REVEALING RAINY-DAY VULNERABILITIES IN VISION-LANGUAGE MODELS: A TWO-STAGE SEMANTIC DECOUPLING APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–Language Models (VLMs) are trained on image–text pairs collected under canonical visual conditions and achieve strong performance on multimodal tasks. However, their robustness to real–world weather conditions, and the stability of cross–modal semantic alignment under such structured perturbations, remain insufficiently studied. In this paper, we focus on rainy scenarios and introduce the first adversarial framework that exploits realistic weather to attack VLMs, using a two-stage, parameterized perturbation model based on semantic decoupling to analyze rain-induced shifts in decision-making. In Stage 1, we model the global effects of rainfall by applying a low-dimensional global modulation to condition the embedding space and gradually weaken the original semantic decision boundaries. In Stage 2, we introduce structured rain variations by explicitly modeling multi-scale raindrop appearance and rainfall-induced illumination changes, and optimize the resulting non-differentiable weather space to induce stable semantic shifts. Operating in a non-pixel parameter space, our framework generates perturbations that are both physically grounded and interpretable. Experiments across multiple tasks show that even physically plausible, highly constrained weather perturbations can induce substantial semantic misalignment in mainstream VLMs, posing potential safety and reliability risks in real-world deployment. Ablations further confirm that illumination modeling and multi-scale raindrop structures are key drivers of these semantic shifts.

1 INTRODUCTION

Vision–Language Models (VLMs) achieve strong performance by aligning visual and linguistic representations through large-scale joint pre-training (Radford et al., 2021; Jia et al., 2021; Li et al., 2021; 2022b). By grounding images in high-level semantic concepts, VLMs enable unified cross-modal reasoning and consistently surpass vision-only models in image classification, captioning, and visual question answering (VQA) (Alayrac et al., 2022; Li et al., 2023; Chen et al., 2024; Liu et al., 2024; 2023). Unlike conventional CNNs relying on local pixel statistics (Radosavovic et al., 2020; Tan & Le, 2021; Liu et al., 2022), VLMs interpret images through semantic abstraction, enhancing generalization across tasks and environments.

However, this semantic-centered design introduces a unique vulnerability. Semantic representations in VLMs depend on the global coherence of visual cues such as shape, color, brightness, texture, and spatial layout (Li et al., 2022a; Zhang et al., 2025; Covert et al., 2025). When these cues are systematically disturbed by changes in appearance or illumination, image–text alignment can deteriorate even if objects remain visually recognizable. Unlike pixel-level noise (Usama et al., 2025; Zhang et al., 2025), such structured perturbations directly distort high-level semantic representations and impair VLM reasoning. Real-world weather conditions such as rain, fog, and snow naturally cause correlated disruptions through interacting physical factors. As shown in Figure 1(a), in image captioning, weather-induced perturbations jointly distort visual appearance and contextual cues, causing substantial misalignment between images and text by corrupting the semantic grounding of the primary subject.

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.

Despite the ubiquity of such conditions in real-world environments, existing robustness evaluations of VLMs mainly target synthetic noise or simple, isolated corruptions, such as uniform brightness shifts or Gaussian noise (Chen et al., 2023; Qiu et al., 2025; Zhao et al., 2023; Liu et al., 2025). These settings fail to reflect the compound nature of real weather, where multiple physical factors jointly shape semantic perception. Consequently, current benchmarks likely underestimate the vulnerability of VLMs in complex environments (Ying et al., 2024; Carlini et al., 2019).

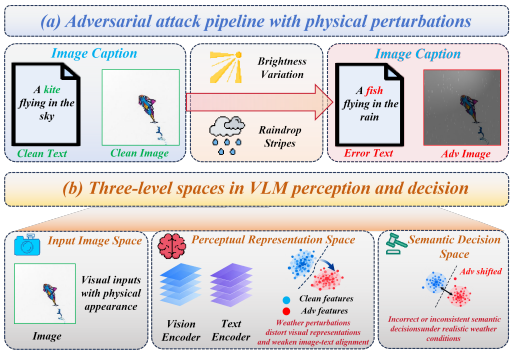


Figure 1: (a) Raindrop streaks and brightness variations disrupt correlated appearance cues, leading to a misalignment between vision and language in image captioning. (b) Vision–language reasoning is decoupled into input, representation, and semantic decision spaces.

Motivated by this gap, we study the semantic robustness of VLMs under realistic rain from an adversarial perspective. Instead of passive testing, we employ adversarial probing to reveal how complex weather disrupts image–text alignment beyond synthetic or pixel-level perturbations. Directly optimizing physical weather parameters is difficult due to the strong coupling between visual and linguistic representations in VLMs, which often yields shallow appearance changes without meaningful semantic shifts (Zhao et al., 2023; Usama et al., 2025). To overcome this, we propose a two-stage Semantic Decoupling framework that separates semantic manipulation from physical appearance synthesis, as illustrated in Figure 1(b). The framework decomposes vision–language reasoning into the input image, perceptual representation, and semantic decision spaces, exposing vulnerabilities at the perceptual level before realizing them through physically grounded weather synthesis.

