Scalable Perturbation-Based Explanations via Tradeoff-Conditioned Smooth Masking

Mehdi Naouar
[1]*University of Freiburg,* [2]*CRIION*
Freiburg, Germany
naouarm@cs.uni-freiburg.de

Jens Rahnfeld
[1]*University of Freiburg*
Freiburg, Germany
rahnfelj@informatik.uni-freiburg.de

Yannick Vogt
[1]*University of Freiburg,* [2]*CRIION*
Freiburg, Germany
vogty@cs.uni-freiburg.de

Joschka Boedecker
[1]*University of Freiburg,* [2]*CRIION,*
[3]*IMBIT//BrainLinks-BrainTools*
Freiburg, Germany
jboedeck@informatik.uni-freiburg.de

Gabriel Kalweit
[2]*CRIION,* [1]*University of Freiburg*
Freiburg, Germany
gabriel.kalweit@criion.org

Maria Kalweit
[1]*University of Freiburg,* [2]*CRIION,*
[3]*IMBIT//BrainLinks-BrainTools*
Freiburg, Germany
kalweitm@cs.uni-freiburg.de

***Abstract*—Understanding and interpreting the predictions of deep image classifiers remains a fundamental challenge in explainable artificial intelligence. Among existing approaches, model-based removal methods train an auxiliary explainer to approximate a classifier's response to masked inputs and have recently achieved remarkable performance in attribution. Despite their effectiveness, these methods rely heavily on large-scale Monte Carlo sampling, which incurs significant computational cost and severely limits scalability as the perturbation space increases. In this work, we first analyze these limitations and show that, as the perturbation resolution grows, sampling-based removal methods exhibit degraded attribution quality. Motivated by this observation, we introduce a sampling-free, perturbation-based training framework based on continuous and differentiable masking. By replacing discrete perturbations with smooth masks, the proposed framework eliminates the need for sampling while preserving a causal interpretation grounded in feature suppression. Finally, we adopt a tradeoff-conditioned formulation that enables a single model to produce explanations at multiple levels of sparsity and granularity. Through experiments on several datasets and classifier architectures, we show that the proposed training strategy achieves competitive or superior attribution faithfulness compared to strong sampling-based baselines, while dramatically reducing computational cost and enabling substantially improved scalability.**

***Index Terms*—XAI, Explainable AI, Attribution methods**

I. Introduction

Deep neural networks have become the dominant paradigm for image classification, achieving state-of-the-art performance across a wide range of visual recognition tasks. Despite their success, the rationale behind their predictions often remains opaque, raising concerns regarding trust, accountability, and the detection of model failures. These challenges have intensified the need for methods that provide faithful and reliable explanations of a classifier's outputs. Feature attribution methods aim to address this issue by identifying which input regions most influence a model's output and are widely used to analyze model behavior, detect spurious correlations, and assess prediction validity [1].

Among attribution techniques, perturbation-based methods have gained attention for their strong empirical performance. They estimate feature importance by measuring how predictions change when input regions are removed or masked, aligning with an intuitive notion of causal influence and often producing more faithful explanations than purely gradient-based approaches. A prominent class of such methods relies on model-based removal, where an auxiliary explainer is trained to approximate the change in the classifier's prediction under masked inputs and infer per-region attribution scores [2], [3]. While these methods outperform classical baselines in attribution quality and robustness, they require sampling large numbers of masks during training, leading to high computational cost and poor scalability, as a fixed sampling budget covers only a small fraction of the combinatorial perturbation space at higher resolutions.

To analyze these limitations, we study how sampling-based removal methods scale with attribution resolution. As the resolution increases, mask sampling provides an increasingly poor approximation of the combinatorial space of possible removals. With a fixed sampling budget, the explainer observes only a small subset of mask configurations, leaving many feature interactions unexplored and degrading the reliability of the learned estimates. This results in a decline in attribution quality and reveals inherent scalability issues, particularly for high-resolution images or fine-grained partitions.

To overcome these limitations, we propose an end-to-end explainer training pipeline based on smooth masking that eliminates the need for mask sampling. The explainer learns a continuous-valued mask and applies smooth, differentiable perturbations, which are optimized directly in an end-to-end manner according to their impact on the classifier's predictions. This formulation enables all regions to be jointly considered during optimization. By removing Monte Carlo sampling, the proposed approach alleviates the computational and scalability issues inherent in sampling-based methods while preserving faithful attribution. Consequently, the ex-

plainer can be trained efficiently at pixel-level resolution without compromising explanation quality.

In summary, this work makes the following contributions:

- We analyze the scalability limitations of sampling-based removal methods and show that their attribution performance degrades as the perturbation space grows.
- We introduce a sampling-free, perturbation-based explainer learning framework that enables end-to-end training through a smooth masking objective, avoiding Monte Carlo sampling. In contrast to sampling-based methods, our approach operates at pixel resolution without requiring coarse superpixel partitions.
- We propose a tradeoff-conditioned formulation between faithfulness and sparsity that enables the visualization of features at different levels of importance.
- Through experiments on Pet [4] and ImageNette [5] and across three classifier backbones (ViT [6], ConvNeXt [7], and EfficientNetV2 [8]), we show that the proposed method achieves improved performance compared to strong perturbation-based baselines, and produces qualitatively coherent attribution maps.

