.7 ±0.2 | 17.5 ±0.3 | 20.8 ±0.2 |
| **Libra Input × Grad** | 37.5 ±0.3 (+42.6%) | 28.2 ±0.3 (+30.4%) | 20.8 ±0.3 (+17.4%) | 23.4 ±0.3 (+33.5%) | 27.5 ±0.3 (+32.2%) |
| AttCAT | 35.6 ±0.3 | 33.4 ±0.3 | 25.3 ±0.2 | 25.7 ±0.3 | 30.0 ±0.3 |
| **Libra AttCAT** | <u>45.0</u> ±0.3 (+26.6%) | **43.9** ±0.3 (+31.4%) | <u>37.5</u> ±0.3 (+47.9%) | <u>40.5</u> ±0.3 (+57.3%) | <u>41.7</u> ±0.3 (+39.0%) |
| GenAtt | 42.7 ±0.3 | 41.3 ±0.3 | 34.2 ±0.3 | 33.2 ±0.3 | 37.8 ±0.3 |
| **Libra GenAtt** | 43.6 ±0.3 (+2.3%) | 43.2 ±0.3 (+4.6%) | 36.6 ±0.3 (+6.8%) | 36.2 ±0.3 (+8.9%) | 39.9 ±0.3 (+5.4%) |
| TokenTM | 41.8 ±0.3 | 40.8 ±0.3 | 33.8 ±0.3 | 31.8 ±0.3 | 37.1 ±0.3 |
| **Libra TokenTM** | 42.6 ±0.3 (+2.0%) | 42.2 ±0.3 (+3.4%) | 35.1 ±0.3 (+4.0%) | 34.2 ±0.3 (+7.4%) | 38.5 ±0.3 (+4.0%) |
| GradCAM+ | 34.1 ±0.3 | 31.5 ±0.3 | 24.8 ±0.2 | 30.0 ±0.3 | 30.1 ±0.3 |
| **Libra GradCAM+** | 39.9 ±0.3 (+16.8%) | 41.2 ±0.3 (+30.7%) | 35.9 ±0.2 (+44.8%) | 36.7 ±0.3 (+22.0%) | 38.4 ±0.3 (+27.5%) |
| HiResCAM | 15.8 ±0.3 | 13.1 ±0.2 | 25.4 ±0.3 | 10.6 ±0.2 | 16.2 ±0.2 |
| **Libra HiResCAM** | 37.8 ±0.3 (+138.5%) | 39.6 ±0.3 (+202.6%) | 33.4 ±0.3 (+31.7%) | 30.2 ±0.3 (+186.3%) | 35.2 ±0.3 (+117.4%) |
| XGradCAM+ | 35.1 ±0.3 | 31.9 ±0.4 | 19.0 ±0.2 | 27.7 ±0.3 | 28.4 ±0.3 |
| **Libra XGradCAM+** | 42.7 ±0.3 (+21.7%) | 42.8 ±0.3 (+34.1%) | **37.7** ±0.2 (+98.6%) | 38.6 ±0.3 (+39.2%) | 40.4 ±0.3 (+42.3%) |
| FullGrad+ | 33.1 ±0.3 | 31.5 ±0.3 | 23.1 ±0.3 | 25.8 ±0.3 | 28.4 ±0.3 |
| **Libra FullGrad+** | **45.6** ±0.3 (+37.8%) | <u>43.8</u> ±0.3 (+39.0%) | 37.2 ±0.3 (+60.9%) | **41.2** ±0.3 (+59.5%) | **41.9** ±0.3 (+47.8%) |

Table 26. How Most-Influential-First Deletion (MIF) AOPC evaluated using ground-truth labels varies with different model sizes.

| Method | ViT-Tiny | ViT-Small | ViT-Base | ViT-Large | *Avg.* |
|---|---|---|---|---|---|
| Random | 58.6 ±0.1 | 66.5 ±0.1 | 73.3 ±0.1 | 70.2 ±0.1 | 67.1 ±0.1 |
| RawAtt | 67.3 ±0.1 | 72.8 ±0.1 | 76.2 ±0.1 | 67.6 ±0.1 | 71.0 ±0.1 |
| Attention Rollout | 65.4 ±0.1 | 67.3 ±0.1 | 73.8 ±0.1 | 68.3 ±0.1 | 68.7 ±0.1 |
| AliLRP | 73.0 ±0.1 | 70.6 ±0.1 | 77.8 ±0.1 | 72.5 ±0.1 | 73.5 ±0.1 |
| AttnLRP | 74.3 ±0.1 | 77.3 ±0.1 | 78.7 ±0.1 | 77.6 ±0.1 | 77.0 ±0.1 |
| DecompX | 74.8 ±0.1 | 75.3 ±0.1 | 79.1 ±0.1 | 75.8 ±0.1 | 76.3 ±0.1 |
| Integrated Gradients | 66.9 ±0.1 | 73.9 ±0.1 | 78.0 ±0.1 | 73.5 ±0.1 | 73.1 ±0.1 |
| Input × Grad | 69.0 ±0.1 | 72.1 ±0.1 | 77.3 ±0.1 | 72.8 ±0.1 | 72.8 ±0.1 |
| **Libra Input × Grad** | 74.8 ±0.1 (+8.4%) | 74.3 ±0.1 (+3.0%) | 80.2 ±0.1 (+3.8%) | 76.7 ±0.1 (+5.4%) | 76.5 ±0.1 (+5.1%) |
| AttCAT | 74.8 ±0.1 | 78.5 ±0.1 | 82.5 ±0.1 | 77.5 ±0.1 | 78.3 ±0.1 |
| **Libra AttCAT** | <u>77.9</u> ±0.1 (+4.2%) | **81.0** ±0.1 (+3.2%) | <u>86.7</u> ±0.1 (+5.1%) | <u>82.2</u> ±0.1 (+6.1%) | <u>81.9</u> ±0.1 (+4.6%) |
| GenAtt | 76.2 ±0.1 | 79.1 ±0.1 | 84.0 ±0.1 | 78.2 ±0.1 | 79.4 ±0.1 |
| **Libra GenAtt** | 74.6 ±0.1 (-2.1%) | 79.0 ±0.1 (-0.1%) | 84.4 ±0.1 (+0.4%) | 78.8 ±0.1 (+0.7%) | 79.2 ±0.1 (-0.2%) |
| TokenTM | 74.2 ±0.1 | 77.2 ±0.1 | 83.1 ±0.1 | 77.3 ±0.1 | 77.9 ±0.1 |
| **Libra TokenTM** | 73.7 ±0.1 (-0.6%) | 77.1 ±0.1 (-0.1%) | 83.2 ±0.1 (+0.1%) | 77.8 ±0.1 (+0.7%) | 78.0 ±0.1 (+0.0%) |
| GradCAM+ | 65.1 ±0.1 | 71.9 ±0.1 | 78.5 ±0.1 | 76.8 ±0.1 | 73.1 ±0.1 |
| **Libra GradCAM+** | 70.2 ±0.1 (+7.8%) | 78.0 ±0.1 (+8.4%) | 84.9 ±0.1 (+8.3%) | 79.1 ±0.1 (+3.0%) | 78.0 ±0.1 (+6.8%) |
| HiResCAM | 48.3 ±0.1 | 62.8 ±0.1 | 79.5 ±0.1 | 59.3 ±0.1 | 62.5 ±0.1 |
| **Libra HiResCAM** | 68.0 ±0.1 (+40.8%) | 76.1 ±0.1 (+21.2%) | 82.7 ±0.1 (+4.0%) | 73.8 ±0.1 (+24.4%) | 75.2 ±0.1 (+20.3%) |
| XGradCAM+ | 66.7 ±0.1 | 73.5 ±0.1 | 73.3 ±0.1 | 75.7 ±0.1 | 72.3 ±0.1 |
| **Libra XGradCAM+** | 72.8 ±0.1 (+9.0%) | 78.5 ±0.1 (+6.8%) | 85.4 ±0.1 (+16.6%) | 80.0 ±0.1 (+5.6%) | 79.2 ±0.1 (+9.5%) |
| FullGrad+ | 73.4 ±0.1 | 77.6 ±0.1 | 81.6 ±0.1 | 76.8 ±0.1 | 77.4 ±0.1 |
| **Libra FullGrad+** | **78.8** ±0.1 (+7.3%) | **81.0** ±0.1 (+4.4%) | **87.0** ±0.0 (+6.6%) | **82.6** ±0.1 (+7.6%) | **82.4** ±0.1 (+6.5%) |

Table 27. How Least-Influential-First Deletion (LIF) Accuracy evaluated using predicted labels varies with different model sizes.

| Method | ViT-Tiny | ViT-Small | ViT-Base | ViT-Large | *Avg.* |
|---|---|---|---|---|---|
| Random | 49.2 ±0.1 | 57.7 ±0.1 | 65.2 ±0.1 | 62.9 ±0.1 | 58.8 ±0.1 |
| RawAtt | 55.2 ±0.1 | 63.9 ±0.1 | 67.5 ±0.1 | 60.3 ±0.1 | 61.7 ±0.1 |
| Attention Rollout | 54.4 ±0.1 | 58.6 ±0.1 | 65.9 ±0.1 | 61.9 ±0.1 | 60.2 ±0.1 |
| AliLRP | 63.1 ±0.1 | 62.5 ±0.1 | 69.9 ±0.1 | 65.4 ±0.1 | 65.2 ±0.1 |
| AttnLRP | 63.2 ±0.1 | 68.4 ±0.1 | 71.0 ±0.1 | 70.3 ±0.1 | 68.2 ±0.1 |
| DecompX | 63.3 ±0.1 | 66.5 ±0.1 | 71.1 ±0.1 | 68.8 ±0.1 | 67.4 ±0.1 |
| Integrated Gradients | 64.0 ±0.1 | 69.6 ±0.1 | 74.4 ±0.1 | 71.1 ±0.1 | 69.8 ±0.1 |
| Input × Grad | 58.6 ±0.1 | 64.2 ±0.1 | 69.9 ±0.1 | 65.8 ±0.1 | 64.6 ±0.1 |
| **Libra Input × Grad** | 65.2 ±0.1 (+11.3%) | 66.4 ±0.1 (+3.4%) | 72.5 ±0.1 (+3.7%) | 70.1 ±0.1 (+6.6%) | 68.5 ±0.1 (+6.1%) |
| AttCAT | 66.5 ±0.1 | 71.9 ±0.1 | 76.8 ±0.1 | 71.8 ±0.1 | 71.7 ±0.1 |
| **Libra AttCAT** | <u>69.5</u> ±0.1 (+4.5%) | <u>74.2</u> ±0.1 (+3.3%) | <u>80.2</u> ±0.1 (+4.5%) | <u>76.3</u> ±0.1 (+6.2%) | <u>75.0</u> ±0.1 (+4.6%) |
| GenAtt | 63.6 ±0.1 | 69.6 ±0.1 | 74.8 ±0.1 | 70.0 ±0.1 | 69.5 ±0.1 |
| **Libra GenAtt** | 62.3 ±0.1 (-2.0%) | 69.5 ±0.1 (-0.2%) | 75.1 ±0.1 (+0.4%) | 70.9 ±0.1 (+1.3%) | 69.4 ±0.1 (-0.1%) |
| TokenTM | 61.2 ±0.1 | 67.4 ±0.1 | 73.5 ±0.1 | 68.9 ±0.1 | 67.7 ±0.1 |
| **Libra TokenTM** | 60.8 ±0.1 (-0.6%) | 67.4 ±0.1 (-0.1%) | 73.6 ±0.1 (+0.1%) | 69.4 ±0.1 (+0.8%) | 67.8 ±0.1 (+0.1%) |
| GradCAM+ | 57.9 ±0.1 | 65.0 ±0.1 | 72.0 ±0.1 | 70.5 ±0.1 | 66.3 ±0.1 |
| **Libra GradCAM+** | 61.9 ±0.1 (+7.0%) | 70.7 ±0.1 (+8.9%) | 78.0 ±0.1 (+8.3%) | 72.6 ±0.1 (+2.9%) | 70.8 ±0.1 (+6.7%) |
| HiResCAM | 42.4 ±0.1 | 55.8 ±0.1 | 71.9 ±0.1 | 53.6 ±0.1 | 55.9 ±0.1 |
| **Libra HiResCAM** | 60.0 ±0.1 (+41.5%) | 68.5 ±0.1 (+22.6%) | 75.6 ±0.1 (+5.1%) | 67.4 ±0.1 (+25.7%) | 67.9 ±0.1 (+21.3%) |
| XGradCAM+ | 59.5 ±0.1 | 66.3 ±0.1 | 67.0 ±0.1 | 69.5 ±0.1 | 65.6 ±0.1 |
| **Libra XGradCAM+** | 64.4 ±0.1 (+8.3%) | 71.4 ±0.1 (+7.7%) | 78.1 ±0.1 (+16.6%) | 73.5 ±0.1 (+5.7%) | 71.9 ±0.1 (+9.6%) |
| FullGrad+ | 64.5 ±0.1 | 70.3 ±0.1 | 75.2 ±0.1 | 71.5 ±0.1 | 70.4 ±0.1 |
| **Libra FullGrad+** | **70.2** ±0.1 (+8.8%) | **74.4** ±0.1 (+5.9%) | **80.6** ±0.1 (+7.1%) | **76.8** ±0.1 (+7.5%) | **75.5** ±0.1 (+7.3%) |

Table 28. How Least-Influential-First Deletion (LIF) Accuracy evaluated using ground-truth labels varies with different model sizes.

| Method | ViT-Tiny | ViT-Small | ViT-Base | ViT-Large | *Avg.* |
|---|---|---|---|---|---|
| Random | 79.0 ±0.2 | 81.8 ±0.2 | 85.8 ±0.2 | 83.7 ±0.2 | 82.6 ±0.2 |
| RawAtt | 85.6 ±0.2 | 87.2 ±0.2 | 87.6 ±0.1 | 81.5 ±0.1 | 85.5 ±0.2 |
| Attention Rollout | 84.3 ±0.2 | 82.4 ±0.3 | 86.0 ±0.2 | 81.9 ±0.2 | 83.7 ±0.2 |
| AliLRP | 92.6 ±0.2 | 85.8 ±0.2 | 89.3 ±0.2 | 85.9 ±0.2 | 88.4 ±0.2 |
| AttnLRP | 93.2 ±0.2 | 92.8 ±0.2 | 90.8 ±0.1 | 91.3 ±0.2 | 92.1 ±0.2 |
| DecompX | 93.1 ±0.2 | 90.3 ±0.2 | 90.6 ±0.1 | 89.3 ±0.2 | 90.8 ±0.2 |
| Integrated Gradients | 88.8 ±0.2 | 90.1 ±0.2 | 91.3 ±0.1 | 88.4 ±0.2 | 89.6 ±0.2 |
| Input × Grad | 88.3 ±0.2 | 87.8 ±0.3 | 90.2 ±0.1 | 86.7 ±0.1 | 88.2 ±0.2 |
| **Libra Input × Grad** | 93.9 ±0.2 (+6.3%) | 89.6 ±0.2 (+2.1%) | 91.3 ±0.2 (+1.2%) | 90.2 ±0.2 (+4.0%) | 91.3 ±0.2 (+3.4%) |
| AttCAT | 95.3 ±0.2 | 95.6 ±0.2 | 96.6 ±0.2 | 92.6 ±0.2 | 95.0 ±0.2 |
| **Libra AttCAT** | <u>98.1</u> ±0.2 (+2.9%) | <u>97.8</u> ±0.2 (+2.3%) | <u>99.2</u> ±0.1 (+2.7%) | <u>97.1</u> ±0.2 (+4.8%) | <u>98.0</u> ±0.2 (+3.2%) |
| GenAtt | 93.5 ±0.2 | 92.9 ±0.2 | 94.6 ±0.1 | 91.5 ±0.2 | 93.1 ±0.2 |
| **Libra GenAtt** | 92.2 ±0.2 (-1.4%) | 92.7 ±0.2 (-0.2%) | 94.8 ±0.1 (+0.2%) | 92.0 ±0.2 (+0.5%) | 92.9 ±0.2 (-0.2%) |
| TokenTM | 91.3 ±0.2 | 90.8 ±0.2 | 93.3 ±0.1 | 90.3 ±0.2 | 91.4 ±0.2 |
| **Libra TokenTM** | 90.5 ±0.2 (-0.9%) | 90.8 ±0.2 (+0.0%) | 93.5 ±0.2 (+0.2%) | 90.8 ±0.2 (+0.6%) | 91.4 ±0.2 (+0.0%) |
| GradCAM+ | 84.3 ±0.2 | 88.1 ±0.2 | 91.5 ±0.2 | 91.0 ±0.2 | 88.7 ±0.2 |
| **Libra GradCAM+** | 89.6 ±0.2 (+6.2%) | 93.7 ±0.2 (+6.4%) | 96.2 ±0.1 (+5.2%) | 92.7 ±0.2 (+1.8%) | 93.0 ±0.2 (+4.9%) |
| HiResCAM | 71.3 ±0.2 | 78.7 ±0.3 | 91.7 ±0.2 | 74.2 ±0.3 | 79.0 ±0.2 |
| **Libra HiResCAM** | 86.2 ±0.2 (+20.8%) | 91.1 ±0.2 (+15.7%) | 94.3 ±0.1 (+2.8%) | 88.0 ±0.2 (+18.6%) | 89.9 ±0.2 (+13.8%) |
| XGradCAM+ | 86.4 ±0.2 | 89.5 ±0.2 | 86.6 ±0.2 | 90.1 ±0.2 | 88.2 ±0.2 |
| **Libra XGradCAM+** | 91.9 ±0.2 (+6.3%) | 94.1 ±0.2 (+5.2%) | 96.6 ±0.1 (+11.6%) | 93.7 ±0.2 (+3.9%) | 94.1 ±0.2 (+6.7%) |
| FullGrad+ | 93.5 ±0.2 | 94.3 ±0.2 | 95.0 ±0.2 | 91.8 ±0.2 | 93.7 ±0.2 |
| **Libra FullGrad+** | **99.2** ±0.2 (+6.1%) | **97.9** ±0.2 (+3.8%) | **99.6** ±0.1 (+4.8%) | **97.4** ±0.2 (+6.0%) | **98.5** ±0.2 (+5.2%) |

Table 29. How Least-Influential-First Deletion (LIF) AOPC evaluated using predicted labels varies with different model sizes.

| Method | ViT-Tiny | ViT-Small | ViT-Base | ViT-Large | *Avg.* |
|---|---|---|---|---|---|
| Random | 82.8 ±0.2 | 84.1 ±0.2 | 87.5 ±0.2 | 85.4 ±0.2 | 84.9 ±0.2 |
| RawAtt | 88.1 ±0.2 | 89.4 ±0.2 | 89.2 ±0.1 | 83.3 ±0.2 | 87.5 ±0.2 |
| Attention Rollout | 87.2 ±0.2 | 84.8 ±0.2 | 87.9 ±0.2 | 84.1 ±0.2 | 86.0 ±0.2 |
| AliLRP | 95.0 ±0.3 | 88.2 ±0.2 | 90.9 ±0.2 | 87.7 ±0.2 | 90.4 ±0.2 |
| AttnLRP | 95.6 ±0.2 | 94.6 ±0.2 | 92.4 ±0.2 | 92.9 ±0.2 | 93.8 ±0.2 |
| DecompX | 95.4 ±0.2 | 92.2 ±0.2 | 92.2 ±0.2 | 91.0 ±0.2 | 92.7 ±0.2 |
| Integrated Gradients | 96.3 ±0.2 | 95.4 ±0.2 | 95.7 ±0.2 | 93.3 ±0.2 | 95.2 ±0.2 |
| Input × Grad | 90.8 ±0.2 | 90.0 ±0.3 | 91.8 ±0.2 | 88.4 ±0.2 | 90.3 ±0.2 |
| **Libra Input × Grad** | 96.4 ±0.2 (+6.2%) | 92.0 ±0.2 (+2.1%) | 93.1 ±0.2 (+1.4%) | 92.0 ±0.2 (+4.0%) | 93.3 ±0.2 (+3.4%) |
| AttCAT | 98.0 ±0.2 | 97.7 ±0.2 | 98.4 ±0.2 | 94.3 ±0.2 | 97.1 ±0.2 |
| **Libra AttCAT** | <u>100.8</u> ±0.2 (+2.9%) | <u>100.0</u> ±0.2 (+2.4%) | <u>100.8</u> ±0.2 (+2.4%) | <u>98.5</u> ±0.2 (+4.5%) | <u>100.0</u> ±0.2 (+3.0%) |
| GenAtt | 95.4 ±0.2 | 94.7 ±0.2 | 95.7 ±0.2 | 92.8 ±0.2 | 94.6 ±0.2 |
| **Libra GenAtt** | 94.3 ±0.2 (-1.1%) | 94.5 ±0.2 (-0.2%) | 96.0 ±0.1 (+0.3%) | 93.2 ±0.2 (+0.5%) | 94.5 ±0.2 (-0.1%) |
| TokenTM | 93.4 ±0.2 | 92.7 ±0.2 | 94.4 ±0.1 | 91.6 ±0.2 | 93.0 ±0.2 |
| **Libra TokenTM** | 92.7 ±0.2 (-0.7%) | 92.8 ±0.2 (+0.1%) | 94.6 ±0.1 (+0.2%) | 92.1 ±0.2 (+0.5%) | 93.0 ±0.2 (+0.0%) |
| GradCAM+ | 88.6 ±0.2 | 90.5 ±0.2 | 93.5 ±0.2 | 92.9 ±0.2 | 91.4 ±0.2 |
| **Libra GradCAM+** | 93.3 ±0.2 (+5.2%) | 96.0 ±0.2 (+6.1%) | 98.1 ±0.2 (+4.9%) | 94.4 ±0.2 (+1.6%) | 95.4 ±0.2 (+4.4%) |
| HiResCAM | 76.6 ±0.3 | 81.8 ±0.3 | 93.3 ±0.2 | 76.7 ±0.2 | 82.1 ±0.3 |
| **Libra HiResCAM** | 90.1 ±0.2 (+17.7%) | 93.3 ±0.2 (+14.1%) | 96.1 ±0.2 (+3.0%) | 90.0 ±0.2 (+17.3%) | 92.4 ±0.2 (+12.5%) |
| XGradCAM+ | 90.5 ±0.3 | 91.8 ±0.2 | 88.9 ±0.3 | 92.1 ±0.2 | 90.8 ±0.2 |
| **Libra XGradCAM+** | 95.5 ±0.2 (+5.6%) | 96.5 ±0.2 (+5.1%) | 98.4 ±0.2 (+10.8%) | 95.3 ±0.2 (+3.5%) | 96.4 ±0.2 (+6.2%) |
| FullGrad+ | 96.3 ±0.2 | 96.3 ±0.2 | 96.6 ±0.2 | 93.8 ±0.2 | 95.7 ±0.2 |
| **Libra FullGrad+** | **101.8** ±0.2 (+5.7%) | **100.1** ±0.2 (+4.0%) | **101.2** ±0.2 (+4.8%) | **98.9** ±0.2 (+5.4%) | **100.5** ±0.2 (+4.9%) |

Table 30. How Least-Influential-First Deletion (LIF) AOPC evaluated using ground-truth labels varies with different model sizes.

| Method | ViT-Tiny | ViT-Small | ViT-Base | ViT-Large | *Avg.* |
|---|---|---|---|---|---|
| Random | 49.6 ±0.1 | 50.2 ±0.1 | 49.9 ±0.1 | 49.8 ±0.1 | 49.9 ±0.1 |
| RawAtt | 68.4 ±0.1 | 65.8 ±0.1 | 60.4 ±0.1 | 53.3 ±0.1 | 62.0 ±0.1 |
| Attention Rollout | 64.8 ±0.1 | 56.2 ±0.1 | 54.6 ±0.1 | 49.9 ±0.1 | 56.3 ±0.1 |
| AliLRP | 68.7 ±0.1 | 56.5 ±0.1 | 55.5 ±0.1 | 52.8 ±0.1 | 58.4 ±0.1 |
| AttnLRP | 72.0 ±0.1 | 64.9 ±0.1 | 58.6 ±0.1 | 59.7 ±0.1 | 63.8 ±0.1 |
| DecompX | 72.6 ±0.1 | 62.9 ±0.1 | 58.5 ±0.1 | 57.4 ±0.1 | 62.8 ±0.1 |
| Integrated Gradients | 62.0 ±0.1 | 59.9 ±0.1 | 56.7 ±0.1 | 54.7 ±0.1 | 58.3 ±0.1 |
| Input × Grad | 62.3 ±0.1 | 57.0 ±0.1 | 55.9 ±0.1 | 53.3 ±0.1 | 57.1 ±0.1 |
| **Libra Input × Grad** | 72.8 ±0.1 (+16.8%) | 61.8 ±0.1 (+8.5%) | 59.4 ±0.1 (+6.3%) | 58.6 ±0.1 (+9.9%) | 63.1 ±0.1 (+10.6%) |
| AttCAT | 72.0 ±0.1 | 68.7 ±0.1 | 64.7 ±0.1 | 61.2 ±0.1 | 66.7 ±0.1 |
| **Libra AttCAT** | <u>79.4</u> ±0.1 (+10.2%) | **75.7** ±0.1 (+10.1%) | **75.1** ±0.1 (+16.1%) | <u>71.8</u> ±0.1 (+17.4%) | <u>75.5</u> ±0.1 (+13.2%) |
| GenAtt | 76.6 ±0.1 | 72.7 ±0.1 | 71.1 ±0.1 | 65.0 ±0.1 | 71.4 ±0.1 |
| **Libra GenAtt** | 76.5 ±0.1 (-0.2%) | 73.6 ±0.1 (+1.3%) | 73.0 ±0.1 (+2.6%) | 67.1 ±0.1 (+3.2%) | 72.5 ±0.1 (+1.6%) |
| TokenTM | 74.6 ±0.1 | 71.2 ±0.1 | 70.0 ±0.1 | 63.6 ±0.1 | 69.8 ±0.1 |
| **Libra TokenTM** | 75.0 ±0.1 (+0.5%) | 71.8 ±0.1 (+0.9%) | 71.1 ±0.1 (+1.7%) | 65.2 ±0.1 (+2.4%) | 70.8 ±0.1 (+1.3%) |
| GradCAM+ | 65.7 ±0.1 | 63.7 ±0.1 | 62.0 ±0.1 | 62.7 ±0.1 | 63.5 ±0.1 |
| **Libra GradCAM+** | 71.5 ±0.1 (+9.0%) | 72.2 ±0.1 (+13.3%) | 73.2 ±0.1 (+18.0%) | 67.8 ±0.1 (+8.1%) | 71.2 ±0.1 (+12.1%) |
| HiResCAM | 43.7 ±0.1 | 46.2 ±0.1 | 62.5 ±0.1 | 42.5 ±0.1 | 48.7 ±0.1 |
| **Libra HiResCAM** | 68.9 ±0.1 (+57.9%) | 69.8 ±0.1 (+51.1%) | 69.7 ±0.1 (+11.6%) | 61.4 ±0.1 (+44.4%) | 67.4 ±0.1 (+38.5%) |
| XGradCAM+ | 67.1 ±0.1 | 64.7 ±0.1 | 55.9 ±0.1 | 60.8 ±0.1 | 62.1 ±0.1 |
| **Libra XGradCAM+** | 74.9 ±0.1 (+11.6%) | 73.5 ±0.1 (+13.5%) | 74.6 ±0.1 (+33.5%) | 69.4 ±0.1 (+14.1%) | 73.1 ±0.1 (+17.6%) |
| FullGrad+ | 69.7 ±0.1 | 66.7 ±0.1 | 62.9 ±0.1 | 60.9 ±0.1 | 65.0 ±0.1 |
| **Libra FullGrad+** | **80.3** ±0.1 (+15.2%) | <u>75.6</u> ±0.1 (+13.3%) | <u>75.0</u> ±0.1 (+19.4%) | **72.5** ±0.1 (+19.0%) | **75.9** ±0.1 (+16.6%) |

Table 31. How Symmetric Relevance Gain (SRG) Accuracy evaluated using predicted labels varies with different model sizes.

| Method | ViT-Tiny | ViT-Small | ViT-Base | ViT-Large | *Avg.* |
|---|---|---|---|---|---|
| Random | 49.6 ±0.1 | 49.7 ±0.1 | 49.9 ±0.1 | 49.9 ±0.1 | 49.8 ±0.1 |
| RawAtt | 64.6 ±0.1 | 63.5 ±0.1 | 58.8 ±0.1 | 52.9 ±0.1 | 60.0 ±0.1 |
| Attention Rollout | 61.6 ±0.1 | 54.9 ±0.1 | 53.9 ±0.1 | 50.4 ±0.1 | 55.2 ±0.1 |
| AliLRP | 66.0 ±0.1 | 55.7 ±0.1 | 54.8 ±0.1 | 52.6 ±0.1 | 57.3 ±0.1 |
| AttnLRP | 68.3 ±0.1 | 63.0 ±0.1 | 57.8 ±0.1 | 58.7 ±0.1 | 62.0 ±0.1 |
| DecompX | 68.6 ±0.1 | 61.3 ±0.1 | 57.6 ±0.1 | 56.6 ±0.1 | 61.0 ±0.1 |
| Integrated Gradients | 66.8 ±0.1 | 63.3 ±0.1 | 60.6 ±0.1 | 58.7 ±0.1 | 62.4 ±0.1 |
| Input × Grad | 59.8 ±0.1 | 56.1 ±0.1 | 55.1 ±0.1 | 53.0 ±0.1 | 56.0 ±0.1 |
| **Libra Input × Grad** | 69.9 ±0.1 (+16.8%) | 60.7 ±0.1 (+8.2%) | 58.6 ±0.1 (+6.3%) | 58.0 ±0.1 (+9.5%) | 61.8 ±0.1 (+10.3%) |
| AttCAT | 69.5 ±0.1 | 67.0 ±0.1 | 63.6 ±0.1 | 60.2 ±0.1 | 65.1 ±0.1 |
| **Libra AttCAT** | <u>76.5</u> ±0.1 (+10.1%) | <u>73.9</u> ±0.1 (+10.3%) | **73.3** ±0.1 (+15.3%) | <u>70.5</u> ±0.1 (+17.0%) | <u>73.6</u> ±0.1 (+13.0%) |
| GenAtt | 72.0 ±0.1 | 69.6 ±0.1 | 68.4 ±0.1 | 63.2 ±0.1 | 68.3 ±0.1 |
| **Libra GenAtt** | 71.9 ±0.1 (-0.1%) | 70.6 ±0.1 (+1.4%) | 70.1 ±0.1 (+2.5%) | 65.3 ±0.1 (+3.3%) | 69.5 ±0.1 (+1.7%) |
| TokenTM | 70.0 ±0.1 | 68.2 ±0.1 | 67.1 ±0.1 | 61.9 ±0.1 | 66.8 ±0.1 |
| **Libra TokenTM** | 70.3 ±0.1 (+0.5%) | 68.8 ±0.1 (+1.0%) | 68.2 ±0.1 (+1.7%) | 63.4 ±0.1 (+2.4%) | 67.7 ±0.1 (+1.4%) |
| GradCAM+ | 64.2 ±0.1 | 62.5 ±0.1 | 61.3 ±0.1 | 62.0 ±0.1 | 62.5 ±0.1 |
| **Libra GradCAM+** | 69.3 ±0.1 (+8.0%) | 70.4 ±0.1 (+12.8%) | 71.7 ±0.1 (+17.0%) | 66.7 ±0.1 (+7.7%) | 69.6 ±0.1 (+11.3%) |
| HiResCAM | 45.2 ±0.1 | 47.1 ±0.1 | 61.2 ±0.1 | 43.2 ±0.1 | 49.2 ±0.1 |
| **Libra HiResCAM** | 67.0 ±0.1 (+48.3%) | 68.0 ±0.1 (+44.2%) | 68.2 ±0.1 (+11.5%) | 60.7 ±0.1 (+40.7%) | 66.0 ±0.1 (+34.2%) |
| XGradCAM+ | 65.6 ±0.1 | 63.3 ±0.1 | 55.5 ±0.1 | 60.2 ±0.1 | 61.2 ±0.1 |
| **Libra XGradCAM+** | 72.5 ±0.1 (+10.5%) | 71.7 ±0.1 (+13.3%) | 72.7 ±0.1 (+31.0%) | 68.2 ±0.1 (+13.3%) | 71.3 ±0.1 (+16.6%) |
| FullGrad+ | 67.1 ±0.1 | 65.0 ±0.1 | 61.7 ±0.1 | 60.3 ±0.1 | 63.5 ±0.1 |
| **Libra FullGrad+** | **77.2** ±0.1 (+15.0%) | **74.0** ±0.1 (+13.9%) | **73.3** ±0.1 (+18.8%) | **71.2** ±0.1 (+18.1%) | **73.9** ±0.1 (+16.4%) |

Table 32. How Symmetric Relevance Gain (SRG) Accuracy evaluated using ground-truth labels varies with different model sizes.

| Method | ViT-Tiny | ViT-Small | ViT-Base | ViT-Large | *Avg.* |
|---|---|---|---|---|---|
| Random | 49.9 ±0.2 | 50.2 ±0.2 | 50.0 ±0.2 | 49.8 ±0.2 | 50.0 ±0.2 |
| RawAtt | 65.2 ±0.2 | 64.2 ±0.3 | 57.8 ±0.2 | 53.4 ±0.2 | 60.1 ±0.2 |
| Attention Rollout | 62.1 ±0.2 | 55.6 ±0.2 | 53.6 ±0.2 | 50.1 ±0.2 | 55.3 ±0.2 |
| AliLRP | 66.0 ±0.2 | 56.0 ±0.3 | 54.2 ±0.2 | 52.5 ±0.2 | 57.2 ±0.2 |
| AttnLRP | 68.8 ±0.2 | 64.0 ±0.2 | 57.1 ±0.2 | 59.5 ±0.2 | 62.3 ±0.2 |
| DecompX | 69.0 ±0.2 | 61.9 ±0.2 | 56.7 ±0.2 | 57.3 ±0.2 | 61.2 ±0.2 |
| Integrated Gradients | 61.1 ±0.2 | 59.7 ±0.3 | 56.3 ±0.2 | 55.1 ±0.2 | 58.1 ±0.2 |
| Input × Grad | 60.0 ±0.2 | 56.4 ±0.3 | 55.2 ±0.2 | 53.2 ±0.2 | 56.2 ±0.2 |
| **Libra Input × Grad** | 68.9 ±0.2 (+14.8%) | 60.9 ±0.2 (+8.0%) | 57.3 ±0.2 (+3.9%) | 58.2 ±0.2 (+9.4%) | 61.3 ±0.2 (+9.1%) |
| AttCAT | 68.6 ±0.2 | 66.9 ±0.2 | 62.7 ±0.2 | 60.8 ±0.2 | 64.8 ±0.2 |
| **Libra AttCAT** | 75.1 ±0.2 (+9.4%) | 73.3 ±0.3 (+9.6%) | **70.4** ±0.2 (+12.2%) | 70.8 ±0.2 (+16.4%) | 72.4 ±0.2 (+11.8%) |
| GenAtt | 71.5 ±0.2 | 69.6 ±0.2 | 66.3 ±0.2 | 64.0 ±0.2 | 67.8 ±0.2 |
| **Libra GenAtt** | 71.4 ±0.2 (-0.1%) | 70.5 ±0.3 (+1.2%) | 67.6 ±0.2 (+2.1%) | 65.8 ±0.3 (+2.8%) | 68.8 ±0.2 (+1.5%) |
| TokenTM | 69.8 ±0.2 | 68.3 ±0.2 | 65.3 ±0.2 | 62.6 ±0.2 | 66.5 ±0.2 |
| **Libra TokenTM** | 69.9 ±0.2 (+0.1%) | 69.1 ±0.3 (+1.0%) | 66.2 ±0.2 (+1.3%) | 64.1 ±0.3 (+2.4%) | 67.3 ±0.2 (+1.2%) |
| GradCAM+ | 62.2 ±0.2 | 62.0 ±0.3 | 59.5 ±0.2 | 62.0 ±0.2 | 61.4 ±0.2 |
| **Libra GradCAM+** | 68.0 ±0.2 (+9.3%) | 69.9 ±0.3 (+12.8%) | 67.9 ±0.2 (+14.1%) | 66.4 ±0.2 (+7.2%) | 68.0 ±0.2 (+10.8%) |
| HiResCAM | 45.4 ±0.2 | 47.0 ±0.3 | 60.1 ±0.2 | 43.2 ±0.2 | 48.9 ±0.2 |
| **Libra HiResCAM** | 65.1 ±0.2 (+43.3%) | 67.7 ±0.2 (+44.1%) | 65.7 ±0.2 (+9.2%) | 60.6 ±0.2 (+40.3%) | 64.8 ±0.2 (+32.4%) |
| XGradCAM+ | 63.8 ±0.2 | 62.8 ±0.3 | 54.1 ±0.2 | 60.3 ±0.2 | 60.3 ±0.2 |
| **Libra XGradCAM+** | 70.7 ±0.2 (+10.8%) | 71.0 ±0.2 (+12.9%) | 69.1 ±0.2 (+27.7%) | 68.0 ±0.3 (+12.6%) | 69.7 ±0.2 (+15.6%) |
| FullGrad+ | 66.3 ±0.2 | 65.2 ±0.3 | 60.7 ±0.2 | 60.4 ±0.2 | 63.1 ±0.2 |
| **Libra FullGrad+** | **76.0** ±0.2 (+14.5%) | **73.4** ±0.3 (+12.5%) | **70.4** ±0.2 (+16.0%) | **71.3** ±0.2 (+18.1%) | **72.8** ±0.2 (+15.2%) |

