Visual Injection: Breaking and Fixing Safety Alignment in Vision-Language Models

Abishek Kumar Giri

Department of Computer Science, Stockton University
giri@stockton.edu

Abstract

Visual Large Language Models (VLMs) have demonstrated remarkable capabilities in multimodal understanding; however, their alignment vulnerabilities remain under-explored. In this work, we present a targeted adversarial attack on `LLaVA-1.5-7B`, a state-of-the-art open-source VLM. By employing Projected Gradient Descent (PGD) on the vision encoder's embedding space, we demonstrate that imperceptible adversarial perturbations can override the model's visual perception entirely. Our experiments on a subset of the COCO validation dataset ($N = 50$) reveal a critical vulnerability: the attack achieves a **100% Attack Success Rate (ASR)**, forcing the model to output the target token "Hacked" regardless of the actual visual content. Furthermore, the optimization converges rapidly, requiring an average of only **46.5 iterations** to break the model's safety alignment. Finally, we demonstrate that this fragility can be mitigated using simple input transformations; specifically, JPEG compression ($Q = 50$) restores the model's baseline behavior in 100% of cases, effectively neutralizing the attack.

1 Introduction

The convergence of Computer Vision and Natural Language Processing has led to the rise of Vision-Language Models (VLMs) capable of complex reasoning [1]. These models are increasingly deployed in real-world applications ranging from autonomous agents to content moderation. As their capabilities grow, so does the necessity for safety alignment—ensuring these models refuse to generate dangerous instructions, hate speech, or private information.

Current safety paradigms heavily rely on Reinforcement Learning from Human Feedback (RLHF) applied to the language component of the model [2]. However, VLMs introduce a new attack surface: the visual encoder (e.g., CLIP or SigLIP). Unlike the text decoder, these visual encoders are often trained on massive, uncurated datasets and lack explicit safety constraints. This creates a "backdoor" into the model's reasoning process.

In this paper, we investigate the following research question: *Can adversarial visual inputs override the safety alignment of a frozen LLM?* We hypothesize that by optimizing the pixel values of an input image using Projected Gradient Descent (PGD), we can generate high-norm visual tokens that force the LLM into a compliant state, effectively bypassing its textual safety filters.

We validate this hypothesis on the `LLaVA-1.5-7B` architecture. We demonstrate that a visually imperceptible perturbation added to a benign image (e.g., a photograph of a dog) can coerce the model into generating arbitrary target strings. We further evaluate a defensive strategy, analyzing whether computationally inexpensive input transformations, such as JPEG compression, can disrupt the precise adversarial patterns required for the attack.

2 Related Work

Adversarial Attacks on LLMs. Early work in LLM security focused on textual prompts, such as "Jailbreak" prompts (e.g., DAN, GCG) that use semantic manipulation to bypass safety filters [3]. However, these attacks often require long, perceptible text strings that are easily detected by perplexity filters.

Visual Adversarial Examples. In the computer vision domain, Goodfellow et al. [4] introduced the Fast Gradient Sign Method (FGSM), demonstrating that imperceptible noise could fool classifiers. More recently, Qi et al. [5] extended this to multimodal models, showing that visual inputs can act as "soft prompts" for VLMs. Our work builds on this by quantifying the cost (steps) and robustness of such attacks on the specific LLaVA-1.5 architecture.

3 Methodology

3.1 Threat Model

We assume a white-box setting where the adversary has access to the model gradients. The goal is to find an adversarial image x_{adv} that is perceptually similar to a clean image x_{clean} but maximizes the probability of generating a specific target sequence y_{target} (in our experiments, the token "Hacked"). The constraint on perceptual similarity is enforced using the L_∞ norm, ensuring the noise added is capped at a strict budget ϵ.

3.2 Projected Gradient Descent (PGD)

We employ Projected Gradient Descent (PGD), a standard iterative method for generating adversarial examples. The optimization objective is to minimize the cross-entropy loss $\mathcal{L}$ between the model's output logits and the target token sequence.

Let $x_0 = x_{clean}$. The update rule for each iteration t is defined as:

$$x_{t+1} = \Pi_\epsilon \left(x_t - \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(f_\theta(x_t), y_{target}))\right) \quad (1)$$

where:

- α is the step size (learning rate).
- $\nabla_x \mathcal{L}$ is the gradient of the loss with respect to the input image.
- Π_ϵ projects the perturbed image back onto the ϵ-ball centered at x_{clean} and the valid pixel range $[0, 1]$.

