WHAT YOU SEE IS NOT WHAT MATTERS: ATTACKING VLM EXPLANATION FAITHFULNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual explanations are increasingly used to understand the decisions of Vision-Language Models (VLMs), yet their robustness under adversarial perturbations remains poorly understood. Existing studies mainly focus on saliency changes, but a substantial explanation shift does not necessarily imply unfaithfulness. We investigate whether prediction-preserving perturbations can redirect saliency toward regions that provide weaker support for the model's prediction. To support this evaluation, we propose the **Faithfulness-Oriented Adversarial (FOA) Attack**, which uses explanation faithfulness as an optimization signal to stress-test highlighted regions. By detaching saliency and mask generation, FOA avoids backpropagation through the explainer while enabling faithfulness-oriented updates. Across five VLM explanation methods and three datasets, our evaluation shows that faithfulness can be substantially degraded while preserving the original prediction, with FOA consistently revealing stronger vulnerabilities than X-Shift, a state-of-the-art attack on VLM explanations that primarily targets saliency shifts rather than explanation faithfulness. FOA perturbations also transfer effectively across explainers, suggesting that these vulnerabilities extend across explanation methods. These results demonstrate that highlighted regions can become less representative of the evidence supporting a prediction even when the prediction is unchanged. Our findings show that explanation robustness should be assessed not only by **where explanations shift**, but also by **whether the resulting explanations remain faithful**.

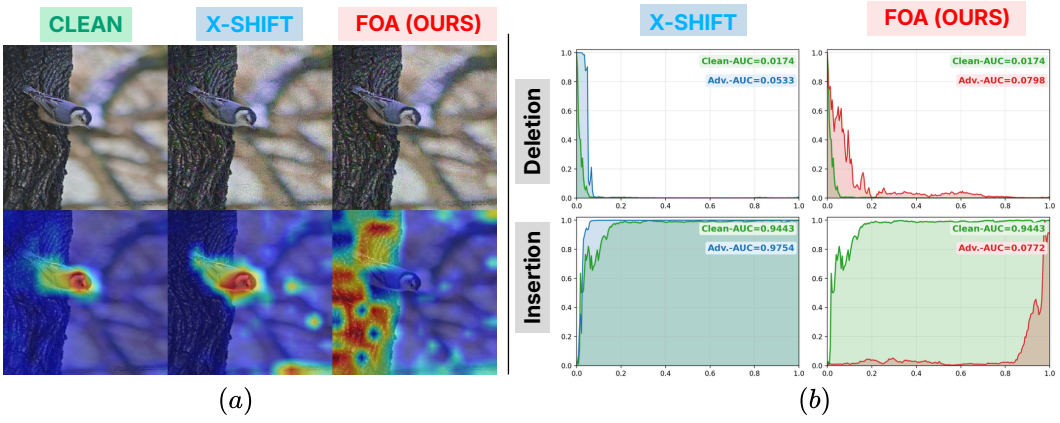

Figure 1: Faithfulness evaluation of GradECLIP (Zhao et al., 2024) explanations under Clean, X-Shift (Babadi & Karimipour, 2026), and our proposed Faithfulness-Oriented Adversarial (FOA) attack. All samples correspond to the text *"A photo of White breasted Nuthatch"* and retain the prediction *"White breasted Nuthatch"*. **a)** Clean and adversarial images (top) and their corresponding explanations (bottom). **b)** Faithfulness evaluation using Deletion and Insertion (Petsiuk et al., 2018). Higher Deletion AUC and lower Insertion AUC indicate lower explanation faithfulness.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Recent explanation methods for Vision-Language Models (VLMs) (Zhao et al., 2024; Wang et al., 2023; Selvaraju et al., 2017; Chefer et al., 2021; Petsiuk et al., 2018; Dong et al., 2023) have demonstrated high explanation faithfulness by highlighting image regions that contribute to the predicted image-text similarity. To quantitatively assess explanation quality, *faithfulness metrics* (Petsiuk et al., 2018), particularly Deletion and Insertion, have become a popular evaluation protocol in the literature. These metrics measure how the model prediction changes as salient regions are progressively removed (*Deletion*) or restored (*Insertion*). Specifically, a faithful explanation should identify features that are truly responsible for the model's decision. Consequently, during the Deletion process, progressively removing the most salient regions should cause the prediction confidence (e.g., logit or probability) to decrease rapidly. In contrast, during the Insertion process, progressively restoring these salient regions should lead to a rapid increase in the prediction confidence.

Recent studies have shown that visual explanations can be highly sensitive to input adversarial perturbations. In particular, adversarial attacks can substantially alter saliency maps while preserving the model's original prediction (Ghorbani et al., 2019; Dombrowski et al., 2019; Babadi & Karimipour, 2026). Such findings raise concerns about the trustworthiness of VLM explanations. However, a change in saliency alone does not necessarily imply that the resulting explanation is misleading. A shifted explanation may still highlight another set of image regions that are genuinely important for the model's prediction. In this case, **although the saliency map has changed substantially, the explanation may remain faithful**, as presented in Figure 1. Therefore, explanation instability and explanation unfaithfulness are fundamentally different phenomena: the former concerns whether the highlighted regions change, whereas the latter concerns whether the newly highlighted regions remain relevant to the model's prediction. This distinction motivates a more fundamental question: *Can adversarial perturbations shift VLM saliency toward regions with weak contribution while preserving the model's original prediction?*

To investigate this question, we propose the **Faithfulness-Oriented Adversarial (FOA) Attack**, a framework that leverages faithfulness information as an optimization signal to induce misleading shifts in VLM saliency. Rather than directly forcing explanations toward a predefined visual pattern, FOA optimizes adversarial perturbations to redirect the explanation toward regions that contribute less to the model's prediction. Across five VLM explainers and three datasets, our method substantially reduces explanation faithfulness while inducing significant changes in saliency maps. FOA consistently outperforms X-Shift (Babadi & Karimipour, 2026), the state-of-the-art attack on VLM explanations, in terms of faithfulness degradation, robustness, and attack transferability across explainers. Our contributions are summarized as follows:

- **Adversarial study of faithfulness.** To the best of our knowledge, we present the first study to investigate the faithfulness of VLM explanations under adversarial perturbations. We highlight the distinction between saliency instability and unfaithfulness, and investigate whether prediction-preserving perturbations can shift saliency toward regions with weaker contributions.

- **Faithfulness-Oriented Adversarial (FOA) Attack.** We propose the **FOA Attack**, which leverages faithfulness as an optimization signal to induce prediction-preserving saliency shifts toward less contributive regions. Unlike approaches that require gradients to be propagated through the saliency generation process and the explainer, FOA avoids back-propagation through these components, which may involve discrete or non-differentiable operations. Instead, we formulate a simpler gradient flow that directly optimizes the input perturbation based on faithfulness feedback, resulting in a more efficient and broadly applicable attack. Our method consistently outperforms X-Shift (Babadi & Karimipour, 2026) in terms of faithfulness degradation, robustness, and attack transferability across explainers.

For ease of reading, we present the structure of our paper as follows: **Section 2** reviews related work on VLM explanations and faithfulness; **Section 3** introduces the proposed FOA Attack and **Section 4** presents the experimental setup, results, and discussion of the findings.

2 Related Work

VLM Explanation Methods. Early studies on visual explanation investigated how individual input pixels or image regions contribute to a model's prediction by generating a heat map, commonly referred to as a *saliency map* (Baehrens et al., 2010; Simonyan et al., 2014; Shrikumar et al., 2017; Sundararajan et al., 2017; Selvaraju et al., 2017; Petsiuk et al., 2018). These methods were initially developed primarily for CNN-based vision models. With the introduction of the self-attention mechanism in Vision Transformers (ViTs) , subsequent studies have extended visual explanation to Transformer-based architectures (Qiang et al., 2022; Xie et al., 2023; Yu & Xiang, 2023; Chefer et al., 2021; Abnar & Zuidema, 2020). More recently, these explanation techniques (Selvaraju et al., 2017; Chefer et al., 2021; Petsiuk et al., 2018) have been adapted to Vision-Language Models (VLMs), particularly CLIP-based models (Radford et al., 2021), while other methods have been specifically designed for VLMs (Zhao et al., 2024; Dong et al., 2023). These approaches aim to identify the image regions that are most relevant to a given textual concept.

Adversarial Attack on Explanation. Existing studies have revealed multiple limitations of saliency maps produced by explanation methods, including failures under sanity checks, inconsistency across attribution techniques, and weak robustness to input perturbations (Hedström et al., 2024; Kindermans et al., 2019; Yeh et al., 2019). These findings suggest that explanation quality should be evaluated not only by visual plausibility but also by its stability under perturbations. They have shown that gradient-based attacks, such as PGD (Madry et al., 2018), can substantially suppress or relocate salient regions while preserving the model prediction (Ghorbani et al., 2019). Other studies further demonstrate that explanations can even be manipulated toward arbitrary target patterns through gradient-based optimization (Dombrowski et al., 2019). More recently, X-Shift (Babadi & Karimipour, 2026) has extended explanation manipulation to VLMs. Rather than directly optimizing the input image to alter saliency, X-Shift perturbs patch-level visual representations and encourages selected patch features to align with a target text representation. This manipulation shifts the patch-text similarity underlying VLM explanation maps, thereby redirecting the resulting saliency toward attacker-specified regions while preserving the original model prediction.

Although there have been studies investigating the robustness of visual explanations in both vision models and VLMs, existing efforts primarily focus on the instability or manipulability of explanation maps, while the faithfulness of these explanations under adversarial perturbations remains largely unexplored.

3 Faithfulness-Oriented Adversarial (FOA) Attack

3.1 Preliminaries

Given a Vision-Language Model (VLM) consisting of a vision encoder f_θ, a text encoder f_ϕ, an input image $x \in [0,1]^{H\times W\times 3}$, and a text prompt y (i.e., "a photo of a <class-name>"), the image-text similarity score is computed as $s(x,y) = \frac{f_\theta(x)^\top f_\phi(y)}{\|f_\theta(x)\|\,\|f_\phi(y)\|}$. For zero-shot classification with a set of class prompts $\mathcal{Y} = \{y_1, \ldots, y_C\}$, the probability of class c is computed by applying a softmax over the image-text similarity scores:

$$p(c,x) = \frac{\exp(s(x,y_c)/\tau)}{\sum_{k=1}^{C}\exp(s(x,y_k)/\tau)}, \tag{1}$$

where τ is the temparature parameter. From that, the predicted class is $\hat{y}(x) = \arg\max_c p(c,x)$. For convenience, we also denote the model output as the probability vector over all classes: $f(x) = [p(c,x)]_{c\in\mathcal{Y}}$.

The explanation method G for VLMs generates a saliency map with respect to the class c, $S = G(c,x) \in \mathbb{R}^{H\times W}$, where each element of S indicates the importance of the corresponding pixel for the prediction probability $p(c,x)$.

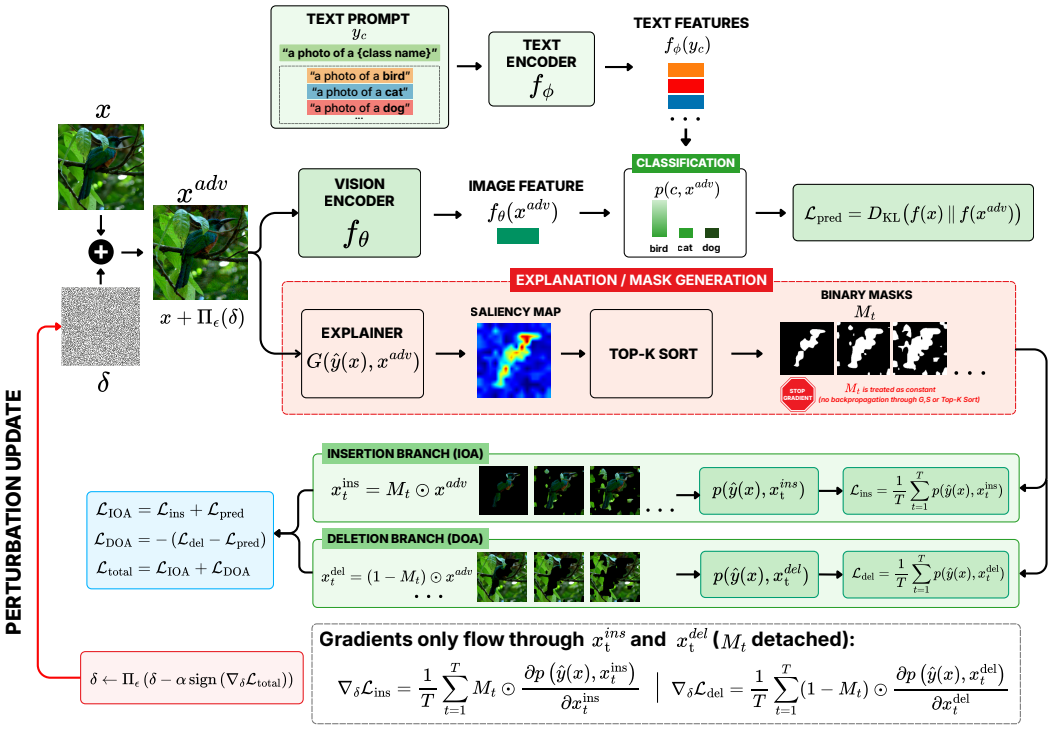

Figure 2: Overview of the proposed **Faithfulness-Oriented Adversarial (FOA)** Attack. FOA iteratively updates an ℓ_∞-bounded perturbation to degrade VLM explanation faithfulness while preserving the original prediction. At each iteration, the adversarial image is passed through an explanation method to obtain a saliency map, which is ranked to construct binary masks for the Insertion-Oriented Adversarial (IOA) and Deletion-Oriented Adversarial (DOA) objectives. The corresponding faithfulness losses are combined with a prediction-preserving loss to guide the perturbation update. The saliency-map and mask-generation branch is detached, allowing gradients to flow only through the insertion/deletion evaluation paths.

3.2 Our Attack

We propose **Faithfulness-Oriented Adversarial (FOA)** Attack, a framework that uses faithfulness as an optimization signal to generate adversarial examples that preserve the original prediction while shifting saliency toward regions with weaker contribution. FOA consists of three variants: **Insertion-Oriented Adversarial (IOA)** Attack, which encourages the newly highlighted regions to provide limited predictive support during the Insertion process; **Deleted-Oriented Adversarial (DOA)** Attack, which encourages the removal of highlighted regions to have limited impact on the prediction; and the combination of these two variants, considered as **IOA+DOA**, which leverages both the Insertion and Deletion processes to guide the attack.

Since three variants share the same paradigm, we first describe IOA in detail and then highlight the key modifications required for DOA and IOA+DOA variants. Figure 2 indicates the overview of FOA, while Algorithm 1 (Appendix A) presents the pseudocode of the devised approach.

IOA. The target of IOA is to generate an adversarial perturbation that minimizes the Insertion AUC while preserving the original prediction. The problem can be formulated as:

$$\delta^* = \arg\min_{\delta} \text{InsertionAUC}(x + \delta) \quad \text{s.t.} \quad \|\delta\|_\infty \leq \epsilon,\ \hat{y}(x + \delta) = \hat{y}(x). \tag{2}$$

To solve this optimization problem, the attack algorithm begins by initializing the adversarial perturbation as $\delta = \mathbf{0}$. Given an optimization budget of K iterations, the adversarial example at the k-th iteration is generated by $x^{adv} = x + \delta$. To preserve the original prediction, we minimize the Kullback-Leibler (KL) (Kullback & Leibler, 1951) divergence between the output distributions of the

clean and adversarial image, thereby raising the following objective function:

$$\mathcal{L}_{\text{pred}} = D_{\text{KL}}\big(f(x) \,\|\, f(x^{adv})\big)\,. \tag{3}$$

Next, we compute the saliency map $S = G(\hat{y}(x), x^{adv})$, where $\hat{y}(x)$ is the original predicted class. The saliency values are sorted in descending order to construct a sequence of binary masks $\mathcal{M} = \{M_t\}_{t=1}^{T}$, where T is the number of causal faithfulness evaluation steps. Each mask progressively reveals the most salient pixels. Accordingly, the reference image at step t is defined as

$$x_t^{\text{ins}} = M_t \odot x^{adv}. \tag{4}$$

Using the generated reference images, we define the insertion objective as

$$\mathcal{L}_{\text{ins}} = \frac{1}{T}\sum_{t=1}^{T} p(\hat{y}(x), x_t^{\text{ins}}), \tag{5}$$

which encourages the prediction confidence to be reduced throughout the insertion process. Therefore, the overall optimization objective for the IOA variant is

$$\mathcal{L}_{\text{IOA}} = \mathcal{L}_{\text{ins}} + \mathcal{L}_{\text{pred}}, \tag{6}$$

To minimize the objective presented in Equation 6, we update the perturbation as follows:

$$\delta \leftarrow \Pi_\epsilon\left(\delta - \alpha\,\text{sign}\left(\nabla_\delta \mathcal{L}_{\text{IOA}}\right)\right), \tag{7}$$

where $\Pi_\epsilon(\cdot)$ projects the perturbation onto the ℓ_∞ ball of radius ϵ.

