EXPOSING VULNERABILITIES IN VISIBLE-INFRARED VLMS: A UNIFIED ADVERSARIAL FRAMEWORK WITH CROSS-TASK TRANSFERABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models (VLMs) have made substantial progress in multimodal understanding and have achieved strong performance across diverse vision–language tasks. However, existing VLM security studies remain largely confined to single-modality settings, predominantly in the visible spectrum, leaving adversarial robustness in visible–infrared (VIS–IR) scenarios underexplored. This omission is consequential for real-world perception systems, where VIS–IR sensing is widely deployed to sustain reliable semantic understanding under challenging imaging conditions. Motivated by this underexplored cross-modal threat setting, we propose a curved-edge fractal adversarial patch (CFPatch), the first unified adversarial patch framework for cross-modal attacks against VIS–IR VLMs. Specifically, CFPatch takes triangular fractal geometry as its base and transforms rigid straight-edged primitives into Bezier-curved elements, preserving fractal self-similarity for multi-scale adversarial effects while enhancing structural expressiveness through smoother contours, richer directional variation, and flexible shape deformation. Complementing this global geometric disruption, we introduce a modality-specific Fraser-spiral rendering mechanism that injects fine-grained texture distortions and misleading perceptual cues tailored to visible and infrared imagery. Through this geometry–appearance coupling, CFPatch forms a dual-level perturbation: the curved fractal geometry disrupts global shape perception, while Fraser-spiral rendering corrupts local appearance interpretation via modality-specific texture interference. We further use expectation over transformation (EOT) to improve patch robustness against common image-level transformations. Extensive experiments show that CFPatch effectively fools VIS–IR VLMs and outperforms standard patch baselines in both attack effectiveness and robustness. Moreover, adversarial samples optimized for the zero-shot classification task transfer well to image captioning and visual question answering (VQA), demonstrating the strong cross-task transferability and generalizability of CFPatch across downstream tasks.

1 INTRODUCTION

VLMs achieve strong performance across zero-shot classification, image captioning, and VQA (Radford et al., 2021; Jia et al., 2021a; Li et al., 2021; 2022b; 2023; Alayrac et al., 2022; Liu et al., 2023; Chen et al., 2024). However, existing robustness studies remain concentrated on single-modality settings, particularly the visible spectrum (Chen et al., 2023; Xie et al., 2025; Shirnin et al., 2024; Zhao et al., 2023; Zhang et al., 2025a; Ying et al., 2024b; Wang et al., 2026). VIS–IR robustness therefore remains insufficiently understood, despite the widespread use of dual-spectrum perception, where infrared sensing complements visible imagery under poor illumination and degraded imaging conditions (Tang et al., 2022; Zhang et al., 2020). Perturbing only

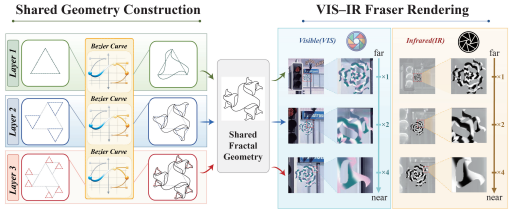

Figure 1: Shared curved-fractal construction and VIS–IR Fraser rendering of CFPatch.

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.

one branch thus provides an incomplete view of system vulnerability, motivating attacks that jointly disrupt both spectra. Existing VIS–IR adversarial patches remain detector-oriented, focusing mainly on detection or tracking rather than VLM semantic reasoning (Wei et al., 2023; Hu et al., 2025; Long et al., 2025).

Moving from detector-oriented VIS–IR attacks to semantic attacks on VIS–IR VLMs introduces intertwined challenges. A unified patch should preserve shared geometry across spectra rather than reduce the setting to independent single-modality attacks. At the same time, this shared structure must remain expressive enough to accommodate the distinct imaging characteristics of visible and infrared branches. Beyond cross-spectral consistency, attacking VLMs also requires perturbations that affect semantic reasoning rather than only local visual evidence. These requirements motivate a compact design that jointly modulates coarse shape and fine-grained appearance cues, consistent with evidence that high-level reasoning depends on globally coherent visual information (Li et al., 2022a; Covert et al., 2025; Zhang et al., 2025b).

To address these challenges, we introduce CFPatch, the first unified adversarial patch framework designed to expose cross-modal vulnerabilities in VIS–IR VLMs. Figure 1 illustrates the two core components of our design: a curved-edge triangular fractal geometry and modality-specific Fraser-spiral rendering. Specifically, we take triangular fractal geometry as the base structure and redesign its originally rigid straight-edged primitives into Bezier-curved elements, producing a compact multi-scale geometric carrier shared across spectra. The Fraser-spiral renderer then instantiates this shared carrier with visible- and infrared-specific appearance patterns. This geometry–appearance coupling yields a dual-level perturbation that disrupts global shape perception and local appearance interpretation within a unified cross-modal geometry.

We optimize CFPatch in a compact parameter space using particle swarm optimization (PSO), and incorporate EOT to improve the robustness of the learned patch against common image-level transformations such as scale changes, rotations, blur, and appearance variations (Kennedy & Eberhart, 1995; Athalye et al., 2018). A paired attack is counted as successful only when the same shared-geometry patch fools both the visible and infrared branches simultaneously. We use zero-shot classification as the primary optimization task and directly reuse the resulting adversarial samples for image captioning and VQA without task-specific re-optimization.

Overall, our work extends VIS–IR adversarial patch research from detector-oriented evasion to a broader semantic threat model for VLMs. Our contributions are summarized as follows:

- To the best of our knowledge, we present CFPatch, the first unified adversarial patch framework for VIS–IR VLMs. The framework is instantiated by transforming triangular fractal geometry into a shared curved-edge geometric carrier and coupling it with Fraser-spiral-based modality-specific rendering, thereby integrating cross-spectral structural consistency with spectrum-adaptive appearance interference.
- Through extensive experiments on zero-shot classification, image captioning, and VQA, we show that CFPatch consistently surpasses representative VIS–IR patch baselines and achieves strong simultaneous attack performance across visible and infrared branches.
- Ablation studies and detailed analysis further reveal the contributions of key design choices, including curved-edge fractal geometry, modality-specific Fraser rendering, fractal depth, and optimization budget, thereby validating the robustness and cross-task transferability of the framework.

2 Related Work

Robustness of vision–language models. VLMs are vulnerable to textual and visual perturbations. Prompt design affects zero-shot behavior (Zhou et al., 2022b;a; Khattak et al., 2023a;b), while image-space attacks degrade multimodal representations (Chen et al., 2023; Xie et al., 2025; Shirnin et al., 2024; Zhao et al., 2023; Zhang et al., 2025a; Ying et al., 2024b). Benchmarks also show limited robustness to realistic visual shifts (Wang et al., 2026; Qiu et al., 2025; Usama et al., 2025; Ying et al., 2024a), yet VIS–IR robustness remains largely unexplored.

