

A Framework for Evaluating Vision-Language Model Safety: Building Trust in AI for Public Sector Applications

Maisha Binte Rashid, Pablo Rivas

Department of Computer Science
School of Engineering and Computer Science
Baylor University, Texas USA
{Maisha_Rashid1, Pablo_Rivas}@Baylor.edu

Abstract

Vision-Language Models (VLMs) are increasingly deployed in public sector missions, necessitating robust evaluation of their safety and vulnerability to adversarial attacks. This paper introduces a novel framework to quantify adversarial risks in VLMs. We analyze model performance under Gaussian, salt-and-pepper, and uniform noise, identifying misclassification thresholds and deriving composite noise patches and saliency patterns that highlight vulnerable regions. These patterns are compared against the Fast Gradient Sign Method (FGSM) to assess their adversarial effectiveness. We propose a new Vulnerability Score that combines the impact of random noise and adversarial attacks, providing a comprehensive metric for evaluating model robustness.

Application Context

The framework we present addresses the pressing need for robust evaluation metrics to ensure the safety and reliability of VLMs in public missions. Our work specifically targets applications where VLMs play a critical role, such as disaster response, medical diagnostics, infrastructure management, and public education. These domains demand systems capable of maintaining performance even under adverse conditions, including adversarial manipulations.

Our proposed Vulnerability Score metric evaluates the resilience of VLMs against both random noise and targeted adversarial attacks. This makes it particularly suitable for government and non-governmental organizations tasked with deploying AI systems in high-stakes environments. For example, in disaster response scenarios, where real-time and accurate decision-making is essential, identifying and mitigating adversarial risks ensures that VLMs can be trusted to process critical visual and textual information reliably. Similarly, in medical imaging or urban planning, robustness against adversarial threats protects against errors that could have significant societal consequences.

By combining comprehensive noise analysis and adversarial impact assessment, this approach aligns with the public sector's emphasis on trustworthy AI. While the method's computational intensity might pose challenges for resource-constrained settings, its ability to quantify vulnerabilities

and provide actionable insights outweighs these limitations in mission-critical scenarios. Consequently, the framework equips stakeholders with a powerful tool to make informed decisions about deploying and governing AI technologies in public service contexts.

Introduction

With the rapid advancement of artificial intelligence, VLMs have become deeply embedded in our daily lives. As of 2023, over six billion people worldwide use smartphones (Aarestad 2023), many of which are enabled with VLM-based features like augmented reality, real-time translation, and intelligent personal assistants. The increasing integration of artificial intelligence (AI) into a wide range of public sector missions follows from disaster response and medical diagnostics to infrastructure management and educational initiatives, there is a growing imperative to ensure that such systems are both reliable and safe. The integration of VLMs into governmental and non-governmental services demands high standards of trust, fairness, and resilience, especially when these models operate under challenging conditions or encounter adversarial manipulations.

An adversarial attack involves manipulating a machine learning model by using carefully crafted data to exploit its vulnerabilities, potentially leading to incorrect predictions. While there has been extensive research on adversarial attacks in computer vision and natural language processing, Binte Rashid, Rahaman, and Rivas (2024) identified a research gap in effectively measuring adversarial attacks and conducting safety assessments for VLMs. Existing studies have primarily focused on developing attack methodologies and their defense mechanisms, with little attention given to reliable measurement of risk assessment and developing evaluation metrics to assess vulnerability.

To ensure safety and trustworthiness in public service AI systems, we must develop comprehensive and standardized metrics that reliably quantify risk, evaluate robustness, and identify vulnerabilities. Public missions often deal with different kinds of stakeholders, diverse data sources, and time constraints. It becomes not only a technical issue but also a public policy necessity to guarantee AI's consistent performance while maintaining equity, transparency, and social well-being.

This paper introduces a novel approach to efficiently

quantifying and evaluating the adversarial vulnerabilities of VLMs, with a specific focus on the CLIP model (Radford et al. 2021). Our approach combines a comparison to well-known adversarial attack approaches with an analysis of model performance under different noise conditions. Our goal is to present a thorough evaluation of VLM vulnerability by looking at the model’s response to Gaussian, Salt and Pepper, and Uniform noise. Our research contributes to the field in several ways:

- We conduct an experiment by applying incremental noise levels to assess the threshold at which misclassification occurs in image classification task.
- We compare the effectiveness of our noise-based perturbations against the FGSM, a well-established adversarial attack technique.
- We propose a new evaluation metric that calculates a Vulnerability Score by combining the impact of both random noise and targeted adversarial attacks.

