Right Predictions, Misleading Explanations: On the Vulnerability of Vision–Language Model Explanations

Narges Babadi [1] **Hadis Karimipour** [1]

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, faithfully reflect model reasoning 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 explanation faithfulness. To study this vulnerability, we introduce *X-Shift*, a novel grey-box attack that perturbs patch-level visual representations to redirect explanation heatmaps toward semantically irrelevant regions 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 and generalizes across multiple CLIP architectures and explanation methods. We evaluate the proposed approach on ImageNet-1k, MS-COCO, and Flickr30K, demonstrating consistent degradation in explanation alignment under imperceptible perturbations while maintaining prediction stability. Furthermore, standard prediction-oriented adversarial attacks fail to reproduce the same explanation-shifting behavior even under substantially larger perturbation budgets. 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]Department of Electrical and Software Engineering, University of Calgary, Calgary, Canada. Correspondence to: Narges Babadi <narges.babadi1@ucalgary.ca>.

Workshop on Trustworthy AI for Good (AI4GOOD) at the 43rd International Conference on Machine Learning (ICML 2026), Seoul, South Korea. Copyright 2026 by the author(s).

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 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 faithful 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 toward clinically irrelevant regions. 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 explanation faithfulness can be manipulated independently of model predictions. We introduce *X-Shift*, a gray-box adversarial attack that operates on patch-level representations to systematically redirect explanation heatmaps toward attacker-specified regions while preserving the original output. 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 requires no access to model parameters or architecture and can be applied at inference time, making it practical in real-world deployment scenarios.

We evaluate X-Shift across multiple datasets and CLIP backbones, demonstrating that it achieves substantial manipulation of explanation maps while maintaining prediction consistency. Furthermore, the attack exhibits strong transferability across architectures and explanation methods, suggesting that the vulnerability is systemic rather than model-specific. Our contributions are summarized as follows:

- We introduce X-Shift, a gray-box adversarial attack that redirects explanation heatmaps at the patch level while preserving model predictions.
- We demonstrate that the resulting perturbations remain imperceptible to human observers, enabling stealthy manipulation in practical settings.
- We show that X-Shift generalizes across multiple VLM architectures and explanation methods, revealing strong transferability and a broad attack surface.

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. In contrast, a growing line of work studies a more subtle threat: 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. (Kindermans et al., 2019) showed that saliency maps are not invariant to simple input transformations, while Ghorbani et al. (Ghorbani et al., 2019) and Dombrowski et al. (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. (Heo et al., 2019) trained models to produce misleading attributions under methods such as Grad-CAM and LRP, while Slack et al. (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).

Subsequent work has developed more targeted and efficient attacks on explanations. Huang et al. (Huang et al., 2023) proposed a focus-shifting attack that redirects saliency toward adversary-specified regions while preserving prediction consistency. Ajalloeian et al. (Ajalloeian et al., 2023) introduced sparse perturbation techniques that achieve similar effects under minimal input changes. In parallel, Kuppa and Le-Khac (Kuppa & Le-Khac, 2020) studied grey-box attacks on LIME and SHAP in cybersecurity settings, highlighting the broader risks of explanation manipulation.

Despite these advances, existing work is largely restricted to unimodal image classifiers. The security of explanation mechanisms in vision–language models (VLMs) remains largely underexplored. Prior studies on models such as CLIP have primarily focused on prediction robustness, including adversarial examples, poisoning attacks, and backdoor vulnerabilities (Yang et al., 2024; Zhang et al.; Huang et al., 2025b; Jia et al., 2022; 2021; Koh et al., 2023; Huang et al., 2025a). However, these works do not consider the integrity of explanation mechanisms, which are increasingly used as signals for trust and downstream decision-making.

In contrast, we study adversarial manipulation of explanation faithfulness in VLMs. 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 model predictions. Our method operates in a grey-box setting, produces imperceptible perturbations, and generalizes across architectures and explanation methods, revealing a previously unrecognized and practically

relevant attack surface in VLM-based systems.

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 explanation heatmaps are systematically redirected toward target regions. The attack jointly optimizes four objectives: (i) explanation manipulation, (ii) prediction preservation, (iii) sparsity of perturbations, and (iv) validity constraints on adversarial examples.

3.1. Threat Model

X-Shift operates under a *grey-box* threat model. The adversary has access to the model's forward pass and gradient information sufficient to optimize inputs, but does not modify or access model parameters or training data. This setting is realistic for publicly available vision–language models such as CLIP, where model architectures and weights are accessible, but training data and internal supervision remain unknown. 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:

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

where τ is a learnable temperature. Our attack perturbs x into $x_{\text{adv}} = x + \delta$, preserving predictions but shifting explanation maps toward a target class.

3.3. X-Shift attack Objectives

We combine the following complementary objectives to achieve explanation-focused adversarial perturbations:

Explanation manipulation. The primary goal is to force patch embeddings toward the target text embedding. Let p denote the normalized embedding of patch p, and t_{target} the target text embedding. Similarity is $s_p = p^\top t_{target}$. We maximize similarity of the top-K patches while suppressing others:

$$\mathcal{L}_{\text{xai}} = -\frac{1}{K}\sum_{i\in\text{TopK}} s_{i,t} + \alpha\cdot\frac{1}{P-K}\sum_{i\notin\text{TopK}} s_{i,t}, \quad (2)$$

where $s_{i,t} = z_i^\top z_{T_{\text{tar}}}$ denotes the similarity between patch embedding z_i and the target text embedding $z_{T_{\text{tar}}}$.

Prediction preservation. To prevent label change, we enforce the clean prediction y^* at the global (CLS) level:

$$\mathcal{L}_{\text{pred}} = -\log\frac{\exp(z_{\text{cls}}^\top t_{y^*})}{\sum_c \exp(z_{\text{cls}}^\top t_c)}. \quad (3)$$

Patch-level margin. For each patch, the target similarity $s_{p,t}$ must dominate over other classes:

$$\mathcal{L}_{\text{patch}} = \frac{1}{P}\sum_{p=1}^{P}\max\big(0, \max_{c\neq t}(s_{p,c} - s_{p,t} + m)\big), \quad (4)$$

where $s_{p,c} = z_p^\top z_{T_c}$ is the similarity between patch embedding z_p and text embedding z_{T_c}.

