RIGHT PREDICTIONS, MISLEADING EXPLANATIONS: ON THE VULNERABILITY OF VISION–LANGUAGE MODEL EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Explanation mechanisms are increasingly used to support transparency and trust in vision–language models (VLMs), particularly in settings where model decisions require human oversight. However, the robustness of these explanations remains insufficiently understood. In this work, we investigate whether explanation heatmaps in VLMs, particularly CLIP-based models, remain reliable under adversarial conditions. We show that explanation maps can be systematically manipulated while preserving the model's original prediction, revealing a disconnect between predictive behavior and the evidence the explanation presents. To study this vulnerability, we introduce *X-Shift*, a white-box attack that perturbs the input image to redirect the explanation heatmap of the predicted class onto the region that the explainer associates with a second, attacker-chosen label, without altering the predicted output. Unlike conventional adversarial attacks that aim to induce misclassification, X-Shift specifically targets the integrity of the explanation process itself. The attack operates without modifying model parameters. We evaluate the proposed approach on MS-COCO, Flickr30k and ImageNet-1k with three CLIP backbones, demonstrating that the explanation is relocated onto the target region under small perturbations on image–backbone pairs while the prediction is preserved, with every result compared against random noise of the same budget. Our findings highlight a fundamental limitation of current explanation mechanisms in VLMs and raise concerns about their use as reliable indicators of model trustworthiness in high-impact applications.

1 INTRODUCTION

Vision–language models (VLMs) have rapidly become foundational components in modern AI systems, enabling joint reasoning over visual and textual modalities. Architectures such as CLIP (Radford et al., 2021) are now widely deployed across a range of applications, including healthcare, autonomous systems, and decision-support pipelines. In these settings, explanation methods, commonly referred to as explainable AI (XAI), are frequently used to provide transparency by highlighting input regions that influence model predictions. Such explanations are often treated as a proxy for model reliability, supporting tasks such as auditing, debugging, and trust calibration (Lipton, 2018; Li et al., 2022; Selvaraju et al., 2017; Li et al., 2025).

However, a growing body of work has shown that explanation mechanisms can be highly sensitive to adversarial manipulation. Prior studies demonstrate that carefully crafted perturbations can preserve model predictions while substantially altering explanation outputs, thereby breaking the assumed link between explanations and model reasoning (Kindermans et al., 2019; Ghorbani et al., 2019; Dombrowski et al., 2019; Heo et al., 2019; Slack et al., 2020; Lakkaraju & Bastani, 2020; Huang et al., 2023; Ajalloeian et al., 2023; Kuppa & Le-Khac, 2020). These attacks have primarily been studied in unimodal vision models, focusing on gradient-based or post-hoc explanation methods such as LIME and SHAP.

In contrast, the robustness of explanation mechanisms in VLMs remains largely unexplored. This gap is particularly concerning because explanations in VLMs are not merely diagnostic tools; they are often directly integrated into downstream processes such as visual grounding, reasoning, and

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.

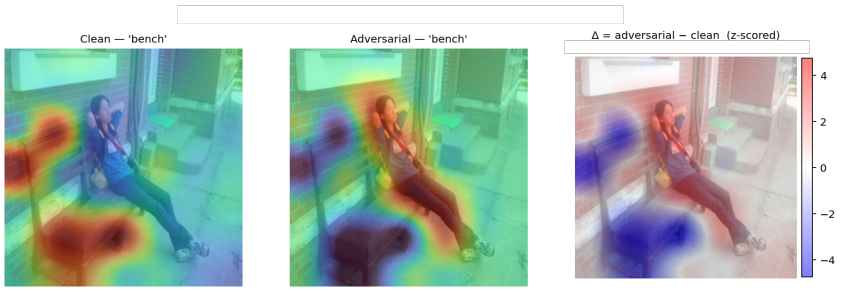

Figure 1: X-Shift on CLIP model (ViT-B/32, $\epsilon = 8/255$). Left: the clean explanation of the predicted class "bench" covers the bench. Middle: after the attack, the explanation of the same, unchanged prediction covers the person, the region the explainer associates with the target label "person". Right, across CLIP backbones. The prediction is unchanged in X-Shift attack scenario

safety verification. In models such as CLIP, explanations are commonly derived from patch–text similarity maps, which serve as a key signal for interpreting model behavior. If these representations are compromised, both human users and automated systems may be systematically misled without any observable degradation in prediction accuracy. Despite this risk, there are currently limited mechanisms to ensure that such explanations remain reliable under adversarial conditions (Baniecki & Biecek, 2024).

To illustrate the potential impact, consider a clinical decision-support scenario in which a VLM-based system assists radiologists in diagnosing pneumonia from chest X-rays. The model correctly predicts *pneumonia* and highlights the affected lung region as supporting evidence. Under adversarial manipulation, an attacker can introduce a subtle perturbation that preserves the prediction while redirecting the explanation heatmap onto a different, clinically unrelated structure in the image. A practitioner relying on the explanation may question the diagnosis or misinterpret the underlying evidence, potentially leading to delayed or incorrect decisions. Notably, such manipulation remains difficult to detect, as standard monitoring mechanisms focus on prediction outputs rather than explanation integrity.

In this paper, we expose a previously underexplored attack surface in VLMs by showing that explanations can be manipulated independently of model predictions. We introduce *X-Shift*, a white-box adversarial attack that perturbs the input image to redirect the explanation heatmap onto the region that the explainer associates with a second label, while preserving the original output. Its design is guided by one observation: explanation maps are normalized per image before they are displayed or evaluated, so X-Shift optimizes only objectives that are unaffected by this normalization. Unlike conventional adversarial attacks that target prediction correctness, X-Shift targets the integrity of explanations, thereby undermining their use as a signal of trust and reliability. The attack does not modify model parameters and is applied at inference time through the input alone, which makes it practical for widely deployed open-weight models such as CLIP. Figure 1 shows an example: the prediction remains "bench" while its explanation moves from the bench to the person.

We evaluate X-Shift across three datasets and three CLIP backbones, demonstrating that it relocates the explanation onto the target region while maintaining prediction consistency, with every result compared against random noise of the same perturbation budget. Our contributions are summarized as follows:

- We introduce X-Shift, a white-box adversarial attack that redirects explanation heatmaps at the patch level while preserving model predictions, using objectives that are invariant to the normalization applied to explanation maps, and we show that this invariance roughly doubles the redirection achieved under the same budget.
- We demonstrate that the attack succeeds on all 45 evaluated image–backbone pairs at $\epsilon = 4/255$ while changing the model's output less than random noise of the same budget, and that most of the effect remains at $\epsilon = 2/255$, where the perturbation is small under SSIM and LPIPS.

- We show that the attack can transfer well across different explanation methods and CLIP backbones.