Specifically, our framework consists of two stages. In Stage 1, we modulate the image–text embedding space to weaken the semantic alignment learned by VLMs. This design is motivated by the observation that realistic rainy conditions primarily affect high-level semantics rather than isolated pixels. By relaxing semantic boundaries in advance, subsequent weather-induced effects can more effectively induce semantic misalignment. In Stage 2, we optimize a parameterized rainy-weather space that jointly models multi-scale raindrops and spatial illumination. Using CMA-ES (Nomura & Shibata, 2024), we generate physically plausible rainy patterns that exploit the loosened semantic structure to maximize semantic misalignment. Unlike approaches that accumulate pixel-level noise, our framework provides an interpretable and physically grounded perspective on how different weather components contribute to semantic degradation in VLMs. Compared to prior work, our method produces realistic perturbations, improves optimization stability through semantic decoupling, and demonstrates strong transferability across tasks and model architectures. Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to systematically investigate the robustness of VLMs under rainy conditions, with an emphasis on semantic degradation rather than mere pixel-level distortions.
- We propose a two-stage, semantics-decoupled framework that closely aligns with the semantic reasoning mechanisms of VLMs and the compound nature of environmental perturbations induced by weather effects.
- We comprehensively evaluate our approach on zero-shot image classification, image captioning, and VQA, providing detailed analyses of semantic shifts and cross-task transferability.

2 Related Work

Text-side Manipulation and Sensitivity Analysis. Text-side variations provide an alternative pathway for influencing vision–language model inference by manipulating linguistic inputs without modifying visual content. Changes in descriptions, prompt templates, or phrasing can substantially affect model predictions, as evidenced by the strong prompt sensitivity observed in CLIP-style zero-shot classification (Radford et al., 2021). Subsequent prompt-learning approaches introduce learnable textual prototypes or contextual vectors, further emphasizing the dominant role of language

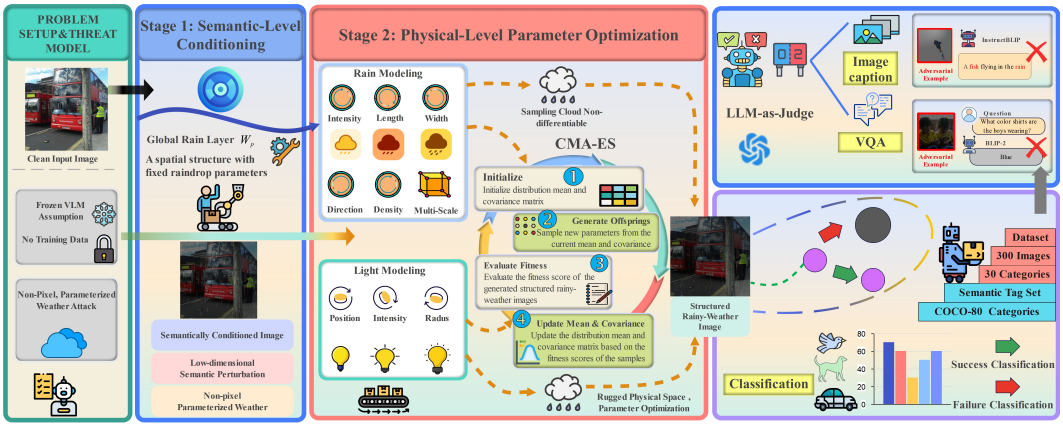


Figure 2: Framework of our method. Stage 1 conditions the embedding space with a global rain layer; Stage 2 uses CMA-ES to optimize parameterized raindrops and illumination to generate structured adversarial rainy images, evaluated on classification and transferred to captioning and VQA.

in shaping cross-modal alignment (Zhou et al., 2022; Khattak et al., 2023). Beyond performance effects, recent studies demonstrate that adversarially crafted prompts can steer model behavior or induce unintended and potentially unsafe outputs without altering the visual input (Jiang et al., 2024; Shayegani et al., 2024).

Visually Grounded Adversarial Attacks without Physical Constraints. Small-magnitude pixel-space perturbations are widely used to attack VLMs by disrupting cross-modal alignment. By transferring CNN-based adversarial techniques to CLIP-style architectures, prior studies show that visual-only perturbations can substantially degrade cross-modal matching and downstream performance (Zhao et al., 2023; Schlarmann et al., 2024). Recent analyses (Zhang et al., 2024) further reveal that such perturbations distort joint embedding geometry and erode decision boundaries, exposing representation-level instability as a key source of adversarial vulnerability in VLMs. Related robustness issues have also been observed in vision foundation models such as SAM under adversarial perturbations (Long et al., 2025).

Physically Consistent Environmental Perturbations with Semantic-level Analysis. To improve real-world relevance, recent adversarial evaluations (Liu et al., 2025; Ruan et al., 2025) increasingly enforce physical consistency by perturbing interpretable environmental factors rather than unconstrained pixels. ITA (Liu et al., 2025) parameterizes illumination and uses black-box optimization to find adversarial relighting that disrupts cross-modal alignment while preserving perceptual naturalness. Similarly, AdvDreamer (Ruan et al., 2025) studies physically grounded adversarial 3D variations, showing that realistic environmental changes can reliably induce semantic failures in VLMs.

3 Approach

3.1 Framework Overview

Figure 2 presents a two-stage, semantics-decoupled rainy-weather attack. Stage 1 conditions the cross-modal embedding space via low-dimensional rain-layer mixing, while Stage 2 employs CMA-ES to optimize multi-scale raindrop and illumination parameters for stable semantic shifts. The generated images are evaluated on zero-shot classification and transferred to captioning and VQA using an LLM-as-Judge (Bavaresco et al., 2025).

3.2 Model Setup and Problem Definition

This section explores the problem of robust adversarial attacks on VLMs under real-world physical consistency constraints. The models considered consist of an image encoder $f_v(\cdot)$ and a text encoder $f_t(\cdot)$, with semantic discrimination achieved through cross-modal similarity.

In the evaluation, we primarily use a text collection consisting of COCO-80 semantic labels for zero-shot image classification, while reusing the same visual embedding representation for image captioning and visual question answering tasks. Since the attack directly targets the embedding space output by the image encoder, the method itself is independent of the specific task format. Given an input image I, the similarity between the image and the text prompt T_k corresponding to the semantic label K is defined as:

$$S_k(I) = \langle f_v(I), f_t(T_k) \rangle \tag{1}$$

The model's prediction is determined by the semantic label with the highest similarity.