Table 33. How Symmetric Relevance Gain (SRG) AOPC evaluated using predicted labels varies with different model sizes.

| Method | ViT-Tiny | ViT-Small | ViT-Base | ViT-Large | *Avg.* |
|---|---|---|---|---|---|
| Random | 49.9 ±0.2 | 49.9 ±0.3 | 49.9 ±0.2 | 49.7 ±0.2 | 49.9 ±0.2 |
| RawAtt | 63.3 ±0.3 | 63.0 ±0.3 | 57.1 ±0.2 | 53.1 ±0.2 | 59.1 ±0.2 |
| Attention Rollout | 60.4 ±0.3 | 54.9 ±0.3 | 53.3 ±0.2 | 50.3 ±0.3 | 54.8 ±0.3 |
| AliLRP | 64.2 ±0.3 | 55.5 ±0.3 | 53.8 ±0.2 | 52.4 ±0.2 | 56.5 ±0.3 |
| AttnLRP | 66.7 ±0.3 | 62.8 ±0.3 | 56.6 ±0.2 | 58.8 ±0.3 | 61.2 ±0.3 |
| DecompX | 66.8 ±0.3 | 60.9 ±0.3 | 56.3 ±0.2 | 56.8 ±0.3 | 60.2 ±0.3 |
| Integrated Gradients | 64.6 ±0.3 | 62.3 ±0.3 | 59.1 ±0.2 | 58.2 ±0.3 | 61.0 ±0.3 |
| Input × Grad | 58.5 ±0.2 | 55.8 ±0.3 | 54.8 ±0.2 | 53.0 ±0.2 | 55.5 ±0.2 |
| **Libra Input × Grad** | 67.0 ±0.3 (+14.4%) | 60.1 ±0.3 (+7.6%) | 56.9 ±0.2 (+4.0%) | 57.7 ±0.3 (+8.9%) | 60.4 ±0.3 (+8.8%) |
| AttCAT | 66.8 ±0.3 | 65.5 ±0.3 | 61.9 ±0.2 | 60.0 ±0.2 | 63.5 ±0.2 |
| **Libra AttCAT** | 72.9 ±0.2 (+9.2%) | 71.9 ±0.3 (+9.8%) | 69.1 ±0.2 (+11.7%) | 69.5 ±0.3 (+15.8%) | 70.9 ±0.3 (+11.5%) |
| GenAtt | 69.1 ±0.2 | 68.0 ±0.3 | 65.0 ±0.2 | 63.0 ±0.2 | 66.2 ±0.2 |
| **Libra GenAtt** | 69.0 ±0.2 (-0.1%) | 68.8 ±0.3 (+1.3%) | 66.3 ±0.2 (+2.0%) | 64.7 ±0.3 (+2.7%) | 67.2 ±0.2 (+1.4%) |
| TokenTM | 67.6 ±0.2 | 66.8 ±0.3 | 64.1 ±0.2 | 61.7 ±0.3 | 65.0 ±0.3 |
| **Libra TokenTM** | 67.7 ±0.2 (+0.1%) | 67.5 ±0.3 (+1.1%) | 64.9 ±0.2 (+1.2%) | 63.1 ±0.3 (+2.3%) | 65.8 ±0.3 (+1.2%) |
| GradCAM+ | 61.4 ±0.3 | 61.0 ±0.3 | 59.2 ±0.2 | 61.5 ±0.3 | 60.8 ±0.3 |
| **Libra GradCAM+** | 66.6 ±0.3 (+8.4%) | 68.6 ±0.3 (+12.4%) | 67.0 ±0.2 (+13.3%) | 65.5 ±0.3 (+6.6%) | 66.9 ±0.3 (+10.2%) |
| HiResCAM | 46.2 ±0.3 | 47.4 ±0.3 | 59.3 ±0.2 | 43.6 ±0.2 | 49.2 ±0.3 |
| **Libra HiResCAM** | 64.0 ±0.2 (+38.4%) | 66.4 ±0.2 (+40.0%) | 64.7 ±0.2 (+9.1%) | 60.1 ±0.2 (+37.7%) | 63.8 ±0.2 (+29.8%) |
| XGradCAM+ | 62.8 ±0.3 | 61.9 ±0.3 | 53.9 ±0.2 | 59.9 ±0.3 | 59.6 ±0.3 |
| **Libra XGradCAM+** | 69.1 ±0.2 (+10.1%) | 69.7 ±0.3 (+12.6%) | 68.1 ±0.2 (+26.2%) | 66.9 ±0.3 (+11.8%) | 68.4 ±0.3 (+14.8%) |
| FullGrad+ | 64.7 ±0.3 | 63.9 ±0.3 | 59.8 ±0.2 | 59.8 ±0.2 | 62.1 ±0.3 |
| **Libra FullGrad+** | **73.7** ±0.2 (+13.9%) | **72.0** ±0.3 (+12.6%) | **69.2** ±0.2 (+15.6%) | **70.0** ±0.3 (+17.1%) | **71.2** ±0.3 (+14.7%) |

Table 34. How Symmetric Relevance Gain (SRG) AOPC evaluated using ground-truth labels varies with different model sizes.

### D.3.1. Segmentation Average Precision (AP)

| Method | ViT-Tiny | ViT-Small | ViT-Base | ViT-Large | *Avg.* |
|---|---|---|---|---|---|
| Random | 42.0 ±0.4 | 41.9 ±0.4 | 41.9 ±0.4 | 42.0 ±0.4 | 41.9 ±0.4 |
| RawAtt | 60.2 ±0.3 | 57.8 ±0.3 | 46.9 ±0.3 | 40.2 ±0.4 | 51.3 ±0.3 |
| Attention Rollout | 61.2 ±0.4 | 47.1 ±0.3 | 45.3 ±0.3 | 39.9 ±0.3 | 48.3 ±0.3 |
| AliLRP | 54.5 ±0.3 | 42.5 ±0.4 | 43.8 ±0.4 | 42.7 ±0.4 | 45.9 ±0.3 |
| AttnLRP | 59.7 ±0.3 | 46.2 ±0.3 | 42.0 ±0.4 | 47.2 ±0.3 | 48.8 ±0.3 |
| DecompX | 60.0 ±0.3 | 47.7 ±0.3 | 44.3 ±0.3 | 54.2 ±0.3 | 51.6 ±0.3 |
| Integrated Gradients | 52.4 ±0.3 | 51.7 ±0.3 | 47.5 ±0.3 | 46.6 ±0.3 | 49.6 ±0.3 |
| Input $\times$ Grad | 50.6 ±0.3 | 48.5 ±0.3 | 44.8 ±0.3 | 43.6 ±0.4 | 46.9 ±0.3 |
| **Libra Input $\times$ Grad** | 57.1 ±0.3 (+12.8%) | 46.0 ±0.3 (-5.1%) | 44.4 ±0.3 (-0.9%) | 53.6 ±0.3 (+22.9%) | 50.3 ±0.3 (+7.3%) |
| AttCAT | 54.7 ±0.3 | 49.8 ±0.3 | 44.5 ±0.3 | 44.9 ±0.3 | 48.5 ±0.3 |
| **Libra AttCAT** | 61.1 ±0.3 (+11.7%) | 56.0 ±0.3 (+12.4%) | 61.5 ±0.3 (+38.3%) | 53.3 ±0.3 (+18.8%) | 58.0 ±0.3 (+19.6%) |
| GenAtt | 71.1 ±0.3 | 65.9 ±0.2 | 71.0 ±0.2 | 50.9 ±0.3 | 64.7 ±0.3 |
| **Libra GenAtt** | **75.0** ±0.3 (+5.5%) | 71.0 ±0.3 (+7.7%) | **77.5** ±0.2 (+9.2%) | 58.6 ±0.3 (+15.1%) | **70.5** ±0.3 (+9.0%) |
| TokenTM | 70.8 ±0.3 | 68.2 ±0.2 | 70.2 ±0.2 | 50.0 ±0.3 | 64.8 ±0.3 |
| **Libra TokenTM** | 73.7 ±0.3 (+4.1%) | **71.4** ±0.2 (+4.7%) | 73.9 ±0.2 (+5.2%) | 53.9 ±0.3 (+7.9%) | 68.2 ±0.3 (+5.3%) |
| GradCAM+ | 48.4 ±0.4 | 46.4 ±0.4 | 50.2 ±0.4 | 52.1 ±0.4 | 49.3 ±0.4 |
| **Libra GradCAM+** | 56.3 ±0.4 (+16.4%) | 60.7 ±0.4 (+30.8%) | 72.1 ±0.3 (+43.6%) | 60.2 ±0.4 (+15.5%) | 62.3 ±0.4 (+26.5%) |
| HiResCAM | 50.6 ±0.4 | 48.4 ±0.4 | 59.0 ±0.3 | 38.5 ±0.4 | 49.1 ±0.4 |
| **Libra HiResCAM** | 63.8 ±0.3 (+26.1%) | 69.4 ±0.3 (+43.2%) | 72.6 ±0.3 (+23.1%) | 48.0 ±0.3 (+24.8%) | 63.4 ±0.3 (+29.1%) |
| XGradCAM+ | 48.8 ±0.4 | 45.4 ±0.4 | 41.0 ±0.4 | 46.9 ±0.4 | 45.5 ±0.4 |
| **Libra XGradCAM+** | 61.4 ±0.4 (+26.0%) | 62.3 ±0.4 (+37.2%) | 75.0 ±0.3 (+82.8%) | 60.3 ±0.4 (+28.6%) | 64.7 ±0.4 (+42.3%) |
| FullGrad+ | 53.2 ±0.3 | 50.0 ±0.3 | 45.2 ±0.3 | 44.2 ±0.3 | 48.1 ±0.3 |
| **Libra FullGrad+** | 65.0 ±0.3 (+22.2%) | 59.6 ±0.3 (+19.2%) | 65.5 ±0.3 (+44.8%) | **64.5** ±0.3 (+46.0%) | 63.6 ±0.3 (+32.2%) |

Table 35. How Segmentation AP varies with different model sizes.

### D.4. Across Datasets

| Method | ImageNet | ImageNet-Hard | MURA | Oxford-IIIT Pet | *Avg.* |
|---|---|---|---|---|---|
| Random | 26.5 ±0.1 | 52.4 ±0.1 | 15.1 ±0.1 | 13.7 ±0.1 | 26.9 ±0.1 |
| RawAtt | 44.6 ±0.1 | 65.9 ±0.1 | 24.8 ±0.1 | 37.2 ±0.1 | 43.1 ±0.1 |
| Attention Rollout | 35.4 ±0.1 | 62.2 ±0.1 | 21.5 ±0.1 | 21.2 ±0.1 | 35.1 ±0.1 |
| AliLRP | 33.3 ±0.1 | 64.1 ±0.1 | 19.2 ±0.1 | 19.0 ±0.1 | 33.9 ±0.1 |
| AttnLRP | 38.5 ±0.1 | 70.8 ±0.1 | 22.8 ±0.1 | 30.3 ±0.1 | 40.6 ±0.1 |
| DecompX | 37.8 ±0.1 | 67.7 ±0.1 | 21.6 ±0.1 | 22.5 ±0.1 | 37.4 ±0.1 |
| Integrated Gradients | 35.4 ±0.1 | 66.6 ±0.1 | 23.8 ±0.1 | 20.7 ±0.1 | 36.6 ±0.1 |
| Input × Grad | 34.4 ±0.1 | 67.6 ±0.1 | 25.5 ±0.1 | 20.4 ±0.1 | 37.0 ±0.1 |
| **Libra Input × Grad** | 38.6 ±0.1 (+12.0%) | 68.8 ±0.1 (+1.8%) | 21.6 ±0.1 (-15.1%) | 23.5 ±0.1 (+15.4%) | 38.1 ±0.1 (+3.1%) |
| AttCAT | 46.9 ±0.1 | 82.3 ±0.1 | 31.1 ±0.1 | 37.3 ±0.1 | 49.4 ±0.1 |
| **Libra AttCAT** | <u>63.5</u> ±0.1 (+35.4%) | <u>87.3</u> ±0.1 (+6.1%) | <u>40.9</u> ±0.1 (+31.6%) | <u>55.3</u> ±0.1 (+48.1%) | <u>61.8</u> ±0.1 (+25.0%) |
| GenAtt | 58.2 ±0.1 | 81.3 ±0.1 | 30.0 ±0.1 | 44.1 ±0.1 | 53.4 ±0.1 |
| **Libra GenAtt** | 61.6 ±0.1 (+5.8%) | 82.8 ±0.1 (+1.8%) | 30.1 ±0.1 (+0.4%) | 46.5 ±0.1 (+5.4%) | 55.2 ±0.1 (+3.4%) |
| TokenTM | 56.8 ±0.1 | 79.3 ±0.1 | 28.0 ±0.1 | 44.0 ±0.1 | 52.0 ±0.1 |
| **Libra TokenTM** | 59.1 ±0.1 (+4.1%) | 80.0 ±0.1 (+0.8%) | 28.0 ±0.1 (+0.0%) | 45.4 ±0.1 (+3.3%) | 53.1 ±0.1 (+2.1%) |
| GradCAM+ | 45.6 ±0.1 | 75.8 ±0.1 | 24.0 ±0.1 | 32.6 ±0.1 | 44.5 ±0.1 |
| **Libra GradCAM+** | 61.4 ±0.1 (+34.8%) | 83.4 ±0.1 (+10.0%) | 34.7 ±0.1 (+44.8%) | 47.8 ±0.1 (+46.6%) | 56.8 ±0.1 (+27.8%) |
| HiResCAM | 45.4 ±0.1 | 74.2 ±0.1 | 22.2 ±0.1 | 18.0 ±0.1 | 39.9 ±0.1 |
| **Libra HiResCAM** | 56.7 ±0.1 (+24.8%) | 79.7 ±0.1 (+7.4%) | 30.1 ±0.1 (+35.7%) | 39.4 ±0.1 (+119.0%) | 51.5 ±0.1 (+28.9%) |
| XGradCAM+ | 38.6 ±0.1 | 72.1 ±0.1 | 23.7 ±0.1 | 33.2 ±0.1 | 41.9 ±0.1 |
| **Libra XGradCAM+** | **63.9** ±0.1 (+65.6%) | 84.7 ±0.1 (+17.3%) | 36.6 ±0.1 (+54.6%) | 52.6 ±0.1 (+58.4%) | 59.4 ±0.1 (+41.8%) |
| FullGrad+ | 44.2 ±0.1 | 80.1 ±0.1 | 32.8 ±0.1 | 35.3 ±0.1 | 48.1 ±0.1 |
| **Libra FullGrad+** | 63.1 ±0.1 (+42.9%) | **87.6** ±0.1 (+9.4%) | **43.2** ±0.1 (+31.7%) | **57.3** ±0.1 (+62.3%) | **62.8** ±0.1 (+30.6%) |

Table 36. Cross-dataset analysis of Most-Influential-First Deletion (MIF) Accuracy evaluated using predicted labels on ViT-B.

| Method | ImageNet | ImageNet-Hard | MURA | Oxford-IIIT Pet | *Avg.* |
|---|---|---|---|---|---|
| Random | 34.5 ±0.1 | 81.7 ±0.1 | 25.4 ±0.1 | 14.6 ±0.1 | 39.1 ±0.1 |
| RawAtt | 50.1 ±0.1 | 85.9 ±0.1 | 33.4 ±0.1 | 37.7 ±0.1 | 51.8 ±0.1 |
| Attention Rollout | 41.9 ±0.1 | 84.7 ±0.1 | 29.9 ±0.1 | 22.2 ±0.1 | 44.7 ±0.1 |
| AliLRP | 39.8 ±0.1 | 85.7 ±0.1 | 28.4 ±0.1 | 19.7 ±0.1 | 43.4 ±0.1 |
| AttnLRP | 44.5 ±0.1 | 88.2 ±0.1 | 31.6 ±0.1 | 30.9 ±0.1 | 48.8 ±0.1 |
| DecompX | 44.0 ±0.1 | 87.1 ±0.1 | 30.7 ±0.1 | 23.2 ±0.1 | 46.3 ±0.1 |
| Integrated Gradients | 46.9 ±0.1 | 89.5 ±0.1 | 35.6 ±0.1 | 27.5 ±0.1 | 49.9 ±0.1 |
| Input × Grad | 40.4 ±0.1 | 87.0 ±0.1 | 33.2 ±0.1 | 20.9 ±0.1 | 45.4 ±0.1 |
| **Libra Input × Grad** | 44.8 ±0.1 (+10.8%) | 87.5 ±0.1 (+0.6%) | 30.7 ±0.1 (-7.5%) | 24.3 ±0.1 (+16.2%) | 46.8 ±0.1 (+3.2%) |
| AttCAT | 50.4 ±0.1 | 91.8 ±0.1 | 37.8 ±0.1 | 37.6 ±0.1 | 54.4 ±0.1 |
| **Libra AttCAT** | <u>66.4</u> ±0.1 (+31.7%) | <u>94.4</u> ±0.1 (+2.9%) | <u>47.1</u> ±0.1 (+24.5%) | <u>55.5</u> ±0.1 (+47.6%) | <u>65.9</u> ±0.1 (+21.0%) |
| GenAtt | 61.9 ±0.1 | 92.0 ±0.1 | 37.8 ±0.1 | 44.5 ±0.1 | 59.1 ±0.1 |
| **Libra GenAtt** | 65.1 ±0.1 (+5.1%) | 92.6 ±0.1 (+0.6%) | 38.0 ±0.1 (+0.7%) | 46.8 ±0.1 (+5.3%) | 60.6 ±0.1 (+2.7%) |
| TokenTM | 60.6 ±0.1 | 90.9 ±0.1 | 36.1 ±0.1 | 44.4 ±0.1 | 58.0 ±0.1 |
| **Libra TokenTM** | 62.8 ±0.1 (+3.6%) | 91.4 ±0.1 (+0.5%) | 36.0 ±0.1 (-0.1%) | 45.9 ±0.1 (+3.3%) | 59.0 ±0.1 (+1.7%) |
| GradCAM+ | 50.5 ±0.1 | 89.2 ±0.1 | 32.6 ±0.1 | 33.1 ±0.1 | 51.4 ±0.1 |
| **Libra GradCAM+** | 65.3 ±0.1 (+29.3%) | 92.7 ±0.1 (+3.9%) | 42.3 ±0.1 (+29.9%) | 48.2 ±0.1 (+45.8%) | 62.1 ±0.1 (+21.0%) |
| HiResCAM | 50.4 ±0.1 | 89.3 ±0.1 | 31.4 ±0.1 | 18.7 ±0.1 | 47.5 ±0.1 |
| **Libra HiResCAM** | 60.8 ±0.1 (+20.6%) | 91.4 ±0.1 (+2.4%) | 37.9 ±0.1 (+20.4%) | 40.2 ±0.1 (+114.4%) | 57.6 ±0.1 (+21.3%) |
| XGradCAM+ | 44.0 ±0.1 | 87.8 ±0.1 | 32.4 ±0.1 | 33.5 ±0.1 | 49.4 ±0.1 |
| **Libra XGradCAM+** | **67.4** ±0.1 (+53.0%) | 93.2 ±0.1 (+6.2%) | 43.4 ±0.1 (+34.1%) | 52.8 ±0.1 (+57.6%) | 64.2 ±0.1 (+29.9%) |
| FullGrad+ | 48.2 ±0.1 | 90.5 ±0.1 | 39.1 ±0.1 | 35.6 ±0.1 | 53.4 ±0.1 |
| **Libra FullGrad+** | 66.1 ±0.1 (+37.1%) | **94.7** ±0.1 (+4.6%) | **48.7** ±0.1 (+24.5%) | **57.5** ±0.1 (+61.6%) | **66.7** ±0.1 (+25.1%) |

Table 37. Cross-dataset analysis of Most-Influential-First Deletion (MIF) Accuracy evaluated using ground-truth labels on ViT-B.

| Method | ImageNet | ImageNet-Hard | MURA | Oxford-IIIT Pet | *Avg.* |
|---|---|---|---|---|---|
| Random | 14.2 ±0.2 | 16.4 ±0.2 | 4.2 ±0.1 | 4.3 ±0.1 | 9.8 ±0.1 |
| RawAtt | 27.9 ±0.3 | 23.8 ±0.2 | 14.9 ±0.2 | 29.6 ±0.3 | 24.0 ±0.2 |
| Attention Rollout | 21.2 ±0.2 | 21.7 ±0.2 | 10.8 ±0.2 | 12.4 ±0.2 | 16.5 ±0.2 |
| AliLRP | 19.1 ±0.2 | 22.6 ±0.2 | 8.5 ±0.2 | 10.0 ±0.2 | 15.1 ±0.2 |
| AttnLRP | 23.4 ±0.2 | 26.3 ±0.2 | 12.1 ±0.2 | 22.5 ±0.2 | 21.1 ±0.2 |
| DecompX | 22.8 ±0.2 | 25.0 ±0.2 | 10.9 ±0.2 | 14.1 ±0.2 | 18.2 ±0.2 |
| Integrated Gradients | 21.4 ±0.2 | 24.6 ±0.3 | 13.8 ±0.2 | 11.5 ±0.2 | 17.8 ±0.2 |
| Input × Grad | 20.2 ±0.2 | 24.3 ±0.2 | 14.1 ±0.2 | 11.6 ±0.2 | 17.6 ±0.2 |
| **Libra Input × Grad** | 23.4 ±0.2 (+15.8%) | 25.3 ±0.2 (+3.7%) | 10.9 ±0.2 (-23.0%) | 15.1 ±0.2 (+30.0%) | 18.6 ±0.2 (+6.2%) |
| AttCAT | 28.8 ±0.2 | 31.9 ±0.2 | 19.6 ±0.1 | 27.3 ±0.4 | 26.9 ±0.2 |
| **Libra AttCAT** | **41.5** ±0.3 (+44.2%) | <u>35.5</u> ±0.2 (+11.2%) | <u>28.9</u> ±0.2 (+47.7%) | <u>44.9</u> ±0.3 (+64.7%) | <u>37.7</u> ±0.3 (+40.2%) |
| GenAtt | 37.9 ±0.2 | 32.2 ±0.2 | 21.2 ±0.2 | 35.7 ±0.3 | 31.8 ±0.3 |
| **Libra GenAtt** | 40.4 ±0.3 (+6.6%) | 33.1 ±0.2 (+2.6%) | 21.1 ±0.2 (-0.6%) | 38.1 ±0.3 (+6.6%) | 33.2 ±0.3 (+4.4%) |
| TokenTM | 37.4 ±0.3 | 31.3 ±0.2 | 19.5 ±0.2 | 36.1 ±0.3 | 31.1 ±0.3 |
| **Libra TokenTM** | 38.9 ±0.3 (+3.8%) | 31.7 ±0.2 (+1.4%) | 19.2 ±0.2 (-1.7%) | 37.5 ±0.3 (+3.9%) | 31.8 ±0.3 (+2.4%) |
| GradCAM+ | 27.6 ±0.2 | 28.4 ±0.2 | 12.8 ±0.2 | 22.8 ±0.3 | 22.9 ±0.2 |
| **Libra GradCAM+** | 39.6 ±0.2 (+43.5%) | 33.2 ±0.2 (+17.0%) | 22.4 ±0.2 (+75.5%) | 38.6 ±0.3 (+69.7%) | 33.5 ±0.2 (+46.3%) |
| HiResCAM | 28.5 ±0.2 | 28.2 ±0.2 | 11.8 ±0.2 | 8.7 ±0.2 | 19.3 ±0.2 |
| **Libra HiResCAM** | 37.0 ±0.2 (+29.6%) | 31.4 ±0.2 (+11.3%) | 19.2 ±0.2 (+63.0%) | 30.9 ±0.4 (+254.4%) | 29.6 ±0.3 (+53.4%) |
| XGradCAM+ | 21.5 ±0.2 | 26.4 ±0.2 | 12.3 ±0.1 | 23.5 ±0.3 | 20.9 ±0.2 |
| **Libra XGradCAM+** | **41.5** ±0.2 (+92.8%) | 33.9 ±0.2 (+28.3%) | 25.2 ±0.3 (+104.9%) | 42.8 ±0.3 (+81.7%) | 35.8 ±0.3 (+71.1%) |
| FullGrad+ | 26.3 ±0.2 | 30.5 ±0.2 | 20.7 ±0.2 | 25.0 ±0.3 | 25.6 ±0.2 |
| **Libra FullGrad+** | 41.2 ±0.3 (+56.7%) | **35.6** ±0.2 (+16.6%) | **30.5** ±0.2 (+47.1%) | **46.7** ±0.3 (+86.8%) | **38.5** ±0.2 (+50.2%) |

Table 38. Cross-dataset analysis of Most-Influential-First Deletion (MIF) AOPC evaluated using predicted labels on ViT-B.

| Method | ImageNet | ImageNet-Hard | MURA | Oxford-IIIT Pet | *Avg.* |
|---|---|---|---|---|---|
| Random | 12.3 ±0.2 | 6.6 ±0.1 | 3.3 ±0.1 | 4.1 ±0.1 | 6.6 ±0.1 |
| RawAtt | 25.0 ±0.3 | 9.5 ±0.1 | 12.6 ±0.3 | 29.1 ±0.3 | 19.0 ±0.3 |
| Attention Rollout | 18.8 ±0.3 | 8.7 ±0.1 | 8.6 ±0.2 | 12.2 ±0.2 | 12.1 ±0.2 |
| AliLRP | 16.7 ±0.2 | 9.1 ±0.1 | 7.1 ±0.2 | 9.7 ±0.2 | 10.7 ±0.2 |
| AttnLRP | 20.8 ±0.3 | 10.8 ±0.2 | 10.4 ±0.2 | 22.1 ±0.2 | 16.0 ±0.2 |
| DecompX | 20.3 ±0.3 | 10.2 ±0.1 | 9.5 ±0.2 | 13.8 ±0.2 | 13.4 ±0.2 |
| Integrated Gradients | 22.5 ±0.2 | 11.7 ±0.1 | 14.5 ±0.2 | 17.6 ±0.3 | 16.6 ±0.2 |
| Input × Grad | 17.7 ±0.2 | 9.8 ±0.1 | 12.0 ±0.2 | 11.2 ±0.2 | 12.7 ±0.2 |
| **Libra Input × Grad** | 20.8 ±0.3 (+17.4%) | 10.3 ±0.1 (+5.3%) | 9.5 ±0.2 (-20.9%) | 14.8 ±0.2 (+32.2%) | 13.8 ±0.2 (+9.3%) |
| AttCAT | 25.3 ±0.2 | 13.1 ±0.1 | 16.7 ±0.2 | 26.7 ±0.3 | 20.5 ±0.2 |
| **Libra AttCAT** | <u>37.5</u> ±0.3 (+47.9%) | <u>14.9</u> ±0.1 (+14.0%) | <u>25.6</u> ±0.3 (+53.0%) | <u>44.3</u> ±0.3 (+65.8%) | <u>30.6</u> ±0.3 (+49.3%) |
| GenAtt | 34.2 ±0.3 | 13.5 ±0.1 | 18.1 ±0.3 | 35.2 ±0.3 | 25.2 ±0.3 |
| **Libra GenAtt** | 36.6 ±0.3 (+6.8%) | 13.8 ±0.1 (+2.8%) | 18.0 ±0.3 (-0.2%) | 37.6 ±0.3 (+6.7%) | 26.5 ±0.3 (+5.0%) |
| TokenTM | 33.8 ±0.3 | 12.9 ±0.1 | 16.4 ±0.3 | 35.6 ±0.3 | 24.7 ±0.3 |
| **Libra TokenTM** | 35.1 ±0.3 (+4.0%) | 13.1 ±0.1 (+2.0%) | 16.2 ±0.3 (-1.5%) | 37.0 ±0.3 (+4.0%) | 25.4 ±0.3 (+2.8%) |
| GradCAM+ | 24.8 ±0.2 | 11.4 ±0.1 | 11.0 ±0.2 | 22.2 ±0.3 | 17.4 ±0.2 |
| **Libra GradCAM+** | 35.9 ±0.2 (+44.8%) | 13.8 ±0.1 (+21.2%) | 20.1 ±0.3 (+82.2%) | 38.0 ±0.3 (+71.0%) | 27.0 ±0.2 (+55.3%) |
| HiResCAM | 25.4 ±0.3 | 11.5 ±0.1 | 10.4 ±0.2 | 8.5 ±0.2 | 13.9 ±0.2 |
| **Libra HiResCAM** | 33.4 ±0.3 (+31.7%) | 12.9 ±0.1 (+12.7%) | 17.0 ±0.2 (+63.5%) | 30.7 ±0.4 (+260.3%) | 23.5 ±0.3 (+68.7%) |
| XGradCAM+ | 19.0 ±0.2 | 10.6 ±0.1 | 10.6 ±0.2 | 23.0 ±0.3 | 15.8 ±0.2 |
| **Libra XGradCAM+** | **37.7** ±0.2 (+98.6%) | 14.1 ±0.1 (+33.7%) | 22.2 ±0.3 (+108.5%) | 42.2 ±0.3 (+83.3%) | 29.0 ±0.3 (+83.8%) |
| FullGrad+ | 23.1 ±0.3 | 12.3 ±0.1 | 17.7 ±0.2 | 24.5 ±0.3 | 19.4 ±0.2 |
| **Libra FullGrad+** | 37.2 ±0.3 (+60.9%) | **15.0** ±0.1 (+22.4%) | **26.9** ±0.3 (+51.7%) | **46.1** ±0.3 (+88.1%) | **31.3** ±0.3 (+61.3%) |

Table 39. Cross-dataset analysis of Most-Influential-First Deletion (MIF) AOPC evaluated using ground-truth labels on ViT-B.

| Method | ImageNet | ImageNet-Hard | MURA | Oxford-IIIT Pet | *Avg.* |
|---|---|---|---|---|---|
| Random | 73.3 ±0.1 | 47.0 ±0.1 | 85.8 ±0.1 | 85.8 ±0.1 | 72.9 ±0.1 |
| RawAtt | 76.2 ±0.1 | 52.0 ±0.1 | 85.7 ±0.1 | 86.0 ±0.1 | 75.0 ±0.1 |
| Attention Rollout | 73.8 ±0.1 | 49.8 ±0.1 | 84.1 ±0.1 | 82.4 ±0.1 | 72.5 ±0.1 |
| AliLRP | 77.8 ±0.1 | 53.5 ±0.1 | 87.0 ±0.1 | 87.7 ±0.0 | 76.5 ±0.1 |
| AttnLRP | 78.7 ±0.1 | 61.6 ±0.1 | 87.0 ±0.1 | 88.7 ±0.0 | 79.0 ±0.1 |
| DecompX | 79.1 ±0.1 | 56.8 ±0.1 | 87.7 ±0.0 | 88.1 ±0.0 | 77.9 ±0.1 |
| Integrated Gradients | 78.0 ±0.1 | 54.2 ±0.1 | 86.3 ±0.1 | 87.0 ±0.1 | 76.4 ±0.1 |
| Input × Grad | 77.3 ±0.1 | 55.9 ±0.1 | 88.2 ±0.0 | 88.7 ±0.0 | 77.5 ±0.1 |
| **Libra Input × Grad** | 80.2 ±0.1 (+3.8%) | 57.9 ±0.1 (+3.5%) | 87.7 ±0.0 (-0.5%) | 88.3 ±0.0 (-0.5%) | 78.5 ±0.1 (+1.3%) |
| AttCAT | 82.5 ±0.1 | 69.2 ±0.1 | 89.1 ±0.0 | 89.3 ±0.0 | 82.5 ±0.1 |
| **Libra AttCAT** | 86.7 ±0.1 (+5.1%) | 75.9 ±0.1 (+9.5%) | 89.4 ±0.0 (+0.3%) | 89.3 ±0.0 (+0.0%) | 85.3 ±0.1 (+3.4%) |
| GenAtt | 84.0 ±0.1 | 65.7 ±0.1 | 88.3 ±0.0 | 88.7 ±0.0 | 81.7 ±0.1 |
| **Libra GenAtt** | 84.4 ±0.1 (+0.4%) | 66.5 ±0.1 (+1.1%) | 88.3 ±0.0 (+0.1%) | 88.4 ±0.0 (-0.3%) | 81.9 ±0.1 (+0.2%) |
| TokenTM | 83.1 ±0.1 | 62.9 ±0.1 | 87.4 ±0.1 | 88.4 ±0.0 | 80.4 ±0.1 |
| **Libra TokenTM** | 83.2 ±0.1 (+0.1%) | 63.0 ±0.1 (+0.3%) | 87.5 ±0.0 (+0.2%) | 88.2 ±0.0 (-0.3%) | 80.5 ±0.1 (+0.0%) |
| GradCAM+ | 78.5 ±0.1 | 61.2 ±0.1 | 85.9 ±0.1 | 84.2 ±0.1 | 77.4 ±0.1 |
| **Libra GradCAM+** | 84.9 ±0.1 (+8.3%) | 68.6 ±0.1 (+12.2%) | 88.6 ±0.0 (+3.1%) | 88.4 ±0.0 (+5.0%) | 82.6 ±0.1 (+6.7%) |
| HiResCAM | 79.5 ±0.1 | 57.7 ±0.1 | 86.1 ±0.1 | 81.6 ±0.1 | 76.2 ±0.1 |
| **Libra HiResCAM** | 82.7 ±0.1 (+4.0%) | 62.0 ±0.1 (+7.6%) | 87.8 ±0.0 (+1.9%) | 86.2 ±0.1 (+5.7%) | 79.7 ±0.1 (+4.5%) |
| XGradCAM+ | 73.3 ±0.1 | 58.9 ±0.1 | 85.8 ±0.1 | 85.2 ±0.1 | 75.8 ±0.1 |
| **Libra XGradCAM+** | 85.4 ±0.1 (+16.6%) | 69.9 ±0.1 (+18.5%) | 88.2 ±0.0 (+2.8%) | 88.6 ±0.0 (+4.0%) | 83.0 ±0.1 (+9.5%) |
| FullGrad+ | 81.6 ±0.1 | 67.3 ±0.1 | **89.5** ±0.0 | 89.3 ±0.0 | 81.9 ±0.1 |
| **Libra FullGrad+** | **87.0** ±0.0 (+6.6%) | **76.2** ±0.1 (+13.4%) | **89.5** ±0.0 (+0.0%) | **89.6** ±0.0 (+0.3%) | **85.6** ±0.1 (+4.5%) |

Table 40. Cross-dataset analysis of Least-Influential-First Deletion (LIF) Accuracy evaluated using predicted labels on ViT-B.