Computing this update requires the gradient $\nabla_\delta \mathcal{L}_{\text{IOA}}$, which consists of two components: $\nabla_\delta \mathcal{L}_{\text{ins}}$ and $\nabla_\delta \mathcal{L}_{\text{pred}}$. The KL Divergence loss $\mathcal{L}_{\text{pred}}$ (see Equation 3) is fully differentiable, allowing $\nabla_\delta \mathcal{L}_{\text{pred}}$ to be computed directly. In contrast, computing $\nabla_\delta \mathcal{L}_{\text{ins}}$ is substantially less straightforward because the binary masks are generated through a ranking operation, which is discrete and non-differentiable. One possible way to backpropagate through this flow is to leverage a differentiable version of Equation 4, which is proposed by Yoshikawa & Iwata (2024). However, this requires a more complex computation due to backpropagation through the saliency map, as well as the explanation method G. This often introduces **second-order derivatives** (Dombrowski et al., 2019; Yoshikawa & Iwata, 2024; Ghorbani et al., 2019), as many VLM explanation methods rely on computing the gradient of the model output with respect to the input (Zhao et al., 2024; Selvaraju et al., 2017; Dong et al., 2023; Chefer et al., 2021). Moreover, each explanation method has its own flow for generating the saliency map, which requires more effort to formulate. From that, although this approach can help us to more accurately compute the gradient, it's not appropriate to use because of its computational cost and complexity, as also noted by Yoshikawa & Iwata (2024).

To address this challenge, we propose **detaching the gradient through the M_t generation branch**, treating M_t as a constant during the attack optimization. In other words, we stop the gradient from propagating through the mask-generation process. We formally derive $\nabla_\delta \mathcal{L}_{\text{ins}}$ under this treatment in Appendix B, leading to the following surrogate gradient:

$$\nabla_\delta \mathcal{L}_{\text{ins}} = \frac{1}{T}\sum_{t=1}^{T} M_t \odot \frac{\partial p\left(\hat{y}(x), x_t^{\text{ins}}\right)}{\partial x_t^{\text{ins}}} \tag{8}$$

Intuitively, Equation 8 directs the gradient toward the regions identified as most salient, measuring how changes to these regions affect the insertion objective.

After completing the predefined K iterations, we select the final adversarial candidate as the class-preserving candidate with the lowest $\mathcal{L}_{\text{IOA}}$.

DOA. The target of DOA is to generate an adversarial perturbation that maximizes the Deletion AUC while preserving the original prediction:

$$\delta^* = \arg\max_{\delta} \text{DeletionAUC}(x+\delta) \quad \text{s.t.} \quad \|\delta\|_\infty \le \epsilon,\ \hat{y}(x+\delta) = \hat{y}(x). \tag{9}$$

DOA follows the same optimization pipeline as IOA, except that the masks are used to construct the deletion reference images:

$$x_t^{\text{del}} = (1 - M_t) \odot x^{adv}. \tag{10}$$

Using the generated reference images, we define the deletion objective as

$$\mathcal{L}_{\text{del}} = \frac{1}{T} \sum_{t=1}^{T} p(\hat{y}(x), x_t^{\text{del}}), \tag{11}$$

which encourages the prediction confidence to remain high after removing the salient regions throughout the deletion process. To maximize the Deletion AUC while preserving the original prediction, the overall optimization objective for the DOA variant is defined as

$$\mathcal{L}_{\text{DOA}} = -\left(\mathcal{L}_{\text{del}} - \mathcal{L}_{\text{pred}}\right). \tag{12}$$

To minimize the objective in Equation 12, we update the perturbation using the same projected gradient descent procedure:

$$\delta \leftarrow \Pi_\epsilon \left(\delta - \alpha \operatorname{sign}\left(\nabla_\delta \mathcal{L}_{\text{DOA}}\right)\right). \tag{13}$$

As with $\mathcal{L}_{\text{ins}}$, we handle the non-differentiable mask generation by detaching M_t when computing $\nabla_\delta \mathcal{L}_{\text{del}}$, yielding:

$$\nabla_\delta \mathcal{L}_{\text{del}} = \frac{1}{T} \sum_{t=1}^{T} (1 - M_t) \odot \frac{\partial p\left(\hat{y}(x), x_t^{\text{del}}\right)}{\partial x_t^{\text{del}}}. \tag{14}$$

IOA+DOA. To jointly manipulate both insertion and deletion faithfulness, we combine the two objectives and minimize their sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{IOA}} + \mathcal{L}_{\text{DOA}}. \tag{15}$$

The perturbation is optimized using the same projected gradient descent update:

$$\delta \leftarrow \Pi_\epsilon \left(\delta - \alpha \operatorname{sign}\left(\nabla_\delta \mathcal{L}_{\text{total}}\right)\right). \tag{16}$$

4 Experiments

4.1 Experimental Setup

Target VLM and explanation methods. Following prior work on VLM explanations (Zhao et al., 2024; Dong et al., 2023), we conduct all experiments on CLIP ViT-B/16 (Radford et al., 2021), which serves as the target VLM throughout our evaluation. We evaluate the attack across five representative VLM explanation methods: GradECLIP (Zhao et al., 2024), GAME (Chefer et al., 2021), Grad-CAM (Selvaraju et al., 2017), MaskCLIP (Dong et al., 2023), and RISE (Petsiuk et al., 2018).

Datasets. We evaluate FOA on three widely used benchmarks for VLM zero-shot classification: ImageNet (Russakovsky et al., 2015), Stanford Pets (Khosla et al., 2011), and CUB-200-2011 (Welinder et al., 2010). For each dataset, we select 200 images that are correctly classified by the target VLM and use them for adversarial evaluation.

Baseline. In addition to comparing our proposed attack, **FOA**, against the clean condition (i.e., *Without Attack*), we compare it with **X-Shift** (Babadi & Karimipour, 2026), the first and only adversarial attack on VLMs' explanations. Since X-Shift generates adversarial perturbations under an ℓ_0 constraint, which only restricts the number of perturbed pixels, we reimplement X-Shift by replacing its ℓ_0 projection with an ℓ_∞ projection using the same perturbation budget as our approach. This ensures that both methods are evaluated under the same constraint on perturbation magnitude.

Attack hyperparameters. We set the perturbation budget to $\epsilon = 16/255$, the step size to $\alpha = 4/255$, and the number of optimization iterations to $K = 50$. For faithfulness evaluation, we use $T = 224$ evaluation steps, where each step inserts or removes 224 pixels for a 224×224 input image, following Zhao et al. (2024) and Petsiuk et al. (2018).

Table 1: Comparison of different adversarial attack strategies across VLM explanations and datasets. Lower IMD (↓) and Spearman correlation (↓), while higher ROAD (↑), indicate more effective attacks. The **blue** values indicate the best score for each metric, dataset, and explainer. The **red** values denote the improvement of the devised approach over the corresponding baseline, using *Without Attack* as the baseline for IMD and ROAD, and *X-Shift* as the baseline for Spearman correlation.

		ImageNet			StanfordPets			CUB		
Explainer	**Attack**	**IMD ↓**	**SRC ↓**	**ROAD ↑**	**IMD ↓**	**SRC ↓**	**ROAD ↑**	**IMD ↓**	**SRC ↓**	**ROAD ↑**
GradECLIP	Without Attack	0.6370	–	0.1466	0.6103	–	0.1137	0.5910	–	0.1578
	X-Shift	0.7085 $_{+.072}$	0.5688	0.1178 $_{-.029}$	0.7348 $_{+.125}$	0.6728	0.0901 $_{-.024}$	0.8074 $_{+.216}$	0.5257	0.0889 $_{-.069}$
	IOA	**0.0957** $_{-.541}$	**0.5235** $_{-.045}$	0.1493 $_{+.003}$	**0.2014** $_{-.409}$	0.6841 $_{+.011}$	0.1211 $_{+.007}$	0.4755 $_{-.115}$	0.6189 $_{+.093}$	0.1085 $_{-.049}$
	DOA	0.2337 $_{-.403}$	0.5372 $_{-.032}$	**0.3688** $_{+.222}$	0.2549 $_{-.355}$	**0.6255** $_{-.047}$	**0.3163** $_{+.203}$	**0.3854** $_{-.206}$	**0.4898** $_{-.036}$	**0.3120** $_{+.154}$
	IOA+DOA	0.3283 $_{-.309}$	0.7662 $_{+.197}$	0.1582 $_{+.012}$	0.5352 $_{-.075}$	0.9019 $_{+.229}$	0.1548 $_{+.041}$	0.4447 $_{-.146}$	0.8818 $_{+.356}$	0.1182 $_{-.040}$
GAME	Without Attack	0.5816	–	0.1819	0.5973	–	0.1478	0.5388	–	0.1690
	X-Shift	0.6770 $_{+.095}$	0.4197	0.1561 $_{-.026}$	0.7316 $_{+.134}$	0.5351	0.0983 $_{-.049}$	0.7859 $_{+.247}$	0.3969	0.1099 $_{-.059}$
	IOA	**0.1115** $_{-.470}$	**0.3511** $_{-.069}$	0.1599 $_{-.022}$	**0.2077** $_{-.390}$	0.4933 $_{-.042}$	0.1339 $_{-.014}$	0.4628 $_{-.076}$	0.4395 $_{+.043}$	0.1255 $_{-.044}$
	DOA	0.2241 $_{-.358}$	0.3942 $_{-.026}$	**0.4022** $_{+.220}$	0.2633 $_{-.334}$	**0.4720** $_{-.063}$	**0.3384** $_{+.191}$	**0.3662** $_{-.173}$	**0.3853** $_{-.012}$	**0.3580** $_{+.189}$
	IOA+DOA	0.2588 $_{-.323}$	0.5972 $_{+.177}$	0.2122 $_{+.030}$	0.4483 $_{-.149}$	0.8303 $_{+.295}$	0.1671 $_{+.019}$	0.4222 $_{-.117}$	0.7993 $_{+.402}$	0.1515 $_{-.018}$
GradCAM	Without Attack	0.4608	–	0.2115	0.4789	–	0.1441	0.1475	–	0.3256
	X-Shift	0.5354 $_{+.075}$	0.3165	0.2050 $_{-.007}$	0.5896 $_{+.111}$	**0.3954**	0.1434 $_{-.001}$	0.4375 $_{+.290}$	**0.1432**	0.2827 $_{-.043}$
	IOA	**-0.0374** $_{-.498}$	**0.3070** $_{-.010}$	0.2210 $_{+.010}$	**0.0887** $_{-.390}$	0.3998 $_{+.004}$	0.1729 $_{+.029}$	-0.0908 $_{-.238}$	0.2799 $_{+.137}$	0.3505 $_{+.025}$
	DOA	0.0071 $_{-.454}$	0.4038 $_{+.087}$	**0.5084** $_{+.297}$	0.1249 $_{-.354}$	0.4448 $_{+.049}$	**0.4196** $_{+.275}$	**-0.1779** $_{-.325}$	0.3816 $_{+.238}$	**0.6078** $_{+.282}$
	IOA+DOA	0.2576 $_{-.203}$	0.6632 $_{+.347}$	0.1885 $_{-.023}$	0.2968 $_{-.182}$	0.4746 $_{+.079}$	0.1934 $_{+.049}$	0.3915 $_{+.244}$	0.7749 $_{+.632}$	0.1306 $_{-.195}$
MaskCLIP	Without Attack	0.5587	–	0.1617	0.5683	–	0.1232	0.5404	–	0.1564
	X-Shift	0.6747 $_{+.116}$	0.4278	0.1416 $_{-.020}$	0.7165 $_{+.148}$	0.5572	0.0937 $_{-.029}$	0.7855 $_{+.245}$	0.4027	0.0977 $_{-.059}$
	IOA	**0.1411** $_{-.418}$	**0.4037** $_{-.024}$	0.1470 $_{-.015}$	**0.2185** $_{-.350}$	0.5470 $_{-.010}$	0.1021 $_{-.021}$	0.4812 $_{-.059}$	0.4691 $_{+.066}$	0.0985 $_{-.058}$
	DOA	0.2255 $_{-.333}$	0.4545 $_{+.027}$	**0.3625** $_{+.201}$	0.2470 $_{-.321}$	**0.5111** $_{-.046}$	**0.3088** $_{+.186}$	**0.3798** $_{-.161}$	**0.3743** $_{-.028}$	**0.3147** $_{+.158}$
	IOA+DOA	0.2723 $_{-.286}$	0.7195 $_{+.292}$	0.1744 $_{+.013}$	0.4788 $_{-.090}$	0.8385 $_{+.281}$	0.1516 $_{+.028}$	0.4277 $_{-.113}$	0.8630 $_{+.460}$	0.1411 $_{-.015}$
RISE	Without Attack	0.3497	–	0.3107	0.3552	–	0.2353	0.3006	–	0.2886
	X-Shift	0.4623 $_{+.113}$	0.0644	0.2905 $_{-.020}$	0.5586 $_{+.203}$	0.0575	0.1844 $_{-.051}$	0.5395 $_{+.239}$	0.0257	0.2557 $_{-.033}$
	IOA	**0.0417** $_{-.308}$	**0.0060** $_{-.058}$	0.2167 $_{-.094}$	**0.0943** $_{-.261}$	0.0528 $_{-.005}$	0.1909 $_{-.044}$	**0.0967** $_{-.204}$	0.0213 $_{-.004}$	0.3204 $_{+.032}$
	DOA	0.1918 $_{-.158}$	0.0090 $_{-.055}$	**0.5424** $_{+.232}$	0.1692 $_{-.186}$	**0.0274** $_{-.030}$	**0.5140** $_{+.279}$	0.2116 $_{-.089}$	**0.0044** $_{-.021}$	**0.5639** $_{+.275}$
	IOA+DOA	0.1741 $_{-.176}$	0.0270 $_{-.037}$	0.2556 $_{-.055}$	0.2535 $_{-.102}$	0.0554 $_{-.002}$	0.2397 $_{+.004}$	0.2536 $_{-.047}$	0.0433 $_{+.018}$	0.2133 $_{-.075}$

Evaluation metrics. We evaluate attack effectiveness using faithfulness degradation and explanation ranking change. We report the **Deletion AUC** and **Insertion AUC** (Petsiuk et al., 2018). A higher Deletion AUC and a lower Insertion AUC indicate weaker faithfulness and, consequently, stronger attacks. We further report the **IMD** score, defined as the Insertion AUC minus the Deletion AUC. A lower IMD indicates weaker overall faithfulness and a stronger attack. We measure changes in saliency ranking using the **Spearman Rank Correlation (SRC)** (Spearman, 1904), where a lower correlation indicates a larger ranking change and a stronger attack. SRC is considered 1.0 under clean condition, meaning that the produced saliency maps are identical to the clean ones. Finally, we employ an additional faithfulness metric, **Remove and Debias (ROAD)** (Rong et al., 2022), which is not included in the FOA optimization process. Conventional deletion evaluation can suffer from *class information leakage*, where imputation artifacts may reveal the removal mask rather than reflecting the model's reliance on the removed features. ROAD addresses this issue by replacing the removed regions with imputed values at each deletion step rather than filling them with a black mask. In our experiments, removed regions are replaced with their channel-wise mean.

