Learn to Rank: Visual Attribution by Learning Importance Ranking

David Schinagl[1], Christian Fruhwirth-Reisinger[1,2], Alexander Prutsch[1], Samuel Schulter[3⋆], and Horst Possegger[1]

[1] Institute of Visual Computing (IVC), Graz University of Technology, Graz, Austria
{david.schinagl,reisinger,alexander.prutsch,possegger}@tugraz.at
[2] Nomadic AI, USA
[3] Amazon, USA

Abstract. Interpreting the decisions of complex computer vision models is crucial to establish trust and accountability, especially in safety-critical domains. An established approach to interpretability is generating visual attribution maps that highlight regions of the input most relevant to the model's prediction. However, existing methods face a three-way trade-off. Propagation-based approaches are efficient, but they can be biased and architecture-specific. Meanwhile, perturbation-based methods are causally grounded, yet they are expensive and for vision transformers often yield coarse, patch-level explanations. Learning-based explainers are fast but usually optimize surrogate objectives or distill from heuristic teachers. We propose a learning scheme that instead optimizes deletion and insertion metrics directly. Since these metrics depend on non-differentiable sorting and ranking, we frame them as permutation learning and replace the hard sorting with a differentiable relaxation using Gumbel-Sinkhorn. This enables end-to-end training through attribution-guided perturbations of the target model. During inference, our method produces dense, pixel-level attributions in a single forward pass with optional, few-step gradient refinement. Our experiments demonstrate consistent quantitative improvements and sharper, boundary-aligned explanations, particularly for transformer-based vision models. Code and pre-trained models are available at `https://github.com/dschinagl/AHA`.

Keywords: Visual Attribution · Amortized Explanation · Permutation

1 Introduction

Over the past decade, deep neural networks [24,31,39,40] have powered the vast majority of vision models. The growing complexity of these inherently opaque models, however, makes interpreting their decisions difficult. Especially in safety-critical domains such as healthcare or autonomous driving, explaining *why* a model predicts a certain outcome can be as crucial as the accuracy of the prediction itself. A well-established approach for explainability in computer vision

⋆ This work is independent of the author's employment at Amazon.

is the creation of visual attribution maps, which highlight the input regions most relevant to a model's prediction [14,46,53,56,70]. Such attribution maps provide an intuitive, human-interpretable way to understand model behavior and can help identify potential biases or failure modes [4, 11].

The majority of attribution methods are *post-hoc*, *i.e.* they explain already trained models without modifying their architecture. These can be broadly categorized into three groups: *Propagation-based* approaches which rely on propagating internal signals through the target model, such as gradients [56, 62, 70], activations [53, 66, 72] or attention weights [1, 14, 15] in transformer-based vision models. They are very efficient and align with the model's internal computations but are architecture-specific, can bias toward low-level features like edges [2, 59] and may under-represent non-attention pathways in transformer models. In contrast, *perturbation-based* methods [28, 41, 46, 50] infer feature relevance from changes in the target model's output when parts of the input are modified. They provide more causally grounded attributions, as they directly measure the effects of input changes, including higher-order interactions. However, they are computationally demanding, requiring many model evaluations to produce a single attribution map. For transformer-based vision models, they usually operate on the native patch representation [25, 64, 67] yielding patch-level explanations that must be upsampled to pixel space, resulting in coarse, spatially imprecise maps, as shown in Figure 1. The third and least explored category, *learning-based* (amortized) explainers shift the computational burden to an offline training phase by learning a separate explainer model, which is then able to generate attribution maps very efficiently during inference [21]. However, the current training strategies typically rely on either supervision from other attribution methods [32, 58] and may adopt their biases, or on self-supervised surrogate objectives that are based on restrictive assumptions like a binary signal to noise decomposition of features [7, 18].

Motivated by these limitations, it is natural to ask whether we can avoid attribution heuristics and instead *optimize attribution quality directly*. Common and widely-used attribution metrics are the *Deletion* and *Insertion* metrics [46]. They quantify the change in a model's output when features are removed or added back in order of importance. Recent metric-driven approaches [20, 64] for vision transformers pursue this direction, but do so via a perturbation-based, per-sample optimization on the patch representation. Consequently, they retain the high computational cost and cannot refine the coarse spatial resolution.

In this work, we propose *Amortized Hybrid Attribution (AHA)*: a learning-based approach that bridges these families of attribution methods by integrating their complementary strengths into a unified framework. We retain the causal grounding of perturbation-based methods by learning from the target model's outputs under attribution-guided perturbations. Since we couple explainer and target model through end-to-end backpropagation, we inherit the model-consistent attribution characteristic of propagation-based approaches. We realize this by directly optimizing the *Deletion* and *Insertion* metrics. To overcome the non-differentiability of these metrics, caused by the sorting operations

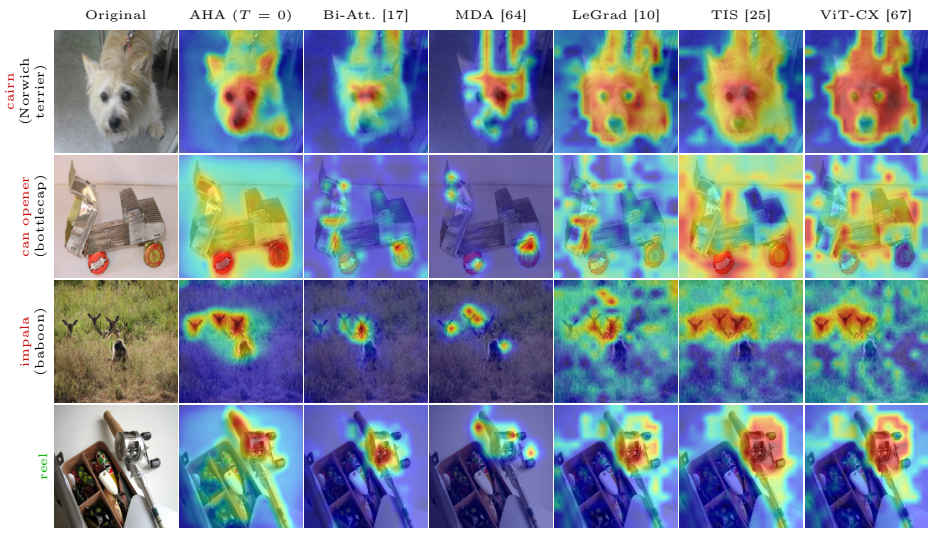

Fig. 1: Qualitative comparison of attribution maps produced by our method (AHA) and recent ViT attribution methods for a ViT-B/16 classifier [24] on ImageNet [22] samples. The green / red labels mark correct / incorrect predictions (ground truth in parentheses). While existing approaches operate on the patch representation and produce coarse, blocky attribution maps, our method generates dense, pixel-level attributions that align closely with object boundaries. Note that in this case no additional backprop refinement steps were applied at inference (denoted $T = 0$).

they depend on, we leverage the observation that they depend solely on the order of pixel importance rather than their exact values, and treat them as permutation problems. By replacing the hard sorting and ranking operations with smooth approximations, we enable end-to-end training of the explainer model. At inference, we preserve the amortized efficiency of learned explainers (single forward pass), but we also keep the door open to the propagation-based regime: we can optionally perform a few additional backward steps through the target model to locally refine the explainer's output on a specific sample. This effectively combines a learned, metric-aligned prior with a cheap, model-specific propagation signal. Due to the end-to-end optimization our explainer is able to generate dense attributions that align closely with object boundaries and visually meaningful structures, even for transformer-based vision models, as shown in Figure 1.

