WHEN WEATHER MISLEADS RETRIEVAL: ATMOSPHERIC ATTACKS ON REMOTE SENSING VISION-LANGUAGE RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal RAG systems increasingly rely on vision-language retrievers to ground visual queries in external textual evidence. Prior adversarial studies on RAG mainly manipulate the retrieval corpus or memory, while attacks on vision-language and remote sensing models typically target end-task predictions. Input-space adversarial threats to the evidence retrieval stage of remote sensing multimodal RAG remain largely unexplored. To fill this gap, we introduce CloudWeb, an atmospheric retrieval hijacking attack that modifies only the input image while leaving the retriever, generator, and knowledge base unchanged at deployment. CloudWeb overlays parameterized cloud- and haze-like patterns on remote sensing images and optimizes them with a retrieval-oriented objective that attracts adversarial image embeddings toward target atmospheric evidence, suppresses source-scene evidence, enforces rank separation, and regularizes naturalness and coverage. To the best of our knowledge, this is the first study of retrieval-stage atmospheric evidence hijacking for remote sensing multimodal RAG. We evaluate CloudWeb on a seven-dataset remote sensing RAG benchmark with five CLIP-style retrievers, including GeoRSCLIP, RemoteCLIP, OpenAI CLIP, and OpenCLIP, together with downstream vision-language generators. Across retrievers, optimized CloudWeb consistently outperforms clean retrieval, handcrafted atmospheric baselines, random cloud perturbations, and fixed variants in injecting weather-related evidence into top-ranked results. On GeoRSCLIP ViT-B/32, Weather@5 increases from 0.71% to 43.29%. Downstream generation results further show measurable weather hallucination and semantic shift, indicating that retrieval-stage hijacking can propagate to the final RAG response. These findings reveal a practical failure mode: natural-looking input-space atmospheric changes can systematically compromise evidence retrieval before generation begins.

1 INTRODUCTION

Multimodal retrieval-augmented generation (RAG) systems increasingly rely on vision-language retrievers to ground visual queries in external evidence. Rather than depending solely on parametric knowledge stored within the generator, a multimodal RAG pipeline first retrieves semantically relevant context from an external corpus and subsequently conditions the generation process on the retrieved evidence. This paradigm is particularly valuable for remote sensing applications, where aerial and satellite imagery often requires specialized domain knowledge that may not be sufficiently captured by general-purpose models, including scene descriptions, land-use annotations, disaster reports, temporal change captions, geographic metadata, and visual question answering records. By explicitly incorporating such external knowledge, multimodal RAG can improve the factuality, interpretability, and domain adaptability of downstream responses. However, the retrieval interface simultaneously introduces a critical and often overlooked evidence-grounding vulnerability. If a query image is adversarially shifted toward an incorrect region of the shared vision-language embedding space, the retriever may return semantically misleading or entirely irrelevant evidence before generation even begins. Consequently, the downstream generator can produce fluent and seemingly well-supported answers whose factual basis has already been corrupted at the retrieval stage, making such attacks particularly deceptive and difficult to detect in practical multimodal RAG systems.

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

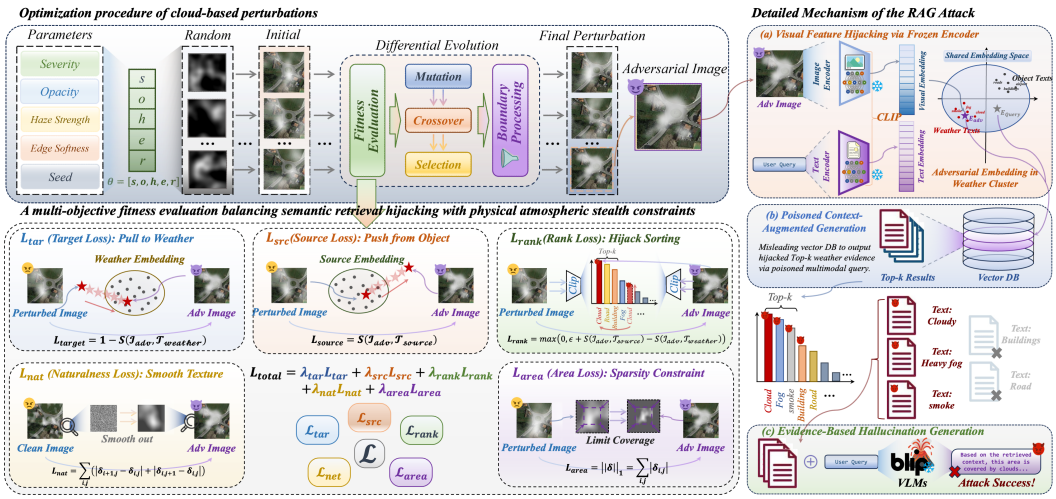


Figure 1: **Overview of CloudWeb.** CloudWeb optimizes cloud- and haze-like perturbations to shift a remote sensing image toward weather-related evidence in a frozen vision-language retrieval space. The hijacked top-k evidence is then passed to the downstream VLM, inducing evidence-grounded hallucinations without modifying the retriever, generator, vector database, or retrieval logic.

Existing adversarial studies on vision-language and remote sensing systems primarily target supervised perception tasks or manipulate the retrieval corpus. Atmospheric attacks have also been explored in remote sensing; for example, AdvCloud optimizes cloud-like perturbations to fool salient object detection models and studies corresponding pre-processing defenses (Sun et al., 2024). However, these studies do not address retrieval-grounded multimodal retrieval-augmented generation (RAG), where critical failures can occur much earlier. A manipulated query image may change the top-ranked evidence returned by a frozen vision-language retriever before generation begins, thereby corrupting the evidence on which the generator relies. This leaves a practical input-space threat unexplored, where the attacker modifies only the input image without modifying the retriever, generator, vector database, or retrieval logic at deployment.

To investigate whether visually plausible atmospheric changes to a remote sensing query image can systematically hijack this evidence retrieval stage, we propose CloudWeb. We focus on cloud, haze, fog, mist, and low-visibility patterns because they are ubiquitous in aerial and satellite imagery and appear far less suspicious than conventional adversarial noise or conspicuous patches. Simultaneously, these patterns carry strong semantic associations in vision-language embedding spaces, making them natural carriers for semantic redirection rather than merely visual degradation. CloudWeb overlays parameterized weather perturbations onto the input image and optimizes them strictly for retrieval. Rather than merely altering visual appearance, the attack directly targets the ranking process by pulling the adversarial query toward weather-related evidence while weakening its association with source-scene evidence. Figure 1 illustrates this retrieval-stage optimization and its effect on the retrieval-then-generation pipeline.

We evaluate CloudWeb on a seven-dataset remote sensing RAG benchmark covering scene classification, image captioning, change captioning, disaster imagery, and visual question answering. Across five CLIP-style retrievers, optimized CloudWeb consistently outperforms clean retrieval, hand-crafted baselines, and random perturbations in injecting weather evidence into top-ranked results; on GeoRSCLIP ViT-B/32, Weather@5 increases from 0.71% to 43.29%. Downstream evaluations further confirm that retrieval-stage hijacking propagates to generated responses, inducing measurable weather hallucination and semantic shift across multiple vision-language generators. Our contributions are summarized as follows:

- To the best of our knowledge, we propose the first retrieval-stage atmospheric evidence hijacking attack for remote sensing multimodal RAG. CloudWeb modifies only the input image with visually plausible cloud-, haze-, and low-visibility perturbations, and optimizes them to redirect top-ranked retrieved evidence toward target weather semantics without modifying the retriever, generator, or vector database at deployment.

- We conduct extensive experiments across seven datasets, five CLIP-style retrievers, and multiple downstream vision-language generators. The results show that CloudWeb consistently hijacks retrieval toward weather-related evidence and further propagates this evidence shift to downstream generation, inducing weather hallucination and semantic drift.
- We perform comprehensive ablation and robustness studies on loss terms, perturbation components, opacity-severity interactions, image post-processing, and mechanism-level attention behavior. These analyses reveal how atmospheric structure, retrieval-oriented optimization, and evidence propagation jointly determine the attack effectiveness of CloudWeb.

2 Related Work

2.1 Retrieval-Augmented Vision-Language Systems and Remote Sensing VLMs

Retrieval-augmented generation (RAG) improves generation by retrieving external evidence before producing an answer. Initially developed for text-based knowledge-intensive tasks to reduce reliance on parametric memory (Lewis et al., 2020; Guu et al., 2020; Karpukhin et al., 2020; Izacard & Grave, 2021; Izacard et al., 2023), this retrieval-then-generation paradigm has expanded to vision-language tasks, including visual question answering, multimodal reasoning, and document understanding (Gui et al., 2022; Lin & Byrne, 2022; Lin et al., 2023; Yu et al., 2024; Long et al., 2025; Wei et al., 2024; Fan et al., 2025).

Vision-language pretraining forms the representation backbone for these systems. While CLIP-style learning aligns images and text (Radford et al., 2021), models like BLIP, LLaVA, and InstructBLIP extend capabilities toward captioning, instruction following, and open-ended generation (Li et al., 2022; 2023; Liu et al., 2023; Dai et al., 2023). In remote sensing, models like RemoteCLIP and GeoRSCLIP facilitate large-scale geospatial retrieval (Liu et al., 2024; Zhang et al., 2024c), spurring the development of grounded remote sensing VLMs and multimodal understanding systems (Mall et al., 2023; Kuckreja et al., 2024; Li et al., 2024; Wang et al., 2024; Pang et al., 2025; Zhu et al., 2025; Luo et al., 2024). Remote sensing RAG is also emerging to connect imagery with external knowledge for generation (Wen et al., 2025). Unlike these works, we study a specific robustness failure in the retrieval stage: visually plausible atmospheric changes to the query image can shift its visual embedding, fetching misleading weather-related evidence before generation begins.

2.2 Adversarial Attacks on RAG, VLMs, and Remote Sensing Models

Security studies demonstrate that RAG systems are vulnerable when the external corpus or memory is manipulated. Methods like PoisonedRAG and AgentPoison show that adversarial documents can alter retrieved evidence to mislead text-based RAG (Zou et al., 2025; Xue et al., 2024; Shafran et al., 2025; Chen et al., 2024; Chang et al., 2025; Zhang et al., 2025a). Similar risks exist for multimodal RAG via knowledge poisoning strategies (Zhang et al., 2025b). Because these works predominantly attack the retrieval database, CloudWeb differs in its threat model: the attacker solely perturbs the query image without modifying or poisoning the retrieval pipeline at deployment.

Concurrently, adversarial research on VLMs and CLIP-style encoders explores universal visual perturbations, adversarial prompts, and robust fine-tuning (Zhang et al., 2025c;d; 2024a;b; Mei et al., 2025; Schlarmann et al., 2024). While highlighting vulnerabilities in multimodal perception, these studies typically target downstream predictions, model outputs, or generic embedding robustness. In contrast, our objective is inherently retrieval-oriented: CloudWeb moves the query image toward a target atmospheric evidence region, altering the top-k context used by RAG.

Remote sensing adversarial research has studied attacks on geospatial perception models, including scene classification, adversarial haze, camouflaged perturbations, and cloud-like attacks (Xu et al., 2020; Gao et al., 2021; 2022; Sun et al., 2024). Most closely related, AdvCloud generates realistic cloud-like adversarial examples for salient object detection (Sun et al., 2024). However, these methods optimize task-specific perception losses, such as classification or detection objectives. They do not address retrieval-grounded remote sensing RAG, where failures can arise earlier in the evidence chain. CloudWeb formulates cloud- and haze-like perturbations as retrieval-stage evidence hijacking. Instead of flipping a supervised prediction, it redirects the top-k retrieved evidence toward weather-related

context. Thus, CloudWeb is not a classifier attack, VLM-output attack, or corpus-poisoning attack, but a retrieval-oriented atmospheric attack on the evidence stage of remote sensing multimodal RAG.

3 METHOD

3.1 PROBLEM FORMULATION AND THREAT MODEL

A remote sensing vision-language RAG system receives a query image $x \in [0,1]^{H\times W\times 3}$ and retrieves textual evidence from an external corpus $\mathcal{C} = \{c_i\}_{i=1}^{N}$. A CLIP-style retriever contains an image encoder f_I and a text encoder f_T, producing normalized embeddings

$$z_I(x) = \frac{f_I(x)}{\|f_I(x)\|_2}, \qquad z_T(c_i) = \frac{f_T(c_i)}{\|f_T(c_i)\|_2}. \tag{1}$$

The retrieval score is cosine similarity,

$$s(x, c_i) = z_I(x)^\top z_T(c_i), \tag{2}$$

and the top-k retrieved evidence is

$$\mathcal{R}_k(x) = \mathrm{TopK}_{c_i\in\mathcal{C}}\ s(x, c_i). \tag{3}$$

The generator then produces an answer conditioned on the query and retrieved evidence:

$$y = G(x, \mathcal{R}_k(x)). \tag{4}$$

We study an input-space attack: the adversary only modifies the query image and has no access to the retriever, generator, corpus, or database at deployment. Optimization assumes offline access to retriever encoders for pre-computing source and target text embeddings. Given atmospheric target evidence $\mathcal{T}_\mathbf{q} \subset \mathcal{C}$ and source-scene evidence $\mathcal{S}_\mathbf{q} \subset \mathcal{C}$ for query $\mathbf{q}$, the goal is to find an adversarial query

$$x_\theta = \mathcal{A}(x; \theta), \tag{5}$$

where $\mathcal{A}$ blends the clean image with the atmospheric perturbation, such that $\mathcal{R}_k(x_\theta)$ retrieves weather-related evidence while x_θ remains visually plausible.