II. Related Work

Research on explaining image classifiers has produced a wide spectrum of attribution methods that differ primarily in how relevance is defined and estimated. Broadly, existing approaches can be grouped into several families: *activation-based methods*, which derive saliency from internal feature maps (e.g. grad-CAM [9], finer-CAM [10]); *attention-based methods*, which interpret attention weights or their compositions as importance scores [11], [12]; *gradient-based methods*, which use local sensitivities of the output with respect to the input [13], [14]; *propagation-based methods*, which redistribute relevance through the network using conservation rules [15]; and *perturbation- or removal-based methods*, which measure relevance by observing how predictions change when parts of the input are suppressed or modified [2], [3], [16]–[18]. While most of these families rely on internal model signals or local approximations, perturbation-based methods are grounded in an explicit causal criterion: a feature is important if removing it changes the model's output. This conceptual clarity has made perturbation-based formulations particularly attractive, despite their higher computational cost.

A. Perturbation-based Attribution

Perturbation-based explainers treat the classifier as a black box and estimate feature importance by systematically suppressing subsets of the input and measuring the induced change in prediction. Early approaches rely on Monte Carlo sampling over the space of binary masks. RISE [16] samples random masks and correlates them with prediction scores to obtain pixel-wise relevance maps. Although conceptually simple, the exponential size of the mask space requires a large number of samples to achieve stable estimates.

To reduce this cost, several methods replace explicit sampling with learned approximations. FastSHAP [17] trains a parametric explainer to approximate Shapley values, amortizing the cost of mask sampling across data points. ViT-Shapley [3] adapts this framework to vision transformers, combining learning Shapley values with masked attention. Salvage [2] further extends the learning-based paradigm by explicitly modeling the classifier's behavior under feature removal. It formulates a probabilistic objective that aligns predictions on masked inputs with the original classifier outputs, and introduces an informed sampling strategy to improve sample-efficiency. Building on this idea, One For All [19] integrates prediction and attribution into a single training stage, yielding an explainable classifier.

B. Smooth Masking for Perturbation-Based Methods

A subset of perturbation-based methods relies on smooth masking, replacing hard feature removal with continuous perturbations. A seminal example is Meaningful Perturbation (MP) [20], which formulates explanation as the optimization of a continuous, spatially smooth mask for a single test image. MP balances prediction suppression with mask sparsity to identify regions whose perturbation most strongly affects the classifier's output. This introduced the idea of using differentiable masks as a surrogate for discrete feature removal.

Subsequent works adopt similar strategies but replace per-image optimization with a learned masking network. Finding NEM-U [21] trains a neural explainer to predict soft masks that highlight regions responsible for a learned representation in an unsupervised setting. The explainer is trained jointly with a representation model using a reconstruction-based objective under masked inputs, enabling explanations in a single forward pass. However, its objective emphasizes reconstruction fidelity rather than predictive behavior.

Our work follows this model-based perturbation paradigm but targets *supervised classification* and explicitly aligns explanations with the classifier's response to perturbations. Unlike Meaningful Perturbation, which optimizes a separate mask for each test sample, we train a single global masking model across training images using a loss defined directly on the classifier's output. In contrast to NEM-U, which explains latent representations via reconstruction, our approach optimizes explanations based on how the classifier's predictions change under smooth perturbations. In contrast to sampling-based approaches such as RISE or FastSHAP, our formulation operates in a continuous mask space, avoiding combinatorial enumeration and enabling end-to-end training via standard gradient-based optimization.

III. Sampling-based Explainer Models under the Curse of Dimensionality

Sampling-based explainer models estimate feature contributions by evaluating the classifier on masked variants of the input. Among recent approaches, ViT-Shapley [3] approximates Shapley values through repeated masked perturbations to estimate marginal contributions. Salvage [2] extends this paradigm by approximating the classifier's prediction distribution under masking, using a probabilistic training objective and informed

sampling to improve sample efficiency. Although these methods have demonstrated strong empirical performance, they fundamentally depend on sampling from an exponentially large perturbation space. For an input with n feature elements, the number of possible masks is 2^n, making exact enumeration infeasible and necessitating Monte Carlo approximations. As n increases (due to higher spatial resolution or finer partitioning) the variance of these estimators rises, and achieving stable training requires substantially more samples. This dependence introduces two core limitations. First, the computational cost scales with the number of sampled masks, directly impacting training time and memory usage. Second, when the sampling budget is fixed, the fraction of the perturbation space it covers decreases rapidly with dimensionality, leading to poorer approximation quality, noisier importance estimates, and reduced stability of the trained explainer.

We quantify these effects by evaluating sampling efficiency across different perturbation dimensionalities using SRG [22], which evaluates the faithfulness of feature rankings, and its relative counterpart R-SRG [2], which additionally accounts for relative attribution magnitudes. Our results presented in Figure 1 show that sampling-based removal methods exhibit diminishing returns as the feature space expands. In particular, while the performance of ViT-Shapley collapses sharply with increasing resolution, Salvage remains relatively stable (e.g. for ImageNette, SRG: 44.80 $\rightarrow$ 40.35 $\rightarrow$ 37.04; R-SRG: 25.63 $\rightarrow$ 21.70 $\rightarrow$ 18.90). This relative robustness can be attributed to the informed sampling strategy used by Salvage to increase its sampling efficiency during training. These observations motivate the need for explainer models that avoid sampling altogether.