Entropy sharpening. To avoid diffuse attention maps, we encourage sharp similarity distributions:

$$\mathcal{L}_{\text{entropy}} = \sum_{p=1}^{P} m_p \log m_p, \qquad m_p = \frac{\exp(s_{p,t})}{\sum_q \exp(s_{q,t})}, \quad (5)$$

which corresponds to the negative Shannon entropy of the normalized similarities. Minimizing this term encourages sharp and peaked similarity distributions rather than diffuse heatmaps.

Sparsity constraint. Perturbations are restricted to k pixels by projecting $\delta = x_{adv} - x$ onto its top-k entries:

$$\delta \leftarrow \text{TopK}(\delta, k). \quad (6)$$

Validity constraint. Ensure the adversarial image remains in the valid input domain:

$$x_{adv} \in [0,1]^d. \quad (7)$$

The total objective combines explanation manipulation with auxiliary constraints:

$$\mathcal{L} = \mathcal{L}_{\text{xai}} + \lambda_{\text{pred}}\mathcal{L}_{\text{pred}} + \lambda_{\text{patch}}\mathcal{L}_{\text{patch}} + \lambda_{\text{ent}}\mathcal{L}_{\text{entropy}} \quad (8)$$

where $\lambda_{\text{pred}}, \lambda_{\text{patch}}$, and λ_{ent} are trade-off coefficients that balance the relative contributions of preserving prediction consistency, enforcing patch-level constraints, and controlling explanation entropy. Tuning these hyperparameters adjusts the strength of each auxiliary objective relative to the main explanation-shifting loss $\mathcal{L}_{\text{xai}}$.

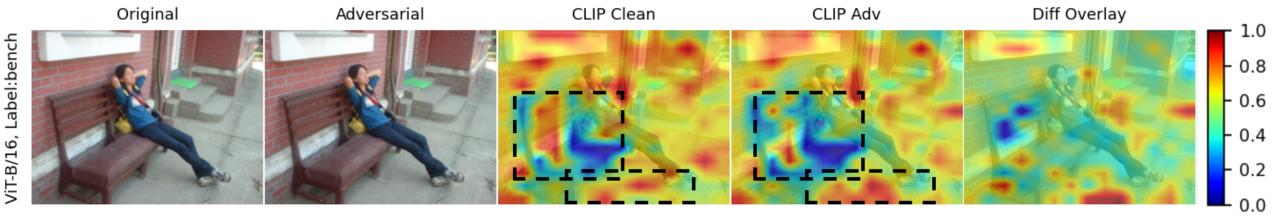


Figure 1. Visualization of a sample image under the X-Shift attack. From left to right, columns show the clean and adversarial images (the adversarial image is optimized to shift CLIP's explanation toward "ground" while preserving the "bench" prediction; see black box), the clean–adversarial difference map, CLIP heatmaps illustrating explanation drift.

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

Algorithm 1 X-Shift Attack: Explanation Manipulation on CLIP

Input: clean image x, text embeddings $\{t_c\}$, target index t, step size η, sparsity k, iterations T
Output: adversarial image x^{adv}
Initialize $x^{(0)} \leftarrow x$
For $i = 1$ to T :
Compute patch embeddings $\{z_p\}$ *and CLS embedding* z_{cls}
Evaluate losses: $\mathcal{L}_{xai}, \mathcal{L}_{pred}, \mathcal{L}_{patch}, \mathcal{L}_{entropy}$
Total loss:

$$\mathcal{L} \leftarrow \mathcal{L}_{xai} + \lambda_{pred}\mathcal{L}_{pred} + \lambda_{patch}\mathcal{L}_{patch} + \lambda_{ent}\mathcal{L}_{\text{entropy}}$$

Gradient update:

$$x^{(i)} \leftarrow x^{(i-1)} + \eta \cdot \text{sign}(\nabla_x \mathcal{L})$$

Sparsity projection:

$$\delta \leftarrow \text{TopK}(x^{(i)} - x^{(0)}, k), \quad x^{(i)} \leftarrow x^{(0)} + \delta$$

Clamp to valid domain:

$$x^{(i)} \leftarrow \text{clip}(x^{(i)}, 0, 1)$$

$x^{adv} = x^{(T)}$

4. Experiments

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

- **RQ1:** How effectively does X-Shift manipulate patch–text explanation maps while preserving the original model predictions?
- **RQ2:** Does X-Shift maintain prediction consistency across different CLIP backbones and large-scale datasets?
- **RQ3:** To what extent is X-Shift transferable across different CLIP architectures and explanation (heatmap) methods?

Models and Datasets. We evaluate our attack at inference time, without requiring additional training data. Experiments are conducted on the validation splits of three benchmark datasets: ImageNet-1k (Deng et al., 2009), Flickr30k (Young et al., 2014), and MS-COCO (Chen et al., 2015), which provide diverse natural images and object-level annotations for assessing VLMs explanations. For models, we utilize the CLIP family of vision–language encoders, specifically ViT-B/16 (Radford et al., 2021), ViT-B/32 (Radford et al., 2021), and ViT-L/14 (Dosovitskiy et al., 2020), which span a range of capacities and input resolutions to assess the generality of our attack and defense across different backbones.

Implementation. We implement the proposed attack on official CLIP models, using patch–text similarity maps that compute cosine similarity between patch and text embeddings. Unlike gradient-based attributions (e.g., Grad-CAM, Integrated Gradients), which often yield unstable ViT heatmaps, similarity maps are faithful, text-conditioned, efficient (single forward pass), and deterministic. CLIP employs attention pooling, yielding a 7×7 grid for 224×224 inputs (datasets resized accordingly). The attack loss follows Section 3, with weights 20.0 for $\mathcal{L}_{\text{xai}}$, λ_{ent} for entropy, λ_{margin} for patch separation, and $0.01\lambda_{\text{pred}}$ for prediction consistency, tuned to balance manipulation and stability.

Metrics. We evaluate global prediction stability and explanation robustness using four quantitative metrics: (i) cosine similarity between clean and adversarial CLS embeddings (CosSim, ↑), (ii) maximum probability deviation (Max ΔProb, ↓), and (iii) spatial overlap of explanation maps via Top-k Intersection-over-Union (IoU, ↓). Formal definitions are provided in Appendix A.