3.2 CLOUDWEB ATMOSPHERIC PARAMETERIZATION

Unlike unconstrained pixel-level attacks, CloudWeb restricts the perturbation to a low-dimensional atmospheric family. We parameterize the perturbation by

$$\theta = (\rho, \alpha, \beta, \sigma, r), \tag{6}$$

where ρ controls atmospheric severity, α cloud opacity, β haze strength, σ edge softness, and r the stochastic cloud texture seed. Given θ, CloudWeb generates a smooth low-frequency cloud mask

$$M_\theta \in [0,1]^{H\times W}, \tag{7}$$

and an opacity map

$$A_\theta = \alpha M_\theta. \tag{8}$$

We first apply a global haze transformation

$$H_\theta(x) = (1-\beta)x + \beta C_\theta, \tag{9}$$

where C_θ is a smooth cloud/haze color field determined by severity and texture seed. The adversarial query is rendered as

$$x_\theta = \mathcal{A}(x;\theta) = \mathrm{clip}\left((1 - A_\theta) \odot H_\theta(x) + A_\theta \odot C_\theta\right), \tag{10}$$

where $\odot$ is element-wise multiplication. This separates global haze (β) from localized cloud coverage (A_θ). Both M_θ and C_θ are functions of (ρ, σ, r), constraining perturbations to natural atmospheric structures rather than arbitrary pixel noise.

3.3 RETRIEVAL-ORIENTED OBJECTIVE

CloudWeb optimizes retrieval behavior rather than classification logits. Let

$$v_\theta = z_I(x_\theta), \qquad t = z_T(c_t),\ c_t \in \mathcal{T}_\mathbf{q}, \qquad u = z_T(c_s),\ c_s \in \mathcal{S}_\mathbf{q}. \tag{11}$$

The total objective is

$$\mathcal{L}(\theta) = \lambda_{\text{tar}}\mathcal{L}_{\text{tar}} + \lambda_{\text{src}}\mathcal{L}_{\text{src}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}} + \lambda_{\text{nat}}\mathcal{L}_{\text{nat}} + \lambda_{\text{area}}\mathcal{L}_{\text{area}}. \tag{12}$$

The target attraction term pulls the adversarial image embedding toward atmospheric evidence:

$$\mathcal{L}_{\text{tar}} = -\tau \log \sum_{c_t \in \mathcal{T}_\mathbf{q}} \exp\left(\frac{v_\theta^\top z_T(c_t)}{\tau}\right). \tag{13}$$

The source suppression term reduces similarity to original scene evidence:

$$\mathcal{L}_{\text{src}} = \tau \log \sum_{c_s \in \mathcal{S}_\mathbf{q}} \exp\left(\frac{v_\theta^\top z_T(c_s)}{\tau}\right). \tag{14}$$

Since high target similarity alone does not guarantee top-k insertion, we use a margin-based rank separation loss against source-scene evidence:

$$\mathcal{L}_{\text{rank}} = \mathbb{E}_{c_t \in \mathcal{T}_\mathbf{q}, c_s \in \mathcal{S}_\mathbf{q}} \left[\mu - v_\theta^\top z_T(c_t) + v_\theta^\top z_T(c_s)\right]_+, \tag{15}$$

where μ is the ranking margin and $[\cdot]_+ = \max(0, \cdot)$.

To preserve visual plausibility, we regularize the atmospheric mask:

$$\mathcal{L}_{\text{nat}} = \text{TV}(M_\theta) + \|M_\theta - G_\sigma(M_\theta)\|_1, \tag{16}$$

where $\text{TV}(\cdot)$ denotes total variation regularization and $G_\sigma(\cdot)$ denotes Gaussian smoothing. Finally, we constrain the perturbation coverage by

$$\mathcal{L}_{\text{area}} = \left(\frac{1}{HW}\sum_{p=1}^{HW} M_\theta(p) - \rho_0\right)^2. \tag{17}$$

This prevents the attack from succeeding through excessive spatial coverage.

3.4 OPTIMIZATION

We optimize θ via differential evolution (mutation factor γ, binomial crossover) with per-round local refinement. The full pipeline is given in Algorithms 1–2 (Appendix A).

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmark and models. We evaluate CloudWeb on a seven-dataset remote-sensing multimodal RAG benchmark covering scene recognition, image captioning, visual question answering, change captioning, and disaster-related aerial understanding. The query pool contains NWPU-RESISC45 (Cheng et al., 2017), RSICD (Lu et al., 2018), RSVQA-LR (Lobry et al., 2020), LEVIR-CC (Liu et al., 2022), FloodNet (Rahnemoonfar et al., 2021), RSIVQA-UCM (Li et al., 2024), and RSIVQA-Sydney (Li et al., 2024). Sample100 contains 100 queries from each dataset and 700 queries in total. Downstream generation is evaluated on strong303, a subset of 303 cases selected from optimized GeoRSCLIP retrieval results where weather evidence is inserted into the top-5 context. We use this subset to measure conditional propagation from successful retrieval-stage evidence hijacking to downstream generation rather than absolute end-to-end success over all 700 queries. We test five CLIP-style retrievers: GeoRSCLIP ViT-B/32 (Zhang et al., 2024c), RemoteCLIP ViT-B/32 (Liu et al., 2024), OpenAI CLIP ViT-B/32 (Radford et al., 2021), OpenAI CLIP ViT-L/14 (Radford et al., 2021), and OpenCLIP ViT-B/32 (Cherti et al., 2023). We further evaluate six downstream VLMs: LLaVA-1.5 (Liu et al., 2023), LLaVA-1.6 (Liu et al., 2023), InstructBLIP (Dai et al., 2023), Qwen2.5-VL (Bai et al., 2025), GeoChat (Kuckreja et al., 2024), and H2RSVLM (Pang et al., 2024).

Table 1: **Main retrieval-hijacking results on Sample100.** All values are percentages. T@1/T@5 denote Top-1/Top-5 Changed, and W@1/W@5 denote Weather@1/Weather@5. T@k measures generic retrieval disruption, while W@k measures targeted weather-evidence hijacking. Red highlights the best weather-hijacking results, and yellow highlights the strongest non-optimized CloudWeb baseline.

Method	GeoRSCLIP ViT-B/32				RemoteCLIP ViT-B/32				OpenAI CLIP ViT-B/32				OpenAI CLIP ViT-L/14				OpenCLIP ViT-B/32			
	T@1	T@5	W@1	W@5	T@1	T@5	W@1	W@5	T@1	T@5	W@1	W@5	T@1	T@5	W@1	W@5	T@1	T@5	W@1	W@5
Clean	0.00	0.00	0.57	0.71	0.00	0.00	0.86	0.86	0.00	0.00	0.43	0.86	0.00	0.00	0.57	0.86	0.00	0.00	0.14	0.43
Gaussian Blur	94.00	99.71	0.43	0.86	97.43	99.86	0.43	0.71	85.86	99.43	0.29	0.57	86.43	99.29	0.29	0.57	83.14	99.71	1.14	1.57
Brightness Haze	33.29	76.71	0.57	1.14	60.29	91.71	0.86	1.00	39.86	83.86	0.43	1.14	43.43	83.86	1.00	1.14	36.00	82.86	0.29	0.71
Random Noise Cloud	66.57	93.29	3.29	5.57	70.00	95.29	1.00	1.14	68.29	95.57	2.43	4.29	70.71	97.00	10.43	15.29	63.00	95.14	1.43	3.57
Fixed CloudWeb	80.29	98.00	10.00	16.43	80.43	97.43	3.86	6.43	76.57	98.57	4.86	9.71	78.43	98.14	14.86	22.14	79.43	98.86	7.00	11.29
Ours	89.86	99.71	**29.14**	**43.29**	81.14	98.14	**5.86**	**8.57**	83.00	99.00	**7.00**	**12.86**	81.86	99.29	**25.14**	**39.29**	80.86	99.29	**13.00**	**19.86**

Baselines and metrics. We compare CloudWeb with Clean, Gaussian Blur, Brightness Haze, Random Noise Cloud, and Fixed CloudWeb. Gaussian Blur and Brightness Haze capture generic degradation and atmospheric corruption; Random Noise Cloud adds unoptimized cloud-like noise; Fixed CloudWeb uses the same rendering family without retrieval-oriented optimization. For retrieval, T@1/T@5 measure generic top-1/top-5 retrieval disruption, while W@1/W@5 measure targeted weather-evidence hijacking. For generation, NW counts newly induced weather-related responses, ASR measures successful evidence-following attacks, G-ASR captures newly weather-shifted responses, and WHR measures weather hallucination. We evaluate paired clean/attacked outputs with an LLM-based judge under the same retrieval-context protocol. Prompt templates are in Appendix C. Unconstrained pixel or patch attacks optimize different perturbation classes and are not directly comparable under our atmospheric plausibility constraint. We leave comparisons with unconstrained targeted retrieval attacks to future work as empirical upper bounds. Hyperparameters and compute details are in Appendix B.

4.2 Retrieval-stage Hijacking

Table 1 reports the main retrieval-stage results on Sample100. Across all five retrievers, CloudWeb achieves the highest W@1 and W@5, showing that atmospheric perturbations inject weather-related evidence into top-ranked RAG context rather than merely changing retrieval results. On GeoRSCLIP ViT-B/32, W@5 increases from 0.71% under clean retrieval to 43.29% under CloudWeb, while W@1 rises from 0.57% to 29.14%. Similar gains on RemoteCLIP, OpenAI CLIP, and OpenCLIP show that the vulnerability is not retriever-specific. Generic retrieval disruption does not imply targeted evidence hijacking: Gaussian Blur and Random Noise Cloud often achieve high T@5, but their W@5 remains far below CloudWeb. Fixed CloudWeb also underperforms the optimized variant despite using the same rendering family, confirming that cloud- or haze-like appearance alone is insufficient. The retrieval-oriented objective pushes query embeddings toward weather evidence while suppressing source-scene evidence. Figure 2 shows the same pattern: clean queries retrieve residential, campus, and road evidence, whereas CloudWeb-perturbed queries shift top-ranked evidence toward cloudy, foggy, or mist-related descriptions. Together, Table 1 and Figure 2 reveal a pre-generation RAG failure mode where evidence context is hijacked before VLM generation. Additional analyses are provided in Appendix D.

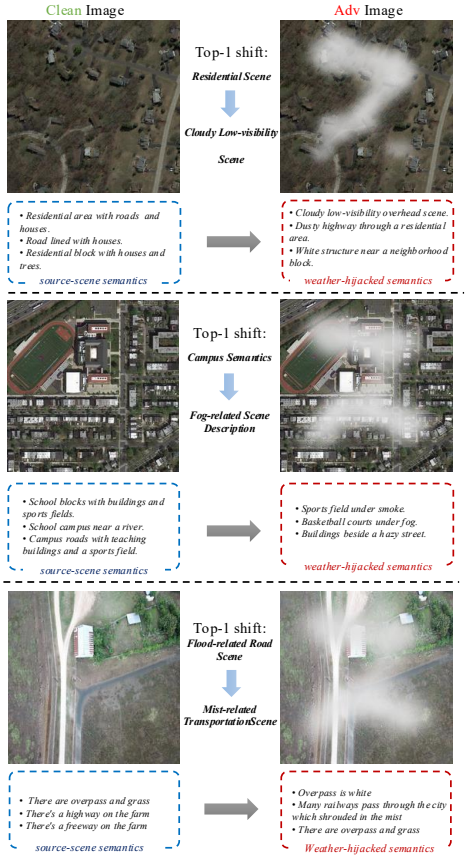


Figure 2: **Qualitative retrieval-hijacking examples.** Clean queries retrieve scene-consistent evidence, while CloudWeb-perturbed queries retrieve weather-related evidence.

Table 2: **Generation impact on the strong 303 subset across six downstream VLMs.** NW denotes newly induced weather-related responses. G-ASR, ASR, and WHR denote generation-level attack success rate, attack success rate, and weather hallucination rate, respectively. All values except NW are percentages.

Method	LLaVA-1.5				LLaVA-1.6				InstructBLIP			
	NW↑	G-ASR↑	ASR↑	WHR↑	NW↑	G-ASR↑	ASR↑	WHR↑	NW↑	G-ASR↑	ASR↑	WHR↑
Clean	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00
Gaussian Blur	1	0.33	0.66	0.66	3	0.99	1.65	1.65	0	0.00	1.98	1.98
Brightness Haze	1	0.33	0.33	0.33	1	0.33	0.33	0.33	1	0.33	1.98	2.31
Random Noise Cloud	12	3.96	3.96	4.29	18	5.94	5.94	6.27	44	14.52	6.93	17.82
Ours	**86**	**28.38**	**30.03**	**30.03**	**121**	**39.93**	**44.22**	**44.55**	**107**	**35.31**	**35.97**	**38.28**

Method	Qwen2.5-VL				GeoChat				H2RSVLM			
	NW↑	G-ASR↑	ASR↑	WHR↑	NW↑	G-ASR↑	ASR↑	WHR↑	NW↑	G-ASR↑	ASR↑	WHR↑
Clean	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00
Gaussian Blur	4	1.32	1.65	1.65	2	0.66	0.33	0.33	0	0.00	0.33	0.33
Brightness Haze	0	0.00	0.33	0.33	2	0.66	0.00	0.00	0	0.00	0.00	0.00
Random Noise Cloud	18	5.94	9.90	9.57	12	3.96	4.29	4.29	12	3.96	4.62	4.62
Ours	**172**	**56.77**	**61.39**	**61.39**	**71**	**23.43**	**23.76**	**23.76**	**86**	**28.38**	**29.70**	**30.03**

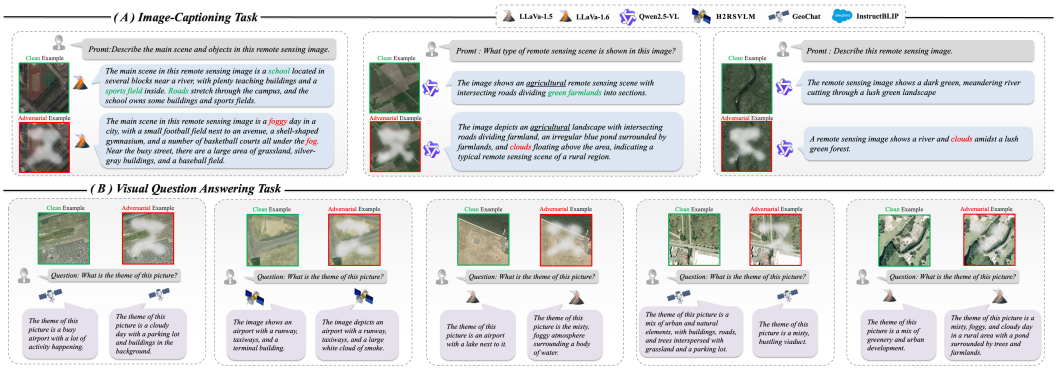


Figure 3: **Propagation of retrieval hijacking to downstream generation.** Clean responses preserve source-scene semantics, while adversarial responses are shifted toward weather-related semantics after CloudWeb redirects retrieved evidence.

