Towards Mitigating Simple and Adversarial Attacks for Large Vision Language Models

Mridul Khurana *
Virginia Tech
PID: mridul
mridul@vt.edu

Harish Babu Manogaran *
Virginia Tech
PID: harishbabu
harishbabu@vt.edu

Abstract

Large Vision Language Models (VLMs) have demonstrated powers in integrating natural language processing with multimodal understanding, excelling in tasks from language translation to visual question answering (VQA). Despite their capabilities, the susceptibility of VLMs to adversarial attacks threatens their reliability, particularly in critical applications. This paper investigates VLMs' robustness, revealing a pronounced positional bias in VQA tasks that significantly impacts performance. We further explore their vulnerability to sophisticated adversarial attacks using Projected Gradient Descent (PGD), which induces harmful textual responses. To combat these vulnerabilities, we introduce a novel mitigation strategy, latent smoothing, inspired by traditional noise reduction techniques. Our findings show this method effectively enhances model resilience against adversarial perturbations. This study highlights critical vulnerabilities in multimodal AI systems and provides actionable strategies for improving their adversarial robustness, ensuring safer and more reliable AI deployments.

1. Introduction

Large Vision Language Models (VLMs) have emerged as a transformative force in artificial intelligence, merging the power of natural language processing with multimodal understanding. These models are typically pre-trained on expansive datasets, enabling them to excel in a variety of tasks, including language translation, summarization, image captioning, and visual question answering (VQA). By harnessing vast amounts of text and image data, VLMs are capable of encoding deep semantic representations that empower them to interpret and produce complex multimodal content with high accuracy.

Despite their remarkable capabilities and broad applicability across sectors such as healthcare, finance, entertainment, and autonomous systems, VLMs exhibit significant vulnerabilities, especially when subjected to adversarial attacks. These attacks manipulate inputs in subtle ways that can lead the models to produce incorrect or misleading outputs, thereby compromising their reliability and utility in critical applications. This paper specifically focuses on evaluating the robustness of VLMs against a spectrum of threats, ranging from simple manipulations to sophisticated adversarial strategies.

Our investigation begins with an evaluation of VLMs in Visual Question Answering (VQA) tasks. Here, models are tested on their ability to select the correct answer from multiple choices given an image and a related question. This assessment reveals a notable positional bias—VLMs often favor answers based on their position in the input sequence, a bias likely ingrained during training on structured datasets like ImageNet [2]. In challenging datasets like MMMU [14], this bias can degrade the model's decision-making to near-random levels.

Building on this foundation, we escalate the complexity of our tests by employing Projected Gradient Descent (PGD) [7] to generate adversarial images. These crafted images are designed to deceive the VLMs into generating harmful or entirely inaccurate textual responses. This stage of our study highlights the models' critical vulnerabilities in processing visual information and their potential for being exploited in adverse scenarios.

To counteract these vulnerabilities, we introduce a mitigation strategy named "Visual Latent Smoothing." This approach, inspired by traditional noise reduction techniques, is shown to significantly mitigate the effects of adversarial attacks. Through empirical testing, latent smoothing proves to be an effective defense, enhancing the models' resilience against visual perturbations and contributing to their robustness.

Our contributions through this work are manifold:

*Equal Contribution

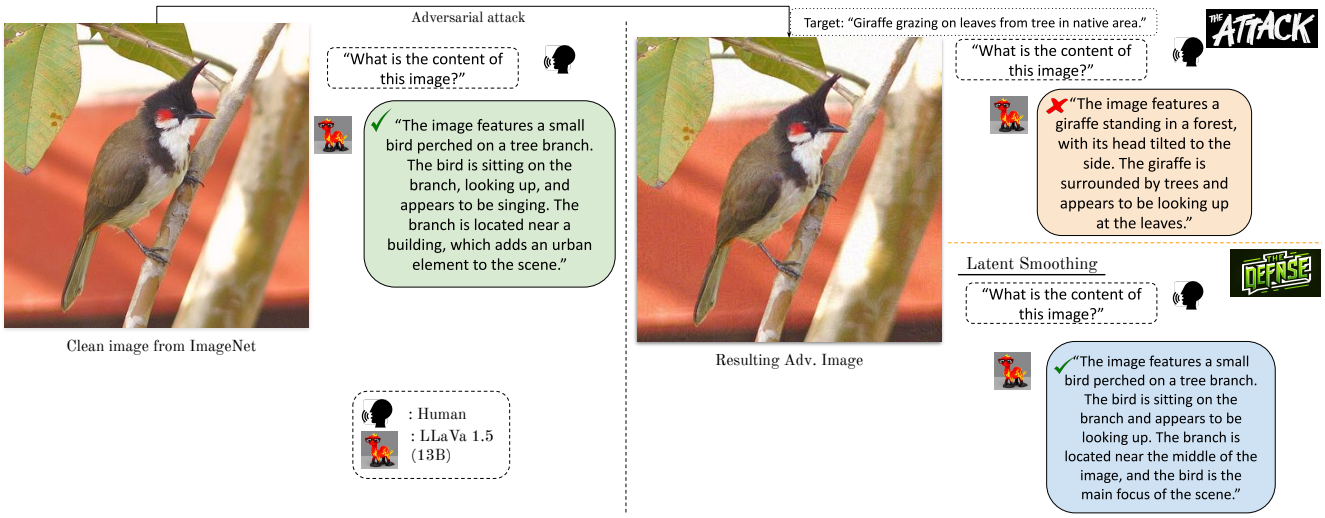

Figure 1: **Analysis of VLM Responses after Adversarial Attack**. This figure illustrates the response of a LLaVa 1.5 under three conditions: (A) Response to a clean image from ImageNet, showing the model's capability in generating good text description (left); (B) Response to the same image after being subjected to an adversarial perturbation, demonstrating the model's vulnerability to adversarial attacks (right); (C) Response following the proposed Visual Latent Smoothing, highlighting the effectiveness of the defense strategy in mitigating adversarial effects and restoring the model's performance towards baseline levels.