4.2 Main Results and Discussion

Table 1 presents a comparison of different attack strategies across VLM explanations and datasets, while Figure 3 provides some qualitative examples. A large number of additional qualitative examples are provided in Appendix D.

FOA effectively degrades the Faithfulness and Robustness of VLM explanations. Compared to the clean condition (i.e., *Without Attack*), FOA effectively reduces the faithfulness of VLM explanations and substantially alters their saliency maps while preserving the original predictions, as indicated by the red numbers in almost all cells corresponding to our proposed attacks. For instance, on the GradECLIP explainer and ImageNet dataset, FOA reduces the IMD from 0.6370 under *Without Attack* to 0.0957 under the IOA variant, while increasing the ROAD score from 0.1466 up to 0.3688 under the DOA variant. This is further evidenced in Figure 3, where the red areas

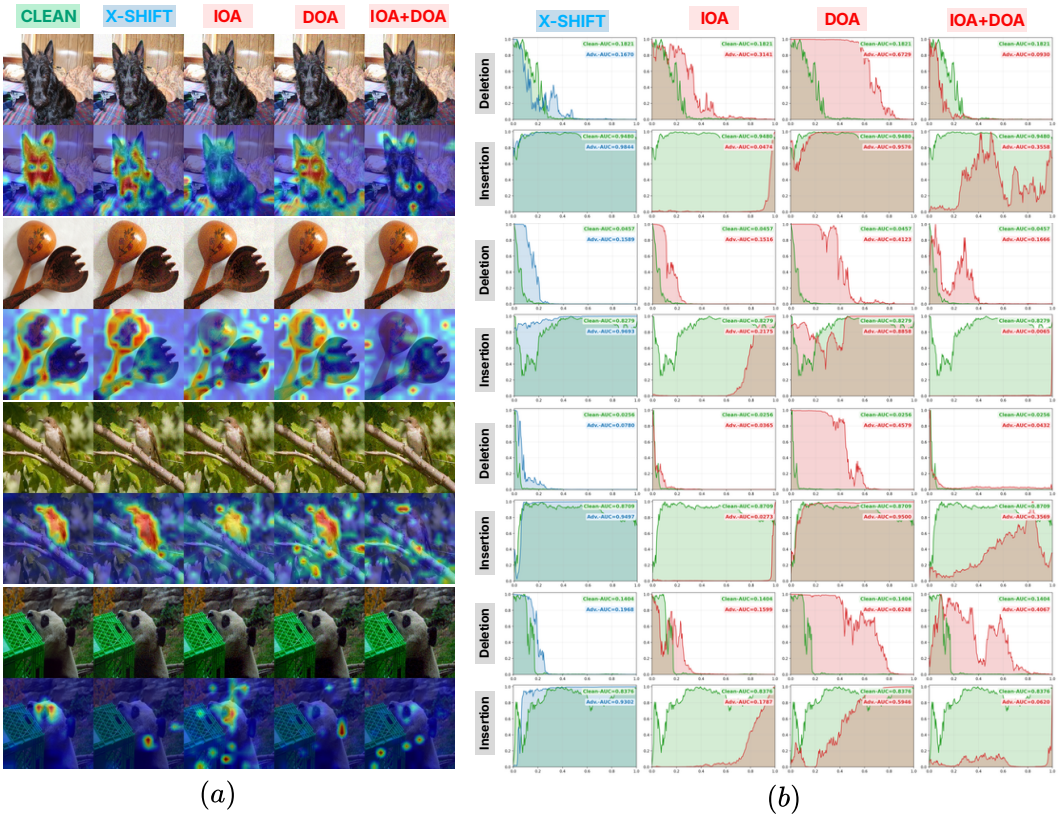

Figure 3: The faithfulness evaluation of Vision-Language Model (VLM) explanations under Clean, X-Shift (Babadi & Karimipour, 2026), and our proposed Faithfulness-Oriented Adversarial (FOA) attack. **a)** Adversarial images and their corresponding visual explanations under different FOA variants. **b)** Faithfulness evaluation based on Deletion and Insertion processes, comparing FOA variants with the *Clean* and X-Shift conditions.

representing the FOA curves dominate the green areas corresponding to the clean condition in the Deletion evaluation, indicating a substantial reduction in explanation faithfulness. Conversely, for the Insertion evaluation, the red areas are substantially larger than the green areas, likewise indicating degraded faithfulness under FOA. In terms of explanation shift, FOA substantially alters the VLM's attention toward different regions, as reflected by the reduced SRC, indicating a significant change in the ranking of regions in the visual explanation after the attack. For example, FOA reduces the SRC with respect to the clean explanation (considered as 1.0) to as low as 0.3743 for the MaskCLIP explainer. Qualitatively, Figure 3 shows that under the clean condition, the visual explanations primarily focus on semantically relevant regions, such as the dog's face. After applying FOA, the attention patterns are changed, with the original salient regions becoming blurred and attention being altered to less relevant regions, such as the dog's ears, body, and surrounding areas.

Moreover, compared to X-Shift (Babadi & Karimipour, 2026), the only existing state-of-the-art attack on VLM explanations, FOA achieves substantially stronger attack performance. As discussed in Section 1, although X-Shift can shift the saliency map away from the clean explanation, it fails to effectively degrade explanation faithfulness. This is because its objective redirects the model's attention toward alternative regions that remain important to the model's prediction. In contrast, by explicitly optimizing for faithfulness degradation, FOA not only induces a substantial explanation shift but also redirects attention toward regions that are less relevant to the model's prediction, thereby simultaneously achieving both objectives. From this perspective, FOA consistently outperforms X-Shift, as shown in Table 1. X-Shift achieves limited degradation in faithfulness, as reflected by the few red values in the IMD and ROAD metrics, whereas FOA consistently achieves lower SRC values, highlighted in blue, indicating stronger explanation shifts. This finding is also supported by Figure 3. For example, in the second sample with the IOA+DOA variant, X-Shift's explanation primarily

Table 2: Transfer Attack (average across three datasets). IMD (↓), Spearman rank correlation (SRC, ↓), and ROAD (↑). **Bold** values indicate direct attacks, where the perturbation is optimized on the same explainer used for evaluation (*Source* = *Target*). **Red** values indicate improvements of our proposed FOA over the corresponding X-Shift baseline.

Attack	Source	IMD ↓ Target Explainer					SRC ↓ Target Explainer					ROAD ↑ Target Explainer				
		GradECLIP	GAME	GradCAM	MaskCLIP	RISE	GradECLIP	GAME	GradCAM	MaskCLIP	RISE	GradECLIP	GAME	GradCAM	MaskCLIP	RISE
X-Shift	–	0.7502	0.7315	0.5208	0.6439	0.5201	0.5891	0.4506	0.2850	0.4626	0.0492	0.0989	0.1214	0.2104	0.1110	0.2435
IOA	GradECLIP	**0.2575**	0.3404 $_{-.391}$	0.1179 $_{-.403}$	0.3292 $_{-.315}$	0.2307 $_{-.289}$	**0.6088**	0.4249 $_{-.026}$	0.3102 $_{-.025}$	0.4676 $_{+.005}$	0.0324 $_{-.017}$	**0.1263**	0.1300 $_{+.009}$	0.2008 $_{-.010}$	0.1195 $_{+.008}$	0.1921 $_{-.051}$
	GAME	0.3557 $_{-.394}$	**0.2607** $_{-.391}$	0.1722 $_{-.403}$	0.3226 $_{-.321}$	0.2401 $_{-.280}$	0.5824 $_{-.007}$	**0.4280** $_{-.026}$	0.2913 $_{-.006}$	0.4605 $_{-.002}$	0.0170 $_{-.032}$	0.1140 $_{+.015}$	**0.1398** $_{+.009}$	0.1787 $_{-.032}$	0.1185 $_{+.007}$	0.1883 $_{-.055}$
	GradCAM	0.3823 $_{-.368}$	0.3994 $_{-.232}$	**-0.0132** $_{-.349}$	0.3877 $_{-.256}$	0.2675 $_{-.253}$	0.5398 $_{-.049}$	0.3954 $_{-.055}$	**0.3289** $_{-.006}$	0.4468 $_{-.016}$	0.0325 $_{-.017}$	0.1189 $_{+.020}$	0.1406 $_{+.019}$	**0.2481** $_{-.032}$	0.1244 $_{+.013}$	0.1812 $_{-.062}$
	MaskCLIP	0.3540 $_{-.396}$	0.3289 $_{-.303}$	0.1564 $_{-.364}$	**0.2803** $_{-.241}$	0.2347 $_{-.285}$	0.5851 $_{-.004}$	0.4376 $_{-.013}$	0.3080 $_{-.023}$	**0.4733** $_{-.016}$	0.0269 $_{-.022}$	0.1135 $_{+.015}$	0.1243 $_{+.003}$	0.1960 $_{-.014}$	**0.1159** $_{-.033}$	0.1795 $_{-.064}$
	RISE	0.4325 $_{-.318}$	0.4143 $_{-.317}$	0.2047 $_{-.316}$	0.3997 $_{-.244}$	**0.0776** $_{-.316}$	0.5694 $_{-.020}$	0.4239 $_{-.027}$	0.2812 $_{-.004}$	0.4676 $_{+.005}$	**0.0267** $_{-.022}$	0.1197 $_{+.021}$	0.1304 $_{+.009}$	0.1770 $_{-.033}$	0.1091 $_{-.002}$	**0.2427** $_{-.064}$
DOA	GradECLIP	**0.2913**	0.3877 $_{-.344}$	0.1515 $_{-.369}$	0.3581 $_{-.286}$	0.2310 $_{-.289}$	**0.5508**	0.3990 $_{-.052}$	0.3827 $_{-.098}$	0.4620 $_{-.001}$	0.0266 $_{-.023}$	**0.3324**	0.3183 $_{+.197}$	0.4167 $_{-.206}$	0.2963 $_{+.183}$	0.4114 $_{+.168}$
	GAME	0.4438 $_{-.306}$	**0.2845** $_{-.344}$	0.2298 $_{-.291}$	0.3995 $_{-.244}$	0.2273 $_{-.293}$	0.5946 $_{+.006}$	**0.4172** $_{-.052}$	0.3777 $_{-.093}$	0.4555 $_{-.007}$	0.0239 $_{-.025}$	0.2445 $_{+.146}$	**0.3662** $_{+.197}$	0.3680 $_{+.158}$	0.2737 $_{+.163}$	0.4121 $_{+.169}$
	GradCAM	0.5688 $_{-.181}$	0.5911 $_{-.140}$	**-0.0153** $_{-.349}$	0.5579 $_{-.086}$	0.3452 $_{-.175}$	0.6445 $_{+.055}$	0.4478 $_{-.003}$	**0.4101** $_{-.098}$	0.5097 $_{+.047}$	0.0245 $_{-.025}$	0.1937 $_{+.095}$	0.2035 $_{+.082}$	**0.5119** $_{-.206}$	0.1971 $_{+.086}$	0.3694 $_{+.126}$
	MaskCLIP	0.3959 $_{-.354}$	0.3751 $_{-.356}$	0.1919 $_{-.329}$	**0.2841** $_{-.241}$	0.2355 $_{-.285}$	0.5832 $_{-.006}$	0.3990 $_{-.052}$	0.3753 $_{-.090}$	**0.4466** $_{-.093}$	0.0297 $_{-.019}$	0.2718 $_{+.173}$	0.3179 $_{+.197}$	0.3832 $_{-.173}$	**0.3287** $_{-.162}$	0.4027 $_{+.159}$
	RISE	0.6028 $_{-.147}$	0.5325 $_{-.199}$	0.2795 $_{-.241}$	0.5670 $_{-.077}$	**0.1909** $_{-.241}$	0.6588 $_{+.070}$	0.4844 $_{+.034}$	0.4340 $_{+.149}$	0.5402 $_{+.078}$	**0.0136** $_{-.034}$	0.2057 $_{+.107}$	0.2836 $_{+.162}$	0.4206 $_{+.210}$	0.2361 $_{+.125}$	**0.5401** $_{+.210}$
IOA+DOA	GradECLIP	**0.4360**	0.4167 $_{-.315}$	0.2321 $_{-.289}$	0.3947 $_{-.249}$	0.2770 $_{-.243}$	**0.8500**	0.8066 $_{+.356}$	0.7931 $_{+.508}$	0.8234 $_{+.361}$	0.0302 $_{-.019}$	**0.1437**	0.1767 $_{+.055}$	0.2255 $_{-.015}$	0.1530 $_{+.042}$	0.2522 $_{+.009}$
	GAME	0.4194 $_{-.331}$	**0.3886** $_{-.281}$	0.2398 $_{-.281}$	0.3516 $_{-.292}$	0.2738 $_{-.246}$	0.8153 $_{+.226}$	**0.7423** $_{-.457}$	0.7421 $_{+.457}$	0.7797 $_{+.317}$	0.0138 $_{-.035}$	0.1494 $_{+.051}$	**0.1769** $_{-.002}$	0.2120 $_{-.002}$	0.1528 $_{+.042}$	0.2440 $_{+.001}$
	GradCAM	0.4928 $_{-.257}$	0.4747 $_{-.257}$	**0.2500** $_{-.349}$	0.4445 $_{-.199}$	0.3271 $_{-.193}$	0.7891 $_{+.200}$	0.7364 $_{+.286}$	**0.6376** $_{-.098}$	0.7473 $_{+.285}$	0.0405 $_{-.009}$	0.1542 $_{+.055}$	0.1925 $_{+.071}$	**0.1698** $_{-.206}$	0.1611 $_{+.050}$	0.2484 $_{+.005}$
	MaskCLIP	0.4580 $_{-.292}$	0.4175 $_{-.214}$	0.2600 $_{-.261}$	**0.3929** $_{-.241}$	0.2796 $_{-.240}$	0.8414 $_{+.252}$	0.7979 $_{+.347}$	0.7818 $_{+.497}$	**0.8219** $_{+.034}$	0.0393 $_{-.010}$	0.1485 $_{+.050}$	0.1740 $_{+.053}$	0.2209 $_{-.011}$	**0.1557** $_{-.162}$	0.2514 $_{+.008}$
	RISE	0.4288 $_{-.321}$	0.3930 $_{-.339}$	0.2348 $_{-.286}$	0.3658 $_{-.278}$	**0.2136** $_{-.316}$	0.7814 $_{+.192}$	0.7034 $_{+.253}$	0.6460 $_{+.361}$	0.7166 $_{+.254}$	**0.0419** $_{-.034}$	0.1479 $_{+.049}$	0.1861 $_{+.065}$	0.2030 $_{-.007}$	0.1564 $_{+.043}$	**0.2362** $_{-.210}$

focuses on the two wooden spoons, whereas FOA shifts the attention toward the background. We further analyze the effect of different adversarial perturbation budgets ϵ in Appendix C.

Transfer-Based Attack Ability. Table 2 presents the transfer attack results, where adversarial examples are optimized using one VLM explainer (denoted as *Source*) and subsequently evaluated on another explainer (denoted as *Target*). Since X-Shift is optimized based on the model's attention rather than a specific VLM explainer, its attack is inherently explainer-agnostic and is therefore reported as a single baseline row. Adversarial perturbations optimized with FOA on one source explainer can effectively transfer to different target explainers, often achieving substantially stronger attacks than X-Shift across the three evaluation metrics. For example, when transferring the DOA perturbation optimized with GradECLIP to GAME, it reduces the IMD from 0.7315 to 0.3887, decreases the SRC from 0.4506 to 0.3990, and substantially increases the ROAD score from 0.1214 to 0.3183. This trend is consistently observed across many source-target combinations in Table 2, as indicated by the numerous red values, demonstrating the strong transferability of FOA across different explainers. This indicates that FOA shifts saliency toward regions that are less important to the model, making the resulting adversarial perturbations easier to transfer across different explainers.