| Method | ImageNet | ImageNet-Hard | MURA | Oxford-IIIT Pet | *Avg.* |
|---|---|---|---|---|---|
| Random | 65.2 ±0.1 | 18.6 ±0.1 | 75.2 ±0.1 | 84.9 ±0.1 | 61.0 ±0.1 |
| RawAtt | 67.5 ±0.1 | 19.5 ±0.1 | 75.9 ±0.1 | 85.0 ±0.1 | 62.0 ±0.1 |
| Attention Rollout | 65.9 ±0.1 | 19.8 ±0.1 | 75.0 ±0.1 | 81.6 ±0.1 | 60.6 ±0.1 |
| AliLRP | 69.9 ±0.1 | 21.5 ±0.1 | 77.8 ±0.1 | 86.9 ±0.1 | 64.0 ±0.1 |
| AttnLRP | 71.0 ±0.1 | 26.4 ±0.1 | 78.2 ±0.1 | 88.0 ±0.0 | 65.9 ±0.1 |
| DecompX | 71.1 ±0.1 | 22.6 ±0.1 | 78.7 ±0.1 | 87.1 ±0.1 | 64.9 ±0.1 |
| Integrated Gradients | 74.4 ±0.1 | 28.6 ±0.1 | 81.2 ±0.1 | 88.1 ±0.0 | 68.1 ±0.1 |
| Input × Grad | 69.9 ±0.1 | 23.4 ±0.1 | 80.4 ±0.1 | 88.2 ±0.0 | 65.5 ±0.1 |
| **Libra Input × Grad** | 72.5 ±0.1 (+3.7%) | 23.8 ±0.1 (+1.6%) | 78.7 ±0.1 (-2.0%) | 87.4 ±0.0 (-0.9%) | 65.6 ±0.1 (+0.2%) |
| AttCAT | 76.8 ±0.1 | 35.1 ±0.1 | 82.4 ±0.1 | 89.0 ±0.0 | 70.8 ±0.1 |
| **Libra AttCAT** | 80.2 ±0.1 (+4.5%) | 35.8 ±0.1 (+1.8%) | 83.2 ±0.1 (+0.9%) | 88.9 ±0.0 (+0.0%) | 72.0 ±0.1 (+1.7%) |
| GenAtt | 74.8 ±0.1 | 25.4 ±0.1 | 78.3 ±0.1 | 87.8 ±0.0 | 66.6 ±0.1 |
| **Libra GenAtt** | 75.1 ±0.1 (+0.4%) | 25.4 ±0.1 (+0.1%) | 78.3 ±0.1 (+0.0%) | 87.6 ±0.0 (-0.3%) | 66.6 ±0.1 (+0.0%) |
| TokenTM | 73.5 ±0.1 | 23.9 ±0.1 | 77.1 ±0.1 | 87.4 ±0.0 | 65.5 ±0.1 |
| **Libra TokenTM** | 73.6 ±0.1 (+0.1%) | 24.3 ±0.1 (+1.8%) | 77.1 ±0.1 (+0.0%) | 87.2 ±0.1 (-0.2%) | 65.6 ±0.1 (+0.1%) |
| GradCAM+ | 72.0 ±0.1 | 27.9 ±0.1 | 77.3 ±0.1 | 83.7 ±0.1 | 65.2 ±0.1 |
| **Libra GradCAM+** | 78.0 ±0.1 (+8.3%) | 30.2 ±0.1 (+8.3%) | 80.9 ±0.1 (+4.7%) | 87.9 ±0.0 (+5.0%) | 69.3 ±0.1 (+6.2%) |
| HiResCAM | 71.9 ±0.1 | 24.7 ±0.1 | 76.8 ±0.1 | 81.1 ±0.1 | 63.6 ±0.1 |
| **Libra HiResCAM** | 75.6 ±0.1 (+5.1%) | 26.6 ±0.1 (+7.7%) | 80.1 ±0.1 (+4.3%) | 85.5 ±0.1 (+5.5%) | 67.0 ±0.1 (+5.2%) |
| XGradCAM+ | 67.0 ±0.1 | 26.9 ±0.1 | 77.1 ±0.1 | 84.8 ±0.1 | 64.0 ±0.1 |
| **Libra XGradCAM+** | 78.1 ±0.1 (+16.6%) | 30.3 ±0.1 (+12.3%) | 81.3 ±0.1 (+5.5%) | 88.2 ±0.0 (+4.0%) | 69.5 ±0.1 (+8.6%) |
| FullGrad+ | 75.2 ±0.1 | 32.3 ±0.1 | 83.2 ±0.1 | 88.9 ±0.0 | 69.9 ±0.1 |
| **Libra FullGrad+** | **80.6** ±0.1 (+7.1%) | **36.1** ±0.1 (+12.0%) | **84.0** ±0.1 (+0.9%) | **89.1** ±0.0 (+0.2%) | **72.4** ±0.1 (+3.6%) |

Table 41. Cross-dataset analysis of Least-Influential-First Deletion (LIF) Accuracy evaluated using ground-truth labels on ViT-B.

| Method | ImageNet | ImageNet-Hard | MURA | Oxford-IIIT Pet | *Avg.* |
|---|---|---|---|---|---|
| Random | 85.8 ±0.2 | 83.1 ±0.2 | 96.2 ±0.1 | 95.4 ±0.1 | 90.1 ±0.2 |
| RawAtt | 87.6 ±0.1 | 85.5 ±0.2 | 96.6 ±0.1 | 96.1 ±0.1 | 91.5 ±0.1 |
| Attention Rollout | 86.0 ±0.2 | 84.1 ±0.2 | 94.9 ±0.1 | 91.7 ±0.2 | 89.2 ±0.2 |
| AliLRP | 89.3 ±0.2 | 87.3 ±0.2 | 98.5 ±0.1 | 98.0 ±0.1 | 93.3 ±0.2 |
| AttnLRP | 90.8 ±0.1 | 93.4 ±0.2 | 98.7 ±0.1 | 99.4 ±0.1 | 95.6 ±0.1 |
| DecompX | 90.6 ±0.1 | 89.4 ±0.2 | 99.4 ±0.1 | 98.3 ±0.1 | 94.4 ±0.2 |
| Integrated Gradients | 91.3 ±0.1 | 89.7 ±0.2 | 99.2 ±0.1 | 97.5 ±0.1 | 94.4 ±0.1 |
| Input × Grad | 90.2 ±0.1 | 89.9 ±0.2 | 100.5 ±0.1 | 99.4 ±0.1 | 95.0 ±0.2 |
| **Libra Input × Grad** | 91.3 ±0.2 (+1.2%) | 90.1 ±0.2 (+0.2%) | 99.4 ±0.1 (-1.1%) | 98.7 ±0.1 (-0.7%) | 94.9 ±0.1 (-0.1%) |
| AttCAT | 96.6 ±0.2 | 102.1 ±0.2 | 102.1 ±0.1 | <u>100.9</u> ±0.1 | 100.4 ±0.1 |
| **Libra AttCAT** | <u>99.2</u> ±0.1 (+2.7%) | <u>105.6</u> ±0.2 (+3.5%) | 102.4 ±0.1 (+0.2%) | 100.8 ±0.1 (-0.1%) | <u>102.0</u> ±0.1 (+1.6%) |
| GenAtt | 94.6 ±0.1 | 93.4 ±0.2 | 99.3 ±0.1 | 99.2 ±0.1 | 96.6 ±0.1 |
| **Libra GenAtt** | 94.8 ±0.1 (+0.2%) | 93.4 ±0.2 (+0.1%) | 99.2 ±0.1 (-0.1%) | 98.6 ±0.1 (-0.6%) | 96.5 ±0.1 (-0.1%) |
| TokenTM | 93.3 ±0.1 | 91.1 ±0.2 | 98.2 ±0.1 | 98.7 ±0.1 | 95.3 ±0.1 |
| **Libra TokenTM** | 93.5 ±0.2 (+0.2%) | 91.3 ±0.2 (+0.2%) | 98.3 ±0.1 (+0.1%) | 98.3 ±0.1 (-0.5%) | 95.3 ±0.1 (+0.0%) |
| GradCAM+ | 91.5 ±0.2 | 93.9 ±0.2 | 96.9 ±0.2 | 94.1 ±0.2 | 94.1 ±0.2 |
| **Libra GradCAM+** | 96.2 ±0.1 (+5.2%) | 98.3 ±0.2 (+4.6%) | 100.6 ±0.1 (+3.8%) | 99.1 ±0.1 (+5.3%) | 98.6 ±0.1 (+4.7%) |
| HiResCAM | 91.7 ±0.2 | 90.6 ±0.2 | 97.6 ±0.1 | 90.9 ±0.2 | 92.7 ±0.2 |
| **Libra HiResCAM** | 94.3 ±0.1 (+2.8%) | 93.3 ±0.2 (+3.0%) | 99.8 ±0.1 (+2.3%) | 96.3 ±0.2 (+5.9%) | 95.9 ±0.1 (+3.5%) |
| XGradCAM+ | 86.6 ±0.2 | 92.5 ±0.2 | 97.0 ±0.1 | 95.5 ±0.2 | 92.9 ±0.2 |
| **Libra XGradCAM+** | 96.6 ±0.1 (+11.6%) | 99.0 ±0.2 (+7.1%) | 100.4 ±0.1 (+3.5%) | 99.5 ±0.1 (+4.2%) | 98.9 ±0.1 (+6.5%) |
| FullGrad+ | 95.0 ±0.2 | 99.8 ±0.2 | **102.7** ±0.1 | 100.8 ±0.1 | 99.6 ±0.1 |
| **Libra FullGrad+** | **99.6** ±0.1 (+4.8%) | **106.3** ±0.2 (+6.5%) | **102.7** ±0.1 (+0.0%) | **101.1** ±0.1 (+0.3%) | **102.4** ±0.1 (+2.8%) |

Table 42. Cross-dataset analysis of Least-Influential-First Deletion (LIF) AOPC evaluated using predicted labels on ViT-B.

| Method | ImageNet | ImageNet-Hard | MURA | Oxford-IIIT Pet | *Avg.* |
|---|---|---|---|---|---|
| Random | 87.5 ±0.2 | 93.5 ±0.1 | 97.0 ±0.1 | 95.6 ±0.1 | 93.4 ±0.2 |
| RawAtt | 89.2 ±0.1 | 94.2 ±0.1 | 97.7 ±0.1 | 96.3 ±0.2 | 94.4 ±0.1 |
| Attention Rollout | 87.9 ±0.2 | 94.0 ±0.1 | 96.4 ±0.1 | 92.0 ±0.2 | 92.6 ±0.2 |
| AliLRP | 90.9 ±0.2 | 95.2 ±0.1 | 99.9 ±0.1 | 98.3 ±0.1 | 96.1 ±0.2 |
| AttnLRP | 92.4 ±0.2 | 98.2 ±0.1 | 100.4 ±0.1 | 99.7 ±0.1 | 97.7 ±0.1 |
| DecompX | 92.2 ±0.2 | 96.2 ±0.1 | 100.9 ±0.1 | 98.4 ±0.1 | 96.9 ±0.1 |
| Integrated Gradients | 95.7 ±0.2 | 99.7 ±0.1 | 103.7 ±0.1 | 100.0 ±0.1 | 99.8 ±0.1 |
| Input × Grad | 91.8 ±0.2 | 96.6 ±0.1 | 102.6 ±0.1 | 99.9 ±0.1 | 97.7 ±0.1 |
| **Libra Input × Grad** | 93.1 ±0.2 (+1.4%) | 96.5 ±0.1 (-0.1%) | 100.9 ±0.1 (-1.7%) | 99.0 ±0.1 (-0.8%) | 97.4 ±0.1 (-0.4%) |
| AttCAT | 98.4 ±0.2 | 104.1 ±0.1 | 105.0 ±0.1 | <u>101.4</u> ±0.1 | 102.2 ±0.1 |
| **Libra AttCAT** | <u>100.8</u> ±0.2 (+2.4%) | <u>104.4</u> ±0.2 (+0.4%) | <u>105.7</u> ±0.2 (+0.7%) | 101.3 ±0.1 (+0.0%) | <u>103.1</u> ±0.1 (+0.8%) |
| GenAtt | 95.7 ±0.2 | 97.8 ±0.1 | 100.4 ±0.1 | 99.4 ±0.1 | 98.4 ±0.1 |
| **Libra GenAtt** | 96.0 ±0.1 (+0.3%) | 97.9 ±0.1 (+0.1%) | 100.3 ±0.1 (-0.1%) | 98.9 ±0.1 (-0.6%) | 98.3 ±0.1 (-0.1%) |
| TokenTM | 94.4 ±0.1 | 96.9 ±0.1 | 99.1 ±0.1 | 99.0 ±0.1 | 97.4 ±0.1 |
| **Libra TokenTM** | 94.6 ±0.1 (+0.2%) | 97.0 ±0.1 (+0.1%) | 99.1 ±0.1 (+0.0%) | 98.5 ±0.1 (-0.4%) | 97.3 ±0.1 (+0.0%) |
| GradCAM+ | 93.5 ±0.2 | 98.8 ±0.1 | 98.7 ±0.2 | 94.6 ±0.2 | 96.4 ±0.2 |
| **Libra GradCAM+** | 98.1 ±0.2 (+4.9%) | 100.4 ±0.2 (+1.6%) | 103.0 ±0.1 (+4.3%) | 99.6 ±0.1 (+5.3%) | 100.3 ±0.1 (+4.0%) |
| HiResCAM | 93.3 ±0.2 | 97.1 ±0.1 | 99.0 ±0.2 | 91.4 ±0.2 | 95.2 ±0.2 |
| **Libra HiResCAM** | 96.1 ±0.2 (+3.0%) | 98.2 ±0.1 (+1.2%) | 102.0 ±0.1 (+3.1%) | 96.6 ±0.2 (+5.7%) | 98.2 ±0.2 (+3.2%) |
| XGradCAM+ | 88.9 ±0.3 | 98.3 ±0.1 | 98.7 ±0.2 | 96.0 ±0.2 | 95.4 ±0.2 |
| **Libra XGradCAM+** | 98.4 ±0.2 (+10.8%) | 100.7 ±0.2 (+2.5%) | 103.4 ±0.2 (+4.8%) | 100.0 ±0.1 (+4.2%) | 100.6 ±0.2 (+5.4%) |
| FullGrad+ | 96.6 ±0.2 | 102.0 ±0.1 | <u>105.7</u> ±0.1 | 101.2 ±0.1 | 101.4 ±0.1 |
| **Libra FullGrad+** | **101.2** ±0.2 (+4.8%) | **104.7** ±0.1 (+2.7%) | **106.3** ±0.2 (+0.5%) | **101.5** ±0.1 (+0.3%) | **103.4** ±0.1 (+2.0%) |

Table 43. Cross-dataset analysis of Least-Influential-First Deletion (LIF) AOPC evaluated using ground-truth labels on ViT-B.

| Method | ImageNet | ImageNet-Hard | MURA | Oxford-IIIT Pet | *Avg.* |
|---|---|---|---|---|---|
| Random | 49.9 ±0.1 | 49.7 ±0.1 | 50.4 ±0.1 | 49.7 ±0.1 | 49.9 ±0.1 |
| RawAtt | 60.4 ±0.1 | 58.9 ±0.1 | 55.3 ±0.1 | 61.6 ±0.1 | 59.1 ±0.1 |
| Attention Rollout | 54.6 ±0.1 | 56.0 ±0.1 | 52.8 ±0.1 | 51.8 ±0.1 | 53.8 ±0.1 |
| AliLRP | 55.5 ±0.1 | 58.8 ±0.1 | 53.1 ±0.1 | 53.3 ±0.1 | 55.2 ±0.1 |
| AttnLRP | 58.6 ±0.1 | 66.2 ±0.1 | 54.9 ±0.1 | 59.5 ±0.1 | 59.8 ±0.1 |
| DecompX | 58.5 ±0.1 | 62.2 ±0.1 | 54.7 ±0.1 | 55.3 ±0.1 | 57.7 ±0.1 |
| Integrated Gradients | 56.7 ±0.1 | 60.4 ±0.1 | 55.1 ±0.1 | 53.8 ±0.1 | 56.5 ±0.1 |
| Input × Grad | 55.9 ±0.1 | 61.8 ±0.1 | 56.8 ±0.1 | 54.5 ±0.1 | 57.2 ±0.1 |
| **Libra Input × Grad** | 59.4 ±0.1 (+6.3%) | 63.3 ±0.1 (+2.6%) | 54.7 ±0.1 (-3.8%) | 55.9 ±0.1 (+2.5%) | 58.3 ±0.1 (+1.9%) |
| AttCAT | 64.7 ±0.1 | 75.8 ±0.1 | 60.1 ±0.1 | 63.3 ±0.1 | 66.0 ±0.1 |
| **Libra AttCAT** | **75.1** ±0.1 (+16.1%) | <u>81.6</u> ±0.1 (+7.7%) | <u>65.2</u> ±0.1 (+8.4%) | <u>72.3</u> ±0.1 (+14.2%) | <u>73.5</u> ±0.1 (+11.5%) |
| GenAtt | 71.1 ±0.1 | 73.5 ±0.1 | 59.1 ±0.1 | 66.4 ±0.1 | 67.5 ±0.1 |
| **Libra GenAtt** | 73.0 ±0.1 (+2.6%) | 74.6 ±0.1 (+1.5%) | 59.2 ±0.1 (+0.1%) | 67.4 ±0.1 (+1.6%) | 68.6 ±0.1 (+1.5%) |
| TokenTM | 70.0 ±0.1 | 71.1 ±0.1 | 57.7 ±0.1 | 66.2 ±0.1 | 66.2 ±0.1 |
| **Libra TokenTM** | 71.1 ±0.1 (+1.7%) | 71.5 ±0.1 (+0.6%) | 57.8 ±0.1 (+0.2%) | 66.8 ±0.1 (+0.9%) | 66.8 ±0.1 (+0.9%) |
| GradCAM+ | 62.0 ±0.1 | 68.5 ±0.1 | 54.9 ±0.1 | 58.4 ±0.1 | 61.0 ±0.1 |
| **Libra GradCAM+** | 73.2 ±0.1 (+18.0%) | 76.0 ±0.1 (+11.0%) | 61.6 ±0.1 (+12.2%) | 68.1 ±0.1 (+16.6%) | 69.7 ±0.1 (+14.4%) |
| HiResCAM | 62.5 ±0.1 | 65.9 ±0.1 | 54.1 ±0.1 | 49.8 ±0.1 | 58.1 ±0.1 |
| **Libra HiResCAM** | 69.7 ±0.1 (+11.6%) | 70.9 ±0.1 (+7.5%) | 58.9 ±0.1 (+8.9%) | 62.8 ±0.1 (+26.2%) | 65.6 ±0.1 (+12.9%) |
| XGradCAM+ | 55.9 ±0.1 | 65.5 ±0.1 | 54.7 ±0.1 | 59.2 ±0.1 | 58.8 ±0.1 |
| **Libra XGradCAM+** | 74.6 ±0.1 (+33.5%) | 77.3 ±0.1 (+17.9%) | 62.4 ±0.1 (+14.0%) | 70.6 ±0.1 (+19.3%) | 71.2 ±0.1 (+21.0%) |
| FullGrad+ | 62.9 ±0.1 | 73.7 ±0.1 | 61.2 ±0.1 | 62.3 ±0.1 | 65.0 ±0.1 |
| **Libra FullGrad+** | <u>75.0</u> ±0.1 (+19.4%) | **81.9** ±0.1 (+11.2%) | **66.4** ±0.1 (+8.5%) | **73.4** ±0.1 (+17.9%) | **74.2** ±0.1 (+14.2%) |

Table 44. Cross-dataset analysis of Symmetric Relevance Gain (SRG) Accuracy evaluated using predicted labels on ViT-B.

| Method | ImageNet | ImageNet-Hard | MURA | Oxford-IIIT Pet | *Avg.* |
|---|---|---|---|---|---|
| Random | 49.9 ±0.1 | 50.1 ±0.1 | 50.3 ±0.1 | 49.7 ±0.1 | 50.0 ±0.1 |
| RawAtt | 58.8 ±0.1 | 52.7 ±0.1 | 54.7 ±0.1 | 61.3 ±0.1 | 56.9 ±0.1 |
| Attention Rollout | 53.9 ±0.1 | 52.3 ±0.1 | 52.5 ±0.1 | 51.9 ±0.1 | 52.6 ±0.1 |
| AliLRP | 54.8 ±0.1 | 53.6 ±0.1 | 53.1 ±0.1 | 53.3 ±0.1 | 53.7 ±0.1 |
| AttnLRP | 57.8 ±0.1 | 57.3 ±0.1 | 54.9 ±0.1 | 59.4 ±0.1 | 57.4 ±0.1 |
| DecompX | 57.6 ±0.1 | 54.8 ±0.1 | 54.7 ±0.1 | 55.1 ±0.1 | 55.6 ±0.1 |
| Integrated Gradients | 60.6 ±0.1 | 59.0 ±0.1 | 58.4 ±0.1 | 57.8 ±0.1 | 59.0 ±0.1 |
| Input × Grad | 55.1 ±0.1 | 55.2 ±0.1 | 56.8 ±0.1 | 54.6 ±0.1 | 55.4 ±0.1 |
| **Libra Input × Grad** | 58.6 ±0.1 (+6.3%) | 55.7 ±0.1 (+0.8%) | 54.7 ±0.1 (-3.6%) | 55.9 ±0.1 (+2.4%) | 56.2 ±0.1 (+1.4%) |
| AttCAT | 63.6 ±0.1 | 63.5 ±0.1 | 60.1 ±0.1 | 63.3 ±0.1 | 62.6 ±0.1 |
| **Libra AttCAT** | **73.3** ±0.1 (+15.3%) | <u>65.1</u> ±0.1 (+2.6%) | <u>65.1</u> ±0.1 (+8.3%) | <u>72.2</u> ±0.1 (+14.1%) | <u>68.9</u> ±0.1 (+10.1%) |
| GenAtt | 68.4 ±0.1 | 58.7 ±0.1 | 58.0 ±0.1 | 66.1 ±0.1 | 62.8 ±0.1 |
| **Libra GenAtt** | 70.1 ±0.1 (+2.5%) | 59.0 ±0.1 (+0.5%) | 58.2 ±0.1 (+0.2%) | 67.2 ±0.1 (+1.6%) | 63.6 ±0.1 (+1.3%) |
| TokenTM | 67.1 ±0.1 | 57.4 ±0.1 | 56.6 ±0.1 | 65.9 ±0.1 | 61.8 ±0.1 |
| **Libra TokenTM** | 68.2 ±0.1 (+1.7%) | 57.9 ±0.1 (+0.8%) | 56.6 ±0.1 (+0.0%) | 66.6 ±0.1 (+1.0%) | 62.3 ±0.1 (+0.9%) |
| GradCAM+ | 61.3 ±0.1 | 58.6 ±0.1 | 54.9 ±0.1 | 58.4 ±0.1 | 58.3 ±0.1 |
| **Libra GradCAM+** | 71.7 ±0.1 (+17.0%) | 61.5 ±0.1 (+5.0%) | 61.6 ±0.1 (+12.1%) | 68.0 ±0.1 (+16.6%) | 65.7 ±0.1 (+12.7%) |
| HiResCAM | 61.2 ±0.1 | 57.0 ±0.1 | 54.1 ±0.1 | 49.9 ±0.1 | 55.5 ±0.1 |
| **Libra HiResCAM** | 68.2 ±0.1 (+11.5%) | 59.0 ±0.1 (+3.5%) | 59.0 ±0.1 (+9.0%) | 62.8 ±0.1 (+26.0%) | 62.3 ±0.1 (+12.1%) |
| XGradCAM+ | 55.5 ±0.1 | 57.4 ±0.1 | 54.7 ±0.1 | 59.1 ±0.1 | 56.7 ±0.1 |
| **Libra XGradCAM+** | 72.7 ±0.1 (+31.0%) | 61.8 ±0.1 (+7.6%) | 62.4 ±0.1 (+14.0%) | 70.5 ±0.1 (+19.2%) | 66.8 ±0.1 (+17.9%) |
| FullGrad+ | 61.7 ±0.1 | 61.4 ±0.1 | 61.2 ±0.1 | 62.2 ±0.1 | 61.6 ±0.1 |
| **Libra FullGrad+** | **73.3** ±0.1 (+18.8%) | **65.4** ±0.1 (+6.5%) | **66.4** ±0.1 (+8.5%) | **73.3** ±0.1 (+17.8%) | **69.6** ±0.1 (+12.9%) |

Table 45. Cross-dataset analysis of Symmetric Relevance Gain (SRG) Accuracy evaluated using ground-truth labels on ViT-B.

| Method | ImageNet | ImageNet-Hard | MURA | Oxford-IIIT Pet | *Avg.* |
|---|---|---|---|---|---|
| Random | 50.0 ±0.2 | 49.8 ±0.2 | 50.2 ±0.1 | 49.8 ±0.1 | 50.0 ±0.2 |
| RawAtt | 57.8 ±0.2 | 54.6 ±0.2 | 55.8 ±0.2 | 62.8 ±0.2 | 57.7 ±0.2 |
| Attention Rollout | 53.6 ±0.2 | 52.9 ±0.2 | 52.9 ±0.2 | 52.0 ±0.2 | 52.9 ±0.2 |
| AliLRP | 54.2 ±0.2 | 55.0 ±0.2 | 53.5 ±0.1 | 54.0 ±0.2 | 54.2 ±0.2 |
| AttnLRP | 57.1 ±0.2 | 59.9 ±0.2 | 55.4 ±0.2 | 60.9 ±0.2 | 58.3 ±0.2 |
| DecompX | 56.7 ±0.2 | 57.2 ±0.2 | 55.1 ±0.1 | 56.2 ±0.2 | 56.3 ±0.2 |
| Integrated Gradients | 56.3 ±0.2 | 57.2 ±0.2 | 56.5 ±0.1 | 54.5 ±0.1 | 56.1 ±0.2 |
| Input × Grad | 55.2 ±0.2 | 57.1 ±0.2 | 57.3 ±0.2 | 55.5 ±0.2 | 56.3 ±0.2 |
| **Libra Input × Grad** | 57.3 ±0.2 (+3.9%) | 57.7 ±0.2 (+1.0%) | 55.1 ±0.1 (-3.8%) | 56.9 ±0.2 (+2.5%) | 56.8 ±0.2 (+0.9%) |
| AttCAT | 62.7 ±0.2 | 67.0 ±0.2 | 60.8 ±0.1 | 64.1 ±0.3 | 63.7 ±0.2 |
| **Libra AttCAT** | **70.4** ±0.2 (+12.2%) | <u>70.5</u> ±0.2 (+5.3%) | <u>65.6</u> ±0.2 (+7.9%) | <u>72.9</u> ±0.2 (+13.7%) | <u>69.9</u> ±0.2 (+9.7%) |
| GenAtt | 66.3 ±0.2 | 62.8 ±0.2 | 60.3 ±0.2 | 67.5 ±0.2 | 64.2 ±0.2 |
| **Libra GenAtt** | 67.6 ±0.2 (+2.1%) | 63.3 ±0.2 (+0.7%) | 60.1 ±0.2 (-0.2%) | 68.3 ±0.2 (+1.3%) | 64.8 ±0.2 (+1.0%) |
| TokenTM | 65.3 ±0.2 | 61.2 ±0.2 | 58.8 ±0.2 | 67.4 ±0.2 | 63.2 ±0.2 |
| **Libra TokenTM** | 66.2 ±0.2 (+1.3%) | 61.5 ±0.2 (+0.5%) | 58.7 ±0.2 (-0.2%) | 67.9 ±0.2 (+0.7%) | 63.6 ±0.2 (+0.6%) |
| GradCAM+ | 59.5 ±0.2 | 61.2 ±0.2 | 54.8 ±0.2 | 58.5 ±0.2 | 58.5 ±0.2 |
| **Libra GradCAM+** | 67.9 ±0.2 (+14.1%) | 65.7 ±0.2 (+7.5%) | 61.5 ±0.2 (+12.2%) | 68.9 ±0.2 (+17.8%) | 66.0 ±0.2 (+12.8%) |
| HiResCAM | 60.1 ±0.2 | 59.4 ±0.2 | 54.7 ±0.1 | 49.8 ±0.2 | 56.0 ±0.2 |
| **Libra HiResCAM** | 65.7 ±0.2 (+9.2%) | 62.3 ±0.2 (+5.0%) | 59.5 ±0.2 (+8.8%) | 63.6 ±0.3 (+27.7%) | 62.8 ±0.2 (+12.1%) |
| XGradCAM+ | 54.1 ±0.2 | 59.4 ±0.2 | 54.6 ±0.1 | 59.5 ±0.2 | 56.9 ±0.2 |
| **Libra XGradCAM+** | 69.1 ±0.2 (+27.7%) | 66.5 ±0.2 (+11.8%) | 62.8 ±0.2 (+14.9%) | 71.1 ±0.2 (+19.6%) | 67.4 ±0.2 (+18.4%) |
| FullGrad+ | 60.7 ±0.2 | 65.2 ±0.2 | 61.7 ±0.1 | 62.9 ±0.2 | 62.6 ±0.2 |
| **Libra FullGrad+** | **70.4** ±0.2 (+16.0%) | **71.0** ±0.2 (+8.9%) | **66.6** ±0.2 (+7.9%) | **73.9** ±0.2 (+17.5%) | **70.5** ±0.2 (+12.5%) |

Table 46. Cross-dataset analysis of Symmetric Relevance Gain (SRG) AOPC evaluated using predicted labels on ViT-B.

| Method | ImageNet | ImageNet-Hard | MURA | Oxford-IIIT Pet | *Avg.* |
|---|---|---|---|---|---|
| Random | 49.9 ±0.2 | 50.0 ±0.1 | 50.2 ±0.1 | 49.9 ±0.1 | 50.0 ±0.1 |
| RawAtt | 57.1 ±0.2 | 51.8 ±0.1 | 55.2 ±0.2 | 62.7 ±0.2 | 56.7 ±0.2 |
| Attention Rollout | 53.3 ±0.2 | 51.3 ±0.1 | 52.5 ±0.2 | 52.1 ±0.2 | 52.3 ±0.2 |
| AliLRP | 53.8 ±0.2 | 52.1 ±0.1 | 53.5 ±0.2 | 54.0 ±0.2 | 53.4 ±0.2 |
| AttnLRP | 56.6 ±0.2 | 54.5 ±0.1 | 55.4 ±0.2 | 60.9 ±0.2 | 56.8 ±0.2 |
| DecompX | 56.3 ±0.2 | 53.2 ±0.1 | 55.2 ±0.2 | 56.1 ±0.2 | 55.2 ±0.2 |
| Integrated Gradients | 59.1 ±0.2 | 55.7 ±0.1 | 59.1 ±0.2 | 58.8 ±0.2 | 58.2 ±0.2 |
| Input × Grad | 54.8 ±0.2 | 53.2 ±0.1 | 57.3 ±0.2 | 55.5 ±0.2 | 55.2 ±0.2 |
| **Libra Input × Grad** | 56.9 ±0.2 (+4.0%) | 53.4 ±0.1 (+0.4%) | 55.2 ±0.2 (-3.7%) | 56.9 ±0.2 (+2.5%) | 55.6 ±0.2 (+0.7%) |
| AttCAT | 61.9 ±0.2 | 58.6 ±0.1 | 60.8 ±0.2 | 64.0 ±0.2 | 61.3 ±0.2 |
| **Libra AttCAT** | <u>69.1</u> ±0.2 (+11.7%) | <u>59.7</u> ±0.2 (+1.9%) | <u>65.6</u> ±0.2 (+7.9%) | <u>72.8</u> ±0.2 (+13.7%) | <u>66.8</u> ±0.2 (+8.9%) |
| GenAtt | 65.0 ±0.2 | 55.6 ±0.1 | 59.2 ±0.2 | 67.3 ±0.2 | 61.8 ±0.2 |
| **Libra GenAtt** | 66.3 ±0.2 (+2.0%) | 55.9 ±0.1 (+0.4%) | 59.2 ±0.2 (-0.1%) | 68.2 ±0.2 (+1.3%) | 62.4 ±0.2 (+0.9%) |
| TokenTM | 64.1 ±0.2 | 54.9 ±0.1 | 57.8 ±0.2 | 67.3 ±0.2 | 61.0 ±0.2 |
| **Libra TokenTM** | 64.9 ±0.2 (+1.2%) | 55.1 ±0.1 (+0.3%) | 57.7 ±0.2 (-0.2%) | 67.8 ±0.2 (+0.8%) | 61.3 ±0.2 (+0.5%) |
| GradCAM+ | 59.2 ±0.2 | 55.1 ±0.1 | 54.9 ±0.2 | 58.4 ±0.2 | 56.9 ±0.2 |
| **Libra GradCAM+** | 67.0 ±0.2 (+13.3%) | 57.1 ±0.1 (+3.7%) | 61.5 ±0.2 (+12.2%) | 68.8 ±0.2 (+17.8%) | 63.6 ±0.2 (+11.9%) |
| HiResCAM | 59.3 ±0.2 | 54.3 ±0.1 | 54.7 ±0.2 | 50.0 ±0.2 | 54.6 ±0.2 |
| **Libra HiResCAM** | 64.7 ±0.2 (+9.1%) | 55.6 ±0.1 (+2.4%) | 59.5 ±0.2 (+8.8%) | 63.6 ±0.3 (+27.4%) | 60.9 ±0.2 (+11.6%) |
| XGradCAM+ | 53.9 ±0.2 | 54.4 ±0.1 | 54.6 ±0.2 | 59.5 ±0.2 | 55.6 ±0.2 |
| **Libra XGradCAM+** | 68.1 ±0.2 (+26.2%) | 57.4 ±0.2 (+5.5%) | 62.8 ±0.3 (+14.9%) | 71.1 ±0.2 (+19.5%) | 64.8 ±0.2 (+16.6%) |
| FullGrad+ | 59.8 ±0.2 | 57.1 ±0.1 | 61.7 ±0.2 | 62.9 ±0.2 | 60.4 ±0.2 |
| **Libra FullGrad+** | **69.2** ±0.2 (+15.6%) | **59.9** ±0.1 (+4.8%) | **66.6** ±0.2 (+7.9%) | **73.8** ±0.2 (+17.4%) | **67.4** ±0.2 (+11.5%) |

Table 47. Cross-dataset analysis of Symmetric Relevance Gain (SRG) AOPC evaluated using ground-truth labels on ViT-B.