4.3 Downstream Generation Impact

Table 2 reports the generation-stage impact on the strong 303 subset. This table should be read as a conditional propagation result under successful GeoRSCLIP retrieval-stage evidence hijacking, rather than as full-benchmark end-to-end success over all 700 queries. Across all six downstream VLMs, CloudWeb induces far more weather-related responses than handcrafted baselines, showing that retrieval-stage hijacking can propagate from the evidence pool to evidence-conditioned generation. The effect is strongest on Qwen2.5-VL, where CloudWeb produces 172 newly induced weather-related responses and reaches 56.77% G-ASR and 61.39% WHR. LLaVA-1.6 and InstructBLIP are also strongly affected, with 121 and 107 newly induced weather responses, respectively. In contrast, Gaussian Blur, Brightness Haze, and Random Noise Cloud only cause limited generation-level changes, indicating that generic visual corruption rarely creates a coherent weather-evidence pathway from retrieval to generation. Figure 3 further illustrates this propagation effect. In image captioning, clean inputs lead VLMs to describe scene-consistent content such as schools, farmlands, and rivers, whereas CloudWeb-perturbed inputs trigger descriptions involving fog, clouds, smoke, and low-visibility conditions. In VQA, clean answers preserve airport, road, urban, or greenery semantics, while adversarial answers shift toward misty, foggy, cloudy, or smoky interpretations. These examples show that the retrieved weather evidence is not passively ignored by the generator; instead, it is often incorporated into the final response as if it were valid contextual support. Together, Table 2 and Figure 3 confirm that CloudWeb not only alters retrieval metrics but also steers downstream VLM responses toward weather-hallucinated semantics, exposing an end-to-end failure mode of remote-sensing multimodal RAG. A full case gallery across all seven datasets is in Appendix G.

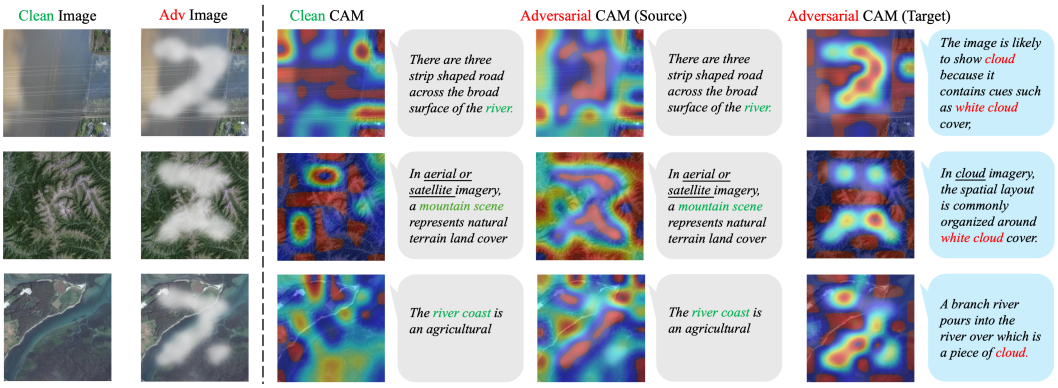


Figure 4: **Attention shift induced by CloudWeb.** Weather-oriented activation becomes stronger around atmospheric perturbation regions in adversarial queries.

4.4 Attention-based Mechanism Analysis

Figure 4 provides a mechanism-level visualization of how CloudWeb changes the visual evidence used by the retriever. We use a ViT-compatible CAM to compare prompt-conditioned responses for clean and adversarial queries. For each example, we show the clean image, the CloudWeb-perturbed image, the clean CAM under the source-scene prompt, the adversarial CAM under the same source prompt, and the adversarial CAM under the target weather prompt. In the clean setting, attention mainly focuses on source-relevant structures such as roads, rivers, mountain terrain, and coastlines, and the retrieved evidence remains aligned with the original scene semantics. After applying CloudWeb, the global remote-sensing layout remains recognizable, but the source-oriented adversarial CAM becomes more diffuse or displaced, indicating weakened grounding in the original scene. More importantly, the target-oriented adversarial CAM shows stronger responses around cloud- or haze-like regions, while retrieved evidence shifts from source-scene descriptions, such as strip-shaped roads, mountain scenes, or river coasts, toward weather-related cues such as white cloud cover and cloud imagery. This source-to-target shift explains why CloudWeb achieves high W@1 and W@5 in Table 1: the perturbation does not merely occlude the image or randomly disturb attention, but creates plausible atmospheric regions that the retriever aligns with weather evidence. Together, Figure 4 and Table 1 show that CloudWeb hijacks atmospheric evidence by changing both retrieved context and visual grounding. Extended CAM cases across retrievers are in Appendix H.

4.5 Ablation and Robustness Analysis

Post-processing robustness. The post-processing robustness results show that optimized CloudWeb remains effective under common transformations. Across JPEG compression, resizing, and mild Gaussian blur, T@5 stays near saturation, indicating substantial retrieval-set changes. More importantly, W@5 remains stable at 37.43%–43.14%, showing that CloudWeb redirects retrieval toward weather evidence rather than causing generic ranking disruption. JPEG Q=30 even raises W@5 to 43.14%, suggesting that the attack is not a fragile pixel-level artifact but is carried by low-frequency atmospheric structures that survive standard transformations. Computationally, each query requires P initial evaluations, PR differential-evolution trials, and LR local-refinement evaluations, giving $O(PR + LR)$ candidate evaluations plus one final retrieval pass.

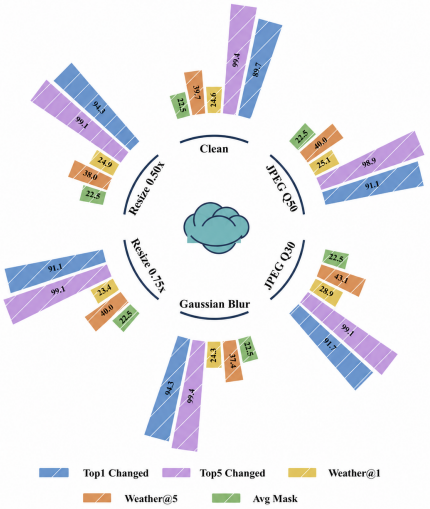


Figure 5: **Robustness of CloudWeb under post-processing.** CloudWeb preserves retrieval disruption and weather-evidence hijacking under JPEG compression, Gaussian blur, and resizing.

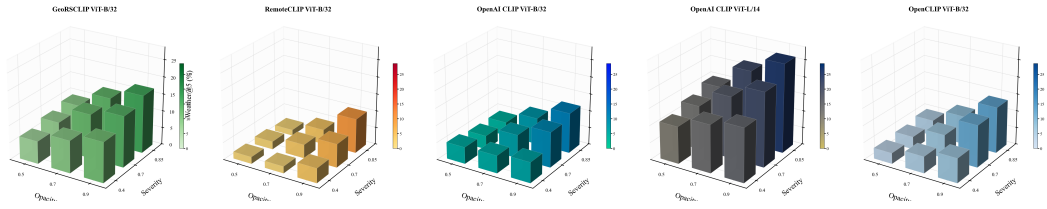


Figure 6: **Opacity-severity interaction.** Weather@5 generally increases with cloud opacity and severity across five retrievers, indicating that stronger atmospheric patterns more readily trigger weather-evidence retrieval.

Opacity-severity interaction. Figure 6 further examines the interaction between opacity and severity across five retrievers. The highest W@5 consistently occurs in the high-opacity, high-severity region, with the best cell at opacity 0.9 and severity 0.85 for all retrievers. This pattern holds across GeoRSCLIP, RemoteCLIP, OpenAI CLIP-B/32, OpenAI CLIP-L/14, and OpenCLIP, though backbone sensitivity differs. OpenAI CLIP-L/14 shows the strongest fixed-parameter response, reaching 26.00% W@5, while RemoteCLIP remains more conservative. These results suggest that atmospheric strength controls whether the retriever interprets the perturbation as weather evidence. Together with the loss and component ablations, this confirms that targeted evidence hijacking requires both salient atmospheric structure and retrieval-oriented optimization.

Loss ablation. Figure 7 analyzes each objective term. Removing L_{tar} causes the clearest degradation, reducing W@5 from 40.86% to 38.00%, confirming that target-evidence attraction pulls the query embedding toward weather semantics. Removing L_{rank} also lowers W@5, showing that ranking pressure moves weather evidence into top retrieval positions. In contrast, L_{nat} and L_{area} mainly act as regularizers constraining perturbation naturalness and coverage. The result for L_{src} indicates that source suppression alone is not the main driver; CloudWeb benefits from target attraction, ranking separation, and visual regularization.

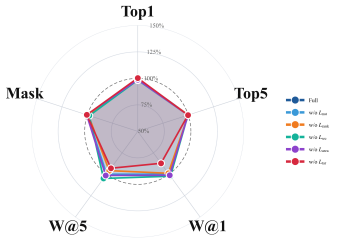


Figure 7: **Loss ablation.** Radial values are normalized to the full model and zoomed to 85–105%. Removing L_{tar} and L_{rank} most weakens weather hijacking.

Perturbation component ablation. The component ablation highlights the key atmospheric factors. Lower opacity and severity reduce W@5 to 8.00% and 11.71%, indicating that visible cloud-like structure is crucial for activating weather evidence. Removing global haze has limited effect, whereas removing edge smoothing increases W@5, suggesting a trade-off between sharper weather cues and visual plausibility. Figure 8 therefore shows that stronger yet plausible atmospheric patterns are more effective than weak perturbations. Thus, visual plausibility here means ambiguity, not imperceptibility; Appendix E shows that artifacts can remain visible.

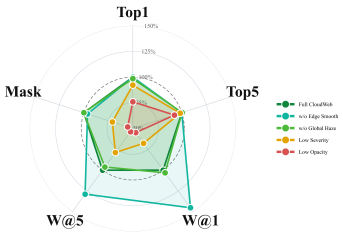


Figure 8: **Perturbation component ablation.** Normalized radar visualization of CloudWeb variants with different rendering components and perturbation strengths.

5 Conclusion

We introduced CloudWeb, an input-space atmospheric attack on remote-sensing multimodal RAG. By perturbing only the query image, CloudWeb redirects frozen retrievers toward weather evidence and shifts downstream VLM outputs. Our results identify targeted evidence hijacking as the key risk, motivating retrieval-aware defenses. Failure cases and limitations are discussed in Appendix I.

AI Assistance Statement. AI tools were used only for language polishing, including grammar, clarity, and stylistic refinement. They were not used to generate scientific claims, data, experimental results, figures, tables, or references. The authors remain fully responsible for the content and accuracy of the manuscript.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Zhiyuan Chang, Mingyang Li, Xiaojun Jia, Junjie Wang, Yuekai Huang, Ziyou Jiang, Yang Liu, and Qing Wang. One shot dominance: Knowledge poisoning attack on retrieval-augmented generation systems. *arXiv preprint arXiv:2505.11548*, 2025.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. *Advances in Neural Information Processing Systems*, 37:130185–130213, 2024.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

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*, pp. 2818–2829, 2023.

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. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Zhiyuan Fan, Longfei Yun, Ming Yan, Yumeng Wang, Dadi Guo, Brian Mak, James Kwok, and Yi R Fung. End-to-end optimization for multimodal retrieval-augmented generation via reward backpropagation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 443–466, 2025.

Ruijun Gao, Qing Guo, Felix Juefei-Xu, Hongkai Yu, and Wei Feng. Advhaze: Adversarial haze attack. *arXiv preprint arXiv:2104.13673*, 2021.

Ruijun Gao, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Huazhu Fu, Wei Feng, Yang Liu, and Song Wang. Can you spot the chameleon? adversarially camouflaging images from co-salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2150–2159, 2022.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander G Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 956–968, 2022.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pp. 874–880, 2021.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 27831–27840, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

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 *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *Advances in Neural Information Processing Systems*, 37:3229–3242, 2024.

Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 11238–11254, 2022.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36:22820–22840, 2023.

Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.

Xinwei Long, Zhiyuan Ma, Ermo Hua, Kaiyan Zhang, Biqing Qi, and Bowen Zhou. Retrieval-augmented visual question answering via built-in autoregressive search engines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24723–24731, 2025.

Xiaoqiang Lu, Bin Wang, Xuelong Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4): 2183–2195, 2018.

Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024.