IV. Tradeoff-Conditioned Smooth Masking for End-to-End Explainer Training

The limitations of sampling-based removal methods motivate the development of an explainer model that does not rely on Monte Carlo sampling. To this end, we introduce a sampling-free, perturbation-based explainer optimization strategy, Tradeoff-Conditioned Smooth Masking (TCSM), which learns a continuous mask over input image pixels and optimizes it directly through an end-to-end objective. An overview of the proposed approach is illustrated in Figure 2.

A. Continuous Masking Mechanism

To avoid discrete perturbations and repeated mask sampling, the explainer predicts a continuous mask $m \in [0,1]^{H\times W}$, where H and W denote the spatial dimensions of the input image $x \in \mathbb{R}^{H\times W\times 3}$. This mask softly modulates the input by interpolating between the original image content and noise:

$$\tilde{x}(m) = m \odot x + (1-m) \odot \eta, \tag{1}$$

where $\eta \in \mathbb{R}^{H\times W\times 3}$ denotes RGB noise, with i.i.d. components drawn from a uniform distribution, and $\odot$ denotes element-wise multiplication. This smooth formulation provides a differentiable approximation of region suppression, allowing gradients to flow directly through the masking operation.

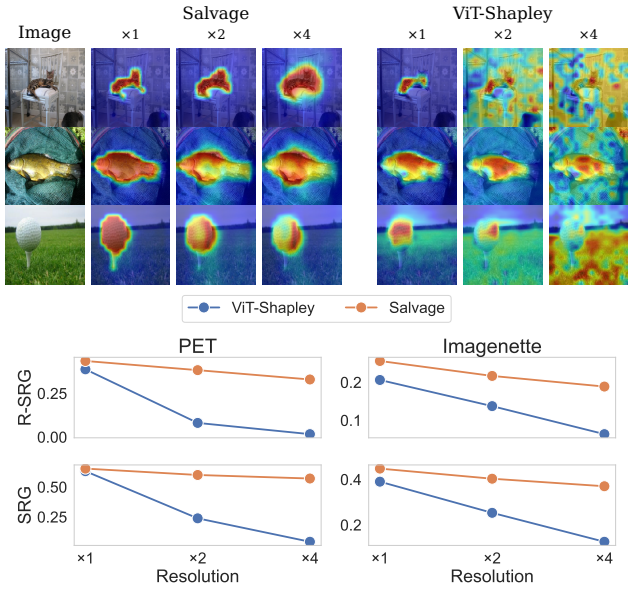

Fig. 1. Impact of perturbation resolution on sampling-based explainer performance. Resolution $\times 1$ corresponds to the original setting used in ViT-Shapley and Salvage, where images are partitioned into superpixels of size 16×16 pixels. Increasing the resolution to $\times 2$ and $\times 4$ reduces the superpixel size to 8×8 and 4×4 pixels, respectively, thereby increasing the dimensionality of the perturbation space. Performance is reported on ImageNette and Pet using SRG [22] and R-SRG [2]. As the resolution increases, the performance of both ViT-Shapley and Salvage consistently degrades, illustrating the loss of sampling efficiency and the effect of the curse of dimensionality on sampling-based removal methods.

B. End-to-End Explanation via Smooth Masking

Building on this continuous masking formulation, the explainer is trained end-to-end to identify regions that are necessary and sufficient for the classifier's prediction. Specifically, the predicted mask is optimized such that the classifier output remains high when informative regions are preserved and low when those regions are suppressed. Based on the continuous masking mechanism above, we now formulate the training objective that guides the explainer toward faithful and compact explanations. Let $p(\cdot)$ denote the classifier's probability of the ground-truth class. We define two complementary loss terms:

$$L^+ = -\log\, p(\tilde{x}(m))\,, \tag{2}$$

$$L^- = -\log(1 - p(\tilde{x}(1-m)))\,, \tag{3}$$

where L^+ promotes *sufficiency* by encouraging the selected region $\tilde{x}(m)$ to be sufficient for correct prediction, while L^- promotes *necessity* by enforcing that the remaining input $\tilde{x}(1-m)$ is insufficient to support the prediction. To avoid trivial solutions that retain nearly the entire image, and to promote compact explanations, we introduce an L_1 regularization term:

$$R = \|m\|_1. \tag{4}$$

The final training objective combines these components:

$$L = \lambda\big(L^+ + L^-\big) + (1-\lambda)R, \tag{5}$$

where $\lambda \in [0,1]$ controls the tradeoff between predictive consistency and regularization.