### D.5. Results Per Model

### D.5.1. MLP-Mixer-L

Since MLP-Mixer is an attention-free architecture, certain attribution methods couldn't be applied and were omitted.

| Method | MIF Deletion (GT) | | MIF Deletion (Predicted) | | Segmentation |
|---|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC | AP |
| Random | 48.7 ±0.1 | 20.3 ±0.3 | 42.0 ±0.1 | 25.8 ±0.2 | 43.2 ±0.4 |
| AliLRP | 64.6 ±0.1 | 33.9 ±0.3 | 60.2 ±0.1 | 41.0 ±0.2 | 58.6 ±0.3 |
| DecompX | 66.0 ±0.1 | 35.6 ±0.3 | 61.8 ±0.1 | 42.8 ±0.2 | 59.6 ±0.3 |
| Integrated Gradients | 62.2 ±0.1 | 30.8 ±0.2 | 53.0 ±0.1 | 34.7 ±0.2 | 54.3 ±0.3 |
| Input × Grad | 59.0 ±0.1 | 28.8 ±0.2 | 54.5 ±0.1 | 35.3 ±0.2 | 52.3 ±0.3 |
| **Libra Input × Grad** | **79.6** ±0.1 (+34.9%) | **43.6** ±0.3 (+51.5%) | **77.0** ±0.1 (+41.3%) | **51.4** ±0.2 (+45.5%) | <u>68.1</u> ±0.3 (+30.2%) |
| GradCAM+ | 62.2 ±0.1 | 31.1 ±0.3 | 57.7 ±0.1 | 37.9 ±0.2 | 52.2 ±0.4 |
| **Libra GradCAM+** | 66.3 ±0.1 (+6.6%) | 34.4 ±0.3 (+10.6%) | 61.9 ±0.1 (+7.2%) | 41.3 ±0.2 (+9.1%) | 57.8 ±0.3 (+10.9%) |
| HiResCAM | 54.2 ±0.1 | 25.3 ±0.3 | 48.2 ±0.1 | 31.3 ±0.2 | 47.4 ±0.4 |
| **Libra HiResCAM** | 55.0 ±0.1 (+1.4%) | 26.1 ±0.3 (+3.4%) | 48.9 ±0.1 (+1.4%) | 32.1 ±0.2 (+2.8%) | 50.5 ±0.3 (+6.5%) |
| XGradCAM+ | 62.8 ±0.1 | 31.7 ±0.3 | 58.3 ±0.1 | 38.4 ±0.2 | 53.3 ±0.4 |
| **Libra XGradCAM+** | 69.1 ±0.1 (+10.2%) | 36.4 ±0.3 (+15.0%) | 65.1 ±0.1 (+11.5%) | 43.5 ±0.2 (+13.3%) | 62.8 ±0.3 (+17.7%) |
| FullGrad+ | 64.0 ±0.1 | 31.7 ±0.3 | 60.2 ±0.1 | 38.6 ±0.3 | 53.3 ±0.3 |
| **Libra FullGrad+** | <u>76.0</u> ±0.1 (+18.8%) | <u>41.3</u> ±0.3 (+30.1%) | <u>73.1</u> ±0.1 (+21.5%) | <u>48.9</u> ±0.2 (+26.4%) | **70.1** ±0.3 (+31.4%) |

Table 48. Comparison of attribution methods and their LibraGrad-enhanced versions on the MLP-Mixer-L model. We report faithfulness metrics using Most-Influential-First Deletion, MIF with ground-truth (GT) and predicted labels, including Accuracy and Area Over Perturbation Curve (AOPC) and Segmentation Average Precision (AP). The results demonstrate that composing existing methods with LibraGrad significantly enhances their performance across all metrics.

| Method | LIF Deletion (GT) | | LIF Deletion (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 51.1 ±0.1 | 79.3 ±0.2 | 57.6 ±0.1 | 73.6 ±0.2 |
| AliLRP | 66.6 ±0.1 | 89.5 ±0.2 | 74.3 ±0.1 | 84.6 ±0.2 |
| DecompX | 66.3 ±0.1 | 89.8 ±0.2 | 74.1 ±0.1 | 84.9 ±0.2 |
| Integrated Gradients | 65.3 ±0.1 | 88.7 ±0.2 | 67.2 ±0.1 | 80.2 ±0.2 |
| Input × Grad | 61.8 ±0.1 | 85.4 ±0.3 | 69.0 ±0.1 | 80.2 ±0.2 |
| **Libra Input × Grad** | **74.2** ±0.1 (+19.9%) | **94.5** ±0.2 (+10.6%) | **81.6** ±0.1 (+18.2%) | **90.0** ±0.2 (+12.2%) |
| GradCAM+ | 63.7 ±0.1 | 87.2 ±0.2 | 70.6 ±0.1 | 81.8 ±0.2 |
| **Libra GradCAM+** | 66.6 ±0.1 (+4.6%) | 89.4 ±0.2 (+2.6%) | 73.7 ±0.1 (+4.5%) | 84.3 ±0.2 (+3.0%) |
| HiResCAM | 56.6 ±0.1 | 82.9 ±0.2 | 63.9 ±0.1 | 77.6 ±0.2 |
| **Libra HiResCAM** | 57.4 ±0.1 (+1.4%) | 83.0 ±0.2 (+0.2%) | 64.4 ±0.1 (+0.8%) | 77.5 ±0.2 (-0.1%) |
| XGradCAM+ | 64.3 ±0.1 | 87.3 ±0.2 | 71.3 ±0.1 | 82.1 ±0.2 |
| **Libra XGradCAM+** | 67.8 ±0.1 (+5.5%) | 90.0 ±0.2 (+3.0%) | 74.8 ±0.1 (+5.0%) | 85.0 ±0.2 (+3.5%) |
| FullGrad+ | 66.4 ±0.1 | 88.7 ±0.3 | 73.4 ±0.1 | 83.8 ±0.2 |
| **Libra FullGrad+** | 72.6 ±0.1 (+9.4%) | 91.1 ±0.2 (+2.6%) | 80.1 ±0.1 (+9.2%) | 86.3 ±0.2 (+3.0%) |

Table 49. Comparison of attribution methods and their LibraGrad-enhanced versions on the MLP-Mixer-L model.

| Method | SRG (GT) | | SRG (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 49.9 ±0.1 | 49.8 ±0.2 | 49.8 ±0.1 | 49.7 ±0.2 |
| AliLRP | 65.6 ±0.1 | 61.7 ±0.3 | 67.3 ±0.1 | 62.8 ±0.2 |
| DecompX | 66.2 ±0.1 | 62.7 ±0.3 | 68.0 ±0.1 | 63.8 ±0.2 |
| Integrated Gradients | 63.8 ±0.1 | 59.7 ±0.2 | 60.1 ±0.1 | 57.4 ±0.2 |
| Input × Grad | 60.4 ±0.1 | 57.1 ±0.2 | 61.8 ±0.1 | 57.8 ±0.2 |
| **Libra Input × Grad** | **76.9** ±0.1 (+27.2%) | **69.0** ±0.2 (+20.9%) | **79.3** ±0.1 (+28.4%) | **70.7** ±0.2 (+22.4%) |
| GradCAM+ | 62.9 ±0.1 | 59.1 ±0.2 | 64.1 ±0.1 | 59.8 ±0.2 |
| **Libra GradCAM+** | 66.5 ±0.1 (+5.6%) | 61.9 ±0.3 (+4.7%) | 67.8 ±0.1 (+5.7%) | 62.8 ±0.2 (+4.9%) |
| HiResCAM | 55.4 ±0.1 | 54.1 ±0.3 | 56.1 ±0.1 | 54.4 ±0.2 |
| **Libra HiResCAM** | 56.2 ±0.1 (+1.4%) | 54.6 ±0.3 (+0.9%) | 56.7 ±0.1 (+1.0%) | 54.8 ±0.2 (+0.7%) |
| XGradCAM+ | 63.5 ±0.1 | 59.5 ±0.2 | 64.8 ±0.1 | 60.2 ±0.2 |
| **Libra XGradCAM+** | 68.5 ±0.1 (+7.8%) | 63.2 ±0.3 (+6.2%) | 69.9 ±0.1 (+7.9%) | 64.2 ±0.2 (+6.6%) |
| FullGrad+ | 65.2 ±0.1 | 60.2 ±0.3 | 66.8 ±0.1 | 61.2 ±0.2 |
| **Libra FullGrad+** | 74.3 ±0.1 (+14.0%) | 66.2 ±0.3 (+9.8%) | 76.6 ±0.1 (+14.7%) | 67.6 ±0.2 (+10.4%) |

Table 50. Comparison of attribution methods and their LibraGrad-enhanced versions on the MLP-Mixer-L model.

### D.5.2. ViT-T

| Method | MIF Deletion (GT) | | MIF Deletion (Predicted) | | Segmentation |
|---|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC | AP |
| Random | 50.1 ±0.1 | 17.0 ±0.2 | 40.5 ±0.1 | 20.7 ±0.2 | 42.0 ±0.4 |
| RawAtt | 74.0 ±0.1 | 38.6 ±0.3 | 69.5 ±0.1 | 44.8 ±0.3 | 60.2 ±0.3 |
| Attention Rollout | 68.7 ±0.1 | 33.6 ±0.3 | 64.1 ±0.1 | 39.8 ±0.3 | 61.2 ±0.4 |
| AliLRP | 68.9 ±0.1 | 33.4 ±0.3 | 64.4 ±0.1 | 39.3 ±0.2 | 54.5 ±0.3 |
| AttnLRP | 73.4 ±0.1 | 37.8 ±0.3 | 69.7 ±0.1 | 44.3 ±0.3 | 59.7 ±0.3 |
| DecompX | 74.0 ±0.1 | 38.2 ±0.3 | 70.4 ±0.1 | 44.8 ±0.3 | 60.0 ±0.3 |
| Integrated Gradients | 69.7 ±0.1 | 32.8 ±0.3 | 57.1 ±0.1 | 33.3 ±0.2 | 52.4 ±0.3 |
| Input × Grad | 61.1 ±0.1 | 26.3 ±0.3 | 55.6 ±0.1 | 31.8 ±0.2 | 50.6 ±0.3 |
| **Libra Input × Grad** | 74.5 ±0.1 (+22.0%) | 37.5 ±0.3 (+42.6%) | 70.8 ±0.1 (+27.2%) | 44.0 ±0.3 (+38.3%) | 57.1 ±0.3 (+12.8%) |
| AttCAT | 72.6 ±0.1 | 35.6 ±0.3 | 69.3 ±0.1 | 42.0 ±0.3 | 54.7 ±0.3 |
| **Libra AttCAT** | <u>83.6</u> ±0.1 (+15.2%) | <u>45.0</u> ±0.3 (+26.6%) | <u>81.0</u> ±0.1 (+16.7%) | <u>52.1</u> ±0.2 (+24.1%) | 61.1 ±0.3 (+11.7%) |
| GenAtt | 80.4 ±0.1 | 42.7 ±0.3 | 77.1 ±0.1 | 49.4 ±0.3 | 71.1 ±0.3 |
| **Libra GenAtt** | 81.6 ±0.1 (+1.4%) | 43.6 ±0.3 (+2.3%) | 78.4 ±0.1 (+1.7%) | 50.5 ±0.2 (+2.2%) | **75.0** ±0.3 (+5.5%) |
| TokenTM | 78.8 ±0.1 | 41.8 ±0.3 | 75.0 ±0.1 | 48.3 ±0.3 | 70.8 ±0.3 |
| **Libra TokenTM** | 79.9 ±0.1 (+1.4%) | 42.6 ±0.3 (+2.0%) | 76.2 ±0.1 (+1.6%) | 49.2 ±0.3 (+1.9%) | <u>73.7</u> ±0.3 (+4.1%) |
| GradCAM+ | 70.5 ±0.1 | 34.1 ±0.3 | 66.2 ±0.1 | 40.1 ±0.2 | 48.4 ±0.4 |
| **Libra GradCAM+** | 76.8 ±0.1 (+8.9%) | 39.9 ±0.3 (+16.8%) | 72.9 ±0.1 (+10.1%) | 46.4 ±0.2 (+15.7%) | 56.3 ±0.4 (+16.4%) |
| HiResCAM | 48.0 ±0.1 | 15.8 ±0.3 | 39.0 ±0.1 | 19.5 ±0.3 | 50.6 ±0.4 |
| **Libra HiResCAM** | 74.1 ±0.1 (+54.3%) | 37.8 ±0.3 (+138.5%) | 69.9 ±0.1 (+79.1%) | 44.0 ±0.2 (+125.6%) | 63.8 ±0.3 (+26.1%) |
| XGradCAM+ | 71.7 ±0.1 | 35.1 ±0.3 | 67.5 ±0.1 | 41.2 ±0.2 | 48.8 ±0.4 |
| **Libra XGradCAM+** | 80.6 ±0.1 (+12.4%) | 42.7 ±0.3 (+21.7%) | 77.0 ±0.1 (+14.1%) | 49.5 ±0.2 (+20.1%) | 61.4 ±0.4 (+26.0%) |
| FullGrad+ | 69.8 ±0.1 | 33.1 ±0.3 | 65.9 ±0.1 | 39.2 ±0.3 | 53.2 ±0.3 |
| **Libra FullGrad+** | **84.2** ±0.1 (+20.8%) | **45.6** ±0.3 (+37.8%) | **81.7** ±0.1 (+24.0%) | **52.7** ±0.2 (+34.7%) | 65.0 ±0.3 (+22.2%) |

Table 51. Comparison of attribution methods and their LibraGrad-enhanced versions on the ViT-T model. We report faithfulness metrics using Most-Influential-First Deletion, MIF with ground-truth (GT) and predicted labels, including Accuracy and Area Over Perturbation Curve (AOPC) and Segmentation Average Precision (AP). The results demonstrate that composing existing methods with LibraGrad significantly enhances their performance across all metrics.

| Method | LIF Deletion (GT) Accuracy | LIF Deletion (GT) AOPC | LIF Deletion (Predicted) Accuracy | LIF Deletion (Predicted) AOPC |
|---|---|---|---|---|
| Random | 49.2 ±0.1 | 82.8 ±0.2 | 58.6 ±0.1 | 79.0 ±0.2 |
| RawAtt | 55.2 ±0.1 | 88.1 ±0.2 | 67.3 ±0.1 | 85.6 ±0.2 |
| Attention Rollout | 54.4 ±0.1 | 87.2 ±0.2 | 65.4 ±0.1 | 84.3 ±0.2 |
| AliLRP | 63.1 ±0.1 | 95.0 ±0.3 | 73.0 ±0.1 | 92.6 ±0.2 |
| AttnLRP | 63.2 ±0.1 | 95.6 ±0.2 | 74.3 ±0.1 | 93.2 ±0.2 |
| DecompX | 63.3 ±0.1 | 95.4 ±0.2 | 74.8 ±0.1 | 93.1 ±0.2 |
| Integrated Gradients | 64.0 ±0.1 | 96.3 ±0.2 | 66.9 ±0.1 | 88.8 ±0.2 |
| Input × Grad | 58.6 ±0.1 | 90.8 ±0.2 | 69.0 ±0.1 | 88.3 ±0.2 |
| **Libra Input × Grad** | 65.2 ±0.1 (+11.3%) | 96.4 ±0.2 (+6.2%) | 74.8 ±0.1 (+8.4%) | 93.9 ±0.2 (+6.3%) |
| AttCAT | 66.5 ±0.1 | 98.0 ±0.2 | 74.8 ±0.1 | 95.3 ±0.2 |
| **Libra AttCAT** | <u>69.5</u> ±0.1 (+4.5%) | <u>100.8</u> ±0.2 (+2.9%) | <u>77.9</u> ±0.1 (+4.2%) | <u>98.1</u> ±0.2 (+2.9%) |
| GenAtt | 63.6 ±0.1 | 95.4 ±0.2 | 76.2 ±0.1 | 93.5 ±0.2 |
| **Libra GenAtt** | 62.3 ±0.1 (-2.0%) | 94.3 ±0.2 (-1.1%) | 74.6 ±0.1 (-2.1%) | 92.2 ±0.2 (-1.4%) |
| TokenTM | 61.2 ±0.1 | 93.4 ±0.2 | 74.2 ±0.1 | 91.3 ±0.2 |
| **Libra TokenTM** | 60.8 ±0.1 (-0.6%) | 92.7 ±0.2 (-0.7%) | 73.7 ±0.1 (-0.6%) | 90.5 ±0.2 (-0.9%) |
| GradCAM+ | 57.9 ±0.1 | 88.6 ±0.2 | 65.1 ±0.1 | 84.3 ±0.2 |
| **Libra GradCAM+** | 61.9 ±0.1 (+7.0%) | 93.3 ±0.2 (+5.2%) | 70.2 ±0.1 (+7.8%) | 89.6 ±0.2 (+6.2%) |
| HiResCAM | 42.4 ±0.1 | 76.6 ±0.3 | 48.3 ±0.1 | 71.3 ±0.2 |
| **Libra HiResCAM** | 60.0 ±0.1 (+41.5%) | 90.1 ±0.2 (+17.7%) | 68.0 ±0.1 (+40.8%) | 86.2 ±0.2 (+20.8%) |
| XGradCAM+ | 59.5 ±0.1 | 90.5 ±0.3 | 66.7 ±0.1 | 86.4 ±0.2 |
| **Libra XGradCAM+** | 64.4 ±0.1 (+8.3%) | 95.5 ±0.2 (+5.6%) | 72.8 ±0.1 (+9.0%) | 91.9 ±0.2 (+6.3%) |
| FullGrad+ | 64.5 ±0.1 | 96.3 ±0.2 | 73.4 ±0.1 | 93.5 ±0.2 |
| **Libra FullGrad+** | **70.2** ±0.1 (+8.8%) | **101.8** ±0.2 (+5.7%) | **78.8** ±0.1 (+7.3%) | **99.2** ±0.2 (+6.1%) |

Table 52. Comparison of attribution methods and their LibraGrad-enhanced versions on the ViT-T model.

| Method | SRG (GT) Accuracy | SRG (GT) AOPC | SRG (Predicted) Accuracy | SRG (Predicted) AOPC |
|---|---|---|---|---|
| Random | 49.6 ±0.1 | 49.9 ±0.2 | 49.6 ±0.1 | 49.9 ±0.2 |
| RawAtt | 64.6 ±0.1 | 63.3 ±0.3 | 68.4 ±0.1 | 65.2 ±0.2 |
| Attention Rollout | 61.6 ±0.1 | 60.4 ±0.3 | 64.8 ±0.1 | 62.1 ±0.2 |
| AliLRP | 66.0 ±0.1 | 64.2 ±0.3 | 68.7 ±0.1 | 66.0 ±0.2 |
| AttnLRP | 68.3 ±0.1 | 66.7 ±0.3 | 72.0 ±0.1 | 68.8 ±0.2 |
| DecompX | 68.6 ±0.1 | 66.8 ±0.3 | 72.6 ±0.1 | 69.0 ±0.2 |
| Integrated Gradients | 66.8 ±0.1 | 64.6 ±0.3 | 62.0 ±0.1 | 61.1 ±0.2 |
| Input × Grad | 59.8 ±0.1 | 58.5 ±0.2 | 62.3 ±0.1 | 60.0 ±0.2 |
| **Libra Input × Grad** | 69.9 ±0.1 (+16.8%) | 67.0 ±0.3 (+14.4%) | 72.8 ±0.1 (+16.8%) | 68.9 ±0.2 (+14.8%) |
| AttCAT | 69.5 ±0.1 | 66.8 ±0.3 | 72.0 ±0.1 | 68.6 ±0.2 |
| **Libra AttCAT** | <u>76.5</u> ±0.1 (+10.1%) | <u>72.9</u> ±0.2 (+9.2%) | <u>79.4</u> ±0.1 (+10.2%) | <u>75.1</u> ±0.2 (+9.4%) |
| GenAtt | 72.0 ±0.1 | 69.1 ±0.2 | 76.6 ±0.1 | 71.5 ±0.2 |
| **Libra GenAtt** | 71.9 ±0.1 (-0.1%) | 69.0 ±0.2 (-0.1%) | 76.5 ±0.1 (-0.2%) | 71.4 ±0.2 (-0.1%) |
| TokenTM | 70.0 ±0.1 | 67.6 ±0.2 | 74.6 ±0.1 | 69.8 ±0.2 |
| **Libra TokenTM** | 70.3 ±0.1 (+0.5%) | 67.7 ±0.2 (+0.1%) | 75.0 ±0.1 (+0.5%) | 69.9 ±0.2 (+0.1%) |
| GradCAM+ | 64.2 ±0.1 | 61.4 ±0.3 | 65.7 ±0.1 | 62.2 ±0.2 |
| **Libra GradCAM+** | 69.3 ±0.1 (+8.0%) | 66.6 ±0.3 (+8.4%) | 71.5 ±0.1 (+9.0%) | 68.0 ±0.2 (+9.3%) |
| HiResCAM | 45.2 ±0.1 | 46.2 ±0.3 | 43.7 ±0.1 | 45.4 ±0.2 |
| **Libra HiResCAM** | 67.0 ±0.1 (+48.3%) | 64.0 ±0.2 (+38.4%) | 68.9 ±0.1 (+57.9%) | 65.1 ±0.2 (+43.3%) |
| XGradCAM+ | 65.6 ±0.1 | 62.8 ±0.3 | 67.1 ±0.1 | 63.8 ±0.2 |
| **Libra XGradCAM+** | 72.5 ±0.1 (+10.5%) | 69.1 ±0.2 (+10.1%) | 74.9 ±0.1 (+11.6%) | 70.7 ±0.2 (+10.8%) |
| FullGrad+ | 67.1 ±0.1 | 64.7 ±0.3 | 69.7 ±0.1 | 66.3 ±0.2 |
| **Libra FullGrad+** | **77.2** ±0.1 (+15.0%) | **73.7** ±0.2 (+13.9%) | **80.3** ±0.1 (+15.2%) | **76.0** ±0.2 (+14.5%) |

Table 53. Comparison of attribution methods and their LibraGrad-enhanced versions on the ViT-T model.

### D.5.3. ViT-S

| Method | MIF Deletion (GT) Accuracy | MIF Deletion (GT) AOPC | MIF Deletion (Predicted) Accuracy | MIF Deletion (Predicted) AOPC | Segmentation AP |
|---|---|---|---|---|---|
| Random | 41.8 ±0.1 | 15.8 ±0.3 | 33.8 ±0.1 | 18.6 ±0.2 | 41.9 ±0.4 |
| RawAtt | 63.1 ±0.1 | 36.5 ±0.3 | 58.7 ±0.1 | 41.2 ±0.3 | 57.8 ±0.3 |
| Attention Rollout | 51.2 ±0.1 | 25.1 ±0.4 | 45.1 ±0.1 | 28.8 ±0.2 | 47.1 ±0.3 |
| AliLRP | 48.9 ±0.1 | 22.8 ±0.3 | 42.3 ±0.1 | 26.2 ±0.3 | 42.5 ±0.4 |
| AttnLRP | 57.7 ±0.1 | 30.9 ±0.3 | 52.4 ±0.1 | 35.2 ±0.2 | 46.2 ±0.3 |
| DecompX | 56.0 ±0.1 | 29.5 ±0.3 | 50.4 ±0.1 | 33.6 ±0.2 | 47.7 ±0.3 |
| Integrated Gradients | 56.9 ±0.1 | 29.1 ±0.3 | 46.0 ±0.1 | 29.3 ±0.3 | 51.7 ±0.3 |
| Input × Grad | 47.9 ±0.1 | 21.6 ±0.3 | 41.8 ±0.1 | 25.0 ±0.3 | 48.5 ±0.3 |
| **Libra Input × Grad** | 54.9 ±0.1 (+14.7%) | 28.2 ±0.3 (+30.4%) | 49.3 ±0.1 (+18.0%) | 32.2 ±0.2 (+28.5%) | 46.0 ±0.3 (-5.1%) |
| AttCAT | 62.1 ±0.1 | 33.4 ±0.3 | 58.9 ±0.1 | 38.2 ±0.3 | 49.8 ±0.3 |
| **Libra AttCAT** | **73.6** ±0.1 (+18.5%) | **43.9** ±0.3 (+31.4%) | **70.3** ±0.1 (+19.3%) | **48.9** ±0.3 (+28.0%) | 56.0 ±0.3 (+12.4%) |
| GenAtt | 69.7 ±0.1 | 41.3 ±0.3 | 66.3 ±0.1 | 46.3 ±0.3 | 65.9 ±0.2 |
| **Libra GenAtt** | 71.7 ±0.1 (+2.9%) | 43.2 ±0.3 (+4.6%) | 68.2 ±0.1 (+2.9%) | 48.2 ±0.3 (+4.2%) | <u>71.0</u> ±0.3 (+7.7%) |
| TokenTM | 68.9 ±0.1 | 40.8 ±0.3 | 65.2 ±0.1 | 45.9 ±0.3 | 68.2 ±0.2 |
| **Libra TokenTM** | 70.3 ±0.1 (+2.1%) | 42.2 ±0.3 (+3.4%) | 66.5 ±0.1 (+2.0%) | 47.3 ±0.3 (+3.0%) | **71.4** ±0.2 (+4.7%) |
| GradCAM+ | 59.9 ±0.1 | 31.5 ±0.3 | 55.5 ±0.1 | 35.8 ±0.3 | 46.4 ±0.4 |
| **Libra GradCAM+** | 70.2 ±0.1 (+17.0%) | 41.2 ±0.3 (+30.7%) | 66.5 ±0.1 (+19.7%) | 46.1 ±0.3 (+28.7%) | 60.7 ±0.4 (+30.8%) |
| HiResCAM | 38.4 ±0.1 | 13.1 ±0.2 | 29.5 ±0.1 | 15.3 ±0.2 | 48.4 ±0.4 |
| **Libra HiResCAM** | 67.4 ±0.1 (+75.5%) | 39.6 ±0.3 (+202.6%) | 63.4 ±0.1 (+114.7%) | 44.4 ±0.2 (+190.6%) | 69.4 ±0.3 (+43.2%) |
| XGradCAM+ | 60.3 ±0.1 | 31.9 ±0.4 | 55.9 ±0.1 | 36.2 ±0.3 | 45.4 ±0.4 |
| **Libra XGradCAM+** | 72.1 ±0.1 (+19.5%) | 42.8 ±0.3 (+34.1%) | 68.5 ±0.1 (+22.4%) | 47.8 ±0.3 (+32.0%) | 62.3 ±0.4 (+37.2%) |
| FullGrad+ | 59.6 ±0.1 | 31.5 ±0.3 | 55.8 ±0.1 | 36.1 ±0.3 | 50.0 ±0.3 |
| **Libra FullGrad+** | <u>73.5</u> ±0.1 (+23.3%) | <u>43.8</u> ±0.3 (+39.0%) | <u>70.1</u> ±0.1 (+25.8%) | **48.9** ±0.3 (+35.3%) | 59.6 ±0.3 (+19.2%) |

Table 54. Comparison of attribution methods and their LibraGrad-enhanced versions on the ViT-S model. We report faithfulness metrics using Most-Influential-First Deletion, MIF with ground-truth (GT) and predicted labels, including Accuracy and Area Over Perturbation Curve (AOPC) and Segmentation Average Precision (AP). The results demonstrate that composing existing methods with LibraGrad significantly enhances their performance across all metrics.

| Method | LIF Deletion (GT) Accuracy | LIF Deletion (GT) AOPC | LIF Deletion (Predicted) Accuracy | LIF Deletion (Predicted) AOPC |
|---|---|---|---|---|
| Random | 57.7 ±0.1 | 84.1 ±0.2 | 66.5 ±0.1 | 81.8 ±0.2 |
| RawAtt | 63.9 ±0.1 | 89.4 ±0.2 | 72.8 ±0.1 | 87.2 ±0.2 |
| Attention Rollout | 58.6 ±0.1 | 84.8 ±0.2 | 67.3 ±0.1 | 82.4 ±0.3 |
| AliLRP | 62.5 ±0.1 | 88.2 ±0.2 | 70.6 ±0.1 | 85.8 ±0.2 |
| AttnLRP | 68.4 ±0.1 | 94.6 ±0.2 | 77.3 ±0.1 | 92.8 ±0.2 |
| DecompX | 66.5 ±0.1 | 92.2 ±0.2 | 75.3 ±0.1 | 90.3 ±0.2 |
| Integrated Gradients | 69.6 ±0.1 | 95.4 ±0.2 | 73.9 ±0.1 | 90.1 ±0.2 |
| Input × Grad | 64.2 ±0.1 | 90.0 ±0.3 | 72.1 ±0.1 | 87.8 ±0.3 |
| **Libra Input × Grad** | 66.4 ±0.1 (+3.4%) | 92.0 ±0.2 (+2.1%) | 74.3 ±0.1 (+3.0%) | 89.6 ±0.2 (+2.1%) |
| AttCAT | 71.9 ±0.1 | 97.7 ±0.2 | 78.5 ±0.1 | 95.6 ±0.2 |
| **Libra AttCAT** | <u>74.2</u> ±0.1 (+3.3%) | <u>100.0</u> ±0.2 (+2.4%) | **81.0** ±0.1 (+3.2%) | <u>97.8</u> ±0.2 (+2.3%) |
| GenAtt | 69.6 ±0.1 | 94.7 ±0.2 | 79.1 ±0.1 | 92.9 ±0.2 |
| **Libra GenAtt** | 69.5 ±0.1 (-0.2%) | 94.5 ±0.2 (-0.2%) | 79.0 ±0.1 (-0.1%) | 92.7 ±0.2 (-0.2%) |
| TokenTM | 67.4 ±0.1 | 92.7 ±0.2 | 77.2 ±0.1 | 90.8 ±0.2 |
| **Libra TokenTM** | 67.4 ±0.1 (-0.1%) | 92.8 ±0.2 (+0.1%) | 77.1 ±0.1 (-0.1%) | 90.8 ±0.2 (+0.0%) |
| GradCAM+ | 65.0 ±0.1 | 90.5 ±0.2 | 71.9 ±0.1 | 88.1 ±0.2 |
| **Libra GradCAM+** | 70.7 ±0.1 (+8.9%) | 96.0 ±0.2 (+6.1%) | 78.0 ±0.1 (+8.4%) | 93.7 ±0.2 (+6.4%) |
| HiResCAM | 55.8 ±0.1 | 81.8 ±0.3 | 62.8 ±0.1 | 78.7 ±0.3 |
| **Libra HiResCAM** | 68.5 ±0.1 (+22.6%) | 93.3 ±0.2 (+14.1%) | 76.1 ±0.1 (+21.2%) | 91.1 ±0.2 (+15.7%) |
| XGradCAM+ | 66.3 ±0.1 | 91.8 ±0.2 | 73.5 ±0.1 | 89.5 ±0.2 |
| **Libra XGradCAM+** | 71.4 ±0.1 (+7.7%) | 96.5 ±0.2 (+5.1%) | 78.5 ±0.1 (+6.8%) | 94.1 ±0.2 (+5.2%) |
| FullGrad+ | 70.3 ±0.1 | 96.3 ±0.2 | 77.6 ±0.1 | 94.3 ±0.2 |
| **Libra FullGrad+** | **74.4** ±0.1 (+5.9%) | **100.1** ±0.2 (+4.0%) | **81.0** ±0.1 (+4.4%) | **97.9** ±0.2 (+3.8%) |

Table 55. Comparison of attribution methods and their LibraGrad-enhanced versions on the ViT-S model.

| Method | SRG (GT) Accuracy | SRG (GT) AOPC | SRG (Predicted) Accuracy | SRG (Predicted) AOPC |
|---|---|---|---|---|
| Random | 49.7 ±0.1 | 49.9 ±0.3 | 50.2 ±0.1 | 50.2 ±0.2 |
| RawAtt | 63.5 ±0.1 | 63.0 ±0.3 | 65.8 ±0.1 | 64.2 ±0.3 |
| Attention Rollout | 54.9 ±0.1 | 54.9 ±0.3 | 56.2 ±0.1 | 55.6 ±0.2 |
| AliLRP | 55.7 ±0.1 | 55.5 ±0.3 | 56.5 ±0.1 | 56.0 ±0.3 |
| AttnLRP | 63.0 ±0.1 | 62.8 ±0.3 | 64.9 ±0.1 | 64.0 ±0.2 |
| DecompX | 61.3 ±0.1 | 60.9 ±0.3 | 62.9 ±0.1 | 61.9 ±0.2 |
| Integrated Gradients | 63.3 ±0.1 | 62.3 ±0.3 | 59.9 ±0.1 | 59.7 ±0.3 |
| Input × Grad | 56.1 ±0.1 | 55.8 ±0.3 | 57.0 ±0.1 | 56.4 ±0.3 |
| **Libra Input × Grad** | 60.7 ±0.1 (+8.2%) | 60.1 ±0.3 (+7.6%) | 61.8 ±0.1 (+8.5%) | 60.9 ±0.2 (+8.0%) |
| AttCAT | 67.0 ±0.1 | 65.5 ±0.3 | 68.7 ±0.1 | 66.9 ±0.2 |
| **Libra AttCAT** | <u>73.9</u> ±0.1 (+10.3%) | <u>71.9</u> ±0.3 (+9.8%) | **75.7** ±0.1 (+10.1%) | <u>73.3</u> ±0.3 (+9.6%) |
| GenAtt | 69.6 ±0.1 | 68.0 ±0.3 | 72.7 ±0.1 | 69.6 ±0.2 |
| **Libra GenAtt** | 70.6 ±0.1 (+1.4%) | 68.8 ±0.3 (+1.3%) | 73.6 ±0.1 (+1.3%) | 70.5 ±0.3 (+1.2%) |
| TokenTM | 68.2 ±0.1 | 66.8 ±0.3 | 71.2 ±0.1 | 68.3 ±0.2 |
| **Libra TokenTM** | 68.8 ±0.1 (+1.0%) | 67.5 ±0.3 (+1.1%) | 71.8 ±0.1 (+0.9%) | 69.1 ±0.3 (+1.0%) |
| GradCAM+ | 62.5 ±0.1 | 61.0 ±0.3 | 63.7 ±0.1 | 62.0 ±0.3 |
| **Libra GradCAM+** | 70.4 ±0.1 (+12.8%) | 68.6 ±0.3 (+12.4%) | 72.2 ±0.1 (+13.3%) | 69.9 ±0.3 (+12.8%) |
| HiResCAM | 47.1 ±0.1 | 47.4 ±0.3 | 46.2 ±0.1 | 47.0 ±0.3 |
| **Libra HiResCAM** | 68.0 ±0.1 (+44.2%) | 66.4 ±0.2 (+40.0%) | 69.8 ±0.1 (+51.1%) | 67.7 ±0.2 (+44.1%) |
| XGradCAM+ | 63.3 ±0.1 | 61.9 ±0.3 | 64.7 ±0.1 | 62.8 ±0.3 |
| **Libra XGradCAM+** | 71.7 ±0.1 (+13.3%) | 69.7 ±0.3 (+12.6%) | 73.5 ±0.1 (+13.5%) | 71.0 ±0.2 (+12.9%) |
| FullGrad+ | 65.0 ±0.1 | 63.9 ±0.3 | 66.7 ±0.1 | 65.2 ±0.3 |
| **Libra FullGrad+** | **74.0** ±0.1 (+13.9%) | **72.0** ±0.3 (+12.6%) | <u>75.6</u> ±0.1 (+13.3%) | **73.4** ±0.3 (+12.5%) |

Table 56. Comparison of attribution methods and their LibraGrad-enhanced versions on the ViT-S model.