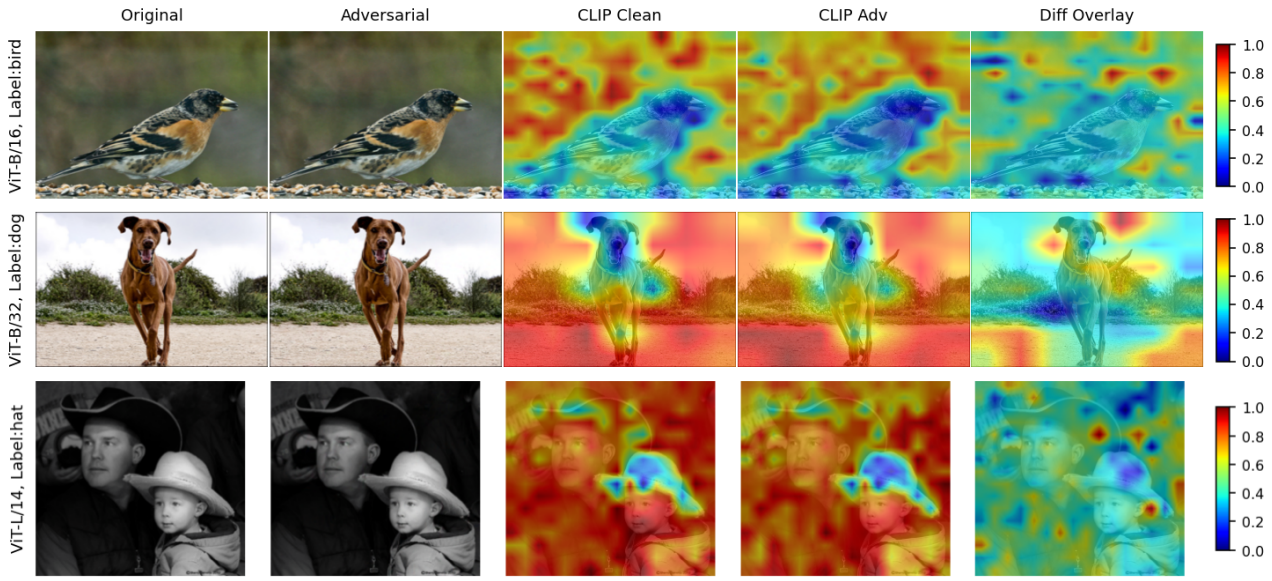


Figure 2. Comparison of CLIP explanations on ImageNet dataset(ViT-B/16, ViT-B/32, ViT-L/14) under X-Shift attack. Columns show original/adversarial images, CLIP heatmaps (clean vs. adversarial).

4.1. Results on Explainability

Proposed Attack Effectiveness. Figure 1 demonstrates that the X-Shift adversarial perturbations successfully shift CLIP's explanation maps while preserving the predicted label. In the clean case, the heatmap correctly attends to the input concept (e.g., "bench"), whereas under the X-Shift attack the attention is redirected toward unrelated regions (e.g., the "wall"), thereby compromising explanation faithfulness.

Furthermore, Figures 2, 3, and 4 visualize additional examples from ImageNet, Flickr30k, and COCO. In each case, the perturbation remains imperceptible to humans yet induces substantial shifts in the explanation maps, highlighting the vulnerability of current XAI methods.

Additionally, Table 1 presents a quantitative evaluation of X-Shift under different datasets and CLIP backbones. The results demonstrate that the attack consistently manipulates explanation outputs while preserving model predictions across all settings. Specifically, the cosine similarity values remain high across datasets and architectures, indicating that the global representation and prediction behavior of the model is largely unchanged under attack. At the same time, the Top-k IoU scores exhibit substantial variation, showing that the spatial distribution of explanation heatmaps is significantly altered despite stable predictive outputs. This discrepancy highlights the core effect of X-Shift: decoupling explanation faithfulness from prediction correctness.

Across ImageNet, Flickr30k, and MS-COCO, X-Shift maintains near-identical CosSim and Max ΔProb values compared to the unperturbed setting, confirming that the attack does not degrade classification confidence or decision consistency. However, the consistent degradation in explanation alignment (as reflected in IoU shifts across backbones) demonstrates that explanation maps can be systematically redirected without triggering changes in model output statistics. Notably, this behavior is preserved across ViT-B/16, ViT-B/32, and ViT-L/14. Overall, these results validate that X-Shift exposes a critical vulnerability in VLMs, where faithful explanations can be adversarially manipulated while maintaining correct predictions, undermining their reliability as a trust signal in safety-critical applications.

Table 1. Quantitative evaluation of **Vanilla CLIP** under X-Shift attack across datasets and backbones. Metrics include cosine similarity (CosSim), maximum probability change under attack (Max ΔProb), and Top-k IoU.

Dataset	Backbone	Vanilla CLIP		
		CosSim	Max ΔProb	IoU
ImageNet	ViT-B/16	0.805	0.004	0.487
	ViT-B/32	0.807	0.004	0.450
	ViT-L/14	0.948	0.000	0.551
Flickr30k	ViT-B/16	0.935	0.000	0.841
	ViT-B/32	0.974	0.000	0.867
	ViT-L/14	0.933	0.000	0.727
MS-COCO	ViT-B/16	0.977	0.000	0.611
	ViT-B/32	0.953	0.000	0.556
	ViT-L/14	0.962	0.000	0.583