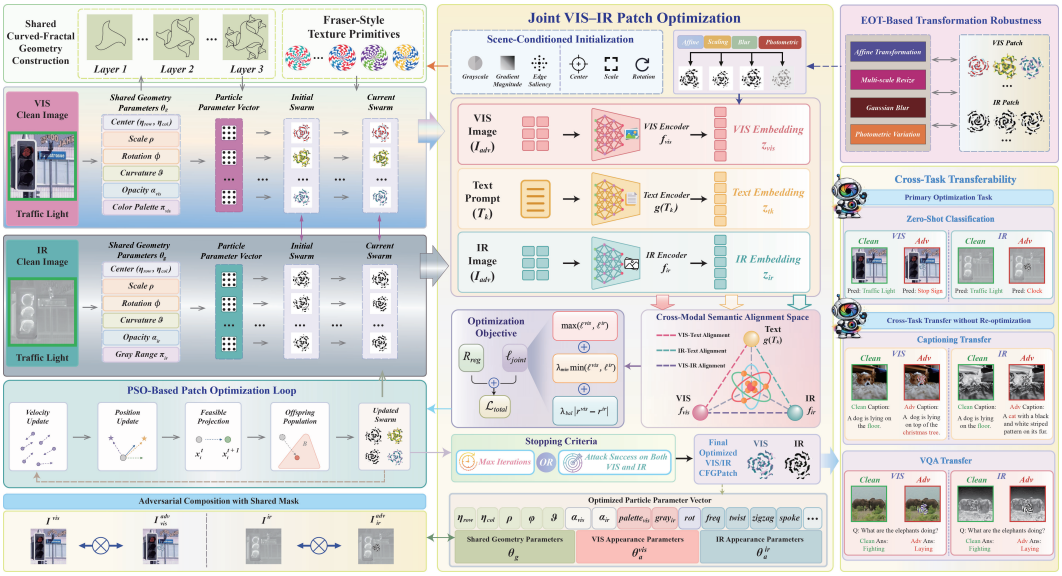

Figure 2: Overview of CFPatch: a shared curved–edge fractal patch for VIS–IR inputs is optimized under PSO with EOT and transferred across tasks.

Geometry priors and structured perturbations. Structured perturbations provide compact priors over pixel-space attacks. Localized adversarial patches alter predictions (Brown et al., 2017; Eykholt et al., 2018; Chen et al., 2019; Liu et al., 2019), while EOT improves transformation robustness (Athalye et al., 2018). Fractal self-similarity yields multi-scale patterns from few parameters (Lutz & Mayordomo, 2008), disrupting coarse and local cues. Related physical attacks exploit shadows, light, illumination, and 3D variation (Zhong et al., 2022; Hsiao et al., 2024; Liu et al., 2025; Ruan et al., 2025).

Cross-modal visible–infrared adversarial attacks. Recent VIS–IR attacks remain detector-centric, using shared shapes or modality-adapted appearance against detectors (Wei et al., 2023; Hu et al., 2025; Long et al., 2025). Infrared VLMs are emerging (Jiang et al., 2024; Cao et al., 2025), and universal patches expose their vulnerability (Hu et al., 2026). We instead use one shared-geometry patch to jointly attack visible and infrared VLMs.

3 Method

3.1 Overview

Given paired visible and infrared images $(I^{\text{vis}}, I^{\text{ir}})$ with ground-truth label y, we aim to construct a single cross-modal patch family that enforces shared geometric support across spectra while allowing modality-specific appearance rendering. Figure 2 summarizes CFPatch, which combines shared geometric disruption with modality-specific appearance interference: a common patch mask provides spectrum-consistent spatial structure, while visible Fraser-style and infrared grayscale textures introduce spectrum-dependent perturbations. The patch is optimized with a joint VIS–IR objective for zero-shot classification, and the resulting adversarial samples are subsequently evaluated on image captioning and VQA.

3.2 Problem Definition

We formulate the attack under the zero-shot VIS–IR classification task, where each paired image sample is associated with a ground-truth label $y \in \mathcal{Y} = \{1, \ldots, K\}$. For each modality $m \in \{\text{vis}, \text{ir}\}$, the VLM performs recognition by comparing the image representation with class-level text prompts.

The image–text similarity score for class k is computed as:

$$s_k^{(m)}(I) = \langle f_m(I),\, g(T_k) \rangle \tag{1}$$

where $f_m(\cdot)$ denotes the image encoder for modality m, $g(\cdot)$ is the shared text encoder, and T_k is the text prompt associated with class k. The predicted label for modality m is then given by:

$$\hat{y}^{(m)}(I) = \arg\max_{k \in \mathcal{Y}} s_k^{(m)}(I) \tag{2}$$

Here, $s_k^{(m)}(I)$ measures the alignment between image I and class prompt T_k, while $\hat{y}^{(m)}(I)$ denotes the resulting zero-shot prediction.

We evaluate only paired samples for which both clean branches are correctly classified, so that attack success is not inflated by pre-existing recognition failures:

$$\hat{y}^{(\mathrm{vis})}\big(I^{\mathrm{vis}}\big) = y, \qquad \hat{y}^{(\mathrm{ir})}\big(I^{\mathrm{ir}}\big) = y \tag{3}$$

After applying the adversarial patch, a paired attack is counted as successful only when both the visible and infrared branches are misclassified:

$$\hat{y}^{(\mathrm{vis})}\big(I^{\mathrm{vis}}_{\mathrm{adv}}\big) \neq y, \qquad \hat{y}^{(\mathrm{ir})}\big(I^{\mathrm{ir}}_{\mathrm{adv}}\big) \neq y \tag{4}$$

This conjunction-based criterion defines a strict VIS–IR attack protocol, where fooling only one spectrum is insufficient for cross-modal success.

3.3 Shared Curved-Triangle Fractal Geometry

The patch support is defined by a recursively constructed curved-triangle fractal layout. Starting from a canonical equilateral triangle, we generate a hierarchy of guide triangles and retain the level-1, level-2, and level-3 structures jointly, resulting in a compact yet multi-scale support region. This recursive design is important because the self-similarity of fractal geometry introduces repeated patterns across scales, enabling the patch to simultaneously interfere with coarse object-level structure and fine-grained local evidence.

To further enhance geometric expressiveness, each triangle edge is parameterized by a cubic Bezier curve instead of a straight segment. For an edge with endpoints $p_0, p_3 \in \mathbb{R}^2$ and control points $p_1, p_2 \in \mathbb{R}^2$, the curve is defined as:

$$B(t) = (1-t)^3 p_0 + 3(1-t)^2 t\, p_1 + 3(1-t) t^2 p_2 + t^3 p_3 \tag{5}$$

where $t \in [0, 1]$ is the curve parameter, and p_1 and p_2 determine the local tangent and curvature. Compared with rigid polygonal edges, this parameterization produces smoother boundaries, richer directional variation, and more expressive shape deformations.

We denote the shared geometry parameters by:

$$\theta_g = \{\eta_{\mathrm{row}}, \eta_{\mathrm{col}}, \rho, \phi, \vartheta, \theta_{\mathrm{shape}}\} \tag{6}$$

where η_{row} and η_{col} denote the patch center coordinates along the image row and column directions, respectively, ρ is the global scale, ϕ is the in-plane rotation, ϑ controls the overall curvature strength, and θ_{shape} collects the remaining shape variables induced by the recursive construction. By rasterizing and unioning all retained curved triangles, we obtain a shared geometry mask:

$$M_g(\theta_g) \in [0, 1]^{H \times W} \tag{7}$$

where H and W denote the image height and width. The same mask $M_g(\theta_g)$ is applied to both visible and infrared modalities, thereby directly enforcing the shared-geometry constraint.