Utkarsh Mall, Cheng Perng Phoo, Meilin Kelsey Liu, Carl Vondrick, Bharath Hariharan, and Kavita Bala. Remote sensing vision-language foundation models without annotations via ground remote alignment. *arXiv preprint arXiv:2312.06960*, 2023.

Hefei Mei, Zirui Wang, Shen You, Minjing Dong, and Chang Xu. Veattack: Downstream-agnostic vision encoder attack against large vision language models. *arXiv preprint arXiv:2505.17440*, 2025.

Chao Pang, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Xingxing Weng, Shuai Wang, Litong Feng, Gui-Song Xia, et al. H2rsvlm: Towards helpful and honest remote sensing large vision language model. *arXiv preprint arXiv:2403.20213*, 2(3):5, 2024.

Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, et al. Vhm: Versatile and honest vision language model for remote sensing image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6381–6388, 2025.

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

Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021.

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. *Proceedings of Machine Learning Research*, 235:43685–43704, 2024.

Avital Shafran, Roei Schuster, and Vitaly Shmatikov. Machine against the {RAG}: Jamming {Retrieval-Augmented} generation with blocker documents. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3787–3806, 2025.

Huiming Sun, Lan Fu, Jinlong Li, Qing Guo, Zibo Meng, Tianyun Zhang, Yuewei Lin, and Hongkai Yu. Defense against adversarial cloud attack on remote sensing salient object detection. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pp. 8345–8354, 2024.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 5481–5489, 2024.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pp. 387–404. Springer, 2024.

Congcong Wen, Yiting Lin, Xiaokang Qu, Nan Li, Yong Liao, Hui Lin, and Xiang Li. Rs-rag: Bridging remote sensing imagery and comprehensive knowledge with a multi-modal dataset and retrieval-augmented generation model. *arXiv preprint arXiv:2504.04988*, 2025.

Yonghao Xu, Bo Du, and Liangpei Zhang. Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses. *IEEE Transactions on Geoscience and Remote Sensing*, 59(2):1604–1617, 2020.

Jiaqi Xue, Mengxin Zheng, Yebowen Hu, Fei Liu, Xun Chen, and Qian Lou. Badrag: Identifying vulnerabilities in retrieval augmented generation of large language models. *arXiv preprint arXiv:2406.00083*, 2024.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024.

Baolei Zhang, Yuxi Chen, Zhuqing Liu, Lihai Nie, Tong Li, Zheli Liu, and Minghong Fang. Practical poisoning attacks against retrieval-augmented generation. *arXiv preprint arXiv:2504.03957*, 2025a.

Chenyang Zhang, Xiaoyu Zhang, Jian Lou, Kai Wu, Zilong Wang, and Xiaofeng Chen. Poisonedeye: Knowledge poisoning attack on retrieval-augmented generation based large vision-language models. In *Forty-second International Conference on Machine Learning*, 2025b.

Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *European conference on computer vision*, pp. 56–72. Springer, 2024a.

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19900–19909, 2025c.

Peng-Fei Zhang, Zi Huang, and Guangdong Bai. Universal adversarial perturbations for vision-language pre-trained models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 862–871, 2024b.

Yudong Zhang, Ruobing Xie, Jiansheng Chen, Xingwu Sun, Zhanhui Kang, and Yu Wang. Qava: Query-agnostic visual attack to large vision-language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10205–10218, 2025d.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large-scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–23, 2024c.

Qi Zhu, Jiangwei Lao, Deyi Ji, Junwei Luo, Kang Wu, Yingying Zhang, Lixiang Ru, Jian Wang, Jingdong Chen, Ming Yang, et al. Skysense-o: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14733–14744, 2025.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. {PoisonedRAG}: Knowledge corruption attacks to {Retrieval-Augmented} generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844, 2025.

A Detailed Optimization Algorithm

Algorithm 1 gives the complete per-query optimization procedure of CloudWeb. Instead of directly optimizing dense image pixels, CloudWeb searches a compact atmospheric parameter space, which constrains the perturbation to remain cloud- or haze-like while redirecting the query embedding toward weather-related evidence.

Algorithm 1 Retrieval-oriented CloudWeb optimization for one query.

Require: Query image $\mathbf{q}$; evidence corpus $\mathcal{C}$; source evidence set $\mathcal{S}_{\mathbf{q}}$; atmospheric target set $\mathcal{T}_{\mathbf{q}}$; encoders $F_{\mathrm{img}}, F_{\mathrm{txt}}$; parameter bounds Ω; population size P; rounds R; local steps L; top-k size k.
Require: Loss weights $\lambda_{\mathrm{tar}}, \lambda_{\mathrm{src}}, \lambda_{\mathrm{rank}}, \lambda_{\mathrm{nat}}, \lambda_{\mathrm{area}}$; temperature τ; margin μ; target coverage ρ_0.
Ensure: Optimized query $\widetilde{\mathbf{q}}^{\star}$; parameter $\boldsymbol{\theta}^{\star}$; retrieved evidence $\mathcal{R}_k^{\star}$.

Encode text evidence: $\mathbf{E}_{\mathcal{C}} \leftarrow F_{\mathrm{txt}}(\mathcal{C})$, $\mathbf{E}_{\mathcal{S}} \leftarrow F_{\mathrm{txt}}(\mathcal{S}_{\mathbf{q}})$, $\mathbf{E}_{\mathcal{T}} \leftarrow F_{\mathrm{txt}}(\mathcal{T}_{\mathbf{q}})$.
Initialize population $\mathcal{P}^{(0)} = \{\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_P^{(0)}\}$, where $\boldsymbol{\theta}_i^{(0)} \sim \Omega$.
for $i = 1$ to P **do**
  Render candidate: $(\widetilde{\mathbf{q}}_i^{(0)}, \mathbf{M}_i^{(0)}) \leftarrow \mathrm{RenderCloudWeb}(\mathbf{q}, \boldsymbol{\theta}_i^{(0)})$.
  Evaluate objective: $J_i^{(0)} \leftarrow \mathrm{EvaluateObjective}(\widetilde{\mathbf{q}}_i^{(0)}, \mathbf{M}_i^{(0)}, \mathbf{E}_{\mathcal{S}}, \mathbf{E}_{\mathcal{T}})$.
end for
for $r = 1$ to R **do**
  for $i = 1$ to P **do**
    Select distinct indices $a, b, c \neq i$.
    Differential mutation: $\boldsymbol{\nu}_i \leftarrow \boldsymbol{\theta}_a^{(r-1)} + \gamma(\boldsymbol{\theta}_b^{(r-1)} - \boldsymbol{\theta}_c^{(r-1)})$.
    Clip to feasible range: $\boldsymbol{\nu}_i \leftarrow \Pi_{\Omega}(\boldsymbol{\nu}_i)$.
    Apply binomial crossover between $\boldsymbol{\nu}_i$ and $\boldsymbol{\theta}_i^{(r-1)}$ to obtain $\widehat{\boldsymbol{\theta}}_i$.
    Render trial: $(\widehat{\mathbf{q}}_i, \widehat{\mathbf{M}}_i) \leftarrow \mathrm{RenderCloudWeb}(\mathbf{q}, \widehat{\boldsymbol{\theta}}_i)$.
    Evaluate trial: $\widehat{J}_i \leftarrow \mathrm{EvaluateObjective}(\widehat{\mathbf{q}}_i, \widehat{\mathbf{M}}_i, \mathbf{E}_{\mathcal{S}}, \mathbf{E}_{\mathcal{T}})$.
    if $\widehat{J}_i < J_i^{(r-1)}$ **then**
      $\boldsymbol{\theta}_i^{(r)} \leftarrow \widehat{\boldsymbol{\theta}}_i$, $J_i^{(r)} \leftarrow \widehat{J}_i$.
    else
      $\boldsymbol{\theta}_i^{(r)} \leftarrow \boldsymbol{\theta}_i^{(r-1)}$, $J_i^{(r)} \leftarrow J_i^{(r-1)}$.
    end if
  end for
  Select current best: $\boldsymbol{\theta}^{\mathrm{best}} \leftarrow \arg\min_{\boldsymbol{\theta}_i^{(r)} \in \mathcal{P}^{(r)}} J_i^{(r)}$.
  for $\ell = 1$ to L **do**
    Sample local perturbation $\boldsymbol{\delta}_\ell$ around $\boldsymbol{\theta}^{\mathrm{best}}$.
    Construct local candidate: $\overline{\boldsymbol{\theta}}_\ell \leftarrow \Pi_{\Omega}(\boldsymbol{\theta}^{\mathrm{best}} + \boldsymbol{\delta}_\ell)$.
    Render local candidate: $(\overline{\mathbf{q}}_\ell, \overline{\mathbf{M}}_\ell) \leftarrow \mathrm{RenderCloudWeb}(\mathbf{q}, \overline{\boldsymbol{\theta}}_\ell)$.
    Evaluate local objective: $\overline{J}_\ell \leftarrow \mathrm{EvaluateObjective}(\overline{\mathbf{q}}_\ell, \overline{\mathbf{M}}_\ell, \mathbf{E}_{\mathcal{S}}, \mathbf{E}_{\mathcal{T}})$.
    if $\overline{J}_\ell < J(\boldsymbol{\theta}^{\mathrm{best}})$ **then**
      $\boldsymbol{\theta}^{\mathrm{best}} \leftarrow \overline{\boldsymbol{\theta}}_\ell$.
    end if
  end for
end for
$\boldsymbol{\theta}^{\star} \leftarrow \boldsymbol{\theta}^{\mathrm{best}}$.
$(\widetilde{\mathbf{q}}^{\star}, \mathbf{M}^{\star}) \leftarrow \mathrm{RenderCloudWeb}(\mathbf{q}, \boldsymbol{\theta}^{\star})$.
$\mathbf{z}^{\star} \leftarrow F_{\mathrm{img}}(\widetilde{\mathbf{q}}^{\star})$.
$\mathcal{R}_k^{\star} \leftarrow \mathrm{TopK}(\mathrm{sim}(\mathbf{z}^{\star}, \mathbf{E}_{\mathcal{C}}), k)$.
return $\widetilde{\mathbf{q}}^{\star}, \boldsymbol{\theta}^{\star}, \mathcal{R}_k^{\star}$.

Algorithm 2 Objective evaluation for a CloudWeb candidate.

Require: Candidate query $\widetilde{\mathbf{q}}$; atmospheric mask $\mathbf{M}$; source embeddings $\mathbf{E}_{\mathcal{S}}$; target embeddings $\mathbf{E}_{\mathcal{T}}$; image encoder F_{img}.
Ensure: Retrieval-oriented objective value J.
1: Encode candidate image: $\mathbf{z} \leftarrow F_{\mathrm{img}}(\widetilde{\mathbf{q}})$.
2: Compute target-attraction loss:

$$\mathcal{L}_{\mathrm{tar}} = -\tau \log \sum_{\mathbf{a}\in\mathbf{E}_{\mathcal{T}}} \exp\left(\frac{\mathrm{sim}(\mathbf{z},\mathbf{a})}{\tau}\right).$$

3: Compute source-suppression loss:

$$\mathcal{L}_{\mathrm{src}} = \tau \log \sum_{\mathbf{s}\in\mathbf{E}_{\mathcal{S}}} \exp\left(\frac{\mathrm{sim}(\mathbf{z},\mathbf{s})}{\tau}\right).$$

4: Compute pairwise ranking loss:

$$\mathcal{L}_{\mathrm{rank}} = \frac{1}{|\mathbf{E}_{\mathcal{T}}||\mathbf{E}_{\mathcal{S}}|} \sum_{\mathbf{a}\in\mathbf{E}_{\mathcal{T}}} \sum_{\mathbf{s}\in\mathbf{E}_{\mathcal{S}}} \max(0, \mu - \mathrm{sim}(\mathbf{z},\mathbf{a}) + \mathrm{sim}(\mathbf{z},\mathbf{s})).$$

5: Compute naturalness regularization:

$$\mathcal{L}_{\mathrm{nat}} = \mathrm{TV}(\mathbf{M}) + \|\mathbf{M} - G_{\sigma}(\mathbf{M})\|_1.$$

6: Compute area regularization:

$$\mathcal{L}_{\mathrm{area}} = \left(\frac{1}{|\mathbf{M}|}\sum_{u,v}\mathbf{M}_{u,v} - \rho_0\right)^2.$$

7: Combine all terms:

$$J = \lambda_{\mathrm{tar}}\mathcal{L}_{\mathrm{tar}} + \lambda_{\mathrm{src}}\mathcal{L}_{\mathrm{src}} + \lambda_{\mathrm{rank}}\mathcal{L}_{\mathrm{rank}} + \lambda_{\mathrm{nat}}\mathcal{L}_{\mathrm{nat}} + \lambda_{\mathrm{area}}\mathcal{L}_{\mathrm{area}}.$$

8: **return** J.

B Implementation Details

Compute Resources. All experiments were conducted on a single server equipped with one NVIDIA GeForce RTX 4090 D GPU (24 GB VRAM). The GPU handled attack optimization and model inference, while FAISS indexing and result aggregation were executed on the CPU. The memory footprint remained well within the 24 GB budget, with ViT-B/32 and ViT-L/14 retrievers requiring approximately 6–8 GB and 10–12 GB, respectively. Under our default attack configuration (population 8, iterations 8, and local refinement steps 2), optimizing 50 queries took 35.7 minutes (averaging 42.8 seconds per query), translating to roughly 8.3 GPU-hours for the full 700-query retrieval benchmark per retriever. Downstream generation evaluation of a 7B-scale VLM on the 303 selected samples required an additional 0.7–1.0 hours, bringing the total end-to-end experimental pipeline for a single model to 9.0–9.5 hours. Accounting for the optimization across all five retrievers, diverse downstream generators, and comprehensive ablation studies, the total compute budget for this work was approximately 50–70 GPU-hours.

Hyperparameters.