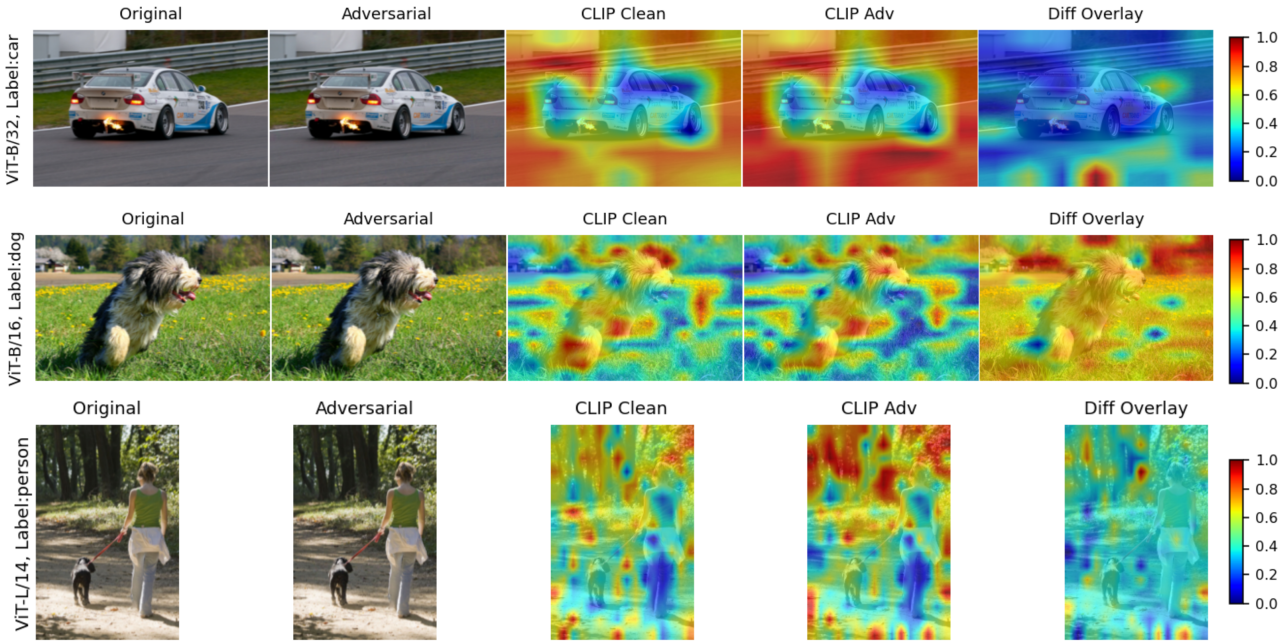


Figure 3. Explanations on Flickr30k samples using CLIP (ViT-B/16, ViT-B/32, ViT-L/14) under X-Shift attack. Shown are original/adversarial images, CLIP heatmaps (clean vs. adversarial).

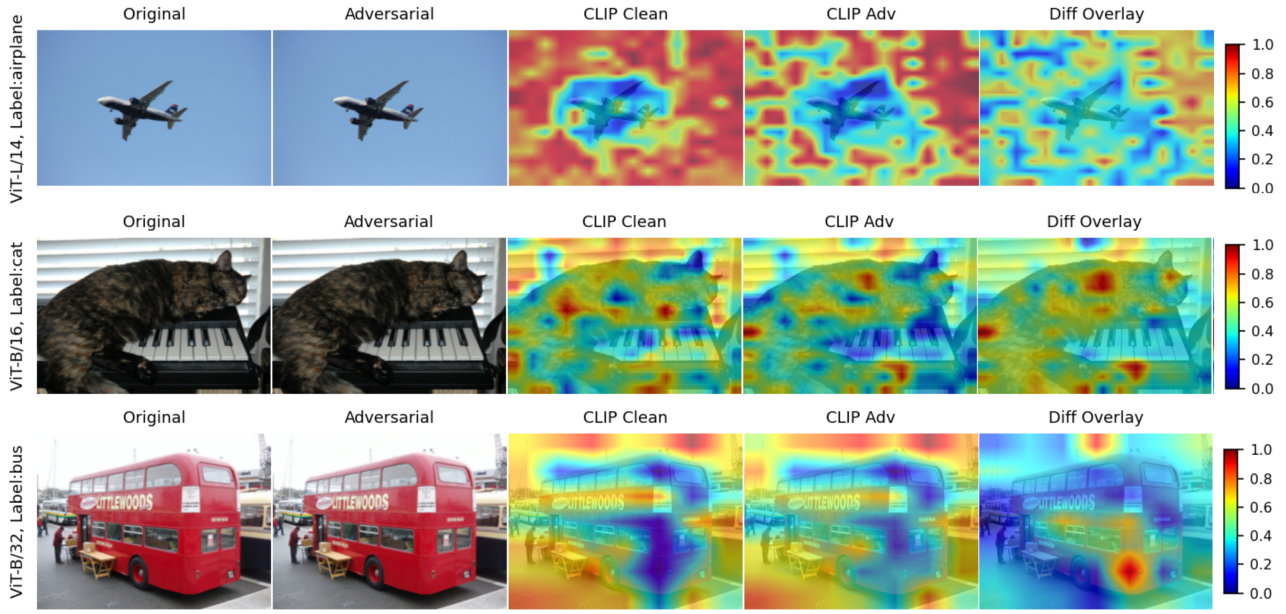


Figure 4. Explanation robustness on COCO samples using CLIP (ViT-B/16, ViT-B/32, ViT-L/14) under X-Shift attack. Columns display original vs. adversarial images, CLIP heatmaps (clean vs. adversarial).

4.2. Transferability of X-Shift Across Vision Transformer Backbones

We evaluate whether explanation-shifting perturbations generated on one CLIP encoder transfer to other CLIP variants with different patch sizes and embedding dimensions. Specifically, we test ViT-B/16, ViT-B/32, and ViT-L/14 models in a source-to-target setting, measuring:

- Cosine similarity between clean and adversarial CLS embeddings ($\text{CosSim}_{\text{CLS}}$)

- Maximum deviation in predicted probabilities across all text prompts ($\mathrm{Max}\Delta\mathrm{Prob}$)
- Patch-level shift in the similarity map for the target concept using $\mathrm{IoU}_{\text{Top-}k}$ (lower is better for measuring explanation manipulation)
- Smooth distributional similarity shift using Soft-IoU (also lower is better)

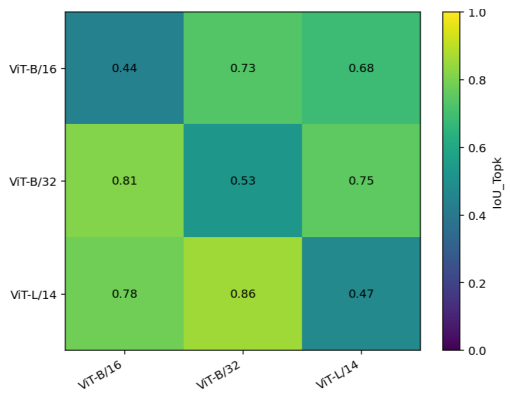


Figure 5. Transfer matrix of IoU-TopK (lower indicates stronger manipulation) across source → target CLIP backbones. ViT-B/32 perturbations transfer most broadly, while ViT-L/14 perturbations remain more model-specific.