3.4 Modality-Specific Fraser Rendering

While the patch geometry is shared across modalities, its appearance is rendered in a modality-specific manner. This design is motivated by a simple physical intuition: a single patch carrier can preserve a common geometric identity while exhibiting different spectral responses under visible and infrared imaging. Based on this observation, we build a Fraser-spiral-based rendering module on top of the shared geometry mask.

Table 1: Attack results (%) for zero-shot classification on paired VIS–IR inputs. Higher values indicate stronger attack success.

Method	OpenAI CLIP ViT-L/14			OpenCLIP ViT-B/16			Meta-CLIP ViT-L/14			EVA-CLIP ViT-G/14		
	VIS	IR	VIS–IR	VIS	IR	VIS–IR	VIS	IR	VIS–IR	VIS	IR	VIS–IR
Unified Patch	21.15	36.54	14.10	24.05	40.51	17.72	10.22	13.44	8.06	16.26	21.18	12.81
TOUAP	38.46	46.79	25.64	43.04	69.62	36.10	29.03	55.37	25.27	39.41	40.87	27.09
Ours (Random)	61.53	65.38	54.49	72.15	88.61	65.19	63.44	66.13	59.14	65.02	73.40	64.53
Ours	**69.54**	**84.48**	**67.82**	**80.38**	**92.41**	**77.22**	**72.58**	**80.65**	**68.28**	**77.34**	**79.80**	**73.89**

Let θ_a^{vis} and θ_a^{ir} denote the appearance parameters for the visible and infrared branches, respectively. Both branches are driven by a common Fraser-style coordinate field, whose structure is controlled by factors such as rotation, phase, ring frequency, twist, zigzag amplitude, spoke number, and sharpness. On this basis, the visible branch maps the field to a color palette, whereas the infrared branch maps it to a grayscale intensity range. The resulting patch textures are defined as:

$$P^{\mathrm{vis}} = R_{\mathrm{vis}}\big(\theta_g, \theta_a^{\mathrm{vis}}\big) \tag{8}$$

$$P^{\mathrm{ir}} = R_{\mathrm{ir}}\big(\theta_g, \theta_a^{\mathrm{ir}}\big) \tag{9}$$

where $R_{\mathrm{vis}}(\cdot)$ and $R_{\mathrm{ir}}(\cdot)$ denote the rendering functions for the visible and infrared modalities, respectively. In both branches, the spatial support is inherited from the shared geometry parameters θ_g, while the texture appearance is controlled by modality-specific parameters. The adversarial visible and infrared images are then obtained by alpha blending:

$$I_{\mathrm{adv}}^{\mathrm{vis}} = (1 - \alpha_{\mathrm{vis}} M_g) \odot I^{\mathrm{vis}} + \alpha_{\mathrm{vis}} M_g \odot P^{\mathrm{vis}} \tag{10}$$

$$I_{\mathrm{adv}}^{\mathrm{ir}} = (1 - \alpha_{\mathrm{ir}} M_g) \odot I^{\mathrm{ir}} + \alpha_{\mathrm{ir}} M_g \odot P^{\mathrm{ir}} \tag{11}$$

where $\odot$ denotes element-wise multiplication and $M_g = M_g(\theta_g)$ is the shared geometry mask. The coefficients $\alpha_{\mathrm{vis}}, \alpha_{\mathrm{ir}} \in [0, 1]$ control modality-specific opacity. This construction yields a two-level perturbation mechanism: the outer curved fractal geometry disrupts global shape perception, while the inner Fraser-style texture interferes with local appearance cues.

3.5 Joint Cross-Modal Objective

Our goal is to drive the visible and infrared branches toward simultaneous failure, rather than allowing the optimization to overfit only one modality. For each modality $m \in \{\mathrm{vis}, \mathrm{ir}\}$, let $s_y^{(m)}$ denote the similarity score of the ground-truth class y, and let the strongest competing score be defined as:

$$s_{\max \setminus y}^{(m)} = \max_{k \neq y} s_k^{(m)} \tag{12}$$

Based on this quantity, we define a margin-based loss for each modality as:

$$\ell^{(m)} = \mathrm{softplus}\Big(s_y^{(m)} - s_{\max \setminus y}^{(m)} + \delta\Big) \tag{13}$$

where $\delta > 0$ is a safety margin. Intuitively, $\ell^{(m)}$ becomes small when the strongest incorrect class already exceeds the ground-truth class by a sufficient margin. To more directly characterize attack progress, we further define the attack reward as:

$$r^{(m)} = s_{\max \setminus y}^{(m)} - s_y^{(m)} \tag{14}$$

which measures the extent to which the most competitive incorrect class dominates the true class.

To couple the two modalities, we construct the following joint objective:

$$\ell_{\mathrm{joint}} = \max\big(\ell^{(\mathrm{vis})}, \ell^{(\mathrm{ir})}\big) + \lambda_{\min} \min\big(\ell^{(\mathrm{vis})}, \ell^{(\mathrm{ir})}\big) + \lambda_{\mathrm{bal}} \big| r^{(\mathrm{vis})} - r^{(\mathrm{ir})} \big| \tag{15}$$

where $\lambda_{\min} > 0$ controls the contribution of the easier modality, and $\lambda_{\mathrm{bal}} > 0$ penalizes imbalanced attack progress across spectra. In our implementation, we set $\lambda_{\min} = 0.55$. The first term emphasizes

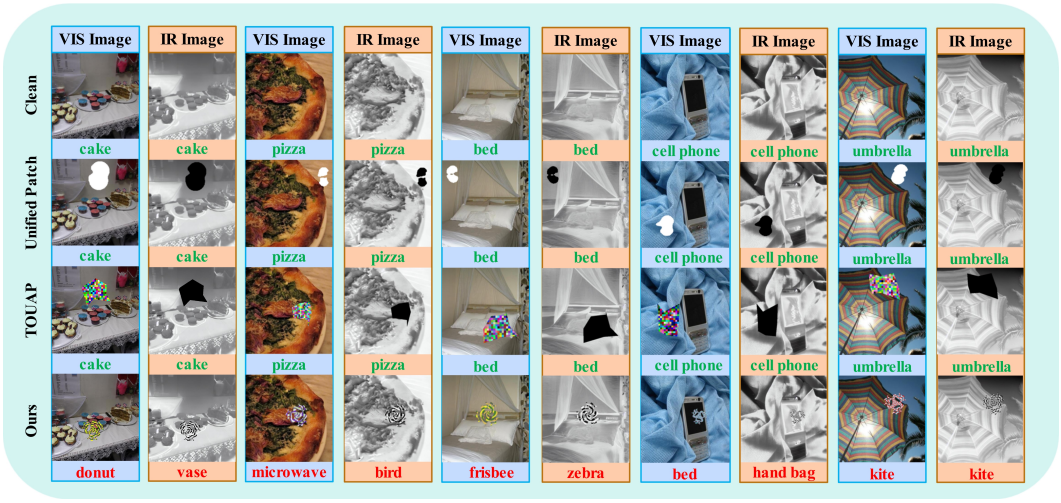

Figure 3: Qualitative classification examples in the VIS–IR setting. From top to bottom: Clean, Unified Patch, TOUAP, and Ours (CFPatch); left: VIS Image, right: IR Image. Green and red text denote correct and adversarial predictions, respectively.

the harder branch, the second term preserves optimization pressure on the easier branch, and the balance term discourages one-sided fooling.