Our contributions can be summarized as follows:

- We propose a learning-based attribution framework that bridges perturbation- and propagation-based explainability.
- Our amortized explainer directly optimizes differentiable surrogates of *Deletion* and *Insertion*, producing dense pixel-space attributions also for transformer-based vision models.

- We formulate metric optimization as a permutation learning problem and enable end-to-end training via a Gumbel-Sinkhorn relaxation of sorting and top-k selection.
- To reduce the computational burden and mitigate pixel-level shortcut artifacts, the training procedure leverages region-level permutation, perturbation-step sampling, and grid augmentation.
- At inference, the method retains single-pass efficiency and optionally adds a few backprop refinement steps for sample-specific adjustments.

2 Related Work

Visual attribution methods highlight the most influential input regions for a given model decision. *Ante-hoc* methods [3] alter the model architecture or training, *e.g.*, prototype-based networks [16, 23, 38, 44], concept bottleneck models [37, 45], or alignment-based approaches [5, 9, 63]. The vast majority, however, are *post-hoc* methods that explain already trained models without modifying them, typically via *propagation-*, *perturbation-* or *learning-based* methods.

Propagation-based Methods propagate attribution signals through the model to determine the relevance of input features. *Backpropagation*-based methods use gradients of the output w.r.t. the input [56, 70]. Various enhancements mitigate noisy gradients, including aggregating multiple gradient signals [35, 36, 61, 62, 68], preventing negative gradients from flowing back [60], or specific backpropagation rules [6, 54, 71]. *Forward propagation*-based methods, on the other hand, utilize the activations of intermediate layers to generate attribution maps [34, 72]. Grad-CAM [53] and its successors [13, 33] use the gradient w.r.t. the feature maps to weight the activations in the last convolutional layers. Specialized approaches to interpret vision transformers [12, 24, 30, 39, 48] typically analyze the self-attention mechanism: Attention-Rollout [1] recursively multiplies attention weights across layers, Attention-Flow [1] models attention as a flow network, and Transition-Attention [69] accumulates attention as a Markov process combined with Integrated Gradients [62]. Based on LRP [6], Transformer-Attribution [15] proposes propagation rules with gradient integration for self-attention layers. To approximate token contributions, Bidirectional Attention [17] jointly analyzes attention and partial derivatives to produce interpretable attribution signals. Instead of using the gradients to weight the attention maps, LeGrad [10] uses the gradient w.r.t. the attention scores as attribution signals.

The major advantages of these approaches are computational efficiency and their model faithfulness as the attribution is directly derived from the model's internal computations. However, they are often tied to specific architectures or layer types and the reliance on local gradients can bias the attributions toward low-level edges and textures [2, 59].

Perturbation-based Methods directly measure output changes under systematic input perturbations to create causally grounded attributions [73]. In their pioneering work, Zeiler and Fergus [70] monitored the model's output

change while a sliding window occludes parts of the input. LIME [50] uses superpixel perturbations to fit an interpretable surrogate model per sample, whereas SHAP [41] leverages cooperative game theory to assign attributions by approximating Shapley values over feature subsets for individual predictions. Fong *et al*. [27,28] formulate attribution as a per-sample optimization problem, searching for the smallest region that maximally influences the output. RISE [46] generates attribution maps as an output score-weighted combination of random input masks, while EVA [26] proposes a more systematic and exhaustive exploration of the perturbation space. While these perturbation-based approaches have predominantly been studied for CNNs, recent adaptations exploit the patch-wise input representation of vision transformers: ViT-CX [67] derives attribution by scoring the causal impact of patch tokens on the model output, while TIS [25] leverages the variable-length token processing of transformers by removing token subsets. Contrary to heuristic attribution generation methods, Metric-Driven Attributions (MDA) [64] aims to directly optimize the metrics used to evaluate the attribution maps. In a per-sample optimization process, MDA aims to find the optimal order and magnitude of patch perturbations to maximize the evaluation metrics. Relatedly, *Less is More* [20] builds on a given attribution map as a prior and refines it via greedy submodular selection of few regions, resulting in a small set of explanatory masks.

Since perturbation-based methods directly measure the model responses, they provide causally grounded attributions and can capture higher-order feature interactions. However, a major drawback is that they are computationally expensive as multiple model evaluations are required per attribution map. Moreover, methods tailored to vision transformers typically yield only patch-level attributions and therefore rely on upsampling to obtain input-resolution maps, often resulting in coarse and spatially imprecise attribution maps (recall Fig. 1).

Learning-based Methods (Amortized Explainers) train a neural explainer model to directly output attributions [21]. This amortizes the cost of per-sample iterative computation by shifting it to an offline training stage performed once per target model, while explanations can be generated efficiently at inference time. *Prior explanation*-driven approaches use other attribution methods as supervisory signals [58] or approximate Shapley-based priors [32]. *Self-supervised amortizers* on the other hand, interact with the target model to optimize a proxy objective. However, these objectives often rely on ideal assumptions like a near-binary signal/noise decomposition to isolate relevant features [7], or assume that only a small, separable feature subset carries the predictive signal [18], which may not hold in practice, especially for complex vision models.

In contrast, our learning-based approach is directly guided by recognized evaluation metrics (*Deletion* & *Insertion*) without any heuristic assumptions about the underlying data distribution or model behavior. Since the supervisory signal is derived from the model's responses to perturbations, our method remains causally grounded. Additionally, our end-to-end optimization schema is not limited to patch-level attribution, resulting in denser and better aligned results than existing ViT-specific methods.

3 Amortized Hybrid Attribution

In the following, we first formalize the problem setting to obtain attribution maps produced for a fixed, pretrained classifier. We then review the non-differentiable perturbation-based *Deletion* and *Insertion* metrics that serve as our target notion of attribution quality, and present our differentiable relaxation. This makes the involved sorting and top-k selection operations differentiable, allowing us to optimize the metrics end-to-end. Finally, we describe practical strategies to make training efficient and robust at high resolution, as well as an optional test-time refinement to adapt the attribution map to sample-specific characteristics for use cases which prioritize explanation quality over throughput.

Problem Setting and Notation. Let $f : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^C$ denote a pretrained image classifier that maps an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ to class probabilities $f(\mathbf{I}) = \mathbf{s} \in \mathbb{R}^C$, where s_c corresponds to the predicted probability of class c. Given a target class $t \in \{1, \dots, C\}$, an attribution method is defined as a mapping $\Phi_f : \mathbb{R}^{H\times W\times 3} \times \{1, \dots, C\} \to \mathbb{R}^{H\times W}$, which produces an attribution map $\mathbf{A} = \Phi_f(\mathbf{I}, t)$ of the same spatial resolution as the input, where each element a_{ij} reflects the importance of spatial location (i, j) to the target probability s_t. Unless otherwise specified $t = \arg\max_c s_c$.