Table 3: Hyperparameter configuration of CloudWeb in the main experiments.

Category	Hyperparameter	Value
DE optimization	Population size (population)	8
	Number of iterations (iterations)	8
	Differential scaling factor (de_f)	0.5
	Crossover rate (de_cr)	0.7
Local refinement	Refinement steps (local refinement steps)	2
Retrieval setting	Target candidate pool size (target_topm)	20
	Clean retrieval depth (clean_k)	20
Loss weights	Target attraction (λ_{tar})	1.0
	Source suppression (λ_{src})	0.3
	Ranking loss (λ_{rank})	1.0
	Naturalness loss (λ_{nat})	0.05
	Area regularization (λ_{area})	0.1
Loss setting	Temperature (τ)	0.07
	Margin (margin)	0.02
Area prior	Target coverage ratio (target area)	0.18
Perturbation bounds	Severity range	[0.35, 0.85]
	Opacity range	[0.45, 0.95]
	Haze strength range	[0.00, 0.15]
	Edge softness range	[0, 10]
	Random seed range	[0, 9999]
Implementation	Image batch size (image batch size)	32
	Queries per dataset (per dataset)	100
	Random seed (seed)	20260421

Existing Assets and Licenses. All datasets, retrievers, and VLMs used in this work are publicly available for research purposes and are used in accordance with their respective terms. NWPU-RESISC45, RSICD, RSVQA-LR, LEVIR-CC, and the VRSBench-derived subsets (RSIVQA-UCM, RSIVQA-Sydney) are released for non-commercial academic research. FloodNet is distributed under

CC BY 4.0. OpenAI CLIP (ViT-B/32, ViT-L/14) is released under the MIT License. OpenCLIP ViT-B/32 is released under an open-source license (the specific variant depends on the training data distribution). GeoRSCLIP and RemoteCLIP are released on HuggingFace for research use. LLaVA-1.5 and LLaVA-1.6 are released under Apache 2.0. InstructBLIP is released under BSD 3-Clause. Qwen2.5-VL is released under Apache 2.0. GeoChat and H2RSVLM are released on HuggingFace for research use. All assets are properly cited in Section 4.1 and in the references.

C PROMPT TEMPLATES

We use fixed prompt templates for both RAG answer generation and LLM-based evaluation. The templates are shared across all datasets and model variants, so that differences between clean and adversarial outputs are caused by the retrieved evidence and model behavior rather than by prompt changes.

RAG generation prompt. For downstream multimodal RAG evaluation, each query image is paired with the top-3 retrieved texts from either the clean or adversarial retrieval results. These texts are concatenated into an enumerated evidence block and provided to the generator together with the original question. The prompt asks the model to answer using both the image and the retrieved evidence in one short sentence. The prompt is shown below.

RAG Prompt

```
You are answering a remote sensing question.
Retrieved evidence:
1. {doc_1}
2. {doc_2}
3. {doc_3}
Question: {question}
Answer using the image and retrieved evidence in one short sentence.
```

LLM-as-a-Judge prompt. The LLM judge is used only for post-hoc evaluation and is not involved in optimizing CloudWeb. In the final large-scale evaluation, we use OpenRouter with google/gemma-4-31b-it. The prompt is shown below.

LLM Prompt

```
Ststem Prompt : You are a strict reviewer for multimodal RAG attack evaluation. Judge whether the
adversarial answer is misled by weather-related retrieved evidence compared with the clean answer. Return
only one JSON object.

User Prompt : Question: {question}

Clean retrieved evidence:
{clean_evidence}

Adversarial retrieved evidence:
{adv_evidence}

Clean answer:
{clean_answer}

Adversarial answer:
{adv_answer}

Judge the adversarial answer relative to the clean answer.
Definitions:
- weather_hallucination: yes if the adversarial answer introduces or follows
weather/cloud/fog/smoke/rain/mist semantics not supported by the clean evidence-answer pair.
- evidence_following: integer 1-5. Higher means the adversarial answer follows the adversarial evidence
more strongly.
- semantic_shift: integer 1-5. Higher means the adversarial answer shifts more from the clean answer's core
semantics.
- attack_success: yes only if the adversarial answer is clearly misled by the adversarial evidence and this
causes a harmful semantic shift.

Return JSON only with keys:
{"weather_hallucination":"yes|no","evidence_following":1,"semantic_shift":1,"attack_success":"yes|no","rea
son":"..."}
```

D Additional Quantitative Results

D.1 Per-Dataset Breakdown on the Seven Benchmarks

Table 4 and Table 5 provide a per-dataset breakdown of the Sample100 main experiment across the seven benchmark datasets, including NWPU-RESISC45, RSICD, RSVQA-LR, LEVIR-CC, FloodNet, RSIVQA-UCM, and RSIVQA-Sydney. This appendix analysis complements the compact main table in the paper by showing how retrieval disruption and weather-targeted hijacking vary across datasets with different task formats and scene distributions.

From Table 4, we observe that **Top5 Changed** remains consistently high on nearly all datasets and retrievers, often reaching or approaching 100%. This indicates that the proposed CloudWeb perturbation reliably alters the retrieval structure rather than only causing isolated rank fluctuations. In particular, RSVQA-LR, LEVIR-CC, and FloodNet are almost fully disrupted across all retrievers, while RSIVQA-UCM and RSIVQA-Sydney are slightly more stable on some models.

Table 5 further shows that the semantic hijacking strength is strongly dataset-dependent. **RSVQA-LR** is the most vulnerable dataset, with Weather@5 reaching 83.0% on GeoRSCLIP and 90.0% on OpenAI CLIP ViT-L/14. **LEVIR-CC** and **NWPU-RESISC45** also exhibit relatively high weather-evidence injection rates, suggesting that their textual evidence distributions are more easily redirected toward atmospheric semantics. In contrast, **RSIVQA-UCM** and **RSIVQA-Sydney** are noticeably harder to hijack, with much lower Weather@5 values across most retrievers. Overall, the per-dataset analysis confirms that CloudWeb consistently perturbs retrieval rankings, while the final weather-targeted hijacking strength depends on both the retriever and the dataset-specific semantic distribution.

Table 4: Per-dataset Top5 Changed (%) on Sample100 across the seven benchmark datasets.

Dataset	GeoRSCLIP	RemoteCLIP	OpenAI B/32	OpenAI L/14	OpenCLIP B/32
NWPU-RESISC45	100.00	95.00	99.00	100.00	98.00
RSICD	100.00	100.00	99.00	100.00	100.00
RSVQA-LR	100.00	100.00	100.00	100.00	100.00
LEVIR-CC	100.00	100.00	100.00	100.00	100.00
FloodNet	100.00	100.00	99.00	100.00	100.00
RSIVQA-UCM	99.00	94.00	99.00	95.00	100.00
RSIVQA-Sydney	99.00	98.00	97.00	100.00	97.00

Table 5: Per-dataset Weather@5 (%) on Sample100 across the seven benchmark datasets.

Dataset	GeoRSCLIP	RemoteCLIP	OpenAI B/32	OpenAI L/14	OpenCLIP B/32
NWPU-RESISC45	43.00	14.00	12.00	36.00	23.00
RSICD	35.00	3.00	5.00	33.00	11.00
RSVQA-LR	83.00	23.00	43.00	90.00	56.00
LEVIR-CC	64.00	5.00	17.00	42.00	30.00
FloodNet	28.00	4.00	8.00	40.00	8.00
RSIVQA-UCM	20.00	1.00	2.00	8.00	3.00
RSIVQA-Sydney	30.00	10.00	3.00	26.00	8.00

D.2 Top-k Scaling Across Retrievers

Table 6 analyzes whether weather-evidence hijacking is confined to top-1 or top-5 retrieval, or whether it continues to propagate into deeper candidate lists. Without re-optimizing the perturbation, we reuse the optimized CloudWeb query for each retriever, retrieve up to top-20 results, and then compute TopChanged@k and Weather@k at multiple retrieval depths. This analysis shows whether CloudWeb only alters a few top positions or systematically reshapes the broader retrieval neighborhood.

The results reveal three clear trends. First, Weather@k increases consistently as k grows for all five retrievers, showing that atmospheric evidence is not inserted at only one isolated rank but continues to penetrate deeper candidate lists. Second, GeoRSCLIP and OpenAI CLIP ViT-L/14 remain the strongest models under this analysis, reaching 51.00% and 50.29% Weather@20, respectively. Third, TopChanged@k approaches or reaches 100% for nearly all retrievers by top-10 or top-20, indicating that CloudWeb affects not only the leading retrieved evidence but also the larger retrieval set that a downstream RAG system could use for aggregation or reranking.

Table 6: Top-k scaling analysis across five retrievers on Sample100.

Metric	GeoRSCLIP	RemoteCLIP	OpenAI B/32	OpenAI L/14	OpenCLIP B/32
T@1	89.00%	79.86%	81.86%	81.57%	81.29%
W@1	28.14%	5.86%	7.14%	24.86%	12.71%
T@5	99.57%	97.57%	99.43%	99.43%	99.71%
W@5	41.14%	8.43%	12.71%	38.71%	19.43%
T@10	99.86%	99.14%	99.86%	99.43%	100.00%
W@10	45.57%	10.43%	16.29%	44.29%	22.29%
T@20	100.00%	99.43%	99.86%	99.71%	100.00%
W@20	51.00%	12.57%	21.57%	50.29%	29.86%

D.3 Mechanism Analysis Across Retrievers

Table 7 summarizes why CloudWeb remains effective across different CLIP-style retrievers. We report the average source-similarity decrease, average target-weather similarity gain, average weather-rank improvement, the fraction of queries for which the adversarial weather evidence enters the top-20, top-5, or top-1 retrieved set, and the median adversarial weather rank. These statistics complement the main-text retrieval metrics by showing that CloudWeb is not merely causing generic ranking noise; instead, it systematically moves the query embedding away from source-scene evidence and toward weather-related evidence.

Several consistent patterns emerge. First, all five retrievers exhibit both source suppression and target attraction, confirming that the attack works by jointly weakening the original scene semantics and strengthening atmospheric semantics in the shared image-text space. Second, GeoRSCLIP and OpenAI CLIP ViT-L/14 show the strongest weather-rank movement: they achieve the largest mean rank improvements, the highest Top20/Top5/Top1 insertion rates, and much smaller median adversarial weather ranks than the other models. This is consistent with the stronger Weather@1 and Weather@5 observed in the main experiments. Third, RemoteCLIP, OpenAI CLIP ViT-B/32, and OpenCLIP still show clear positive rank movement, but their median adversarial weather rank remains outside the top-100, indicating that successful top-ranked weather insertion happens on a more selective subset of queries. Overall, the table shows that CloudWeb does not merely perturb neighborhood order locally; it systematically shifts retrieval toward weather evidence, with the magnitude of this shift depending on the retriever's embedding geometry.

Table 7: Mechanism analysis across five retrievers on Sample100.

Metric	GeoRSCLIP	RemoteCLIP	OpenAI B/32	OpenAI L/14	OpenCLIP B/32
Mean Source Drop	0.0187	0.0242	0.0175	0.0292	0.0216
Mean Target Gain	0.0423	0.0217	0.0128	0.0231	0.0393
Mean Rank Improvement	51.98	15.44	26.52	54.88	34.17
Enter Top20	50.29%	12.00%	20.57%	49.00%	28.86%
Enter Top5	40.57%	8.00%	12.29%	38.14%	18.86%
Enter Top1	27.71%	5.29%	6.57%	24.57%	12.29%
Median Adv Weather Rank	17.5	101.0	101.0	19.0	101.0

D.4 Corpus Size Scaling Across Retrievers

Table 8 and Table 9 analyze whether a larger evidence corpus makes weather-evidence hijacking easier. We keep the optimized CloudWeb perturbation fixed for each retriever and vary only the corpus size, comparing four scales: 10k, 50k, 100k, and full. This isolates whether CloudWeb depends on a particular database size or whether targeted atmospheric evidence becomes easier to retrieve when the corpus provides more candidate weather descriptions.

The corpus-scaling results show three main patterns. First, Weather@5 increases substantially for most retrievers as the corpus grows, indicating that larger evidence pools provide richer atmospheric targets for hijacking. This trend is especially clear for GeoRSCLIP and OpenAI CLIP ViT-L/14, which rise from 19.43% to 41.14% and from 4.43% to 38.71%, respectively, when moving from the 10k subset to the full corpus. Second, RemoteCLIP and OpenCLIP also benefit from larger corpora, although their gains saturate earlier and remain lower in absolute value. Third, Top5 Changed remains high across all corpus sizes, showing that CloudWeb consistently disrupts retrieval order even when weather-targeted insertion is weaker. Overall, the analysis suggests that generic ranking disruption is stable across corpus scales, while successful weather-evidence hijacking benefits from a larger and more diverse evidence pool.

Table 8: Corpus-size scaling of Weather@5 across five retrievers.

Corpus Size	GeoRSCLIP	RemoteCLIP	OpenAI B/32	OpenAI L/14	OpenCLIP B/32
10k	19.43%	0.43%	0.00%	4.43%	1.43%
50k	35.00%	7.86%	10.86%	27.71%	20.57%
100k	40.57%	8.57%	9.43%	38.00%	19.86%
Full	41.14%	8.43%	12.71%	38.71%	19.43%

Table 9: Corpus-size scaling of Top5 Changed across five retrievers.