2 Related Work

The vulnerability of deep neural networks to adversarial perturbations is well established (Huang et al., 2021; Szegedy et al., 2013; Goodfellow et al., 2014; Carlini & Wagner, 2017; Ilyas et al., 2018; 2019; Modas et al., 2019; Babadi et al., 2023; Wang et al., 2024; Croce & Hein, 2019; Madry et al., 2017). Most of this literature focuses on degrading predictive performance through misclassification attacks. Other attacks leave the prediction intact and target a different property of the system, such as its energy consumption and latency through sponge examples, which has in turn motivated dedicated detection methods, including meta-learning-based detectors for sponge attacks in social applications (Babadi & Karimipour, 2026). In the same spirit, a growing line of work studies manipulating model explanations without altering predictions (Baniecki & Biecek, 2024), challenging the assumption that explanations faithfully reflect model reasoning.

Early studies demonstrated the fragility of explanation methods. Kindermans et al. (2019) showed that saliency maps are not invariant to simple input transformations, while Ghorbani et al. (2019) and Dombrowski et al. (2019) demonstrated that imperceptible perturbations can substantially alter attribution maps without affecting predictions. Beyond input-level attacks, model-level manipulation has also been explored. Heo et al. (2019) trained models to produce misleading attributions under methods such as Grad-CAM and LRP, while Slack et al. (2020) showed that grey-box wrappers can arbitrarily control LIME and SHAP outputs, enabling deceptive practices such as fairwashing (Lakkaraju & Bastani, 2020).

More targeted attacks followed. Huang et al. (2023) redirect saliency toward adversary-specified regions while preserving the prediction, Ajalloeian et al. (2023) achieve similar effects with sparse perturbations, and Kuppa & Le-Khac (2020) attack LIME and SHAP in cybersecurity settings.

These attacks are largely restricted to unimodal image classifiers. For vision–language models such as CLIP, whose explanations are typically patch–text similarity maps, their CLIP-Surgery refinement (Li et al., 2025), or attention-based relevance (Chefer et al., 2021; Abnar & Zuidema, 2020), prior security work has focused on prediction robustness, including adversarial examples, poisoning and backdoors (Yang et al., 2024; Zhang et al.; Huang et al., 2025b; Jia et al., 2022; 2021; Koh et al., 2023; Huang et al., 2025a), and has not considered the integrity of these explanations.

To the best of our knowledge, we provide the first systematic demonstration that CLIP-based explanation maps can be manipulated in a targeted manner while preserving the prediction. Our work differs from focus-shifting attacks (Huang et al., 2023) in three respects: the destination is the region the explainer itself associates with a second label, so the manipulated map has the form of a genuine explanation; the attacked explanations are forward-computed patch–text maps rather than gradient saliency; and our white-box objective is invariant to the per-image normalization of explanation maps, which we show matters for attack strength.

3 The Proposed Attack: X-Shift

We introduce **X-Shift**, an explanation-focused adversarial attack that perturbs input images such that model predictions remain unchanged while the explanation heatmap of the predicted class is redirected onto the region associated with a second label. The attack jointly optimizes two groups of objectives, (i) explanation redirection and (ii) prediction preservation, subject to (iii) a perturbation budget and (iv) validity constraints on adversarial examples.

3.1 Threat Model

X-Shift operates under a *white-box* threat model. The adversary has access to the model's forward pass and gradients with respect to the input, for both the deployed model and the explanation method, but does not modify model parameters or access training data. This setting is realistic for publicly available vision–language models such as CLIP and for explanation methods built on them, whose architectures and weights are openly released. The adversary observes the predicted

class and chooses a second, target label whose explanation region lies elsewhere in the image; the goal is that the explanation shown for the unchanged prediction covers that region. The attack is purely input-driven and does not require retraining or fine-tuning, making it applicable in deployed systems.

3.2 Baseline: CLIP Model

CLIP (Radford et al., 2021) aligns an image encoder f_I and text encoder f_T in a shared embedding space. Given an image x and text t, their normalized embeddings are $z_I = f_I(x)/\|f_I(x)\|_2,\ z_T = f_T(t)/\|f_T(t)\|_2$, with similarity $s(x,t) = z_I^\top z_T$. Training minimizes a symmetric contrastive loss over N image–text pairs:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2N}\sum_{i=1}^{N}\left[-\log\frac{\exp(s(x_i,t_i)/\tau)}{\sum_{j=1}^{N}\exp(s(x_i,t_j)/\tau)} - \log\frac{\exp(s(x_i,t_i)/\tau)}{\sum_{j=1}^{N}\exp(s(x_j,t_i)/\tau)}\right], \tag{1}$$

where τ is a learnable temperature. For zero-shot classification over C labels with text embeddings $\{z_{T_c}\}_{c=1}^{C}$, the logits are $\ell_c(x) = s(x,t_c)$ and the prediction is $\hat{y} = \arg\max_c \ell_c(x)$. Our attack perturbs x into $x_{\text{adv}} = x + \delta$, preserving $\hat{y}$ while moving the explanation of $\hat{y}$ onto the region of a target class y_t.

3.3 Explanation Maps and Target Regions

Let $E_c(x) \in \mathbb{R}^P$ denote the explanation map of class c over the P image patches. With normalized patch tokens $z_1, \ldots, z_P$ and class token z_0 from the encoder, it computes

$$E_c(x)_i = w_c\, z_i^\top z_{T_c} - \frac{1}{C}\sum_{c'=1}^{C} w_{c'}\, z_i^\top z_{T_{c'}}, \qquad w_c = C\cdot\text{softmax}_c\big(\gamma\, z_0^\top z_{T_c}\big), \tag{2}$$

where the subtracted term removes features shared by all classes and γ is a fixed temperature. The same attack applies to any explainer whose map is differentiable in x, including raw patch–text similarity and gradient-weighted attention rollout (Abnar & Zuidema, 2020; Chefer et al., 2021).

On the clean image we define the target region $R_t = \text{TopK}_q(E_{y_t}(x))$ and the source region $R_s = \text{TopK}_q(E_{\hat{y}}(x)) \setminus R_t$, where TopK_q returns the indices of the top q fraction of patches. Both regions are fixed before optimization. Because R_t is where the explainer itself places the target label, the redirected map has the form of a genuine explanation of that label rather than an arbitrary pattern.

3.4 Why Normalization Matters

Explanation maps are almost never shown or scored in raw units. They are min–max normalized for display, thresholded at a percentile, or compared by top-k overlap, and all of these operations are unchanged when the map is shifted by a constant or rescaled by a positive factor. An objective that can be improved by such changes therefore spends part of the perturbation budget on differences no auditor sees. A natural raw-score objective, of the kind used in earlier explanation attacks, raises the source-class map on the target region and lowers it elsewhere:

$$\mathcal{L}_{\text{naive}} = -\frac{1}{|R_t|}\sum_{i\in R_t} E_{\hat{y}}(x')_i + \alpha\cdot\frac{1}{P-|R_t|}\sum_{i\notin R_t} E_{\hat{y}}(x')_i. \tag{3}$$

Its gradient with respect to the map has inner product $\alpha - 1 \neq 0$ with the all-ones vector $\mathbf{1}$ for any $\alpha \neq 1$, so part of every step raises or lowers the whole map uniformly, which leaves the displayed map unchanged. X-Shift instead builds every explanation term on the z-scored map $\tilde{E} = (E - \bar{E}\mathbf{1})/\sigma(E)$.