### D.5.4. ViT-B

| Method | MIF Deletion (GT) Accuracy | MIF Deletion (GT) AOPC | MIF Deletion (Predicted) Accuracy | MIF Deletion (Predicted) AOPC | Segmentation AP |
|---|---|---|---|---|---|
| Random | 34.5 ±0.1 | 12.3 ±0.2 | 26.5 ±0.1 | 14.2 ±0.2 | 41.9 ±0.4 |
| RawAtt | 50.1 ±0.1 | 25.0 ±0.3 | 44.6 ±0.1 | 27.9 ±0.3 | 46.9 ±0.3 |
| Attention Rollout | 41.9 ±0.1 | 18.8 ±0.3 | 35.4 ±0.1 | 21.2 ±0.2 | 45.3 ±0.3 |
| AliLRP | 39.8 ±0.1 | 16.7 ±0.2 | 33.3 ±0.1 | 19.1 ±0.2 | 43.8 ±0.4 |
| AttnLRP | 44.5 ±0.1 | 20.8 ±0.3 | 38.5 ±0.1 | 23.4 ±0.2 | 42.0 ±0.4 |
| DecompX | 44.0 ±0.1 | 20.3 ±0.3 | 37.8 ±0.1 | 22.8 ±0.2 | 44.3 ±0.3 |
| Integrated Gradients | 46.9 ±0.1 | 22.5 ±0.2 | 35.4 ±0.1 | 21.4 ±0.2 | 47.5 ±0.3 |
| Input × Grad | 40.4 ±0.1 | 17.7 ±0.2 | 34.4 ±0.1 | 20.2 ±0.2 | 44.8 ±0.3 |
| **Libra Input × Grad** | 44.8 ±0.1 (+10.8%) | 20.8 ±0.3 (+17.4%) | 38.6 ±0.1 (+12.0%) | 23.4 ±0.2 (+15.8%) | 44.4 ±0.3 (-0.9%) |
| AttCAT | 50.4 ±0.1 | 25.3 ±0.2 | 46.9 ±0.1 | 28.8 ±0.2 | 44.5 ±0.3 |
| **Libra AttCAT** | <u>66.4</u> ±0.1 (+31.7%) | <u>37.5</u> ±0.3 (+47.9%) | <u>63.5</u> ±0.1 (+35.4%) | **41.5** ±0.3 (+44.2%) | 61.5 ±0.3 (+38.3%) |
| GenAtt | 61.9 ±0.1 | 34.2 ±0.3 | 58.2 ±0.1 | 37.9 ±0.2 | 71.0 ±0.2 |
| **Libra GenAtt** | 65.1 ±0.1 (+5.1%) | 36.6 ±0.3 (+6.8%) | 61.6 ±0.1 (+5.8%) | 40.4 ±0.3 (+6.6%) | **77.5** ±0.2 (+9.2%) |
| TokenTM | 60.6 ±0.1 | 33.8 ±0.3 | 56.8 ±0.1 | 37.4 ±0.3 | 70.2 ±0.2 |
| **Libra TokenTM** | 62.8 ±0.1 (+3.6%) | 35.1 ±0.3 (+4.0%) | 59.1 ±0.1 (+4.1%) | 38.9 ±0.3 (+3.8%) | 73.9 ±0.2 (+5.2%) |
| GradCAM+ | 50.5 ±0.1 | 24.8 ±0.2 | 45.6 ±0.1 | 27.6 ±0.2 | 50.2 ±0.4 |
| **Libra GradCAM+** | 65.3 ±0.1 (+29.3%) | 35.9 ±0.2 (+44.8%) | 61.4 ±0.1 (+34.8%) | 39.6 ±0.2 (+43.5%) | 72.1 ±0.3 (+43.6%) |
| HiResCAM | 50.4 ±0.1 | 25.4 ±0.3 | 45.4 ±0.1 | 28.5 ±0.2 | 59.0 ±0.3 |
| **Libra HiResCAM** | 60.8 ±0.1 (+20.6%) | 33.4 ±0.3 (+31.7%) | 56.7 ±0.1 (+24.8%) | 37.0 ±0.2 (+29.6%) | 72.6 ±0.3 (+23.1%) |
| XGradCAM+ | 44.0 ±0.1 | 19.0 ±0.2 | 38.6 ±0.1 | 21.5 ±0.2 | 41.0 ±0.4 |
| **Libra XGradCAM+** | **67.4** ±0.1 (+53.0%) | **37.7** ±0.2 (+98.6%) | **63.9** ±0.1 (+65.6%) | **41.5** ±0.2 (+92.8%) | <u>75.0</u> ±0.3 (+82.8%) |
| FullGrad+ | 48.2 ±0.1 | 23.1 ±0.3 | 44.2 ±0.1 | 26.3 ±0.2 | 45.2 ±0.3 |
| **Libra FullGrad+** | 66.1 ±0.1 (+37.1%) | 37.2 ±0.3 (+60.9%) | 63.1 ±0.1 (+42.9%) | 41.2 ±0.3 (+56.7%) | 65.5 ±0.3 (+44.8%) |

Table 57. Comparison of attribution methods and their LibraGrad-enhanced versions on the ViT-B model. We report faithfulness metrics using Most-Influential-First Deletion, MIF with ground-truth (GT) and predicted labels, including Accuracy and Area Over Perturbation Curve (AOPC) and Segmentation Average Precision (AP). The results demonstrate that composing existing methods with LibraGrad significantly enhances their performance across all metrics.

| Method | LIF Deletion (GT) | | LIF Deletion (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 65.2 ±0.1 | 87.5 ±0.2 | 73.3 ±0.1 | 85.8 ±0.2 |
| RawAtt | 67.5 ±0.1 | 89.2 ±0.1 | 76.2 ±0.1 | 87.6 ±0.1 |
| Attention Rollout | 65.9 ±0.1 | 87.9 ±0.2 | 73.8 ±0.1 | 86.0 ±0.2 |
| AliLRP | 69.9 ±0.1 | 90.9 ±0.2 | 77.8 ±0.1 | 89.3 ±0.2 |
| AttnLRP | 71.0 ±0.1 | 92.4 ±0.2 | 78.7 ±0.1 | 90.8 ±0.1 |
| DecompX | 71.1 ±0.1 | 92.2 ±0.2 | 79.1 ±0.1 | 90.6 ±0.1 |
| Integrated Gradients | 74.4 ±0.1 | 95.7 ±0.2 | 78.0 ±0.1 | 91.3 ±0.1 |
| Input $\times$ Grad | 69.9 ±0.1 | 91.8 ±0.2 | 77.3 ±0.1 | 90.2 ±0.1 |
| **Libra Input $\times$ Grad** | 72.5 ±0.1 (+3.7%) | 93.1 ±0.2 (+1.4%) | 80.2 ±0.1 (+3.8%) | 91.3 ±0.2 (+1.2%) |
| AttCAT | 76.8 ±0.1 | 98.4 ±0.2 | 82.5 ±0.1 | 96.6 ±0.2 |
| **Libra AttCAT** | <u>80.2</u> ±0.1 (+4.5%) | <u>100.8</u> ±0.2 (+2.4%) | <u>86.7</u> ±0.1 (+5.1%) | <u>99.2</u> ±0.1 (+2.7%) |
| GenAtt | 74.8 ±0.1 | 95.7 ±0.2 | 84.0 ±0.1 | 94.6 ±0.1 |
| **Libra GenAtt** | 75.1 ±0.1 (+0.4%) | 96.0 ±0.1 (+0.3%) | 84.4 ±0.1 (+0.4%) | 94.8 ±0.1 (+0.2%) |
| TokenTM | 73.5 ±0.1 | 94.4 ±0.1 | 83.1 ±0.1 | 93.3 ±0.1 |
| **Libra TokenTM** | 73.6 ±0.1 (+0.1%) | 94.6 ±0.1 (+0.2%) | 83.2 ±0.1 (+0.1%) | 93.5 ±0.2 (+0.2%) |
| GradCAM+ | 72.0 ±0.1 | 93.5 ±0.2 | 78.5 ±0.1 | 91.5 ±0.2 |
| **Libra GradCAM+** | 78.0 ±0.1 (+8.3%) | 98.1 ±0.2 (+4.9%) | 84.9 ±0.1 (+8.3%) | 96.2 ±0.1 (+5.2%) |
| HiResCAM | 71.9 ±0.1 | 93.3 ±0.2 | 79.5 ±0.1 | 91.7 ±0.2 |
| **Libra HiResCAM** | 75.6 ±0.1 (+5.1%) | 96.1 ±0.2 (+3.0%) | 82.7 ±0.1 (+4.0%) | 94.3 ±0.1 (+2.8%) |
| XGradCAM+ | 67.0 ±0.1 | 88.9 ±0.3 | 73.3 ±0.1 | 86.6 ±0.2 |
| **Libra XGradCAM+** | 78.1 ±0.1 (+16.6%) | 98.4 ±0.2 (+10.8%) | 85.4 ±0.1 (+16.6%) | 96.6 ±0.1 (+11.6%) |
| FullGrad+ | 75.2 ±0.1 | 96.6 ±0.2 | 81.6 ±0.1 | 95.0 ±0.2 |
| **Libra FullGrad+** | **80.6** ±0.1 (+7.1%) | **101.2** ±0.2 (+4.8%) | **87.0** ±0.0 (+6.6%) | **99.6** ±0.1 (+4.8%) |

Table 58. Comparison of attribution methods and their LibraGrad-enhanced versions on the ViT-B model.

| Method | SRG (GT) | | SRG (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 49.9 ±0.1 | 49.9 ±0.2 | 49.9 ±0.1 | 50.0 ±0.2 |
| RawAtt | 58.8 ±0.1 | 57.1 ±0.2 | 60.4 ±0.1 | 57.8 ±0.2 |
| Attention Rollout | 53.9 ±0.1 | 53.3 ±0.2 | 54.6 ±0.1 | 53.6 ±0.2 |
| AliLRP | 54.8 ±0.1 | 53.8 ±0.2 | 55.5 ±0.1 | 54.2 ±0.2 |
| AttnLRP | 57.8 ±0.1 | 56.6 ±0.2 | 58.6 ±0.1 | 57.1 ±0.2 |
| DecompX | 57.6 ±0.1 | 56.3 ±0.2 | 58.5 ±0.1 | 56.7 ±0.2 |
| Integrated Gradients | 60.6 ±0.1 | 59.1 ±0.2 | 56.7 ±0.1 | 56.3 ±0.2 |
| Input $\times$ Grad | 55.1 ±0.1 | 54.8 ±0.2 | 55.9 ±0.1 | 55.2 ±0.2 |
| **Libra Input $\times$ Grad** | 58.6 ±0.1 (+6.3%) | 56.9 ±0.2 (+4.0%) | 59.4 ±0.1 (+6.3%) | 57.3 ±0.2 (+3.9%) |
| AttCAT | 63.6 ±0.1 | 61.9 ±0.2 | 64.7 ±0.1 | 62.7 ±0.2 |
| **Libra AttCAT** | **73.3** ±0.1 (+15.3%) | <u>69.1</u> ±0.2 (+11.7%) | **75.1** ±0.1 (+16.1%) | **70.4** ±0.2 (+12.2%) |
| GenAtt | 68.4 ±0.1 | 65.0 ±0.2 | 71.1 ±0.1 | 66.3 ±0.2 |
| **Libra GenAtt** | 70.1 ±0.1 (+2.5%) | 66.3 ±0.2 (+2.0%) | 73.0 ±0.1 (+2.6%) | 67.6 ±0.2 (+2.1%) |
| TokenTM | 67.1 ±0.1 | 64.1 ±0.2 | 70.0 ±0.1 | 65.3 ±0.2 |
| **Libra TokenTM** | 68.2 ±0.1 (+1.7%) | 64.9 ±0.2 (+1.2%) | 71.1 ±0.1 (+1.7%) | 66.2 ±0.2 (+1.3%) |
| GradCAM+ | 61.3 ±0.1 | 59.2 ±0.2 | 62.0 ±0.1 | 59.5 ±0.2 |
| **Libra GradCAM+** | 71.7 ±0.1 (+17.0%) | 67.0 ±0.2 (+13.3%) | 73.2 ±0.1 (+18.0%) | 67.9 ±0.2 (+14.1%) |
| HiResCAM | 61.2 ±0.1 | 59.3 ±0.2 | 62.5 ±0.1 | 60.1 ±0.2 |
| **Libra HiResCAM** | 68.2 ±0.1 (+11.5%) | 64.7 ±0.2 (+9.1%) | 69.7 ±0.1 (+11.6%) | 65.7 ±0.2 (+9.2%) |
| XGradCAM+ | 55.5 ±0.1 | 53.9 ±0.2 | 55.9 ±0.1 | 54.1 ±0.2 |
| **Libra XGradCAM+** | 72.7 ±0.1 (+31.0%) | 68.1 ±0.2 (+26.2%) | 74.6 ±0.1 (+33.5%) | 69.1 ±0.2 (+27.7%) |
| FullGrad+ | 61.7 ±0.1 | 59.8 ±0.2 | 62.9 ±0.1 | 60.7 ±0.2 |
| **Libra FullGrad+** | **73.3** ±0.1 (+18.8%) | **69.2** ±0.2 (+15.6%) | <u>75.0</u> ±0.1 (+19.4%) | **70.4** ±0.2 (+16.0%) |

Table 59. Comparison of attribution methods and their LibraGrad-enhanced versions on the ViT-B model.

### D.5.5. ImageNet-Hard ViT-B

| Method | MIF Deletion (GT) | | MIF Deletion (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 81.7 ±0.1 | 6.6 ±0.1 | 52.4 ±0.1 | 16.4 ±0.2 |
| RawAtt | 85.9 ±0.1 | 9.5 ±0.1 | 65.9 ±0.1 | 23.8 ±0.2 |
| Attention Rollout | 84.7 ±0.1 | 8.7 ±0.1 | 62.2 ±0.1 | 21.7 ±0.2 |
| AliLRP | 85.7 ±0.1 | 9.1 ±0.1 | 64.1 ±0.1 | 22.6 ±0.2 |
| AttnLRP | 88.2 ±0.1 | 10.8 ±0.2 | 70.8 ±0.1 | 26.3 ±0.2 |
| DecompX | 87.1 ±0.1 | 10.2 ±0.1 | 67.7 ±0.1 | 25.0 ±0.2 |
| Integrated Gradients | 89.5 ±0.1 | 11.7 ±0.1 | 66.6 ±0.1 | 24.6 ±0.3 |
| Input × Grad | 87.0 ±0.1 | 9.8 ±0.1 | 67.6 ±0.1 | 24.3 ±0.2 |
| **Libra Input × Grad** | 87.5 ±0.1 (+0.6%) | 10.3 ±0.1 (+5.3%) | 68.8 ±0.1 (+1.8%) | 25.3 ±0.2 (+3.7%) |
| AttCAT | 91.8 ±0.1 | 13.1 ±0.1 | 82.3 ±0.1 | 31.9 ±0.2 |
| **Libra AttCAT** | <u>94.4</u> ±0.1 (+2.9%) | <u>14.9</u> ±0.1 (+14.0%) | <u>87.3</u> ±0.1 (+6.1%) | <u>35.5</u> ±0.2 (+11.2%) |
| GenAtt | 92.0 ±0.1 | 13.5 ±0.1 | 81.3 ±0.1 | 32.2 ±0.2 |
| **Libra GenAtt** | 92.6 ±0.1 (+0.6%) | 13.8 ±0.1 (+2.8%) | 82.8 ±0.1 (+1.8%) | 33.1 ±0.2 (+2.6%) |
| TokenTM | 90.9 ±0.1 | 12.9 ±0.1 | 79.3 ±0.1 | 31.3 ±0.2 |
| **Libra TokenTM** | 91.4 ±0.1 (+0.5%) | 13.1 ±0.1 (+2.0%) | 80.0 ±0.1 (+0.8%) | 31.7 ±0.2 (+1.4%) |
| GradCAM+ | 89.2 ±0.1 | 11.4 ±0.1 | 75.8 ±0.1 | 28.4 ±0.2 |
| **Libra GradCAM+** | 92.7 ±0.1 (+3.9%) | 13.8 ±0.1 (+21.2%) | 83.4 ±0.1 (+10.0%) | 33.2 ±0.2 (+17.0%) |
| HiResCAM | 89.3 ±0.1 | 11.5 ±0.1 | 74.2 ±0.1 | 28.2 ±0.2 |
| **Libra HiResCAM** | 91.4 ±0.1 (+2.4%) | 12.9 ±0.1 (+12.7%) | 79.7 ±0.1 (+7.4%) | 31.4 ±0.2 (+11.3%) |
| XGradCAM+ | 87.8 ±0.1 | 10.6 ±0.1 | 72.1 ±0.1 | 26.4 ±0.2 |
| **Libra XGradCAM+** | 93.2 ±0.1 (+6.2%) | 14.1 ±0.1 (+33.7%) | 84.7 ±0.1 (+17.3%) | 33.9 ±0.2 (+28.3%) |
| FullGrad+ | 90.5 ±0.1 | 12.3 ±0.1 | 80.1 ±0.1 | 30.5 ±0.2 |
| **Libra FullGrad+** | **94.7** ±0.1 (+4.6%) | **15.0** ±0.1 (+22.4%) | **87.6** ±0.1 (+9.4%) | **35.6** ±0.2 (+16.6%) |

Table 60. Comparison of attribution methods and their LibraGrad-enhanced versions on the ImageNet-Hard ViT-B model.

| Method | LIF Deletion (GT) | | LIF Deletion (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 18.6 ±0.1 | 93.5 ±0.1 | 47.0 ±0.1 | 83.1 ±0.2 |
| RawAtt | 19.5 ±0.1 | 94.2 ±0.1 | 52.0 ±0.1 | 85.5 ±0.2 |
| Attention Rollout | 19.8 ±0.1 | 94.0 ±0.1 | 49.8 ±0.1 | 84.1 ±0.2 |
| AliLRP | 21.5 ±0.1 | 95.2 ±0.1 | 53.5 ±0.1 | 87.3 ±0.2 |
| AttnLRP | 26.4 ±0.1 | 98.2 ±0.1 | 61.6 ±0.1 | 93.4 ±0.2 |
| DecompX | 22.6 ±0.1 | 96.2 ±0.1 | 56.8 ±0.1 | 89.4 ±0.2 |
| Integrated Gradients | 28.6 ±0.1 | 99.7 ±0.1 | 54.2 ±0.1 | 89.7 ±0.2 |
| Input × Grad | 23.4 ±0.1 | 96.6 ±0.1 | 55.9 ±0.1 | 89.9 ±0.2 |
| **Libra Input × Grad** | 23.8 ±0.1 (+1.6%) | 96.5 ±0.1 (-0.1%) | 57.9 ±0.1 (+3.5%) | 90.1 ±0.2 (+0.2%) |
| AttCAT | 35.1 ±0.1 | 104.1 ±0.1 | 69.2 ±0.1 | 102.1 ±0.2 |
| **Libra AttCAT** | <u>35.8</u> ±0.1 (+1.8%) | <u>104.4</u> ±0.2 (+0.4%) | <u>75.9</u> ±0.1 (+9.5%) | <u>105.6</u> ±0.2 (+3.5%) |
| GenAtt | 25.4 ±0.1 | 97.8 ±0.1 | 65.7 ±0.1 | 93.4 ±0.2 |
| **Libra GenAtt** | 25.4 ±0.1 (+0.1%) | 97.9 ±0.1 (+0.1%) | 66.5 ±0.1 (+1.1%) | 93.4 ±0.2 (+0.1%) |
| TokenTM | 23.9 ±0.1 | 96.9 ±0.1 | 62.9 ±0.1 | 91.1 ±0.2 |
| **Libra TokenTM** | 24.3 ±0.1 (+1.8%) | 97.0 ±0.1 (+0.1%) | 63.0 ±0.1 (+0.3%) | 91.3 ±0.2 (+0.2%) |
| GradCAM+ | 27.9 ±0.1 | 98.8 ±0.1 | 61.2 ±0.1 | 93.9 ±0.2 |
| **Libra GradCAM+** | 30.2 ±0.1 (+8.3%) | 100.4 ±0.2 (+1.6%) | 68.6 ±0.1 (+12.2%) | 98.3 ±0.2 (+4.6%) |
| HiResCAM | 24.7 ±0.1 | 97.1 ±0.1 | 57.7 ±0.1 | 90.6 ±0.2 |
| **Libra HiResCAM** | 26.6 ±0.1 (+7.7%) | 98.2 ±0.1 (+1.2%) | 62.0 ±0.1 (+7.6%) | 93.3 ±0.2 (+3.0%) |
| XGradCAM+ | 26.9 ±0.1 | 98.3 ±0.1 | 58.9 ±0.1 | 92.5 ±0.2 |
| **Libra XGradCAM+** | 30.3 ±0.1 (+12.3%) | 100.7 ±0.2 (+2.5%) | 69.9 ±0.1 (+18.5%) | 99.0 ±0.2 (+7.1%) |
| FullGrad+ | 32.3 ±0.1 | 102.0 ±0.1 | 67.3 ±0.1 | 99.8 ±0.2 |
| **Libra FullGrad+** | **36.1** ±0.1 (+12.0%) | **104.7** ±0.1 (+2.7%) | **76.2** ±0.1 (+13.4%) | **106.3** ±0.2 (+6.5%) |

Table 61. Comparison of attribution methods and their LibraGrad-enhanced versions on the ImageNet-Hard ViT-B model.

| Method | SRG (GT) | | SRG (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 50.1 ±0.1 | 50.0 ±0.1 | 49.7 ±0.1 | 49.8 ±0.2 |
| RawAtt | 52.7 ±0.1 | 51.8 ±0.1 | 58.9 ±0.1 | 54.6 ±0.2 |
| Attention Rollout | 52.3 ±0.1 | 51.3 ±0.1 | 56.0 ±0.1 | 52.9 ±0.2 |
| AliLRP | 53.6 ±0.1 | 52.1 ±0.1 | 58.8 ±0.1 | 55.0 ±0.2 |
| AttnLRP | 57.3 ±0.1 | 54.5 ±0.1 | 66.2 ±0.1 | 59.9 ±0.2 |
| DecompX | 54.8 ±0.1 | 53.2 ±0.1 | 62.2 ±0.1 | 57.2 ±0.2 |
| Integrated Gradients | 59.0 ±0.1 | 55.7 ±0.1 | 60.4 ±0.1 | 57.2 ±0.2 |
| Input × Grad | 55.2 ±0.1 | 53.2 ±0.1 | 61.8 ±0.1 | 57.1 ±0.2 |
| **Libra Input × Grad** | 55.7 ±0.1 (+0.8%) | 53.4 ±0.1 (+0.4%) | 63.3 ±0.1 (+2.6%) | 57.7 ±0.2 (+1.0%) |
| AttCAT | 63.5 ±0.1 | 58.6 ±0.1 | 75.8 ±0.1 | 67.0 ±0.2 |
| **Libra AttCAT** | <u>65.1</u> ±0.1 (+2.6%) | <u>59.7</u> ±0.2 (+1.9%) | <u>81.6</u> ±0.1 (+7.7%) | <u>70.5</u> ±0.2 (+5.3%) |
| GenAtt | 58.7 ±0.1 | 55.6 ±0.1 | 73.5 ±0.1 | 62.8 ±0.2 |
| **Libra GenAtt** | 59.0 ±0.1 (+0.5%) | 55.9 ±0.1 (+0.4%) | 74.6 ±0.1 (+1.5%) | 63.3 ±0.2 (+0.7%) |
| TokenTM | 57.4 ±0.1 | 54.9 ±0.1 | 71.1 ±0.1 | 61.2 ±0.2 |
| **Libra TokenTM** | 57.9 ±0.1 (+0.8%) | 55.1 ±0.1 (+0.3%) | 71.5 ±0.1 (+0.6%) | 61.5 ±0.2 (+0.5%) |
| GradCAM+ | 58.6 ±0.1 | 55.1 ±0.1 | 68.5 ±0.1 | 61.2 ±0.2 |
| **Libra GradCAM+** | 61.5 ±0.1 (+5.0%) | 57.1 ±0.1 (+3.7%) | 76.0 ±0.1 (+11.0%) | 65.7 ±0.2 (+7.5%) |
| HiResCAM | 57.0 ±0.1 | 54.3 ±0.1 | 65.9 ±0.1 | 59.4 ±0.2 |
| **Libra HiResCAM** | 59.0 ±0.1 (+3.5%) | 55.6 ±0.1 (+2.4%) | 70.9 ±0.1 (+7.5%) | 62.3 ±0.2 (+5.0%) |
| XGradCAM+ | 57.4 ±0.1 | 54.4 ±0.1 | 65.5 ±0.1 | 59.4 ±0.2 |
| **Libra XGradCAM+** | 61.8 ±0.1 (+7.6%) | 57.4 ±0.2 (+5.5%) | 77.3 ±0.1 (+17.9%) | 66.5 ±0.2 (+11.8%) |
| FullGrad+ | 61.4 ±0.1 | 57.1 ±0.1 | 73.7 ±0.1 | 65.2 ±0.2 |
| **Libra FullGrad+** | **65.4** ±0.1 (+6.5%) | **59.9** ±0.1 (+4.8%) | **81.9** ±0.1 (+11.2%) | **71.0** ±0.2 (+8.9%) |

Table 62. Comparison of attribution methods and their LibraGrad-enhanced versions on the ImageNet-Hard ViT-B model.

### D.5.6. MURA ViT-B

| Method | MIF Deletion (GT) Accuracy | MIF Deletion (GT) AOPC | MIF Deletion (Predicted) Accuracy | MIF Deletion (Predicted) AOPC |
|---|---|---|---|---|
| Random | 25.4 ±0.1 | 3.3 ±0.1 | 15.1 ±0.1 | 4.2 ±0.1 |
| RawAtt | 33.4 ±0.1 | 12.6 ±0.3 | 24.8 ±0.1 | 14.9 ±0.2 |
| Attention Rollout | 29.9 ±0.1 | 8.6 ±0.2 | 21.5 ±0.1 | 10.8 ±0.2 |
| AliLRP | 28.4 ±0.1 | 7.1 ±0.2 | 19.2 ±0.1 | 8.5 ±0.2 |
| AttnLRP | 31.6 ±0.1 | 10.4 ±0.2 | 22.8 ±0.1 | 12.1 ±0.2 |
| DecompX | 30.7 ±0.1 | 9.5 ±0.2 | 21.6 ±0.1 | 10.9 ±0.2 |
| Integrated Gradients | 35.6 ±0.1 | 14.5 ±0.2 | 23.8 ±0.1 | 13.8 ±0.2 |
| Input $\times$ Grad | 33.2 ±0.1 | 12.0 ±0.2 | 25.5 ±0.1 | 14.1 ±0.2 |
| **Libra Input $\times$ Grad** | 30.7 ±0.1 (-7.5%) | 9.5 ±0.2 (-20.9%) | 21.6 ±0.1 (-15.1%) | 10.9 ±0.2 (-23.0%) |
| AttCAT | 37.8 ±0.1 | 16.7 ±0.2 | 31.1 ±0.1 | 19.6 ±0.1 |
| **Libra AttCAT** | <u>47.1</u> ±0.1 (+24.5%) | <u>25.6</u> ±0.3 (+53.0%) | <u>40.9</u> ±0.1 (+31.6%) | <u>28.9</u> ±0.2 (+47.7%) |
| GenAtt | 37.8 ±0.1 | 18.1 ±0.3 | 30.0 ±0.1 | 21.2 ±0.2 |
| **Libra GenAtt** | 38.0 ±0.1 (+0.7%) | 18.0 ±0.3 (-0.2%) | 30.1 ±0.1 (+0.4%) | 21.1 ±0.2 (-0.6%) |
| TokenTM | 36.1 ±0.1 | 16.4 ±0.3 | 28.0 ±0.1 | 19.5 ±0.2 |
| **Libra TokenTM** | 36.0 ±0.1 (-0.1%) | 16.2 ±0.3 (-1.5%) | 28.0 ±0.1 (+0.0%) | 19.2 ±0.2 (-1.7%) |
| GradCAM+ | 32.6 ±0.1 | 11.0 ±0.2 | 24.0 ±0.1 | 12.8 ±0.2 |
| **Libra GradCAM+** | 42.3 ±0.1 (+29.9%) | 20.1 ±0.3 (+82.2%) | 34.7 ±0.1 (+44.8%) | 22.4 ±0.2 (+75.5%) |
| HiResCAM | 31.4 ±0.1 | 10.4 ±0.2 | 22.2 ±0.1 | 11.8 ±0.2 |
| **Libra HiResCAM** | 37.9 ±0.1 (+20.4%) | 17.0 ±0.2 (+63.5%) | 30.1 ±0.1 (+35.7%) | 19.2 ±0.2 (+63.0%) |
| XGradCAM+ | 32.4 ±0.1 | 10.6 ±0.2 | 23.7 ±0.1 | 12.3 ±0.1 |
| **Libra XGradCAM+** | 43.4 ±0.1 (+34.1%) | 22.2 ±0.3 (+108.5%) | 36.6 ±0.1 (+54.6%) | 25.2 ±0.3 (+104.9%) |
| FullGrad+ | 39.1 ±0.1 | 17.7 ±0.2 | 32.8 ±0.1 | 20.7 ±0.2 |
| **Libra FullGrad+** | **48.7** ±0.1 (+24.5%) | **26.9** ±0.3 (+51.7%) | **43.2** ±0.1 (+31.7%) | **30.5** ±0.2 (+47.1%) |

Table 63. Comparison of attribution methods and their LibraGrad-enhanced versions on the MURA ViT-B model.

| Method | LIF Deletion (GT) | | LIF Deletion (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 75.2 ±0.1 | 97.0 ±0.1 | 85.8 ±0.1 | 96.2 ±0.1 |
| RawAtt | 75.9 ±0.1 | 97.7 ±0.1 | 85.7 ±0.1 | 96.6 ±0.1 |
| Attention Rollout | 75.0 ±0.1 | 96.4 ±0.1 | 84.1 ±0.1 | 94.9 ±0.1 |
| AliLRP | 77.8 ±0.1 | 99.9 ±0.1 | 87.0 ±0.1 | 98.5 ±0.1 |
| AttnLRP | 78.2 ±0.1 | 100.4 ±0.1 | 87.0 ±0.1 | 98.7 ±0.1 |
| DecompX | 78.7 ±0.1 | 100.9 ±0.1 | 87.7 ±0.0 | 99.4 ±0.1 |
| Integrated Gradients | 81.2 ±0.1 | 103.7 ±0.1 | 86.3 ±0.1 | 99.2 ±0.1 |
| Input × Grad | 80.4 ±0.1 | 102.6 ±0.1 | 88.2 ±0.0 | 100.5 ±0.1 |
| **Libra Input × Grad** | 78.7 ±0.1 (-2.0%) | 100.9 ±0.1 (-1.7%) | 87.7 ±0.0 (-0.5%) | 99.4 ±0.1 (-1.1%) |
| AttCAT | 82.4 ±0.1 | 105.0 ±0.1 | 89.1 ±0.0 | 102.1 ±0.1 |
| **Libra AttCAT** | <u>83.2</u> ±0.1 (+0.9%) | <u>105.7</u> ±0.2 (+0.7%) | 89.4 ±0.0 (+0.3%) | 102.4 ±0.1 (+0.2%) |
| GenAtt | 78.3 ±0.1 | 100.4 ±0.1 | 88.3 ±0.0 | 99.3 ±0.1 |
| **Libra GenAtt** | 78.3 ±0.1 (+0.0%) | 100.3 ±0.1 (-0.1%) | 88.3 ±0.0 (+0.1%) | 99.2 ±0.1 (-0.1%) |
| TokenTM | 77.1 ±0.1 | 99.1 ±0.1 | 87.4 ±0.1 | 98.2 ±0.1 |
| **Libra TokenTM** | 77.1 ±0.1 (+0.0%) | 99.1 ±0.1 (+0.0%) | 87.5 ±0.0 (+0.2%) | 98.3 ±0.1 (+0.1%) |
| GradCAM+ | 77.3 ±0.1 | 98.7 ±0.2 | 85.9 ±0.1 | 96.9 ±0.2 |
| **Libra GradCAM+** | 80.9 ±0.1 (+4.7%) | 103.0 ±0.1 (+4.3%) | 88.6 ±0.0 (+3.1%) | 100.6 ±0.1 (+3.8%) |
| HiResCAM | 76.8 ±0.1 | 99.0 ±0.2 | 86.1 ±0.1 | 97.6 ±0.1 |
| **Libra HiResCAM** | 80.1 ±0.1 (+4.3%) | 102.0 ±0.1 (+3.1%) | 87.8 ±0.0 (+1.9%) | 99.8 ±0.1 (+2.3%) |
| XGradCAM+ | 77.1 ±0.1 | 98.7 ±0.2 | 85.8 ±0.1 | 97.0 ±0.1 |
| **Libra XGradCAM+** | 81.3 ±0.1 (+5.5%) | 103.4 ±0.2 (+4.8%) | 88.2 ±0.0 (+2.8%) | 100.4 ±0.1 (+3.5%) |
| FullGrad+ | <u>83.2</u> ±0.1 | <u>105.7</u> ±0.1 | **89.5** ±0.0 | **102.7** ±0.1 |
| **Libra FullGrad+** | **84.0** ±0.1 (+0.9%) | **106.3** ±0.2 (+0.5%) | **89.5** ±0.0 (+0.0%) | **102.7** ±0.1 (+0.0%) |

Table 64. Comparison of attribution methods and their LibraGrad-enhanced versions on the MURA ViT-B model.

| Method | SRG (GT) | | SRG (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 50.3 ±0.1 | 50.2 ±0.1 | 50.4 ±0.1 | 50.2 ±0.1 |
| RawAtt | 54.7 ±0.1 | 55.2 ±0.2 | 55.3 ±0.1 | 55.8 ±0.2 |
| Attention Rollout | 52.5 ±0.1 | 52.5 ±0.2 | 52.8 ±0.1 | 52.9 ±0.2 |
| AliLRP | 53.1 ±0.1 | 53.5 ±0.2 | 53.1 ±0.1 | 53.5 ±0.1 |
| AttnLRP | 54.9 ±0.1 | 55.4 ±0.2 | 54.9 ±0.1 | 55.4 ±0.2 |
| DecompX | 54.7 ±0.1 | 55.2 ±0.2 | 54.7 ±0.1 | 55.1 ±0.1 |
| Integrated Gradients | 58.4 ±0.1 | 59.1 ±0.2 | 55.1 ±0.1 | 56.5 ±0.1 |
| Input × Grad | 56.8 ±0.1 | 57.3 ±0.2 | 56.8 ±0.1 | 57.3 ±0.2 |
| **Libra Input × Grad** | 54.7 ±0.1 (-3.6%) | 55.2 ±0.2 (-3.7%) | 54.7 ±0.1 (-3.8%) | 55.1 ±0.1 (-3.8%) |
| AttCAT | 60.1 ±0.1 | 60.8 ±0.2 | 60.1 ±0.1 | 60.8 ±0.1 |
| **Libra AttCAT** | <u>65.1</u> ±0.1 (+8.3%) | <u>65.6</u> ±0.2 (+7.9%) | <u>65.2</u> ±0.1 (+8.4%) | <u>65.6</u> ±0.2 (+7.9%) |
| GenAtt | 58.0 ±0.1 | 59.2 ±0.2 | 59.1 ±0.1 | 60.3 ±0.2 |
| **Libra GenAtt** | 58.2 ±0.1 (+0.2%) | 59.2 ±0.2 (-0.1%) | 59.2 ±0.1 (+0.1%) | 60.1 ±0.2 (-0.2%) |
| TokenTM | 56.6 ±0.1 | 57.8 ±0.2 | 57.7 ±0.1 | 58.8 ±0.2 |
| **Libra TokenTM** | 56.6 ±0.1 (+0.0%) | 57.7 ±0.2 (-0.2%) | 57.8 ±0.1 (+0.2%) | 58.7 ±0.2 (-0.2%) |
| GradCAM+ | 54.9 ±0.1 | 54.9 ±0.2 | 54.9 ±0.1 | 54.8 ±0.2 |
| **Libra GradCAM+** | 61.6 ±0.1 (+12.1%) | 61.5 ±0.2 (+12.2%) | 61.6 ±0.1 (+12.2%) | 61.5 ±0.2 (+12.2%) |
| HiResCAM | 54.1 ±0.1 | 54.7 ±0.2 | 54.1 ±0.1 | 54.7 ±0.1 |
| **Libra HiResCAM** | 59.0 ±0.1 (+9.0%) | 59.5 ±0.2 (+8.8%) | 58.9 ±0.1 (+8.9%) | 59.5 ±0.2 (+8.8%) |
| XGradCAM+ | 54.7 ±0.1 | 54.6 ±0.2 | 54.7 ±0.1 | 54.6 ±0.1 |
| **Libra XGradCAM+** | 62.4 ±0.1 (+14.0%) | 62.8 ±0.3 (+14.9%) | 62.4 ±0.1 (+14.0%) | 62.8 ±0.2 (+14.9%) |
| FullGrad+ | 61.2 ±0.1 | 61.7 ±0.2 | 61.2 ±0.1 | 61.7 ±0.1 |
| **Libra FullGrad+** | **66.4** ±0.1 (+8.5%) | **66.6** ±0.2 (+7.9%) | **66.4** ±0.1 (+8.5%) | **66.6** ±0.2 (+7.9%) |

Table 65. Comparison of attribution methods and their LibraGrad-enhanced versions on the MURA ViT-B model.