3.3 Physical Perturbation Modeling

Parametric Modeling of Raindrop Patterns. In contrast to pixel-level perturbations, we model raindrops using explicit physical parameters, yielding physically interpretable perturbations. Specifically, raindrop generation is governed by six optimizable variables: intensity α_r, controlling raindrop brightness contribution; density ρ_r, determining the number of raindrops per unit area; length l_r and width b_r, capturing motion-induced streaking effects; direction φ_r, simulating wind-driven rainfall; and blur kernel size, approximating scale-dependent blur caused by depth and focus variations. These parameters are collectively represented as a raindrop control vector, defined as:

$$\theta_r = (\alpha_r, \rho_r, l_r, b_r, \varphi_r, k_r) \tag{2}$$

To capture heterogeneous raindrop scales observed in real rainy conditions, we introduce a multi-scale superposition mechanism:

$$R(\theta_r) = \sum_{s \in \mathcal{S}} R_s(\theta_r) \tag{3}$$

Here, $R(\theta_r)$ denotes the overall rain perturbation layer, constructed by linearly superposing scale-specific raindrop layers $R_s(\cdot)$ over a set of scales S. This multi-scale design enables the perturbation to simultaneously influence fine-grained local textures and mid-scale structures, thereby more effectively modulating visual feature extraction.

Illumination Modulation Model. Considering that real rainfall is often accompanied by localized or non-uniform illumination variations, we introduce an illumination field composed of multiple local light sources:

$$L(x, y) = \sum_{i=1}^{N} \alpha_i \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2r_i^2} \right) \tag{4}$$

Here, (x_i, y_i), α_i, and r_i denote the position, intensity, and range of the i-th light source, respectively.

The illumination perturbation is applied multiplicatively to image brightness, where $\odot$ denotes element-wise multiplication across spatial locations and color channels. Specifically, the illumination gain map is defined as:

$$G = w_\ell L + (1 - w_\ell) \tag{5}$$

where L denotes the illumination map and $w_\ell \in [0, 1]$ controls the modulation strength. The modulated image is then obtained by:

$$I_{out} = I_{in} \odot G \tag{6}$$

This illumination modulation is physically grounded in real imaging pipelines and is naturally compatible with the raindrop perturbation. To characterize global and local illumination variations, we further derive the spatial average $\overline{G}$ and the maximum value $G_{\max}$ of the gain map G, together with a reference level G_0 and an upper-bound threshold G_{thr}. Deviations of $\overline{G}$ from G_0 are constrained by a global illumination regularizer weighted by $\lambda^{(2)}_{light}$, while excessive local amplification beyond G_{thr} is penalized by a range regularizer weighted by $\lambda^{(2)}_{range}$.

3.4 Stage 1: Global Raindrop Layer Mixing

In the first stage, we fix the spatial structure of the raindrops and control the perturbation intensity by adjusting the mixing weight between the perturbation and the original image. This allows us to

rapidly compress the semantic discrimination margin within a low-dimensional parameter space. The adversarial samples in Stage 1 are generated through linear mixing as follows:

$$I^{(1)} = (1 - w_p)I + w_p R_{fix} \tag{7}$$

Here, w_p is the optimizable global raindrop mixing weight that controls raindrop visibility. R_{fix} denotes the fixed global rain layer used in Stage 1.

For the synthesized image $I^{(1)}$, we define the attack objective as the similarity difference between the real semantic and the most competitive non-real semantic:

$$\mathcal{L}_{atk}^{(1)} = S_y\big(I^{(1)}\big) - \max_{k \neq y} S_k\big(I^{(1)}\big) \tag{8}$$

Minimizing $\mathcal{L}_{atk}^{(1)}$ progressively reduces the margin between the similarity to the ground-truth text T_y and the strongest competing text T_k in the joint embedding space, thereby eroding the discriminative advantage of the true semantics and moving the sample toward the semantic decision boundary.

To prevent excessive mixing weights from causing significant structural damage, Stage 1 explicitly introduces VGG perceptual consistency (Simonyan & Zisserman, 2015), where $\phi_\ell(\cdot)$ denotes the feature response of the pretrained VGG network at the ℓ-th layer.

$$\mathcal{L}_{perc}^{(1)} = \sum_{\ell \in \mathcal{L}} \Big\| \phi_\ell(I^{(1)}) - \phi_\ell(I) \Big\|_2^2 \tag{9}$$

Additionally, quadratic regularization is applied to the mixing weights to ensure the stability of the optimization process:

$$\mathcal{L}_{reg} = (w_p - w_p^0)^2 \tag{10}$$

The overall objective for Stage 1 is as follows:

$$\mathcal{L}_{stage1} = \mathcal{L}_{atk}^{(1)} + \lambda_{perc}^{(1)} \mathcal{L}_{perc}^{(1)} + \lambda_{reg}^{(1)} \mathcal{L}_{reg} \tag{11}$$

Here, $\lambda_{perc}^{(1)}$ and $\lambda_{reg}^{(1)}$ control the strengths of the perceptual and regularization terms, respectively.

3.5 Stage 2: Joint Weather Optimization

Building upon the fact that Stage 1 has pushed the sample closer to the semantic decision boundary, Stage 2 further introduces multi-scale raindrop and illumination perturbations. It uses a gradient-free optimization algorithm to optimize for more expressive perturbation combinations in the high-dimensional physical parameter space. The final synthesized image $I^{(2)}$ is expressed as:

$$I^{(2)} = I_r^{(2)} \odot G \tag{12}$$

Here, $I_r^{(2)}$ denotes the intermediate rain-perturbed image obtained after Stage 2 raindrop modeling.