Proposition 1. *Let $g(E) = h(\tilde{E})$ for a differentiable h, at a map E with $\sigma(E) > 0$. Then $\langle\nabla g(E), \mathbf{1}\rangle = 0$ and $\langle\nabla g(E), E\rangle = 0$. Consequently, the descent direction $-\nabla g(E)$ has no component along $\text{span}\{\mathbf{1}, E\}$, the directions along which every normalization that is invariant to positive affine rescaling leaves the displayed map unchanged.*

The proof is given in Section A. The contrast and shape terms below are functions of $\tilde{E}$ only, so every descent step of X-Shift on the explanation acts on the part of the map that normalization preserves.

Algorithm 1 X-Shift Attack: Explanation Redirection on CLIP

Input: clean image x, predicted class $\hat{y}$, target class y_t, budget ϵ, learning rate η, iterations N
Output: adversarial image x^{adv}
$R_t \leftarrow \mathrm{TopK}_q(E_{y_t}(x)), \quad R_s \leftarrow \mathrm{TopK}_q(E_{\hat{y}}(x)) \setminus R_t$
$p_{\mathrm{ref}} \leftarrow \mathrm{softmax}(\tilde{E}_{y_t}(x)/T)$; record $\ell(x), p(x), z_I(x), \Delta(x)$
Initialize $\delta \leftarrow 0$
for $i = 1$ to N **do**
  $x' \leftarrow \mathrm{clip}(x + \delta, 0, 1)$
  $\tilde{E}' \leftarrow \text{z-score}\big(E_{\hat{y}}(x')\big)$
  $\mathcal{L} \leftarrow \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}} + \lambda_{\mathrm{shape}}\mathcal{L}_{\mathrm{shape}} + \mathcal{L}_{\mathrm{pres}}$
  $\delta \leftarrow \mathrm{Adam}(\delta, \nabla_\delta \mathcal{L}, \eta_i)$ ▷ cosine schedule on η_i
  $\delta \leftarrow \mathrm{clip}(\delta, -\epsilon, \epsilon)$
end for
return $x^{adv} = \mathrm{clip}(x + \delta, 0, 1)$

3.5 X-Shift Attack Objectives

We combine the following complementary objectives to achieve explanation-focused adversarial perturbations. Let $x' = \mathrm{clip}(x + \delta, 0, 1)$ and $\tilde{E}' = \tilde{E}_{\hat{y}}(x')$ the z-scored explanation of the predicted class on the perturbed image.

Explanation redirection. A contrast term moves explanation mass from the source region to the target region:

$$\mathcal{L}_{\mathrm{con}} = \frac{1}{|R_s|} \sum_{i \in R_s} \tilde{E}'_i \; - \; \frac{1}{|R_t|} \sum_{i \in R_t} \tilde{E}'_i. \tag{4}$$

A shape term matches the full spatial distribution of the redirected map to the clean explanation of the target label:

$$\mathcal{L}_{\mathrm{shape}} = \mathrm{KL}\Big(p_{\mathrm{ref}} \,\Big\|\, \mathrm{softmax}(\tilde{E}'/T)\Big), \qquad p_{\mathrm{ref}} = \mathrm{softmax}\big(\tilde{E}_{y_t}(x)/T\big), \tag{5}$$

with temperature T. The contrast term only moves mass into R_t and leaves its shape unconstrained, whereas the shape term makes the redirected map look like the explanation of the target label; Section D shows that the shape term carries most of the effect.

Prediction preservation. Keeping only the argmax is not sufficient for stealth: the output probabilities, the image embedding and the decision margin can all change visibly while $\hat{y}$ is unchanged. We therefore anchor all three to their clean values:

$$\begin{aligned}\mathcal{L}_{\mathrm{pres}} = \lambda_\ell \, \|\ell(x') - \ell(x)\|_2^2 + \lambda_{\mathrm{KL}} \, \mathrm{KL}\big(p(x) \,\|\, p(x')\big) + \lambda_{\cos} \left(1 - \cos(z_I(x), z_I(x'))\right) \\ + \lambda_m \left[\kappa \, \Delta(x) - \Delta(x')\right]_+,\end{aligned} \tag{6}$$

where $p(\cdot) = \mathrm{softmax}(\ell(\cdot)/\tau)$ uses CLIP's logit scale, $\Delta(x) = \ell_{\hat{y}}(x) - \max_{c \neq \hat{y}} \ell_c(x)$ is the decision margin, and $\kappa < 1$ allows a small reduction of the margin.

Budget and validity constraints. The perturbation is bounded in the ℓ_∞ norm, $\|\delta\|_\infty \leq \epsilon$, and the adversarial image must remain in the valid input domain, $x_{\text{adv}} \in [0, 1]^d$.

The total objective combines explanation redirection with prediction preservation:

$$\mathcal{L} = \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}} + \lambda_{\mathrm{shape}}\mathcal{L}_{\mathrm{shape}} + \mathcal{L}_{\mathrm{pres}}, \tag{7}$$

where λ_{con} and λ_{shape} weight the two redirection terms and the coefficients inside $\mathcal{L}_{\mathrm{pres}}$ weight the preservation terms. All explanation terms act on z-scored maps and are therefore invariant to the normalization discussed in Section 3.4. Hyperparameter values are listed in Section E.

Explainability Attack Algorithm. We generate adversarial examples via iterative gradient-based optimization to redirect explanation maps while preserving the original prediction. The procedure is summarized in Algorithm 1.

4 EXPERIMENTS

Our experimental evaluation is designed to answer the following research questions:

- **RQ1:** Does X-Shift move the explanation of the predicted class onto the target region while preserving the prediction, and by how much more than random noise of the same budget?
- **RQ2:** Does this hold across CLIP backbones and datasets, and what is the perceptual cost of the perturbation?
- **RQ3:** Does the perturbation transfer to explanation methods and backbones it was not optimized against?
- **RQ4:** Which components of the objective are responsible for the effect, and how does X-Shift differ from standard prediction-targeted attacks?

4.1 EXPERIMENTAL SETUP

Models and datasets. We attack three CLIP backbones, ViT-B/16, ViT-B/32 and ViT-L/14 (Radford et al., 2021) from CLIP model structure. All inputs are resized to 224×224, which gives patch grids of 14×14, 7×7 and 16×16, respectively. We use three image sources: the 2017 unlabeled split of MS-COCO (Chen et al., 2015), the 31,783 images of Flickr30k (Young et al., 2014), and the validation split of ImageNet-1k (Deng et al., 2009). For MS-COCO and Flickr30k the label set is a fixed vocabulary of 66 object and scene categories; for ImageNet-1k it is the 1,000 ImageNet class names. Text embeddings use CLIP's prompt ensemble. The attack uses no labels, so the unlabeled MS-COCO split is sufficient.