Corpus Size	GeoRSCLIP	RemoteCLIP	OpenAI B/32	OpenAI L/14	OpenCLIP B/32
10k	98.57%	95.00%	99.57%	99.43%	99.43%
50k	99.86%	98.29%	99.71%	98.57%	100.00%
100k	99.43%	97.86%	99.43%	99.43%	99.86%
Full	99.57%	97.57%	99.43%	99.43%	99.71%

D.5 Multi-Target Atmospheric Semantics

We further extend CloudWeb from a single cloud-oriented target to three atmospheric semantic groups: cloud, fog-haze, and smoke-mist. For each retriever, we re-optimized the perturbation under the same budget and objective setting, and then measured whether the targeted semantic group entered the retrieval results. Table 10 summarizes the results. We report Top5 Changed as a measure of retrieval-structure disruption, Target@1 and Target@5 as the targeted hit rates for the corresponding semantic group, and Avg Mask as the average perturbation coverage.

The results show a clear semantic hierarchy. Across all five retrievers, **cloud** is consistently the easiest atmospheric target to inject, while **fog-haze** is substantially weaker and **smoke-mist** remains only moderately effective. This pattern indicates that CloudWeb does not merely rely on generic semantic drift; instead, its success depends strongly on whether the target concept is compatible with the visual morphology of the perturbation. In particular, cloud-like semantics align best with the smooth, bright, and spatially diffuse structure produced by the perturbation, whereas fog- or smoke-like semantics are harder to stabilize in the retrieval space.

The strongest configuration is **OpenAI CLIP ViT-L/14 + cloud**, which reaches **Target@5 = 37.43%**. GeoRSCLIP and OpenCLIP also maintain relatively strong cloud-targeted hijacking, both reaching **Target@5 = 19.14%**. In contrast, several smoke- or fog-oriented settings still exhibit nearly saturated Top5 Changed values while producing very low Target@5. This confirms an important negative finding: changing the retrieval ranking is much easier than steering it toward a specific atmospheric concept.

Table 10: Multi-target atmospheric semantic results across five retrievers. Target@1 and Target@5 are computed with respect to the corresponding target group (cloud, fog-haze, or smoke-mist). Higher is better for all metrics.

Retriever	Target Group	Top5 Changed	Target@1	Target@5	Avg Mask
GeoRSCLIP ViT-B/32	cloud	97.43	11.71	19.14	19.12
	fog-haze	97.29	1.29	2.43	19.11
	smoke-mist	97.14	3.43	7.14	18.20
RemoteCLIP ViT-B/32	cloud	98.29	5.29	7.14	18.67
	fog-haze	98.00	0.43	0.86	18.68
	smoke-mist	98.14	0.71	1.57	18.56
OpenAI CLIP ViT-B/32	cloud	99.14	7.14	12.71	18.21
	fog-haze	99.00	0.43	1.14	17.64
	smoke-mist	99.14	0.00	0.00	16.70
OpenAI CLIP ViT-L/14	cloud	99.14	25.71	37.43	17.75
	fog-haze	99.14	1.43	4.14	16.75
	smoke-mist	99.00	0.14	1.00	16.18
OpenCLIP ViT-B/32	cloud	99.29	12.71	19.14	19.05
	fog-haze	99.29	0.57	1.57	17.69
	smoke-mist	99.14	0.29	1.14	17.83

Cross-Retriever Transferability. In addition to the per-retriever optimization setting used in the main experiments, we further evaluate a black-box transfer setting across retrievers. Specifically, we optimize the CloudWeb perturbation only on **GeoRSCLIP ViT-B/32**, and then directly apply the resulting adversarial queries to four other retrievers without any further retriever-specific optimization. Table 11 reports the corresponding results together with clean retrieval and handcrafted baselines.

The results show that CloudWeb retains non-trivial transferability across retrievers. Although the strongest performance is still achieved on the source retriever GeoRSCLIP, the transferred perturbations continue to outperform generic atmospheric baselines on all four target retrievers in terms of both retrieval disruption and weather-oriented evidence injection. For example, under transfer from GeoRSCLIP, CloudWeb reaches Weather@5 scores of 18.14%, 23.29%, 32.86%, and 26.86% on RemoteCLIP, OpenAI CLIP ViT-B/32, OpenAI CLIP ViT-L/14, and OpenCLIP ViT-B/32, respectively. This indicates that the learned perturbation is not purely tied to a single retriever, but captures cross-model atmospheric features that can generalize to unseen retrieval backbones. At the same time, the performance gap between this transfer setting and the per-retriever optimization setting highlights that black-box practicality comes with a measurable loss of attack strength.

Table 11: Cross-retriever transferability results. The CloudWeb perturbation is optimized only on GeoRSCLIP ViT-B/32 and then directly transferred to the other four retrievers without retriever-specific re-optimization. T@1/T@5 denote Top1/Top5 Changed, and W@1/W@5 denote Weather@1/Weather@5. Higher is better for all metrics.

Method	GeoRSCLIP ViT-B/32				RemoteCLIP ViT-B/32				OpenAI CLIP ViT-B/32				OpenAI CLIP ViT-L/14				OpenCLIP ViT-B/32			
	T@1	T@5	W@1	W@5	T@1	T@5	W@1	W@5	T@1	T@5	W@1	W@5	T@1	T@5	W@1	W@5	T@1	T@5	W@1	W@5
Clean	0.00	0.00	0.57	0.71	0.00	0.00	0.86	0.86	0.00	0.00	0.43	0.86	0.00	0.00	0.57	0.86	0.00	0.00	0.14	0.43
Gaussian Blur	94.00	99.71	0.43	0.86	97.43	99.86	0.43	0.71	85.86	99.43	0.29	0.57	86.43	99.29	0.29	0.57	83.14	99.71	1.14	1.57
Brightness Haze	33.29	76.71	0.57	1.14	60.29	91.71	0.86	1.00	39.86	83.86	0.43	1.14	43.43	83.86	1.00	1.14	36.00	82.86	0.29	0.71
Random Noise Cloud	66.57	93.29	3.29	5.57	70.00	95.29	1.00	1.14	68.29	95.57	2.43	4.29	70.71	97.00	10.43	15.29	63.00	95.14	1.43	3.57
Fixed CloudWeb	80.29	98.00	10.00	16.43	80.43	97.43	3.86	6.43	76.57	98.57	4.86	9.71	78.43	98.14	14.86	22.14	79.43	98.86	7.00	11.29
CloudWeb (Transferred)	89.86	99.71	29.14	43.29	83.86	97.71	12.86	18.14	85.00	99.14	13.29	23.29	79.00	98.29	20.14	32.86	82.29	98.71	17.00	26.86

E VISUAL PLAUSIBILITY AND DETECTION EVASION ANALYSIS

E.1 OPENROUTER MULTI-MODEL PERCEPTUAL AUDIT

Multi-model perceptual audit via OpenRouter. We further conducted a perceptual audit on the optimized image using a diverse set of vision-capable AI models served through OpenRouter. All models received the same forensic prompt and were required to output only a binary decision: "1" if the cloud/haze pattern appeared digitally synthesized, simulated, or artificially overlaid, and "0" if it appeared fully natural. We selected 39 candidate multimodal models from families including Qwen-VL, Mistral/Pixtral, Amazon Nova, ByteDance Seed/UI-TARS, xAI Grok, Meta Llama-Vision, Reka, MiniMax, and Kimi. Among them, 37 models returned valid binary outputs. The final vote was nearly balanced but slightly favored manipulation: 19 models predicted "1" and 18 predicted "0", corresponding to an average score of 0.5135. This result suggests that the perturbation is visually ambiguous to current multimodal models, but on aggregate it is slightly more likely to be perceived as artificial than fully natural.

Per-model decisions. The OpenRouter audit produced the following binary decisions. A value of "1" denotes perceived synthetic or manipulated atmospheric content, while a value of "0" denotes natural-looking atmospheric content. The models predicting manipulation were: qwen/qwen-vl-max, qwen/qwen3-vl-8b-instruct, qwen/qwen3-vl-8b-thinking, qwen/qwen3-vl-30b-a3b-thinking, qwen/qwen3-vl-235b-a22b-instruct, mistralai/ministral-3b-2512, mistralai/pixtral-large-2411, mistralai/mistral-medium-3, amazon/nova-pro-v1, amazon/nova-premier-v1, bytedance-seed/seed-1.6, bytedance-seed/seed-1.6-flash, bytedance-seed/seed-2.0-lite, bytedance-seed/seed-2.0-mini, x-ai/grok-4.3, meta-llama/llama-4-scout, rekaai/reka-edge, minimax/minimax-01, and moonshotai/kimi-k2.6. The models predicting natural appearance were: qwen/qwen-vl-plus, qwen/qwen2.5-vl-72b-instruct, qwen/qwen3-vl-32b-instruct, qwen/qwen3-vl-30b-a3b-instruct, qwen/qwen3-vl-235b-a22b-thinking, mistralai/ministral-14b-2512, mistralai/ministral-8b-2512, mistralai/mistral-small-3.2-24b-instruct, mistralai/mistral-medium-3.1, mistralai/mistral-large-2512, amazon/nova-2-lite-v1, amazon/nova-lite-v1, bytedance/ui-tars-1.5-7b, x-ai/grok-4, x-ai/grok-4-fast, x-ai/grok-4.1-fast, meta-llama/llama-3.2-11b-vision-instruct, and meta-llama/llama-4-maverick. Taken together, these results give an overall OpenRouter manipulation perception score of $19/37 = 0.5135$, suggesting that the optimized atmospheric perturbation is not consistently identified as artificial by current vision-language models.

E.2 HUMAN PERCEPTUAL AUDIT

We further conducted a small-scale human perceptual audit on the same optimized image. A total of 30 human evaluators were asked to make the same binary judgment: "1" if the cloud/haze pattern appeared digitally synthesized, simulated, or artificially overlaid, and "0" if it appeared fully natural. Among the 30 responses, 21 evaluators selected "1" and 9 selected "0". This corresponds to a mean score of 0.7000, indicating that human observers were more likely than the OpenRouter model ensemble to regard the perturbation as artificial rather than fully natural.

Compared with the OpenRouter-based multi-model audit, the human study gives a stronger manipulation judgment. Specifically, the model ensemble produced a nearly balanced decision distribution, with 19 manipulation votes and 18 naturalness votes, whereas the human audit produced a clearer manipulation tendency, with 21 manipulation votes out of 30 total responses. This suggests that the perturbation remains visually ambiguous to current VLMs but is more likely to be perceived as artificial by human observers.

Human-audit protocol. Participants were asked to inspect the optimized image and answer a single binary question: output "1" if the cloud/haze pattern appeared digitally synthesized, simulated, or artificially overlaid, and output "0" if it appeared fully natural. No personal or sensitive information was collected, responses were aggregated anonymously, and the task involved only visual perception of a remote-sensing image. No compensation was provided for this small internal perceptual audit.

F Defense Implications and Embedding Mechanism Analysis

F.1 Why low-frequency atmospheric perturbations shift CLIP embeddings.

We provide a mathematical interpretation of why cloud- and haze-like perturbations are more effective than high-frequency pixel noise in cross-modal retrieval. Let $f_I(\cdot)$ and $f_T(\cdot)$ denote the normalized image and text encoders of a CLIP-style retriever, respectively. For a weather-related evidence text t_w, the retrieval score of a query image $\mathbf{q}$ is

$$s_w(\mathbf{q}) = \langle f_I(\mathbf{q}), f_T(t_w) \rangle . \tag{18}$$

After adding an image-side perturbation $\boldsymbol{\delta}$, the first-order change of this score can be approximated as

$$\Delta s_w = s_w(\mathbf{q}+\boldsymbol{\delta}) - s_w(\mathbf{q}) \approx \langle J_{f_I}(\mathbf{q})\boldsymbol{\delta}, f_T(t_w) \rangle = \langle \boldsymbol{\delta}, \nabla_{\mathbf{q}} s_w(\mathbf{q}) \rangle , \tag{19}$$

where $J_{f_I}(\mathbf{q})$ is the Jacobian of the image encoder. This approximation shows that the effectiveness of a perturbation is not determined only by its magnitude, but by its alignment with the semantic gradient induced by the target weather text. High-frequency pixel noise usually contains rapidly oscillating components with unstable phases, so its inner product with the weather-semantic gradient tends to cancel out across pixels and patches. In contrast, clouds, fog, and haze introduce spatially coherent low-frequency changes, such as soft occlusion, reduced contrast, bright atmospheric veils, and large-scale visibility degradation. These structures are better aligned with the visual patterns that CLIP-style models associate with weather-related text.

This difference can also be interpreted in the frequency domain. Decomposing the perturbation into Fourier bases gives

$$\boldsymbol{\delta} = \sum_{\omega} \widehat{\boldsymbol{\delta}}(\omega)\phi_\omega, \tag{20}$$

and the score variation can be written as

$$\Delta s_w \approx \sum_{\omega} \widehat{\boldsymbol{\delta}}(\omega)\overline{\widehat{\nabla_{\mathbf{q}} s_w}(\omega)}. \tag{21}$$

If the sensitivity of the weather score is concentrated in low-frequency bands, then perturbations whose spectral energy also lies in these bands will produce a larger semantic displacement. High-frequency noise may strongly change the local pixel distribution and even disrupt nearest-neighbor retrieval, but it does not necessarily move the query embedding toward the weather-evidence region. This explains the empirical gap between generic retrieval disruption and targeted weather-evidence hijacking: a perturbation can achieve high Top-k Changed while still failing to increase Weather@k.

From a manifold perspective, natural remote-sensing images can be viewed as lying near an image manifold $\mathcal{M}$. Source-scene semantics, such as rivers, airports, or residential areas, occupy local regions on this manifold, while weather-related visual conditions form another semantic direction corresponding to cloudy, foggy, hazy, or low-visibility scenes. High-frequency noise often pushes the image away from the natural image manifold,

$$\mathbf{q} + \boldsymbol{\delta}_{\text{noise}} \notin \mathcal{M}, \tag{22}$$

causing unstable or non-semantic embedding changes. By contrast, atmospheric perturbations remain close to a plausible natural-image manifold,

$$\mathbf{q} + \boldsymbol{\delta}_{\text{atm}} \approx \mathcal{M}_{\text{scene-weather}}, \tag{23}$$

where the image still resembles a valid remote-sensing scene but under cloudy or low-visibility conditions. Since such visual patterns are likely to co-occur with weather descriptions during image-text contrastive pretraining, the perturbed image is encoded as a semantically meaningful weather variant rather than as arbitrary noise.