Metrics. Attribution methods are commonly evaluated by perturbation metrics measuring changes in model output under attribution-guided input perturbations. Petsiuk *et al.* [46] introduced the *Deletion* and *Insertion* metrics, which are widely used for attribution assessment [10, 14, 15, 19] and optimization [64]. *Deletion* quantifies the drop in the target probability as high-attribution pixels are removed, while *Insertion* measures the probability increase as these pixels are added to a blank reference image. Formally this is defined as follows: Let $\mathbf{a} = \mathrm{vec}(\mathbf{A}) \in \mathbb{R}^N$ with $N = H \cdot W$ denote the vectorized attribution map. First, a permutation $\pi_{\mathbf{A}}$ over the pixel indices is computed such that it sorts the entries of $\mathbf{a}$ in descending order, $a_{\pi_{\mathbf{A}}(1)} \geq a_{\pi_{\mathbf{A}}(2)} \geq \cdots \geq a_{\pi_{\mathbf{A}}(N)}$. Based on this order a sequence of binary masks $\{\mathbf{M}_{\mathbf{A}}^{(k)}\}_{k=1}^{N}$ is constructed, where each mask $\mathbf{M}_{\mathbf{A}}^{(k)} \in \{0, 1\}^{H\times W}$ selects the top-k pixels. Next, the input image $\mathbf{I}$ is progressively blended with a blank reference image $\mathbf{I}_0$ (e.g., black image, mean colored image or a blurred version of the original image) using these masks, yielding two sequences of perturbed images $\{\mathbf{I}_{del}^{(k)}\}_{k=1}^{N}$ and $\{\mathbf{I}_{ins}^{(k)}\}_{k=1}^{N}$ where,

$$\mathbf{I}_{del}^{(k)}(\mathbf{A}, \mathbf{I}) = \mathbf{I} \odot (1 - \mathbf{M}_{\mathbf{A}}^{(k)}) + \mathbf{I}_0 \odot \mathbf{M}_{\mathbf{A}}^{(k)}, \tag{1}$$

$$\mathbf{I}_{ins}^{(k)}(\mathbf{A}, \mathbf{I}) = \mathbf{I} \odot \mathbf{M}_{\mathbf{A}}^{(k)} + \mathbf{I}_0 \odot (1 - \mathbf{M}_{\mathbf{A}}^{(k)}), \tag{2}$$

with $\odot$ denoting element-wise multiplication. In $\mathbf{I}_{del}^{(k)}$, the top-k pixels of the original image are replaced by the blank reference image, while in $\mathbf{I}_{ins}^{(k)}$, the top-k pixels are inserted into the blank reference image. The attribution quality is then evaluated by computing the area under the curve (AUC) of the target class

probability s_t across the sequence of perturbed images:

$$del_{\text{AUC}}(\mathbf{A}, \mathbf{I}, t) = \frac{1}{N} \sum_{k=1}^{N} f(\mathbf{I}_{del}^{(k)}(\mathbf{A}, \mathbf{I}))_t, \tag{3}$$

$$ins_{\text{AUC}}(\mathbf{A}, \mathbf{I}, t) = \frac{1}{N} \sum_{k=1}^{N} f(\mathbf{I}_{ins}^{(k)}(\mathbf{A}, \mathbf{I}))_t. \tag{4}$$

Meaningful attribution maps lead to a rapid score drop under *Deletion* (low del_{AUC}) and a rapid score increase under *Insertion* (high ins_{AUC}), indicating that highly ranked pixels indeed have strong influence on the prediction.

3.1 Learn to Attribute

In our method we aim to train an attribution model $\Phi_{f,\theta}$ with parameters θ for a given model f, that is able to create dense attribution maps $\mathbf{A}$ which follow the definition of importance as measured by the *Deletion* and *Insertion* metrics. However, the metrics defined in Equations (3) and (4) are non-differentiable w.r.t. the attribution map $\mathbf{A}$ due to the discrete pixel sorting (permutation) $\pi_{\mathbf{A}}$ and the corresponding binary top-k masks $\mathbf{M}_{\mathbf{A}}^{(k)}$, which introduce discontinuities whenever the order changes. Nevertheless, a key observation is that both metrics depend exclusively on the *relative order* of the attribution values, not their absolute values. This fact allows us to bypass the non-differentiability by employing continuous relaxations of the sorting and selection operations, using differentiable sorting techniques [8, 42, 47] to obtain smooth approximations. This relaxation yields soft permutation matrices and, in turn, soft top-k masks that enable end-to-end optimization of del_{AUC} and ins_{AUC} with respect to the attributions.

Soft Permutation Matrices. The hard permutation $\pi_{\mathbf{A}}$ induced by sorting the attribution scores $\mathbf{a}$ can be represented as permutation matrix $\mathbf{P}_{\pi_{\mathbf{A}}} \in \{0,1\}^{N \times N}$ such that,

$$\mathbf{P}_{\pi_{\mathbf{A}}} \mathbf{a} = \left[a_{\pi_{\mathbf{A}}(1)},\, a_{\pi_{\mathbf{A}}(2)},\, \ldots,\, a_{\pi_{\mathbf{A}}(N)}\right]^{\top}. \tag{5}$$

Our approach to obtain a differentiable approximation of this permutation matrix is based on the *Gumbel-Sinkhorn* algorithm [42]. First, we compute a similarity matrix $\mathbf{L} \in \mathbb{R}^{N \times N}$, which measures the similarity between the attribution scores $\mathbf{a}$ produced by the explainer model and a set of N target positions $\mathbf{p} = \frac{1}{N}[N, N-1, \ldots, 1]^{\top}$ defined as linearly decreasing values, representing the ideal sorted distribution. Each entry $\mathbf{L}_{i,j}$ is computed as the negative squared distance between the attribution score a_i and the target position p_j,

$$\mathbf{L}_{i,j} = -(a_i - p_j)^2, \tag{6}$$

where larger values correspond to smaller assignment costs and therefore better matches between attribution scores and target ranks. Subsequently, we add noise

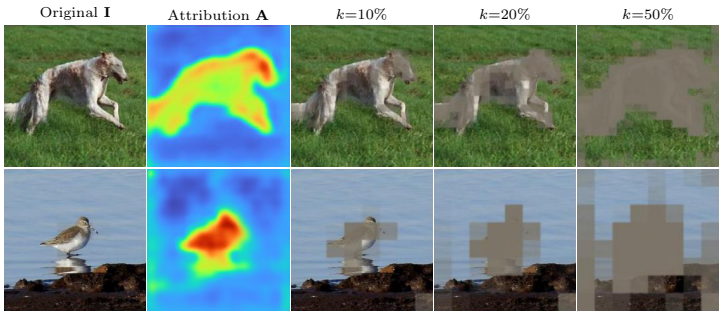

Fig. 2: Soft deletion masks. Given an input image $\mathbf{I}$ and its attribution map $\mathbf{A}$, the soft top-k masks $\mathbf{m}_{\mathbf{A}}^{\text{soft},(k)}$ progressively replace the most important regions with the reference image. Unlike hard binary masks, the soft relaxation (Sec. 3.1) assigns continuous importance values to each region, enabling gradient-based optimization of the deletion and insertion metrics. Note the varying grid resolution G across examples top *vs.* bottom (Sec. 3.2) which prevents overfitting to a fixed spatial partitioning.