Image and target selection. Images are not hand-picked. For each dataset and backbone we shuffle the image folder with a fixed seed and accept the first five images that pass three screening rules. (i) The clean prediction must be decisive: the gap between the top two class probabilities under CLIP's logit scale must be at least 0.5, so that preserving the prediction is a meaningful constraint rather than a near-tie that noise would also preserve. (ii) The target label is chosen automatically as the label, among clean-ranking positions 4 to 25, whose clean explanation region is most disjoint from the source region; the top three positions are excluded so that the target is not a near competitor of the prediction. (iii) The clean source and target regions must overlap by at most 0.25 IoU, so that any measured shift was not already present before the attack. In total we screened 103 images to obtain the 45 image–backbone pairs used below; every rejection was caused by rule (i). Per-cell pass rates and runtimes are given in Table 5 (Section F).

Explainers. The attacked explainer is designed from (Equation (2)). To measure transfer, we additionally compute raw patch–text similarity, gradient-weighted attention rollout (Abnar & Zuidema, 2020; Chefer et al., 2021) and ScoreCAM (Wang et al., 2020) on the same images. ScoreCAM is query-based and is never differentiated through by any variant of the attack.

Implementation. Unless stated otherwise, $\epsilon = 4/255$ in pixel space, the region fraction is $q = 0.15$, and the optimizer is Adam with learning rate $2/255$ and a cosine schedule for 300 steps. The loss weights are $\lambda_{\text{con}} = 1$, $\lambda_{\text{shape}} = 1$ with $T = 0.5$, $\lambda_\ell = 20$, $\lambda_{\text{KL}} = 5$, $\lambda_{\cos} = 5$, $\lambda_m = 50$ and $\kappa = 0.95$; the CLIP explanation temperature is $\gamma = 2$. The same hyperparameters are used for every dataset and backbone. One attack takes 12–13 s on ViT-B/32, 23–27 s on ViT-B/16 and 55–64 s on ViT-L/14 on a single NVIDIA A100 GPU (Table 5).

Metrics. Let Top_k denote the top 25% of patches of a map. We report $\text{IoU}_{\to t}$, the IoU between Top_k of the adversarial source-class map and Top_k of the clean target-class map (higher means the explanation moved onto the target), and $\text{IoU}_{\to s}$, the IoU with the clean source-class map (lower means it left the original region). Mass_{R_t} is the share of the softmax-normalized adversarial map inside R_t; a uniform map gives $|R_t|/P \approx 0.14$–0.15. Prediction preservation is measured by the flip rate, the cosine similarity of clean and adversarial image embeddings (CosSim), and the largest change in any class probability (MaxΔP). Perceptibility is measured by SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018) and PSNR. Formal definitions are given in Section B.

Controls and success. Every attacked image is compared with five uniform random perturbations of the same ℓ_∞ budget, evaluated with the same metrics. An attack counts as a *success* only if the prediction is unchanged and $\text{IoU}_{\to t}$ exceeds both the clean overlap of the source and target maps

Table 1: X-Shift on CLIP-Surgery. Success requires an unchanged prediction and $\mathrm{IoU}_{\to t}$ above both the clean overlap and the noise mean plus one standard deviation. $\mathrm{IoU}_{\to t}$ and $\mathrm{IoU}_{\to s}$ compare the top-25% patches of the adversarial source-class map with the clean target and clean source maps. Mass_{R_t} is the share of the softmax-normalized map inside the target region.

Dataset	Backbone	$\mathrm{IoU}_{\to t}\uparrow$	$\mathrm{IoU}_{\to s}\downarrow$	$\mathrm{Mass}_{R_t}\uparrow$	CosSim↑	MaxΔP↓	SSIM↑	LPIPS↓
MS-COCO	ViT-B/16	0.92±0.11 (0.01)	0.01 (0.86)	0.55 (0.06)	0.995 (0.980)	0.0004 (0.0342)	0.954 (0.973)	0.027 (0.004)
	ViT-B/32	0.85±0.10 (0.05)	0.04 (0.78)	0.50 (0.08)	0.989 (0.983)	0.0012 (0.0313)	0.954 (0.973)	0.028 (0.003)
	ViT-L/14	0.82±0.09 (0.02)	0.01 (0.75)	0.54 (0.06)	0.986 (0.962)	0.0017 (0.0615)	0.949 (0.970)	0.041 (0.008)
Flickr30k	ViT-B/16	0.93±0.03 (0.04)	0.04 (0.84)	0.53 (0.06)	0.995 (0.977)	0.0004 (0.0345)	0.967 (0.981)	0.016 (0.003)
	ViT-B/32	0.97±0.07 (0.06)	0.06 (0.91)	0.48 (0.08)	0.991 (0.986)	0.0013 (0.0292)	0.972 (0.984)	0.026 (0.002)
	ViT-L/14	0.75±0.21 (0.02)	0.03 (0.83)	0.56 (0.05)	0.986 (0.975)	0.0019 (0.0754)	0.967 (0.980)	0.021 (0.003)
ImageNet-1k	ViT-B/16	0.84±0.15 (0.07)	0.06 (0.86)	0.54 (0.07)	0.992 (0.982)	0.0004 (0.0260)	0.948 (0.969)	0.027 (0.006)
	ViT-B/32	0.85±0.14 (0.08)	0.05 (0.92)	0.51 (0.08)	0.992 (0.981)	0.0007 (0.0548)	0.952 (0.971)	0.036 (0.006)
	ViT-L/14	0.87±0.08 (0.03)	0.03 (0.85)	0.52 (0.06)	0.982 (0.980)	0.0010 (0.0145)	0.954 (0.973)	0.023 (0.005)

and the mean plus one standard deviation of $\mathrm{IoU}_{\to t}$ under the random controls. With five images per cell, we report counts and means and do not make claims of statistical significance.

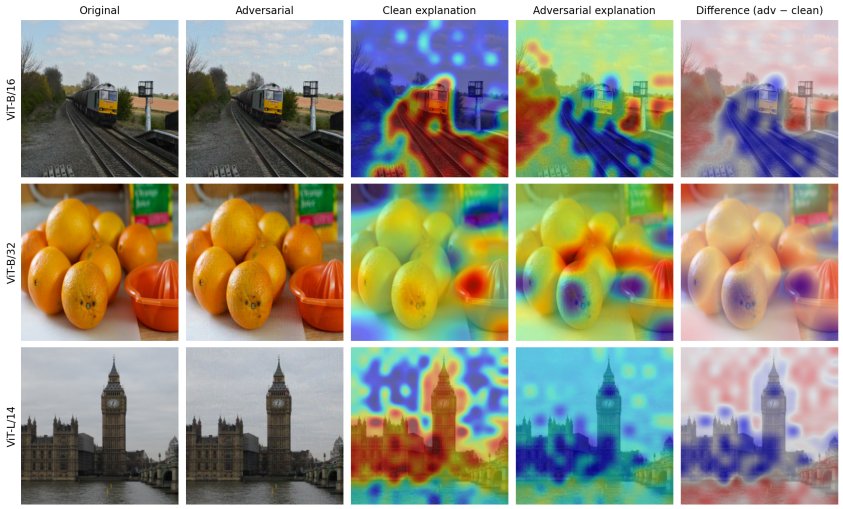

Figure 2: X-Shift on MS-COCO across CLIP backbones. Columns: original and adversarial image, explanation of the predicted class on each, and their difference. The prediction is unchanged in X-Shift attack scenario.