### D.5.7. Oxford Pet ViT-B

| Method | MIF Deletion (GT) | | MIF Deletion (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 14.6 ±0.1 | 4.1 ±0.1 | 13.7 ±0.1 | 4.3 ±0.1 |
| RawAtt | 37.7 ±0.1 | 29.1 ±0.3 | 37.2 ±0.1 | 29.6 ±0.3 |
| Attention Rollout | 22.2 ±0.1 | 12.2 ±0.2 | 21.2 ±0.1 | 12.4 ±0.2 |
| AliLRP | 19.7 ±0.1 | 9.7 ±0.2 | 19.0 ±0.1 | 10.0 ±0.2 |
| AttnLRP | 30.9 ±0.1 | 22.1 ±0.2 | 30.3 ±0.1 | 22.5 ±0.2 |
| DecompX | 23.2 ±0.1 | 13.8 ±0.2 | 22.5 ±0.1 | 14.1 ±0.2 |
| Integrated Gradients | 27.5 ±0.1 | 17.6 ±0.3 | 20.7 ±0.1 | 11.5 ±0.2 |
| Input × Grad | 20.9 ±0.1 | 11.2 ±0.2 | 20.4 ±0.1 | 11.6 ±0.2 |
| **Libra Input × Grad** | 24.3 ±0.1 (+16.2%) | 14.8 ±0.2 (+32.2%) | 23.5 ±0.1 (+15.4%) | 15.1 ±0.2 (+30.0%) |
| AttCAT | 37.6 ±0.1 | 26.7 ±0.3 | 37.3 ±0.1 | 27.3 ±0.4 |
| **Libra AttCAT** | <u>55.5</u> ±0.1 (+47.6%) | <u>44.3</u> ±0.3 (+65.8%) | <u>55.3</u> ±0.1 (+48.1%) | <u>44.9</u> ±0.3 (+64.7%) |
| GenAtt | 44.5 ±0.1 | 35.2 ±0.3 | 44.1 ±0.1 | 35.7 ±0.3 |
| **Libra GenAtt** | 46.8 ±0.1 (+5.3%) | 37.6 ±0.3 (+6.7%) | 46.5 ±0.1 (+5.4%) | 38.1 ±0.3 (+6.6%) |
| TokenTM | 44.4 ±0.1 | 35.6 ±0.3 | 44.0 ±0.1 | 36.1 ±0.3 |
| **Libra TokenTM** | 45.9 ±0.1 (+3.3%) | 37.0 ±0.3 (+4.0%) | 45.4 ±0.1 (+3.3%) | 37.5 ±0.3 (+3.9%) |
| GradCAM+ | 33.1 ±0.1 | 22.2 ±0.3 | 32.6 ±0.1 | 22.8 ±0.3 |
| **Libra GradCAM+** | 48.2 ±0.1 (+45.8%) | 38.0 ±0.3 (+71.0%) | 47.8 ±0.1 (+46.6%) | 38.6 ±0.3 (+69.7%) |
| HiResCAM | 18.7 ±0.1 | 8.5 ±0.2 | 18.0 ±0.1 | 8.7 ±0.2 |
| **Libra HiResCAM** | 40.2 ±0.1 (+114.4%) | 30.7 ±0.4 (+260.3%) | 39.4 ±0.1 (+119.0%) | 30.9 ±0.4 (+254.4%) |
| XGradCAM+ | 33.5 ±0.1 | 23.0 ±0.3 | 33.2 ±0.1 | 23.5 ±0.3 |
| **Libra XGradCAM+** | 52.8 ±0.1 (+57.6%) | 42.2 ±0.3 (+83.3%) | 52.6 ±0.1 (+58.4%) | 42.8 ±0.3 (+81.7%) |
| FullGrad+ | 35.6 ±0.1 | 24.5 ±0.3 | 35.3 ±0.1 | 25.0 ±0.3 |
| **Libra FullGrad+** | **57.5** ±0.1 (+61.6%) | **46.1** ±0.3 (+88.1%) | **57.3** ±0.1 (+62.3%) | **46.7** ±0.3 (+86.8%) |

Table 66. Comparison of attribution methods and their LibraGrad-enhanced versions on the Oxford Pet ViT-B model.

| Method | LIF Deletion (GT) | | LIF Deletion (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 84.9 ±0.1 | 95.6 ±0.1 | 85.8 ±0.1 | 95.4 ±0.1 |
| RawAtt | 85.0 ±0.1 | 96.3 ±0.2 | 86.0 ±0.1 | 96.1 ±0.1 |
| Attention Rollout | 81.6 ±0.1 | 92.0 ±0.2 | 82.4 ±0.1 | 91.7 ±0.2 |
| AliLRP | 86.9 ±0.1 | 98.3 ±0.1 | 87.7 ±0.0 | 98.0 ±0.1 |
| AttnLRP | 88.0 ±0.0 | 99.7 ±0.1 | 88.7 ±0.0 | 99.4 ±0.1 |
| DecompX | 87.1 ±0.1 | 98.4 ±0.1 | 88.1 ±0.0 | 98.3 ±0.1 |
| Integrated Gradients | 88.1 ±0.0 | 100.0 ±0.1 | 87.0 ±0.1 | 97.5 ±0.1 |
| Input × Grad | 88.2 ±0.0 | 99.9 ±0.1 | 88.7 ±0.0 | 99.4 ±0.1 |
| **Libra Input × Grad** | 87.4 ±0.0 (-0.9%) | 99.0 ±0.1 (-0.8%) | 88.3 ±0.0 (-0.5%) | 98.7 ±0.1 (-0.7%) |
| AttCAT | <u>89.0</u> ±0.0 | <u>101.4</u> ±0.1 | <u>89.3</u> ±0.0 | <u>100.9</u> ±0.1 |
| **Libra AttCAT** | 88.9 ±0.0 (+0.0%) | 101.3 ±0.1 (+0.0%) | <u>89.3</u> ±0.0 (+0.0%) | 100.8 ±0.1 (-0.1%) |
| GenAtt | 87.8 ±0.0 | 99.4 ±0.1 | 88.7 ±0.0 | 99.2 ±0.1 |
| **Libra GenAtt** | 87.6 ±0.0 (-0.3%) | 98.9 ±0.1 (-0.6%) | 88.4 ±0.0 (-0.3%) | 98.6 ±0.1 (-0.6%) |
| TokenTM | 87.4 ±0.0 | 99.0 ±0.1 | 88.4 ±0.0 | 98.7 ±0.1 |
| **Libra TokenTM** | 87.2 ±0.1 (-0.2%) | 98.5 ±0.1 (-0.4%) | 88.2 ±0.0 (-0.3%) | 98.3 ±0.1 (-0.5%) |
| GradCAM+ | 83.7 ±0.1 | 94.6 ±0.2 | 84.2 ±0.1 | 94.1 ±0.2 |
| **Libra GradCAM+** | 87.9 ±0.0 (+5.0%) | 99.6 ±0.1 (+5.3%) | 88.4 ±0.0 (+5.0%) | 99.1 ±0.1 (+5.3%) |
| HiResCAM | 81.1 ±0.1 | 91.4 ±0.2 | 81.6 ±0.1 | 90.9 ±0.2 |
| **Libra HiResCAM** | 85.5 ±0.1 (+5.5%) | 96.6 ±0.2 (+5.7%) | 86.2 ±0.1 (+5.7%) | 96.3 ±0.2 (+5.9%) |
| XGradCAM+ | 84.8 ±0.1 | 96.0 ±0.2 | 85.2 ±0.1 | 95.5 ±0.2 |
| **Libra XGradCAM+** | 88.2 ±0.0 (+4.0%) | 100.0 ±0.1 (+4.2%) | 88.6 ±0.0 (+4.0%) | 99.5 ±0.1 (+4.2%) |
| FullGrad+ | 88.9 ±0.0 | 101.2 ±0.1 | <u>89.3</u> ±0.0 | 100.8 ±0.1 |
| **Libra FullGrad+** | **89.1** ±0.0 (+0.2%) | **101.5** ±0.1 (+0.3%) | **89.6** ±0.0 (+0.3%) | **101.1** ±0.1 (+0.3%) |

Table 67. Comparison of attribution methods and their LibraGrad-enhanced versions on the Oxford Pet ViT-B model.

| Method | SRG (GT) | | SRG (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 49.7 ±0.1 | 49.9 ±0.1 | 49.7 ±0.1 | 49.8 ±0.1 |
| RawAtt | 61.3 ±0.1 | 62.7 ±0.2 | 61.6 ±0.1 | 62.8 ±0.2 |
| Attention Rollout | 51.9 ±0.1 | 52.1 ±0.2 | 51.8 ±0.1 | 52.0 ±0.2 |
| AliLRP | 53.3 ±0.1 | 54.0 ±0.2 | 53.3 ±0.1 | 54.0 ±0.2 |
| AttnLRP | 59.4 ±0.1 | 60.9 ±0.2 | 59.5 ±0.1 | 60.9 ±0.2 |
| DecompX | 55.1 ±0.1 | 56.1 ±0.2 | 55.3 ±0.1 | 56.2 ±0.2 |
| Integrated Gradients | 57.8 ±0.1 | 58.8 ±0.2 | 53.8 ±0.1 | 54.5 ±0.1 |
| Input × Grad | 54.6 ±0.1 | 55.5 ±0.2 | 54.5 ±0.1 | 55.5 ±0.2 |
| **Libra Input × Grad** | 55.9 ±0.1 (+2.4%) | 56.9 ±0.2 (+2.5%) | 55.9 ±0.1 (+2.5%) | 56.9 ±0.2 (+2.5%) |
| AttCAT | 63.3 ±0.1 | 64.0 ±0.2 | 63.3 ±0.1 | 64.1 ±0.3 |
| **Libra AttCAT** | <u>72.2</u> ±0.1 (+14.1%) | <u>72.8</u> ±0.2 (+13.7%) | <u>72.3</u> ±0.1 (+14.2%) | <u>72.9</u> ±0.2 (+13.7%) |
| GenAtt | 66.1 ±0.1 | 67.3 ±0.2 | 66.4 ±0.1 | 67.5 ±0.2 |
| **Libra GenAtt** | 67.2 ±0.1 (+1.6%) | 68.2 ±0.2 (+1.3%) | 67.4 ±0.1 (+1.6%) | 68.3 ±0.2 (+1.3%) |
| TokenTM | 65.9 ±0.1 | 67.3 ±0.2 | 66.2 ±0.1 | 67.4 ±0.2 |
| **Libra TokenTM** | 66.6 ±0.1 (+1.0%) | 67.8 ±0.2 (+0.8%) | 66.8 ±0.1 (+0.9%) | 67.9 ±0.2 (+0.7%) |
| GradCAM+ | 58.4 ±0.1 | 58.4 ±0.2 | 58.4 ±0.1 | 58.5 ±0.2 |
| **Libra GradCAM+** | 68.0 ±0.1 (+16.6%) | 68.8 ±0.2 (+17.8%) | 68.1 ±0.1 (+16.6%) | 68.9 ±0.2 (+17.8%) |
| HiResCAM | 49.9 ±0.1 | 50.0 ±0.2 | 49.8 ±0.1 | 49.8 ±0.2 |
| **Libra HiResCAM** | 62.8 ±0.1 (+26.0%) | 63.6 ±0.3 (+27.4%) | 62.8 ±0.1 (+26.2%) | 63.6 ±0.3 (+27.7%) |
| XGradCAM+ | 59.1 ±0.1 | 59.5 ±0.2 | 59.2 ±0.1 | 59.5 ±0.2 |
| **Libra XGradCAM+** | 70.5 ±0.1 (+19.2%) | 71.1 ±0.2 (+19.5%) | 70.6 ±0.1 (+19.3%) | 71.1 ±0.2 (+19.6%) |
| FullGrad+ | 62.2 ±0.1 | 62.9 ±0.2 | 62.3 ±0.1 | 62.9 ±0.2 |
| **Libra FullGrad+** | **73.3** ±0.1 (+17.8%) | **73.8** ±0.2 (+17.4%) | **73.4** ±0.1 (+17.9%) | **73.9** ±0.2 (+17.5%) |

Table 68. Comparison of attribution methods and their LibraGrad-enhanced versions on the Oxford Pet ViT-B model.

### D.5.8. ViT-L

| Method | MIF Deletion (GT) Accuracy | MIF Deletion (GT) AOPC | MIF Deletion (Predicted) Accuracy | MIF Deletion (Predicted) AOPC | Segmentation AP |
|---|---|---|---|---|---|
| Random | 36.9 ±0.1 | 14.1 ±0.2 | 29.5 ±0.1 | 15.8 ±0.2 | 42.0 ±0.4 |
| RawAtt | 45.4 ±0.1 | 22.9 ±0.3 | 39.1 ±0.1 | 25.3 ±0.2 | 40.2 ±0.4 |
| Attention Rollout | 39.0 ±0.1 | 16.5 ±0.3 | 31.4 ±0.1 | 18.3 ±0.3 | 39.9 ±0.3 |
| AliLRP | 39.8 ±0.1 | 17.2 ±0.3 | 33.2 ±0.1 | 19.2 ±0.2 | 42.7 ±0.4 |
| AttnLRP | 47.1 ±0.1 | 24.8 ±0.3 | 41.8 ±0.1 | 27.6 ±0.3 | 47.2 ±0.3 |
| DecompX | 44.4 ±0.1 | 22.6 ±0.3 | 38.9 ±0.1 | 25.3 ±0.3 | 54.2 ±0.3 |
| Integrated Gradients | 46.3 ±0.1 | 23.1 ±0.3 | 35.9 ±0.1 | 21.9 ±0.2 | 46.6 ±0.3 |
| Input $\times$ Grad | 40.1 ±0.1 | 17.5 ±0.3 | 33.9 ±0.1 | 19.6 ±0.2 | 43.6 ±0.4 |
| **Libra Input $\times$ Grad** | 45.9 ±0.1 (+14.4%) | 23.4 ±0.3 (+33.5%) | 40.5 ±0.1 (+19.6%) | 26.1 ±0.3 (+33.1%) | 53.6 ±0.3 (+22.9%) |
| AttCAT | 48.7 ±0.1 | 25.7 ±0.3 | 44.8 ±0.1 | 29.0 ±0.3 | 44.9 ±0.3 |
| **Libra AttCAT** | <u>64.7</u> ±0.1 (+33.0%) | <u>40.5</u> ±0.3 (+57.3%) | <u>61.3</u> ±0.1 (+36.9%) | <u>44.5</u> ±0.3 (+53.6%) | 53.3 ±0.3 (+18.8%) |
| GenAtt | 56.4 ±0.1 | 33.2 ±0.3 | 51.8 ±0.1 | 36.5 ±0.3 | 50.9 ±0.3 |
| **Libra GenAtt** | 59.7 ±0.1 (+5.9%) | 36.2 ±0.3 (+8.9%) | 55.4 ±0.1 (+6.8%) | 39.6 ±0.3 (+8.7%) | 58.6 ±0.3 (+15.1%) |
| TokenTM | 54.9 ±0.1 | 31.8 ±0.3 | 50.0 ±0.1 | 34.9 ±0.3 | 50.0 ±0.3 |
| **Libra TokenTM** | 57.3 ±0.1 (+4.5%) | 34.2 ±0.3 (+7.4%) | 52.5 ±0.1 (+5.0%) | 37.4 ±0.3 (+7.1%) | 53.9 ±0.3 (+7.9%) |
| GradCAM+ | 53.4 ±0.1 | 30.0 ±0.3 | 48.6 ±0.1 | 33.0 ±0.2 | 52.1 ±0.4 |
| **Libra GradCAM+** | 60.9 ±0.1 (+14.0%) | 36.7 ±0.3 (+22.0%) | 56.5 ±0.1 (+16.2%) | 40.1 ±0.3 (+21.8%) | 60.2 ±0.4 (+15.5%) |
| HiResCAM | 32.7 ±0.1 | 10.6 ±0.2 | 25.7 ±0.1 | 12.2 ±0.2 | 38.5 ±0.4 |
| **Libra HiResCAM** | 54.0 ±0.1 (+65.2%) | 30.2 ±0.3 (+186.3%) | 49.0 ±0.1 (+90.7%) | 33.2 ±0.3 (+171.8%) | 48.0 ±0.3 (+24.8%) |
| XGradCAM+ | 50.9 ±0.1 | 27.7 ±0.3 | 45.9 ±0.1 | 30.5 ±0.3 | 46.9 ±0.4 |
| **Libra XGradCAM+** | 63.0 ±0.1 (+23.6%) | 38.6 ±0.3 (+39.2%) | 58.8 ±0.1 (+28.1%) | 42.2 ±0.3 (+38.3%) | <u>60.3</u> ±0.4 (+28.6%) |
| FullGrad+ | 49.1 ±0.1 | 25.8 ±0.3 | 45.1 ±0.1 | 28.9 ±0.3 | 44.2 ±0.3 |
| **Libra FullGrad+** | **65.5** ±0.1 (+33.5%) | **41.2** ±0.3 (+59.5%) | **62.4** ±0.1 (+38.5%) | **45.3** ±0.3 (+56.5%) | **64.5** ±0.3 (+46.0%) |

Table 69. Comparison of attribution methods and their LibraGrad-enhanced versions on the ViT-L model. We report faithfulness metrics using Most-Influential-First Deletion, MIF with ground-truth (GT) and predicted labels, including Accuracy and Area Over Perturbation Curve (AOPC) and Segmentation Average Precision (AP). The results demonstrate that composing existing methods with LibraGrad significantly enhances their performance across all metrics.

| Method | LIF Deletion (GT) Accuracy | LIF Deletion (GT) AOPC | LIF Deletion (Predicted) Accuracy | LIF Deletion (Predicted) AOPC |
|---|---|---|---|---|
| Random | 62.9 ±0.1 | 85.4 ±0.2 | 70.2 ±0.1 | 83.7 ±0.2 |
| RawAtt | 60.3 ±0.1 | 83.3 ±0.2 | 67.6 ±0.1 | 81.5 ±0.1 |
| Attention Rollout | 61.9 ±0.1 | 84.1 ±0.2 | 68.3 ±0.1 | 81.9 ±0.2 |
| AliLRP | 65.4 ±0.1 | 87.7 ±0.2 | 72.5 ±0.1 | 85.9 ±0.2 |
| AttnLRP | 70.3 ±0.1 | 92.9 ±0.2 | 77.6 ±0.1 | 91.3 ±0.2 |
| DecompX | 68.8 ±0.1 | 91.0 ±0.2 | 75.8 ±0.1 | 89.3 ±0.2 |
| Integrated Gradients | 71.1 ±0.1 | 93.3 ±0.2 | 73.5 ±0.1 | 88.4 ±0.2 |
| Input × Grad | 65.8 ±0.1 | 88.4 ±0.2 | 72.8 ±0.1 | 86.7 ±0.1 |
| **Libra Input × Grad** | 70.1 ±0.1 (+6.6%) | 92.0 ±0.2 (+4.0%) | 76.7 ±0.1 (+5.4%) | 90.2 ±0.2 (+4.0%) |
| AttCAT | 71.8 ±0.1 | 94.3 ±0.2 | 77.5 ±0.1 | 92.6 ±0.2 |
| **Libra AttCAT** | <u>76.3</u> ±0.1 (+6.2%) | <u>98.5</u> ±0.2 (+4.5%) | <u>82.2</u> ±0.1 (+6.1%) | <u>97.1</u> ±0.2 (+4.8%) |
| GenAtt | 70.0 ±0.1 | 92.8 ±0.2 | 78.2 ±0.1 | 91.5 ±0.2 |
| **Libra GenAtt** | 70.9 ±0.1 (+1.3%) | 93.2 ±0.2 (+0.5%) | 78.8 ±0.1 (+0.7%) | 92.0 ±0.2 (+0.5%) |
| TokenTM | 68.9 ±0.1 | 91.6 ±0.2 | 77.3 ±0.1 | 90.3 ±0.2 |
| **Libra TokenTM** | 69.4 ±0.1 (+0.8%) | 92.1 ±0.2 (+0.5%) | 77.8 ±0.1 (+0.7%) | 90.8 ±0.2 (+0.6%) |
| GradCAM+ | 70.5 ±0.1 | 92.9 ±0.2 | 76.8 ±0.1 | 91.0 ±0.2 |
| **Libra GradCAM+** | 72.6 ±0.1 (+2.9%) | 94.4 ±0.2 (+1.6%) | 79.1 ±0.1 (+3.0%) | 92.7 ±0.2 (+1.8%) |
| HiResCAM | 53.6 ±0.1 | 76.7 ±0.2 | 59.3 ±0.1 | 74.2 ±0.3 |
| **Libra HiResCAM** | 67.4 ±0.1 (+25.7%) | 90.0 ±0.2 (+17.3%) | 73.8 ±0.1 (+24.4%) | 88.0 ±0.2 (+18.6%) |
| XGradCAM+ | 69.5 ±0.1 | 92.1 ±0.2 | 75.7 ±0.1 | 90.1 ±0.2 |
| **Libra XGradCAM+** | 73.5 ±0.1 (+5.7%) | 95.3 ±0.2 (+3.5%) | 80.0 ±0.1 (+5.6%) | 93.7 ±0.2 (+3.9%) |
| FullGrad+ | 71.5 ±0.1 | 93.8 ±0.2 | 76.8 ±0.1 | 91.8 ±0.2 |
| **Libra FullGrad+** | **76.8** ±0.1 (+7.5%) | **98.9** ±0.2 (+5.4%) | **82.6** ±0.1 (+7.6%) | **97.4** ±0.2 (+6.0%) |

Table 70. Comparison of attribution methods and their LibraGrad-enhanced versions on the ViT-L model.

| Method | SRG (GT) Accuracy | SRG (GT) AOPC | SRG (Predicted) Accuracy | SRG (Predicted) AOPC |
|---|---|---|---|---|
| Random | 49.9 ±0.1 | 49.7 ±0.2 | 49.8 ±0.1 | 49.8 ±0.2 |
| RawAtt | 52.9 ±0.1 | 53.1 ±0.2 | 53.3 ±0.1 | 53.4 ±0.2 |
| Attention Rollout | 50.4 ±0.1 | 50.3 ±0.3 | 49.9 ±0.1 | 50.1 ±0.2 |
| AliLRP | 52.6 ±0.1 | 52.4 ±0.2 | 52.8 ±0.1 | 52.5 ±0.2 |
| AttnLRP | 58.7 ±0.1 | 58.8 ±0.3 | 59.7 ±0.1 | 59.5 ±0.2 |
| DecompX | 56.6 ±0.1 | 56.8 ±0.3 | 57.4 ±0.1 | 57.3 ±0.2 |
| Integrated Gradients | 58.7 ±0.1 | 58.2 ±0.3 | 54.7 ±0.1 | 55.1 ±0.2 |
| Input × Grad | 53.0 ±0.1 | 53.0 ±0.2 | 53.3 ±0.1 | 53.2 ±0.2 |
| **Libra Input × Grad** | 58.0 ±0.1 (+9.5%) | 57.7 ±0.3 (+8.9%) | 58.6 ±0.1 (+9.9%) | 58.2 ±0.2 (+9.4%) |
| AttCAT | 60.2 ±0.1 | 60.0 ±0.2 | 61.2 ±0.1 | 60.8 ±0.2 |
| **Libra AttCAT** | <u>70.5</u> ±0.1 (+17.0%) | <u>69.5</u> ±0.3 (+15.8%) | <u>71.8</u> ±0.1 (+17.4%) | <u>70.8</u> ±0.2 (+16.4%) |
| GenAtt | 63.2 ±0.1 | 63.0 ±0.2 | 65.0 ±0.1 | 64.0 ±0.2 |
| **Libra GenAtt** | 65.3 ±0.1 (+3.3%) | 64.7 ±0.3 (+2.7%) | 67.1 ±0.1 (+3.2%) | 65.8 ±0.3 (+2.8%) |
| TokenTM | 61.9 ±0.1 | 61.7 ±0.3 | 63.6 ±0.1 | 62.6 ±0.2 |
| **Libra TokenTM** | 63.4 ±0.1 (+2.4%) | 63.1 ±0.3 (+2.3%) | 65.2 ±0.1 (+2.4%) | 64.1 ±0.3 (+2.4%) |
| GradCAM+ | 62.0 ±0.1 | 61.5 ±0.3 | 62.7 ±0.1 | 62.0 ±0.2 |
| **Libra GradCAM+** | 66.7 ±0.1 (+7.7%) | 65.5 ±0.3 (+6.6%) | 67.8 ±0.1 (+8.1%) | 66.4 ±0.2 (+7.2%) |
| HiResCAM | 43.2 ±0.1 | 43.6 ±0.2 | 42.5 ±0.1 | 43.2 ±0.2 |
| **Libra HiResCAM** | 60.7 ±0.1 (+40.7%) | 60.1 ±0.2 (+37.7%) | 61.4 ±0.1 (+44.4%) | 60.6 ±0.2 (+40.3%) |
| XGradCAM+ | 60.2 ±0.1 | 59.9 ±0.3 | 60.8 ±0.1 | 60.3 ±0.2 |
| **Libra XGradCAM+** | 68.2 ±0.1 (+13.3%) | 66.9 ±0.3 (+11.8%) | 69.4 ±0.1 (+14.1%) | 68.0 ±0.3 (+12.6%) |
| FullGrad+ | 60.3 ±0.1 | 59.8 ±0.2 | 60.9 ±0.1 | 60.4 ±0.2 |
| **Libra FullGrad+** | **71.2** ±0.1 (+18.1%) | **70.0** ±0.3 (+17.1%) | **72.5** ±0.1 (+19.0%) | **71.3** ±0.2 (+18.1%) |

Table 71. Comparison of attribution methods and their LibraGrad-enhanced versions on the ViT-L model.

### D.5.9. EVA2-S

| Method | MIF Deletion (GT) Accuracy | MIF Deletion (GT) AOPC | MIF Deletion (Predicted) Accuracy | MIF Deletion (Predicted) AOPC | Segmentation AP |
|---|---|---|---|---|---|
| Random | 29.9 ±0.1 | 6.6 ±0.2 | 21.2 ±0.1 | 8.2 ±0.2 | 37.7 ±0.3 |
| RawAtt | 55.4 ±0.1 | 30.3 ±0.3 | 50.8 ±0.1 | 33.9 ±0.3 | 59.0 ±0.3 |
| Attention Rollout | 47.0 ±0.1 | 22.0 ±0.4 | 41.1 ±0.1 | 24.9 ±0.3 | 45.3 ±0.3 |
| AliLRP | 52.8 ±0.1 | 27.7 ±0.4 | 48.0 ±0.1 | 31.3 ±0.3 | 58.7 ±0.3 |
| AttnLRP | 66.6 ±0.1 | 39.6 ±0.3 | 63.5 ±0.1 | 44.2 ±0.2 | 73.1 ±0.2 |
| DecompX | 51.6 ±0.1 | 27.0 ±0.4 | 46.8 ±0.1 | 30.7 ±0.3 | 60.0 ±0.3 |
| Integrated Gradients | 46.2 ±0.1 | 21.0 ±0.3 | 34.8 ±0.1 | 19.3 ±0.2 | 51.2 ±0.3 |
| Input × Grad | 37.9 ±0.1 | 14.1 ±0.2 | 32.3 ±0.1 | 17.0 ±0.2 | 42.5 ±0.3 |
| **Libra Input × Grad** | 67.0 ±0.1 (+76.8%) | 39.6 ±0.3 (+180.9%) | 64.1 ±0.1 (+98.5%) | 44.4 ±0.3 (+161.4%) | 72.1 ±0.3 (+69.5%) |
| AttCAT | 56.9 ±0.1 | 30.9 ±0.2 | 54.1 ±0.1 | 35.3 ±0.3 | 58.9 ±0.3 |
| **Libra AttCAT** | <u>72.1</u> ±0.1 (+26.8%) | <u>43.8</u> ±0.3 (+41.8%) | <u>69.5</u> ±0.1 (+28.4%) | <u>48.7</u> ±0.2 (+38.1%) | 75.1 ±0.3 (+27.6%) |
| GenAtt | 46.3 ±0.1 | 21.2 ±0.2 | 40.7 ±0.1 | 24.3 ±0.2 | 42.3 ±0.3 |
| **Libra GenAtt** | 47.7 ±0.1 (+3.1%) | 22.5 ±0.3 (+6.5%) | 42.1 ±0.1 (+3.6%) | 25.6 ±0.2 (+5.4%) | 44.3 ±0.3 (+4.7%) |
| TokenTM | 50.4 ±0.1 | 25.1 ±0.3 | 44.7 ±0.1 | 28.3 ±0.3 | 45.5 ±0.3 |
| **Libra TokenTM** | 51.6 ±0.1 (+2.4%) | 25.6 ±0.3 (+1.9%) | 46.0 ±0.1 (+2.8%) | 28.8 ±0.3 (+1.6%) | 46.7 ±0.3 (+2.7%) |
| GradCAM+ | 50.6 ±0.1 | 25.1 ±0.3 | 47.1 ±0.1 | 29.0 ±0.3 | 49.3 ±0.4 |
| **Libra GradCAM+** | 69.9 ±0.1 (+38.0%) | 41.4 ±0.3 (+65.2%) | 67.0 ±0.1 (+42.1%) | 46.1 ±0.2 (+58.9%) | <u>79.8</u> ±0.3 (+62.1%) |
| HiResCAM | 63.1 ±0.1 | 36.1 ±0.2 | 59.1 ±0.1 | 40.1 ±0.2 | 73.2 ±0.3 |
| **Libra HiResCAM** | 65.9 ±0.1 (+4.4%) | 38.6 ±0.3 (+6.8%) | 62.6 ±0.1 (+6.0%) | 42.9 ±0.2 (+7.1%) | 76.5 ±0.3 (+4.5%) |
| XGradCAM+ | 53.7 ±0.1 | 27.9 ±0.2 | 50.2 ±0.1 | 31.9 ±0.2 | 55.2 ±0.4 |
| **Libra XGradCAM+** | 71.9 ±0.1 (+34.1%) | 43.3 ±0.3 (+54.9%) | 69.3 ±0.1 (+38.0%) | 48.1 ±0.2 (+50.8%) | **82.7** ±0.3 (+49.9%) |
| FullGrad+ | 50.9 ±0.1 | 25.7 ±0.2 | 48.0 ±0.1 | 30.0 ±0.2 | 51.5 ±0.3 |
| **Libra FullGrad+** | **74.1** ±0.1 (+45.5%) | **45.5** ±0.3 (+77.0%) | **71.7** ±0.1 (+49.4%) | **50.5** ±0.2 (+68.5%) | 79.4 ±0.3 (+54.2%) |

Table 72. Comparison of attribution methods and their LibraGrad-enhanced versions on the EVA2-S model. We report faithfulness metrics using Most-Influential-First Deletion, MIF with ground-truth (GT) and predicted labels, including Accuracy and Area Over Perturbation Curve (AOPC) and Segmentation Average Precision (AP). The results demonstrate that composing existing methods with LibraGrad significantly enhances their performance across all metrics.

| Method | LIF Deletion (GT) Accuracy | LIF Deletion (GT) AOPC | LIF Deletion (Predicted) Accuracy | LIF Deletion (Predicted) AOPC |
|---|---|---|---|---|
| Random | 70.0 ±0.1 | 93.5 ±0.2 | 79.0 ±0.1 | 92.3 ±0.2 |
| RawAtt | 73.3 ±0.1 | 96.9 ±0.1 | 82.7 ±0.1 | 95.7 ±0.1 |
| Attention Rollout | 70.1 ±0.1 | 93.6 ±0.2 | 78.8 ±0.1 | 91.9 ±0.2 |
| AliLRP | 79.7 ±0.1 | 102.6 ±0.2 | 87.2 ±0.1 | 100.9 ±0.1 |
| AttnLRP | 78.8 ±0.1 | 103.0 ±0.1 | 87.5 ±0.0 | 101.9 ±0.1 |
| DecompX | 76.3 ±0.1 | 100.4 ±0.1 | 85.8 ±0.1 | 99.5 ±0.1 |
| Integrated Gradients | 82.0 ±0.1 | 105.3 ±0.2 | 83.5 ±0.1 | 99.8 ±0.2 |
| Input × Grad | 76.5 ±0.1 | 100.0 ±0.2 | 84.0 ±0.1 | 98.9 ±0.2 |
| **Libra Input × Grad** | 82.0 ±0.1 (+7.2%) | 104.7 ±0.1 (+4.7%) | <u>88.3</u> ±0.0 (+5.1%) | 102.5 ±0.1 (+3.7%) |
| AttCAT | **82.7** ±0.1 | **107.2** ±0.2 | 87.8 ±0.0 | **105.3** ±0.1 |
| **Libra AttCAT** | 82.2 ±0.1 (-0.6%) | 105.0 ±0.1 (-2.0%) | <u>88.3</u> ±0.0 (+0.5%) | 102.8 ±0.1 (-2.4%) |
| GenAtt | 71.9 ±0.1 | 95.3 ±0.2 | 80.7 ±0.1 | 94.0 ±0.2 |
| **Libra GenAtt** | 72.7 ±0.1 (+1.1%) | 95.9 ±0.2 (+0.6%) | 81.6 ±0.1 (+1.1%) | 94.5 ±0.2 (+0.6%) |
| TokenTM | 73.3 ±0.1 | 96.6 ±0.2 | 82.1 ±0.1 | 95.2 ±0.1 |
| **Libra TokenTM** | 72.9 ±0.1 (-0.6%) | 96.3 ±0.2 (-0.3%) | 81.9 ±0.1 (-0.2%) | 94.8 ±0.2 (-0.4%) |
| GradCAM+ | 77.3 ±0.1 | 100.8 ±0.3 | 82.8 ±0.1 | 98.6 ±0.2 |
| **Libra GradCAM+** | 80.1 ±0.1 (+3.6%) | 102.6 ±0.1 (+1.8%) | 86.4 ±0.1 (+4.4%) | 100.3 ±0.1 (+1.8%) |
| HiResCAM | 79.3 ±0.1 | 103.1 ±0.2 | 86.1 ±0.1 | 101.0 ±0.1 |
| **Libra HiResCAM** | 79.4 ±0.1 (+0.1%) | 102.4 ±0.2 (-0.7%) | 86.3 ±0.1 (+0.3%) | 100.5 ±0.1 (-0.6%) |
| XGradCAM+ | 78.3 ±0.1 | 101.9 ±0.3 | 83.8 ±0.1 | 99.7 ±0.2 |
| **Libra XGradCAM+** | 80.1 ±0.1 (+2.3%) | 102.6 ±0.1 (+0.7%) | 86.6 ±0.1 (+3.4%) | 100.3 ±0.1 (+0.6%) |
| FullGrad+ | 82.1 ±0.1 | <u>106.6</u> ±0.3 | 86.8 ±0.1 | <u>104.5</u> ±0.2 |
| **Libra FullGrad+** | <u>82.6</u> ±0.1 (+0.7%) | 105.3 ±0.2 (-1.2%) | **88.5** ±0.0 (+1.9%) | 103.0 ±0.1 (-1.4%) |

Table 73. Comparison of attribution methods and their LibraGrad-enhanced versions on the EVA2-S model.

| Method | SRG (GT) Accuracy | SRG (GT) AOPC | SRG (Predicted) Accuracy | SRG (Predicted) AOPC |
|---|---|---|---|---|
| Random | 50.0 ±0.1 | 50.0 ±0.2 | 50.1 ±0.1 | 50.2 ±0.2 |
| RawAtt | 64.3 ±0.1 | 63.6 ±0.2 | 66.8 ±0.1 | 64.8 ±0.2 |
| Attention Rollout | 58.5 ±0.1 | 57.8 ±0.3 | 59.9 ±0.1 | 58.4 ±0.3 |
| AliLRP | 66.2 ±0.1 | 65.1 ±0.3 | 67.6 ±0.1 | 66.1 ±0.2 |
| AttnLRP | 72.7 ±0.1 | 71.3 ±0.2 | 75.5 ±0.1 | 73.1 ±0.2 |
| DecompX | 64.0 ±0.1 | 63.7 ±0.3 | 66.3 ±0.1 | 65.1 ±0.2 |
| Integrated Gradients | 64.1 ±0.1 | 63.1 ±0.2 | 59.2 ±0.1 | 59.6 ±0.2 |
| Input × Grad | 57.2 ±0.1 | 57.1 ±0.2 | 58.2 ±0.1 | 57.9 ±0.2 |
| **Libra Input × Grad** | 74.5 ±0.1 (+30.3%) | 72.2 ±0.3 (+26.5%) | 76.2 ±0.1 (+31.0%) | 73.4 ±0.2 (+26.8%) |
| AttCAT | 69.8 ±0.1 | 69.0 ±0.2 | 71.0 ±0.1 | 70.3 ±0.2 |
| **Libra AttCAT** | <u>77.2</u> ±0.1 (+10.6%) | <u>74.4</u> ±0.2 (+7.8%) | <u>78.9</u> ±0.1 (+11.2%) | <u>75.7</u> ±0.2 (+7.8%) |
| GenAtt | 59.1 ±0.1 | 58.2 ±0.2 | 60.7 ±0.1 | 59.1 ±0.2 |
| **Libra GenAtt** | 60.2 ±0.1 (+1.9%) | 59.2 ±0.2 (+1.7%) | 61.9 ±0.1 (+1.9%) | 60.0 ±0.2 (+1.6%) |
| TokenTM | 61.8 ±0.1 | 60.9 ±0.2 | 63.4 ±0.1 | 61.7 ±0.2 |
| **Libra TokenTM** | 62.2 ±0.1 (+0.6%) | 61.0 ±0.3 (+0.1%) | 63.9 ±0.1 (+0.8%) | 61.8 ±0.2 (+0.1%) |
| GradCAM+ | 64.0 ±0.1 | 62.9 ±0.3 | 65.0 ±0.1 | 63.8 ±0.2 |
| **Libra GradCAM+** | 75.0 ±0.1 (+17.2%) | 72.0 ±0.2 (+14.4%) | 76.7 ±0.1 (+18.1%) | 73.2 ±0.2 (+14.7%) |
| HiResCAM | 71.2 ±0.1 | 69.6 ±0.2 | 72.6 ±0.1 | 70.6 ±0.2 |
| **Libra HiResCAM** | 72.6 ±0.1 (+2.0%) | 70.5 ±0.2 (+1.3%) | 74.5 ±0.1 (+2.6%) | 71.7 ±0.2 (+1.6%) |
| XGradCAM+ | 66.0 ±0.1 | 64.9 ±0.3 | 67.0 ±0.1 | 65.8 ±0.2 |
| **Libra XGradCAM+** | 76.0 ±0.1 (+15.2%) | 72.9 ±0.2 (+12.3%) | 78.0 ±0.1 (+16.4%) | 74.2 ±0.2 (+12.8%) |
| FullGrad+ | 66.5 ±0.1 | 66.2 ±0.3 | 67.4 ±0.1 | 67.2 ±0.2 |
| **Libra FullGrad+** | **78.3** ±0.1 (+17.8%) | **75.4** ±0.3 (+14.0%) | **80.1** ±0.1 (+18.8%) | **76.8** ±0.2 (+14.2%) |

Table 74. Comparison of attribution methods and their LibraGrad-enhanced versions on the EVA2-S model.

