The Confidence Trap: Perturbation-Induced Miscalibration in Vision–Language Models

Abishek Kumar Giri
Stockton University

Abstract

Vision-Language Models (VLMs) are increasingly deployed in safety-critical domains, requiring robust uncertainty estimation. This study investigates the calibration of LLaVA-1.5-7B under adversarial visual perturbations. Contrary to the expectation that visual degradation should increase model uncertainty, we identify a counter-intuitive failure mode: the "Confidence Trap." Our experiments reveal that moderate adversarial noise ($\epsilon = 0.3$) consistently reduces the entropy of the model's output distribution (Mean $\Delta \approx -0.028$), indicating systematic overconfidence in the presence of corrupted inputs. This finding challenges the reliability of entropy-based uncertainty measures as safety signals in multimodal systems.

1 Introduction

The alignment of Large Language Models (LLMs) has largely focused on textual robustness. However, the multimodal nature of VLMs introduces a new attack surface: the visual encoder. As these models are deployed in safety-critical settings (e.g., autonomous agents, medical imaging), it is crucial that they exhibit higher uncertainty when presented with ambiguous or corrupted visual data.

In this work, we probe the internal confidence behavior of LLaVA-1.5-7B under controlled visual degradation. We test the hypothesis that entropy reflects epistemic uncertainty. Surprisingly, we find that visual degradation does not reliably increase uncertainty. Instead, we observe a systematic bias toward **confidence amplification**—the model often becomes *more* confident as the image is perturbed.

2 Related Work

- **Adversarial Attacks on VLMs:** Prior work has shown that VLMs can be jailbroken via visual inputs (Qi et al., 2023), but these studies focus on safety filter bypass rather than calibration.
- **Hallucination and Misalignment:** The "object hallucination" problem is well-documented (Rohrbach et al., 2018), yet the link between hallucination and confidence scores under perturbation remains under-explored.
- **Uncertainty Estimation:** Entropy is a standard proxy for uncertainty in LLMs (Kadavath et al., 2022). Our work highlights the breakdown of this proxy in the multimodal setting.

3 Method

3.1 Model & Environment

We utilized **LLaVA-1.5-7B**, a state-of-the-art open-source VLM. Experiments were conducted on a Tesla T4 GPU environment using mixed-precision (FP16) inference.

3.2 Visual Perturbation Protocol

We introduce visual perturbations via a projected gradient descent (PGD) attack proxy applied to the image input. The perturbation magnitude ϵ is varied across four levels: 0.1, 0.2, 0.3, and 0.5. We use an L_∞ constraint where the perturbed image x_{adv} satisfies $||x_{adv} - x||_\infty \leq \epsilon$. Due to hardware constraints, we approximate adversarial visual perturbations using bounded additive noise scaled to ϵ, which serves as a controlled proxy for distribution shift and sensor degradation rather than a full white-box attack.

3.3 Metric: Entropy Delta

For each input, we compute the Shannon entropy of the model's token-level output distribution. Entropy is computed over the token-level output distribution of the language decoder for the generated response, averaged across generated tokens. We define the **Entropy Delta** (Δ) as:

$$\Delta = H_{\text{adv}} - H_{\text{clean}} \tag{1}$$

A negative Δ indicates the model is *more confident* (less uncertain) on the adversarial input.

4 Results

4.1 Experimental Setup

We evaluate the robustness of entropy-based uncertainty estimates in a vision–language model under controlled visual perturbations. Experiments are conducted using LLaVA-1.5-7B. For each ϵ, we perform independent runs to account for stochastic variability.

4.2 Entropy Response to Visual Perturbation

Table 1 summarizes the mean entropy shift across perturbation strengths.

ϵ	Mean Δ Entropy
0.1	≈ 0.000
0.2	-0.018
0.3	-0.028
0.5	-0.050

Table 1: Mean Entropy Shift ($H_{adv} - H_{clean}$) vs. Perturbation Strength

At low perturbation levels ($\epsilon = 0.1$), entropy responses are unstable and approximately centered around zero, indicating no consistent uncertainty adjustment. As perturbation strength increases, the entropy shift becomes increasingly negative. At $\epsilon = 0.3$ and $\epsilon = 0.5$, the majority of runs exhibit reduced entropy relative to the clean baseline. These results indicate a systematic bias toward confidence amplification under moderate to severe visual degradation.