4.2 Explanation Redirection with Preserved Predictions

Table 1 reports the main results. X-Shift succeeds on all 45 image–backbone pairs, and no prediction changes. In every cell the adversarial explanation of the unchanged prediction covers the target region ($\mathrm{IoU}_{\to t}$ between 0.75 and 0.97), while random noise of the same budget leaves it where it was ($\mathrm{IoU}_{\to t}$ between 0.01 and 0.08). The original region is almost entirely vacated: $\mathrm{IoU}_{\to s}$ falls to 0.01–0.06, compared with 0.75–0.92 under noise. The redirected map is also concentrated rather than spread out, with about half of its mass inside R_t (0.48–0.56), more than three times the uniform level. Figure 2 shows one clean and adversarial image with their explanation maps for each backbone on MS-COCO, and Figures 3 and 4 show the same for Flickr30k and ImageNet-1k. In each case the example is the image closest to the median $\mathrm{IoU}_{\to t}$ of its cell rather than the best case.

The attack changes the model's output less than random noise does. Across all nine cells, the largest change in any class probability is between 0.0004 and 0.0019 under X-Shift and between 0.0145 and 0.0754 under noise, and the image embedding stays at least as close to the clean one (CosSim 0.982–0.995 versus 0.962–0.986). A monitor that watches the model's outputs or embeddings therefore has less to detect under X-Shift than under unstructured noise of the same size.

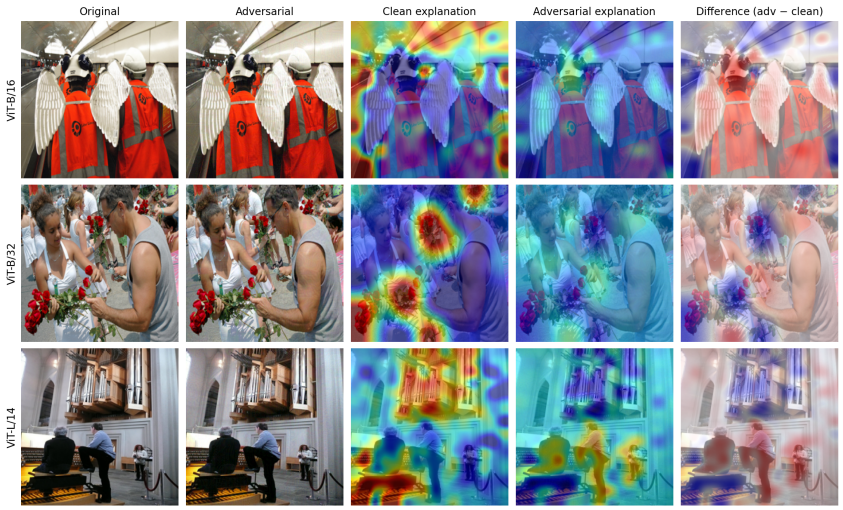

Figure 3: X-Shift on Flickr30k across CLIP backbones. Columns: original and adversarial image, explanation of the predicted class on each, and their difference. The prediction is unchanged in X-Shift attack scenario.

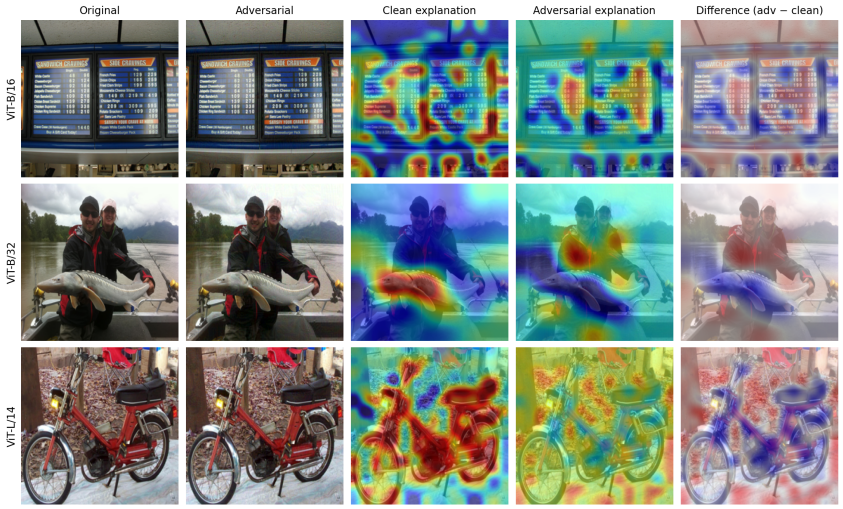

Figure 4: X-Shift on ImageNet-1k across CLIP backbones. Columns: original and adversarial image, explanation of the predicted class on each, and their difference. The prediction is unchanged in X-Shift attack scenario.

The perturbation is small but not free. At $\epsilon = 4/255$, SSIM is 0.948–0.972 and LPIPS is 0.016–0.041, compared with 0.969–0.984 and 0.002–0.008 for noise of the same budget, and PSNR is about 38 dB. The attack is thus measurably more visible than random noise under both perceptual metrics, although absolute values remain low. Section D shows that most of the effect survives at $\epsilon = 2/255$, where LPIPS drops to 0.008.

The effect is consistent across backbones and datasets. The weakest cell is Flickr30k on ViT-L/14 ($\mathrm{IoU}_{\to t} = 0.75 \pm 0.21$), where one image was only partially redirected but still exceeded both floors. Screening removed more Flickr30k images than MS-COCO or ImageNet-1k images (Table 5), because many Flickr30k scenes are not decisively described by a single label of the vocabulary. On ImageNet-1k, the automatically selected target is usually a class absent from the image, so the attack removes the explanation from the predicted object and places it on background rather than on a second object (Figure 4).

4.3 Transferability and Ablation

A perturbation crafted against CLIP transfers: the explanations of other explainers moved onto the target region on 6 (patch similarity), 4 (attention rollout) and 9 (ScoreCAM) of the 45 pairs, and those of other backbones on 20 of 85 cases (Section C). The ablation (Section D) shows that the shape term drives the redirection ($\mathrm{IoU}_{\to t}$ 0.46 without it versus 0.86), that the non-invariant objective of Equation (3) reaches only 0.45 under the same budget, and that protecting only the argmax keeps the prediction but lowers CosSim from 0.990 to 0.768. The attack still reaches 0.81 at $\epsilon = 2/255$, and same-budget PGD flips every prediction without redirecting the explanation ($\mathrm{IoU}_{\to t} \leq 0.07$).