It is worth noting that X-Shift is inherently explainer-agnostic, as its perturbation is optimized based on the VLM's model-level attention rather than a specific explanation function. Therefore, X-Shift is reported as a single baseline row rather than under a source-target explainer setting. This does not provide X-Shift with an easier or incomparable setting: FOA is explicitly evaluated in a cross-explainer setting, where perturbations are optimized using one explainer and transferred to another. Moreover, constructing an explainer-specific version of X-Shift would instead require differentiating through the corresponding explanation function during optimization, introducing additional computational and implementation complexity, as discussed in Section 3.2.

5 Conclusion

Our results show that explanation robustness cannot be characterized by saliency stability alone. A VLM explanation may change substantially yet remain faithful, while prediction-preserving perturbations can redirect saliency toward regions that provide weak support for the model's decision. The proposed **Faithfulness-Oriented Adversarial (FOA)** attack provides a practical way to expose this failure mode by stress-testing explanation faithfulness without backpropagating through the explainer. Its effectiveness across multiple explanation methods and cross-explainer transferability suggest that this vulnerability is not tied to a particular saliency mechanism. More broadly, our findings motivate a shift in robustness evaluation from asking *whether an explanation changes* to asking *whether it still reflects the evidence underlying the model's prediction*. Ultimately, robust explanation should mean more than preserving where a model appears to look; it should preserve whether those highlighted regions actually matter to the prediction.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks:

- Editing the manuscript for spelling, grammar, readability, and presentation, including refinement of mathematical notation and formulations developed by the authors.
- Assisting with literature search and summarization, and providing feedback on the clarity and presentation of the manuscript.
- Assisting with the development of auxiliary code and visualization utilities. For example, generative AI tools were used to develop functions for saving saliency maps with specified color maps, reorganizing file and directory paths used by the main experimental pipeline, and implementing a Streamlit interface for qualitative visualization of the results.

Generative AI tools were not used to originate the proposed method, formulate the central research hypotheses, develop mathematical proofs or derivations from scratch, or implement the proposed method from scratch. Any AI-assisted writing was based on research ideas, technical content, and formulations first developed by the authors.

All AI-assisted text, code, and other outputs were reviewed by the authors, and AI-assisted code was verified before use in the experiments. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personal data, or private information, and all experiments are conducted on publicly released datasets and research models. Our study investigates the robustness and faithfulness of visual explanations for vision-language models (VLMs) under adversarial perturbations. The proposed attack is intended as a diagnostic tool for stress-testing explanation systems and identifying failure modes that may otherwise remain hidden under standard evaluation.

As with adversarial robustness research more broadly, the proposed method has a potential dual-use aspect, since similar techniques could be misused to manipulate model explanations. However, understanding whether explanations can be made unfaithful while model predictions remain unchanged is important for assessing their reliability, particularly when explanations are used to support human interpretation or decision making. Identifying such vulnerabilities is a necessary step toward developing more reliable evaluation protocols and more robust explanation methods.

Our experiments do not target individual users, deployed systems, or real-world services. The goal of this work is to expose and characterize weaknesses in current VLM explanation methods so that future research can better assess and improve their robustness and faithfulness.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 4.1 provides detailed descriptions of the datasets, target VLM, explanation methods, baseline, attack configurations, and evaluation metrics. The pseudocode of the proposed FOA attack is provided in Appendix A, and the detailed gradient derivation is given in Appendix B. Our source code and a **README.md** guide for reproducing all experiments are included in the supplementary material.

REFERENCES

Samira Abnar and Willem H. Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4190–4197. Association for Computational Linguistics, 2020.

Narges Babadi and Hadis Karimipour. Right predictions, misleading explanations: On the vulnerability of vision-language model explanations, 2026.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11: 1803–1831, 2010.

Hila Chefer, Shir Gur, and Lior Wolf. Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 387–396, Los Alamitos, CA, USA, October 2021. IEEE Computer Society.

Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *NeurIPS*, 2019.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. MaskCLIP: Masked Self-Distillation Advances Contrastive Language-Image Pretraining . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10995–11005, Los Alamitos, CA, USA, June 2023. IEEE Computer Society.

Amirata Ghorbani, Abubakar Abid, and James Y. Zou. Interpretation of neural networks is fragile. In *AAAI*, 2019.

Anna Hedström, Leander Weber, Sebastian Lapuschkin, and Marina M.-C. Höhne. A fresh look at sanity checks for saliency maps. In *xAI 2024*, Communications in Computer and Information Science, pp. 403–420. Springer, 2024.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pp. 267–280. Springer, 2019.

Solomon Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

Yao Qiang, Deng Pan, Chengyin Li, Xin Li, Rhongho Jang, and Dongxiao Zhu. Attcat: Explaining transformers via attentive class activation tokens. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong

Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 18770–18795, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*, 2014.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. ISSN 00029556. URL `http://www.jstor.org/stable/1412159`.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

Ying Wang, Tim G. J. Rudner, and Andrew Gordon Wilson. Visual explanations of image-text representations via multi-modal information bottleneck attribution. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-ucsd birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Weiyan Xie, Xiao-Hui Li, Caleb Chen Cao, and Nevin L. Zhang. Vit-cx: Causal explanation of vision transformers. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pp. 1569–1577. ijcai.org, 2023.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *NeurIPS*, 2019.

Yuya Yoshikawa and Tomoharu Iwata. Explanation-based training with differentiable insertion/deletion metric-aware regularizers. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 370–378. PMLR, 02–04 May 2024.

Lu Yu and Wei Xiang. X-pruner: explainable pruning for vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 24355–24363. IEEE, 2023.

Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (eds.), *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pp. 5005–5013. ACM, 2022.

Chenyang Zhao, Kun Wang, Xingyu Zeng, Rui Zhao, and Antoni B. Chan. Gradient-based visual explanation for transformer-based clip. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

APPENDIX

This appendix provides supplementary material for our paper, including the attack algorithm, detailed mathematical derivations, and additional experimental analyses. The contents of the appendix are organized as follows.

- **Section A:** Pseudocode of the proposed Faithfulness-Oriented Adversarial (FOA) Attack.
- **Section B:** Detailed derivation of the gradients used to optimize the adversarial perturbation.
- **Section C:** Analysis of the impact of different perturbation budgets ϵ.
- **Section D:** Additional qualitative results across different datasets and VLM explanation methods.

A ATTACK ALGORITHM

The pseudocode of the Faithfulness-Oriented Adversarial (FOA) Attack is shown in Algorithm 1. We consider three variants of FOA: Insertion-Oriented Adversarial (IOA) Attack, Deletion-Oriented Adversarial (DOA) Attack, and their combination variant IOA+DOA.

The attack starts by initializing the perturbation δ to zero tensor and iteratively updates it for a predefined budget of K iterations. At each iteration, the adversarial example is constructed as $x^{adv} = x + \delta$. We first compute the prediction-preserving loss $\mathcal{L}_{\text{pred}}$ as the KL divergence between the output distributions of the clean and adversarial images. We then generate a saliency map $S = G(c, x^{adv})$ with respect to the original prediction c, and sort its saliency values to construct a sequence of binary masks $\mathcal{M} = \{M_t\}_{t=1}^{T}$. Depending on the selected attack variant, these masks are used to construct insertion or deletion reference images. Specifically, IOA uses $x_t^{ins} = M_t \odot x^{adv}$ to compute the insertion loss $\mathcal{L}_{\text{ins}}$, while DOA uses $x_t^{del} = (1 - M_t) \odot x^{adv}$ to compute the deletion loss $\mathcal{L}_{\text{del}}$. The corresponding attack objectives are then formed by combining the faithfulness loss with $\mathcal{L}_{\text{pred}}$. For IOA, we minimize $\mathcal{L}_{\text{IOA}} = \mathcal{L}_{\text{ins}} + \mathcal{L}_{\text{pred}}$. For DOA, we minimize $\mathcal{L}_{\text{DOA}} = -(\mathcal{L}_{\text{del}} - \mathcal{L}_{\text{pred}})$. For IOA+DOA, the two objectives are jointly minimized as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{IOA}} + \mathcal{L}_{\text{DOA}}$.

Finally, the perturbation is updated using a projected gradient step: the gradient of the selected total objective with respect to δ is computed and used to update δ within the ℓ_∞ perturbation budget. The masks M_t are treated as constants during this optimization. Thus, gradients from the faithfulness objectives propagate through the corresponding reference images and the VLM, but do not backpropagate through the saliency-map generation, ranking, or Top-K mask construction process. This stop-gradient treatment avoids the need to differentiate through the discrete ranking operation and the internal computational graph of the explainer.

After each update, the prediction-preserving adversarial candidate is evaluated and the best valid candidate across all iterations is retained as the final adversarial example x_*^{adv}.

Algorithm 1: Faithfulness-Oriented Adversarial (FOA) Attack

Input: Image x, original prediction c, VLM f, explanation method G, attack type $A \in \{\text{IOA}, \text{DOA}, \text{IOA} + \text{DOA}\}$, perturbation budget ϵ, step size α, iterations K, causal steps T.

Output: Adversarial example x^{adv}.

$\delta \leftarrow \mathbf{0}, \quad \mathcal{L}^\star \leftarrow +\infty, \quad x_\star^{adv} \leftarrow \varnothing;$

for $k = 1, \ldots, K$ **do**

; // Attack optimization

$x^{adv} \leftarrow x + \delta;$

$\mathcal{L}_{pred} \leftarrow D_{\mathrm{KL}}(f(x) \| f(x^{adv}));$

$S \leftarrow G(c, x^{adv})$ and sort S in descending order to obtain $\mathcal{M} = \{M_t\}_{t=1}^{T};$

if $A = \text{IOA}$ ***or*** $A = \text{IOA} + \text{DOA}$ **then**

$\mathcal{L}_{ins} \leftarrow 0;$

for $t = 1, \ldots, T$ **do**

; // Insertion evaluation

$x_t^{ins} \leftarrow M_t \odot x^{adv};$

$\mathcal{L}_{ins} \leftarrow \mathcal{L}_{ins} + p(c, x_t^{ins});$

$\mathcal{L}_{ins} \leftarrow \mathcal{L}_{ins}/T;$

$\mathcal{L}_{IOA} \leftarrow \mathcal{L}_{ins} + \mathcal{L}_{pred};$

if $A = \text{DOA}$ ***or*** $A = \text{IOA} + \text{DOA}$ **then**

$\mathcal{L}_{del} \leftarrow 0;$

for $t = 1, \ldots, T$ **do**

; // Deletion evaluation

$x_t^{del} \leftarrow (1 - M_t) \odot x^{adv};$

$\mathcal{L}_{del} \leftarrow \mathcal{L}_{del} + p(c, x_t^{del});$

$\mathcal{L}_{del} \leftarrow \mathcal{L}_{del}/T;$

$\mathcal{L}_{DOA} \leftarrow -(\mathcal{L}_{del} - \mathcal{L}_{pred});$

if $A = \text{IOA}$ **then**

$\mathcal{L}_{total} \leftarrow \mathcal{L}_{IOA};$

else if $A = \text{DOA}$ **then**

$\mathcal{L}_{total} \leftarrow \mathcal{L}_{DOA};$

else

$\mathcal{L}_{total} \leftarrow \mathcal{L}_{IOA} + \mathcal{L}_{DOA};$

; // FOA perturbation update

$\delta \leftarrow \Pi_\epsilon \left(\delta - \alpha \ \text{sign}(\nabla_\delta \mathcal{L}_{total})\right);$

if $\hat{y}(x^{adv}) = c \ \wedge \ \mathcal{L}_{total} < \mathcal{L}^\star$ **then**

$\mathcal{L}^\star \leftarrow \mathcal{L}_{total}, \quad x_\star^{adv} \leftarrow x^{adv};$

if $x_\star^{adv} \neq \varnothing$ **then**

$x^{adv} \leftarrow x_\star^{adv};$

else

$x^{adv} \leftarrow x + \delta;$

return $x^{adv};$

B DERIVATION OF (8)

We provide the detailed derivation of the gradient calculation used to optimize the adversarial perturbation δ in the proposed FOA attack. We present the formulation for $\nabla_\delta \mathcal{L}_{\text{ins}}$, the derivation for $\nabla_\delta \mathcal{L}_{\text{del}}$ follows analogously. The notation used in this derivation is introduced in Section 3.2

$$
\begin{aligned}
\nabla_\delta \mathcal{L}_{\text{ins}} &= \frac{1}{T} \nabla_\delta \sum_{t=1}^{T} p\left(\hat{y}(x), x_t^{\text{ins}}\right) \\
&= \frac{1}{T} \sum_{t=1}^{T} \nabla_\delta p\left(\hat{y}(x), x_t^{\text{ins}}\right) \\
&= \frac{1}{T} \sum_{t=1}^{T} \frac{\partial p\left(\hat{y}(x), x_t^{\text{ins}}\right)}{\partial x_t^{\text{ins}}} \frac{\partial x_t^{\text{ins}}}{\partial \delta} \\
&= \frac{1}{T} \sum_{t=1}^{T} \frac{\partial p\left(\hat{y}(x), x_t^{\text{ins}}\right)}{\partial x_t^{\text{ins}}} \frac{\partial \left(M_t \odot x^{\text{adv}}\right)}{\partial \delta} \\
&= \frac{1}{T} \sum_{t=1}^{T} \frac{\partial p\left(\hat{y}(x), x_t^{\text{ins}}\right)}{\partial x_t^{\text{ins}}} \left(\text{Diag}\left(x^{\text{adv}}\right) \frac{\partial M_t}{\partial \delta} + \text{Diag}\left(M_t\right) \frac{\partial x^{\text{adv}}}{\partial \delta} \right) \\
&= \frac{1}{T} \sum_{t=1}^{T} \frac{\partial p\left(\hat{y}(x), x_t^{\text{ins}}\right)}{\partial x_t^{\text{ins}}} \text{Diag}\left(x^{\text{adv}}\right) \frac{\partial M_t}{\partial \delta} + \frac{1}{T} \sum_{t=1}^{T} \frac{\partial p\left(\hat{y}(x), x_t^{\text{ins}}\right)}{\partial x_t^{\text{ins}}} \text{Diag}\left(M_t\right) \frac{\partial x^{\text{adv}}}{\partial \delta} \\
&= \frac{1}{T} \sum_{t=1}^{T} \frac{\partial p\left(\hat{y}(x), x_t^{\text{ins}}\right)}{\partial x_t^{\text{ins}}} \text{Diag}\left(x^{\text{adv}}\right) \frac{\partial M_t}{\partial \delta} + \frac{1}{T} \sum_{t=1}^{T} \frac{\partial p\left(\hat{y}(x), x_t^{\text{ins}}\right)}{\partial x_t^{\text{ins}}} \text{Diag}\left(M_t\right) \\
&= \frac{1}{T} \sum_{t=1}^{T} \frac{\partial p\left(\hat{y}(x), x_t^{\text{ins}}\right)}{\partial x_t^{\text{ins}}} \underbrace{\text{Diag}\left(x^{\text{adv}}\right) \frac{\partial M_t}{\partial \delta}}_{\text{stop-gradient}} + \frac{1}{T} \sum_{t=1}^{T} M_t \odot \frac{\partial p\left(\hat{y}(x), x_t^{\text{ins}}\right)}{\partial x_t^{\text{ins}}} \\
&= \frac{1}{T} \sum_{t=1}^{T} M_t \odot \frac{\partial p\left(\hat{y}(x), x_t^{\text{ins}}\right)}{\partial x_t^{\text{ins}}}
\end{aligned}
\tag{17}
$$

C PERTURBATION BUDGET ϵ ANALYSIS

The Effect of the Perturbation Budget ϵ on Attack Effectiveness. We further conduct an experiment to evaluate attack effectiveness on a subset of 50 random samples from each dataset under different perturbation budgets, $\epsilon \in \{2/255, 4/255, 8/255, 16/255\}$, as presented in Figure 4. We also report the average confidence of the adversarial examples generated by different FOA variants across datasets and explainers for different values of ϵ in Table 3.