Experiment analysis. Table 2 shows that self-attacks produce the lowest IoU-TopK values (0.44–0.47), indicating strong spatial manipulation of the similarity map without altering model predictions (CosSim >0.94, MaxΔProb$<$ 4×10^{-4}). Cross-architecture transfer is moderate but consistent: for example, perturbations crafted on ViT-B/32 transfer to ViT-L/14 with IoU = 0.63, demonstrating that the attack generalizes across patch sizes (14–32) and embedding widths. Soft-IoU remains high because CLIP map distributions are smooth, but localized top-k patch ordering is reliably perturbed. Overall, the results confirm that X-Shift attacks preserve classification while inducing model-invariant explanation shifts. Figure 5 visualizes the mean IoU-TopK transfer matrix (lower is better), highlighting asymmetric transfer patterns: perturbations from ViT-B/32 transfer more strongly to other backbones than those from ViT-L/14. The heatmap corroborates the numerical results in Table 2 and illustrates which source architectures most reliably induce cross-model explanation shifts.

4.3. Transferability Across Explainability Methods

To evaluate whether X-Shift attacks generated on vanilla CLIP transfer across attribution methods, datasets, and architectures, we compute a grid of heatmaps using our visualization pipeline. The adversary optimizes the X-Shift perturbation directly on CLIP ViT-B/16, relocating the patch–text similarity mass from the clean concept (e.g., "cat") toward an adversarial concept (e.g., "background") while preserving the original prediction. Thus, any changes observed in Figures 6–8 reflect explanation drift rather than classification errors.

Choice of XAI methods. We include ScoreCAM, RISE, and gradient-based explanation (GAE) *solely* to assess attack transferability. These attribution methods were originally designed for single-stream CNN classifiers and do not model CLIP's multimodal text-image alignment or transformer attention. Consequently, they tend to produce diffuse and low-fidelity maps on both clean and adversarial CLIP inputs. Their purpose here is diagnostic: to show that VLMs such as CLIP require dedicated, reliable, and modality-aware explanation tools, and that CNN-based attribution methods lack the grounding needed to produce trustworthy heatmaps for multimodal models.

We evaluate this effect across COCO (Figure 6), Flickr30K (Figure 7), and ImageNet (Figure 8), plotting similarity maps and their differences for both clean and adversarial images. In addition to CLIP's native patch–text similarity heatmaps, we generate ScoreCAM, RISE, and Gradient-based Explanation (GAE) maps to test cross-method transferability. For vanilla CLIP, the X-Shift perturbation consistently alters the spatial attribution structure: the clean similarity map highlights the true object regions, whereas the adversarial map redirects attention toward background patches aligned with the attacker's target text. This drift appears across all datasets and attribution methods, and the difference-overlay visualizations clearly reveal large, structured regions of displaced saliency.

To assess the contribution of each loss component in the X-Shift objective (Eq. 8), we conduct a controlled ablation study reported in full in Appendix B. Briefly, removing $\mathcal{L}_{\text{xai}}$ (`pred_only` variant) yields negligible explanation drift ($\mathrm{IoU}_{\text{Topk}} = 0.7241$, TargetSim $= 0.2364$), confirming that prediction-stability losses alone cannot drive heatmap manipulation. Conversely, removing prediction-preservation terms (`xai_only`) increases MaxΔProb by nearly an order of magnitude ($0.000060 \rightarrow 0.000540$), producing perturbations detectable by prediction-monitoring systems and violating the core stealth requirement of X-Shift. The full objective achieves the best balance: strong explanation drift ($\mathrm{IoU}_{\text{Topk}} = 0.7857$, TargetSim $= 0.2406$) with a stable CLS representation ($\mathrm{CosSim}_{\text{CLS}} = 0.977$) and negligible classification change (MaxΔProb $= 0.000066$), demonstrating that all four loss components are necessary and non-substitutable for a stable, targeted, and stealthy attack.

Comparison with Other Adversarial Attacks. To contextualize the behavior of X-Shift, we compare it against standard prediction-targeted adversarial attacks, including FGSM and PGD, under both targeted and untargeted set-

Table 2. Transferability of explanation-shifting perturbations across CLIP architectures. We report cosine similarity of CLS tokens (CosSim↑), maximum change in predicted probability (MaxΔProb↓), and patch-overlap metrics IoU-TopK and Soft-IoU (both "lower is better" for capturing successful heatmap manipulation). Self-attacks achieve the lowest IoU (largest shift), while cross-model transfer remains moderate but consistent across backbones.

Source	Target	CosSim↑	MaxΔProb↓	IoU-TopK↓	Soft-IoU↓	Spearman	EMD
ViT-L/14	ViT-L/14	0.9421	0.00044	**0.4713**	0.9837	0.7710	0.0062
	ViT-B/16	**0.9928**	**0.00007**	0.7818	0.9962	0.9496	0.0010
	ViT-B/32	0.9180	0.00023	0.8571	**0.9973**	**0.9914**	0.0017
ViT-B/16	ViT-L/14	0.9805	0.00013	0.6842	0.9915	0.8940	0.0024
	ViT-B/16	0.7628	0.00039	**0.4412**	0.9891	0.7755	0.0104
	ViT-B/32	0.9721	0.00029	0.7059	0.9907	0.9194	0.0032
ViT-B/32	ViT-L/14	0.9520	0.00017	0.6316	0.9910	0.8743	0.0175
	ViT-B/16	0.9933	0.00026	**0.5882**	0.9902	0.9283	0.0060
	ViT-B/32	0.9278	**0.00009**	0.5750	0.9896	0.8442	0.0128

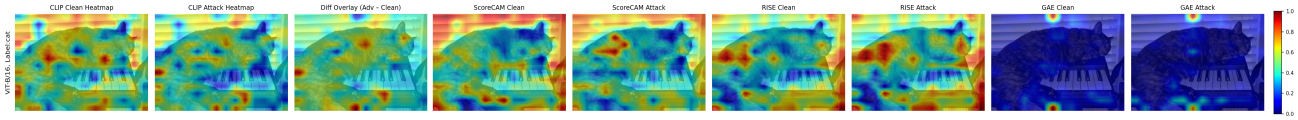


Figure 6. XAI transferability results for the X-Shift adversarial perturbation on COCO dataset with CLIP ViT-B/16. Similarity maps, ScoreCAM, RISE, and GAE all exhibit explanation drift under the attack when applied to vanilla CLIP.