5 Conclusion

We studied whether the explanation maps used to inspect CLIP-based models can be manipulated independently of their predictions. X-Shift, a white-box attack with objectives that are invariant to the normalization applied to explanation maps, moves the CLIP explanation of the predicted class onto the region associated with a second label on all 45 evaluated image–backbone pairs across three datasets and three backbones without changing any prediction. The attack changes the model's output less than random noise of the same budget, and normalization invariance roughly doubles the redirection achieved by a raw-score objective. Standard prediction-targeted attacks, in contrast, flip the prediction without redirecting the explanation.

These results indicate that an explanation that looks plausible, and that accompanies a correct and confident prediction, is not evidence that the explanation reflects the input faithfully when an adversary controls the input. Evaluations of explanation methods for VLMs should therefore include adversarial settings in which the prediction is held fixed, and defenses need to be tested against adaptive attackers that target the explanation directly.

Ethics Statement

This work studies a vulnerability in explanation methods for vision–language models, in which an adversary can redirect explanation outputs while preserving model predictions. We consider disclosure of this vulnerability important for evaluating the reliability of explanations used in model auditing and analysis. The proposed attack assumes white-box access to the model and is evaluated on publicly available datasets. We will release the code to facilitate reproducibility and to support the evaluation of future explanation methods and defenses under this threat model. No human subjects were involved in this work, and no new personally identifiable or sensitive data were collected.

Reproducibility Statement

ode, configurations and the list of screened images will be released. All hyperparameters are in Section E, the screening procedure in Section 4.1, and every table is generated directly from the logged per-image results.

AI Use Statement

A large language model was used to assist with editing and improving the clarity of the manuscript. The authors independently reviewed and verified the manuscript, including the experimental design, implementation, results, interpretations, and claims, and take full responsibility for the content of the paper.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4190–4197, 2020.

Ahmad Ajalloeian, Seyed Mohsen Moosavi-Dezfooli, Michalis Vlachos, and Pascal Frossard. Sparse attacks for manipulating explanations in deep neural network models. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 918–923. IEEE, 2023.

Narges Babadi and Hadis Karimipour. Metaspongedetect: A meta-learning approach to securing ai systems against sponge attacks in social applications. *IEEE Transactions on Computational Social Systems*, 2026.

Narges Babadi, Hadis Karimipour, and Anik Islam. An ensemble learning to detect decision-based adversarial attacks in industrial control systems. In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 879–884. IEEE, 2023.

Hubert Baniecki and Przemyslaw Biecek. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, 107:102303, 2024.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bimodal and encoder-decoder transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 387–396. IEEE, 2021.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Francesco Croce and Matthias Hein. Sparse and imperceivable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4724–4732, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *Advances in neural information processing systems*, 32, 2019.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3681–3688, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. *Advances in neural information processing systems*, 32, 2019.

Hanxun Huang, Yisen Wang, Sarah Erfani, Quanquan Gu, James Bailey, and Xingjun Ma. Exploring architectural ingredients of adversarially robust deep neural networks. *Advances in neural information processing systems*, 34:5545–5559, 2021.

Hanxun Huang, Sarah Erfani, Yige Li, Xingjun Ma, and James Bailey. Detecting backdoor samples in contrastive language image pretraining. *arXiv preprint arXiv:2502.01385*, 2025a.

Hanxun Huang, Sarah Erfani, Yige Li, Xingjun Ma, and James Bailey. X-transfer attacks: Towards super transferable adversarial attacks on clip. *arXiv preprint arXiv:2505.05528*, 2025b.

Qi-Xian Huang, Lin-Kuan Chiang, Min-Yi Chiu, and Hung-Min Sun. Focus-shifting attack: An adversarial attack that retains saliency map information and manipulates model explanations. *IEEE Transactions on Reliability*, 73(2):808–819, 2023.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pp. 2137–2146. PMLR, 2018.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 2043–2059. IEEE, 2022.

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*, pp. 267–280. Springer, 2019.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*, pp. 17283–17300. PMLR, 2023.

Aditya Kuppa and Nhien-An Le-Khac. Black box attacks on explainable artificial intelligence (xai) methods in cyber security. In *2020 International Joint Conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Himabindu Lakkaraju and Osbert Bastani. " how do i fool you?" manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 79–85, 2020.

Yi Li, Hualiang Wang, Yiqun Duan, Hang Xu, and Xiaomeng Li. Exploring visual interpretability for contrastive language-image pre-training. *arXiv preprint arXiv:2209.07046*, 2022.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Sparsefool: a few pixels make a big difference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9087–9096, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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*, pp. 618–626, 2017.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 111–119. IEEE, 2020.

Xinhe Wang, Pingbang Hu, Junwei Deng, and Jiaqi W Ma. Adversarial attacks on data attribution. *arXiv preprint arXiv:2409.05657*, 2024.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Suorong Yang, Peng Ye, Wanli Ouyang, Dongzhan Zhou, and Furao Shen. A clip-powered framework for robust and generalizable data selection. *arXiv preprint arXiv:2410.11215*, 2024.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

Jingfeng Zhang, Prashanth Krishnamurthy, Naman Patel, Anthony Tzes, and Farshad Khorrami. Mp-nav: Enhancing data poisoning attacks against multimodal learning. In *Forty-second International Conference on Machine Learning*.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

A Normalization Invariance

Let $E \in \mathbb{R}^P$ be an explanation map with mean $\bar{E}$ and standard deviation $\sigma(E) > 0$, and let

$$\tilde{E} = \frac{E - \bar{E}\,\mathbf{1}}{\sigma(E)} \tag{8}$$

denote its z-score. We call an operator D on $\mathbb{R}^P$ *affine-invariant* if $D(aE + b\mathbf{1}) = D(E)$ for all $a > 0$ and $b \in \mathbb{R}$. Min–max normalization, percentile thresholding, top-k selection and z-scoring are affine-invariant.

Proposition 2. *Let $g(E) = h(\tilde{E})$ for a differentiable function h. Then*

$$\langle \nabla g(E), \mathbf{1} \rangle = 0 \quad \textit{and} \quad \langle \nabla g(E), E \rangle = 0, \tag{9}$$

so $\nabla g(E)$ is orthogonal to $\mathcal{S}(E) = \operatorname{span}\{\mathbf{1}, E\}$. Moreover, for any affine-invariant operator D, a change of E within $\mathcal{S}(E)$ leaves $D(E)$ unchanged. Consequently, a descent step on g spends no part of its update on directions that an affine-invariant display or metric cannot register.

Proof. For $a > 0$ and $b \in \mathbb{R}$, the map $aE + b\mathbf{1}$ has mean $a\bar{E} + b$ and standard deviation $a\,\sigma(E)$, so its z-score equals $\tilde{E}$. Hence

$$g(aE + b\mathbf{1}) = g(E) \qquad \text{for all } a > 0,\ b \in \mathbb{R}. \tag{10}$$

Differentiating with respect to b at $(a, b) = (1, 0)$ yields $\langle \nabla g(E), \mathbf{1} \rangle = 0$, and differentiating with respect to a at the same point yields $\langle \nabla g(E), E \rangle = 0$.