To achieve stable and controllable semantic flipping, we introduce an attack margin constraint in Stage 2, requiring the erroneous semantic to exceed the real semantic by at least a given margin:

$$\mathcal{L}_{atk}^{(2)} = \max\big(0,\, \delta - \big(\max_{k \neq y} S_k(I^{(2)}) - S_y(I^{(2)})\big)\big) \tag{13}$$

In terms of perceptual consistency, Stage 2 still explicitly incorporates the VGG perceptual constraint, but adopts a Top-K selective constraint strategy to improve optimization efficiency. Specifically, in each iteration of the gradient-free optimization, candidate solutions are first ranked based on the attack objective and low-level real perceptual constraints. The VGG perceptual loss (Simonyan & Zisserman, 2015) is then computed only for the Top-K candidates with the highest attack potential:

$$\mathcal{L}_{perc}^{(2)} = \frac{1}{K} \sum_{j \in \mathrm{Top}-K} \sum_{\ell \in \mathcal{L}} \Big\| \phi_\ell(I^{(2,j)}) - \phi_\ell(I) \Big\|_2^2 \tag{14}$$

Weighted by $\lambda_{perc}^{(2)}$, this design preserves visual naturalness without prematurely constraining the high-dimensional parameter space.

To further enforce visual realism at the image-structure level, we introduce a regularization term based on structural similarity (SSIM) (Wang et al., 2004), defined as:

$$\mathcal{L}_{SSIM} = 1 - \frac{(2\mu_{I^{(2)}}\mu_I + C_1)(2\sigma_{I^{(2)}I} + C_2)}{(\mu_{I^{(2)}}^2 + \mu_I^2 + C_1)(\sigma_{I^{(2)}}^2 + \sigma_I^2 + C_2)} \tag{15}$$

where μ, σ^2, and $\sigma_{I^{(2)}I}$ denote local means, variances, and covariance, respectively, and C_1, C_2 are constants. This term penalizes structural deviations and complements the feature-level perceptual constraint, with strength governed by $\lambda_{real}^{(2)}$.

The overall optimization objective for Stage 2 is as follows:

$$\mathcal{L}_{stage2} = \mathcal{L}_{atk}^{(2)} + \lambda_{perc}^{(2)}\mathcal{L}_{perc}^{(2)} + \lambda_{real}^{(2)}\mathcal{L}_{SSIM} + \lambda_{light}^{(2)}(\bar{G} - G_0)^2 + \lambda_{range}^{(2)}\max(0, G_{max} - G_{thr}) \tag{16}$$

Owing to the high-dimensional, non-convex raindrop and illumination parameter space with unstable gradients, we adopt CMA-ES for joint black-box optimization. This model-agnostic method is detailed in Appendix A.

4 Experiments

4.1 Experimental Settings

Datasets. We follow the evaluation protocol of ITA (Liu et al., 2025) and use the same test set of 300 MS COCO images (Chen et al., 2015) from 30 semantic categories. These categories serve as labels for image classification, with category descriptions constructed using standard prompt templates (Zhou et al., 2022; Khattak et al., 2023). For image captioning and VQA, we adopt identical visual and textual inputs as ITA to ensure fair comparisons. Detailed dataset statistics and category definitions are provided in the Appendix B.1.

Threat Models. To ensure a fair and direct comparison with ITA, we adopt the same experimental protocol across all tasks. For zero-shot image classification, we evaluate multiple representative CLIP variants, including OpenCLIP ViT-B/16 (Cherti et al., 2023; Ilharco et al., 2021), 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 image captioning and VQA, we consider several widely used vision–language models: LLaVA-1.5 (Liu et al., 2023), LLaVA-1.6 (Liu et al., 2024), OpenFlamingo (Awadalla et al., 2023), BLIP-2 ViT-L and BLIP-2 FlanT5-XL (Li et al., 2023), and InstructBLIP (Dai et al., 2023).

Baseline Methods. We compare our approach against state-of-the-art physically consistent environmental attacks, including: Illumination Transformation Attack (ITA) (Liu et al., 2025), Shadows Attack (Zhong et al., 2022), and Natural Light Attack (Hsiao et al., 2024).

Implementation Details. In Stage 1, rain structure and illumination are fixed, and perturbations are generated via multi-scale rain layer superposition. Only the rain–image mixing weight w_p is optimized within $[0.02, 0.7]$, with a VGG-based perceptual loss imposed as regularization. The resulting w_p is then fixed for the subsequent stage. In Stage 2, CMA-ES is employed to jointly optimize the physical parameters of rainy weather, including rain intensity, geometric attributes, orientation, blur, and density, yielding structurally diverse rain patterns. A continuous illumination model based on two-dimensional Gaussian light sources is incorporated, with illumination parameters optimized while the mixing weight w_ℓ is fixed to 0.5.

4.2 Zero-shot Classification Evaluation

For zero-shot image classification, we report the Top-1 accuracy drop relative to clean inputs, following ITA for a fair comparison. As shown in Table 1, our approach causes greater performance degradation than ITA on OpenCLIP ViT-B/16, Meta-CLIP ViT-L/14, and EVA-CLIP ViT-G/14, indicating that the joint modeling of multi-scale raindrops and illumination more effectively disrupts visual semantic representations. On OpenAI CLIP ViT-L/14, ITA slightly outperforms our method, likely due to its comparatively higher robustness against local structural perturbations. Qualitative examples of the attacks are illustrated in Figure 3, and Appendix B.2 provides a detailed analysis of misclassification distributions across object categories.