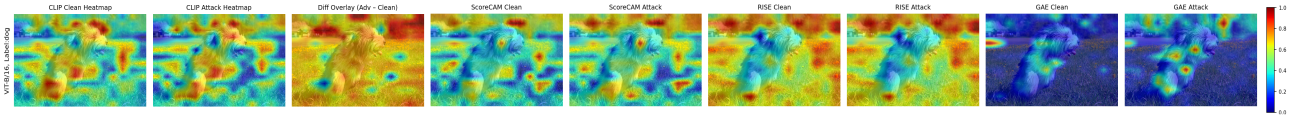


Figure 7. XAI transferability results for the X-Shift adversarial perturbation on the Flickr30k dataset with CLIP ViT-B/16. Similarity maps, ScoreCAM, RISE, and GAE all exhibit explanation drift under the attack when applied to vanilla CLIP.

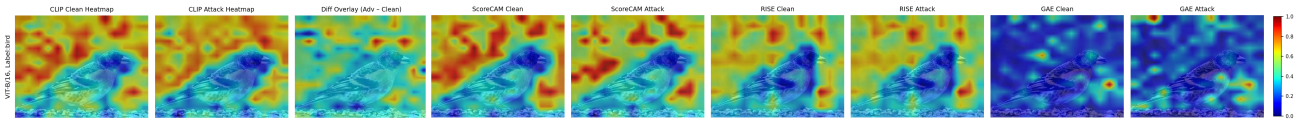


Figure 8. XAI transferability results for the X-Shift adversarial perturbation on the ImageNet dataset with CLIP ViT-B/16. Similarity maps, ScoreCAM, RISE, and GAE all exhibit explanation drift under the attack when applied to vanilla CLIP.

tings. As detailed in Appendix C, these baselines are given significantly larger perturbation budgets and stronger optimization settings to maximize their effectiveness. The results show that while conventional attacks succeed in altering model predictions, they do so by substantially distorting the global representation, leading to noticeable degradation in cosine similarity and prediction stability. In contrast, X-Shift operates in a fundamentally different regime: it preserves prediction consistency while achieving substantial manipulation of explanation maps under a much smaller perturbation budget. This comparison highlights that explanation-targeted attacks introduce a distinct and previously overlooked threat model that is not captured by standard adversarial evaluation protocols.

5. Conclusion

In this work, we investigated the vulnerability of vision–language models, with a particular focus on CLIP, to adversarial manipulation of explanation mechanisms. We proposed *X-Shift*, a targeted grey-box attack that perturbs inputs to systematically distort patch–text similarity heatmaps while preserving model predictions. Our findings demonstrate that explanation outputs can be arbitrarily manipulated without affecting model decisions, exposing a critical disconnect between prediction correctness and explanation faithfulness in current VLMs.

These results highlight the need for faithfulness-aware robustness evaluations and defenses in vision–language systems, particularly in safety-critical applications where explanations are used as proxies for model trustworthiness. Future work will explore the development of defense mechanisms and evaluation frameworks that improve the robustness of explanation methods against such adversarial manipulations, ensuring more reliable deployment of foundation models in real-world settings.

References

Ajalloeian, A., Moosavi-Dezfooli, S. M., Vlachos, M., and Frossard, P. Sparse attacks for manipulating explanations in deep neural network models. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 918–923. IEEE, 2023.

Babadi, N., Karimipour, H., and Islam, A. 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.

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

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

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

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

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dombrowski, A.-K., Alber, M., Anders, C., Ackermann, M., Müller, K.-R., and Kessel, P. Explanations can be manipulated and geometry is to blame. *Advances in neural information processing systems*, 32, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Heo, J., Joo, S., and Moon, T. Fooling neural network interpretations via adversarial model manipulation. *Advances in neural information processing systems*, 32, 2019.

Huang, H., Wang, Y., Erfani, S., Gu, Q., Bailey, J., and Ma, X. Exploring architectural ingredients of adversarially robust deep neural networks. *Advances in neural information processing systems*, 34:5545–5559, 2021.

Huang, H., Erfani, S., Li, Y., Ma, X., and Bailey, J. Detecting backdoor samples in contrastive language image pretraining. *arXiv preprint arXiv:2502.01385*, 2025a.

Huang, H., Erfani, S., Li, Y., Ma, X., and Bailey, J. X-transfer attacks: Towards super transferable adversarial attacks on clip. *arXiv preprint arXiv:2505.05528*, 2025b.

Huang, Q.-X., Chiang, L.-K., Chiu, M.-Y., and Sun, H.-M. Focus-shifting attack: An adversarial attack that retains saliency map information and manipulates model explanations. *IEEE Transactions on Reliability*, 73(2):808–819, 2023.

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

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Jia, J., Liu, Y., and Gong, N. Z. 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.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, explaining and visualizing deep learning*, pp. 267–280. Springer, 2019.

Koh, J. Y., Salakhutdinov, R., and Fried, D. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*, pp. 17283–17300. PMLR, 2023.

Kuppa, A. and Le-Khac, N.-A. 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.

Lakkaraju, H. and Bastani, O. " 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.

Li, Y., Wang, H., Duan, Y., Xu, H., and Li, X. Exploring visual interpretability for contrastive language-image pre-training. *arXiv preprint arXiv:2209.07046*, 2022.

Li, Y., Wang, H., Duan, Y., Zhang, J., and Li, X. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025.

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

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Modas, A., Moosavi-Dezfooli, S.-M., and Frossard, P. 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.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. 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.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. 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.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Wang, X., Hu, P., Deng, J., and Ma, J. W. Adversarial attacks on data attribution. *arXiv preprint arXiv:2409.05657*, 2024.

Yang, S., Ye, P., Ouyang, W., Zhou, D., and Shen, F. A clip-powered framework for robust and generalizable data selection. *arXiv preprint arXiv:2410.11215*, 2024.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. 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.

Zhang, J., Krishnamurthy, P., Patel, N., Tzes, A., and Khorrami, F. Mp-nav: Enhancing data poisoning attacks against multimodal learning. In *Forty-second International Conference on Machine Learning*.