sampled from the Gumbel distribution $\mathbf{G} \sim \text{Gumbel}(0, 1)$ to introduce stochasticity, encouraging the exploration of different permutations during training,

$$\widetilde{\mathbf{L}} = \frac{\mathbf{L} + \mathbf{G}}{\tau}, \tag{7}$$

where $\tau > 0$ is a temperature parameter that controls the smoothness of the resulting approximation. This formulation frames the sorting task as an *optimal transport (OT)* problem, seeking the minimum-cost assignment between the attribution scores and the target positions. By applying the *Sinkhorn-Knopp* algorithm [57] to the matrix $\widetilde{\mathbf{L}}$, which performs iterative row and column normalization, we obtain a doubly-stochastic matrix $\mathbf{P}_{\pi_{\mathbf{A}}}^{\text{soft}} \in [0, 1]^{N \times N}$ that serves as a continuous relaxation of the permutation matrix $\mathbf{P}_{\pi_{\mathbf{A}}}$.

Soft Top-k Masks. A hard top-k mask $\mathbf{M}_{\mathbf{A}}^{(k)} \in \{0, 1\}^{H \times W}$ selects all pixels whose rank is between 1 and k. The vectorized form of this mask $\mathbf{m}_{\mathbf{A}}^{(k)} \in \{0, 1\}^{N}$ is created by summing the first k rows of the permutation matrix $\mathbf{P}_{\pi_{\mathbf{A}}}$. Similarly, given our soft permutation matrix $\mathbf{P}_{\pi_{\mathbf{A}}}^{\text{soft}}$, we create a sequence of N soft top-k masks $\{\mathbf{m}_{\mathbf{A}}^{\text{soft},(k)}\}_{k=1}^{N}$, where $\mathbf{m}_{\mathbf{A}}^{\text{soft},(k)} \in [0, 1]^{N}$ is obtained by computing the cumulative sum along the rank dimensions (rows) of the soft permutation matrix

$$\mathbf{m}_{\mathbf{A}}^{\text{soft},(k)} = \sum_{i=1}^{k} \mathbf{P}_{\pi_{\mathbf{A}}}^{\text{soft}}[i, :]^{\top}. \tag{8}$$

This results in masks with continuous values between 0 and 1, indicating the degree to which each pixel is included in the top-k selection, illustrated in Fig. 2.

Differentiable *Deletion & Insertion*. Given the sequence of soft top-k masks, we now replace the hard masks $\mathbf{M}_{\mathbf{A}}^{(k)}$ in Equations (1) and (2) with the

reshaped soft masks $\mathbf{m}_{\mathbf{A}}^{\text{soft},(k)}$ to obtain $\mathbf{I}_{del}^{\text{soft},(k)}$ and $\mathbf{I}_{ins}^{\text{soft},(k)}$. Consequently, we get differentiable approximations of the *Deletion* and *Insertion* metrics as:

$$del_{\text{AUC}}^{\text{soft}}(\mathbf{A},\, \mathbf{I},\, t) = \frac{1}{N}\sum_{k=1}^{N} f\left(\mathbf{I}_{del}^{\text{soft},(k)}(\mathbf{A},\mathbf{I})\right)_t, \tag{9}$$

$$ins_{\text{AUC}}^{\text{soft}}(\mathbf{A},\, \mathbf{I},\, t) = \frac{1}{N}\sum_{k=1}^{N} f\left(\mathbf{I}_{ins}^{\text{soft},(k)}(\mathbf{A},\mathbf{I})\right)_t. \tag{10}$$

The loss function for training the attribution model can then be defined as a weighted combination of these two metrics, encouraging the model to produce attribution maps guided directly by attribution quality measures:

$$\mathcal{L}(\theta) = \lambda_{\text{del}} \cdot del_{\text{AUC}}^{\text{soft}}(\mathbf{A},\, \mathbf{I},\, t) - \lambda_{\text{ins}} \cdot ins_{\text{AUC}}^{\text{soft}}(\mathbf{A},\, \mathbf{I},\, t), \tag{11}$$

where $\lambda_{\text{del}}, \lambda_{\text{ins}}$ are weighting hyperparameters.

3.2 Efficient and Robust Optimization

Although the differentiable relaxations allow us to compute gradients with respect to the attribution maps, a naive implementation of the training process that involves computing the full sequence of N perturbations for each image would be computationally prohibitive, especially for high-resolution images. On the other hand, optimizing the loss described in Equation (11) on pixel-level would be prone to overfitting, as the attribution model could exploit high-frequency artifacts, resulting in directions targeting the architecture-dependent weaknesses of the model under examination. To mitigate these issues, we propose strategies making the training process practical and robust:

- **Region-Based Permutation.** Instead of sorting all N pixels individually, we partition the image into $K \ll N$ disjoint regions using a regular grid of size $K = G \cdot G$ and mean-pool the attribution values within each region. After the differentiable sorting and soft top-k selection on these region-level scores, the resulting masks are upsampled to the original image resolution for perturbation. This significantly reduces the number of perturbations during training and reduces the risk of exploiting individual pixel artifacts.
- **Sampling Perturbation Steps.** To further reduce the computational burden during training, we do not evaluate the full sequence of K perturbation steps for each image. Instead, we uniformly sample a smaller set of $S \ll K$ steps from the range $\{1, \ldots, K\}$, *i.e.*, we perturb multiple grid regions at once instead of adding them one by one. This approximation of the AUC metrics decreases the number of needed forward passes through the model to a fixed budget.
- **Grid Augmentation.** While the grid-based region partitioning enhances the training efficiency and robustness, a fixed grid alignment may introduce bias towards specific spatial patterns. To alleviate this effect we randomly

vary the grid size G within a predefined range for each training iteration and apply random offsets to the grid's position. These augmentations encourage the attribution model to focus on generalizable importance patterns that are not tied to specific grid configurations or scales.
- **Regularization.** In addition to the objective in Equation (11), we regularize the attribution model to discourage spurious, isolated peaks, *i.e.* individual pixels or small clusters of pixels with high scores that are not supported by their neighboring regions. Concretely, we convolve the attribution map $\mathbf{A}$ with an averaging filter and penalize large deviations from the original map using an L_2 loss.

Taken together, these strategies do not merely make training tractable, they also shape which attribution patterns the objective rewards. A natural concern is that the order-based metric could be satisfied by diffuse maps, since only the relative ranking matters in Equations (3) and (4). The optimization, however, does not act on the ordering alone. The Sinkhorn assignment matches the continuous attribution values to the target ranks through the distance costs in Equation (6), so the degree of separation between region scores controls the sharpness of the resulting soft top-k masks. Near-tied scores thus yield ambiguous masks that mix foreground and background, whereas clearly separated scores yield decisive masks. This separation is what the AUC objectives reward: *Deletion* and *Insertion* are driven by a steep score change already at small k, which only arises when the perturbation is concentrated on the few truly causal regions. Decisive masks therefore induce low *Deletion* and high *Insertion*, while diffuse masks perturb the whole image weakly and flatten both curves. Together with the region-level perturbations, the randomized grid sizes and offsets, and the regularizer penalizing isolated peaks, this drives the explainer toward compact, object-aligned attributions.