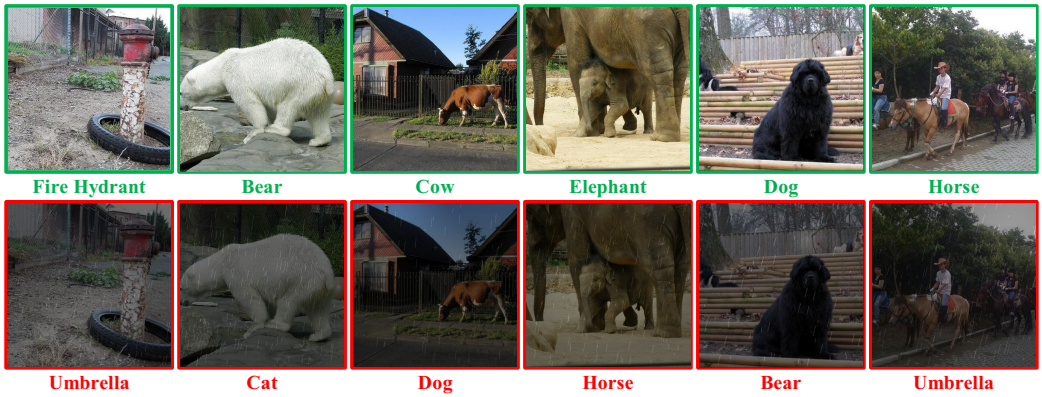


Figure 3: Clean and adversarial samples under weather perturbations. The top row shows clean images with correct labels, while the bottom row shows weather-corrupted adversarial samples with misclassifications.

Table 1: Classification accuracy (%) of different CLIP models under various attack methods.

Methods	OpenCLIP ViT-B/16	Meta-CLIP ViT-L/14	EVA-CLIP ViT-G/14	OpenAI CLIP ViT-L/14
Clean	97	98	98	93
Natural Light Attack	94 (↓ 3)	97 (↓ 1)	97 (↓ 1)	93 (–)
Shadow Attack	84 (↓ 13)	82 (↓ 16)	95 (↓ 3)	79 (↓ 14)
ITA	46 (↓ 51)	64 (↓ 34)	84 (↓ 14)	51 (↓ **42**)
Ours	35 (↓ **62**)	42 (↓ **56**)	58 (↓ **40**)	54 (↓ 39)

4.3 Image Captioning Evaluation

Our method is highly effective on the image captioning task, even though adversarial samples are directly reused from zero-shot image classification without task-specific optimization. We employ GPT-4o as an LLM-as-Judge to assess caption semantic consistency, following prior work, with prompt templates provided in Appendix C. As shown in Table 2 and Figure 4(a), our approach induces significantly larger performance drops than ITA under the cross-task transfer setting, demonstrating stronger generalization across both smaller (e.g., OpenFlamingo, BLIP-2) and larger models (e.g., LLaVA-1.5/1.6). These results indicate that multi-scale raindrops and non-uniform illumination cause stable high-level semantic shifts, leading to systematic multimodal misalignment.

4.4 Visual Question Answering Evaluation

For VQA, we evaluate our method under weather perturbations using the same cross-task transfer setting, where adversarial examples from zero-shot image classification are reused without additional optimization. We employ GPT-4o as an LLM-as-Judge to assess answer correctness and reasoning consistency (see Appendix D). As shown in Table 3 and Figure 4(b), our method induces a significantly larger accuracy drop than ITA. Unlike image captioning, errors primarily arise from incorrect multimodal reasoning, and this effect persists even in large models such as LLaVA-1.5/1.6, revealing fundamental robustness limitations.

4.5 Ablation Study

Ablation on Stage 1 Hyperparameters. We conduct an ablation study on two key hyperparameters in Stage 1: the scaling factor of the regularization on the global raindrop mixing weight $\lambda_{reg}^{(1)}$ and the scaling factor of the perceptual constraint $\lambda_{perc}^{(1)}$, to evaluate their influence on attack performance and subsequent optimization. The experimental results are presented in Figure 5. For Meta-CLIP ViT-L/14, when $\lambda_{reg}^{(1)}$ is large (e.g., 10^0 or 10^{-1}) and $\lambda_{perc}^{(1)}$ is small (e.g., 10^{-3} or 10^{-2}), Stage 1 achieves the highest attack success rate of up to 12.7%. This indicates that moderate rain intensity constraints combined with relaxed perceptual consistency enable faster erosion of decision boundaries.

Table 2: Comparison of attack results (%) on image captioning tasks.

Image Encoder	Models	Params	Clean	Natural Light	Shadow	ITA	**Ours**
OpenAI CLIP ViT-L/14	LLaVA-1.5	7B	78.60	77.00 (↓ 1.63)	74.60 (↓ 4.00)	63.73 (↓ 14.87)	**56.14** (↓ 22.46)
	LLaVA-1.6	7B	72.10	71.70 (↓ 0.39)	71.17 (↓ 0.93)	61.60 (↓ 10.51)	**46.58** (↓ 25.52)
	OpenFlamingo	3B	70.20	69.53 (↓ 0.67)	67.80 (↓ 2.40)	53.93 (↓ 16.27)	**46.10** (↓ 24.10)
	BLIP-2 (FlanT5$_{XL}$ ViT-L)	3.4B	75.10	70.77 (↓ 4.33)	68.57 (↓ 6.53)	60.93 (↓ 14.17)	**51.77** (↓ 23.33)
EVA-CLIP ViT-G/14	BLIP-2 (FlanT5$_{XL}$)	4.1B	74.96	71.27 (↓ 3.69)	68.80 (↓ 6.17)	62.01 (↓ 11.88)	**45.33** (↓ 29.63)
	InstructBLIP (FlanT5$_{XL}$)	4.1B	76.50	72.07 (↓ 4.43)	69.77 (↓ 6.73)	63.20 (↓ 13.30)	**42.86** (↓ 33.64)

Table 3: Visual question answering task results (%).