For the second claim, consider a change $\alpha\mathbf{1} + \beta E$ with $\alpha \in \mathbb{R}$ and $|\beta| < 1$. It maps E to $(1 + \beta)E + \alpha\mathbf{1}$ with $1 + \beta > 0$, so $D\big((1 + \beta)E + \alpha\mathbf{1}\big) = D(E)$ by affine invariance. Components of an update that lie in $\mathcal{S}(E)$ therefore leave $D(E)$ unchanged, and by the first claim the gradient of g has no such component. □

The raw-score objective is not invariant. For the objective of Equation (3), with $K = |R_t|$, the gradient with respect to the scored map s is

$$\nabla_s \mathcal{L}_{\text{naive}} = -\frac{1}{K}\,\mathbf{1}_{R_t} + \frac{\alpha}{P-K}\,\mathbf{1}_{\overline{R_t}}, \qquad \big\langle \nabla_s \mathcal{L}_{\text{naive}}, \mathbf{1} \big\rangle = \alpha - 1, \tag{11}$$

where $\mathbf{1}_R$ is the indicator vector of R and $\overline{R_t}$ the complement of R_t. For $\alpha = 0.1$, as used in our experiments, this inner product is $-0.9 \neq 0$: part of every descent step raises all scores uniformly, which no affine-invariant display or metric registers.

B Evaluation Metrics

Let E_s and E_t denote the CLIP maps of the predicted class $\hat{y}$ and the target class y_t, let x be a clean image and x' its adversarial counterpart, and let $\mathrm{Top}_k(\cdot)$ return the indices of the $k = \mathrm{round}(0.25P)$ highest-scoring patches. For index sets A and B, $\mathrm{IoU}(A, B) = |A \cap B|/|A \cup B|$.

Redirection. Redirection toward the target and departure from the source are measured by

$$\mathrm{IoU}_{\to t} = \mathrm{IoU}\big(\mathrm{Top}_k(E_s(x')),\, \mathrm{Top}_k(E_t(x))\big), \tag{12}$$

$$\mathrm{IoU}_{\to s} = \mathrm{IoU}\big(\mathrm{Top}_k(E_s(x')),\, \mathrm{Top}_k(E_s(x))\big). \tag{13}$$

The concentration of the redirected map in the target region R_t of Section 3.3 is

$$\mathrm{Mass}_{R_t} = \sum_{i \in R_t} \mathrm{softmax}\big(\tilde{E}_s(x')\big)_i, \tag{14}$$

which equals $|R_t|/P$ for a uniform map.

Floors and success. The *clean overlap* is $\mathrm{IoU}\big(\mathrm{Top}_k(E_s(x)), \mathrm{Top}_k(E_t(x))\big)$, the part of any measured shift already present before the attack. The *noise floor* is $\mu_{\text{rand}} + \sigma_{\text{rand}}$, the mean plus one standard deviation of $\mathrm{IoU}_{\to t}$ over five perturbations $\delta_j \sim \mathcal{U}[-\epsilon, \epsilon]^d$ applied to x. An attack is a *success* if

$$\hat{y}(x') = \hat{y}(x) \quad \text{and} \quad \mathrm{IoU}_{\to t} > \max\big(\mathrm{IoU}\big(\mathrm{Top}_k(E_s(x)), \mathrm{Top}_k(E_t(x))\big),\, \mu_{\text{rand}} + \sigma_{\text{rand}}\big). \tag{15}$$

In the transfer experiments (Tables 2 and 6), an explainer or backbone is counted as *moved* if its own $\mathrm{IoU}_{\to t}$ exceeds both its own clean overlap and its own $\mathrm{IoU}_{\to t}$ under the first random perturbation.

Prediction preservation. A *flip* occurs when $\hat{y}(x') \neq \hat{y}(x)$. Global stealth is measured by the cosine similarity of the image embeddings and by the largest change in any class probability,

$$\mathrm{CosSim} = \frac{z_I(x)^\top z_I(x')}{\|z_I(x)\|_2\, \|z_I(x')\|_2}, \qquad \mathrm{Max}\Delta\mathrm{P} = \max_c \big|p_c(x') - p_c(x)\big|, \tag{16}$$

where $p = \mathrm{softmax}(100 \cdot \ell)$ applies CLIP's logit scale to the cosine logits ℓ.

Perceptibility. SSIM (Wang et al., 2004) is computed on 224×224 RGB images in $[0, 1]$ with a data range of 1. LPIPS (Zhang et al., 2018) uses the AlexNet backbone with inputs rescaled to $[-1, 1]$. PSNR is computed on images in $[0, 1]$.

C Transferability Across Explainers and Backbones

Across explanation methods. Table 2 applies the adversarial images crafted against CLIP to explainers the attack did not use, each scored against its own clean maps and its own noise floor. Transfer is limited: over the 45 pairs, the explanation moved onto the target region for raw patch similarity on 6, for attention rollout on 4 and for ScoreCAM on 9. A perturbation crafted against one explainer is therefore largely specific to that explainer.

Across backbones. Table 6 in Section F feeds each adversarial image, unchanged, to the other two backbones and their own CLIP explainers, counting only images on which the receiving model makes the same clean prediction. The explanation moved onto the target on 20 of these 85 cases, and the receiving model's prediction changed on 2. Cross-backbone transfer is thus limited, and the attack is best understood as a white-box attack on a specific model and explainer.

Table 2: Transfer to explainers the attack was not optimized against: number of images on which each explainer's source-class map moved onto the target region above both of its own floors. X-Shift-Ens optimizes CLIP-Surgery, patch similarity and rollout jointly (analysis cells only); ScoreCAM is query-based and never differentiated through.

		X-Shift (CLIP-Surgery only)			X-Shift-Ens		
Dataset	Backbone	Patch sim.	Rollout	ScoreCAM	Patch sim.	Rollout	ScoreCAM
MS-COCO	ViT-B/16	1/5	0/5	0/5	4/5	1/5	0/5
MS-COCO	ViT-B/32	2/5	1/5	2/5	5/5	3/5	2/5
MS-COCO	ViT-L/14	0/5	0/5	2/5	5/5	4/5	1/5
Flickr30k	ViT-B/16	0/5	0/5	0/5	–	–	–
Flickr30k	ViT-B/32	1/5	1/5	1/5	–	–	–
Flickr30k	ViT-L/14	1/5	0/5	1/5	–	–	–
ImageNet-1k	ViT-B/16	0/5	0/5	1/5	–	–	–
ImageNet-1k	ViT-B/32	0/5	2/5	0/5	–	–	–
ImageNet-1k	ViT-L/14	1/5	0/5	2/5	–	–	–

Table 3: Ablation and prediction-targeted baselines on MS-COCO, pooled over backbones (15 image–backbone pairs). One change per row relative to the full attack; same success rule and same-budget noise controls.