In situations where public trust and mission success are crucial, we hope that these measurements can help policy-makers, developers, and community organizations make better decisions concerning the deployment and regulation of AI systems.

Background Study

Adversarial attacks on neural networks often involve crafting input perturbations to cause model misclassifications. One of the first and best known attack is the FGSM, which adds noise aligned with the gradient of the loss to the input data (Goodfellow, Shlens, and Szegedy 2014). Building on this, Projected Gradient Descent (PGD) applies iterative gradient-based perturbations, projecting the adversarial example back onto the valid input space at each step (Madry et al. 2017). Other notable attacks include the Jacobian Saliency Map Adversary (JSMA), which leverages a saliency map derived from the Jacobian matrix to identify and perturb key pixels (Papernot et al. 2016), and the Carlini and Wagner (C&W) attack, an optimization-based approach designed to produce minimal, nearly invisible perturbations that still cause reliable misclassification (Carlini and Wagner 2017).

Dong et al. (2020) highlighted the importance of evaluating robustness using robustness curves that shows model performance across different level of perturbation and attack strengths which provides a more comprehensive assessment of adversarial robustness compared to traditional point-wise accuracy. The authors also discussed about the effectiveness of adversarial training and the robustness of randomization-based defenses against query-based black-box attacks. Zhao et al. (2024) proposed comprehensive evaluation of adversarial robustness in VLMs. The results show that advanced models like MiniGPT-4 (Zhu et al. 2023), LLaVA (Liu et al. 2024), BLIP-2 (Li et al. 2023) trained on large datasets are vulnerable to adversarial noise, which results in generating erroneous or targeted output. By evaluating black-box attacks, the authors concluded that adversarial attacks crafted against one VLM can often be transferred to others, showing the widespread vulnerability across different architectures.

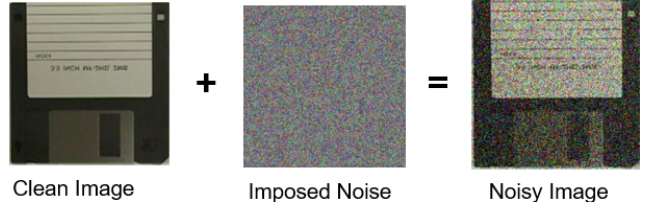


Figure 1: Imposing noise to clean image

Noise Level	Original Class	Misclassified
0.75	Telephone	Soda-Can
0.29	Playing-Card	Sheet-music
0.10	Airplane	Hotdog
0.46	Saturn	Mussels
0.56	Penguin	Chimp

Table 1: Sample of noise levels and model performance for Gaussian noise.

Methodology

Assessing the vulnerability of VLMs to adversarial attacks is essential for ensuring their reliability and safety in real-world applications. Existing metrics like CLEVER (Weng et al. 2018), AutoAttack (Croce and Hein 2020) often fail to capture the combined impact of random noise and targeted adversarial perturbations on model performance. Developing a comprehensive evaluation metric that integrates the effects of various perturbations can efficiently assess a model’s robustness

In our experiment, illustrated in Figure 2, we conducted a classification task using the CLIP model, a vision-language model, on the Caltech-256 dataset. We randomly selected 300 images from the dataset, that is approximately 1% of the entire dataset ensuring that all the classes were included. We applied each of the three previously mentioned noise types to our sample set. The implementation of the noises in as follows:

- For each noise type, starting with a clean image, we increased the noise level by 0.01 in each iteration as shown in Figure 1. We can denote it as, for any noise $N(i)$ applied to an image at iteration i we add 0.01 noise level at each iteration. $N(i + 1) = N(i) + 0.01$.
- At each noise level, we evaluated the CLIP model’s classification performance on the perturbed image.
- We tracked the specific noise level at which the model first misclassified the image, marking the threshold of the model’s robustness for that particular image and noise type. Let $N_{final} = N(i)$ and suppose $F(I, N(i)) \neq True_Label$, where I is the original image and $F(I, N(i))$ is the model function for class prediction. In Table 1, we show a sample of noise levels and the corresponding model performance for Gaussian noise.