From these results, we observe that a higher ϵ generally leads to a larger explanation shift, as indicated by the lower SRC values. However, unlike conventional attack studies Zhao et al. (2023); Zhang et al. (2022); Madry et al. (2018); Ghorbani et al. (2019); Dombrowski et al. (2019), where a larger perturbation budget ϵ typically leads to stronger attack performance, our results show that this is not necessary for faithfulness degradation. For IOA, increasing ϵ decreases SRC but increases IMD, whereas for DOA and IOA+DOA, both SRC and IMD decrease as ϵ increases. Since our attack is designed to manipulate the explanation shift through faithfulness degradation, we report the results at $\epsilon = 16/255$ as the main results in the main text, where the explanation shift is more pronounced.

This behavior can be explained by the interaction between the perturbation budget and the prediction-preserving loss. As ϵ increases, the larger perturbation magnitude makes it easier to reduce the model's confidence in the original class on the adversarial image x_{adv}, as presented in Table 3. For IOA, this tendency is partially aligned with its objective of reducing the confidence of the inserted images x_t^{ins}. Consequently, $\mathcal{L}_{\text{pred}}$ needs to exert stronger pressure to preserve the original prediction distribution, limiting the extent to which the additional perturbation budget can further degrade explanation faithfulness. In contrast, DOA aims to increase the confidence of the deleted images x_t^{del}. As shown in Table 3, their confidence generally increases with ϵ, indicating that the DOA objective becomes increasingly compatible with preserving the original prediction distribution. Consequently, $\mathcal{L}_{pred}$ imposes a weaker effective constraint on the optimization, allowing the larger perturbation budget to be more effectively used to degrade explanation faithfulness, resulting in a lower IMD.

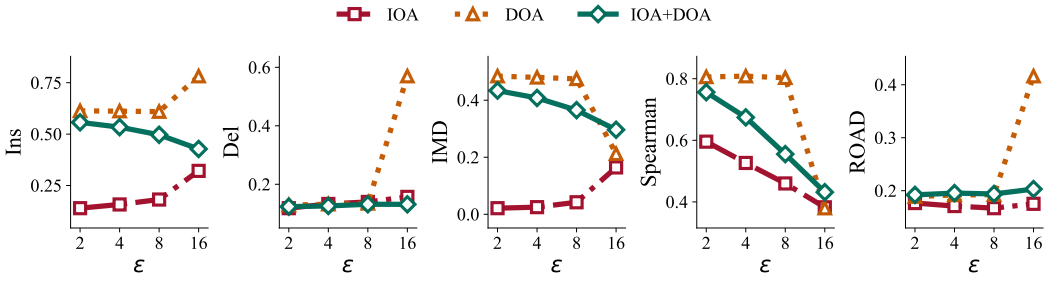

Figure 4: Attack effectiveness under different perturbation budgets ϵ. We report the IMD, Spearman Rank correlation, and ROAD scores, averaged across datasets and explainers.

Table 3: Confidence scores across perturbation strengths, datasets, explainers, and approaches.

		ImageNet				CUB				StanfordPet			
Explainer	**Approach**	$\epsilon=2$	$\epsilon=4$	$\epsilon=8$	$\epsilon=16$	$\epsilon=2$	$\epsilon=4$	$\epsilon=8$	$\epsilon=16$	$\epsilon=2$	$\epsilon=4$	$\epsilon=8$	$\epsilon=16$
GradCAM	IOA	0.90	0.91	0.87	0.60	0.82	0.75	0.75	0.37	0.85	0.85	0.84	0.44
	DOA	0.69	0.69	0.69	1.00	0.49	0.49	0.49	1.00	0.59	0.59	0.59	1.00
	IOA+DOA	0.47	0.34	0.24	0.13	0.44	0.34	0.26	0.29	0.52	0.33	0.25	0.12
GradECLIP	IOA	0.86	0.85	0.87	0.46	0.68	0.67	0.58	0.30	0.82	0.84	0.80	0.60
	DOA	0.69	0.69	0.69	1.00	0.49	0.49	0.49	1.00	0.59	0.59	0.58	1.00
	IOA+DOA	0.44	0.34	0.22	0.18	0.45	0.36	0.27	0.26	0.52	0.38	0.22	0.16
GAME	IOA	0.86	0.86	0.85	0.64	0.64	0.61	0.54	0.33	0.80	0.81	0.81	0.47
	DOA	0.69	0.69	0.69	1.00	0.49	0.49	0.49	1.00	0.59	0.59	0.58	1.00
	IOA+DOA	0.43	0.30	0.19	0.12	0.46	0.30	0.26	0.24	0.51	0.36	0.24	0.17
MaskCLIP	IOA	0.86	0.87	0.87	0.56	0.66	0.62	0.54	0.34	0.82	0.82	0.82	0.56
	DOA	0.69	0.69	0.69	1.00	0.49	0.49	0.49	1.00	0.59	0.59	0.58	0.99
	IOA+DOA	0.44	0.29	0.16	0.16	0.46	0.37	0.25	0.23	0.48	0.39	0.28	0.13
RISE	IOA	0.85	0.83	0.84	0.47	0.63	0.55	0.49	0.30	0.77	0.78	0.76	0.51
	DOA	0.69	0.69	0.70	1.00	0.49	0.49	0.50	1.00	0.59	0.59	0.59	1.00
	IOA+DOA	0.47	0.33	0.21	0.15	0.44	0.38	0.27	0.18	0.52	0.43	0.21	0.15

D MORE QUALITATIVE RESULTS

We provide a large number of qualitative results comparing our proposed method with the baseline approaches. We present samples across three datasets and five VLM explainers, where each figure corresponds to a specific dataset-explainer combination.

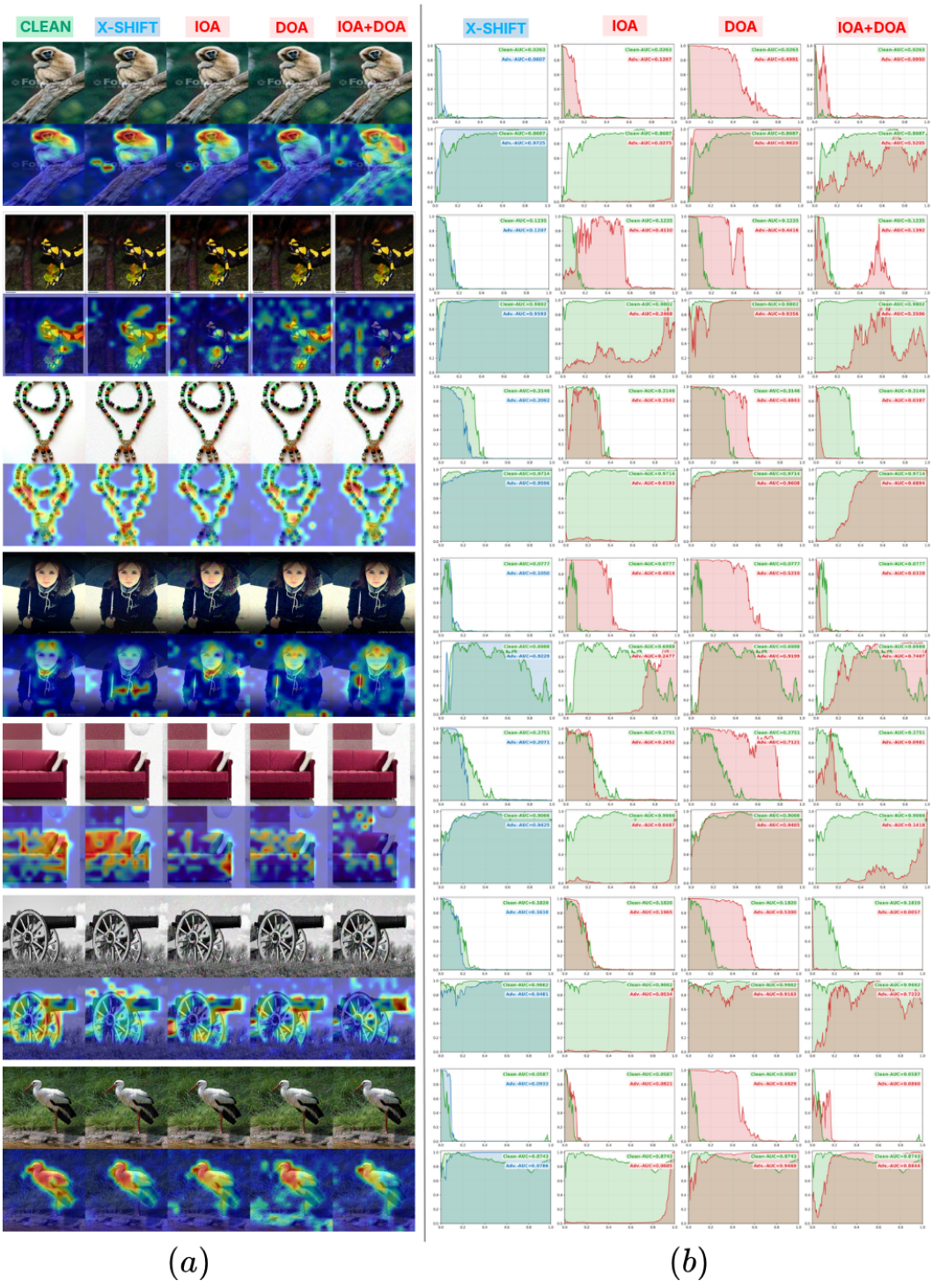

Figure 5: GradECLIP Zhao et al. (2024), ImageNet Russakovsky et al. (2015)

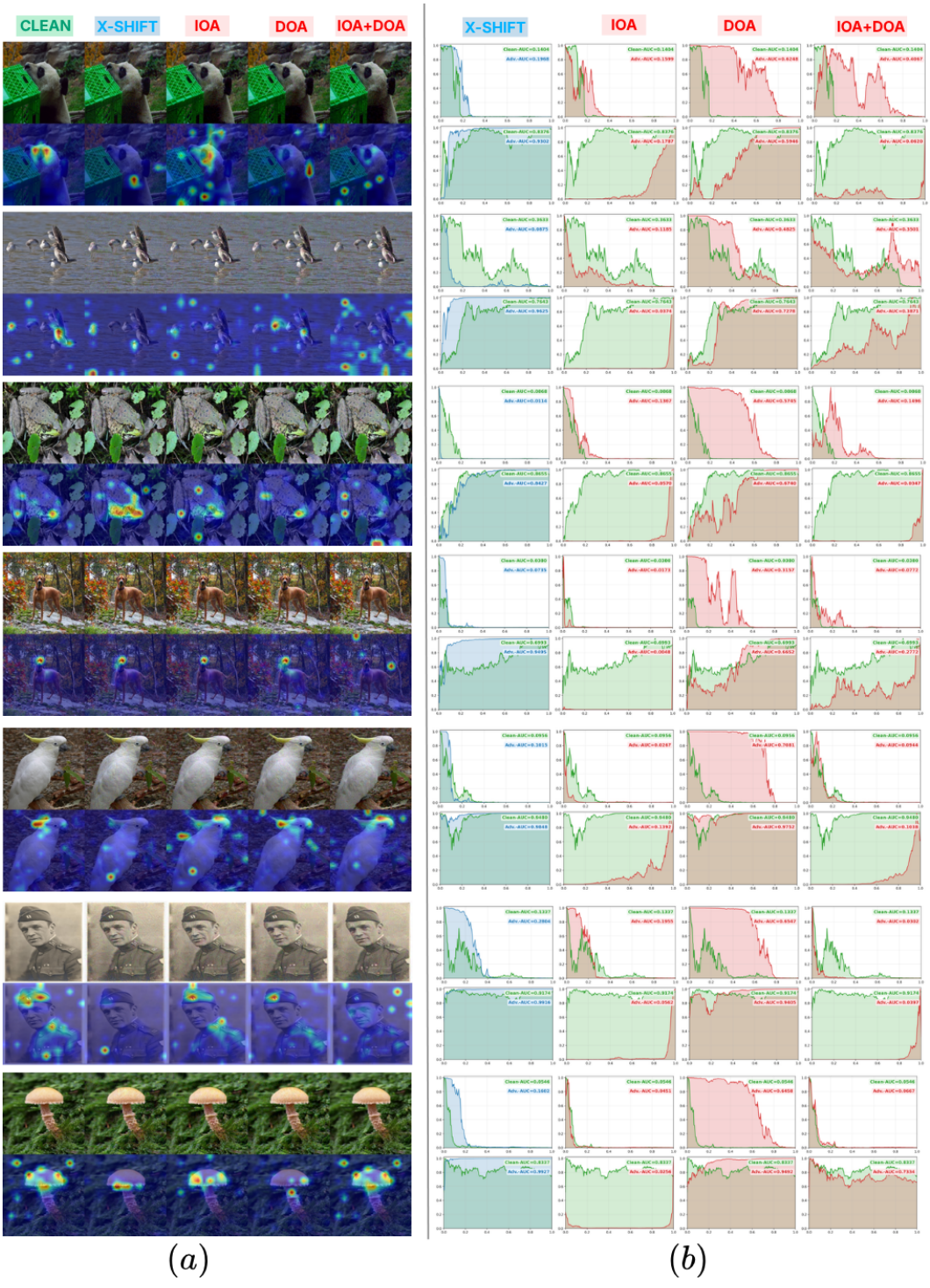

Figure 6: GAME Chefer et al. (2021), ImageNet Russakovsky et al. (2015)

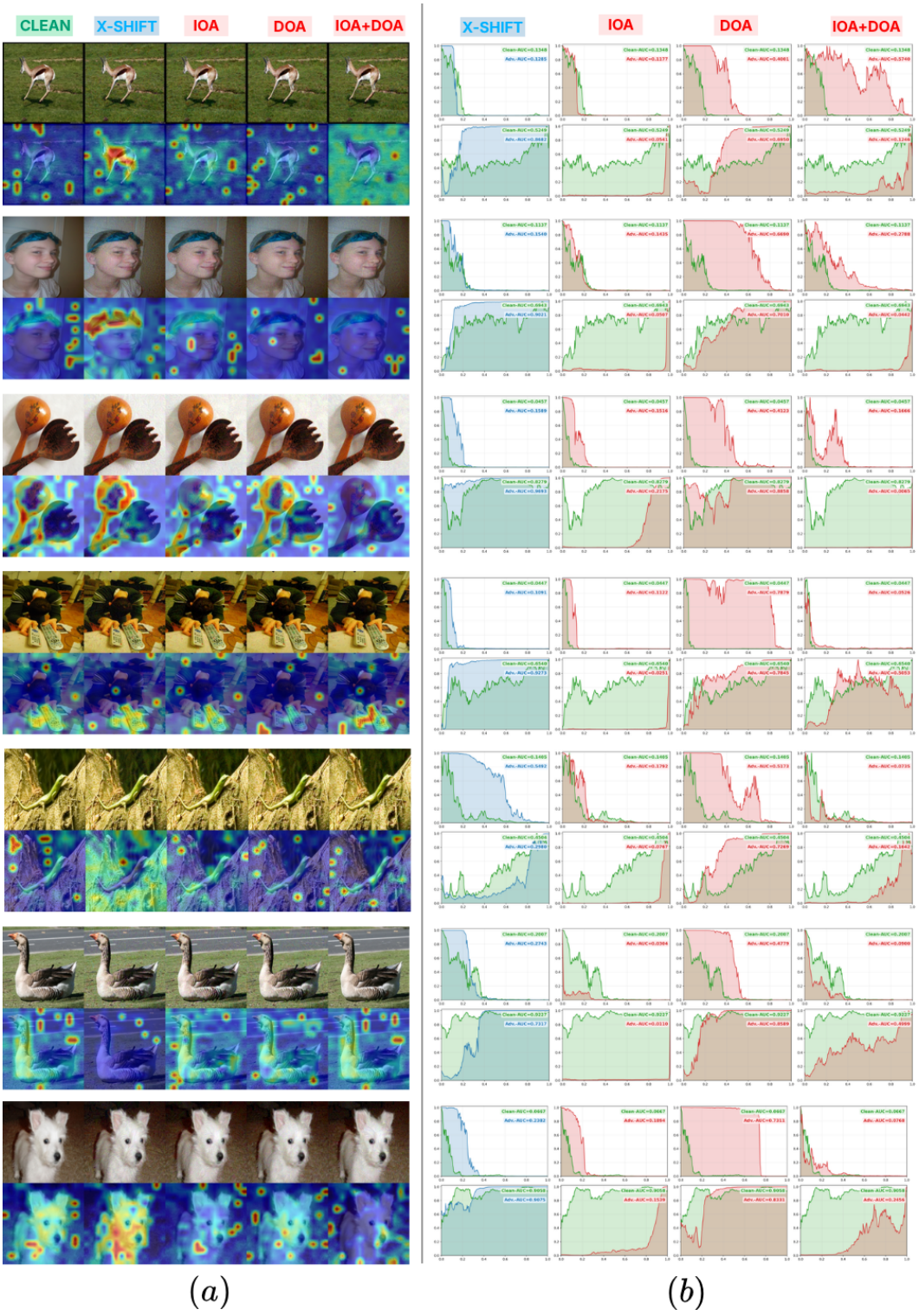

Figure 7: GradCAM Selvaraju et al. (2017), ImageNet Russakovsky et al. (2015)