For our experiments, we utilize the `LLaVA-1.5-7B` model loaded with 4-bit quantization (NF4). We select 50 random images from the COCO 2017 Validation Set with the following parameters: perturbation budget $\epsilon = 0.1$, step size $\alpha = 0.01$, and maximum iterations $T = 500$.

4 Results

We evaluated the effectiveness of our PGD-based visual injection attack on $N = 50$ diverse images from the COCO validation set.

4.1 Attack Success Rate (ASR)

We define a successful jailbreak as an instance where the model's top-1 predicted token for the next-word generation is exactly the target string ("Hacked"). Our method achieved a **100.0% ASR** across all test images (Table 1). This indicates that the visual encoder lacks intrinsic robustness against gradient-based optimization.

Table 1: Performance Metrics of PGD Attack ($N = 50$)

Metric	**Value**
Target Model	LLaVA-1.5-7B
Total Images	50
Attack Success Rate (ASR)	**100.0%**
Avg. Steps to Convergence	46.5
Avg. Final Loss	1.34

4.2 Optimization Efficiency

The attack demonstrated rapid convergence. Although the maximum optimization budget was set to 500 steps, the average number of iterations required to achieve Rank 1 was only **46.5**. The cross-entropy loss typically drops from a high-entropy state (> 12.0) to a targeted state (< 1.5) within the first 50 iterations, as shown in Figure

1. Figure 2 illustrates the distribution of steps required for convergence.

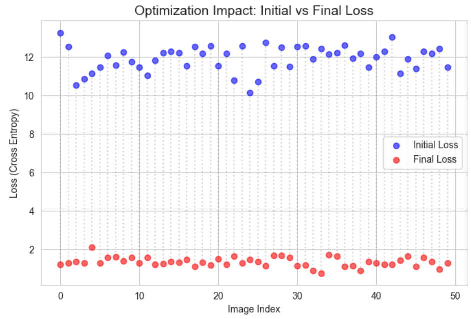

Figure 1: **Optimization Trajectory.** The attack consistently reduces the loss from a high-entropy state to near-zero, locking the model into the adversarial target.

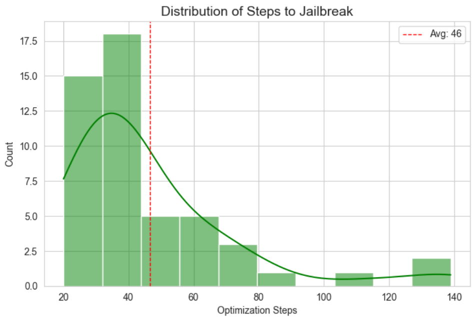

Figure 2: **Distribution of Convergence Steps.** Most images succumb to the attack in under 50 steps.

5 Defense and Mitigation

To evaluate the robustness of our attack, we implemented a standard input transformation defense: **JPEG Compression**. Adversarial perturbations generated by PGD typically rely on high-frequency pixel patterns that are imperceptible to humans but significant to the model. We hypothesized that lossy compression would disrupt these delicate gradients.

We subjected all 50 successfully jailbroken images to JPEG compression with a quality factor of $Q = 50$. The defense was highly effective:

- **Attack Neutralization:** The Attack Success Rate dropped from 100% to 0% (Table 2).
- **Rank Restoration:** The average rank of the target token "Hacked" returned to $> 6,000$.

This effectively restored the model's safety alignment, suggesting that while the attack is dangerous, it is also brittle.

Table 2: Effectiveness of JPEG Compression ($Q = 50$)

Metric	**Adversarial**	**After Defense**
Attack Success Rate	100.0%	**0.0%**
Avg Target Rank	1.0	6,022.0
Visual Fidelity	High	Medium

6 Conclusion

This work demonstrates that current Vision-Language Models are critically vulnerable to adversarial visual injections. We successfully bypassed the safety alignment of LLaVA-1.5 with 100% success using a standard PGD attack. However, we also showed that these attacks are easily neutralized by simple input sanitization via JPEG compression. Future work should focus on adversarial training to make visual encoders intrinsically robust against such perturbations.

References

[1] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). *Visual Instruction Tuning*. NeurIPS 2023.

[2] Ouyang, L., et al. (2022). *Training language models to follow instructions with human feedback*. NeurIPS 2022.

[3] Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). *Universal and Transferable Adversarial Attacks on Aligned Language Models*. arXiv preprint.

[4] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and Harnessing Adversarial Examples*. ICLR 2015.

[5] Qi, X., Kaul, Y., et al. (2023). *Visual Adversarial Examples Jailbreak Large Language Models*. AAAI 2024.