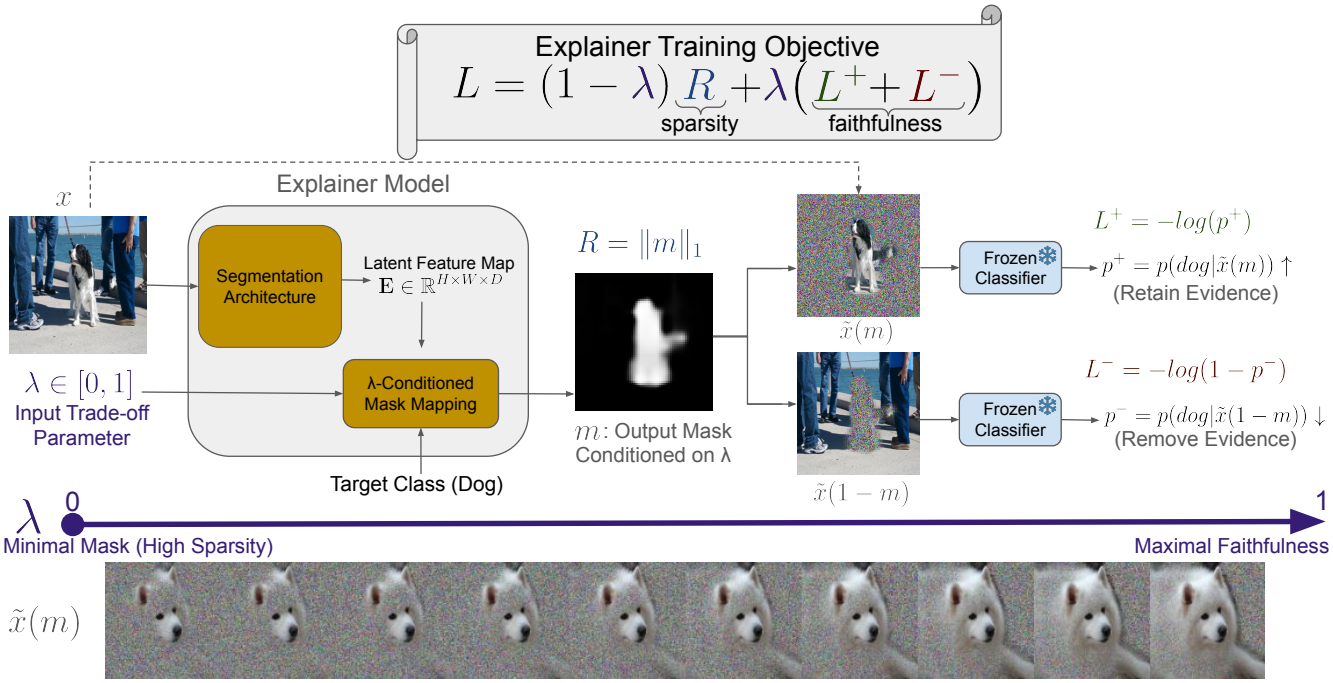

Fig. 2. Overview of TCSM: The explainer predicts a continuous mask $m \in [0,1]^{H\times W}$ using a segmentation backbone that produces a latent feature map $E \in \mathbb{R}^{H\times W\times D}$, followed by a λ-conditioned mask mapping head. The tradeoff parameter $\lambda \in [0,1]$ controls the balance between sparsity and faithfulness in the training objective L, where L^+ enforces evidence preservation and L^- enforces evidence removal. The mask is applied by smooth interpolation between the input image x and noise, yielding $\tilde{x}(m)$ and $\tilde{x}(1-m)$. By conditioning on λ, a single explainer learns a continuum of explanations ranging from minimal, highly sparse masks ($\lambda \approx 0$) to maximally faithful masks ($\lambda \approx 1$), illustrated at the bottom.

C. Tradeoff Conditioning for Continuous Importance Estimation

The training objective in Equation (5) defines an explicit tradeoff between predictive consistency and sparsity. However, the appropriate balance between these two objectives is inherently ambiguous and depends on the visual structure of the input. Objects of interest can vary significantly in size, shape, and contextual dependence, and different instances may require different amounts of evidence to explain a prediction.

Fixing λ therefore enforces a single, global notion of sparsity across all inputs, which may be suboptimal for images with different levels of visual complexity. This motivates learning a family of solutions spanning the entire tradeoff range, rather than committing to a single operating point. By conditioning the explainer on λ, we allow the model to adapt the level of detail in the explanation to the characteristics of each input, and to represent relevance at multiple levels of granularity.

1) Conditioning on the Tradeoff Parameter: To address these limitations, we condition the explainer model on the tradeoff parameter itself. Concretely, the scalar $\lambda \in [0,1]$ is provided as an additional input to the explainer alongside the image representation. During training, for each image x_i, a separate value λ_i is independently sampled from a uniform distribution, $\lambda_i \sim \mathcal{U}(0,1)$. The sampled tradeoffs are applied to each sample to the training objective in Equation (5) and are provided as an additional input to the explainer at the same time. The explainer is therefore trained to adapt its predicted mask to the specific tradeoff value associated with each training sample. This conditioning encourages the model to learn a continuous mapping from the tradeoff parameter λ to its corresponding mask, enabling it to approximate optimal explanations across the entire tradeoff range within a single model.

2) Aggregation at Inference Time: At test time, the conditioned explainer is inferred over a sequence of tradeoff values. Given a set $\{\lambda_i\}_{i=1}^T$ sampled from a linear grid over $[0,1]$, we obtain a corresponding family of masks:

$$m_i = f_\theta(x, \lambda_i), \tag{6}$$

where f_θ denotes the explainer model. Rather than selecting a single operating point, we aggregate the masks across tradeoffs:

$$\hat{m} = \frac{1}{T}\sum_{i=1}^{T} m_i. \tag{7}$$

This integration yields a single attribution map corresponding to the expectation of the mask over λ, combining information from multiple sparsity levels and producing continuous importance scores. By pooling explanations across the entire tradeoff range, the final attribution becomes less sensitive to any specific choice of λ and more robust to variations in object scale and visual complexity. At the same time, the intermediate masks $\{m_i\}$ remain accessible at inference,

enabling inspection of relevance at different levels of detail when desired.