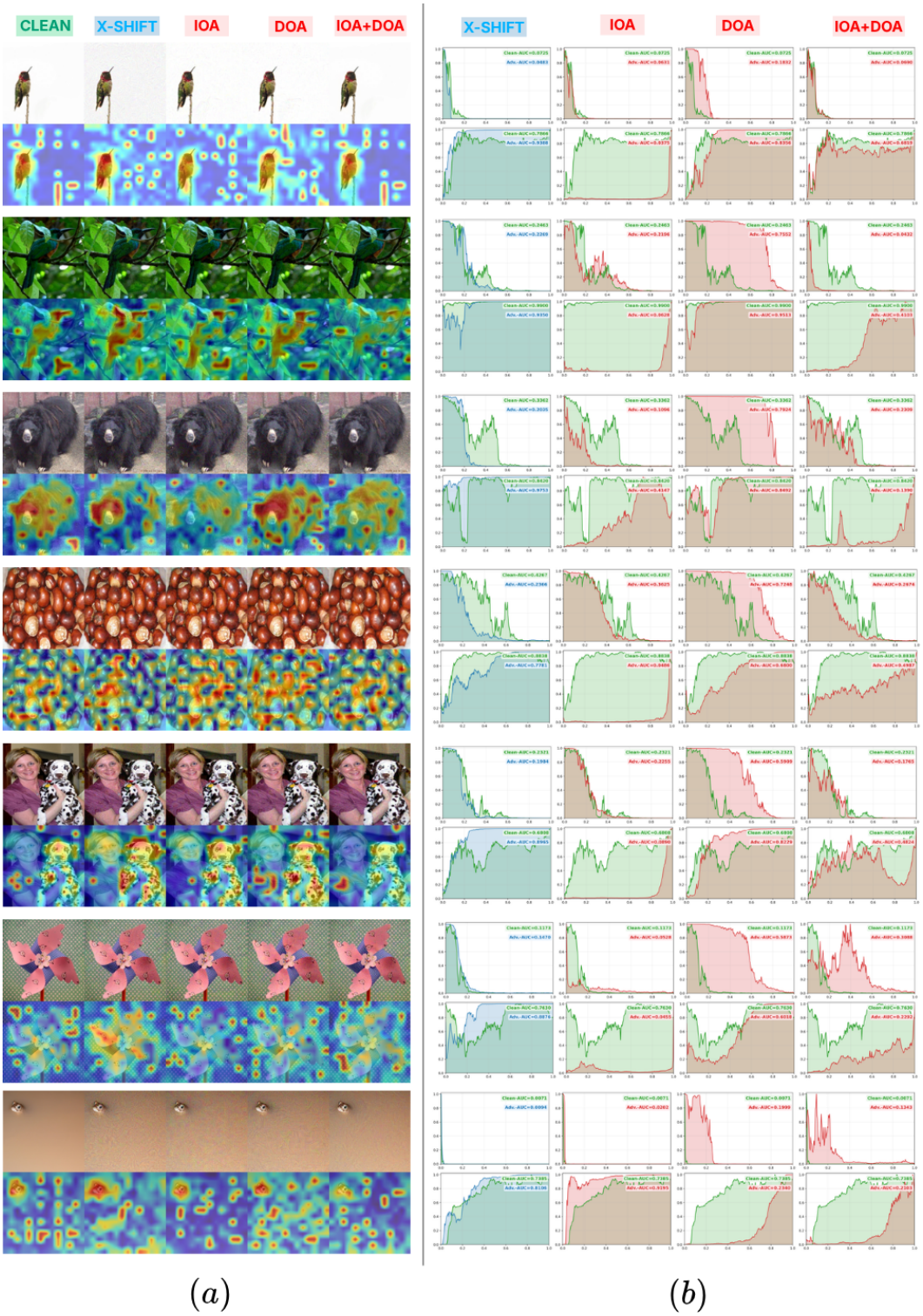

Figure 8: MaskCLIP Dong et al. (2023), ImageNet Russakovsky et al. (2015)

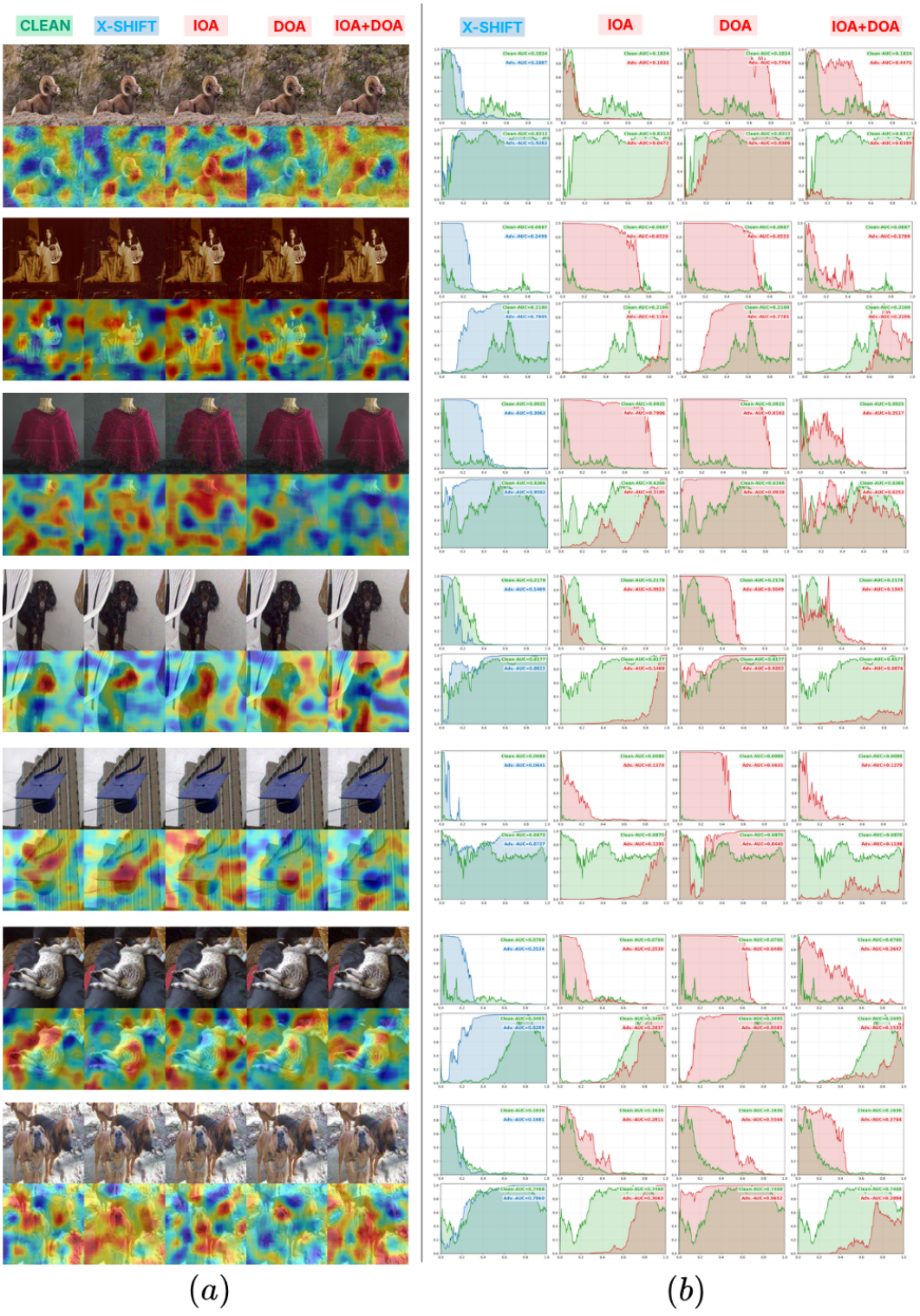

Figure 9: RISE Petsiuk et al. (2018), ImageNet Russakovsky et al. (2015)

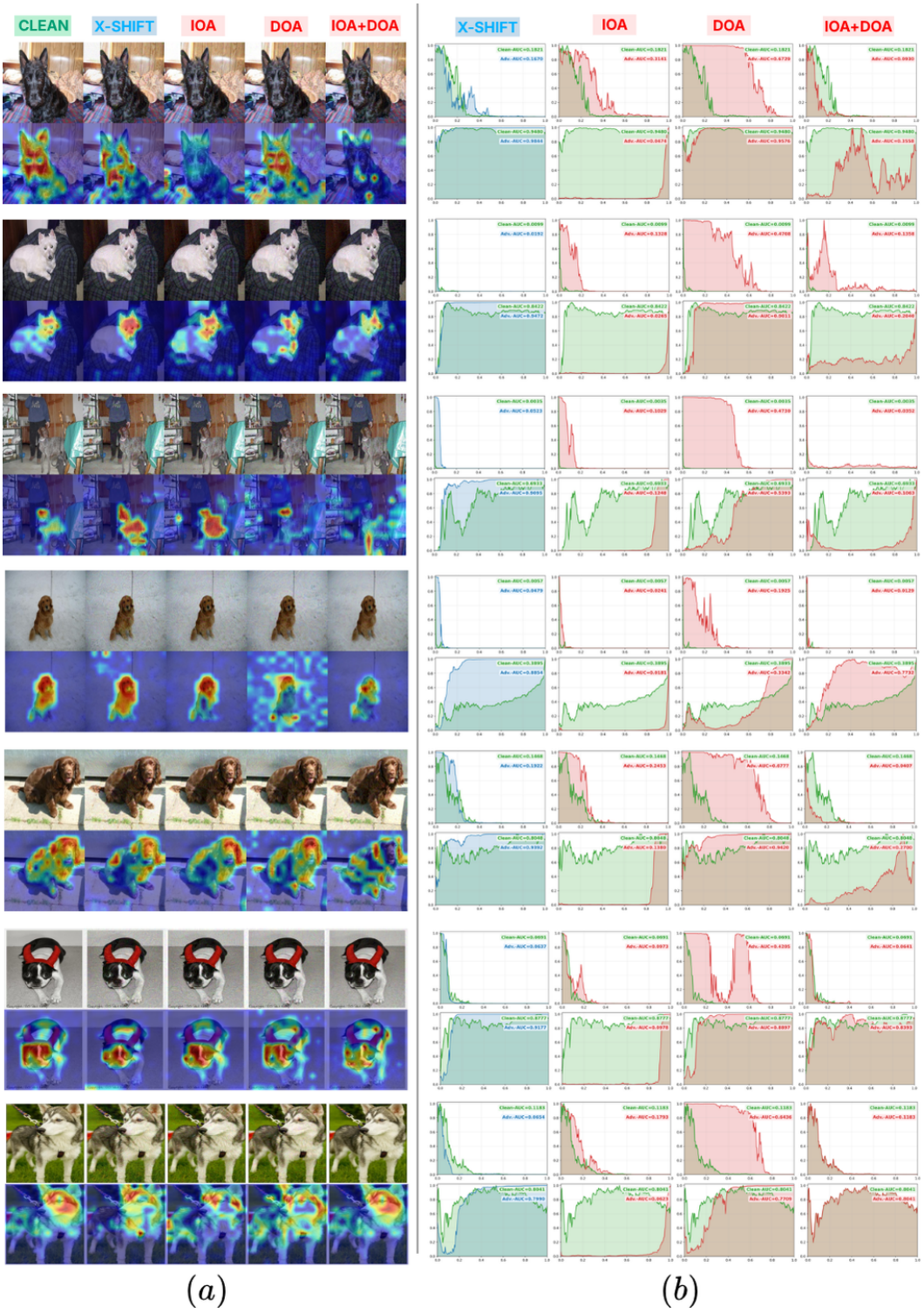

Figure 10: GradECLIP Zhao et al. (2024), Stanford Pet Khosla et al. (2011)

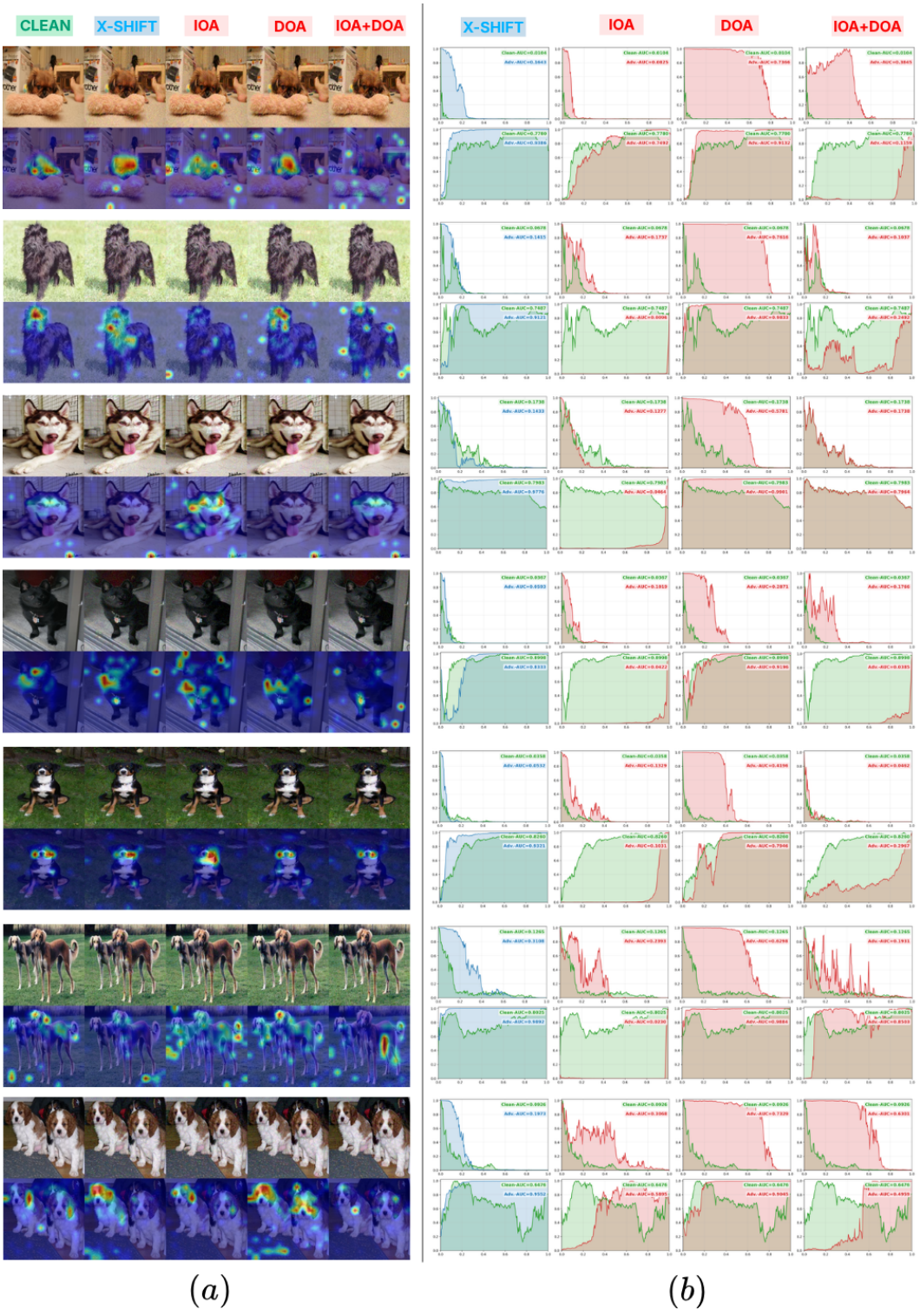

Figure 11: GAME Chefer et al. (2021), Stanford Pet Khosla et al. (2011)