To improve robustness under realistic deployment variations, we optimize the patch under expectation over transformation. Let $\tau \sim \mathcal{T}$ denote a random transformation sampled from a distribution $\mathcal{T}$ over affine perturbations, scale changes, blur, and photometric variations. The final optimization problem is formulated as:

$$\min_{\theta_g, \theta_a^{\text{vis}}, \theta_a^{\text{ir}}} \mathbb{E}_{\tau \sim \mathcal{T}} \left[\ell_{\text{joint}} \Big(\tau(I_{\text{adv}}^{\text{vis}}), \tau(I_{\text{adv}}^{\text{ir}}) \Big) \right] + \lambda_{\text{reg}} \mathcal{R}_{\text{reg}} \tag{16}$$

where $\mathcal{R}_{\text{reg}}$ denotes the overall regularization term and $\lambda_{\text{reg}} > 0$ controls its strength. Specifically, $\mathcal{R}_{\text{reg}}$ consists of lightweight penalties on patch area, mask smoothness, geometry stabilization, and texture smoothness, which help constrain the patch while preserving sufficient attack flexibility. We use the notation $\mathcal{R}_{\text{reg}}$ instead of $\mathcal{R}$ to avoid ambiguity with the rendering operators R_{vis} and R_{ir}.

3.6 Scene-Conditioned Initialization and PSO

Since our patch involves recursive rasterized geometry, discrete palette choices, and non-smooth Fraser rendering, gradient-based optimization is unreliable in this parameter space. We therefore optimize the shared geometry and modality-specific appearance parameters with PSO.

To provide a stable and informative starting point, we use a scene-conditioned initialization strategy. The visible and infrared images are first converted to grayscale, their gradient magnitudes are averaged to capture cross-modal edge saliency, and the initial patch center is estimated from regions with concentrated structural responses. A lightweight coarse search over candidate centers, scales, and rotations further refines this state, biasing the patch toward structurally informative regions that are consistently salient across both modalities.

Each particle encodes the full parameter vector, including shared geometry and modality-specific appearance parameters. For particle i at iteration t, let x_i^t and v_i^t denote its position and velocity, b_i^t its personal best, and $b_\star^t$ the global best. The PSO update is defined as:

$$v_i^{t+1} = \omega\, v_i^t + c_1 \xi_{1,i}^t \odot \left(b_i^t - x_i^t\right) + c_2 \xi_{2,i}^t \odot \left(b_\star^t - x_i^t\right) \tag{17}$$

$$x_i^{t+1} = \Pi_{\mathcal{B}}\left(x_i^t + v_i^{t+1}\right) \tag{18}$$

where ω is the inertia weight, c_1 and c_2 are the cognitive and social coefficients, $\xi_{1,i}^t$ and $\xi_{2,i}^t$ are element-wise random vectors sampled from $[0, 1]$, and $\Pi_{\mathcal{B}}$ projects parameters onto the feasible

Table 2: Attack results (%) on image captioning for paired VIS–IR inputs. Values in parentheses indicate the performance drop relative to the corresponding clean baseline.

Image Encoder	Models	Clean		Unified Patch		TOUAP		Ours	
		VIS	IR	VIS↓	IR↓	VIS↓	IR↓	VIS↓	IR↓
OpenAI CLIP ViT-L/14	LLaVA-1.5 (7B)	78.53	65.40	76.00 (2.53)	58.87 (6.53)	71.95 (6.58)	55.15 (10.25)	**65.39 (13.14)**	**50.25 (15.15)**
	LLaVA-1.6 (7B)	82.37	68.00	81.12 (1.25)	57.17 (10.83)	71.25 (11.12)	56.45 (11.55)	**67.20 (15.17)**	**51.54 (16.46)**
	OpenFlamingo (3B)	79.46	67.36	77.94 (1.52)	61.60 (5.76)	71.20 (8.26)	56.94 (10.42)	**64.53 (14.93)**	**47.31 (20.05)**
	Blip-2 FlanT5XL ViT-L (3.4B)	85.68	62.19	83.95 (1.73)	57.09 (5.10)	79.40 (6.28)	54.26 (7.93)	**75.65 (10.03)**	**49.12 (13.07)**
EVA-CLIP ViT-G/14	Blip-2 FlanT5XL (4.1B)	75.38	65.03	71.89 (3.49)	57.38 (7.65)	63.84 (11.54)	56.83 (8.20)	**60.92 (14.46)**	**48.97 (16.06)**
	InstructBLIP FlanT5XL (4.1B)	77.56	68.44	75.64 (1.92)	65.73 (2.71)	72.11 (5.45)	59.37 (9.07)	**69.81 (7.75)**	**52.91 (15.53)**

domain $\mathcal{B}$. At each iteration, particles are rendered under the shared-geometry constraint, evaluated with sampled EOT transformations, and updated according to Equations 17–18. The search stops early once both modalities are successfully fooled. Pseudo-code is provided in the supplementary material.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate CFPatch on a paired VIS–IR benchmark of 300 MS COCO images spanning 30 semantic categories (Lin et al., 2014; Chen et al., 2015). These categories define the zero-shot label space using standard prompt templates (Zhou et al., 2022b; Khattak et al., 2023b). Following (Han et al., 2024), each visible image is converted into a semantically aligned infrared counterpart while preserving scene semantics. For captioning and VQA, we reuse adversarial samples generated from zero-shot classification together with the corresponding annotations. Prompt templates are provided in the supplementary material. We also evaluate real paired LLVIP pedestrian images (Jia et al., 2021b); see Appendix G.

Models and Infrared Adaptation Protocol. For zero-shot classification, we evaluate OpenCLIP ViT-B/16 (Cherti et al., 2023), Meta-CLIP ViT-L/14 (Xu et al., 2024), EVA-CLIP ViT-G/14 (Sun et al., 2023), and OpenAI CLIP ViT-L/14 (Radford et al., 2021); for captioning and VQA, we use LLaVA-1.5 (Liu et al., 2023), LLaVA-1.6 (Liu et al., 2024), OpenFlamingo (Awadalla et al., 2023), BLIP-2 (Li et al., 2023), and InstructBLIP (Dai et al., 2023). Infrared backbones are adapted from their visible counterparts using an infrared semantic corpus (Jiang et al., 2024) with lightweight LoRA (Hu et al., 2022). Generative models use next-token prediction, while contrastive models use InfoNCE (van den Oord et al., 2018), preserving architectural comparability. See Appendix I for adaptation details.

Evaluation Metrics. For zero-shot classification, we report VIS, IR, and joint VIS–IR attack success rates (ASR). For captioning and VQA, we follow an LLM-as-Judge protocol (Bavaresco et al., 2025), using GPT-5 to assess caption consistency and answer correctness and reporting drops from the clean baseline.

Baselines. For zero-shot classification, we compare CFPatch and its random-search and PSO variants with Unified Patch (Wei et al., 2023) and TOUAP (Hu et al., 2025). All baselines use the same protocol for captioning and VQA.

Implementation Details. All attacks are optimized using particle swarm optimization (PSO) under EOT (Kennedy & Eberhart, 1995; Athalye et al., 2018). In our experiments, we set the swarm size to 20, the optimization budget to 150 iterations, and the number of EOT samples to 2. The patch is parameterized as a depth-3 curved fractal structure, with optimization variables covering shared geometry, opacity, visible palette, infrared grayscale range, and Fraser-texture parameters.



Figure 4: Qualitative examples. (a) Captioning under clean and adversarial VIS–IR inputs. (b) VQA under clean and adversarial conditions for both visible and infrared branches.