3) Efficient Conditioning Integration: As discussed above, the aggregated attribution map $\hat{m}$ is obtained by integrating explanations across multiple tradeoff values, which in principle requires multiple forward passes of the explainer—one per conditioning value λ. A naive implementation would incur a substantial computational overhead at inference time. To mitigate this cost, we adopt an efficient conditioning strategy by injecting the tradeoff parameter exclusively at the final mapping layer of the segmentation network using a Feature-wise Linear Modulation (FiLM) mechanism [23], as described in Algorithm 1. In this design, the backbone feature extraction and intermediate processing are performed once for a given input x, producing a shared latent representation E. The conditioning variable λ is then mapped to a pair of scaling and shifting parameters $(\gamma(\lambda), \beta(\lambda))$, which modulate the prediction mapping through learned affine feature-wise transformations. As a result, aggregation over multiple tradeoffs requires recomputing only the lightweight FiLM-modulated final mapping, while the computationally expensive components of the network are reused. This design enables efficient integration over tradeoffs at inference time, making the proposed aggregation scheme practical without compromising the expressiveness of the conditioned explainer.

Algorithm 1 λ-Conditioned Mask Mapping

Inputs:
$E \in \mathbb{R}^{H \times W \times D}$: latent feature map
$\lambda \in [0, 1]$: input tradeoff factor
$c \in \{1, ...C\}$: Target class ▷ C: the number of classes
Model Parameters:
$emb \in \mathbb{R}^{C \times D}$: learned vector embedding
$\gamma : \mathbb{R}^1 \to \mathbb{R}^D$: learned affine mapping
$\beta : \mathbb{R}^1 \to \mathbb{R}^D$: learned affine mapping
Return:
$m \in \mathbb{R}^{H \times W}$: output mask conditioned on λ

$cond\_mapping \leftarrow emb \odot \gamma(\lambda) + \beta(\lambda)$ ▷ $\odot$: Hadamard product
▷ γ and β are broad-casted over class dimension C
$cond\_mapping \leftarrow \frac{cond\_mapping}{\|cond\_mapping\|_2}$ ▷ L2-normalization over D
$mask\_logits \leftarrow E \cdot cond\_mapping^\top$ ▷ dot product over feature dimension D
$all\_masks \leftarrow \sigma(mask\_logits)$ ▷ σ: sigmoid mapping
$m \leftarrow all\_masks[c]$ ▷ indexing the target class
return m

V. Results

A. Experimental setup

In our setting, we compared our method to Integrated Gradients [14], Finer-CAM [10], RISE [16], Mask [20], ViT-Shapley [3], and Salvage [2]. All classifier models are trained for 25 epochs and subsequently fine-tuned for an additional 50 epochs on masked images, where the masked regions are replaced with uniformly sampled image noise.

We consider three backbone architectures: ViT [6], ConvNeXt [7], and EfficientNetV2 [8]. The ViT and ConvNeXt classifiers are initialized with weights pretrained using DINOv3 [24], while EfficientNetV2 is initialized from ImageNet [25]-pretrained weights. All three classifiers are trained using the same optimization schedule and hyperparameters. For all model-based methods, we adopt the same explainer architecture used in ViT-Shapley and Salvage [2], consisting of a ViT encoder followed by an additional transformer block acting as a decoder. This choice is made solely to ensure a fair comparison with prior work by evaluating all methods in their original architectural setting and isolating performance differences to the explanation mechanisms rather than to differences in network design. Salvage and ViT-Shapley are trained using 32 sampled masks per image for approximately 18k optimization steps, corresponding to 50 epochs on ImageNette [5] and 100 epochs on Pet [4]. All models are trained with a batch size of 64 using the AdamW optimizer [26], a learning rate of 10^{-5}, and a weight decay of 10^{-5}. The only exception is the explainer model of ViT-Shapley, trained with a learning rate of 10^{-4}, following the original setting.

B. Quantitative Results

a) Evaluation Metrics: MIF (Most Influential First) and *LIF (Least Influential First)* (also known as *Insertion* and *Deletion* [16]) measure explanation quality by computing the area under the confidence curve obtained by progressively inserting features into a baseline input or deleting them from the original input, following the attribution ranking. Note that these metrics depend only on the ordering of attribution scores. SRG [22] combines both metrics through their difference, defined as SRG = MIF − LIF, and has been shown to yield more consistent rankings across different masking strategies [22]. To additionally account for the relative magnitudes of attribution scores, R-SRG replaces the deterministic ranking with a weighted sampling scheme proportional to the attribution values [2]. Denoting the corresponding scores by R-MIF and R-LIF, it is defined as R-SRG = R-MIF − R-LIF.

b) Results and Discussion: We report quantitative results on Pet and ImageNette using SRG and R-SRG for three backbone classifiers (ViT, ConvNeXt, and EfficientNetV2). Results are summarized in Table I.

Across all datasets and backbones, TCSM consistently achieves the highest R-SRG scores, indicating substantially improved faithfulness in terms of relative attribution magnitudes. For example, with a ViT backbone, TCSM improves R-SRG from 43.59 to 57.53 on Pet and from 25.63 to 36.25 on ImageNette. Similar gains are observed for ConvNeXt (37.46 → 54.98 on Pet) and EfficientNetV2 (36.14 → 54.74 on Pet), demonstrating robustness across architectures. In terms of SRG, TCSM also attains the best or second-best performance in all settings. While Salvage slightly outperforms TCSM on ViT–Pet, TCSM surpasses all competing methods on Ima-

TABLE I

QUANTITATIVE COMPARISON ON PET AND IMAGENETTE ACROSS THREE BACKBONE CLASSIFIERS (VIT, CONVNEXT, AND EFFICIENTNETV2). FAITHFULNESS IS MEASURED USING SRG AND R-SRG; INFERENCE TIME AND PEAK MEMORY USAGE ARE REPORTED FOR REFERENCE.