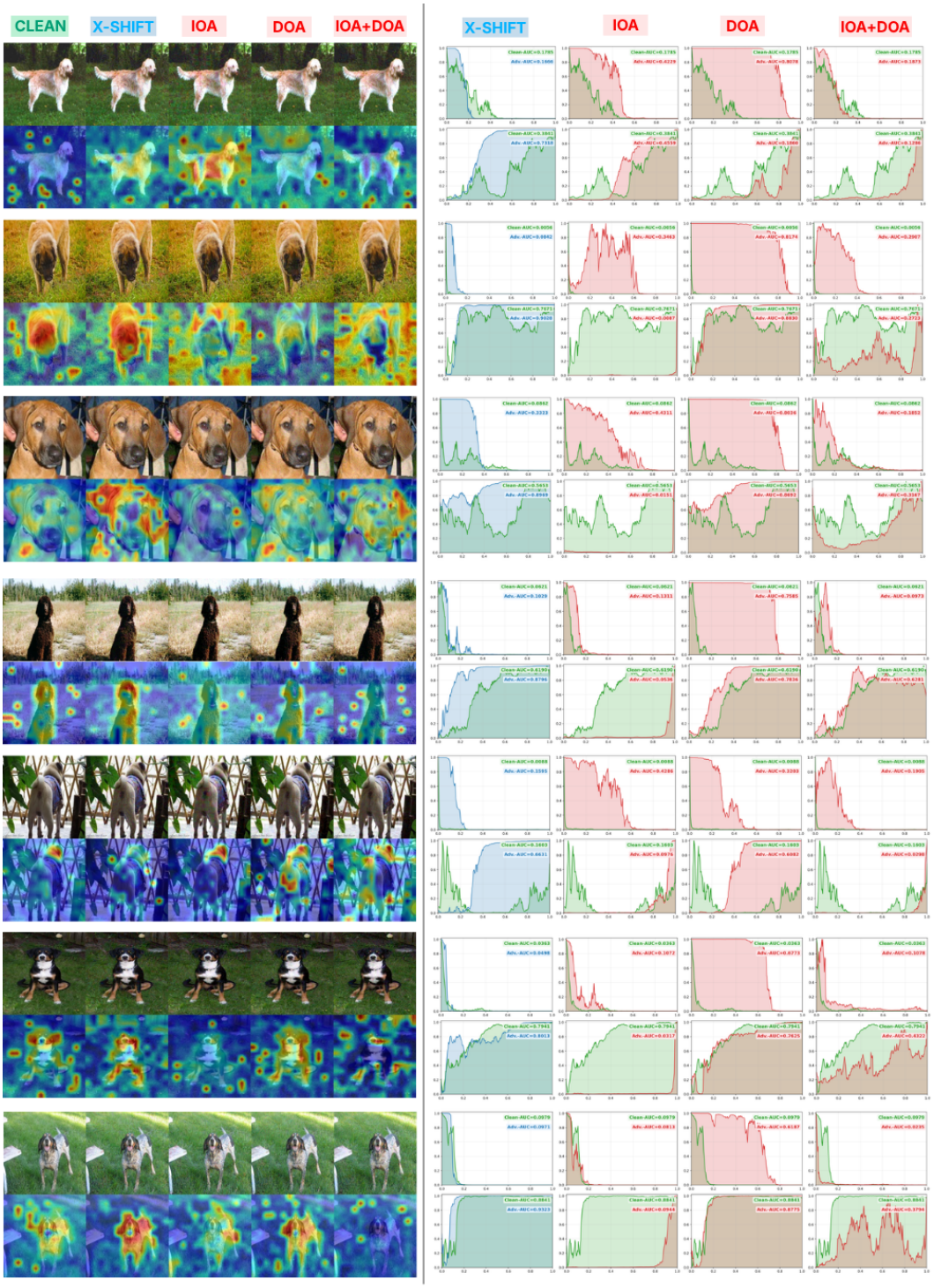

Figure 12: GradCAM Selvaraju et al. (2017), Stanford Pet Khosla et al. (2011)

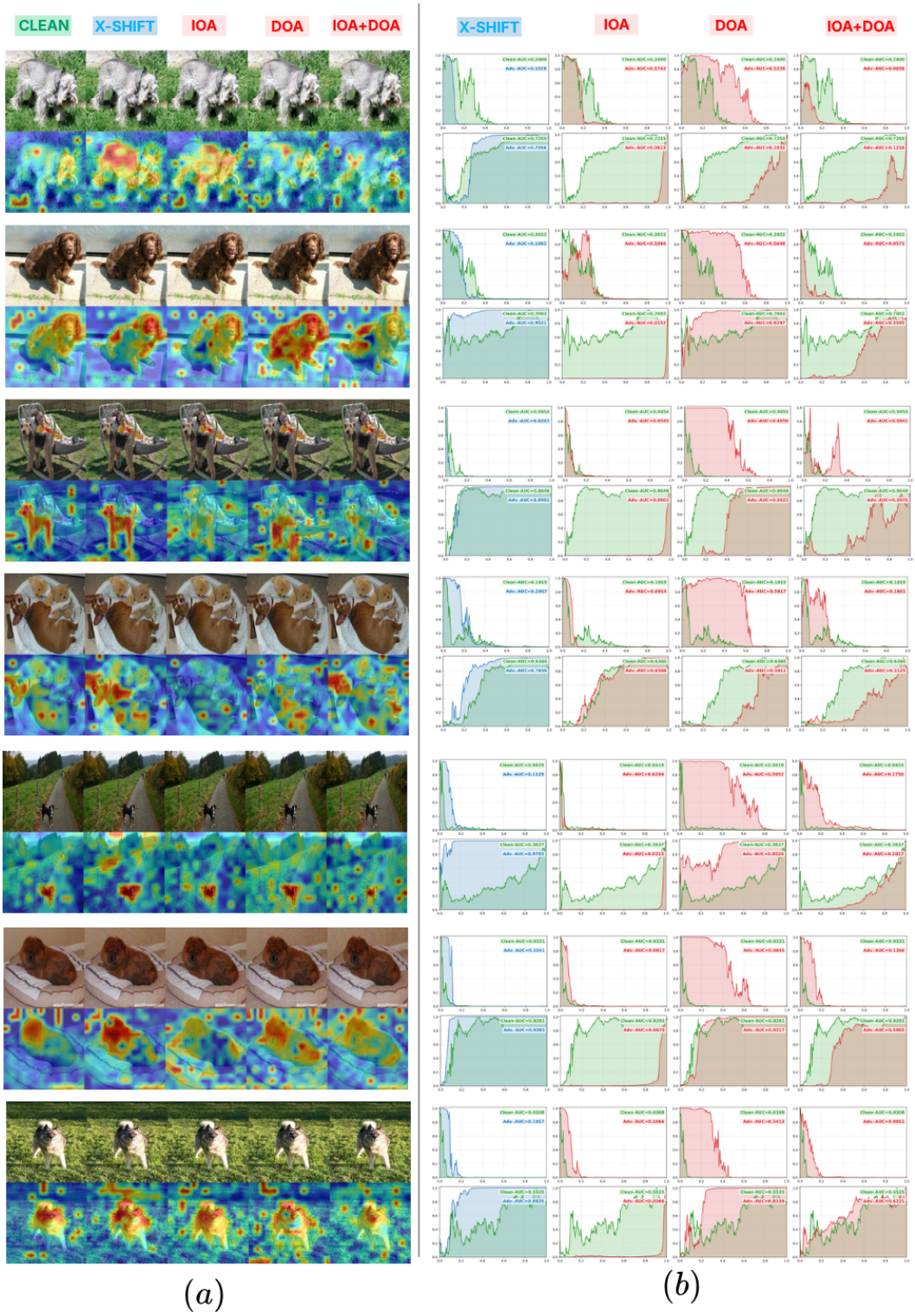

Figure 13: MaskCLIP Dong et al. (2023), Stanford Pet Khosla et al. (2011)

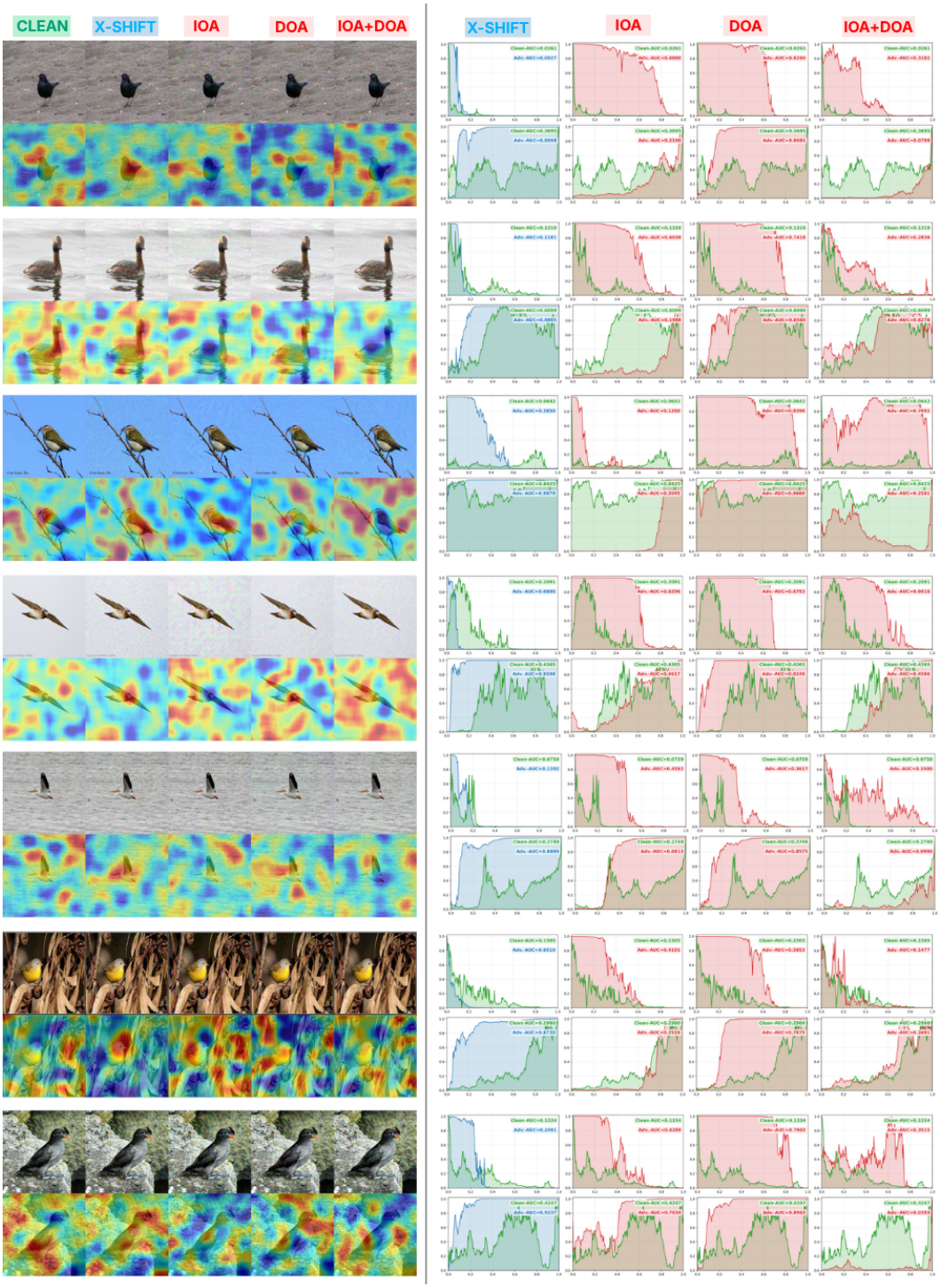

Figure 14: RISE Petsiuk et al. (2018), Stanford Pet Khosla et al. (2011)

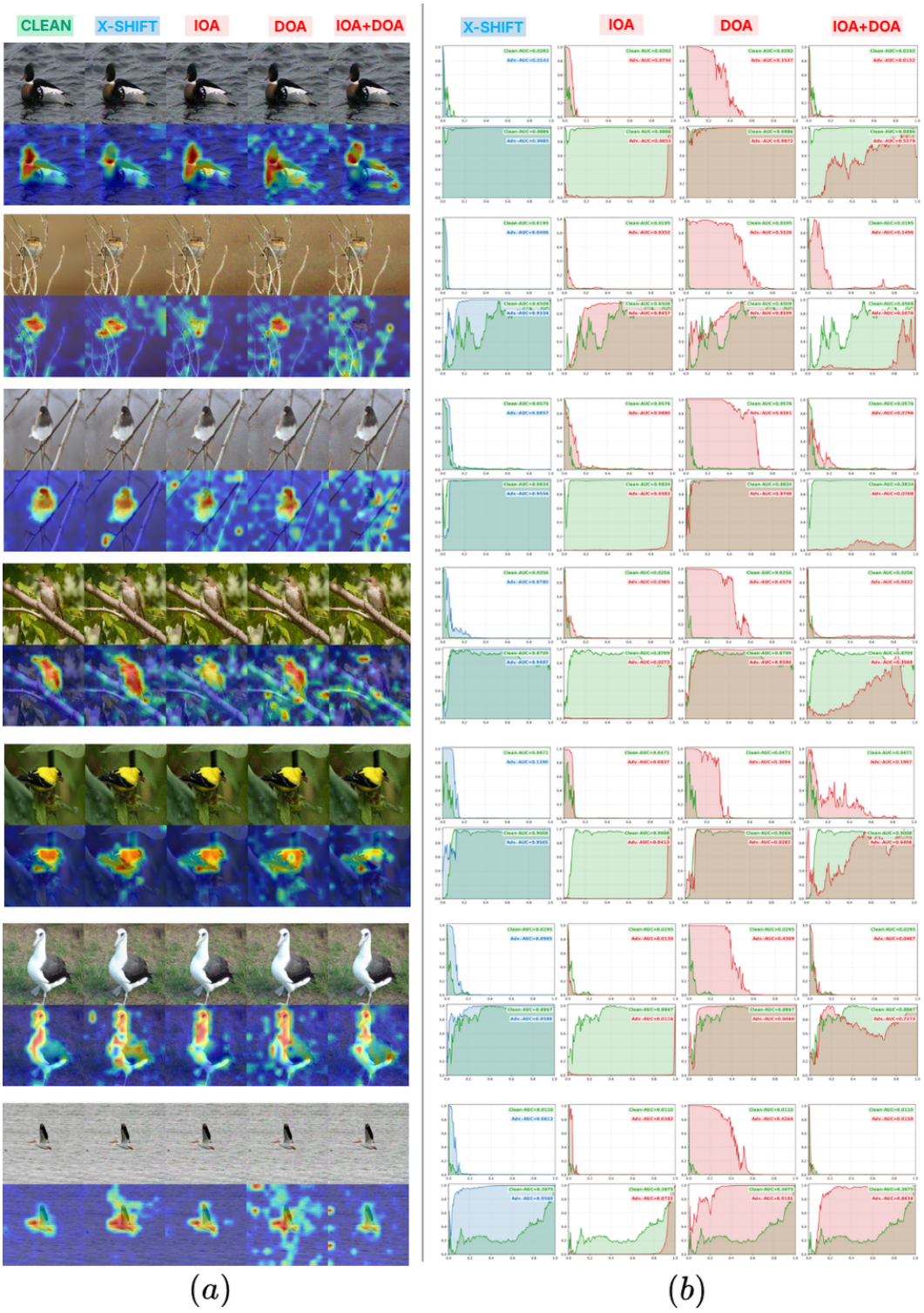

Figure 15: GradECLIP Zhao et al. (2024), CUB Welinder et al. (2010)