3.3 Optional Test-Time Refinement

The design of our attribution framework allows for an optional, fast test-time refinement step at low computational cost, that can further enhance the quality of the generated heatmaps on a per-image basis. The key idea is to briefly fine-tune the explainer for a few gradient steps on the specific test image using the same differentiable AUC objectives used during training. To make refinement more stable and deterministic, we use soft sorting without Gumbel noise and fix the sorting temperature to the final training value, *i.e.* $\tau = \tau_{\text{final}}$. After T optimization steps, a final forward pass produces the refined attribution map. This refinement is particularly useful when explanation quality is prioritized over throughput, with the compute-quality trade-off being adjustable via the number of refinement steps. Moreover, if inputs are expected to significantly deviate from the training distribution, this test-time refinement can help to adapt the explainer to the specific characteristics of the new input, improving attribution quality in such cases.

4 Experiments

4.1 Implementation Details

Unless stated otherwise, we use the following setup for all experiments throughout this section. We train our attribution model on the ImageNet-1K training split [22] to explain a frozen ImageNet ViT-B/16 classifier [24]. Our explainer is an encoder-decoder model with a frozen DINOv3 [55] ViT-L/16 backbone and a DPT-inspired decoder [49] conditioned on a learned class embedding. We optimize only the explainer head parameters and keep the DINOv3 backbone frozen. To create perturbation masks during training, we average-pool the heatmap to a $G\times G$ grid and rank regions by mean attribution using Gumbel-Sinkhorn soft sorting with 30 Sinkhorn iterations. During training, G is sampled uniformly from $\{7, \ldots, 28\}$ to reduce grid-alignment bias (region sizes 32–8 px for 224×224 inputs). The *Deletion* and *Insertion* curves are approximated using $S = 16$ uniformly spaced perturbation steps during training. We equally weight the corresponding losses $\lambda_{\text{del}} = \lambda_{\text{ins}} = 1$ and use $\lambda_{\text{reg}} = 2.5\times10^{-3}$ for the pixel regularization term to penalize isolated high-amplitude activations. The explainer is trained for one epoch with AdamW (lr 3×10^{-4}, weight decay 10^{-3}), gradient clipping (norm 1.0), and a one-cycle cosine schedule. In the optional refinement variant, we update only the explainer head for T steps (lr 10^{-4}) on the test image using the same loss, fixed $\tau{=}\tau_{\text{final}}$, and deterministic soft sorting (Gumbel noise disabled).

For evaluation we use the full corresponding validation split (50K images, 1,000 classes) without Test-Time Refinement ($T = 0$), unless explicitly denoted. We explain the *top-1 predicted class* regardless of the output probability, as well as the *ground truth class* if different, and report results separately for both targets. All metrics are averaged over three different types of reference images $\mathbf{I}_0$ used for the perturbation-based evaluations to replace masked-out pixels: a black image, a uniform gray image (ImageNet mean), and a blurred version of the input image. The detailed evaluation results for all metrics, individual reference images, and architectures are reported in the supplementary material.

4.2 Faithfulness Evaluation

Besides the *Deletion* and *Insertion* described in Section 3, we evaluate the faithfulness of the generated attribution maps also using the *Positive* & *Negative* perturbation evaluation [15]: across the evaluation set, the most-attributed regions (*Positive*) or the least-attributed regions (*Negative*) are progressively removed and the target model's top-1 accuracy is tracked as a function of the percentage of perturbed pixels. Low *Positive* AUC values and high *Negative* AUC values indicate that highly ranked regions capture important features for the target model, while low-ranked regions are non-informative. We also report the *Average Drop Percentage* (ADP) and the *Percentage of Increase in Confidence* (PIC) [13]: for each sample an element-wise product between the normalized attribution map ($\in [0, 1]$) and the input image is created to mask the background regions based

Table 1: Quantitative comparison of attribution methods on ImageNet using a ViT-B/16 classifier. All metrics are averaged over three reference image modes $\mathbf{I}_0$ (black, mean, blur). Top-3 results per metric are highlighted (dark to light).

	Method	Deletion↓	Insertion↑	Ins. - Del.↑	Positive↓	Negative↑	Neg. - Pos.↑	ADP↓	PIC↑	Runtime
Predicted Class	Grad-CAM [53]	0.2858	0.4982	0.2124	0.3346	0.5788	0.2442	84.3339	4.3577	0.014s
	IG [62]	0.1857	0.5544	0.3686	0.2248	0.6347	0.4099	22.9340	33.4353	0.251s
	Trans-Att. [69]	0.1654	0.5833	0.4178	0.2022	0.6654	0.4633	25.6994	30.8467	0.254s
	Bi-Att. [17]	0.1600	0.5958	0.4358	0.1961	0.6789	0.4828	28.4572	28.2614	0.254s
	TIS [25]	0.1420	0.6381	0.4961	0.1731	0.7222	0.5492	**11.2479**	**45.5658**	1.167s
	ViT-CX [67]	0.1750	0.5714	0.3964	0.2149	0.6478	0.4328	12.5472	43.4971	0.860s
	MDA [64]	0.1556	0.6112	0.4556	0.1933	0.6983	0.5050	48.6653	13.2691	13.645s
	LeGrad [10]	0.1666	0.5666	0.4000	0.2036	0.6461	0.4425	20.3850	34.3440	**0.006s**
	AHA (Ours)	0.1381	0.6112	0.4731	0.1714	0.6947	0.5233	15.8713	36.8246	0.016s
	AHA $T = 3$ (Ours)	**0.1215**	**0.6485**	**0.5271**	**0.1491**	**0.7378**	**0.5887**	12.4658	42.2698	0.547s
Ground Truth Class	Grad-CAM [53]	0.2690	0.4638	0.1948	0.3068	0.5244	0.2176	80.4883	7.7366	0.014s
	IG [62]	0.1731	0.5171	0.3440	0.2028	0.5765	0.3737	20.0260	38.5759	0.251s
	Trans-Att. [69]	0.1541	0.5428	0.3887	0.1819	0.6020	0.4200	22.5843	36.0353	0.254s
	Bi-Att. [17]	0.1492	0.5534	0.4042	0.1766	0.6120	0.4354	25.1370	33.5553	0.254s
	TIS [25]	0.1324	0.5966	0.4641	0.1553	0.6595	0.5042	**9.2014**	**50.9796**	1.167s
	ViT-CX [67]	0.1694	0.5231	0.3537	0.2024	0.5748	0.3724	15.2074	42.2452	0.860s
	MDA [64]	0.1461	0.5628	0.4168	0.1756	0.6216	0.4460	44.8870	18.2510	13.645s
	LeGrad [10]	0.1552	0.5280	0.3728	0.1833	0.5859	0.4027	17.6998	39.4492	**0.006s**
	AHA (Ours)	0.1280	0.5657	0.4377	0.1529	0.6222	0.4693	14.0053	41.2158	0.016s
	AHA $T = 3$ (Ours)	**0.1126**	**0.6003**	**0.4877**	**0.1326**	**0.6630**	**0.5303**	10.8509	46.7211	0.547s