	ViT				Convnext				Effnetv2				Inference	
	Pet		ImageNette		Pet		ImageNette		Pet		ImageNette		Speed	Memory
Method	SRG	R-SRG	SRG	R-SRG	SRG	R-SRG	SRG	R-SRG	SRG	R-SRG	SRG	R-SRG	(s/it)	(MiB)
Int. Grad	49.09	30.66	24.08	14.06	43.82	22.74	17.27	8.16	44.12	25.59	22.33	11.95	10.52	1308
Fin. Cam	20.65	20.91	14.79	13.99	16.23	18.44	6.13	11.61	52.39	46.53	33.51	33.44	0.0601	1724
RISE	48.33	14.06	15.36	3.93	38.02	9.77	8.26	1.83	49.09	16.09	19.12	5.45	2.91	1138
Mask	57.54	16.67	28.24	7.69	46.67	10.48	11.52	2.61	60.73	11.03	32.40	5.93	25.66	**958**
ViT-Shap	63.26	38.81	39.05	20.63	53.05	25.71	19.31	11.18	54.62	28.70	37.82	15.61	0.0345	1136
Salvage	**65.44**	43.59	44.80	25.63	63.34	37.46	37.86	14.79	63.36	36.14	46.18	18.19	0.0182	1136
TCSM	64.96	**57.53**	**50.42**	**36.25**	**64.00**	**54.98**	**44.73**	**28.70**	**64.49**	**54.74**	**54.09**	**41.23**	**0.0181**	1172

geNette and consistently dominates under R-SRG, which is more sensitive to relative importance structure.

ViT-Shapley and Salvage achieve strong SRG values, confirming that sampling-based explainers can recover meaningful rankings. However, their substantially lower R-SRG scores reveal that the relative importance structure of the explanations is poorly preserved (as illustrated in Figure 3). This gap is especially pronounced on ImageNette, where R-SRG drops below 20 for ViT-Shapley across all backbones. By contrast, TCSM maintains high R-SRG values while also matching or exceeding the SRG performance of sampling-based methods, showing that it preserves both the ranking and the relative differences between features.

Integrated Gradients, RISE, Finer-CAM, and Mask perform significantly worse on both metrics. Although Mask achieves relatively high SRG, its very low R-SRG indicates that it fails to preserve relative attribution magnitudes.

C. Runtime Analysis

a) Inference efficiency: TCSM matches the inference efficiency of Salvage while remaining substantially faster than most perturbation-based baselines. It achieves an inference time of 0.0181 s/it, comparable to Salvage and faster than ViT-Shapley, with moderate memory usage. This demonstrates that the proposed sampling-free formulation improves explanation faithfulness without introducing additional computational overhead at inference time.

b) Training cost: We further investigate the training time and memory consumption of ViT-Shapley, Salvage, and TCSM under identical batch and hardware configurations, as reported in Table II. TCSM reduces the time per training iteration by more than a factor of $\times 3.5$ compared to both ViT-Shapley and Salvage (0.79 s vs. $\sim$2.9 s), and decreases peak GPU memory usage by nearly half. This gain is primarily explained by the elimination of multiple forward passes through the classifier induced by Monte Carlo sampling in existing removal-based methods. As a result, the proposed training framework significantly improves computational efficiency and makes removal-based attribution practical at higher perturbation resolutions.

TABLE II

TRAINING TIME AND MEMORY CONSUMPTION FOR VIT-SHAPLEY, SALVAGE, AND TCSM UNDER IDENTICAL TRAINING SETTINGS.

Metric	**ViT-Shap**	**Salvage**	**TCSM**
Time per training iteration (s)	2.97	2.90	**0.79**
Peak GPU memory (MiB)	40070	40192	**21668**
Total training time (h)	8.25	8.05	**2.19**

D. Ablation study: Optimization framework

In this subsection, we analyze the contribution of each component of the training objective of our method by selectively disabling individual terms. Specifically, we evaluate all combinations of the regularization term R, the evidence-preservation loss L^+, and the evidence-removal loss L^-. Results are reported in Table III using SRG and R-SRG on Pet and ImageNette on a ViT classifier.

a) Effect of asymmetric objectives: Using only one of the two predictive terms leads to suboptimal behavior. When L^+ is removed, the explainer is only encouraged to suppress evidence through L^-, which results in overly dense masks and weaker explanations. Conversely, removing L^- forces the explainer to retain evidence but provides no constraint on suppressing spurious regions, which limits its discriminative power. Removing L^+ has a smaller impact on performance than removing L^-. This reflects the different roles of the two terms: L^+ encourages the explainer to identify a compact and highly informative subset of regions that is sufficient to support the prediction, whereas L^- enforces that the complementary regions contain no predictive evidence at all. In a removal-

TABLE III

ABLATION OF THE OPTIMIZATION OBJECTIVE. EACH ROW CORRESPONDS TO A VARIANT OF THE LOSS $L = \lambda(L^+ + L^-) + (1-\lambda)R$, WHERE INDIVIDUAL TERMS ARE SELECTIVELY ENABLED OR DISABLED. RESULTS ARE REPORTED USING SRG AND R-SRG ON PET AND IMAGENETTE.