A. Evaluation Metrics

We measure four complementary aspects of model behavior under X-Shift perturbations: (i) embedding-level stealth, (ii) classifier stability, (iii) spatial attribution consistency at the patch level, and (iv) distributional similarity of the full explanation map. Below we summarize the exact formulations.

Cosine Similarity of CLS Tokens (CosSim ↑). This metric quantifies how close the clean and adversarial global embeddings remain. A high value indicates a *stealthy* attack that preserves high-level semantics. Given the CLS embeddings z_{clean} and z_{adv}:

$$\text{CosSim}_{\text{CLS}} = \frac{z_{\text{clean}} \cdot z_{\text{adv}}}{\|z_{\text{clean}}\|_2 \, \|z_{\text{adv}}\|_2}. \tag{9}$$

Maximum Probability Deviation (MaxΔProb ↓). This term measures the largest change in predicted probability across all text prompts. Low values imply that classification remains unchanged even though the explanation map shifts:

$$\text{Max } \Delta\text{Prob} = \max_j \left| P(y_j \mid x_{\text{clean}}) - P(y_j \mid x_{\text{adv}}) \right|. \tag{10}$$

Intersection-over-Union of Top-k Patches (IoU-Top$k$ ↓). We extract the top-k highest-scoring patches in the similarity map for a target concept, compute the corresponding binary masks M_{clean} and M_{adv}, and evaluate. Lower IoU indicates *stronger spatial manipulation*, as fewer top patches are preserved under the adversarial perturbation. We use either a fixed k or a percentage $k = \alpha HW$ of all patches:

$$\text{IoU}_{\text{Top-}k} = \frac{|M_{\text{clean}} \cap M_{\text{adv}}|}{|M_{\text{clean}} \cup M_{\text{adv}}|}. \tag{11}$$

Soft Intersection-over-Union (Soft-IoU ↓). To capture distributional differences beyond hard top-k sets, we compute a soft approximation using a temperature τ. This measures global distributional drift, complementing IoU-Top$k$:

$$p_{\text{clean}} = \text{softmax}(s_{\text{clean}}/\tau), \qquad p_{\text{adv}} = \text{softmax}(s_{\text{adv}}/\tau), \tag{12}$$

$$\text{Soft-IoU} = \frac{\sum_i \min(p_{\text{clean},i}, p_{\text{adv},i})}{\sum_i \max(p_{\text{clean},i}, p_{\text{adv},i})}. \tag{13}$$

Spearman Rank Correlation (Spearman). We compute the rank correlation between the flattened similarity maps. Low correlation indicates large reordering of influential patches:

$$\rho = \text{Spearman}(s_{\text{clean}}, s_{\text{adv}}). \tag{14}$$

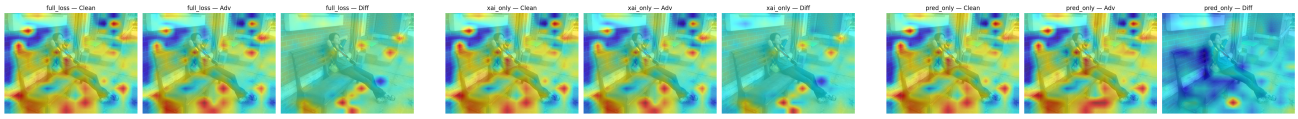
Wasserstein Distance (EMD). We compute the Earth Mover's Distance between flattened similarity scores. EMD captures how much "work" is needed to transform the clean explanation distribution into its adversarial counterpart:

$$\text{EMD}(s_{\text{clean}}, s_{\text{adv}}) = W_1(s_{\text{clean}},\ s_{\text{adv}}). \tag{15}$$

Together, these metrics provide a multi-dimensional characterization of explanation-shifting behavior: *stealth* (CosSim, MaxΔProb), *local spatial reordering* (IoU-Top$k$), and *global distributional drift* (Soft-IoU, Spearman, EMD).

Table 3. Ablation of X-Shift loss components. The full objective achieves the best balance of explanation drift, prediction stability, and stealthiness. ↑/↓ denote preferred direction for attack quality.

Variant	$\text{CosSim}_{\text{CLS}}$↑	MaxΔP↓	IoU_{Topk}↓	TargetSim↑
`full_loss`	0.977	0.000066	0.7857	**0.2406**
`xai_only`	**0.988**	**0.000060**	**0.8182**	0.2405
`pred_only`	0.908	0.000540	0.7241	0.2364



(a) `full_loss`: controlled, targeted drift. *(b)* `xai_only`: strong but unstable drift. *(c)* `pred_only`: minimal explanation shift.

Figure 9. Spatial heatmap ablation of X-Shift loss components. The full objective produces a balanced, targeted explanation shift while preserving the clean prediction. `xai_only` yields stronger but spatially aggressive drift; `pred_only` barely alters the explanation map, confirming that both objective families are necessary for a stable and stealthy attack.

B. Ablation Study: X-Shift Attack Objective

To validate that each component of the X-Shift objective (Eq. 8) is necessary, we perform a controlled ablation over the four loss terms, evaluating three variants: `full_loss` (all terms active), `xai_only` ($\mathcal{L}_{\text{xai}}$ alone), and `pred_only` ($\mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{margin}}$ only). Each variant optimizes toward the same target text embedding under identical hyperparameter settings.

We report four complementary metrics: CLS embedding deviation ($\text{CosSim}_{\text{CLS}}$), maximum class probability change (MaxΔProb), top-k patch overlap between clean and adversarial maps (IoU_{Topk}), and final similarity to the target text embedding (TargetSim). Together, these metrics isolate each term's contribution along the three dimensions that define a high-quality X-Shift attack: *stealthiness*, *prediction stability*, and *explanation drift*.

Effect of $\mathcal{L}_{\text{xai}}$. The `xai_only` variant achieves the strongest patch-level drift ($\text{IoU}_{\text{Topk}} = 0.8182$) and highest TargetSim (0.2405), confirming that $\mathcal{L}_{\text{xai}}$ is the primary driver of explanation manipulation. However, removing prediction-preservation terms increases MaxΔProb by nearly an order of magnitude ($0.000060 \rightarrow 0.000540$ in `pred_only`), producing perturbations that are potentially detectable by prediction-monitoring systems, undermining the core stealth requirement of X-Shift.