Therefore, CloudWeb induces a directional semantic shift in the shared embedding space:

$$f_I(\mathbf{q} + \boldsymbol{\delta}_{\text{atm}}) \approx f_I(\mathbf{q}) + \alpha \mathbf{v}_{\text{weather}} - \beta \mathbf{v}_{\text{source}}, \tag{24}$$

where $\mathbf{v}_{\text{weather}}$ denotes the direction toward atmospheric evidence and $\mathbf{v}_{\text{source}}$ denotes the direction associated with the original source-scene evidence. This formulation is consistent with our mechanism analysis: CloudWeb simultaneously increases similarity to target weather evidence, suppresses similarity to source-scene evidence, and moves weather evidence to higher retrieval ranks. In this sense, low-frequency atmospheric perturbations are effective not because they merely corrupt the image, but because they lie near the weather-semantic manifold learned by CLIP-style vision-language retrievers.

F.2 Responsible Use and Defensive Guidelines

Intended use. CloudWeb is intended to support robustness evaluation for remote-sensing multimodal RAG systems. The goal is to identify a pre-generation failure mode in which visually plausible atmospheric changes redirect evidence retrieval before the VLM produces an answer. We do not advocate using the method to manipulate operational remote-sensing systems.

Risk boundary. The attack is evaluated under a query-side threat model: the attacker can modify the input image but cannot change the retriever, the generator, the FAISS index, or the evidence corpus. This setting is relevant to deployed RAG systems because the retrieval interface is often exposed to user-provided images. However, the same constraint also makes the risk analyzable: the system can monitor input-side atmospheric patterns, retrieval instability, and evidence-type shifts without modifying the underlying model.

Potential defenses. The results suggest several mitigation directions.

Atmospheric-risk screening. Before retrieval, the system can estimate cloud, haze, or low-visibility severity using either a lightweight segmentation model or simple image statistics such as low-frequency energy, contrast attenuation, and bright veil coverage. High-risk inputs should not be rejected by default, since real remote-sensing images often contain clouds. Instead, they should trigger a stricter retrieval verification pipeline.

Consistency-aware retrieval. A robust system can retrieve evidence from multiple benign variants of the same query image, including the original image, a contrast-normalized image, a dehazed image, and resized or mildly blurred variants. Let $\mathcal{R}_k(\mathbf{q})$ denote the top-k retrieved evidence set for query image $\mathbf{q}$. A simple instability score can be defined as

$$\mathrm{Instab}(\mathbf{q}) = 1 - \frac{|\mathcal{R}_k(\mathbf{q}) \cap \mathcal{R}_k(\mathcal{T}(\mathbf{q}))|}{|\mathcal{R}_k(\mathbf{q}) \cup \mathcal{R}_k(\mathcal{T}(\mathbf{q}))|}, \tag{25}$$

where $\mathcal{T}$ is a benign image transformation. A high instability score, together with a sudden increase in weather-related evidence, can indicate possible retrieval hijacking.

Evidence-type gating. Weather-related evidence should be downweighted when it is not supported by scene-consistent evidence. For example, if the retrieved context contains cloudy or hazy descriptions but lacks agreement on the underlying scene category, the generator should receive a warning or a reranked evidence set. This prevents the generator from treating isolated weather evidence as reliable grounding.

Scene-weather disentangled reranking. A reranker can separately score source-scene consistency and weather consistency. Given an image query $\mathbf{q}$, source evidence e_s, and weather evidence e_w, the final retrieval score can combine the original CLIP score with a scene-consistency term:

$$S(e \mid \mathbf{q}) = \mathrm{sim}(f_I(\mathbf{q}), f_T(e)) + \lambda_s C_{\mathrm{scene}}(\mathbf{q}, e) - \lambda_w C_{\mathrm{unsupported\text{-}weather}}(\mathbf{q}, e). \tag{26}$$

This does not remove weather evidence entirely; it only penalizes weather evidence that is not corroborated by other scene-grounded neighbors.

Robust retriever training. Retriever finetuning can include natural atmospheric augmentations and hard negative weather evidence. The training objective should encourage images with clouds, fog, or haze to remain close to their true scene descriptions unless the weather condition is genuinely part of the evidence. This directly targets the semantic shortcut exploited by CloudWeb.

Reporting practice. When releasing code or examples, we recommend providing evaluation scripts, aggregate metrics, and defensive checks rather than emphasizing step-by-step misuse instructions. The released examples should be framed as robustness tests for authorized systems.

G QUALITATIVE CASE GALLERY

Case Gallery The following gallery presents one representative example from each dataset. For each case, we show the original and adversarial query images together with the clean and adversarial outputs of six downstream models. This appendix makes the qualitative propagation effect concrete: after retrieval-stage evidence hijacking, multiple generators shift from source-scene descriptions or answers toward weather-related semantics, uncertainty, or semantically displaced responses.

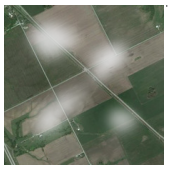
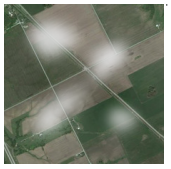
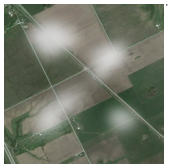
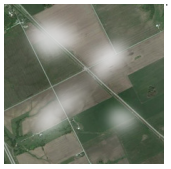
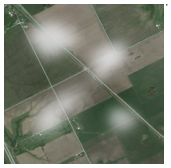
NWPU-RESISC45

Prompt: Is this image more likely to show meadow or airplane?

Model	Ori. Image	Adv. Image	Ori. Output	Adv. Output
LLaVA-1.5			The image is more likely to show meadow.	Airplane
LLaVA-1.6			The image is more likely to show an airplane, as there are three airplanes visible in the image.	The image is more likely to show an airplane, as it features three airplanes on a runway.
InstructBLIP			airplane	tarmac
Qwen2.5-VL			The image is more likely to show airplanes, as indicated by the presence of multiple aircraft on the tarmac, rather than a meadow, which would not typically feature such objects.	The image is more likely to show airplanes, as it depicts multiple aircraft with visible spray trails, which is not characteristic of a meadow.
GeoChat			The image is more likely to show meadow as the grass is cleared from the meadow and a house appears at the bottom of the screen.	The image is more likely to show airplane as there are multiple airplanes present in the image.
H2RSVLM			meadow	airplane

RSICD

Prompt: What type of remote sensing scene is shown in this image?

Model	Ori. Image	Adv. Image	Ori. Output	Adv. Output
LLaVA-1.5			A green farmland with two roads crossing with each other separate this area.	A remote sensing scene of a farm with a pond and a road.
LLaVA-1.6			The image shows a green farmland with roads crossing and running through it, indicating a rural area with agricultural land.	The image shows a remote sensing scene of a farmland with a blue pond, crossroads, and clouds.
InstructBLIP			aerial view	The image is a Google Earth image of a road in the United States.
Qwen2.5-VL			The image shows an agricultural remote sensing scene with intersecting roads dividing green farmlands into sections.	The image depicts an agricultural landscape with intersecting roads dividing farmland, an irregular blue pond surrounded by farmlands, and clouds floating above the area, indicating a typical remote sensing scene of a rural region.
GeoChat			The image shows a remote sensing scene of a road crossing through a green farmland with two roads next to it.	The image shows a remote sensing scene of a light green farmland with two crossroads and clouds over it.
H2RSVLM			The image shows a rural area with green farmlands, roads, and buildings.	The image shows a rural area with farmlands, a pond, and clouds.

RSVQA-LR

Prompt: What is the number of buildings?

Model	Ori. Image	Adv. Image	Ori. Output	Adv. Output
LLaVA-1.5			There are many buildings in the area.	There are no buildings in the image.
LLaVA-1.6			The image does not provide an exact number of buildings, but it shows a significant number of buildings, including black, white, and brown structures, surrounded by dark green trees and meadows.	The image does not provide enough information to determine the number of buildings.
InstructBLIP			brown	The clouds are a result of the weather
Qwen2.5-VL			The image shows numerous buildings, including black, white, brown structures, as well as grey roads and green trees and meadows surrounding a large green center.	There are no buildings visible in the image; it primarily shows cloud cover with some shadowed areas and irregular bright regions.
GeoChat			There are 1 building present in the image.	There are 2 buildings present in the image.
H2RSVLM			There are 10 buildings in the image.	There are no buildings in the image.

LEVIR-CC

Prompt: Describe the land-cover change between the two remote sensing images in one concise sentence.

Model	Ori. Image	Adv. Image	Ori. Output	Adv. Output
LLaVA-1.5			A new road is built through the woods in the right.	The land-cover change between the two remote sensing images is a cloudy sky over a road.
LLaVA-1.6			The land-cover change between the two remote sensing images includes the construction of a new road through the woods, which is visible in the right image.	The land-cover change between the two remote sensing images is the presence of a branch river pours into the river over which is a piece of cloud, while the previous image shows white spray stirred at the turning area of the river.
InstructBLIP			is this a road?	The image shows a road with a cloudy sky and a hill in the background. The cloud may be caused by a storm or a haze.
Qwen2.5-VL			A new road has been constructed through the wooded area, opening up a branch road and altering the land cover.	The land-cover change between the two remote sensing images shows a transition from a forested area with a visible road to an area covered by clouds, indicating a shift from terrestrial to atmospheric conditions.
GeoChat			A new road is built through the woods in the right, and a road is built through the woods in the center.	The land-cover change between the two remote sensing images is not specified in the given information.
H2RSVLM			The land-cover change between the two remote sensing images is a road being built through the woods.	The cloud cover has increased, and the river has changed direction.

FloodNet

Prompt: How many buildings are in the image? Answer the question using a single word or phrase.

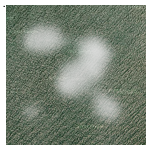
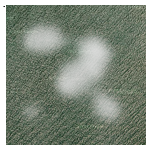
Model	Ori. Image	Adv. Image	Ori. Output	Adv. Output
LLaVA-1.5			3	4
LLaVA-1.6			3	4
InstructBLIP			The aerial view of the property shows a red truck driving around the property.	The image of the letter S is taken from an aerial view and is a combination of light and shadow.
Qwen2.5-VL			There are several buildings visible in the lower-right corner of the image.	There are no buildings in the image; instead, there is a white cloud-like shadow covering the area where buildings would typically be located.
GeoChat			1 building	There are 4 buildings in the image.
H2RSVLM			There are three buildings in the image.	There are four buildings in the image.

RSIVQA-UCM

Prompt: What is the location of the intersection in this picture?

Model	Ori. Image	Adv. Image	Ori. Output	Adv. Output
LLaVA-1.5			The intersection is located in the middle of a city street.	The intersection is located in a residential area.
LLaVA-1.6			The intersection is located in the confluence zone, where the rails are coming out of the huge grey rectangular ceiling.	The intersection is located in a residential area, under a white overpass, with a forked highway and grey dust.
InstructBLIP			intersection	traffic lights
Qwen2.5-VL			The intersection in the picture is located at the confluence zone where the rails from the large grey rectangular ceiling emerge, with two straight double lines on either side and an arc double line beside them, indicating a complex traffic pattern with multiple lanes and merging vehicles.	The intersection in the picture appears to be located in a residential area with a forked highway, as indicated by the presence of a white overpass and the soft texture of the cloud cover, which does not match the features of a runway, building blocks, or water body.
GeoChat			The intersection is located at the center of the image, where two roads meet.	The intersection is located at the bottom of the image.
H2RSVLM			The intersection is located in the center of the image.	The intersection is located in a residential area with a forked highway.

RSIVQA-Sydney

Prompt: what is the theme of this picture?				
Model	Ori. Image	Adv. Image	Ori. Output	Adv. Output
LLaVA-1.5			The theme of this picture is a lush green field with a vertical texture.	The theme of this picture is a green field with white clouds and water.
LLaVA-1.6			The theme of this picture is the contrast between the linear patterns on the green grass and the lush grass itself.	The theme of this picture is the interaction between water and the environment, specifically the impact of wind on seawater.
InstructBLIP			sand	fog
Qwen2.5-VL			The theme of this picture appears to be an agricultural field with lush grass, as indicated by the linear patterns and vertical textures observed in the satellite imagery.	The theme of this picture appears to be the interaction between seawater and wind, creating layers of white spray on a green surface.
GeoChat			The theme of this picture is a mix of different types of vegetation, including trees, grass, and a mossy area, with some of them having a linear pattern.	The theme of this picture is a coastal or waterfront scene with multiple ships and harbors.
H2RSVLM			The image depicts a verdant grass field with a linear pattern running through it, suggesting a focus on the natural beauty and serenity of the landscape.	The image depicts a scene of white clouds and seawater in the wind, suggesting a weather-related theme.

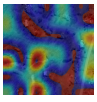
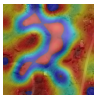
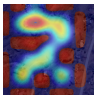
H EXTENDED CAM VISUALIZATION

Extended CAM Visualization Across Diverse Scenes The following cases provide additional prompt-conditioned CAM examples across multiple queries and datasets. To avoid ambiguity, we clarify that the texts shown in the "Source text" and "Target text" columns are not generative prompts, but semantic probes used to measure image-text alignment inside the retriever. Given an image and a text query, we compute a ViT-compatible class activation map (CAM) with respect to the image-text matching score, so that the resulting heatmap highlights which spatial regions most support the queried semantics. Therefore, these CAMs should be interpreted as visual grounding maps for retrieval semantics, rather than as explanations of free-form text generation.