Configuration			Pet		ImageNette	
R	L^+	L^-	SRG	R-SRG	SRG	R-SRG
✓	✓	✗	43.09	41.40	17.91	14.58
✓	✗	✓	62.65	55.77	49.02	33.59
✗	✓	✓	18.35	6.84	15.06	4.27
✓	✓	✓	**64.96**	**57.53**	**50.42**	**36.25**

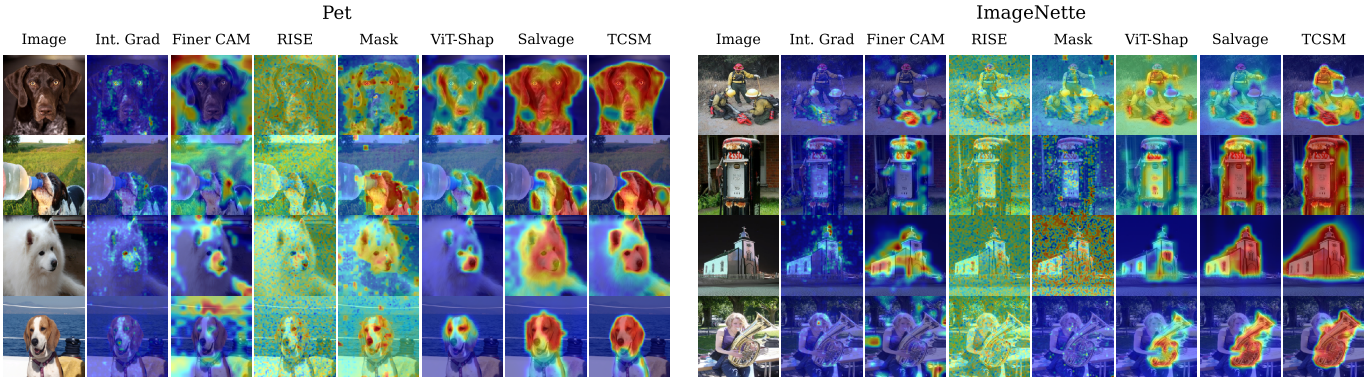

Fig. 3. Qualitative comparison of attribution maps on Pet (left) and ImageNette (right) for different attribution methods.

based setting, this completeness constraint is more critical, as it prevents explanations from focusing only on a minimal core while leaving other relevant features unaccounted for.

b) Importance of symmetric masking: Disabling the regularizer (R) while keeping both predictive terms results in a severe performance drop, particularly in R-SRG. This confirms that the sparsity constraint is essential to avoid degenerate solutions where the mask trivially preserves or removes large regions without meaningful structure.

c) Full objective: Combining all three terms yields the best performance on both datasets and metrics. This validates the proposed symmetric formulation, which jointly enforces evidence preservation, evidence removal, and mask compactness, producing faithful and well-calibrated explanations.

E. Qualitative Results

a) Comparison with existing methods: Figure 3 shows representative attribution maps on Pet and ImageNette for several explanation methods. Gradient-based approaches such as Integrated Gradients and Finer-CAM tend to produce diffuse or fragmented responses, often highlighting large background regions in addition to the object of interest. RISE and Mask generate more localized responses, but exhibit substantial noise and sensitivity to the masking process. Sampling-based methods, ViT-Shapley and Salvage, generally produce more coherent maps and capture the main object regions more consistently. TCSM yields attribution maps that are visually comparable to those of Salvage and ViT-Shapley. In many cases, the highlighted regions overlap with the primary object and suppress irrelevant background areas. While the qualitative differences between TCSM and Salvage are often subtle, TCSM exhibits slightly smoother and more spatially coherent responses in several examples.

b) Mask progression across tradeoffs: Figure 4 illustrates the evolution of the predicted masks for different values of the tradeoff parameter λ. For a fixed input, varying λ at inference time produces a continuous family of intermediate masks spanning different sparsity levels. Lower values of λ emphasize a small set of highly influential regions, whereas higher values progressively include additional features that contribute to the classifier's prediction.

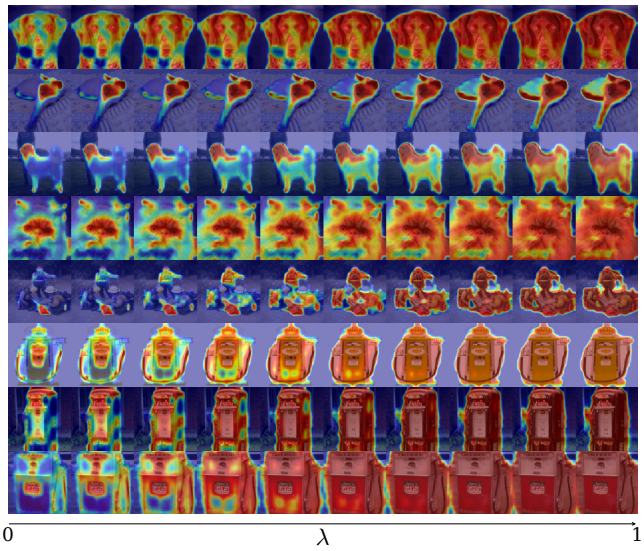

Fig. 4. Mask progression across tradeoff values. Each row shows the attribution maps produced by TCSM as the conditioning parameter λ is varied from 0 (left) to 1 (right). Small λ values yield sparse masks that highlight only the most critical regions, while larger λ values progressively include additional contextual regions.

While the final attribution used in our previous evaluations is obtained by aggregating masks across all tradeoffs, this visualization highlights that the conditioned explainer can also be queried at specific λ values to inspect features at different levels of importance. This enables a more fine-grained analysis of how evidence is distributed across the input.