Image Encoder	Models	Params	Clean	Natural Light	Shadow	ITA	**Ours**
OpenAI CLIP ViT-L/14	LLaVA-1.5	7B	68.00	68.00 (-)	67.00 (↓ 1.00)	48.00 (↓ 20.00)	**38.00** (↓ 30.00)
	LLaVA-1.6	7B	64.00	63.00 (↓ 1.00)	64.00 (-)	43.00 (↓ 21.00)	**37.00** (↓ 27.00)
	OpenFlamingo	3B	45.00	39.00 (↓ 6.00)	44.00 (↓ 1.00)	19.00 (↓ 26.00)	**14.00** (↓ 31.00)
	BLIP-2 (FlanT5$_{XL}$ ViT-L)	3.4B	63.00	58.00 (↓ 5.00)	50.00 (↓ 13.00)	38.00 (↓ 25.00)	**30.00** (↓ 33.00)
EVA-CLIP ViT-G/14	BLIP-2 (FlanT5$_{XL}$)	4.1B	54.00	53.00 (↓ 1.00)	54.00 (-)	33.00 (↓ 21.00)	**22.00** (↓ 32.00)
	InstructBLIP (FlanT5$_{XL}$)	4.1B	68.00	64.00 (↓ 4.00)	62.00 (↓ 6.00)	44.00 (↓ 24.00)	**38.00** (↓ 30.00)



Figure 4: (a) Examples of adversarial sample attacks on the image captioning task; (b) Examples of adversarial sample attacks on the VQA task.

In contrast, excessively large $\lambda_{perc}^{(1)}$ limits perturbation freedom, reducing attack success. Notably, the Stage 1 success rate does not fully determine the final attack outcome, as overall success can still range from 26% to 57.7% even when Stage 1 performance is low. These findings demonstrate that Stage 1 mainly provides favorable initialization for Stage 2, substantially improving the effectiveness of CMA-ES optimization. Furthermore, Stage 2 achieves the greatest performance gain when $\lambda_{reg}^{(1)}$ is large and $\lambda_{perc}^{(1)}$ is small, reaching 45.0%, suggesting that looser constraints in Stage 1 yield more exploratory initial solutions that unlock the optimization potential of Stage 2.

Ablation on Stage 2 Hyperparameters. We analyze four critical Stage-2 hyperparameters: the perceptual consistency weight $\lambda_{perc}^{(2)}$, realism regularization weight $\lambda_{real}^{(2)}$, range constraint weight $\lambda_{range}^{(2)}$, and illumination modulation weight $\lambda_{light}^{(2)}$. These parameters primarily enforce visual naturalness and maintain physical plausibility, rather than directly enhancing attack efficacy. As

Table 4: Performance across CLIP backbones (%).

Methods	OpenCLIP ViT-B/16	Meta-CLIP ViT-L/14	EVA-CLIP ViT-G/14	OpenAI CLIP ViT-L/14
w/o Multi-scale Raindrops	12	17	13	20
w/o Illumination Modulation	36	45	31	27
ITA	51	34	14	**42**
Ours	**62**	**56**	**40**	39

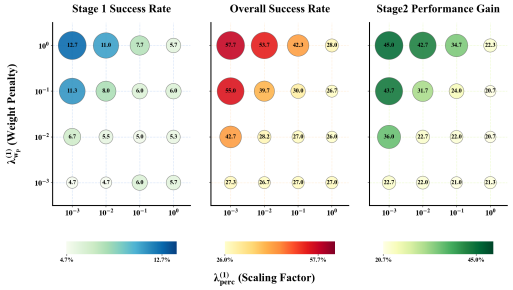


Figure 5: Stage 1 hyperparameter ablation results.

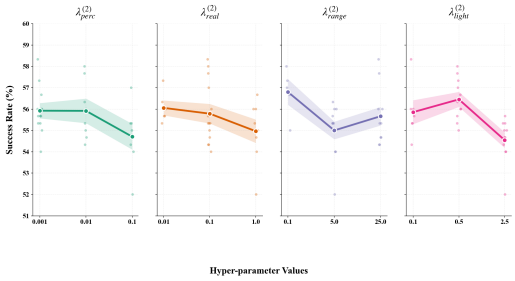


Figure 6: Stage 2 hyperparameter ablation results.

shown in Figure 6, attack success remains robust across a wide range of configurations, indicating that Stage 2's gradient-free optimization is largely hyperparameter-insensitive. Overall performance is predominantly dictated by the physical parameter space and Stage-1 semantic initialization, with strong constraints only moderately restricting the optimization space. Collectively, Stage 2 achieves a balanced trade-off between attack effectiveness and physical realism without requiring fine-tuning.

Component-level Ablation. Table 4 shows that removing either multi-scale raindrops or illumination modulation degrades performance, with raindrop removal causing the largest drop and illumination modulation mainly affecting larger models. This confirms their complementary roles in enabling stable semantic flipping.

Population Size Ablation. Figure 7 illustrates the effect of population size on attack success across several CLIP-based models. As the population size increases from 5 to 15, performance consistently improves, indicating more effective and stable optimization. However, further increases beyond 15 result in diminishing returns, with performance gains gradually plateauing and showing marginal improvement beyond that point.

5 Discussion and Conclusion

We propose a semantics-disentangled, two-stage adversarial framework for systematically evaluating VLM robustness under rainy conditions. By operating in a physically consistent, non-pixel parameter space, our method generates stable, interpretable, and task-transferable semantic perturbations, thereby exposing intrinsic vulnerabilities and latent limitations of VLMs under structured environmental perturbations.