We gathered the noise levels that caused misclassification for each of the 300 images. By averaging these noise lev-

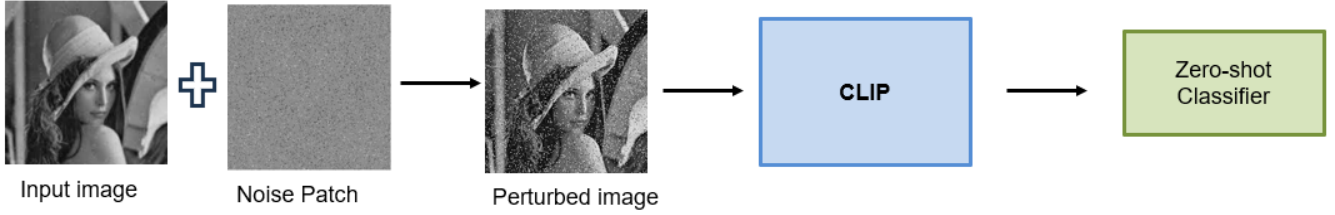


Figure 2: Methodology Process.

els, we created a composite noise image patch representing the average noise pattern that leads to misclassification in the model. We performed this experiment for each of the three noises and collected the average noise patch as shown in Figure 3.

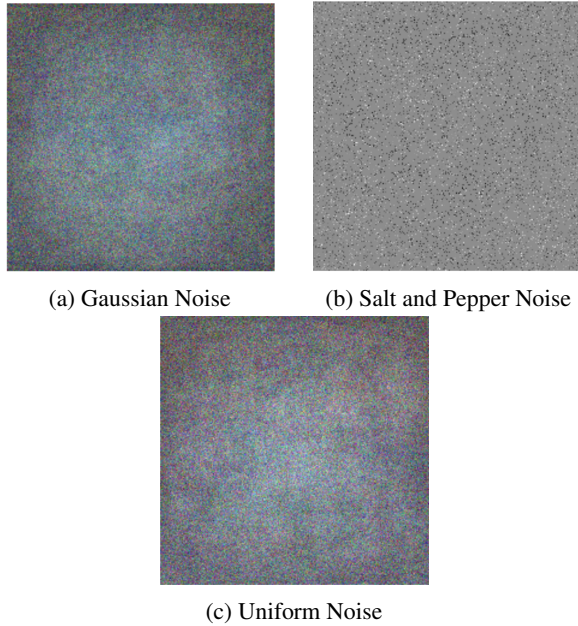


Figure 3: Noise patches of three types of noises.

Results and Performance Analysis

Once we established both the noise patches and the saliency pattern, our next goal was to rigorously assess their potential as adversarial perturbations in comparison to a well-known benchmark, FGSM. We approached this comparison systematically. First, we introduced image-level perturbations derived from each of our identified noise patches. These patches, generated by averaging the noise levels at which misclassifications first occurred, were designed to capture the “average vulnerability” of the model to each type of noise. By applying these noise patches to previously unseen images, we aimed to determine how well they generalized as universal adversarial perturbation. In other words, we wanted to see if a single, pre-computed noise pattern could reliably induce misclassification in images it had never “seen” during the noise derivation phase.

Attack Name	Accuracy(%)
Baseline	95.00
Gaussian Noise	67.54
Salt and Pepper	66.80
Uniform Noise	66.56
FGSM	9.35

Table 2: Comparison of Model Accuracy under Different Adversarial Attacks

We wanted to gain a deeper understanding of which pixels or regions of the image seems to be more vulnerable towards adversarial perturbation. Using the same noise patches, we generated a saliency pattern. This saliency pattern highlights image regions that, when altered, cause the model’s predictive confidence and accuracy to degrade. By applying this saliency pattern as a perturbation, we effectively tested the ability of focusing our adversarial attack on critical, high-impact areas within the image. This approach allowed us to move from broad-spectrum noise attacks to more targeted, strategic perturbations.