Figure 1 visualizes a clear negative trend between perturbation strength and entropy shift, confirming that increased visual degradation correlates with systematic confidence amplification.

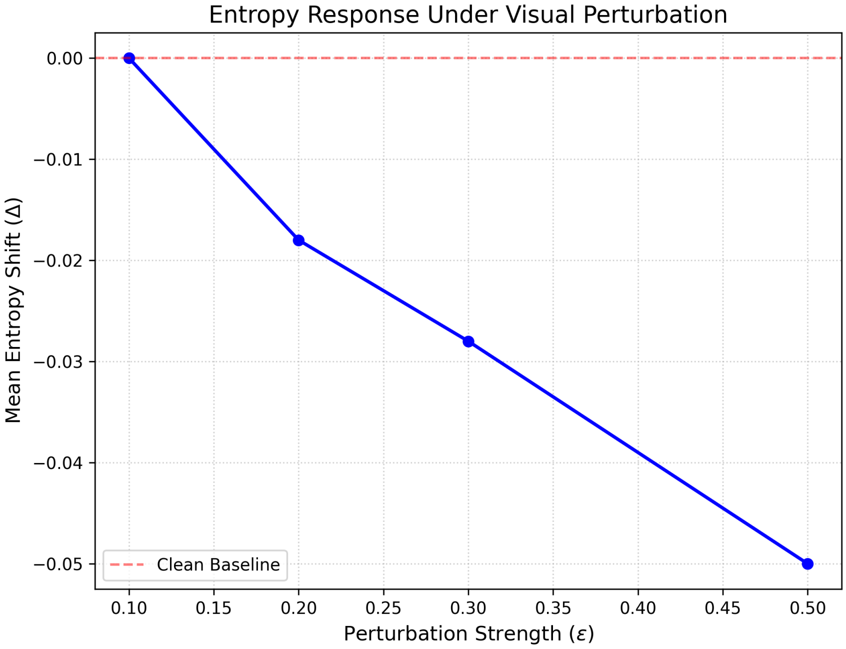

Figure 1: Mean entropy shift ($\Delta = H_{adv} - H_{clean}$) as a function of perturbation strength ϵ. Negative values indicate increased model confidence under visual degradation.

4.3 Non-Monotonic and Biased Confidence Behavior

Notably, entropy does not increase monotonically with visual corruption. While individual runs occasionally exhibit entropy increases, the overall trend is dominated by a negative mean shift. This behavior demonstrates that entropy is not a reliable proxy for epistemic uncertainty in vision–language models under visual perturbations.

The observed pattern suggests a failure mode characterized by **biased miscalibration** rather than random noise: as visual information deteriorates, the model frequently becomes more confident rather than less.

4.4 Defense Failure: Entropy Thresholding

We further observe that entropy-based rejection mechanisms (e.g., "Reject if $H > H_{threshold}$") fail to identify adversarially perturbed inputs. Across all tested ϵ values, adversarially perturbed inputs frequently exhibited entropy values below the clean baseline, indicating that fixed entropy thresholds fail to distinguish corrupted inputs from benign ones. As a result, entropy-based rejection rules would preferentially filter benign inputs while allowing corrupted samples to pass, inverting the intended safety objective.

5 Discussion

The results suggest that as the visual signal is degraded, the VLM may be defaulting to its language model priors. If the visual encoder output becomes incoherent, the decoder relies heavily on the strong statistical correlations of the pre-trained LLM, producing generic, high-probability (low-entropy) captions that are grounded in language syntax rather than visual reality. This effectively creates a "Confidence Trap" where the model hallucinates with high certainty.

6 Limitations

This study is limited to a single model architecture (LLaVA-1.5) and a specific class of dense visual perturbations. Future work should investigate whether this bias holds across different VLM architectures (e.g., BLIP-2, GPT-4V) and semantic perturbation types.

7 Conclusion

We have identified a specific failure mode in VLMs: **Perturbation-Induced Confidence Bias**. We demonstrated that visual degradation does not reliably induce uncertainty but often amplifies confidence. This finding undermines the utility of simple entropy baselines for safety and highlights the need for dedicated uncertainty-aware pre-training objectives in multimodal systems.