VI. Conclusion

In this work, we investigated the scalability limitations of sampling-based removal explainers and showed that their reliance on Monte Carlo mask sampling leads to increased computational cost and degraded attribution quality as the dimensionality of the perturbation space grows. Through a controlled analysis of perturbation resolution, we demonstrated that even state-of-the-art sampling-based methods suffer from diminishing performance as the number of possible feature removals increases.

To address these challenges, we introduced TCSM, a sampling-free, perturbation-based explainer trained end-to-end via a smooth masking formulation. By replacing discrete perturbations with continuous, differentiable masks, the proposed approach avoids combinatorial mask sampling while remaining causally grounded in the classifier's response under feature suppression. This design enables efficient pixel-level training and eliminates the need for superpixel discretization or Monte Carlo approximations. We further proposed a tradeoff-conditioned formulation that models the balance between evidence preservation and compactness, allowing the explainer to represent relevance at multiple levels of granularity. This conditioning enables flexible inspection of explanations across different operating points within a single trained model.

Extensive experiments on two benchmark datasets and three backbone architectures showed that TCSM achieves competitive or improved faithfulness compared to strong sampling-based baselines, particularly when relative attribution magnitudes are considered. At the same time, the method substantially reduces training time and memory consumption, demonstrating improved scalability without sacrificing explanation quality. Overall, our results suggest that continuous, sampling-free formulations provide a promising direction for scalable and faithful perturbation-based explanation, and we hope this work encourages further exploration of differentiable masking strategies as an alternative to combinatorial sampling in explainable image classification.

ACKNOWLEDGMENT

This project was funded by the Mertelsmann Foundation. This work is part of BrainLinks-BrainTools, which is funded by the Federal Ministry of Economics, Science and Arts of Baden-Württemberg within the sustainability program for projects of the excellence initiative II. The authors used ChatGPT [27] to assist in improving the clarity and language of the manuscript. The AI tool was used solely for linguistic refinement, and all scientific content, analysis, and conclusions were developed by the authors.

REFERENCES

[1] J. Rahnfeld, M. Naouar, G. Kalweit, J. Boedecker, E. Dubruc, and M. Kalweit, "A comparative study of explainability methods for whole slide classification of lymph node metastases using vision transformers," *PLOS Digital Health*, vol. 4, no. 4, pp. 1–21, 04 2025. [Online]. Available: https://doi.org/10.1371/journal.pdig.0000792

[2] M. Naouar, H. Raum, J. Rahnfeld, Y. Vogt, J. Boedecker, G. Kalweit, and M. Kalweit, "Salvage: Shapley-distribution approximation learning via attribution guided exploration for explainable image classification," in *ICLR 2025*, 2025.

[3] I. C. Covert, C. Kim, and S. Lee, "Learning to estimate shapley values with vision transformers," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[4] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3498–3505.

[5] J. Howard and S. Gugger, "Fastai: A layered api for deep learning," *Information*, vol. 11, no. 2, p. 108, Feb. 2020.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[7] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *CVPR 2022*. IEEE, 2022, pp. 11 966–11 976.

[8] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," in *ICML 2021*, 2021, pp. 10 096–10 106.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, Oct. 2019.

[10] Z. Zhang, J. Gu, A. Chowdhury, Z. Mai, D. Carlyn, T. Y. Berger-Wolf, Y. Su, and W. Chao, "Finer-cam: Spotting the difference reveals finer details for visual explanation," in *CVPR 2025*. Computer Vision Foundation / IEEE, 2025, pp. 9611–9620.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[12] S. Abnar and W. H. Zuidema, "Quantifying attention flow in transformers," in *ACL 2020*, 2020, pp. 4190–4197.

[13] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *ICLR 2014, Workshop Track Proceedings*, 2014.

[14] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *ICML 2017*, 2017, pp. 3319–3328.

[15] S. Lapuschkin, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, p. e0130140, 07 2015.

[16] V. Petsiuk, A. Das, and K. Saenko, "RISE: randomized input sampling for explanation of black-box models," in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 151.

[17] N. Jethani, M. Sudarshan, I. C. Covert, S. Lee, and R. Ranganath, "Fastshap: Real-time shapley value estimation," in *ICLR 2022*, 2022.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[19] M. Naouar, Y. Vogt, J. Boedecker, G. Kalweit, and M. Kalweit, " One For All: A Unified Approach to Classification and Self-Explanation ," in *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, vol. LNCS 15973. Springer Nature Switzerland, September 2025.

[20] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 3449–3457.

[21] B. L. Møller, C. Igel, K. K. Wickstrøm, J. Sporring, R. Jenssen, and B. Ibragimov, "Finding NEM-U: explaining unsupervised representation learning through neural network generated explanation masks," in *ICML 2024*, 2024.

[22] S. Bluecher, J. Vielhaben, and N. Strodthoff, "Decoupling pixel flipping and occlusion strategy for consistent XAI benchmarks," *Trans. Mach. Learn. Res.*, vol. 2024, 2024.

[23] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press, 2018, pp. 3942–3951.

[24] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. E. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski, "Dinov3," *CoRR*, vol. abs/2508.10104, 2025.

[25] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR 2009*. IEEE Computer Society, 2009, pp. 248–255.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR 2019*, 2019.

[27] OpenAI, "ChatGPT (January 2026 version)," https://chat.openai.com/, 2026.