Variant	Success	$\mathrm{IoU}_{\to t}\uparrow$	$\mathrm{IoU}_{\to s}\downarrow$	Flip↓	CosSim↑	MaxΔP↓
X-Shift (full, $\epsilon = 4/255$)	15/15	0.86	0.02	0/15	0.990	0.0011
w/o $\mathcal{L}_{\mathrm{shape}}$	15/15	0.46	0.07	0/15	0.994	0.0008
w/o $\mathcal{L}_{\mathrm{con}}$	15/15	0.92	0.04	0/15	0.994	0.0007
margin hinge only (no logit/KL/cos)	15/15	0.91	0.02	0/15	0.768	0.1059
w/o $\mathcal{L}_{\mathrm{pres}}$	6/15	0.92	0.02	9/15	0.737	0.5824
naive objective (Eq. 3)	15/15	0.45	0.17	0/15	0.913	0.0332
$\epsilon = 2/255$	15/15	0.81	0.02	0/15	0.982	0.0031
$\epsilon = 8/255$	15/15	0.88	0.02	0/15	0.992	0.0008
PGD, untargeted (prediction attack)	0/15	0.06	0.37	15/15	0.520	0.9311
PGD, targeted to y_t (prediction attack)	0/15	0.07	0.46	15/15	0.544	0.9900

D Ablation and Comparison with Prediction Attacks

Table 3 changes one component at a time on the 15 MS-COCO pairs, using the same success rule and the same noise controls. Because the success rule only checks for a flip and for exceeding the floors, we compare variants mainly on $\mathrm{IoU}_{\to t}$ and on the stealth metrics.

Explanation terms. The shape term drives the redirection. Removing it lowers $\mathrm{IoU}_{\to t}$ from 0.86 to 0.46, whereas removing the contrast term does not reduce it (0.92). The contrast term is therefore not needed in our setting: the shape term alone reproduces, and slightly exceeds, the effect of the full objective. We report the full objective as specified in Section 3.5 for consistency with the main results.

Normalization invariance. Replacing the z-scored objective with the non-invariant objective of Equation (3), under the same budget, optimizer and preservation terms, lowers $\mathrm{IoU}_{\to t}$ from 0.86 to 0.45 and leaves more of the explanation in the source region ($\mathrm{IoU}_{\to s} = 0.17$ versus 0.02). It also changes the output more (MaxΔP 0.0332 versus 0.0011). This is consistent with the argument of Section 3.4: an objective that can be improved without changing the normalized map spends part of the budget on changes that neither the metric nor the auditor sees.

Prediction preservation. Keeping only the margin hinge, so that only the argmax is protected, leaves the redirection intact ($\mathrm{IoU}_{\to t} = 0.91$) but lowers CosSim from 0.990 to 0.768 and raises MaxΔP from 0.0011 to 0.1059. The prediction survives while the output distribution visibly changes. Removing all preservation terms flips the prediction on 9 of 15 pairs. Anchoring the full logit vector, the probabilities and the embedding is what makes the attack stealthy, not the argmax constraint alone.

Budget. At $\epsilon = 2/255$, X-Shift still succeeds on all 15 pairs with $\mathrm{IoU}_{\to t} = 0.81$, at SSIM 0.986 and LPIPS 0.008. Doubling the budget to $8/255$ adds little redirection (0.88) but raises LPIPS to 0.102, so $4/255$ is a reasonable operating point and $2/255$ is sufficient for most of the effect.

Prediction-targeted attacks. Standard ℓ_∞ PGD at the same budget (Madry et al., 2017), untargeted or targeted to y_t, flips the prediction on all 15 pairs and changes the output almost completely (MaxΔP 0.93 and 0.99). It does not move the explanation of the source class onto the target region ($\mathrm{IoU}_{\to t} = 0.06$ and 0.07). Explanation redirection is therefore not a by-product of attacking the prediction: it requires an objective on the explanation itself, and the two attacks operate in different regimes.

E HYPERPARAMETERS

All experiments use the same hyperparameters for every dataset and backbone:

Table 4: Hyperparameters used for all datasets and backbones.

Parameter	Value
Budget ϵ (ℓ_∞, pixel space)	$4/255$ (ablation: $2/255$, $8/255$)
Optimizer	Adam, learning rate $2/255$, cosine schedule, 300 steps
Region fraction q	0.15
λ_{con}, λ_{shape}, T	1, 1, 0.5
λ_ℓ, λ_{KL}, $\lambda_{\cos}$, λ_m, κ	20, 5, 5, 50, 0.95
CLIP heatmap temperature γ	2
Evaluation top-k fraction	0.25
Screening: min. probability gap; target ranks; max. clean region IoU	0.5; 4–25; 0.25
Random controls per image	5
Naive objective (Equation (3))	$\alpha = 0.1$, loss scale 20
PGD baselines	$\epsilon = 4/255$, 20 steps, step size $\epsilon/4$

F SCREENING STATISTICS AND TRANSFER ACROSS BACKBONES

Table 5 reports, for each dataset and backbone, how many images were screened to obtain five that pass the rules of Section 4.1, the most frequent rejection reason, and the runtime of one attack. Table 6 reports transfer of the adversarial images across backbones.

Table 5: Screening and cost. Images are taken in a fixed-seed random order; screening stops once 5 images pass, so the pass rate is estimated from the images seen up to that point. Runtime is wall-clock time of one attack (mean over the kept images).

Dataset	Backbone	Screened	Pass rate	Main rejection	Time/attack (s)
MS-COCO	ViT-B/16	8	5/8	prob_gap (3)	22.6
MS-COCO	ViT-B/32	8	5/8	prob_gap (3)	12.5
MS-COCO	ViT-L/14	8	5/8	prob_gap (3)	55.4
Flickr30k	ViT-B/16	16	5/16	prob_gap (11)	22.7
Flickr30k	ViT-B/32	17	5/17	prob_gap (12)	12.6
Flickr30k	ViT-L/14	17	5/17	prob_gap (12)	55.6
ImageNet-1k	ViT-B/16	10	5/10	prob_gap (5)	27.2
ImageNet-1k	ViT-B/32	11	5/11	prob_gap (6)	12.4
ImageNet-1k	ViT-L/14	8	5/8	prob_gap (3)	64.0

Table 6: Transfer across backbones, pooled over datasets. Entry: images on which the perturbation crafted on the source backbone moved the receiving backbone's own CLIP-Surgery map onto its target region, out of images where the receiving backbone makes the same clean prediction; flips of the receiving model in parentheses.

Source ↓ / Target →	ViT-B/16	ViT-B/32	ViT-L/14
ViT-B/16	15/15 (self)	2/15 (1)	4/15 (0)
ViT-B/32	6/15 (0)	15/15 (self)	4/14 (0)
ViT-L/14	2/13 (1)	2/13 (0)	15/15 (self)