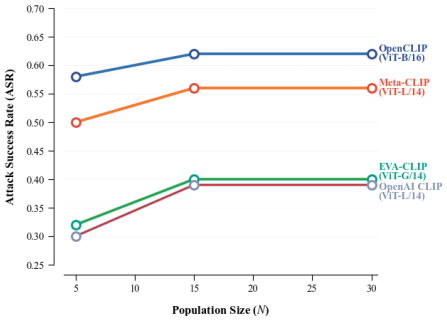


Figure 7: Impact of population size on optimization

Limitations and Future Work. Our study focuses on rainy conditions and ignores other factors such as fog, haze, or snow. Extending the framework to broader weather scenarios would enable a more comprehensive evaluation of VLM robustness. Additionally, analyzing finer-grained perturbations and their interactions could clarify how complex environmental conditions affect cross-modal semantic representations.

AI USE STATEMENT

Generative AI tools were used in two ways in this work. First, GPT-4o was used as an LLM-as-Judge to evaluate semantic consistency and answer correctness in the image captioning and visual question answering experiments, following the evaluation protocol described in the paper. Second, generative AI tools were used to assist with language polishing, LaTeX formatting, code development and debugging, and presentation refinement. All research hypotheses, methodological decisions, experimental protocols, results, and scientific conclusions were determined and validated by the authors. AI-assisted evaluations, code, and text were manually reviewed and checked for correctness. The authors take full responsibility for the final content of this work.

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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022.

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. doi: 10.18653/v1/2025.acl-short.20.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness, 2019.

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 *Advances in Neural Information Processing Systems*, volume 36, pp. 51758–51777, 2023. Datasets and Benchmarks Track.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 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 Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N. Fung, and Steven C. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267, 2023.

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.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021.

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 (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 18–24 Jul 2021.

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. ArtPrompt: ASCII art-based jailbreak attacks against aligned LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15157–15173, Bangkok, Thailand, aug 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.809.

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 2023.

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 *Advances in Neural Information Processing Systems*, volume 35, pp. 7290–7303, 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 M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. 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.

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 *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 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.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, June 2022.

Jiahuan Long, Zhengqin Xu, Tingsong Jiang, Wen Yao, Shuai Jia, Chao Ma, and Xiaoqian Chen. Robust sam: On the adversarial robustness of vision foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 5775–5783, 2025. doi: 10.1609/aaai.v39i6.32616.

Masahiro Nomura and Masashi Shibata. cmaes: A simple yet practical python library for CMA-ES, 2024.

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.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10428–10436, 2020.

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. IEEE, October 2025.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43685–43704. PMLR, 2024.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *International Conference on Learning Representations*, 2024.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

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

Mingxing Tan and Quoc Le. EfficientNetV2: Smaller models and faster training. 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. 10096–10106. PMLR, 18–24 Jul 2021.

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

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57116–57198. PMLR, 2024.

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 2025.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27456–27466, June 2024.

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 *Advances in Neural Information Processing Systems*, volume 36, pp. 54111–54138, 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. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. doi: 10.1007/s11263-022-01653-1.

A Algorithmic Details of the Two-Stage Weather Attack

Algorithm 1 provides the complete pseudo-code of our two-stage physically consistent attack. Stage 1 optimizes a low-dimensional global mixing weight to condition the cross-modal embedding space under a perceptual constraint, and Stage 2 applies CMA-ES to optimize the non-differentiable physical parameter space of multi-scale raindrops and illumination to obtain the final adversarial image.

B Image Classification Task

B.1 Selected COCO Categories

The 30 categories included in the dataset used for the zero-shot classification task are summarized in detail in Table 5.

Table 5: The selected 30 categories in the COCO dataset.

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.2 Specific Statistics on Attack Categories

The Figure 8 illustrates the specific distribution of classification errors across the four models tested in the image classification task.

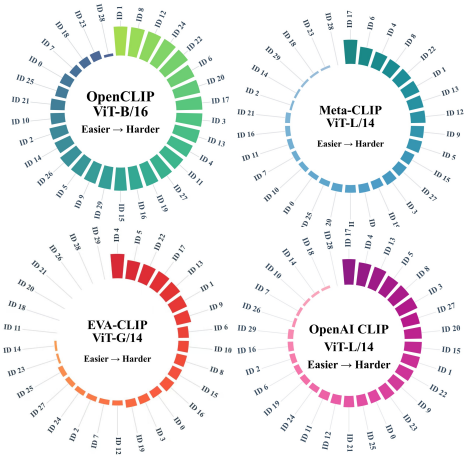


Figure 8: Specific Statistics on Attack Categories.

B.3 Larger-Scale Classification Evaluation

We include larger-scale classification results on 1,200 MS COCO images from 60 categories, with 20 images per category. The additional 30 categories are listed in Table 6. As shown in Table 7, the

Algorithm 1 Two-Stage Physically Consistent Adversarial Attack

Require: Original image I; visual encoder f_v; text encoder f_t; semantic prompts $\{T_k\}$; Stage-1 iterations T_1; Stage-2 iterations T_2
Ensure: Adversarial image I_{adv}