Table 3: Attack results (%) on visual question answering (VQA) for paired VIS–IR inputs. Values in parentheses indicate the performance drop relative to the corresponding clean baseline.

Image Encoder	Models	Clean		Unified Patch		TOUAP		Ours	
		VIS	IR	VIS↓	IR↓	VIS↓	IR↓	VIS↓	IR↓
OpenAI CLIP ViT-L/14	LLaVA-1.5 (7B)	65.38	62.27	61.97 (3.41)	58.73 (3.54)	59.24 (6.14)	54.74 (7.53)	**52.81 (12.57)**	**49.80 (12.47)**
	LLaVA-1.6 (7B)	61.75	54.64	59.11 (2.64)	53.01 (1.63)	56.24 (5.51)	49.12 (5.52)	**50.12 (11.63)**	**42.58 (12.06)**
	OpenFlamingo (3B)	68.26	63.58	66.74 (1.52)	60.73 (2.85)	64.11 (4.15)	57.46 (6.12)	**55.91 (12.35)**	**53.02 (10.56)**
	Blip-2 FlanT5XL ViT-L (3.4B)	57.07	52.46	53.92 (3.15)	48.70 (3.76)	50.65 (6.42)	45.72 (6.74)	**42.77 (14.30)**	**39.26 (13.20)**
EVA-CLIP ViT-G/14	Blip-2 FlanT5XL (4.1B)	67.98	63.15	65.68 (2.30)	59.40 (3.75)	62.09 (5.89)	54.58 (8.57)	**51.97 (16.01)**	**45.90 (17.25)**
	InstructBLIP FlanT5XL (4.1B)	68.41	64.72	66.73 (1.68)	61.45 (3.27)	62.78 (5.63)	56.87 (7.85)	**56.90 (11.51)**	**51.60 (13.12)**

4.2 Zero-shot Classification

Zero-shot classification serves as the primary evaluation task for assessing whether a single shared patch can jointly fool both visible and infrared recognition. Table 1 reports the ASR for the VIS and IR branches individually, together with the strict joint VIS–IR success rate that requires both modalities to be fooled for the same paired sample. Across all four CLIP backbones, CFPatch consistently achieves the highest attack success and outperforms Unified Patch and TOUAP, with particularly strong gains on the joint VIS–IR metric. These results indicate that the proposed shared geometry and modality-specific rendering can effectively coordinate attacks across the two spectral branches rather than overfitting to only one modality. Qualitative examples are shown in Figure 3.

4.3 Image Captioning Evaluation

The adversarial VIS–IR pairs optimized for zero-shot classification are directly transferred to image captioning without task-specific re-optimization. As shown in Table 2, CFPatch causes the largest semantic degradation across all evaluated captioning models for both visible and infrared inputs, consistently exceeding Unified Patch and TOUAP. This result suggests that the attack effect is not confined to closed-set category decisions, but transfers to free-form generation by disrupting the visual semantics required for coherent caption construction. The consistent degradation across different multimodal architectures further supports the cross-model and cross-task generalizability of the learned adversarial structure. Representative examples are shown in Figure 4 (a).



Figure 5: Patch variants in component ablation.

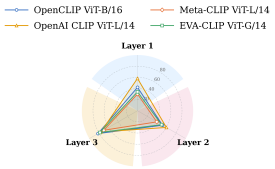

Figure 6: Fractal-depth ablation results.

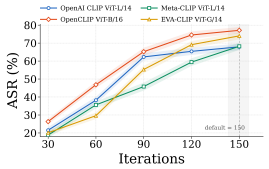

Figure 7: Iteration-budget ablation results.

4.4 Visual Question Answering Evaluation

We further evaluate visual question answering using the same adversarial VIS–IR pairs. Table 3 shows that CFPatch again produces the largest performance degradation across models and modalities, outperforming Unified Patch and TOUAP under both visible and infrared inputs. Unlike classification, VQA requires image-conditioned semantic grounding and reasoning, so the observed degradation indicates that CFPatch disrupts higher-level multimodal evidence rather than only category-level recognition cues. Together with the captioning results, these findings demonstrate that adversarial samples optimized on zero-shot classification transfer effectively to downstream language reasoning tasks. Representative examples are shown in Figure 4 (b).

4.5 Ablation and Analysis

Component-level Ablation. We examine the contributions of Fraser rendering and Bezier boundary deformation. As shown in Table 4, removing either component consistently reduces joint VIS–IR attack success across all four backbones, highlighting the complementary roles of modality-specific texture interference and expressive boundary geometry. Figure 5 further visualizes these component variants.

Methods	OpenAI CLIP ViT-L/14	OpenCLIP ViT-B/16	Meta-CLIP ViT-L/14	EVA-CLIP ViT-G/14
w/o Fraser	56.82	55.62	50.54	64.38
w/o Bezier	57.19	60.45	62.73	67.12
CFPatch	**67.82**	**77.22**	**68.28**	**73.89**

Table 4: Component-level ablation results.

Fractal Depth Ablation. We next study the effect of recursive fractal depth. Figure 6 shows that increasing the recursion depth generally improves attack effectiveness across the evaluated backbones. This trend suggests that deeper fractal structures provide richer multi-scale geometric patterns, which can interfere with both coarse semantic shape cues and finer local evidence. The gains are not identical across models, indicating that different backbones exhibit different sensitivities to recursive geometric complexity.

Iteration Budget Ablation. We further analyze the influence of the PSO iteration budget while keeping the remaining settings fixed. Figure 7 shows that attack success improves markedly as the budget increases from small values and continues to rise toward the default setting of 150 iterations. The improvement becomes progressively smaller at larger budgets, suggesting that the optimization benefits from sufficient exploration but exhibits diminishing returns once effective cross-modal patch configurations have been identified.

Additional Analysis. Additional attention-shift visualizations and defense evaluations are provided in Appendices E and F, respectively.

5 Discussion and Conclusion

We presented CFPatch, a unified adversarial framework for VIS–IR VLMs combining curved fractal geometry, modality-specific Fraser rendering, and PSO with EOT. Experiments across zero-shot classification, image captioning, and VQA demonstrate joint VIS–IR attack performance and cross-task transferability, revealing persistent vulnerabilities to coupled geometric and appearance perturbations.

AI USE STATEMENT

Generative AI was used for two purposes: GPT-5 served as an LLM-as-Judge for caption consistency and VQA correctness, and AI tools assisted with language polishing, LaTeX formatting, code debugging, and presentation refinement. All hypotheses, methodological decisions, experimental protocols, results, and conclusions were determined and validated by the authors. AI-assisted outputs were manually reviewed, and the authors take full responsibility for the final content.

ETHICS STATEMENT

This work studies adversarial robustness in VIS–IR VLMs using a unified geometric patch attack. The method is intended to diagnose cross-modal vulnerabilities and support robustness research. Although adversarial patches could be misused, our experiments are limited to research models and benchmark data and do not target deployed or safety-critical systems.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In Sanmi Koyejo, Shakir Mohamed, Anima Agarwal, Danielle Belgrave, Kyunghyun Cho, and Alice Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 284–293. PMLR, 10–15 Jul 2018.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models, 2023.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–255, Vienna, Austria, jul 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.20.

Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch, 2017.