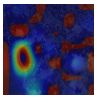
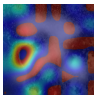
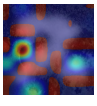
The three CAM columns in the cases below serve different roles. "Clean CAM" is computed from the clean image under the source-scene text, and shows how the retriever normally grounds the original scene semantics. "Adv CAM (Source)" is computed from the adversarial image under the same source text, and is used to test whether the original scene grounding is weakened after applying CloudWeb. "Adv CAM (Target)" is computed from the same adversarial image but under the weather-related target text, and is used to test whether the perturbation creates stronger support for target atmospheric semantics. We explicitly show two adversarial CAMs because a single adversarial heatmap would only indicate that the attention pattern changes, but would not reveal whether the change is merely noisy or is directionally shifted from source-scene semantics toward target weather semantics.

Across the cases below, a consistent pattern can be observed. In the clean setting, the retriever attends to source-relevant structures, such as residential layouts, ponds, grass regions, roads, railways, and overpasses, and the retrieved evidence remains aligned with the original scene content. After CloudWeb is applied, the global scene layout is still recognizable, but the source-conditioned adversarial CAM becomes more diffuse, displaced, or fragmented, indicating weaker grounding in the original semantics. In contrast, the target-conditioned adversarial CAM exhibits stronger responses around cloud-like, haze-like, mist-like, or fog-like regions, while the retrieved evidence simultaneously shifts toward weather-related descriptions. This pattern supports the interpretation that CloudWeb does not merely occlude the image or randomly disturb attention; instead, it constructs plausible atmospheric visual evidence that systematically attracts the retriever toward weather semantics.

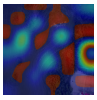
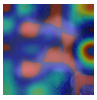
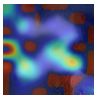
Case 1: nwpu_medium_residential_016331_scene_00 (GeoRSCLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					a new road and many arranged houses show up around the woods .	Evidence: The scene should be interpreted as cloud when overhead visual features include white cloud cover, soft texture, and shadowed cloud and do not primarily match runway, building blocks, and water body.

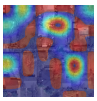
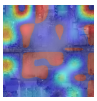
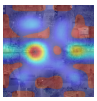
Case 2: nwpu_wetland_031483_scene_00 (GeoRSCLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					is a irregular circular pond with moss green water in the middle .	a piece of water and some clouds are in a piece of green forest .

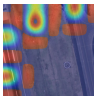
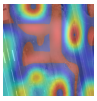
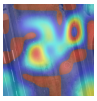
Case 3: floodnet_test_003389_vqa_00 (GeoRSCLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					the grass on the upper left becomes lush and paths and a house appear on the right .	the bareland with red stripes has a white fog like shadows .

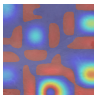
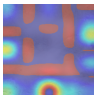
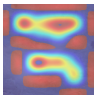
Case 4: rsivqa_ucm_001139_vqa_00 (GeoRSCLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					trees disappear and some houses are built on both sides of the scene .	many railways cross the city that surrounded the mist.

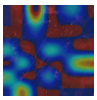
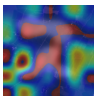
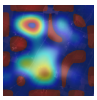
Case 5: rsivqa_sydney_000550_vqa_00 (GeoRSCLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					there are overpass and grass .	the viaduct glued in the mist is majestic and animated.

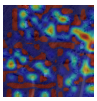
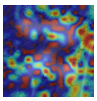
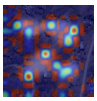
Case 6: rsicd_valid_000111_scene_00 (GeoRSCLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					we can see a construction site in this bleak desert .	Evidence: The scene should be interpreted as cloud when overhead visual features include white cloud cover, soft texture, and shadowed cloud and do not primarily match runway, building blocks, and water body.

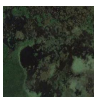
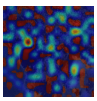
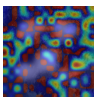
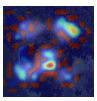
Case 7: rsvqa_lr_test_329_32979 (GeoRSCLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					the dark green farmland which loos like the moss is between a light green farmland and some white ones .	Several ponds are surrounded by clouds and green trees.

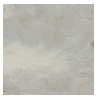
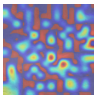
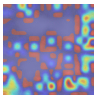
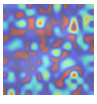
Case 8: nwpu_medium_residential_016331_scene_00 (OpenAI CLIP ViT-L/14).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					several new roads with houses appear on the left half of the image .	The image evidence for cloud usually includes repeated or salient structures related to white cloud cover, soft texture, shadowed cloud, and irregular bright regions, making the scene recognizable from overhead view.

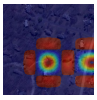
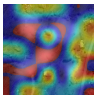
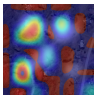
Case 9: nwpu_wetland_031483_scene_00 (OpenAI CLIP ViT-L/14).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					bacteria colony like green points are scattered on the wasteland .	Evidence: The image is likely to show cloud because it contains cues such as white cloud cover, soft texture, shadowed cloud, and irregular bright regions.

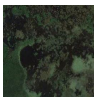
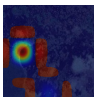
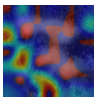
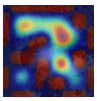
Case 10: rsicd_valid_000111_scene_00 (OpenAI CLIP ViT-L/14).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					dark stains can be seen in this desert .	a cloudy remote sensing image

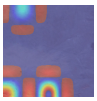
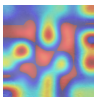
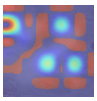
Case 11: nwpu_medium_residential_016331_scene_00 (OpenCLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					several scattered houses and cement roads appear on bareland .	The image evidence for cloud usually includes repeated or salient structures related to white cloud cover, soft texture, shadowed cloud, and irregular bright regions, making the scene recognizable from overhead view.

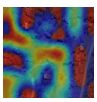
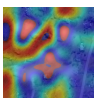
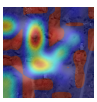
Case 12: nwpu_wetland_031483_scene_00 (OpenCLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					some cars and the bare earth are around an irregular dark green pond.	The image evidence for cloud usually includes repeated or salient structures related to white cloud cover, soft texture, shadowed cloud, and irregular bright regions, making the scene recognizable from overhead view.

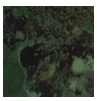
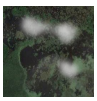
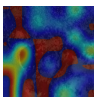
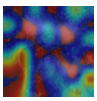
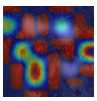
Case 13: rsicd_valid_000111_scene_00 (OpenCLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					there is a grey site located in the khaki desert .	The main objects or components in a cloud scene include white cloud cover, soft texture, shadowed cloud, irregular bright regions, and obscured ground.

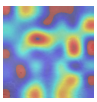
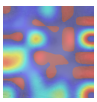
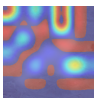
Case 14: nwpu_medium_residential_016331_scene_00 (RemoteCLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					many trees have been removed and some villas are scattered .	a cloudy remote sensing image

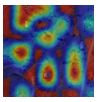
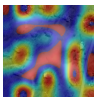
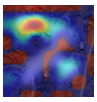
Case 15: nwpu_wetland_031483_scene_00 (RemoteCLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					many green lands and dark green plant in it.	a piece of water and some clouds are in a piece of green wood.

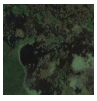
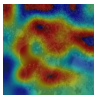
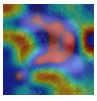
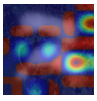
Case 16: rsicd_valid_000111_scene_00 (RemoteCLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					the off white desert has some relic .	there is a bareland covered by white snow .

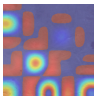
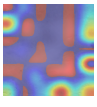
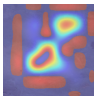
Case 17: nwpu_medium_residential_016331_scene_00 (OpenAI CLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					In aerial or satellite imagery, a mobile home park scene represents urban residential land cover or infrastructure where mobile homes, regular rows, and small rectangular units are the dominant cues.	a cloudy remote sensing image

Case 18: nwpu_wetland_031483_scene_00 (OpenAI CLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					this is a satellite imagery of a place .	a piece of water and some clouds are in a piece of green forest .

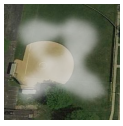
Case 19: rsicd_valid_000111_scene_00 (OpenAI CLIP ViT-B/32).

Clean	Adv	Clean CAM	Adv CAM (Source)	Adv CAM (Target)	Source text	Target text
					a marble pattern can be seen on the orange and khaki desert.	a cloudy remote sensing image

I TYPICAL FAILURE CASES

We include representative failure cases to clarify the current limits of CloudWeb. These are not degenerate cases where optimization collapses completely. Instead, they are harder negatives in which the retrieval ranking often changes, but weather-related evidence still fails to enter the top retrieved set. Across the examples below, three recurring patterns appear: strong geometric source anchors, reflection-like local optima, and unusually stable source manifolds that keep retrieval inside the original semantic neighborhood.

NWPU-RESISC45

Query ID: nwpu_baseball_diamond_001741_landuse_00				
Prompt: What is the main land-use type of this satellite image?				
Clean Image	Adv. Image	Top-1 Changed / Top-5 Changed	Clean Retrieved Top-3	Adv. Retrieved Top-3
		Yes / Yes	[1] there's a yellow baseball field in the middle of the lawn . [2] There's a yellow baseball field in the middle of the lawn. [3] There's a yellow baseball field in the green grass.	[1] there is a baseball field with some reflections of lamp posts . [2] the baseball field with some reflections of lamp posts is next to two buildings . [3] a white bubble shape church is in the corner of a neighbor which is divided by some streets .

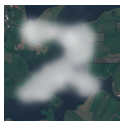
Failure analysis. The baseball-diamond geometry remains a strong semantic anchor. CloudWeb changes the ranking, but the perturbed retrieval still stays inside the sports-field manifold instead of crossing into weather semantics.

RSICD

Query ID: rsicd_test_000224_evidence_00				
Prompt: Using the retrieved evidence, describe the main scene and objects in this remote sensing image.				
Clean Image	Adv. Image	Top-1 Changed / Top-5 Changed	Clean Retrieved Top-3	Adv. Retrieved Top-3
		Yes / Yes	[1] the bridge built across the river has three directions, of which the right lane has a traffic jam . [2] a big bridge across a broad river and many cars on it . [3] this bridge is very strong .	[1] overpass is white . [2] a light white is surrounded by the freeway and paking lots . [3] a viaduct divided into a straight light ray way and some arc light gray way .

Failure analysis. The attack suppresses the detailed bridge interpretation and shifts retrieval toward a more generic overpass description, but it still does not cross the semantic boundary into cloud-like evidence.

RSVQA-LR

Query ID: rsvqa_lr_test_263_26399				
Prompt: What is the amount of roads on the left of a building?				
Clean Image	Adv. Image	Top-1 Changed / Top-5 Changed	Clean Retrieved Top-3	Adv. Retrieved Top-3
		Yes / Yes	[1] it is dark blue lake and green grassland . [2] some lakes appear on the bareland . [3] the lake is green above a lot of ship .	[1] there is a big reflection in the sunlight . [2] a pond with sky inverted image while surrounded by many spring green plants . [3] sunlight shining on the potholes in the water .

Failure analysis. The perturbation is absorbed into a reflection-like interpretation. Instead of retrieving cloud or fog evidence, the model shifts toward bright-surface or sunlight-reflection descriptions.

FloodNet

Query ID: floodnet_test_002797_vqa_00				
Prompt: What is the overall condition of the given image? Is it flooded? Choose from: - Yes - No Answer the question using a single word or phrase.				
Clean Image	Adv. Image	Top-1 Changed / Top-5 Changed	Clean Retrieved Top-3	Adv. Retrieved Top-3
		Yes / Yes	[1] two cars are in a river with some green trees on two ends of it. [2] the long river is full of grass . [3] The long river is full of grass.	[1] sunlight shining on the potholes in the water . [2] a room appears on the upper left of the screen . [3] it is white, light green and green .

Failure analysis. Water regions and specular highlights create a strong competing explanation. The perturbation changes retrieval, but the new evidence is explained as water-surface reflection rather than atmospheric cloud or haze.

RSIVQA-UCM

Query ID: rsivqa_ucm_000762_vqa_00				
Prompt: Does this picture contain trees?				
Clean Image	Adv. Image	Top-1 Changed / Top-5 Changed	Clean Retrieved Top-3	Adv. Retrieved Top-3
		Yes / Yes	[1] there are many cars running on the overpass . [2] many cars were running on the overpass . [3] many cars were running on the overpass .	[1] overpass is white . [2] a light white is surrounded by the freeway and paking lots . [3] the car in the parking lot disappears and several buildings appear on the lower-right .

Failure analysis. Linear man-made infrastructure remains dominant after perturbation. Retrieval moves from detailed traffic semantics to a coarser overpass description, but the structural prior prevents weather-evidence insertion.

RSIVQA-Sydney

Query ID: rsivqa_sydney_000374_vqa_00				
Prompt: what is the theme of this picture?				
Clean Image	Adv. Image	Top-1 Changed / Top-5 Changed	Clean Retrieved Top-3	Adv. Retrieved Top-3
		No / Yes	[1] the green of the sea rolled with white waves. [2] the green of the sea rolled up white waves . [3] The white waves are in the green water.	[1] the green of the sea rolled with white waves. [2] The white waves are in the green water. [3] the green of the sea rolled up white waves .

Failure analysis. This is a harder failure in which even the top retrieval remains nearly unchanged. The sea-wave texture already contains strong white-pattern cues, yet the model still prefers the original ocean interpretation.