on the attribution map. The ADP is then computed as the average percentage drop in the target model's output score between the original image and the masked image. The lower the ADP, the better the attribution map differentiates between the unimportant background and the important foreground. The PIC, on the other hand, is defined as the percentage of cases in which this background-masking leads to an increase in the target model's confidence. Finally, we also report the differences between *Insertion* and *Deletion* AUCs, as well as between *Positive* and *Negative* AUCs, as a measure of the *Faithfulness* [43,52]. To demonstrate that the findings are not mere artefacts of the optimization objective, we provide additional evaluations on MAS [65], ROAD [51] metrics and the ImageNet segmentation benchmark [29] in the supplemental material, which also confirm the performance benefits of AHA.

Table 1 (ViT-B/16) and Table 2 (ViT-B/32) summarize the results of our method compared to a wide range of baselines. Across both analyzed classifiers and both target definitions, AHA without additional refinement steps (T=0) achieves competitive or superior performance to the current state-of-the-art, with particularly strong results on the *Deletion* and *Positive* metrics. We argue that the reason why our *zero-shot* method can compete with, or even outperform, per-sample-optimized methods is because it aligns attribution maps more precisely with the input image. This enables important areas to be localized more accurately than with methods that only operate at patch level. Using only three refinement steps ($T = 3$) at test time, AHA further improves on all metrics, outperforming the baselines in the majority of cases and achieving the best overall *Faithfulness* scores (*Ins.* - *Del.* and *Pos.* - *Neg.*).

Table 2: Quantitative comparison of attribution methods on ImageNet using a ViT-B/32 classifier. All metrics are averaged over three reference image modes $\mathbf{I}_0$ (black, mean, blur). Top-3 results per metric are highlighted (dark to light).

	Method	Deletion↓	Insertion↑	Ins. - Del.↑	Positive↓	Negative↑	Neg. - Pos.↑	ADP↓	PIC↑	Runtime
Predicted Class	Grad-CAM [53]	0.2337	0.5294	0.2956	0.2930	0.6328	0.3398	66.9314	8.5072	0.010s
	IG [62]	0.1897	0.5542	0.3645	0.2426	0.6567	0.4141	14.2831	39.0279	0.157s
	Trans-Att. [69]	0.2024	0.5507	0.3483	0.2567	0.6534	0.3967	17.1553	34.9780	0.157s
	Bi-Att. [17]	0.1887	0.5726	0.3839	0.2416	0.6782	0.4366	17.6613	34.3086	0.158s
	TIS [25]	0.1702	**0.6094**	0.4392	0.2144	0.7181	0.5037	9.7104	48.3937	0.416s
	ViT-CX [67]	0.1982	0.5604	0.3622	0.2525	0.6622	0.4098	9.3272	**48.9950**	0.763s
	MDA [64]	0.1973	0.5939	0.3966	0.2511	0.7043	0.4533	25.2806	26.1375	1.803s
	LeGrad [10]	0.1922	0.5502	0.3580	0.2457	0.6521	0.4065	13.4748	39.4919	**0.005s**
	AHA (Ours)	0.1471	0.5649	0.4178	0.1907	0.6703	0.4796	11.5998	39.7140	0.016s
	AHA $T = 3$ (Ours)	**0.1261**	0.6080	**0.4819**	**0.1600**	**0.7224**	**0.5624**	**7.8339**	47.8573	0.212s
Ground Truth Class	Grad-CAM [53]	0.2185	0.4931	0.2746	0.2647	0.5712	0.3065	63.3098	12.5057	0.010s
	IG [62]	0.1768	0.5157	0.3389	0.2181	0.5910	0.3729	12.7448	43.2325	0.157s
	Trans-Att. [69]	0.1886	0.5126	0.3240	0.2306	0.5883	0.3576	15.2576	39.6312	0.157s
	Bi-Att. [17]	0.1762	0.5316	0.3554	0.2177	0.6071	0.3894	15.7758	38.8666	0.158s
	TIS [25]	0.1589	**0.5691**	0.4102	0.1930	**0.6535**	0.4605	8.0883	**53.2509**	0.416s
	ViT-CX [67]	0.1924	0.5111	0.3187	0.2388	0.5806	0.3418	13.6714	45.4918	0.763s
	MDA [64]	0.1866	0.5443	0.3577	0.2298	0.6191	0.3893	24.5261	29.0981	1.803s
	LeGrad [10]	0.1792	0.5124	0.3331	0.2207	0.5876	0.3669	12.0416	43.7111	**0.005s**
	AHA (Ours)	0.1363	0.5250	0.3887	0.1694	0.6000	0.4306	10.4213	43.8113	0.016s
	AHA $T = 3$ (Ours)	**0.1171**	0.5650	**0.4480**	**0.1420**	0.6497	**0.5077**	**7.0080**	51.5807	0.212s

4.3 Qualitative Evaluation

Figure 1 shows qualitative examples of attribution maps generated by AHA compared to best-performing baselines across the evaluated metrics for the ViT-B/16 classifier. Even without test-time refinement ($T = 0$), AHA produces attribution maps that are more precisely aligned with the input image and better capture object boundaries than the baselines. This is particularly evident for the image classified as *impala*, where our attribution map correctly highlights the animal heads, while the baselines produce more diffuse, blocky maps that also cover the background because of the patch-level attribution. Moreover, in the case of the misclassified *can opener* our method clearly highlights the bottle caps as the most important features for the model's decision, while the baselines produce more noisy maps that are less interpretable. Additional examples for all methods and classifiers are included in the supplementary material.

4.4 Runtime Analysis

The runtime comparison in Tables 1 and 2, measured over 100 iterations on a single NVIDIA RTX 4090, confirms the efficiency advantage of the amortized design. At T=0, AHA requires only a single explainer forward pass and is therefore comparable to the runtime of propagation-based methods like Grad-CAM or LeGrad, while achieving competitive or superior faithfulness compared to methods that are orders of magnitude slower. When refining over T=3 gradient steps, the runtime increases depending on the complexity of the target model, but generally remains below the token/patch perturbation-based methods and significantly below the per-sample metric optimization of MDA. A detailed amortization study is provided in the supplemental material.