Effect of $\mathcal{L}_{\text{pred}}$ and $\mathcal{L}_{\text{margin}}$. The `pred_only` variant yields minimal heatmap drift (lowest IoU_{Topk}) and reduced TargetSim (0.2364), demonstrating that prediction-stability losses alone cannot drive meaningful explanation manipulation. Without $\mathcal{L}_{\text{xai}}$, patch-text alignment remains largely unchanged, confirming that both objective families are necessary and non-substitutable.

Full Objective. The full loss achieves the best balance across all three dimensions: strong explanation drift, a stable CLS embedding ($\text{CosSim}_{\text{CLS}} = 0.977$), and negligible classification change (MaxΔProb $= 0.000066$). Quantitative results are reported in Table 3; qualitative heatmap comparisons are shown in Figure 9.

Spatially, `xai_only` produces concentrated but aggressively redistributed saliency, while `pred_only` preserves most of the clean map structure with minimal drift. The full objective integrates both behaviors: it yields a controlled yet significant shift in saliency while maintaining coherent spatial structure and preserving the original prediction. These results confirm that all components of Eq. (8) jointly contribute to a stable, targeted, and stealthy adversarial perturbation.

C. Comparison of X-Shift Against Standard Adversarial Attacks

Standard adversarial attacks, such as the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014) and Projected Gradient Descent (PGD), are designed to corrupt model predictions by maximizing classification loss. While these methods are effective in inducing misclassification, they do not explicitly target the model's explanation mechanisms, such as patch-level similarity maps in vision–language models.

Table 4. Comparison of X-Shift against prediction-targeted adversarial baselines on CLIP ViT-L/14 (airplane→ground, COCO) and ViT-B/32 (cat→ground, COCO). Baselines use ε=128/255 and 500 PGD steps; X-Shift uses ε=4/255 and 400 Adam steps. **Flip = 1**: prediction changed (red); **Flip = 0**: prediction preserved (green). X-Shift rows shaded. ↑/↓ denote preferred direction.

	ViT-L/14				ViT-B/32			
Attack	IoU↓	CosSim↑	MaxΔP↓	Flip	IoU↓	CosSim↑	MaxΔP↓	Flip
FGSM-T (ε=128/255)	0.1807	0.4732	0.0015	**1**	0.3846	0.4343	0.0014	**1**
FGSM-U (ε=128/255)	0.1951	0.4447	0.0016	**1**	0.0588	0.4304	0.0014	**1**
PGD-T (ε=128/255)	0.1529	0.1853	0.0061	**1**	0.1250	0.1154	0.0041	**1**
PGD-U (ε=128/255)	0.1807	0.2908	0.0045	**1**	0.0588	0.4886	0.0029	**1**
X-Shift (ours, ε=4/255)	**0.2250**	**0.8736**	**0.0003**	**0**	**0.5000**	**0.9549**	**0.0003**	**0**

Figure 10. Clean and adversarial examples generated by FGSM and PGD (targeted and untargeted), compared with X-Shift on the COCO dataset using CLIP ViT-B/32.

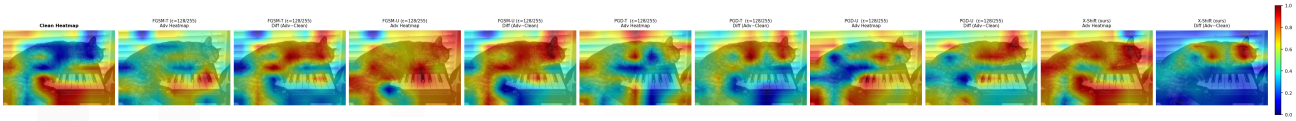


Figure 11. Corresponding explanation (XAI) maps for FGSM, PGD, and X-Shift under targeted and untargeted settings on COCO with CLIP ViT-B/32.

In contrast, X-Shift targets a fundamentally different objective: it manipulates explanation heatmaps while explicitly preserving the model's top-1 prediction. This property renders the attack undetectable by standard prediction-monitoring systems. To empirically validate this distinction, we compare X-Shift against four baselines: FGSM targeted (FGSM-T), FGSM untargeted (FGSM-U), PGD targeted (PGD-T), and PGD untargeted (PGD-U), across two CLIP backbones (ViT-L/14 and ViT-B/32).

All methods are evaluated on COCO 2017 images using CLIP's standard preprocessing pipeline. Baselines are given a substantially larger perturbation budget ($\varepsilon = 128/255$) and 500 optimization steps to maximize their ability to flip predictions. In contrast, X-Shift operates under a much smaller budget ($\varepsilon = 4/255$) with 400 optimization steps. We report four metrics: IoU (for explanation overlap), cosine similarity (CosSim) of global representations, maximum probability change (MaxΔP), and a binary Flip indicator (1 if prediction changes, 0 otherwise). By design, X-Shift enforces Flip$= 0$.

Results. Table 4 summarizes the results across both backbones. The results reveal a clear separation between prediction-targeted and explanation-targeted attacks. Baseline methods achieve prediction flips (Flip$= 1$) but at the cost of significantly degrading the global representation, as reflected in low cosine similarity and higher probability shifts. In contrast, X-Shift preserves prediction consistency (Flip$= 0$) while maintaining high cosine similarity and minimal probability change, indicating negligible impact on the model's decision process.

Orthogonal Threat Surfaces. These findings demonstrate that prediction-targeted and explanation-targeted attacks operate on fundamentally different threat surfaces. Conventional attacks rely on disrupting the global representation to induce misclassification, which can be detected by monitoring prediction confidence or embedding shifts. In contrast,

X-Shift preserves the global representation while selectively manipulating patch-level explanations, bypassing such defenses entirely. This reveals a structural blind spot in current robustness evaluations, which focus exclusively on prediction integrity.

Implications. Overall, Table 4 confirms that X-Shift is not a variant of existing adversarial methods. Also, Figures 10 and 11 demonstrate that X-Shift is not a variant of existing adversarial methods. It is the only approach that simultaneously (i) achieves targeted manipulation of explanation maps, (ii) preserves model predictions, and (iii) operates under a significantly smaller perturbation budget. These properties define a new class of stealthy attacks that challenge the reliability of explanations as a security and trust signal in vision–language systems.