Zhe Cao, Jin Zhang, and Ruiheng Zhang. Irgpt: Understanding real-world infrared image with bi-cross-modal curriculum on large-scale benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 166–176, October 2025.

Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng (Polo) Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley, and Georgiana Ifrim (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 52–68, Cham, 2019. Springer International Publishing. ISBN 978-3-030-10925-7.

Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip Torr, and Volker Tresp. Benchmarking robustness of adaptation methods on pre-trained vision-language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 51758–51777. Curran Associates, Inc., 2023.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, June 2024.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, June 2023.

Ian Covert, Tony Sun, James Y. Zou, and Tatsunori Hashimoto. Locality alignment improves vision-language models. In *International Conference on Learning Representations*, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267. Curran Associates, Inc., 2023.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1625–1634, June 2018.

Zonghao Han, Shun Zhang, Yuru Su, Xiaoning Chen, and Shaohui Mei. Dr-avit: Toward diverse and realistic aerial visible-to-infrared image translation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. doi: 10.1109/TGRS.2024.3405989.

Teng-Fang Hsiao, Bo-Lun Huang, Zi-Xiang Ni, Yan-Ting Lin, Hong-Han Shuai, Yung-Hui Li, and Wen-Huang Cheng. Natural light can also be dangerous: Traffic sign misinterpretation under adversarial natural light attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3915–3924, January 2024.

Chengyin Hu, Weiwen Shi, Wen Yao, Tingsong Jiang, Ling Tian, and Wen Li. Two-stage optimized unified adversarial patch for attacking visible-infrared cross-modal detectors in the physical world. *Applied Soft Computing*, 171:112818, 2025. ISSN 1568-4946. doi: 10.1016/j.asoc.2025.112818.

Chengyin Hu, Yuxian Dong, Yikun Guo, Xiang Chen, Junqi Wu, Jiahuan Long, Yiwei Wei, Tingsong Jiang, and Wen Yao. Revealing physical-world semantic vulnerabilities: Universal adversarial patches for infrared vision-language models, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*, 2022.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 18–24 Jul 2021a.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. LLVIP: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 3496–3504, October 2021b. doi: 10.1109/ICCVW54120.2021.00389.

Shixin Jiang, Zerui Chen, Jiafeng Liang, Yanyan Zhao, Ming Liu, and Bing Qin. Infrared-LLaVA: Enhancing understanding of infrared images in multi-modal large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8573–8591, Miami, Florida, USA, nov 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.501.

James Kennedy and Russell C. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pp. 1942–1948, 1995. doi: 10.1109/ICNN.1995.488968.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122, June 2023a.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15190–15200, October 2023b.

Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. In Sanmi Koyejo, Shakir Mohamed, Anima Agarwal, Danielle Belgrave, Kyunghyun Cho, and Alice Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 7290–7303. Curran Associates, Inc., 2022a.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9694–9705. Curran Associates, Inc., 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 17–23 Jul 2022b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Hanqing Liu, Shouwei Ruan, Yao Huang, Shiji Zhao, and Xingxing Wei. When lighting deceives: Exposing vision-language models' illumination vulnerability through illumination transformation attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10485–10495, October 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. DPatch: An adversarial patch attack on object detectors. In *Proceedings of the AAAI Workshop on Artificial Intelligence Safety (SafeAI 2019)*, volume 2301 of *CEUR Workshop Proceedings*. CEUR-WS, 2019.

Jiahuan Long, Wen Yao, Tingsong Jiang, Jiacheng Hou, Shuai Jia, Junqi Wu, Xiaoya Zhang, Xiaohu Zheng, and Chao Ma. Cdupatch: Color-driven universal adversarial patch attack for dual-modal visible-infrared detectors. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1462–1470, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755188.

Jack H. Lutz and Elvira Mayordomo. Dimensions of points in self-similar fractals. *SIAM Journal on Computing*, 38(3):1080–1112, 2008. doi: 10.1137/070684689.

Muzammal Naseer, Salman H. Khan, and Fatih Porikli. Local gradients smoothing: Defense against localized adversarial attacks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1300–1307. IEEE, 2019. doi: 10.1109/WACV.2019.00143.

Xinkuan Qiu, Meina Kan, Yongbin Zhou, and Shiguang Shan. Benchmarking multimodal large language models against image corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9014–9023, October 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Shouwei Ruan, Hanqing Liu, Yao Huang, Xiaoqi Wang, Caixin Kang, Hang Su, Yinpeng Dong, and Xingxing Wei. Advdreamer unveils: Are vision-language models truly ready for real-world 3d variations? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7894–7904, October 2025.

Alexander Shirnin, Nikita Andreev, Sofia Potapova, and Ekaterina Artemova. Analyzing the robustness of vision & language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2751–2763, 2024. doi: 10.1109/TASLP.2024.3399061.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023.

Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83-84:79–92, 2022. ISSN 1566-2535. doi: 10.1016/j.inffus.2022.03.007.

Muhammad Usama, Syeda Aishah Asim, Syed Bilal Ali, Syed Talal Wasim, and Umair Bin Mansoor. Analysing the robustness of vision-language-models to common corruptions, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

Huanchen Wang, Wencheng Zhang, Zhiqiang Wang, Zhicong Lu, and Yuxin Ma. Vismodai: Visual analytics for evaluating and improving corruption robustness of vision-language models. *IEEE Transactions on Visualization and Computer Graphics*, 32(1):615–625, 2026. doi: 10.1109/TVCG.2025.3634257.

Xingxing Wei, Yao Huang, Yitong Sun, and Jie Yu. Unified adversarial patch for cross-modal attacks in the physical world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4445–4454, October 2023.

Chong Xiang, Saeed Mahloujifar, and Prateek Mittal. PatchCleanser: Certifiably robust defense against adversarial patches for any image classifier. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 2065–2082, Boston, MA, August 2022. USENIX Association. ISBN 978-1-939133-31-1.

Peng Xie, Yequan Bie, Jianda Mao, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14679–14689, June 2025.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *International Conference on Learning Representations*, 2024.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMT-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57116–57198. PMLR, 21–27 Jul 2024a.

Zonghao Ying, Aishan Liu, Tianyuan Zhang, Zhengmin Yu, Siyuan Liang, Xianglong Liu, and Dacheng Tao. Jailbreak vision language models via bi-modal adversarial prompt, 2024b.

Cong Zhang, Shuhui Wang, Xiaodan Li, Yao Zhu, Honggang Qi, and Qingming Huang. Enhancing the robustness of vision-language foundation models by alignment perturbation. *IEEE Transactions on Information Forensics and Security*, 20:7091–7105, 2025a. doi: 10.1109/TIFS.2025.3586430.

Le Zhang, Qian Yang, and Aishwarya Agrawal. Assessing and learning alignment of unimodal vision and language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14604–14614, June 2025b.

Xingchen Zhang, Ping Ye, Henry Leung, Ke Gong, and Gang Xiao. Object fusion tracking based on visible and infrared images: A comprehensive review. *Information Fusion*, 63:166–187, 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2020.05.002.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man (Man) Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 54111–54138. Curran Associates, Inc., 2023.

Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15345–15354, June 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16816–16825, June 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, jul 2022b. doi: 10.1007/s11263-022-01653-1.

A SELECTED COCO CATEGORIES

We evaluate zero-shot image classification on a subset of 30 categories from the COCO dataset, selected to cover diverse semantic groups such as animals, vehicles, and common objects. This subset balances visually distinct and semantically similar classes, providing a compact yet representative testbed for assessing robustness and cross-category generalization of vision–language models under adversarial perturbations. All evaluations are conducted using text prompts without task-specific fine-tuning. The full category list is provided in Table 5.