Stage 1: Global Semantic Perturbation
Initialize: global rain layer R (fixed spatial structure) and mixing coefficient $w_p \leftarrow 0$
for $t = 1$ to T_1 **do**
  Generate intermediate image $I_t \leftarrow (1 - w_p)I + w_p R$
  Extract visual embedding $z_v \leftarrow f_v(I_t)$
  Extract text embedding $z_t \leftarrow f_t(T_k)$
  Compute semantic margin loss $\mathcal{L}_{\mathrm{atk}}^{(1)}$
  Compute perceptual consistency loss $\mathcal{L}_{\mathrm{perc}}^{(1)}$
  Compute overall Stage-1 objective:
    $\mathcal{L}_{\mathrm{stage1}} = \mathcal{L}_{\mathrm{atk}}^{(1)} + \lambda_{\mathrm{perc}}^{(1)}\mathcal{L}_{\mathrm{perc}}^{(1)} + \lambda_{\mathrm{reg}}^{(1)}\mathcal{L}_{\mathrm{reg}}$
  Update w_p by minimizing $\mathcal{L}_{\mathrm{stage1}}$
end for
Obtain intermediate image I_{stage1}

Stage 2: Physically Consistent Optimization
Initialize: multi-scale rain parameters Θ_r, illumination parameters Θ_ℓ (gain map G), and CMA-ES optimize distribution
for $t = 1$ to T_2 **do**
  Sample candidate parameters $\{\Theta_r^{(i)}, \Theta_\ell^{(i)}\}$
  Generate physically consistent candidate images
  Evaluate semantic margin and realism constraints
  Rank candidates based on attack objective
  Select Top-K candidates
  Compute perceptual loss $\mathcal{L}_{\mathrm{perc}}^{(2)}$ for Top-K candidates
  Update CMA-ES distribution
end for
return I_{adv}

main trends remain consistent with those observed under the original evaluation protocol, further confirming stable and strong attack performance on the expanded evaluation set.

Table 6: The additional 30 categories selected from the remaining COCO-80 classes for the larger-scale 60-category evaluation.

ID	Category	ID	Category	ID	Category
30	person	31	bicycle	32	car
33	truck	34	bench	35	cat
36	backpack	37	handbag	38	tie
39	suitcase	40	frisbee	41	skateboard
42	surfboard	43	bottle	44	cup
45	bowl	46	apple	47	orange
48	carrot	49	chair	50	couch
51	potted plant	52	dining table	53	tv
54	laptop	55	keyboard	56	sink
57	refrigerator	58	book	59	scissors

Table 7: Classification accuracy (%) of different CLIP models under clean conditions and under our attack on the larger-scale evaluation set (1,200 images).

Model	Clean	Ours	Drop
OpenCLIP ViT-B/16	81	**21**	↓ 60
Meta-CLIP ViT-L/14	87	**29**	↓ 58
EVA-CLIP ViT-G/14	81	**39**	↓ 42
OpenAI CLIP ViT-L/14	77	**39**	↓ 38

C LLM-AS-JUDGE PROMPTS FOR IMAGE CAPTIONING

Figure 9 presents the prompt template used for LLM-based evaluation of image captioning consistency in our experiments. Given a predicted caption and the reference caption, the judge scores semantic accuracy, tone confidence, and overall consistency, and returns a standardized total score for reproducible comparison.

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 9: Prompt Template for Consistency in Image Captioning Tasks.

D LLM-AS-JUDGE PROMPTS FOR VISUAL QUESTION ANSWERING

Figure 10 shows the prompt template used for LLM-based correctness evaluation in VQA in our experiments. Given a predicted answer and the set of human reference answers, the judge outputs a binary score (1/0) indicating whether the prediction agrees with the high-confidence references under a standardized format.

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 10: Prompt Template for Correctness in Visual Question Answering Tasks.

E HUMAN PERCEPTUAL EVALUATION OF IMPERCEPTIBILITY

We further conduct a human perceptual evaluation on the final 300 adversarial examples selected after the zero-shot classification evaluations on the four CLIP backbones used in the main experiments. The examples are divided into five disjoint subsets of 60 images each, and 40 participants are split into five groups of eight, so that each example receives eight independent judgments, yielding 2,400 binary judgments in total. In each trial, the clean image and its adversarial counterpart are displayed side by side in randomized left–right order, and participants judge whether the adversarial image exhibits a visible perturbation relative to the clean reference. A response is counted as imperceptible when no obvious visible difference is reported. Table 8 summarizes the protocol, and Table 9 reports the aggregated results. The average imperceptibility rate is 87.8%, with all subset-level values falling in the 85%–90% range, indicating consistently high visual imperceptibility across the full set of evaluated adversarial examples.

Table 8: Protocol summary for the human perceptual evaluation.

Item	**Setting**
Example source	Selected from classification attacks on four CLIP backbones
Participants	40
Examples	300 adversarial examples
Partition	5 subsets $\times$ 60 images
Participants/subset	8
Judgments/image	8
Total judgments	2,400
Display	Clean and adversarial images, side by side
Order	Randomized left–right order
Task	Judge visible perturbation
Metric	Imperceptibility rate (%)

Table 9: Aggregated results of the human perceptual evaluation.

Subset	# Images	Imperceptibility Rate (%)
Subset 1	60	88.3
Subset 2	60	87.5
Subset 3	60	89.0
Subset 4	60	86.8
Subset 5	60	87.4
Overall	300	87.8

F Compute Resources

All experiments were conducted on a single server equipped with one NVIDIA A100 GPU (40 GB VRAM) and 83.5 GB system RAM. The GPU was used for attack optimization and model inference, while auxiliary data processing and result aggregation were performed on the CPU. Under the default attack configuration, the proposed rainy-weather attack was evaluated on 300 images spanning 30 categories, with a subset of valid samples retained after filtering for final evaluation. Since different vision encoders exhibit varying computational costs, the runtime differed across OpenAI CLIP ViT-L/14, OpenCLIP ViT-B/16, Meta-CLIP ViT-L/14, and EVA-CLIP ViT-G/14. In practice, the complete evaluation pipeline for a single model typically required approximately 7–8 GPU-hours.