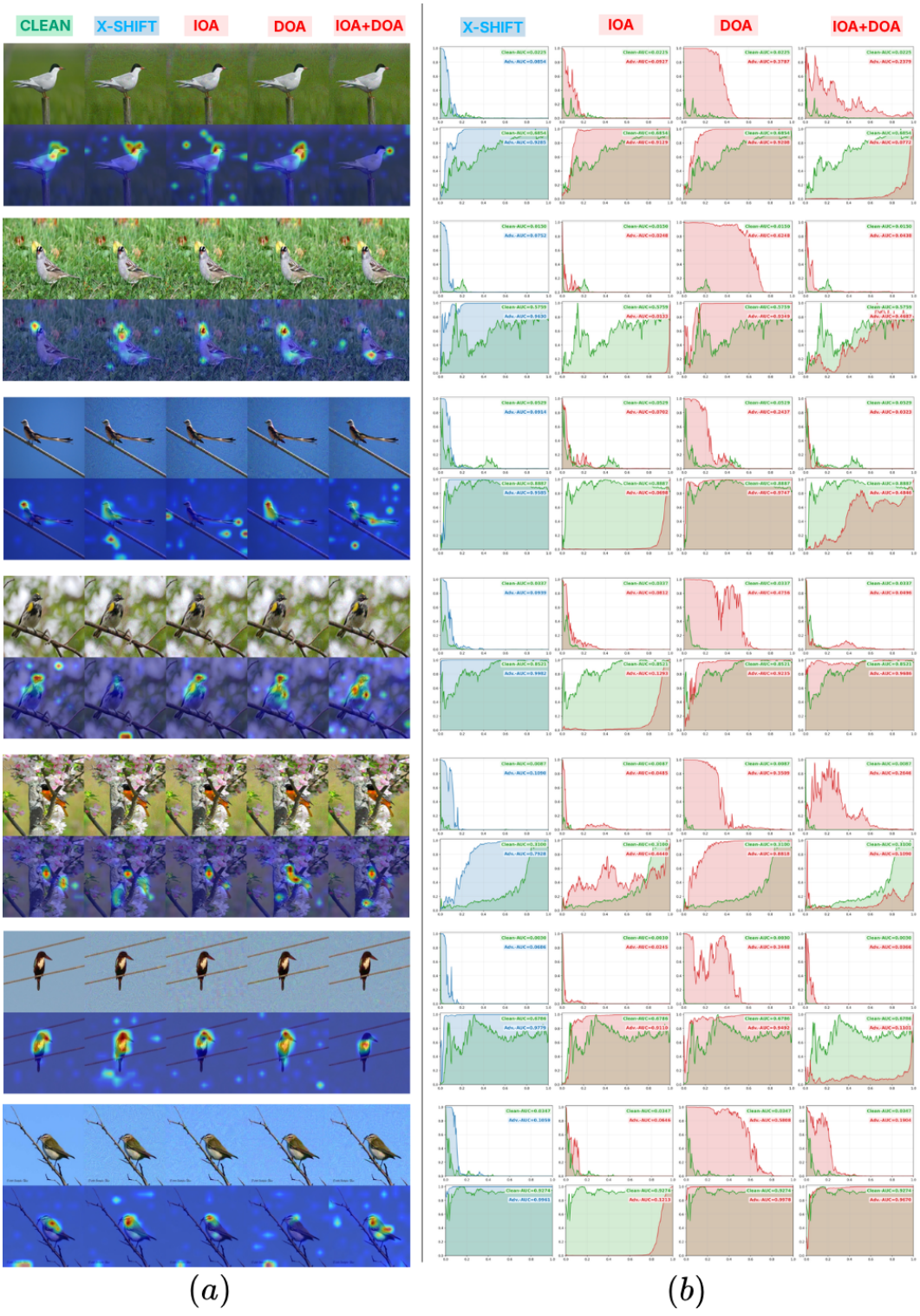

Figure 16: GAME Chefer et al. (2021), CUB Welinder et al. (2010)

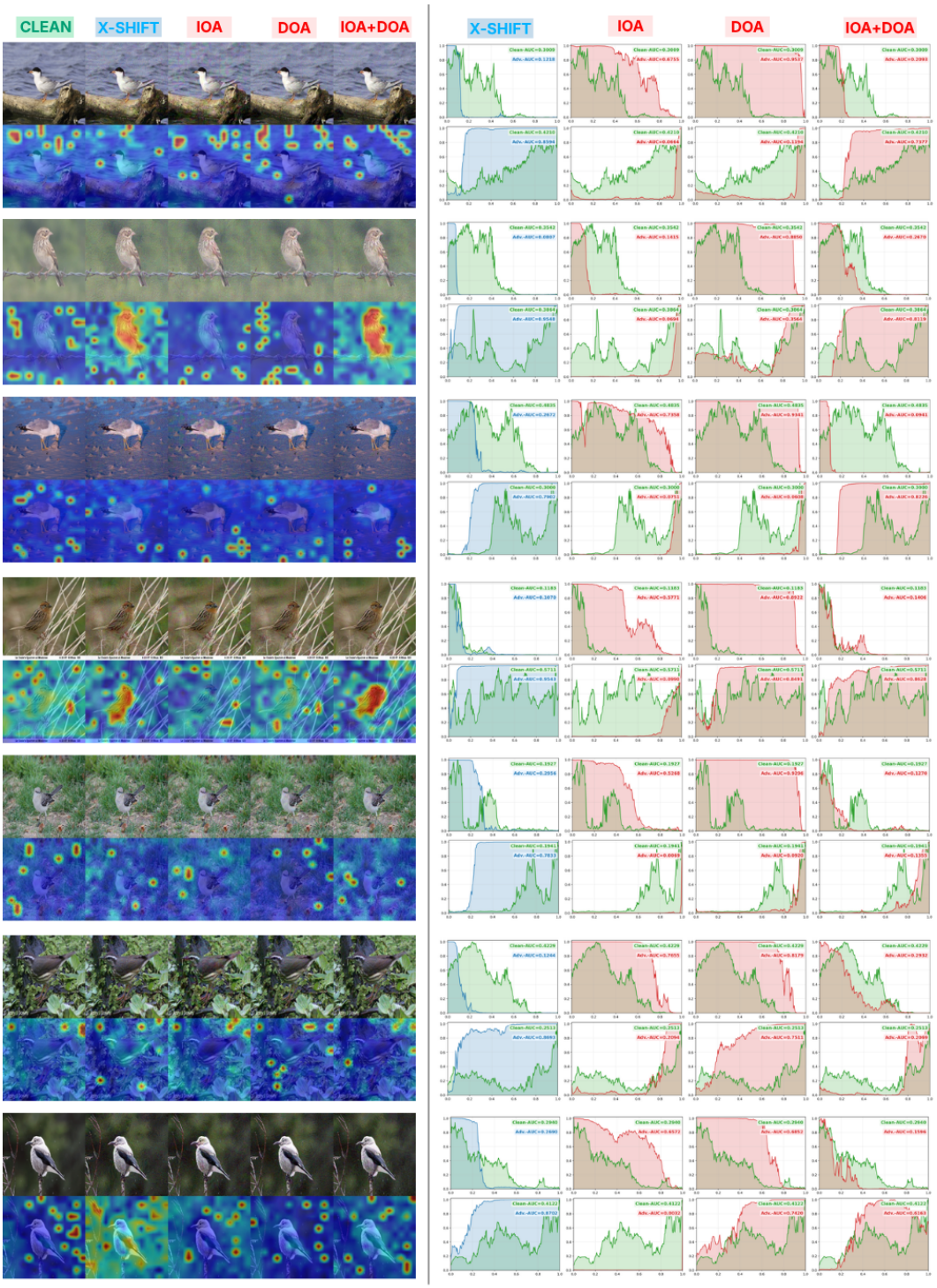

Figure 17: GradCAM Selvaraju et al. (2017), CUB Welinder et al. (2010)

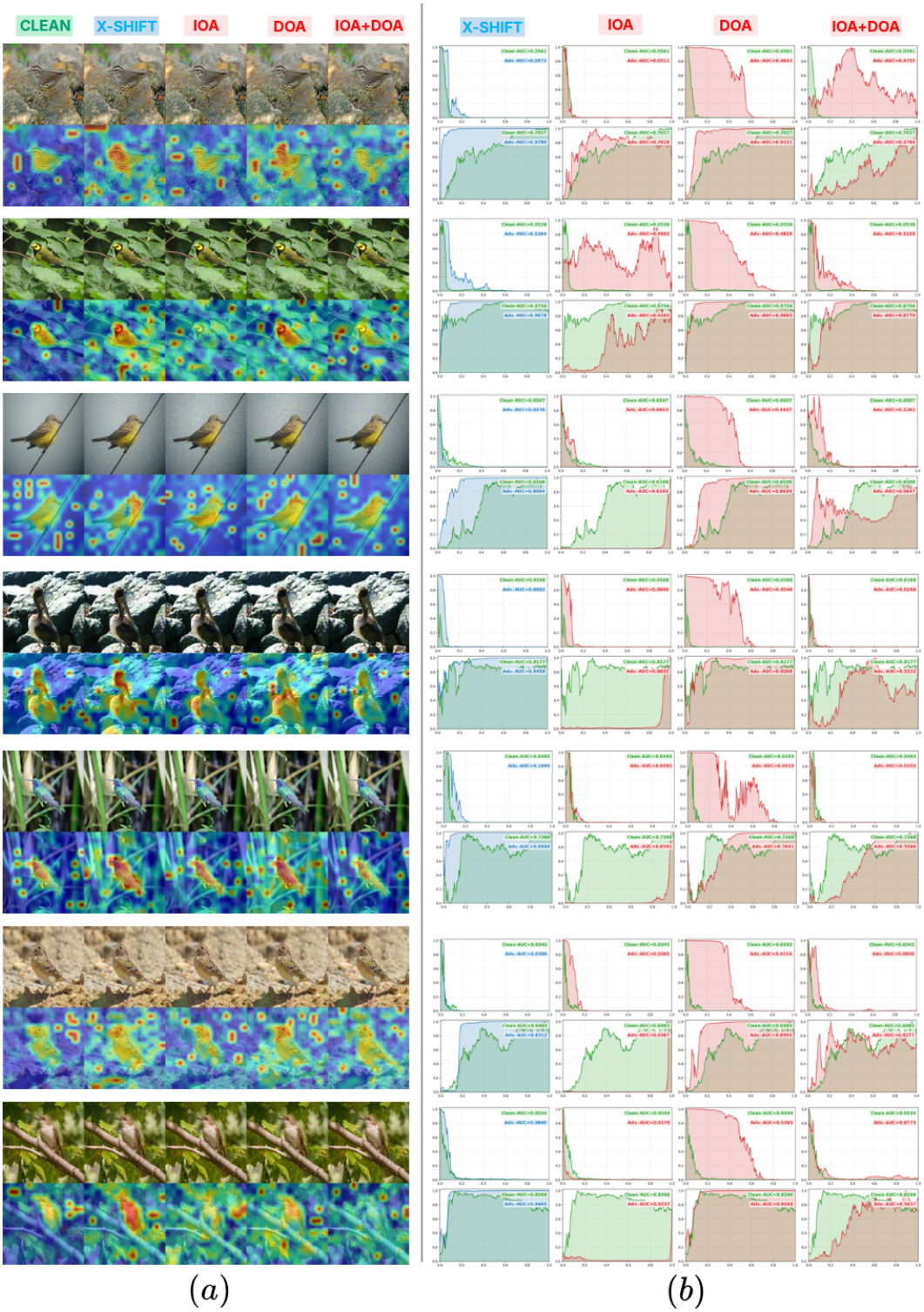

Figure 18: MaskCLIP Dong et al. (2023), CUB Welinder et al. (2010)

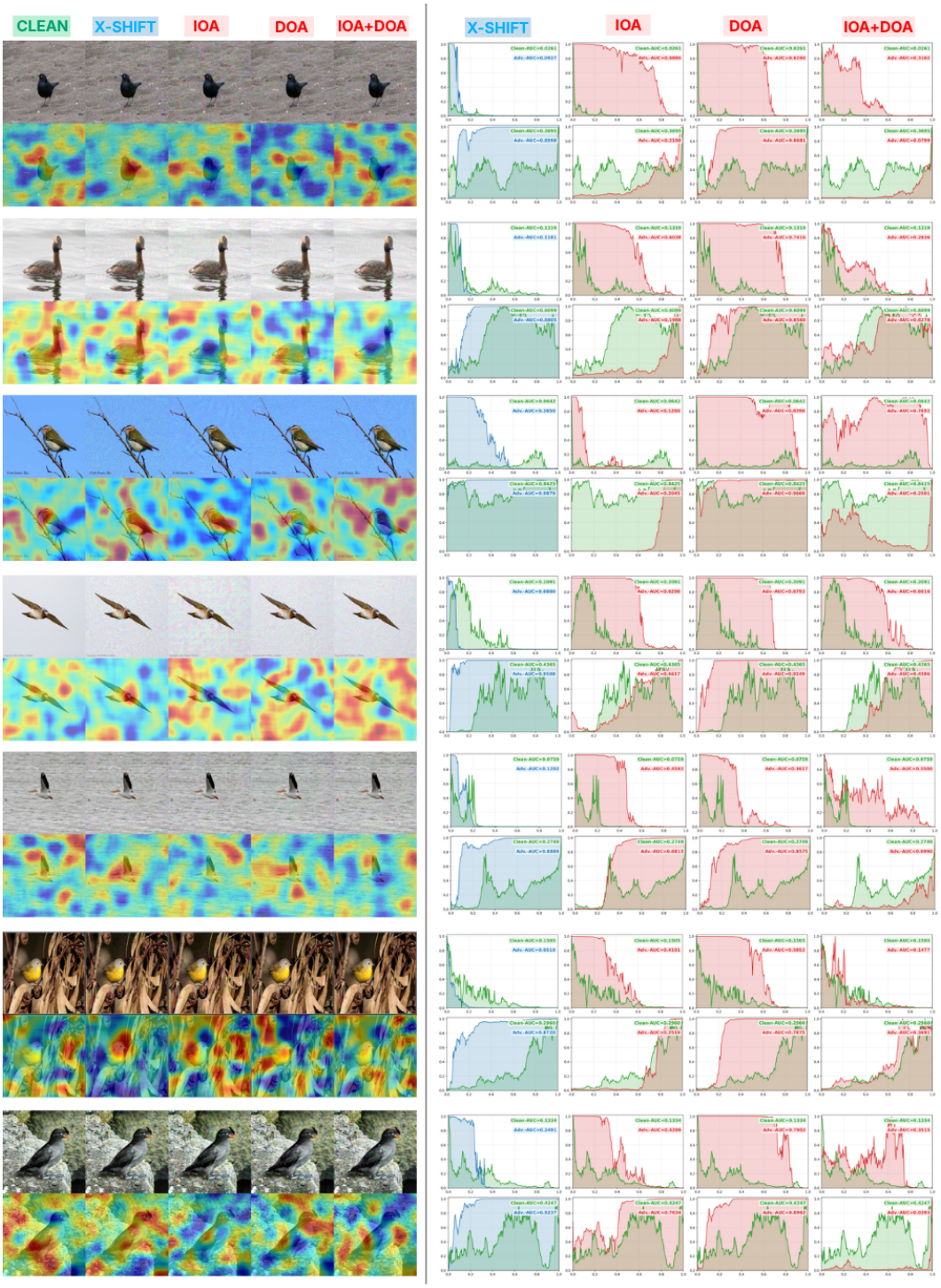

Figure 19: RISE Petsiuk et al. (2018), CUB Welinder et al. (2010)