#### D.5.10. FlexiViT-L

| Method | MIF Deletion (GT) | | MIF Deletion (Predicted) | | Segmentation |
|---|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC | AP |
| Random | 28.8 ±0.1 | 5.2 ±0.2 | 19.2 ±0.1 | 6.4 ±0.2 | 39.8 ±0.4 |
| RawAtt | 47.3 ±0.1 | 23.6 ±0.3 | 41.7 ±0.1 | 26.5 ±0.3 | 49.8 ±0.3 |
| Attention Rollout | 31.7 ±0.1 | 8.2 ±0.2 | 23.2 ±0.1 | 9.7 ±0.2 | 42.2 ±0.3 |
| AliLRP | 32.5 ±0.1 | 8.8 ±0.2 | 24.9 ±0.1 | 10.5 ±0.2 | 49.6 ±0.3 |
| AttnLRP | 30.3 ±0.1 | 6.6 ±0.2 | 21.8 ±0.1 | 8.3 ±0.2 | 43.4 ±0.4 |
| DecompX | 42.0 ±0.1 | 18.1 ±0.2 | 35.5 ±0.1 | 20.7 ±0.2 | 59.2 ±0.3 |
| Integrated Gradients | 31.4 ±0.1 | 8.3 ±0.2 | 22.3 ±0.1 | 9.4 ±0.2 | 41.3 ±0.4 |
| Input × Grad | 28.5 ±0.1 | 5.1 ±0.2 | 19.9 ±0.1 | 6.5 ±0.2 | 41.4 ±0.4 |
| **Libra Input × Grad** | 42.6 ±0.1 (+49.6%) | 18.6 ±0.2 (+263.5%) | 36.4 ±0.1 (+82.8%) | 21.3 ±0.2 (+227.8%) | 60.4 ±0.3 (+45.9%) |
| AttCAT | 45.3 ±0.1 | 18.9 ±0.3 | 41.9 ±0.1 | 22.6 ±0.3 | 45.1 ±0.3 |
| **Libra AttCAT** | <u>61.8</u> ±0.1 (+36.5%) | <u>35.5</u> ±0.3 (+87.9%) | <u>58.4</u> ±0.1 (+39.3%) | <u>39.6</u> ±0.3 (+75.3%) | 74.4 ±0.3 (+65.1%) |
| GenAtt | 57.2 ±0.1 | 31.4 ±0.3 | 53.0 ±0.1 | 35.1 ±0.3 | 75.1 ±0.2 |
| **Libra GenAtt** | 58.3 ±0.1 (+1.9%) | 32.9 ±0.3 (+4.8%) | 54.1 ±0.1 (+2.0%) | 36.7 ±0.3 (+4.5%) | <u>79.4</u> ±0.2 (+5.7%) |
| TokenTM | 54.3 ±0.1 | 29.3 ±0.3 | 49.3 ±0.1 | 32.7 ±0.3 | 72.2 ±0.2 |
| **Libra TokenTM** | 55.7 ±0.1 (+2.5%) | 30.9 ±0.3 (+5.4%) | 51.0 ±0.1 (+3.4%) | 34.4 ±0.3 (+5.4%) | 76.2 ±0.2 (+5.5%) |
| GradCAM+ | 35.8 ±0.1 | 10.9 ±0.2 | 28.7 ±0.1 | 13.1 ±0.2 | 40.5 ±0.4 |
| **Libra GradCAM+** | 40.2 ±0.1 (+12.6%) | 15.7 ±0.2 (+44.3%) | 33.7 ±0.1 (+17.3%) | 18.4 ±0.3 (+40.6%) | 50.2 ±0.4 (+23.7%) |
| HiResCAM | 31.2 ±0.1 | 7.2 ±0.2 | 23.8 ±0.1 | 9.0 ±0.2 | 43.7 ±0.3 |
| **Libra HiResCAM** | 60.1 ±0.1 (+92.8%) | 34.2 ±0.3 (+372.2%) | 56.5 ±0.1 (+137.7%) | 38.1 ±0.3 (+322.1%) | **81.6** ±0.3 (+86.6%) |
| XGradCAM+ | 33.4 ±0.1 | 7.8 ±0.2 | 26.6 ±0.1 | 9.9 ±0.2 | 38.5 ±0.4 |
| **Libra XGradCAM+** | 49.7 ±0.1 (+48.9%) | 24.1 ±0.3 (+207.6%) | 44.3 ±0.1 (+66.5%) | 27.2 ±0.3 (+174.3%) | 63.3 ±0.4 (+64.4%) |
| FullGrad+ | 43.0 ±0.1 | 17.5 ±0.3 | 38.9 ±0.1 | 20.8 ±0.3 | 44.1 ±0.3 |
| **Libra FullGrad+** | **62.4** ±0.1 (+45.2%) | **35.8** ±0.3 (+104.2%) | **59.1** ±0.1 (+51.9%) | **39.8** ±0.3 (+91.6%) | 75.1 ±0.3 (+70.3%) |

Table 75. Comparison of attribution methods and their LibraGrad-enhanced versions on the FlexiViT-L model. We report faithfulness metrics using Most-Influential-First Deletion, MIF with ground-truth (GT) and predicted labels, including Accuracy and Area Over Perturbation Curve (AOPC) and Segmentation Average Precision (AP). The results demonstrate that composing existing methods with LibraGrad significantly enhances their performance across all metrics.

| Method | LIF Deletion (GT) | | LIF Deletion (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 70.7 ±0.1 | 94.7 ±0.2 | 80.7 ±0.1 | 93.7 ±0.1 |
| RawAtt | 72.8 ±0.1 | 96.6 ±0.1 | 82.6 ±0.1 | 95.4 ±0.1 |
| Attention Rollout | 65.0 ±0.1 | 88.2 ±0.2 | 72.7 ±0.1 | 86.2 ±0.2 |
| AliLRP | 75.8 ±0.1 | 98.9 ±0.1 | 84.7 ±0.1 | 97.8 ±0.1 |
| AttnLRP | 68.9 ±0.1 | 92.2 ±0.2 | 77.9 ±0.1 | 90.9 ±0.2 |
| DecompX | 76.7 ±0.1 | 100.6 ±0.1 | 86.2 ±0.1 | 99.6 ±0.1 |
| Integrated Gradients | 70.2 ±0.1 | 93.8 ±0.2 | 77.7 ±0.1 | 91.9 ±0.2 |
| Input × Grad | 69.6 ±0.1 | 92.8 ±0.2 | 78.3 ±0.1 | 91.4 ±0.2 |
| **Libra Input × Grad** | 78.2 ±0.1 (+12.4%) | 101.6 ±0.1 (+9.5%) | 86.9 ±0.1 (+10.9%) | 100.4 ±0.1 (+9.9%) |
| AttCAT | **83.1** ±0.1 | **106.1** ±0.2 | <u>88.3</u> ±0.0 | **104.5** ±0.2 |
| **Libra AttCAT** | 81.4 ±0.1 (-2.0%) | 104.4 ±0.1 (-1.5%) | **88.5** ±0.0 (+0.2%) | 103.0 ±0.1 (-1.4%) |
| GenAtt | 77.3 ±0.1 | 100.8 ±0.1 | 87.0 ±0.1 | 99.7 ±0.1 |
| **Libra GenAtt** | 77.0 ±0.1 (-0.4%) | 100.5 ±0.1 (-0.3%) | 86.6 ±0.1 (-0.4%) | 99.4 ±0.1 (-0.3%) |
| TokenTM | 76.0 ±0.1 | 99.6 ±0.1 | 86.0 ±0.1 | 98.5 ±0.1 |
| **Libra TokenTM** | 75.7 ±0.1 (-0.4%) | 99.6 ±0.1 (+0.0%) | 85.8 ±0.1 (-0.2%) | 98.6 ±0.1 (+0.1%) |
| GradCAM+ | 64.7 ±0.1 | 87.5 ±0.2 | 72.3 ±0.1 | 85.7 ±0.2 |
| **Libra GradCAM+** | 72.9 ±0.1 (+12.8%) | 95.5 ±0.1 (+9.1%) | 80.6 ±0.1 (+11.5%) | 93.8 ±0.1 (+9.5%) |
| HiResCAM | 70.0 ±0.1 | 92.6 ±0.2 | 78.7 ±0.1 | 91.2 ±0.2 |
| **Libra HiResCAM** | 80.7 ±0.1 (+15.3%) | 103.4 ±0.1 (+11.6%) | 87.3 ±0.0 (+11.0%) | 101.6 ±0.1 (+11.3%) |
| XGradCAM+ | 65.0 ±0.1 | 86.6 ±0.3 | 72.3 ±0.1 | 84.7 ±0.3 |
| **Libra XGradCAM+** | 77.5 ±0.1 (+19.3%) | 100.4 ±0.1 (+15.9%) | 85.3 ±0.1 (+18.1%) | 99.0 ±0.1 (+16.8%) |
| FullGrad+ | 81.4 ±0.1 | 104.3 ±0.2 | 87.8 ±0.0 | <u>103.2</u> ±0.2 |
| **Libra FullGrad+** | <u>81.5</u> ±0.1 (+0.1%) | <u>104.5</u> ±0.1 (+0.2%) | <u>88.3</u> ±0.0 (+0.6%) | 103.0 ±0.1 (-0.2%) |

Table 76. Comparison of attribution methods and their LibraGrad-enhanced versions on the FlexiViT-L model.

| Method | SRG (GT) | | SRG (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 49.8 ±0.1 | 50.0 ±0.2 | 49.9 ±0.1 | 50.0 ±0.2 |
| RawAtt | 60.1 ±0.1 | 60.1 ±0.2 | 62.1 ±0.1 | 60.9 ±0.2 |
| Attention Rollout | 48.3 ±0.1 | 48.2 ±0.2 | 48.0 ±0.1 | 48.0 ±0.2 |
| AliLRP | 54.1 ±0.1 | 53.9 ±0.1 | 54.8 ±0.1 | 54.2 ±0.1 |
| AttnLRP | 49.6 ±0.1 | 49.4 ±0.2 | 49.9 ±0.1 | 49.6 ±0.2 |
| DecompX | 59.3 ±0.1 | 59.3 ±0.2 | 60.9 ±0.1 | 60.1 ±0.1 |
| Integrated Gradients | 50.8 ±0.1 | 51.1 ±0.2 | 50.0 ±0.1 | 50.7 ±0.2 |
| Input × Grad | 49.0 ±0.1 | 49.0 ±0.2 | 49.1 ±0.1 | 48.9 ±0.2 |
| **Libra Input × Grad** | 60.4 ±0.1 (+23.2%) | 60.1 ±0.2 (+22.8%) | 61.6 ±0.1 (+25.5%) | 60.8 ±0.1 (+24.3%) |
| AttCAT | 64.2 ±0.1 | 62.5 ±0.3 | 65.1 ±0.1 | 63.5 ±0.3 |
| **Libra AttCAT** | <u>71.6</u> ±0.1 (+11.6%) | <u>70.0</u> ±0.2 (+12.0%) | <u>73.4</u> ±0.1 (+12.8%) | <u>71.3</u> ±0.2 (+12.2%) |
| GenAtt | 67.3 ±0.1 | 66.1 ±0.2 | 70.0 ±0.1 | 67.4 ±0.2 |
| **Libra GenAtt** | 67.6 ±0.1 (+0.5%) | 66.7 ±0.2 (+0.9%) | 70.4 ±0.1 (+0.5%) | 68.0 ±0.2 (+1.0%) |
| TokenTM | 65.2 ±0.1 | 64.4 ±0.2 | 67.6 ±0.1 | 65.6 ±0.2 |
| **Libra TokenTM** | 65.7 ±0.1 (+0.8%) | 65.3 ±0.2 (+1.3%) | 68.4 ±0.1 (+1.1%) | 66.5 ±0.2 (+1.4%) |
| GradCAM+ | 50.2 ±0.1 | 49.2 ±0.2 | 50.5 ±0.1 | 49.4 ±0.2 |
| **Libra GradCAM+** | 56.6 ±0.1 (+12.7%) | 55.6 ±0.2 (+13.0%) | 57.2 ±0.1 (+13.2%) | 56.1 ±0.2 (+13.6%) |
| HiResCAM | 50.6 ±0.1 | 49.9 ±0.2 | 51.2 ±0.1 | 50.1 ±0.2 |
| **Libra HiResCAM** | 70.4 ±0.1 (+39.2%) | 68.8 ±0.3 (+37.8%) | 71.9 ±0.1 (+40.4%) | 69.8 ±0.2 (+39.3%) |
| XGradCAM+ | 49.2 ±0.1 | 47.2 ±0.3 | 49.4 ±0.1 | 47.3 ±0.3 |
| **Libra XGradCAM+** | 63.6 ±0.1 (+29.3%) | 62.3 ±0.2 (+31.8%) | 64.8 ±0.1 (+31.1%) | 63.1 ±0.2 (+33.3%) |
| FullGrad+ | 62.2 ±0.1 | 60.9 ±0.3 | 63.3 ±0.1 | 62.0 ±0.2 |
| **Libra FullGrad+** | **71.9** ±0.1 (+15.7%) | **70.1** ±0.2 (+15.1%) | **73.7** ±0.1 (+16.3%) | **71.4** ±0.2 (+15.2%) |

Table 77. Comparison of attribution methods and their LibraGrad-enhanced versions on the FlexiViT-L model.

### D.5.11. BEiT2-L

| Method | MIF Deletion (GT) Accuracy | MIF Deletion (GT) AOPC | MIF Deletion (Predicted) Accuracy | MIF Deletion (Predicted) AOPC | Segmentation AP |
|---|---|---|---|---|---|
| Random | 25.1 ±0.1 | 5.6 ±0.2 | 18.3 ±0.1 | 6.8 ±0.1 | 39.8 ±0.4 |
| RawAtt | 34.2 ±0.1 | 15.3 ±0.2 | 29.5 ±0.1 | 17.5 ±0.2 | 47.6 ±0.3 |
| Attention Rollout | 26.0 ±0.1 | 7.2 ±0.1 | 19.7 ±0.1 | 8.6 ±0.1 | 42.2 ±0.3 |
| AliLRP | 31.9 ±0.1 | 12.4 ±0.2 | 26.2 ±0.1 | 13.9 ±0.2 | 43.9 ±0.3 |
| AttnLRP | 42.1 ±0.1 | 22.6 ±0.3 | 37.7 ±0.1 | 25.0 ±0.2 | 66.0 ±0.3 |
| DecompX | 36.5 ±0.1 | 17.3 ±0.3 | 31.7 ±0.1 | 19.4 ±0.2 | 55.6 ±0.3 |
| Integrated Gradients | 31.7 ±0.1 | 12.5 ±0.2 | 23.2 ±0.1 | 11.9 ±0.1 | 46.7 ±0.3 |
| Input × Grad | 28.2 ±0.1 | 9.0 ±0.1 | 21.8 ±0.1 | 10.3 ±0.1 | 39.6 ±0.4 |
| **Libra Input × Grad** | 37.7 ±0.1 (+33.6%) | 18.0 ±0.2 (+100.2%) | 33.0 ±0.1 (+51.4%) | 20.2 ±0.2 (+96.6%) | 54.8 ±0.3 (+38.4%) |
| AttCAT | 38.4 ±0.1 | 18.9 ±0.2 | 33.9 ±0.1 | 21.0 ±0.2 | 52.2 ±0.3 |
| **Libra AttCAT** | <u>52.5</u> ±0.1 (+36.6%) | <u>31.6</u> ±0.3 (+66.8%) | <u>48.9</u> ±0.1 (+44.4%) | <u>34.6</u> ±0.2 (+64.9%) | 65.5 ±0.3 (+25.4%) |
| GenAtt | 35.6 ±0.1 | 17.0 ±0.3 | 30.8 ±0.1 | 19.2 ±0.2 | 47.9 ±0.3 |
| **Libra GenAtt** | 37.6 ±0.1 (+5.6%) | 18.4 ±0.3 (+8.4%) | 32.9 ±0.1 (+6.8%) | 20.7 ±0.3 (+7.9%) | 48.8 ±0.3 (+1.8%) |
| TokenTM | 43.9 ±0.1 | 24.3 ±0.3 | 39.6 ±0.1 | 26.8 ±0.3 | 56.0 ±0.3 |
| **Libra TokenTM** | 42.6 ±0.1 (-2.8%) | 23.1 ±0.3 (-4.8%) | 38.3 ±0.1 (-3.4%) | 25.5 ±0.3 (-5.0%) | 54.2 ±0.3 (-3.3%) |
| GradCAM+ | 38.4 ±0.1 | 18.2 ±0.2 | 33.4 ±0.1 | 20.1 ±0.2 | 53.5 ±0.4 |
| **Libra GradCAM+** | 42.3 ±0.1 (+10.2%) | 22.0 ±0.2 (+21.0%) | 37.5 ±0.1 (+12.4%) | 24.3 ±0.2 (+20.5%) | <u>69.4</u> ±0.4 (+29.9%) |
| HiResCAM | 40.3 ±0.1 | 20.1 ±0.2 | 35.8 ±0.1 | 22.3 ±0.2 | 60.8 ±0.3 |
| **Libra HiResCAM** | 41.5 ±0.1 (+2.8%) | 21.2 ±0.2 (+5.7%) | 37.2 ±0.1 (+4.1%) | 23.6 ±0.2 (+5.9%) | 69.0 ±0.3 (+13.4%) |
| XGradCAM+ | 35.6 ±0.1 | 16.0 ±0.2 | 30.6 ±0.1 | 17.9 ±0.2 | 49.0 ±0.4 |
| **Libra XGradCAM+** | 49.5 ±0.1 (+39.0%) | 28.6 ±0.3 (+78.5%) | 45.6 ±0.1 (+49.2%) | 31.4 ±0.3 (+75.3%) | **71.4** ±0.3 (+45.7%) |
| FullGrad+ | 34.4 ±0.1 | 14.9 ±0.2 | 29.0 ±0.1 | 16.6 ±0.2 | 47.4 ±0.3 |
| **Libra FullGrad+** | **53.4** ±0.1 (+55.5%) | **32.4** ±0.3 (+118.0%) | **50.0** ±0.1 (+72.3%) | **35.5** ±0.3 (+113.7%) | 67.9 ±0.3 (+43.2%) |

Table 78. Comparison of attribution methods and their LibraGrad-enhanced versions on the BEiT2-L model. We report faithfulness metrics using Most-Influential-First Deletion, MIF with ground-truth (GT) and predicted labels, including Accuracy and Area Over Perturbation Curve (AOPC) and Segmentation Average Precision (AP). The results demonstrate that composing existing methods with LibraGrad significantly enhances their performance across all metrics.

| Method | LIF Deletion (GT) | | LIF Deletion (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 74.6 ±0.1 | 94.1 ±0.1 | 81.7 ±0.1 | 93.2 ±0.1 |
| RawAtt | 76.6 ±0.1 | 95.8 ±0.1 | 83.7 ±0.1 | 94.9 ±0.1 |
| Attention Rollout | 69.9 ±0.1 | 89.1 ±0.2 | 75.6 ±0.1 | 87.4 ±0.2 |
| AliLRP | 78.0 ±0.1 | 96.7 ±0.1 | 84.5 ±0.1 | 95.5 ±0.1 |
| AttnLRP | 78.4 ±0.1 | 97.8 ±0.1 | 85.7 ±0.1 | 96.8 ±0.1 |
| DecompX | 77.6 ±0.1 | 97.2 ±0.1 | 84.9 ±0.1 | 96.3 ±0.1 |
| Integrated Gradients | 79.4 ±0.1 | 98.8 ±0.1 | 84.2 ±0.1 | 96.5 ±0.2 |
| Input × Grad | 75.5 ±0.1 | 94.5 ±0.1 | 82.0 ±0.1 | 93.5 ±0.1 |
| **Libra Input × Grad** | 79.2 ±0.1 (+4.9%) | 98.1 ±0.1 (+3.9%) | 85.7 ±0.1 (+4.5%) | 96.9 ±0.1 (+3.6%) |
| AttCAT | **81.8** ±0.1 | **101.1** ±0.1 | **87.5** ±0.0 | **100.0** ±0.1 |
| **Libra AttCAT** | <u>80.8</u> ±0.1 (-1.2%) | 99.2 ±0.1 (-1.8%) | <u>87.0</u> ±0.1 (-0.6%) | 97.9 ±0.1 (-2.0%) |
| GenAtt | 75.6 ±0.1 | 95.2 ±0.1 | 83.2 ±0.1 | 94.4 ±0.1 |
| **Libra GenAtt** | 75.5 ±0.1 (-0.2%) | 95.2 ±0.2 (+0.0%) | 83.2 ±0.1 (+0.0%) | 94.3 ±0.2 (-0.1%) |
| TokenTM | 76.8 ±0.1 | 96.2 ±0.1 | 84.6 ±0.1 | 95.5 ±0.1 |
| **Libra TokenTM** | 76.2 ±0.1 (-0.8%) | 95.5 ±0.1 (-0.8%) | 83.8 ±0.1 (-1.0%) | 94.6 ±0.1 (-0.9%) |
| GradCAM+ | 79.2 ±0.1 | 98.5 ±0.2 | 85.1 ±0.1 | 97.1 ±0.1 |
| **Libra GradCAM+** | 78.4 ±0.1 (-1.1%) | 97.1 ±0.1 (-1.3%) | 84.2 ±0.1 (-0.9%) | 95.6 ±0.1 (-1.5%) |
| HiResCAM | 79.4 ±0.1 | 98.3 ±0.1 | 85.5 ±0.1 | 97.0 ±0.1 |
| **Libra HiResCAM** | 80.0 ±0.1 (+0.8%) | 98.4 ±0.1 (+0.2%) | 86.0 ±0.1 (+0.6%) | 97.1 ±0.1 (+0.1%) |
| XGradCAM+ | 78.9 ±0.1 | 97.9 ±0.2 | 84.3 ±0.1 | 96.4 ±0.1 |
| **Libra XGradCAM+** | 79.5 ±0.1 (+0.7%) | 98.0 ±0.1 (+0.1%) | 85.6 ±0.1 (+1.6%) | 96.6 ±0.1 (+0.2%) |
| FullGrad+ | 79.9 ±0.1 | 98.9 ±0.1 | 86.0 ±0.1 | <u>98.0</u> ±0.1 |
| **Libra FullGrad+** | <u>80.8</u> ±0.1 (+1.2%) | <u>99.3</u> ±0.1 (+0.3%) | 86.9 ±0.1 (+1.0%) | <u>98.0</u> ±0.1 (+0.0%) |

Table 79. Comparison of attribution methods and their LibraGrad-enhanced versions on the BEiT2-L model.

| Method | SRG (GT) | | SRG (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 49.8 ±0.1 | 49.8 ±0.1 | 50.0 ±0.1 | 50.0 ±0.1 |
| RawAtt | 55.4 ±0.1 | 55.6 ±0.2 | 56.6 ±0.1 | 56.2 ±0.2 |
| Attention Rollout | 47.9 ±0.1 | 48.1 ±0.2 | 47.7 ±0.1 | 48.0 ±0.2 |
| AliLRP | 55.0 ±0.1 | 54.6 ±0.2 | 55.3 ±0.1 | 54.7 ±0.1 |
| AttnLRP | 60.3 ±0.1 | 60.2 ±0.2 | 61.7 ±0.1 | 60.9 ±0.2 |
| DecompX | 57.0 ±0.1 | 57.3 ±0.2 | 58.3 ±0.1 | 57.8 ±0.2 |
| Integrated Gradients | 55.6 ±0.1 | 55.7 ±0.2 | 53.7 ±0.1 | 54.2 ±0.2 |
| Input × Grad | 51.9 ±0.1 | 51.7 ±0.1 | 51.9 ±0.1 | 51.9 ±0.1 |
| **Libra Input × Grad** | 58.4 ±0.1 (+12.7%) | 58.1 ±0.2 (+12.3%) | 59.3 ±0.1 (+14.4%) | 58.5 ±0.2 (+12.8%) |
| AttCAT | 60.1 ±0.1 | 60.0 ±0.2 | 60.7 ±0.1 | 60.5 ±0.1 |
| **Libra AttCAT** | <u>66.6</u> ±0.1 (+10.9%) | <u>65.4</u> ±0.2 (+9.0%) | <u>67.9</u> ±0.1 (+12.0%) | <u>66.3</u> ±0.2 (+9.6%) |
| GenAtt | 55.6 ±0.1 | 56.1 ±0.2 | 57.0 ±0.1 | 56.8 ±0.2 |
| **Libra GenAtt** | 56.6 ±0.1 (+1.7%) | 56.8 ±0.2 (+1.3%) | 58.1 ±0.1 (+1.9%) | 57.5 ±0.2 (+1.3%) |
| TokenTM | 60.3 ±0.1 | 60.3 ±0.2 | 62.1 ±0.1 | 61.2 ±0.2 |
| **Libra TokenTM** | 59.4 ±0.1 (-1.6%) | 59.3 ±0.3 (-1.6%) | 61.0 ±0.1 (-1.7%) | 60.0 ±0.2 (-1.8%) |
| GradCAM+ | 58.8 ±0.1 | 58.3 ±0.2 | 59.2 ±0.1 | 58.6 ±0.2 |
| **Libra GradCAM+** | 60.3 ±0.1 (+2.6%) | 59.6 ±0.2 (+2.1%) | 60.9 ±0.1 (+2.8%) | 59.9 ±0.2 (+2.3%) |
| HiResCAM | 59.9 ±0.1 | 59.2 ±0.2 | 60.6 ±0.1 | 59.6 ±0.2 |
| **Libra HiResCAM** | 60.8 ±0.1 (+1.5%) | 59.8 ±0.2 (+1.1%) | 61.6 ±0.1 (+1.6%) | 60.4 ±0.2 (+1.2%) |
| XGradCAM+ | 57.3 ±0.1 | 56.9 ±0.2 | 57.4 ±0.1 | 57.2 ±0.1 |
| **Libra XGradCAM+** | 64.5 ±0.1 (+12.6%) | 63.3 ±0.2 (+11.1%) | 65.6 ±0.1 (+14.2%) | 64.0 ±0.2 (+12.0%) |
| FullGrad+ | 57.1 ±0.1 | 56.9 ±0.2 | 57.5 ±0.1 | 57.3 ±0.2 |
| **Libra FullGrad+** | **67.1** ±0.1 (+17.5%) | **65.8** ±0.2 (+15.7%) | **68.5** ±0.1 (+19.0%) | **66.8** ±0.2 (+16.5%) |

Table 80. Comparison of attribution methods and their LibraGrad-enhanced versions on the BEiT2-L model.

### D.5.12. SigLIP-L

Since SigLIP does not have a CLS token, certain attribution methods couldn't be applied and were omitted.

| Method | MIF Deletion (GT) Accuracy | MIF Deletion (GT) AOPC | MIF Deletion (Predicted) Accuracy | MIF Deletion (Predicted) AOPC | Segmentation AP |
|---|---|---|---|---|---|
| Random | 39.0 ±0.1 | 17.3 ±0.2 | 32.8 ±0.1 | 19.1 ±0.2 | 33.0 ±0.3 |
| AliLRP | 58.8 ±0.1 | 36.6 ±0.3 | 55.4 ±0.1 | 40.0 ±0.3 | 33.5 ±0.3 |
| AttnLRP | 64.7 ±0.1 | 42.4 ±0.3 | 62.2 ±0.1 | 46.2 ±0.3 | 36.0 ±0.3 |
| DecompX | 54.5 ±0.1 | 32.6 ±0.2 | 51.1 ±0.1 | 35.7 ±0.2 | 40.5 ±0.3 |
| Integrated Gradients | 52.7 ±0.1 | 30.0 ±0.2 | 44.0 ±0.1 | 28.8 ±0.2 | 41.6 ±0.3 |
| Input × Grad | 44.4 ±0.1 | 23.2 ±0.2 | 40.8 ±0.1 | 26.0 ±0.2 | 35.5 ±0.3 |
| **Libra Input × Grad** | 54.7 ±0.1 (+23.4%) | 32.4 ±0.2 (+40.0%) | 51.1 ±0.1 (+25.4%) | 35.6 ±0.2 (+36.9%) | 39.9 ±0.3 (+12.3%) |
| AttCAT | 48.3 ±0.1 | 27.4 ±0.3 | 45.9 ±0.1 | 30.9 ±0.2 | 37.6 ±0.3 |
| **Libra AttCAT** | **79.0** ±0.1 (+63.4%) | **55.0** ±0.3 (+100.5%) | **77.4** ±0.1 (+68.6%) | **59.7** ±0.2 (+93.1%) | 46.8 ±0.3 (+24.2%) |
| GradCAM+ | 47.6 ±0.1 | 25.4 ±0.3 | 43.5 ±0.1 | 28.1 ±0.2 | 44.3 ±0.4 |
| **Libra GradCAM+** | 51.0 ±0.1 (+7.2%) | 28.8 ±0.3 (+13.6%) | 47.4 ±0.1 (+9.0%) | 31.9 ±0.3 (+13.5%) | 41.7 ±0.3 (-5.7%) |
| HiResCAM | 37.1 ±0.1 | 15.7 ±0.2 | 31.4 ±0.1 | 17.5 ±0.2 | 36.3 ±0.3 |
| **Libra HiResCAM** | 50.0 ±0.1 (+34.8%) | 27.5 ±0.3 (+75.7%) | 46.1 ±0.1 (+46.7%) | 30.4 ±0.2 (+73.7%) | <u>47.5</u> ±0.3 (+30.8%) |
| XGradCAM+ | 54.8 ±0.1 | 34.5 ±0.3 | 51.4 ±0.1 | 37.8 ±0.2 | 43.0 ±0.3 |
| **Libra XGradCAM+** | 66.3 ±0.1 (+21.0%) | 42.5 ±0.3 (+23.2%) | 63.6 ±0.1 (+23.7%) | 46.3 ±0.3 (+22.6%) | 44.3 ±0.4 (+3.1%) |
| FullGrad+ | 46.6 ±0.1 | 25.8 ±0.3 | 43.6 ±0.1 | 29.0 ±0.2 | 37.7 ±0.3 |
| **Libra FullGrad+** | <u>75.3</u> ±0.1 (+61.7%) | <u>50.7</u> ±0.3 (+96.6%) | <u>73.5</u> ±0.1 (+68.5%) | <u>55.1</u> ±0.2 (+89.7%) | **51.7** ±0.3 (+37.1%) |

Table 81. Comparison of attribution methods and their LibraGrad-enhanced versions on the SigLIP-L model. We report faithfulness metrics using Most-Influential-First Deletion, MIF with ground-truth (GT) and predicted labels, including Accuracy and Area Over Perturbation Curve (AOPC) and Segmentation Average Precision (AP). The results demonstrate that composing existing methods with LibraGrad significantly enhances their performance across all metrics.

| Method | LIF Deletion (GT) | | LIF Deletion (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 61.1 ±0.1 | 82.7 ±0.2 | 67.1 ±0.1 | 81.0 ±0.1 |
| AliLRP | 70.8 ±0.1 | 91.2 ±0.2 | 77.0 ±0.1 | 89.8 ±0.2 |
| AttnLRP | 75.0 ±0.1 | 96.0 ±0.2 | 82.2 ±0.1 | 95.0 ±0.1 |
| DecompX | 71.3 ±0.1 | 91.8 ±0.2 | 78.1 ±0.1 | 90.5 ±0.2 |
| Integrated Gradients | 75.9 ±0.1 | 97.0 ±0.3 | 75.6 ±0.1 | 91.2 ±0.2 |
| Input $\times$ Grad | 67.4 ±0.1 | 89.7 ±0.3 | 71.6 ±0.1 | 87.6 ±0.2 |
| **Libra Input $\times$ Grad** | 71.8 ±0.1 (+6.5%) | 91.9 ±0.3 (+2.4%) | 78.3 ±0.1 (+9.4%) | 90.6 ±0.2 (+3.5%) |
| AttCAT | 73.8 ±0.1 | 95.3 ±0.3 | 76.6 ±0.1 | 92.4 ±0.2 |
| **Libra AttCAT** | **80.0** ±0.1 (+8.4%) | **99.6** ±0.2 (+4.5%) | **85.9** ±0.1 (+12.2%) | **98.4** ±0.1 (+6.4%) |
| GradCAM+ | 45.8 ±0.1 | 64.3 ±0.4 | 49.0 ±0.1 | 60.7 ±0.3 |
| **Libra GradCAM+** | 62.8 ±0.1 (+37.1%) | 83.0 ±0.3 (+29.0%) | 67.5 ±0.1 (+37.8%) | 80.8 ±0.2 (+33.1%) |
| HiResCAM | 48.1 ±0.1 | 69.4 ±0.4 | 51.9 ±0.1 | 66.3 ±0.3 |
| **Libra HiResCAM** | 63.7 ±0.1 (+32.2%) | 84.8 ±0.3 (+22.1%) | 68.2 ±0.1 (+31.5%) | 82.6 ±0.2 (+24.6%) |
| XGradCAM+ | 57.3 ±0.1 | 78.4 ±0.4 | 60.6 ±0.1 | 75.2 ±0.3 |
| **Libra XGradCAM+** | 70.5 ±0.1 (+23.2%) | 89.8 ±0.3 (+14.6%) | 76.4 ±0.1 (+26.1%) | 88.4 ±0.2 (+17.5%) |
| FullGrad+ | 70.4 ±0.1 | 92.2 ±0.3 | 73.3 ±0.1 | 89.3 ±0.2 |
| **Libra FullGrad+** | <u>79.8</u> ±0.1 (+13.4%) | <u>99.4</u> ±0.2 (+7.8%) | <u>85.8</u> ±0.1 (+17.0%) | <u>98.2</u> ±0.1 (+10.0%) |

Table 82. Comparison of attribution methods and their LibraGrad-enhanced versions on the SigLIP-L model.

| Method | SRG (GT) | | SRG (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 50.0 ±0.1 | 50.0 ±0.2 | 50.0 ±0.1 | 50.0 ±0.2 |
| AliLRP | 64.8 ±0.1 | 63.9 ±0.3 | 66.2 ±0.1 | 64.9 ±0.2 |
| AttnLRP | 69.8 ±0.1 | 69.2 ±0.3 | 72.2 ±0.1 | 70.6 ±0.2 |
| DecompX | 62.9 ±0.1 | 62.2 ±0.2 | 64.6 ±0.1 | 63.1 ±0.2 |
| Integrated Gradients | 64.3 ±0.1 | 63.5 ±0.3 | 59.8 ±0.1 | 60.0 ±0.2 |
| Input $\times$ Grad | 55.9 ±0.1 | 56.4 ±0.3 | 56.2 ±0.1 | 56.8 ±0.2 |
| **Libra Input $\times$ Grad** | 63.3 ±0.1 (+13.2%) | 62.2 ±0.3 (+10.1%) | 64.7 ±0.1 (+15.2%) | 63.1 ±0.2 (+11.1%) |
| AttCAT | 61.0 ±0.1 | 61.4 ±0.3 | 61.2 ±0.1 | 61.7 ±0.2 |
| **Libra AttCAT** | **79.5** ±0.1 (+30.2%) | **77.3** ±0.3 (+26.0%) | **81.6** ±0.1 (+33.3%) | **79.0** ±0.2 (+28.2%) |
| GradCAM+ | 46.7 ±0.1 | 44.9 ±0.3 | 46.2 ±0.1 | 44.4 ±0.3 |
| **Libra GradCAM+** | 56.9 ±0.1 (+21.9%) | 55.9 ±0.3 (+24.6%) | 57.4 ±0.1 (+24.2%) | 56.4 ±0.3 (+26.9%) |
| HiResCAM | 42.6 ±0.1 | 42.5 ±0.3 | 41.7 ±0.1 | 41.9 ±0.2 |
| **Libra HiResCAM** | 56.8 ±0.1 (+33.4%) | 56.1 ±0.3 (+32.0%) | 57.2 ±0.1 (+37.2%) | 56.5 ±0.2 (+34.9%) |
| XGradCAM+ | 56.0 ±0.1 | 56.4 ±0.3 | 56.0 ±0.1 | 56.5 ±0.2 |
| **Libra XGradCAM+** | 68.4 ±0.1 (+22.1%) | 66.2 ±0.3 (+17.2%) | 70.0 ±0.1 (+25.0%) | 67.3 ±0.2 (+19.2%) |
| FullGrad+ | 58.5 ±0.1 | 59.0 ±0.3 | 58.4 ±0.1 | 59.2 ±0.2 |
| **Libra FullGrad+** | <u>77.6</u> ±0.1 (+32.7%) | <u>75.0</u> ±0.3 (+27.2%) | <u>79.6</u> ±0.1 (+36.2%) | <u>76.7</u> ±0.2 (+29.5%) |

Table 83. Comparison of attribution methods and their LibraGrad-enhanced versions on the SigLIP-L model.