Table 5: COCO categories selected for zero-shot image classification.

ID	Category	ID	Category	ID	Category
0	airplane	1	banana	2	bear
3	bed	4	bird	5	boat
6	broccoli	7	bus	8	cake
9	cell phone	10	clock	11	cow
12	dog	13	donut	14	elephant
15	fire hydrant	16	horse	17	kite
18	motorcycle	19	pizza	20	sandwich
21	teddy bear	22	traffic light	23	stop sign
24	toilet	25	train	26	umbrella
27	vase	28	zebra	29	sheep

B LLM-AS-JUDGE PROMPTS FOR IMAGE CAPTIONING

We adopt an LLM-as-judge framework to evaluate caption consistency under adversarial perturbations. Given an image and its generated caption, the LLM is prompted to assess whether the caption faithfully reflects the visual content, focusing on key objects, attributes, and relationships rather than linguistic fluency. This provides a scalable and reproducible alternative to human evaluation while maintaining strong semantic alignment. The exact prompt template is shown in Figure 8.

Prompt Template for Consistency in Image Captioning Tasks

I will provide a predicted caption along with a standard caption for an image. Please evaluate the quality of the predicted caption according to the following criteria:

1. **Semantic Accuracy**: Does the caption describe the same object(s) and attributes as in the true captions? The score ranges from 0 to 10, where a higher score means greater semantic similarity and relevance.
2. **Tone Confidence**: Does the caption sound confident? A lower score should be given if the caption contains uncertainty or words like "maybe" or "might". The score ranges from 0 to 10.
3. **Consistency**: Is the predicted caption consistent with the reference captions in an overall sence? The score ranges from 0 to 10.

You just need to answer such as "Total Score: 22"
Predicted Caption: "{predicted_caption}"
Reference Captions: "{reference_captions}"

Figure 8: Prompt template for evaluating semantic consistency in image captioning.

C LLM-AS-JUDGE PROMPTS FOR VISUAL QUESTION ANSWERING

For visual question answering, we extend the LLM-as-judge paradigm to assess answer correctness given an image, a question, and a predicted answer. The LLM is instructed to verify consistency

Algorithm 1 CFPatch: Unified VIS–IR Adversarial Patch Optimization

Require: $(I_{\text{vis}}, I_{\text{ir}}), y, N, T, \mathcal{T}, \theta_g, (\theta^a_{\text{vis}}, \theta^a_{\text{ir}})$
Ensure: $(\theta^*_g, \theta^{a*}_{\text{vis}}, \theta^{a*}_{\text{ir}})$
Compute saliency regions and initialize patch pose
Initialize swarm $\{x_i^0, v_i^0, b_i^0\}_{i=1}^N$
for $t = 0$ to $T-1$ **do**
  for $i = 1$ to N **do**
    Construct shared geometry mask $M_g(\theta_g^i)$ and render patches $P_{\text{vis}}, P_{\text{ir}}$
    $I_{\text{vis}}^{\text{adv}} \leftarrow (1 - \alpha_{\text{vis}} M_g) \odot I_{\text{vis}} + \alpha_{\text{vis}} M_g \odot P_{\text{vis}}$
    $I_{\text{ir}}^{\text{adv}} \leftarrow (1 - \alpha_{\text{ir}} M_g) \odot I_{\text{ir}} + \alpha_{\text{ir}} M_g \odot P_{\text{ir}}$
    Sample $\tau \sim \mathcal{T}$
    $\ell_i \leftarrow \ell_{\text{joint}}(\tau(I_{\text{vis}}^{\text{adv}}), \tau(I_{\text{ir}}^{\text{adv}}))$
    $b_i^t \leftarrow \arg\min_{x \in \{b_i^t, x_i^t\}} \ell(x)$
  end for
  $b_\star^t \leftarrow \arg\min_i \ell(b_i^t)$
  for $i = 1$ to N **do**
    $v_i^{t+1} \leftarrow \omega v_i^t + c_1 \xi_1 (b_i^t - x_i^t) + c_2 \xi_2 (b_\star^t - x_i^t)$
    $x_i^{t+1} \leftarrow \Pi_B(x_i^t + v_i^{t+1})$
  end for
  if $\hat{y}_{\text{vis}} \neq y$ **and** $\hat{y}_{\text{ir}} \neq y$ **then**
    break
  end if
end for
return $b_\star^T$

between the answer and the visual evidence, taking into account both question intent and relevant image regions. This unified evaluation protocol enables consistent measurement of adversarial effects across multimodal reasoning tasks. The prompt template used for evaluation is illustrated in Figure 9.

Prompt Template for Correctness in Visual Question Answering Tasks

I will provide a predicted answer to a visual question answering (VQA) task, along with 10 human-provided reference answers with varying confidence levels (e.g., "yes", "maybe", "no").

Please evaluate if the predicted answer is correct based on the reference answers and their confidence levels. If the predicted answer aligns with most high-confidence reference answers ("yes" or multiple "maybe"), mark it as "1" (correct). If it deviates significantly from these high-confidence answers, mark it as "0" (incorrect).

Please provide the score including an explanation for each score. Return a single integer in the format "Score: (1 or 0)" based on your evaluation.
Predicted Answer: "{predicted_answer}"
Reference Answers: "{reference_answers_text}"

Figure 9: Prompt template for correctness evaluation in visual question answering.

D CFPatch Pseudocode

Algorithm 1 presents the unified optimization procedure of CFPatch. Given paired visible (VIS) and infrared (IR) images, the method jointly optimizes a shared geometric structure and modality-specific appearance parameters using particle swarm search under EOT. At each iteration, candidate patches are rendered, composited onto both modalities, and evaluated with a joint loss to enforce simultaneous misclassification.

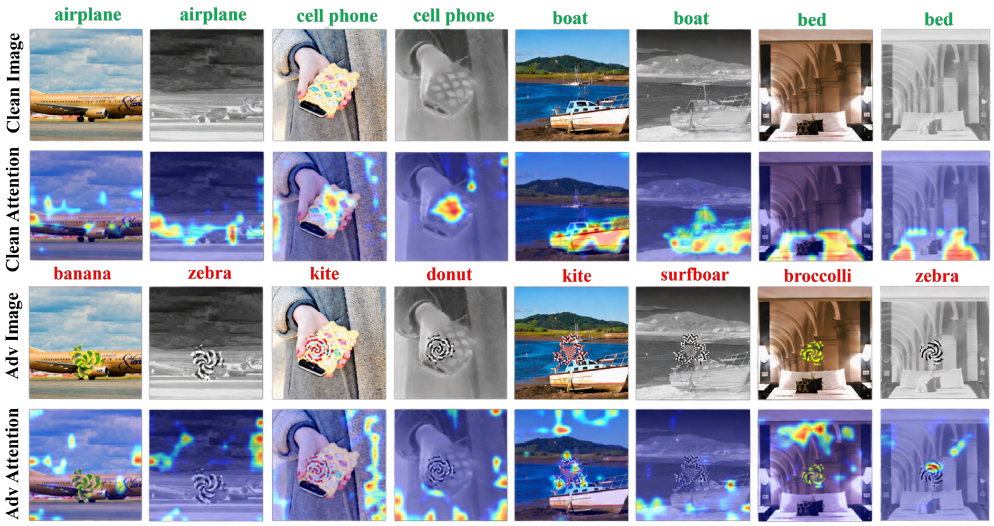

Figure 10: Attention responses before and after applying CFPatch on paired VIS–IR inputs, illustrating how the attack redistributes model attention from semantically relevant regions toward patch-induced geometric and texture cues.

