Breaking the Weakest Link to Evade Vision–Language Models

Ilan Zini[1], Boussad Addad[2], and Katarzyna Kapusta[2]
[1] ESILV student and Thales intern, ilanzini.pro@gmail.com
[2] Thales, Thales cortAIx Labs, France, {boussad.addad, katarzyna.kapusta}@thalesgroup.com

Identification of Use Cases

Vision–Language Models (VLMs) are rapidly becoming a core component of multimodal AI systems, combining a visual encoder with a large language model to jointly reason over imagery and text. In a defence context, this capability supports emerging applications in intelligence, surveillance and reconnaissance (ISR), multi-sensor data fusion, automated scene description and target recognition, mission support, and decision aids for command and control (C2). Such systems are increasingly expected to operate on imagery acquired by drones, satellites, embedded sensors or open sources, and to turn it directly into actionable textual interpretations.

The capability to improve is the trustworthiness and robustness of these systems when their visual inputs cannot be assumed to be reliable. The problem we address is that the robustness of VLMs against adversarial evasion remains poorly characterised: an adversary can apply small, human-imperceptible perturbations to an image and silently steer the model's textual output. This directly threatens safety-critical and contested operations. In fact, a hostile asset could be described as a civilian vehicle, or a benign scene reframed as a threat. The consequences range from biased or misleading decision support to large-scale disinformation, all while the system appears to behave normally.



Figure 1: Source image (left) and adversarial image (right) on Granite-Vision-3.2-2B ($\varepsilon = 0.05$). The perturbation is imperceptible, yet the model's caption switches from "a military tank" to "an ambulance".

This work is therefore positioned as a trusted-AI and resilience-in-contested-environments capability. The intended users are: defence AI developers and integrators who embed VLMs into ISR or autonomy pipelines and need to assess their exposure before fielding; evaluation, qualification and certification authorities (e.g. AMIAD) seeking a concrete, repeatable way to audit the adversarial robustness of multimodal models; and red teams and operational units that must understand how VLM-based tools can fail under deliberate manipulation.

The use case is directly transposable from the civilian domain: our experiments rely on civilian imagery (ImageNet) and reproduce a vulnerability already demonstrated on autonomous-driving VLMs, where adversarial perturbations cause real trajectory deviations. The same mechanism applies to defence-grade imagery analysis. The natural next step is to integrate the method as a standardised robustness-audit harness within an AI qualification process, evaluated on defence-representative datasets and sensor modalities.

Proposed Technical Approach

We propose a gradient-based evasion attack on the visual input of a VLM. The central methodological choice is to optimise the perturbation *exclusively through the vision encoder*, the "weakest link", rather than backpropagating through the full multimodal architecture as in prior work. Because the vision encoder accounts for only 1.6% to 18% of total model parameters (and as little as 0.9% for larger LLM backbones sharing the same encoder), this design drastically reduces the computational cost and resource footprint of the attack while preserving strong effectiveness. We assume a white-box threat model with access to the vision encoder, which is realistic given that many competitive VLMs are released as open source.

Two settings are addressed. In the **targeted** attack, the perturbation is optimised to minimise the mean squared error (MSE) between the adversarial image embedding and a chosen target image embedding, forcing the model toward a specific description. In the **untargeted** attack, the optimisation starts from a lightly noised image and maximises the divergence from the original embedding (negated MSE), simply disrupting the model's interpretation. Perturbations are generated with I-FGSM over 50 iterations under a bounded budget ε.

The attack exploits the differentiable nature of the vision encoder: because gradients can be propagated back to the input pixels, perturbations are optimised with iterative gradient sign descent (I-FGSM), and attack success is scored at scale through an automated LLM-as-a-Judge protocol (Granite-4.0-micro). A key technical barrier is the heterogeneity of vision-encoder architectures across VLMs: they are not all Vision Transformers (ViT), FastVLM for instance relying on a hybrid convolution-transformer encoder, so layer access and projection modules differ and the attack must be adapted per model. The other barrier, the prohibitive cost of full-pipeline attacks, is removed by the encoder-only formulation.

Validation. We evaluated the attack on four open-source VLMs over 1000 image pairs, on a single NVIDIA H100 (80 GB). The headline results below are mean success rates over four runs:

Table 1: Attack success rate (%) across four open-source VLMs.

VLM	**Untargeted** ($\varepsilon = 0.05$)	**Targeted** ($\varepsilon = 0.10$)
Qwen2.5-VL-3B	93.7%	20.9%
Granite-Vision-3.2-2B	99.8%	47.1%
FastVLM-7B	99.8%	27.7%
Phi-3.5-Vision	93.0%	1.6%

Untargeted attacks are extremely effective: every model exceeds 93% success even at the smallest perturbation, and Granite-Vision and FastVLM reach 99–100%, with the effect saturating at low ε. Targeted attacks are far more demanding. Success varies widely across architectures, with Phi-3.5-Vision standing out as a robust outlier ($\sim$2%). The key finding is that disrupting a model's semantic interpretation is fundamentally easier than steering it toward a chosen output, and that no evaluated model offers meaningful resistance to untargeted manipulation.

Beyond attack effectiveness, we evaluated the computational footprint on Qwen2.5-VL and Granite-Vision, the two models with the highest vision-encoder-to-VLM parameter ratios. On an NVIDIA H100 80 GB, encoder-only optimization reduces VRAM usage from approximately 44–47% to 13–16% and reduces GPU utilization. This translates into a major reduction in attack generation time: on Qwen2.5-VL, more than 20 minutes were required with the full VLM without success, compared with approximately 100 seconds using our encoder-based method.

Maturity level: research prototype, validated experimentally on four representative open-source VLMs. The next steps are to characterise cross-architecture transferability (toward black-box settings) and to investigate defence mechanisms, with a view to turning the method into a robustness-qualification tool for trusted multimodal AI.