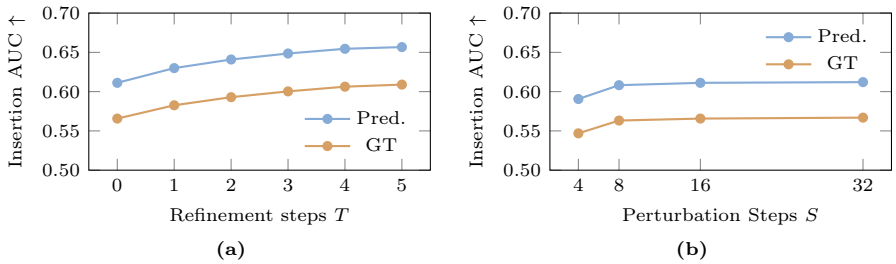

Fig. 3: Effect of (a) test-time refinement steps T and (b) training perturbation steps S on the Insertion AUC. Results are shown for the predicted class (Pred.) and the ground-truth class (GT) on the ImageNet validation set (ViT-B/16).

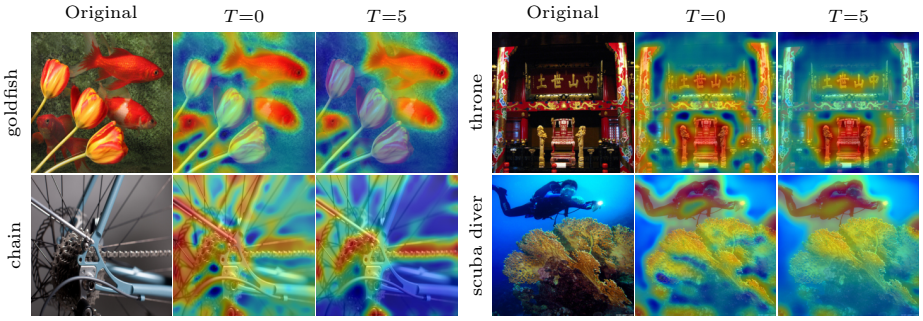

Fig. 4: Exemplary test-time refinement. Attribution maps produced by AHA without (T=0) and with (T=5) test-time refinement for the predicted class using ViT-B/16. Refinement progressively sharpens the heatmaps, concentrating attribution on the relevant object parts while suppressing diffuse background activations.

4.5 Test-Time Refinement Influence

We analyze the effect of applying T gradient steps of test-time refinement on the attribution quality. As shown in Figure 3a, performance improves monotonically with increasing T, with the largest gains concentrated in the first few steps and diminishing returns beyond T=3. Specifically, three refinement steps increase the Insertion AUC from 0.611 to 0.649 (predicted class). The Deletion AUC (where lower is better) follows a similar monotonically improving trend and is reduced from 0.138 to 0.121 for $T = 0 \rightarrow 3$. The qualitative examples in Figure 4 confirm this: refinement progressively sharpens the attribution maps, setting the focus onto the relevant regions and suppressing background activations most clearly visible for the *goldfish* and *scuba diver* examples. The number of refinement steps T thus provides a flexible compute-quality tradeoff, making AHA suitable both for high-throughput settings (T=0) and for scenarios where explanation fidelity is the primary interest.

4.6 Perturbation Steps Influence

During training, the *Deletion* and *Insertion* AUC objectives are approximated by sampling S uniformly spaced perturbation steps instead of evaluating the full sequence. We ablate the effect of this hyperparameter by training models with $S \in \{4, 8, 16, 32\}$. As shown in Figure 3b, the performance improves consistently from S=4 to S=16, after which gains become marginal. This plateau indicates that S=16 already provides a sufficiently dense approximation of the AUC curve.

4.7 Limitations

Despite the state-of-the-art results of our approach, few limitations should be acknowledged. First, each new target model requires a retraining of the explainer. While inference is efficient once trained, this amortization step introduces additional computational cost and limits immediate applicability to arbitrary models. Second, our training objective optimizes approximations of deletion and insertion metrics. While they are widely used to assess attribution faithfulness, they remain proxy measures of explanation quality and have known shortcomings, like sensitivity to the choice of reference image and perturbation strategy (detailed results provided in the supplemental). And third, our method requires access to a representative training dataset for the target model, which may be unavailable for proprietary or closed-source models. However, our optional refinement ($T > 0$) can be employed to resolve such data issues.

5 Conclusion

We introduced AHA, a metric-driven attribution framework that directly optimizes differentiable surrogates of the Deletion and Insertion metrics via a Gumbel–Sinkhorn relaxation of sorting. The proposed approach combines the causal grounding of perturbation-based methods with the efficiency of amortized explainers, producing dense pixel-level attribution maps in a single forward pass. Experiments on ImageNet with Vision Transformers demonstrate improved faithfulness metrics and sharper, boundary-aligned explanations compared to existing methods. These results highlight the potential of learning attribution through direct optimization of evaluation metrics for scalable and faithful model explainability.

Acknowledgements

This work received funding from the AI Ecosystems programme of the Republic of Austria, Federal Ministry of Innovation, Mobility and Infrastructure (BMIMI) under the project DOMINO (923575). The Austrian Research Promotion Agency (FFG) has been authorised for the programme management.

References

1. Abnar, S., Zuidema, W.: Quantifying Attention Flow in Transformers. In: ACL (2020)
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity Checks for Saliency Maps. In: NeurIPS (2018)
3. Alshami, E., Agnihotri, S., Schiele, B., Keuper, M.: AIM: Amending Inherent Interpretability via Self-Supervised Masking. In: ICCV (2025)
4. Anders, C.J., Weber, L., Neumann, D., Samek, W., Müller, K.R., Lapuschkin, S.: Finding and Removing Clever Hans: Using Explanation Methods to Debug and Improve Deep Models. Inf. Fusion **77**, 261–295 (Jan 2022)
5. Arya, S., Rao, S., Böhle, M., Schiele, B.: B-cosification: Transforming Deep Neural Networks to be Inherently Interpretable. In: NeurIPS (2024)
6. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE **10**(7), 1–46 (2015)
7. Bhalla, U., Srinivas, S., Lakkaraju, H.: Verifiable Feature Attributions: A Bridge between Post Hoc Explainability and Inherent Interpretability. In: ICML Workshops (2023)
8. Blondel, M., Teboul, O., Berthet, Q., Djolonga, J.: Fast differentiable sorting and ranking. In: ICML (2020)
9. Böhle, M., Fritz, M., Schiele, B.: B-Cos Networks: Alignment Is All We Need for Interpretability. In: CVPR (2022)
10. Bousselham, W., Boggust, A., Chaybouti, S., Strobelt, H., Kuehne, H.: LeGrad: An Explainability Method for Vision Transformers via Feature Formation Sensitivity. In: ICCV (2025)
11. Boyd, A., Bowyer, K.W., Czajka, A.: Human-Aided Saliency Maps Improve Generalization of Deep Learning. In: WACV (2022)
12. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. In: ICCV (2021)
13. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In: WACV (2018)
14. Chefer, H., Gur, S., Wolf, L.: Generic Attention-Model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. In: ICCV (2021)
15. Chefer, H., Gur, S., Wolf, L.: Transformer Interpretability Beyond Attention Visualization. In: CVPR (2021)
16. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This Looks Like That: Deep Learning for Interpretable Image Recognition. In: NeurIPS (2019)
17. Chen, J., Li, X., Yu, L., Dou, D., Xiong, H.: Beyond Intuition: Rethinking Token Attributions inside Transformers. TMLR (2023)
18. Chen, J., Song, L., Wainwright, M.J., Jordan, M.I.: Learning to explain: An information-theoretic perspective on model interpretation. In: ICML (2018)
19. Chen, R., Liang, S., Li, J., Liu, S., Li, M., Huang, Z., Zhang, H., Cao, X.: Interpreting Object-level Foundation Models via Visual Precision Search. In: CVPR (2025)
20. Chen, R., Zhang, H., Liang, S., Li, J., Cao, X.: Less is More: Fewer Interpretable Region via Submodular Subset Selection. In: ICLR (2024)
21. Covert, I., Kim, C., Lee, S.I., Zou, J., Hashimoto, T.: Stochastic Amortization: A Unified Approach to Accelerate Feature and Data Attribution. In: Advances in Neural Information Processing Systems (2024)

22. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009)
23. Donnelly, J., Barnett, A.J., Chen, C.: Deformable ProtoPNet: An Interpretable Image Classifier Using Deformable Prototypes. In: CVPR (2022)
24. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
25. Englebert, A., Stassin, S., Nanfack, G., Mahmoudi, S.A., Siebert, X., Cornu, O., De Vleeschouwer, C.: Explaining Through Transformer Input Sampling. In: ICCV Workshops (2023)
26. Fel, T., Ducoffe, M., Vigouroux, D., Cadène, R., Capelle, M., Nicodème, C., Serre, T.: Don't Lie to Me! Robust and Efficient Explainability With Verified Perturbation Analysis. In: CVPR (2023)
27. Fong, R., Patrick, M., Vedaldi, A.: Understanding Deep Networks via Extremal Perturbations and Smooth Masks. In: ICCV (2019)
28. Fong, R.C., Vedaldi, A.: Interpretable Explanations of Black Boxes by Meaningful Perturbation. In: ICCV (2017)
29. Guillaumin, M., Küttel, D., Ferrari, V.: ImageNet Auto-Annotation with Segmentation Propagation. IJCV **110**(3), 328—-348 (2014)
30. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked Autoencoders Are Scalable Vision Learners. In: CVPR (2022)
31. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR (2016)
32. Jethani, N., Sudarshan, M., Covert, I.C., Lee, S.I., Ranganath, R.: FastSHAP: Real-Time Shapley Value Estimation. In: ICLR (2022)
33. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: LayerCAM: Exploring Hierarchical Class Activation Maps for Localization. IEEE Transactions on Image Processing **30**, 5875–5888 (2021)
34. Jung, H., Oh, Y.: Towards Better Explanations of Class Activation Mapping. In: ICCV (2021)
35. Kapishnikov, A., Bolukbasi, T., Viegas, F., Terry, M.: XRAI: Better Attributions Through Regions. In: ICCV (2019)
36. Kapishnikov, A., Venugopalan, S., Avci, B., Wedin, B., Terry, M., Bolukbasi, T.: Guided Integrated Gradients: An Adaptive Path Method for Removing Noise. In: CVPR (2021)
37. Koh, Pang Wei and Nguyen, Thao and Tang, Yew Siang and Mussmann, Stephen and Pierson, Emma and Kim, Been and Liang, Percy: Concept Bottleneck Models. In: ICML (2020)
38. Li, O., Liu, H., Chen, C., Rudin, C.: Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions. In: AAAI (2018)
39. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: ICCV (2021)
40. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: CVPR (2022)
41. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: NeurIPS (2017)
42. Mena, G., Belanger, D., Linderman, S., Snoek, J.: Learning Latent Permutations with Gumbel-Sinkhorn Networks. In: ICLR (2018)

43. Muzellec, S., Fel, T., Boutin, V., Andéol, L., VanRullen, R., Serre, T.: Saliency strikes back: how filtering out high frequencies improves white-box explanations. In: ICML (2024)
44. Nauta, M., van Bree, R., Seifert, C.: Neural Prototype Trees for Interpretable Fine-Grained Image Recognition. In: CVPR (2021)
45. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free Concept Bottleneck Models. In: ICLR (2023)
46. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models. In: BMVC (2018)
47. Prillo, S., Eisenschlos, J.M.: SoftSort: A Continuous Relaxation for the argsort Operator. In: ICML (2020)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML (2021)
49. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision Transformers for Dense Prediction. In: ICCV (2021)
50. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: KDD (2016)
51. Rong, Y., Leemann, T., Borisov, V., Kasneci, G., Kasneci, E.: A Consistent and Efficient Evaluation Strategy for Attribution Methods. In: ICML (2022)
52. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the visualization of what a Deep Neural Network has learned. IEEE Transactions on Neural Networks and Learning Systems **28**(11), 2660–2673 (2016)
53. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In: ICCV (2017)
54. Shrikumar, A., Greenside, P., Kundaje, A.: Learning Important Features Through Propagating Activation Differences. In: ICML (2017)
55. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3. arXiv:2508.10104 (2025)
56. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034 (2013)
57. Sinkhorn, R., Knopp, P.: Concerning nonnegative matrices and doubly stochastic matrices. Pacific Journal of Mathematics **21**(2), 343–348 (1967)
58. Situ, X., Zukerman, I., Paris, C., Maruf, S., Haffari, G.: Learning to Explain: Generating Stable Explanations Fast". In: ACL (2021)
59. Sixt, L., Granz, M., Landgraf, T.: When explanations lie: why many modified BP attributions fail. In: ICML (2020)
60. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A.: Striving for Simplicity: The All Convolutional Net. In: ICLR (2015)
61. Srinivas, S., Fleuret, F.: Full-Gradient Representation for Neural Network Visualization. In: NeurIPS (2019)
62. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. In: ICML (2017)
63. Tran, M., Lahiani, A., Dicente Cid, Y., Boxberg, M., Lienemann, P., Matek, C., Wagner, S.J., Theis, F.J., Klaiman, E., Peng, T.: B-Cos Aligned Transformers Learn Human-Interpretable Features. In: MICCAI (2023)

64. Walker, C., Jha, S.K., Ewetz, R.: Metric-Driven Attributions for Vision Transformers. In: ICLR (2025)
65. Walker, C., Simon, D., Chen, K., Ewetz, R.: Attribution Quality Metrics with Magnitude Alignment. In: IJCAI (2024)
66. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: CVPR Workshops (2020)
67. Xie, W., Li, X.H., Cao, C.C., Zhang, N.L.: ViT-CX: Causal Explanation of Vision Transformers. In: IJCAI (2023)
68. Xu, S., Venugopalan, S., Sundararajan, M.: Attribution in Scale and Space. In: CVPR (2020)
69. Yuan, T., Li, X., Xiong, H., Cao, H., Dou, D.: Explaining Information Flow Inside Vision Transformers Using Markov Chain. In: NeurIPS Workshop XAI4Debugging (2021)
70. Zeiler, M.D., Fergus, R.: Visualizing and Understanding Convolutional Networks. In: ECCV (2014)
71. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-Down Neural Attention by Excitation Backprop. IJCV **126**(10), 1084–1102 (2018)
72. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization. In: CVPR (2016)
73. Zintgraf, L.M., Cohen, T.S., Adel, T., Welling, M.: Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. In: ICLR (2017)