E Attention Shift Analysis

To better understand how CFPatch changes visual reasoning, we further analyze the model attention distributions before and after applying the adversarial patch. Figure 10 compares clean and adversarial attention responses for both visible and infrared inputs, showing how attention is redistributed from semantically relevant object regions toward patch-induced geometric and texture cues. This representation-level analysis complements the quantitative attack results and provides additional insight into how the shared curved-fractal geometry and modality-specific Fraser rendering alter the evidence used by VIS–IR VLMs.

F Defense Evaluation

We further evaluate the robustness of CFPatch under a diverse set of input-level and patch-oriented defenses. Specifically, we consider JPEG compression, Gaussian blur, resize–pad preprocessing, LGS-style smoothing (Naseer et al., 2019), mask-guided inpainting, and PatchCleanser-style processing (Xiang et al., 2022). All defenses are applied under the same paired VIS–IR protocol used in the main paper, and attack success is measured using the strict joint criterion that requires both visible and infrared branches to be simultaneously fooled. As summarized in Table 6, all defenses reduce the attack success rate to different extents, with mask-guided inpainting and PatchCleanser-style processing yielding the strongest suppression. Nevertheless, CFPatch consistently retains substantially higher joint VIS–IR attack success than Unified Patch and TOUAP across all four backbones, indicating that the proposed shared geometry and modality-specific rendering remain comparatively robust under common defensive preprocessing.

G Evaluation on Real Paired VIS–IR Images

To evaluate whether CFPatch remains effective on genuinely captured dual-spectrum data, we further conduct experiments on 300 real paired visible–infrared pedestrian images from the LLVIP dataset (Jia et al., 2021b). Unlike the transformed infrared benchmark used in the main experiments, LLVIP provides naturally aligned visible and infrared observations of the same scenes. We formulate this evaluation as an untargeted pedestrian-classification attack: the source semantic is *person*, and only paired samples for which the clean VIS and IR branches are both classified as *person* are treated as valid. After applying CFPatch, a VIS (or IR) attack is successful when that branch is misclassified as any non-person category in the evaluation label set, while strict joint VIS–IR success requires both

Table 6: Defense evaluation on paired VIS–IR inputs. Values denote joint VIS–IR attack success rate (%). Higher values indicate stronger attack persistence under defense.

Defense	Method	OpenAI CLIP ViT-L/14	OpenCLIP ViT-B/16	Meta-CLIP ViT-L/14	EVA-CLIP ViT-G/14
No defense	Unified Patch	14.10	17.72	8.06	12.81
	TOUAP	25.64	36.10	25.27	27.09
	CFPatch	**67.82**	**77.22**	**68.28**	**73.89**
JPEG-70	Unified Patch	10.23	13.41	5.88	9.84
	TOUAP	20.94	30.68	21.32	22.61
	CFPatch	**61.43**	**71.86**	**62.48**	**65.19**
Gaussian blur	Unified Patch	7.34	10.09	4.82	6.63
	TOUAP	16.38	23.82	15.14	18.29
	CFPatch	**50.81**	**59.58**	**49.93**	**54.06**
Resize–pad	Unified Patch	8.91	11.64	5.18	8.13
	TOUAP	18.74	26.36	17.83	19.58
	CFPatch	**55.86**	**63.52**	**57.38**	**60.31**
LGS-style smoothing	Unified Patch	6.12	8.86	3.64	6.39
	TOUAP	13.53	20.58	14.16	15.34
	CFPatch	**50.59**	**53.07**	**46.21**	**48.20**
Mask-guided inpainting	Unified Patch	4.57	5.82	2.14	3.49
	TOUAP	10.36	14.18	8.84	10.62
	CFPatch	**30.76**	**38.92**	**28.43**	**32.64**
PatchCleanser-style	Unified Patch	2.83	3.61	1.32	2.48
	TOUAP	6.86	9.42	5.63	6.21
	CFPatch	**11.39**	**27.15**	**18.61**	**23.08**



Figure 11: Evaluation of CFPatch on 300 real paired visible–infrared samples from LLVIP across four CLIP backbones. Averaged over the four models, the VIS ASR is 84.71%, the IR ASR is 61.18%, and the strict joint VIS–IR ASR is 52.94%.

branches of the same pair to be changed away from *person*. We retain the same shared-geometry attack protocol and evaluate all four CLIP backbones used in the main paper.

Averaged over the four CLIP models, CFPatch achieves a VIS attack success rate (ASR) of 84.71%, an IR ASR of 61.18%, and a strict joint VIS–IR ASR of 52.94%. Because the joint metric requires simultaneous non-person predictions for both branches of the same clean-valid pedestrian pair, it is substantially stricter than evaluating either modality independently. Despite this criterion and the use of real paired sensor data, CFPatch maintains a joint ASR above 50%, indicating that the shared geometry and modality-specific appearance rendering transfer beyond the synthetic VIS-to-IR construction used in the primary benchmark.

Figure 11 summarizes the LLVIP evaluation results across the four CLIP backbones and reports the corresponding average VIS, IR, and joint attack success rates.

H Compute Resources

All experiments were conducted on a single NVIDIA A100 GPU with 40 GB VRAM and 83.5 GB system RAM. Attack optimization and model inference were executed on the GPU, while auxiliary data processing and result aggregation were performed on the CPU. Under the default attack configuration, the shared dual-modality attack was evaluated on 300 paired VIS–IR samples, with a subset of valid samples included in the final evaluation after filtering. Since different vision encoders exhibit varying computational costs, the runtime differed across models, including OpenAI CLIP ViT-L/14, OpenCLIP ViT-B/16, Meta-CLIP ViT-L/14, and EVA-CLIP ViT-G/14. For a single model, the complete evaluation pipeline typically required approximately 11–12 GPU-hours.

I Infrared-VLM Adaptation Details

All infrared victim models are adapted from their corresponding natural-image-pretrained checkpoints before adversarial evaluation. We use a shared lightweight LoRA adaptation protocol across model families and keep the adapted checkpoints frozen throughout attack optimization. For the contrastive CLIP-family models (OpenAI CLIP, OpenCLIP, Meta-CLIP, and EVA-CLIP), infrared adaptation is performed with an InfoNCE objective so that infrared image representations remain aligned with the corresponding text representations. For the generative and instruction-following models (BLIP-2, InstructBLIP, OpenFlamingo, LLaVA-1.5, and LLaVA-1.6), adaptation uses next-token prediction on infrared-style image–text pairs. Each model retains its native image processor, tokenizer, and task interface after adaptation.

This protocol follows the same IR-VLM construction used in our evaluation setup: natural-image-pretrained weights are first adapted to infrared-style data with LoRA, after which no victim-model or text-encoder parameters are updated by CFPatch. Consequently, the reported attack results measure the vulnerability of fixed infrared-adapted checkpoints rather than robustness changes induced during attack-time training. The same adapted checkpoints are used consistently for clean evaluation, adversarial optimization, baseline comparison, and downstream captioning/VQA transfer experiments.