#### D.5.13. CLIP-H

| Method | MIF Deletion (GT) Accuracy | MIF Deletion (GT) AOPC | MIF Deletion (Predicted) Accuracy | MIF Deletion (Predicted) AOPC | Segmentation AP |
|---|---|---|---|---|---|
| Random | 34.3 ±0.1 | 11.2 ±0.2 | 28.0 ±0.1 | 12.7 ±0.2 | 37.8 ±0.3 |
| RawAtt | 46.9 ±0.1 | 21.0 ±0.2 | 42.5 ±0.1 | 23.3 ±0.2 | 41.6 ±0.3 |
| Attention Rollout | 46.4 ±0.1 | 20.5 ±0.3 | 41.3 ±0.1 | 22.5 ±0.3 | 51.7 ±0.4 |
| AliLRP | 40.0 ±0.1 | 15.7 ±0.2 | 34.4 ±0.1 | 17.3 ±0.2 | 38.1 ±0.3 |
| AttnLRP | 50.8 ±0.1 | 24.0 ±0.3 | 46.7 ±0.1 | 26.4 ±0.2 | 50.9 ±0.3 |
| DecompX | 46.7 ±0.1 | 21.3 ±0.2 | 42.4 ±0.1 | 23.5 ±0.2 | 55.0 ±0.3 |
| Integrated Gradients | 37.1 ±0.1 | 13.5 ±0.2 | 31.0 ±0.1 | 15.0 ±0.2 | 36.9 ±0.3 |
| Input × Grad | 37.5 ±0.1 | 13.7 ±0.2 | 31.4 ±0.1 | 15.2 ±0.2 | 36.8 ±0.3 |
| **Libra Input × Grad** | 47.5 ±0.1 (+26.8%) | 21.8 ±0.2 (+59.4%) | 43.1 ±0.1 (+37.3%) | 24.0 ±0.2 (+57.9%) | 54.2 ±0.3 (+47.3%) |
| AttCAT | 42.5 ±0.1 | 18.8 ±0.2 | 39.0 ±0.1 | 21.3 ±0.1 | 38.9 ±0.3 |
| **Libra AttCAT** | <u>61.5</u> ±0.1 (+44.8%) | <u>31.7</u> ±0.3 (+68.9%) | <u>58.5</u> ±0.1 (+49.8%) | <u>34.7</u> ±0.2 (+62.8%) | 61.7 ±0.3 (+58.6%) |
| GenAtt | 54.4 ±0.1 | 26.8 ±0.2 | 51.0 ±0.1 | 29.6 ±0.2 | 55.9 ±0.3 |
| **Libra GenAtt** | 61.0 ±0.1 (+12.2%) | 31.5 ±0.3 (+17.5%) | 58.1 ±0.1 (+14.0%) | 34.5 ±0.2 (+16.7%) | **76.2** ±0.2 (+36.1%) |
| TokenTM | 55.4 ±0.1 | 27.4 ±0.3 | 51.9 ±0.1 | 30.1 ±0.2 | 58.6 ±0.3 |
| **Libra TokenTM** | 60.6 ±0.1 (+9.3%) | 31.2 ±0.3 (+14.0%) | 57.4 ±0.1 (+10.6%) | 34.1 ±0.2 (+13.5%) | 71.5 ±0.3 (+22.1%) |
| GradCAM+ | 38.6 ±0.1 | 14.5 ±0.2 | 33.0 ±0.1 | 16.2 ±0.2 | 43.0 ±0.4 |
| **Libra GradCAM+** | 41.8 ±0.1 (+8.4%) | 16.8 ±0.2 (+15.6%) | 36.2 ±0.1 (+9.8%) | 18.6 ±0.2 (+14.9%) | 47.4 ±0.4 (+10.2%) |
| HiResCAM | 42.3 ±0.1 | 17.6 ±0.2 | 37.6 ±0.1 | 19.7 ±0.2 | 45.9 ±0.3 |
| **Libra HiResCAM** | 52.8 ±0.1 (+24.8%) | 25.4 ±0.2 (+44.3%) | 48.9 ±0.1 (+29.9%) | 27.9 ±0.2 (+41.9%) | 56.8 ±0.3 (+23.7%) |
| XGradCAM+ | 44.2 ±0.1 | 19.2 ±0.2 | 39.4 ±0.1 | 21.3 ±0.2 | 47.7 ±0.4 |
| **Libra XGradCAM+** | 60.8 ±0.1 (+37.4%) | 31.1 ±0.2 (+62.0%) | 57.7 ±0.1 (+46.4%) | 34.1 ±0.2 (+59.7%) | <u>73.3</u> ±0.3 (+53.8%) |
| FullGrad+ | 41.4 ±0.1 | 18.1 ±0.2 | 37.6 ±0.1 | 20.5 ±0.2 | 38.5 ±0.3 |
| **Libra FullGrad+** | **63.8** ±0.1 (+54.3%) | **33.6** ±0.3 (+85.9%) | **61.1** ±0.1 (+62.3%) | **36.8** ±0.2 (+79.1%) | 71.5 ±0.3 (+85.7%) |

Table 84. Comparison of attribution methods and their LibraGrad-enhanced versions on the CLIP-H model. We report faithfulness metrics using Most-Influential-First Deletion, MIF with ground-truth (GT) and predicted labels, including Accuracy and Area Over Perturbation Curve (AOPC) and Segmentation Average Precision (AP). The results demonstrate that composing existing methods with LibraGrad significantly enhances their performance across all metrics.

| Method | LIF Deletion (GT) | | LIF Deletion (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 65.8 ±0.1 | 88.8 ±0.2 | 72.4 ±0.1 | 87.5 ±0.2 |
| RawAtt | 68.7 ±0.1 | 91.1 ±0.1 | 76.0 ±0.1 | 90.0 ±0.2 |
| Attention Rollout | 68.1 ±0.1 | 90.7 ±0.2 | 74.6 ±0.1 | 89.4 ±0.2 |
| AliLRP | 69.1 ±0.1 | 91.3 ±0.1 | 75.3 ±0.1 | 89.9 ±0.1 |
| AttnLRP | 76.8 ±0.1 | 97.3 ±0.2 | 83.3 ±0.1 | 96.1 ±0.1 |
| DecompX | 74.8 ±0.1 | 95.4 ±0.2 | 81.7 ±0.1 | 94.2 ±0.2 |
| Integrated Gradients | 63.3 ±0.1 | 87.2 ±0.1 | 69.4 ±0.1 | 85.7 ±0.1 |
| Input × Grad | 62.7 ±0.1 | 86.5 ±0.2 | 68.8 ±0.1 | 84.9 ±0.1 |
| **Libra Input × Grad** | 76.0 ±0.1 (+21.2%) | 96.3 ±0.2 (+11.3%) | 82.2 ±0.1 (+19.4%) | 94.7 ±0.2 (+11.5%) |
| AttCAT | 72.3 ±0.1 | 96.3 ±0.2 | 76.9 ±0.1 | 94.8 ±0.2 |
| **Libra AttCAT** | <u>78.1</u> ±0.1 (+7.9%) | <u>98.1</u> ±0.1 (+1.8%) | <u>83.8</u> ±0.1 (+9.0%) | <u>96.4</u> ±0.1 (+1.6%) |
| GenAtt | 73.4 ±0.1 | 95.3 ±0.2 | 80.8 ±0.1 | 94.3 ±0.2 |
| **Libra GenAtt** | 75.0 ±0.1 (+2.2%) | 95.5 ±0.1 (+0.3%) | 82.5 ±0.1 (+2.0%) | 94.5 ±0.1 (+0.2%) |
| TokenTM | 73.1 ±0.1 | 94.5 ±0.1 | 80.6 ±0.1 | 93.4 ±0.1 |
| **Libra TokenTM** | 74.1 ±0.1 (+1.3%) | 94.6 ±0.1 (+0.2%) | 81.7 ±0.1 (+1.4%) | 93.6 ±0.1 (+0.2%) |
| GradCAM+ | 63.8 ±0.1 | 87.4 ±0.2 | 69.4 ±0.1 | 85.6 ±0.2 |
| **Libra GradCAM+** | 65.6 ±0.1 (+2.9%) | 88.4 ±0.3 (+1.1%) | 70.7 ±0.1 (+1.9%) | 86.4 ±0.2 (+0.9%) |
| HiResCAM | 72.4 ±0.1 | 94.3 ±0.2 | 77.9 ±0.1 | 92.7 ±0.2 |
| **Libra HiResCAM** | 74.7 ±0.1 (+3.3%) | 95.6 ±0.1 (+1.4%) | 80.9 ±0.1 (+3.9%) | 94.1 ±0.1 (+1.5%) |
| XGradCAM+ | 69.7 ±0.1 | 92.1 ±0.2 | 75.4 ±0.1 | 90.6 ±0.2 |
| **Libra XGradCAM+** | 75.5 ±0.1 (+8.5%) | 95.8 ±0.1 (+3.9%) | 81.0 ±0.1 (+7.4%) | 93.9 ±0.2 (+3.7%) |
| FullGrad+ | 71.4 ±0.1 | 95.0 ±0.2 | 76.2 ±0.1 | 93.4 ±0.2 |
| **Libra FullGrad+** | **79.1** ±0.1 (+10.9%) | **98.4** ±0.1 (+3.6%) | **84.9** ±0.1 (+11.4%) | **96.8** ±0.2 (+3.6%) |

Table 85. Comparison of attribution methods and their LibraGrad-enhanced versions on the CLIP-H model.

| Method | SRG (GT) | | SRG (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 50.0 ±0.1 | 50.0 ±0.2 | 50.2 ±0.1 | 50.1 ±0.2 |
| RawAtt | 57.8 ±0.1 | 56.1 ±0.2 | 59.2 ±0.1 | 56.7 ±0.2 |
| Attention Rollout | 57.2 ±0.1 | 55.6 ±0.3 | 58.0 ±0.1 | 55.9 ±0.2 |
| AliLRP | 54.5 ±0.1 | 53.5 ±0.2 | 54.8 ±0.1 | 53.6 ±0.2 |
| AttnLRP | 63.8 ±0.1 | 60.6 ±0.2 | 65.0 ±0.1 | 61.3 ±0.2 |
| DecompX | 60.8 ±0.1 | 58.3 ±0.2 | 62.1 ±0.1 | 58.9 ±0.2 |
| Integrated Gradients | 50.2 ±0.1 | 50.3 ±0.2 | 50.2 ±0.1 | 50.3 ±0.1 |
| Input × Grad | 50.1 ±0.1 | 50.1 ±0.2 | 50.1 ±0.1 | 50.1 ±0.1 |
| **Libra Input × Grad** | 61.7 ±0.1 (+23.3%) | 59.0 ±0.2 (+17.8%) | 62.6 ±0.1 (+25.0%) | 59.4 ±0.2 (+18.6%) |
| AttCAT | 57.4 ±0.1 | 57.5 ±0.2 | 58.0 ±0.1 | 58.1 ±0.2 |
| **Libra AttCAT** | <u>69.8</u> ±0.1 (+21.6%) | <u>64.9</u> ±0.2 (+12.8%) | <u>71.2</u> ±0.1 (+22.7%) | <u>65.5</u> ±0.2 (+12.9%) |
| GenAtt | 63.9 ±0.1 | 61.1 ±0.2 | 65.9 ±0.1 | 62.0 ±0.2 |
| **Libra GenAtt** | 68.0 ±0.1 (+6.5%) | 63.5 ±0.2 (+4.1%) | 70.3 ±0.1 (+6.7%) | 64.5 ±0.2 (+4.1%) |
| TokenTM | 64.3 ±0.1 | 60.9 ±0.2 | 66.2 ±0.1 | 61.8 ±0.2 |
| **Libra TokenTM** | 67.3 ±0.1 (+4.7%) | 62.9 ±0.2 (+3.3%) | 69.5 ±0.1 (+5.0%) | 63.9 ±0.2 (+3.5%) |
| GradCAM+ | 51.2 ±0.1 | 51.0 ±0.2 | 51.2 ±0.1 | 50.9 ±0.2 |
| **Libra GradCAM+** | 53.7 ±0.1 (+4.9%) | 52.6 ±0.2 (+3.2%) | 53.5 ±0.1 (+4.5%) | 52.5 ±0.2 (+3.1%) |
| HiResCAM | 57.3 ±0.1 | 55.9 ±0.2 | 57.8 ±0.1 | 56.2 ±0.2 |
| **Libra HiResCAM** | 63.8 ±0.1 (+11.2%) | 60.5 ±0.2 (+8.1%) | 64.9 ±0.1 (+12.3%) | 61.0 ±0.2 (+8.6%) |
| XGradCAM+ | 56.9 ±0.1 | 55.7 ±0.2 | 57.4 ±0.1 | 56.0 ±0.2 |
| **Libra XGradCAM+** | 68.2 ±0.1 (+19.7%) | 63.4 ±0.2 (+13.9%) | 69.3 ±0.1 (+20.8%) | 64.0 ±0.2 (+14.4%) |
| FullGrad+ | 56.4 ±0.1 | 56.5 ±0.2 | 56.9 ±0.1 | 57.0 ±0.2 |
| **Libra FullGrad+** | **71.5** ±0.1 (+26.8%) | **66.0** ±0.2 (+16.7%) | **73.0** ±0.1 (+28.2%) | **66.8** ±0.2 (+17.2%) |

Table 86. Comparison of attribution methods and their LibraGrad-enhanced versions on the CLIP-H model.

### D.5.14. DeiT3-H

| Method | MIF Deletion (GT) Accuracy | MIF Deletion (GT) AOPC | MIF Deletion (Predicted) Accuracy | MIF Deletion (Predicted) AOPC | Segmentation AP |
|---|---|---|---|---|---|
| Random | 35.6 ±0.1 | 16.6 ±0.2 | 29.0 ±0.1 | 19.2 ±0.2 | 37.8 ±0.3 |
| RawAtt | 56.1 ±0.1 | 33.3 ±0.3 | 52.0 ±0.1 | 37.2 ±0.2 | 49.7 ±0.3 |
| Attention Rollout | 37.1 ±0.1 | 19.0 ±0.2 | 31.2 ±0.1 | 21.9 ±0.2 | 34.1 ±0.3 |
| AliLRP | 59.6 ±0.1 | 37.3 ±0.3 | 56.3 ±0.1 | 41.7 ±0.2 | 52.2 ±0.3 |
| AttnLRP | 45.4 ±0.1 | 28.1 ±0.3 | 40.7 ±0.1 | 31.7 ±0.2 | 36.0 ±0.3 |
| DecompX | 51.6 ±0.1 | 32.2 ±0.3 | 47.2 ±0.1 | 35.9 ±0.2 | 49.5 ±0.3 |
| Integrated Gradients | 43.7 ±0.1 | 24.9 ±0.3 | 33.2 ±0.1 | 22.8 ±0.2 | 38.9 ±0.3 |
| Input × Grad | 40.4 ±0.1 | 21.9 ±0.3 | 35.1 ±0.1 | 25.1 ±0.2 | 39.6 ±0.3 |
| **Libra Input × Grad** | 52.1 ±0.1 (+29.2%) | 32.4 ±0.3 (+48.1%) | 47.7 ±0.1 (+36.0%) | 36.3 ±0.2 (+44.7%) | 49.0 ±0.3 (+23.8%) |
| AttCAT | 48.2 ±0.1 | 28.6 ±0.3 | 44.0 ±0.1 | 32.3 ±0.3 | 41.7 ±0.3 |
| **Libra AttCAT** | <u>72.6</u> ±0.1 (+50.6%) | <u>47.6</u> ±0.3 (+66.5%) | <u>70.5</u> ±0.1 (+60.2%) | <u>52.8</u> ±0.2 (+63.4%) | 60.1 ±0.3 (+44.1%) |
| GenAtt | 67.2 ±0.1 | 43.3 ±0.3 | 64.6 ±0.1 | 48.1 ±0.2 | 66.2 ±0.2 |
| **Libra GenAtt** | 69.1 ±0.1 (+2.9%) | 45.0 ±0.3 (+3.8%) | 66.5 ±0.1 (+3.0%) | 49.7 ±0.2 (+3.5%) | **76.5** ±0.2 (+15.5%) |
| TokenTM | 66.2 ±0.1 | 42.6 ±0.3 | 63.3 ±0.1 | 47.2 ±0.2 | 61.7 ±0.2 |
| **Libra TokenTM** | 68.1 ±0.1 (+2.8%) | 44.1 ±0.3 (+3.6%) | 65.2 ±0.1 (+3.0%) | 48.8 ±0.2 (+3.3%) | 70.8 ±0.2 (+14.7%) |
| GradCAM+ | 49.5 ±0.1 | 28.3 ±0.3 | 44.5 ±0.1 | 31.8 ±0.2 | 60.3 ±0.4 |
| **Libra GradCAM+** | 52.6 ±0.1 (+6.2%) | 31.4 ±0.3 (+10.9%) | 48.7 ±0.1 (+9.6%) | 35.5 ±0.2 (+11.7%) | 46.7 ±0.4 (-22.5%) |
| HiResCAM | 32.5 ±0.1 | 15.0 ±0.2 | 25.8 ±0.1 | 17.4 ±0.2 | 41.3 ±0.3 |
| **Libra HiResCAM** | 57.4 ±0.1 (+76.7%) | 35.4 ±0.3 (+136.8%) | 53.8 ±0.1 (+108.5%) | 39.7 ±0.2 (+127.5%) | <u>76.3</u> ±0.3 (+84.9%) |
| XGradCAM+ | 49.1 ±0.1 | 27.9 ±0.3 | 45.1 ±0.1 | 31.8 ±0.2 | 48.9 ±0.4 |
| **Libra XGradCAM+** | 68.8 ±0.1 (+40.2%) | 44.2 ±0.3 (+58.3%) | 66.1 ±0.1 (+46.7%) | 49.0 ±0.2 (+54.2%) | 59.4 ±0.3 (+21.5%) |
| FullGrad+ | 45.8 ±0.1 | 26.2 ±0.3 | 41.9 ±0.1 | 30.0 ±0.3 | 40.6 ±0.3 |
| **Libra FullGrad+** | **73.5** ±0.1 (+60.4%) | **48.5** ±0.3 (+84.8%) | **71.5** ±0.1 (+70.7%) | **53.7** ±0.2 (+78.8%) | 65.1 ±0.3 (+60.4%) |

Table 87. Comparison of attribution methods and their LibraGrad-enhanced versions on the DeiT3-H model. We report faithfulness metrics using Most-Influential-First Deletion, MIF with ground-truth (GT) and predicted labels, including Accuracy and Area Over Perturbation Curve (AOPC) and Segmentation Average Precision (AP). The results demonstrate that composing existing methods with LibraGrad significantly enhances their performance across all metrics.

| Method | LIF Deletion (GT) | | LIF Deletion (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 64.2 ±0.1 | 83.5 ±0.2 | 70.7 ±0.1 | 81.1 ±0.1 |
| RawAtt | 70.9 ±0.1 | 86.3 ±0.2 | 78.4 ±0.1 | 84.3 ±0.2 |
| Attention Rollout | 59.0 ±0.1 | 77.9 ±0.3 | 64.5 ±0.1 | 74.8 ±0.2 |
| AliLRP | 79.5 ±0.1 | 97.5 ±0.2 | 86.1 ±0.1 | 96.0 ±0.1 |
| AttnLRP | 74.1 ±0.1 | 94.2 ±0.2 | 80.8 ±0.1 | 92.2 ±0.2 |
| DecompX | 75.8 ±0.1 | 95.2 ±0.2 | 83.1 ±0.1 | 93.4 ±0.1 |
| Integrated Gradients | 71.5 ±0.1 | 91.2 ±0.3 | 74.6 ±0.1 | 84.9 ±0.2 |
| Input × Grad | 71.9 ±0.1 | 90.5 ±0.2 | 77.7 ±0.1 | 87.8 ±0.2 |
| **Libra Input × Grad** | 77.1 ±0.1 (+7.2%) | 95.8 ±0.2 (+5.9%) | 83.7 ±0.1 (+7.7%) | 94.0 ±0.2 (+7.1%) |
| AttCAT | 75.3 ±0.1 | 93.6 ±0.2 | 80.5 ±0.1 | 90.6 ±0.2 |
| **Libra AttCAT** | <u>81.7</u> ±0.1 (+8.4%) | <u>100.0</u> ±0.2 (+6.9%) | **87.7** ±0.0 (+8.9%) | <u>98.6</u> ±0.1 (+8.8%) |
| GenAtt | 76.9 ±0.1 | 94.9 ±0.2 | 85.7 ±0.1 | 93.6 ±0.1 |
| **Libra GenAtt** | 77.3 ±0.1 (+0.5%) | 95.3 ±0.2 (+0.5%) | 86.0 ±0.1 (+0.3%) | 94.0 ±0.1 (+0.5%) |
| TokenTM | 76.2 ±0.1 | 94.2 ±0.2 | 85.0 ±0.1 | 93.0 ±0.1 |
| **Libra TokenTM** | 76.4 ±0.1 (+0.4%) | 94.6 ±0.2 (+0.4%) | 85.4 ±0.1 (+0.4%) | 93.3 ±0.2 (+0.4%) |
| GradCAM+ | 69.3 ±0.1 | 86.7 ±0.2 | 75.8 ±0.1 | 84.4 ±0.2 |
| **Libra GradCAM+** | 74.1 ±0.1 (+6.9%) | 91.9 ±0.2 (+6.0%) | 80.7 ±0.1 (+6.4%) | 89.8 ±0.2 (+6.4%) |
| HiResCAM | 68.1 ±0.1 | 86.2 ±0.2 | 75.5 ±0.1 | 84.2 ±0.1 |
| **Libra HiResCAM** | 75.2 ±0.1 (+10.4%) | 89.8 ±0.3 (+4.1%) | 80.6 ±0.1 (+6.8%) | 86.9 ±0.2 (+3.3%) |
| XGradCAM+ | 71.4 ±0.1 | 89.9 ±0.3 | 77.1 ±0.1 | 87.3 ±0.2 |
| **Libra XGradCAM+** | 79.6 ±0.1 (+11.5%) | 97.0 ±0.2 (+7.9%) | 86.4 ±0.1 (+12.0%) | 95.3 ±0.1 (+9.1%) |
| FullGrad+ | 74.6 ±0.1 | 92.8 ±0.2 | 79.9 ±0.1 | 90.1 ±0.2 |
| **Libra FullGrad+** | **81.8** ±0.1 (+9.7%) | **100.4** ±0.2 (+8.2%) | <u>87.6</u> ±0.0 (+9.6%) | **98.8** ±0.1 (+9.7%) |

Table 88. Comparison of attribution methods and their LibraGrad-enhanced versions on the DeiT3-H model.

| Method | SRG (GT) | | SRG (Predicted) | |
|---|---|---|---|---|
| | Accuracy | AOPC | Accuracy | AOPC |
| Random | 49.9 ±0.1 | 50.0 ±0.2 | 49.8 ±0.1 | 50.1 ±0.2 |
| RawAtt | 63.5 ±0.1 | 59.8 ±0.2 | 65.2 ±0.1 | 60.7 ±0.2 |
| Attention Rollout | 48.1 ±0.1 | 48.4 ±0.3 | 47.8 ±0.1 | 48.3 ±0.2 |
| AliLRP | 69.6 ±0.1 | 67.4 ±0.2 | 71.2 ±0.1 | 68.8 ±0.2 |
| AttnLRP | 59.7 ±0.1 | 61.2 ±0.3 | 60.8 ±0.1 | 62.0 ±0.2 |
| DecompX | 63.7 ±0.1 | 63.7 ±0.2 | 65.1 ±0.1 | 64.7 ±0.2 |
| Integrated Gradients | 57.6 ±0.1 | 58.1 ±0.3 | 53.9 ±0.1 | 53.8 ±0.2 |
| Input × Grad | 56.1 ±0.1 | 56.2 ±0.2 | 56.4 ±0.1 | 56.4 ±0.2 |
| **Libra Input × Grad** | 64.6 ±0.1 (+15.1%) | 64.1 ±0.2 (+14.1%) | 65.7 ±0.1 (+16.5%) | 65.1 ±0.2 (+15.4%) |
| AttCAT | 61.8 ±0.1 | 61.1 ±0.3 | 62.3 ±0.1 | 61.5 ±0.2 |
| **Libra AttCAT** | <u>77.1</u> ±0.1 (+24.9%) | <u>73.8</u> ±0.2 (+20.8%) | <u>79.1</u> ±0.1 (+27.0%) | <u>75.7</u> ±0.2 (+23.1%) |
| GenAtt | 72.1 ±0.1 | 69.1 ±0.2 | 75.2 ±0.1 | 70.8 ±0.2 |
| **Libra GenAtt** | 73.2 ±0.1 (+1.6%) | 70.2 ±0.2 (+1.5%) | 76.2 ±0.1 (+1.4%) | 71.9 ±0.2 (+1.5%) |
| TokenTM | 71.2 ±0.1 | 68.4 ±0.2 | 74.2 ±0.1 | 70.1 ±0.2 |
| **Libra TokenTM** | 72.3 ±0.1 (+1.5%) | 69.3 ±0.2 (+1.4%) | 75.3 ±0.1 (+1.6%) | 71.0 ±0.2 (+1.4%) |
| GradCAM+ | 59.4 ±0.1 | 57.5 ±0.2 | 60.1 ±0.1 | 58.1 ±0.2 |
| **Libra GradCAM+** | 63.4 ±0.1 (+6.6%) | 61.6 ±0.2 (+7.2%) | 64.7 ±0.1 (+7.6%) | 62.6 ±0.2 (+7.8%) |
| HiResCAM | 50.3 ±0.1 | 50.6 ±0.2 | 50.7 ±0.1 | 50.8 ±0.2 |
| **Libra HiResCAM** | 66.3 ±0.1 (+31.8%) | 62.6 ±0.3 (+23.7%) | 67.2 ±0.1 (+32.7%) | 63.3 ±0.2 (+24.6%) |
| XGradCAM+ | 60.2 ±0.1 | 58.9 ±0.3 | 61.1 ±0.1 | 59.6 ±0.2 |
| **Libra XGradCAM+** | 74.2 ±0.1 (+23.2%) | 70.6 ±0.2 (+19.8%) | 76.3 ±0.1 (+24.8%) | 72.2 ±0.2 (+21.2%) |
| FullGrad+ | 60.2 ±0.1 | 59.5 ±0.3 | 60.9 ±0.1 | 60.1 ±0.3 |
| **Libra FullGrad+** | **77.6** ±0.1 (+29.0%) | **74.4** ±0.2 (+25.1%) | **79.6** ±0.1 (+30.6%) | **76.3** ±0.2 (+27.0%) |

Table 89. Comparison of attribution methods and their LibraGrad-enhanced versions on the DeiT3-H model.

## E. Related Work

Input attribution methods are techniques designed to quantify the influence of individual input features, or groups of them, on a model's output [12, 44, 48, 49, 67, 74, 75, 94]. Input attribution methods can assist in understanding a model's decision locally for a single input considered in isolation. They also act as foundational elements for more advanced explanation techniques. For instance, in concept-based explanation methods like CRAFT [31], attribution methods are employed for two main purposes: to quantify the impact of each activated concept and to identify the specific input features responsible for activating these concepts.

Attribution methods have a wide array of applications beyond merely explaining model outputs to humans [27, 69, 84, 87]. They are useful for enhancing the robustness of models against out-of-distribution data, spurious correlations, and adversarial inputs [5, 18, 56, 91]. Additionally, attribution methods have been employed to improve the performance of text-to-image models [19, 43, 58]. Furthermore, adapting forward-mode attribution methods has been explored for on-the-fly feature pruning [30, 52] and model quantization [9]. Attribution methods have been utilized to construct more effective adversarial attacks against models [40, 89, 95].

Given a multi-output neural model, let $f : \mathbb{R}^n \to \mathbb{R}$ be a selected output function. For instance, if $\text{Model}(x) = (p_1, ..., p_k)$ represents class probabilities, we might choose $f(x) = p_i$ to analyze the model's prediction for the $i$-th class. An attribution method $A$ generates relevance scores $A(f)(x)_i$ for each feature $x_i$.

### E.1. Gradient-Based Attribution Methods

**Input $\times$ Grad.** IxG [4, 72, 73] assigns feature relevance by $\text{IxG}\,(f)(x) = x \odot \nabla_x f(x)$, where $\odot$ denotes element-wise multiplication.

**FullGrad.** Expanding on Input $\times$ Grad, FullGrad [76] includes not only the input features but also the bias terms of each layer in the neural network. The FullGrad attribution map is calculated as:

$$\text{FullGrad}(f)(x_0) = \text{IxG}\,(f)(x_0) + \sum_{l=0}^{L-1} \sum_{b \in B_l} \text{IxG}\,(f_b)(b)$$

where $\text{IxG}\,(f)(x_0)$ denotes the Input $\times$ Grad for the input $x_0$ (the input to the first layer), and $\text{IxG}\,(f_b)(b)$ is the Input $\times$ Grad attribution map of the sub-network $f_b$ with a bias term $b$ from layer $l$ as the input. Also, $f_b$ is the sub-network of $f$ starting from the bias term $b$ and going until the end of the model, whereas $B_l$ denotes the set of all bias terms in layer $l$. FullGrad+ $\circ$ PLUS (henceforth FullGrad+) [50] is defined as follows:

$$\text{FullGrad+}\,(f)(x_0) = \sum_{l=0}^{L-1} \text{IxG}\,(f_l)(x_l) + \sum_{l=0}^{L-1} \sum_{b \in B_l} \text{IxG}\,(f_b)(b)$$

where $\text{IxG}\,(f_l)(x_l)$ is the Input $\times$ Grad attribution map of the sub-network $f_l$ with input $x_l$ (the input to the $l$th layer). FullGrad+ aggregates the input attribution maps of each layer along with the attribution maps of all bias terms in each layer.

**Integrated Gradients.** IG [78] computes attributions w.r.t. a baseline input $\bar{x}$ (*e.g.*, zero):

$$\text{IG}\,(f)(x) = (x - \bar{x}) \odot \int_{\alpha=0}^{1} \nabla_x f(\bar{x} + \alpha(x - \bar{x}))d\alpha$$

In practice, we approximate the integral using a 50-step Riemann summation.

**GradCAM.** GradCAM [68] averages the gradient signal across each channel before multiplying it with the input, and operates on the last layer of the network:

- $A^k$: the k-th channel of the feature map in the final layer
- $c$: the class w.r.t. which the attribution map is computed
- $y^c$: the class score (logit)
- Gradients are averaged over the width and height dimensions (indexed by i and j respectively) to obtain the neuron (channel) importance weights $\alpha_k^c$:

$$\alpha_k^c = \overbrace{\frac{1}{Z} \sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

**XGradCAM+.** XGradCAM weights the gradients by their corresponding activation value when computing the spatial average [33]. XGradCAM was proposed on ReLU CNNs where the activations were always positive, hence they did not specify using the absolute value of the activations in the above computation, as is more intuitive. The variant with absolute activations is named XGradCAM+ [50].

**HiResCAM.** HiResCAM [26] is equivalent to Input $\times$ Grad on the last layer of the model. (Standard Input $\times$ Grad is applied on the first layer of the model.)

**PLUS.** PLUS [50] is a way for attribution methods to better aggregate information across layers.

### E.1.1. Gradient-Attention Hybrids

**AttCAT.** AttCAT [62] combines attention weights with Input × Grad to create a hybrid attribution method. The approach operates by first computing the input-times-gradient attribution at each layer, then weighting these attributions using the attention weights from the corresponding attention heads. The method addresses the limitations of pure attention-based or pure gradient-based approaches by leveraging both sources of information. By incorporating both attention patterns and gradient information, AttCAT can better capture the model's decision-making process, particularly in cases where either attention or gradient alone might miss important feature interactions. The final attribution map is computed by aggregating these weighted scores across all layers and attention heads.

**TransAtt.** TransAtt [17] employs the Deep Taylor Decomposition technique [54] to attribute local relevance and subsequently propagates these relevance scores through the entire architecture of a Transformer model. This process effectively enables the backward propagation of information across all layers, starting from the output and extending back to the input. Additionally, this method incorporates gradients of attention weights. The method's functioning can be summarized as follows:

$$\mathit{Rollout}\left(\mathbb{E}_{H:=\text{Heads}}\left[(\mathtt{R}\odot\mathtt{AttnGrad})^{+}\right]\right),$$

where $\mathtt{R}$ stands for the relevancy scores of attention weights. The Rollout technique is a method to aggregate the layer-wise attribution maps. We refer the reader to [1] for a detailed overview.

**GenAtt.** The dependence of TransAtt on specific rules for the propagation of relevance scores imposes limitations on its capacity to furnish explanations for various types of Transformer architectures. To cope with this issue, GenAtt [16] attempts to explain predictions for any Transformer-based architecture by using the attention weights in each block to update the relevancy maps, as demonstrated by the following expression:

$$\mathit{Rollout}\left(\mathbb{E}_{H:=\text{Heads}}\left[(\mathtt{Attn}\odot\mathtt{AttnGrad})^{+}\right]\right).$$

The notation $()^{+}$ denotes a filtering through the ReLU function. [16] show that GenAtt is at least as effective as TransAtt, if not better.

**TokenTM.** TokenTM [88] further improves GenAtt by taking token transformations into account.

## E.2. LRP Methods

Layer-wise Relevance Propagation (LRP) is a principled attribution method that propagates relevance scores backward through a neural network by following specific propagation rules.

**AliLRP.** AliLRP [3] extends traditional LRP for Transformer architectures by introducing specialized propagation rules that offer better numerical stability.

**AttnLRP.** AttnLRP [2] extends LRP to handle attention layers.

## E.3. Forward Attention-Based Token Attribution Methods

**Attention×Input_Norm (AttIN).** Kobayashi et al. [45] multiply the attention weights by the norms of the vectors corresponding to each attention weight. Kobayashi et al. [46] extends AttIN to also incorporate the residual connections.

**GlobEnc & ALTI.** AttIN assumes that tokens retain their original identity. As each self-attention module mixes all the tokens, this assumption might not necessarily hold. Using gradient-based techniques, Brunner et al. [14] studies contextual information aggregation across the model. Following Brunner et al. [14] work, the global token attribution analysis method GlobEnc [51] further extends AttIN by including the Transformer block's second normalization layer in its analysis. In parallel with GlobEnc, the Aggregation of Layer-Wise Token-to-Token Interactions method ALTI [32] was introduced. ALTI shares core concepts with GlobEnc, but the two differ in certain mathematical specifics.

**DecompX.** DecompX [53] enhances GlobEnc by integrating the one element previously overlooked by GlobEnc: the MLP module in the Encoder Transformer layer. This inclusion enables DecompX to generate a set of decomposed vectors that collectively sum up to the actual output vector. Unlike GlobEnc and ALTI, which require computing and aggregating layer-wise attribution maps using techniques like Rollout, DecompX facilitates the direct propagation of these decomposed vectors across layers. This capability allows for the direct computation of attribution maps from any layer to any other layer.

## E.4. Architectural Limitations of Previous Methods

Previous methods often had architectural limitations. DecompX presents equations for an encoder-only model which are non-trivial to extend to the encoder-decoder case. GenAtt is a simplified version of TransAtt that is supposed to support more Transformer architectures, but its Section

3.2 ("Adaptation to attention types") still does not cover many models, *e.g.*, SigLIP with a global average pooling classifier. Such a model lacks the CLS token used in GenAtt, Attention Rollout, and RawAtt. TokenTM is similarly only presented for encoder-only models with a CLS token. LRP methods need specialized rules for any new modules, while LibraGrad can naturally fall back on the standard gradients of these modules. Furthermore, LibraGrad utilizes the automatic differentiation capabilities of modern deep learning frameworks, which makes a bug-free, optimized implementation straightforward.

### E.5. Black-Box Methods

Black-box attribution methods treat the model as an opaque entity, (partially) disregarding its internal structure and gradients. These methods typically involve perturbing the input and observing the corresponding changes in the model's output to infer the importance of each input feature. However, this approach often comes with significant computational costs due to the need for multiple model evaluations. In contrast, white-box methods leverage the internal structure and gradients of the model, providing a more efficient and fine-grained understanding of the model's behavior.

In this paper, we focus on white-box methods for several reasons. Firstly, they offer a more computationally efficient approach compared to black-box methods. Secondly, and more importantly, black-box methods can be seen as directly optimizing the faithfulness metrics on which we evaluate the attribution methods. This raises concerns related to Goodhart's law, which states that when a measure becomes a target, it ceases to be a good measure. In other words, the faithfulness metrics we use are merely proxies for the ultimate desirable properties we seek in attribution methods. By directly optimizing these metrics, black-box methods may inadvertently introduce biases or artifacts that undermine the true faithfulness of the attributions. Therefore, to avoid this potential pitfall and maintain a more objective evaluation, we refrain from including comparisons with black-box methods in this study, acknowledging that they have different trade-offs and use cases.

**LIME** [66] explains the predictions of any classifier by learning a local interpretable model around the prediction.

**RISE** [61] is a black-box approach that generates an importance map indicating the saliency of each pixel for the model's prediction by probing the model with randomly masked versions of the input image and obtaining the corresponding outputs.

**PAMI** [71] masks the majority of the input and uses the corresponding model output as the relative contribution of the preserved input part to the original model prediction.

**ScoreCAM** [86] is a post-hoc visual explanation method based on class activation mapping that eliminates the dependence on gradients by obtaining the weight of each activation map through its forward passing score on the target class.

**ViT-CX** [90] adapts ScoreCAM for ViTs.

**AtMan** [23] is a perturbation method that manipulates the attention mechanisms of transformers to produce relevance maps for the input with respect to the output prediction.

**HSIC** [57] is a black-box attribution method based on the Hilbert-Schmidt Independence Criterion, measuring the dependence between regions of an input image and the model's output using kernel embeddings of distributions.