After experimenting with both the noise patches and the saliency pattern, we then introduced the conventional FGSM attack for comparison. While FGSM is known for its simplicity but as it calculates the gradients of loss to mislead the model, it is computationally expensive. By taking into account the result of FGSM, we could ascertain how well our noise patches and saliency-based perturbations measure up against a classic, widely recognized adversarial method. The outcomes of these experiments are summarized in Table 2, which provides a direct comparison of the model’s accuracy for each type of attack. With the help of the results of Table 2, we developed a new metric that jointly considers the impact of random noise and targeted adversarial perturbations on the model’s performance. This metric offers a more comprehensive measure of the model’s vulnerability. Our evaluation metric is defined as:

$$\text{Noise Impact Score} = \frac{\text{Acc}_{\text{Baseline}} - \text{Acc}_{\text{Noise}}}{\text{Acc}_{\text{Baseline}}} \times 100,$$

$$\text{FGSM Impact Score} = \frac{\text{Acc}_{\text{Baseline}} - \text{Acc}_{\text{FGSM}}}{\text{Acc}_{\text{Baseline}}} \times 100,$$

where $\text{Acc}_{\text{Baseline}}$ is the baseline accuracy, $\text{Acc}_{\text{Noise}}$ is the accuracy under random noise, and Acc_{FGSM} is the accuracy under FGSM attack.

Vulnerability Score = $w_1 \times \text{Noise Impact Score} + w_2 \times \text{FGSM Impact Score}$ where $w_1 + w_2 = 1$ and $w_1, w_2 \geq 0$.

By utilizing our proposed evaluation metric, we aim to systematically assess the vulnerability of VLMs to adversarial perturbations. This metric provides an efficient and scalable method for researchers to quantify their models' robustness, as demonstrated in our experiments where we worked with only 1% of the dataset. Despite the limited sample size, the results effectively highlighted the sensitivity of the model to both random noise and targeted adversarial attacks, underscoring the metric's reliability.

A key advantage of our approach lies in its flexibility: the weights w_1 and w_2 can be adjusted to emphasize either random noise impact or adversarial perturbations, depending on the specific evaluation needs. This adaptability allows researchers to tailor the metric to suit different scenarios, such as prioritizing robustness against natural perturbations in real-world environments or focusing on resilience to deliberate adversarial attacks (Rashid and Rivas 2024).

Conclusions

In this work, we introduced a novel framework to evaluate the adversarial vulnerabilities of VLMs, focusing on public mission applications. Our method integrates the analysis of random noise and adversarial attacks, providing a comprehensive Vulnerability Score that highlights model weaknesses under diverse perturbation conditions. By leveraging incremental noise application and saliency-based analysis, we demonstrated the effectiveness of our approach in quantifying risks and identifying critical areas of vulnerability.

The proposed framework addresses a pressing need in public sector AI, ensuring that systems deployed in high-stakes environments maintain trust and reliability. While computational complexity poses a potential challenge, the insights gained from this evaluation provide actionable metrics for developers, policymakers, and organizations seeking to deploy robust AI systems. Future work will optimize the framework's computational efficiency and extend its applicability to other multimodal AI architectures.

Our findings contribute to the broader effort of ensuring trustworthy AI in public service, offering a practical and scalable solution for assessing and improving the safety of VLMs in mission-critical applications.

Acknowledgments

Part of this work was funded by the National Science Foundation under grants CNS-2210091 and CNS-2136961, and by the U.S. Department of Education under grant P116Z230151.

References

- Aarestad, S. 2023. Smartphone addiction and subjective withdrawal effects: a three-day experimental study. *Sage Open*, 13.
- Binte Rashid, M.; Rahaman, M. S.; and Rivas, P. 2024. Navigating the Multimodal Landscape: A Review on Integration of Text and Image Data in Machine Learning Architectures. *Machine Learning and Knowledge Extraction*, 6(3): 1545–1563.
- Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57.
- Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, 2206–2216. PMLR.
- Dong, Y.; Fu, Q.-A.; Yang, X.; Pang, T.; Su, H.; Xiao, Z.; and Zhu, J. 2020. Benchmarking adversarial robustness on image classification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 321–331.
- Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.
- Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z. B.; and Swami, A. 2016. The Limitations of Deep Learning in Adversarial Settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 372–387.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Rashid, M. B.; and Rivas, P. 2024. AI Safety in Practice: Enhancing Adversarial Robustness in Multimodal Image Captioning. In *Proceedings of the KDD 2024 workshop on Ethical AI*. ArXiv:2407.21174 [cs.CV].
- Weng, T.-W.; Zhang, H.; Chen, P.-Y.; Yi, J.; Su, D.; Gao, Y.; Hsieh, C.-J.; and Daniel, L. 2018. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*.
- Zhao, Y.; Pang, T.; Du, C.; Yang, X.; Li, C.; Cheung, N.-M. M.; and Lin, M. 2024. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36.
- Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.