1. We highlight the influence of simple text-based perturbations, such as the random ordering of choices in VQA tasks, and illustrate how these can drastically affect the performance of VLMs in Section 3.

2. In Section 4, we conduct targeted adversarial attacks on the visual modality, demonstrating that these can significantly mislead VLMs with minimal, nearly imperceptible changes to the input images.

3. We propose and validate a novel defense strategy in Section 5, visual latent smoothing, which proves effective in shielding VLMs from sophisticated vision-based adversarial attacks.

This work not only exposes the current limitations of VLMs when facing adversarial scenarios but also advances the discussion on improving their robustness. By enhancing their ability to withstand such attacks, we help ensure their reliability and trustworthiness in real-world applications across diverse sectors. Our findings and methodologies contribute significantly to the field of multimodal AI, providing a roadmap for future research aimed at fortifying these powerful systems against emerging threats.

2. Related Works

Vision language models (VLMs) represent a significant advancement in the integration of visual and linguistic data processing. These models vary widely in architecture, capability, and accessibility, ranging from proprietary, closed systems developed by large corporations such as ChatGPT cite to open-source models that facilitate academic research like blip cite and innovation. This diversity affects not only the potential applications for each model but also influences the robustness and generalizability of the technologies. Open-source models, such as those built on platforms like Hugging Face cite or shared in academic repositories, allow for broader scrutiny and iterative improvements by the community

Recent research, notably by [16], have highlighted a significant vulnerability in vision language models (VLMs)—a susceptibility to relatively simple text-based attacks, such as the random reordering of answer options in visual question-answering tasks. These findings indicate a possible inherent positional bias within these models, casting doubt on their robustness and reliability for real-world applications. Building upon this groundwork, our research further explores the resilience of VLMs across image classification and logical reasoning tasks. Our empirical results demonstrate that minor perturbations in the answer choices can significantly affect the models' performance.

Moreover, recent studies have also highlighted that the visual component is frequently the most vulnerable aspect of many vision language models (VLMs). An evaluation led by [15] pinpointed this susceptibility through a variety of adversarial attack methods. Operating under black-box

conditions [15] demonstrated that VLMs can be manipulated to generate text that is entirely unconnected to the accompanying image by employing open-source CLIP [9] and BLIP [5] models as proxies to introduce adversarial perturbations. Expanding on these findings, [8] and [12] have shown that it is possible to compromise VLMs through the visual modality alone, causing them to produce harmful, illegal, or offensive content. These methods involve the adversarial insertion of damaging content into input images, underscoring the potential risks associated with the misuse of VLMs. We have replicated these experiments on various VLMs of different sizes to evaluate the robustness of the visual modality in multimodal systems.

Additionally, to counter vision-based adversarial attacks, we draw on traditional random smoothing techniques commonly utilized in image classification to enhance the adversarial robustness of vision encoders. Random smoothing has been proven to mitigate the impact of adversarial attacks [3]. Researches have developed a comprehensive theory demonstrating that random smoothing can effectively ensure adversarial robustness in real-world settings [13]. Also, applying smoothing as a method to protect against text-based adversarial attacks in Large Language Models (LLM) has been tried [10]. In our study, we implement latent smoothing on the vision encoder to neutralize the effects of adversarial intrusions. Preliminary assessments indicate that this approach does not adversely affect the overall performance of the model, maintaining efficacy while enhancing security.

3. Text Based Adversarial Attacks

This section evaluates the robustness of the LLaVa 1.5 13B [6] model applied to Visual Question Answering (VQA) within a Multiple-Choice Question Answering (MCQA) framework. Our aim is to determine how the fixed versus randomized positioning of correct answers influences the model's decision-making process, exposing inherent biases and assessing its generalization capabilities.

3.1. Positional Bias in VQA Tasks

To assess positional bias, we orchestrated a controlled experimental design where the correct answer is statically assigned to one of four designated positions (Options A, B, C, or D) across multiple trials. Each trial tasks the VLM with selecting the correct option from four given choices for each image-based question, ensuring that variables such as question complexity and image content are standardized.

Table 1 Shows results of LLava 1.5 13B that displayed a pronounced positional bias. Notably, when the correct answer was placed in Position A, the model achieved an exceptionally high accuracy of 99.3%. In contrast, the accuracy significantly decreased when the correct answer was located in Positions B, C, and D, with recorded accuracies of 52.5%, 46.4%, and 47.9% respectively.

Choices	**Accuracy (%)** $\uparrow$
Answer is always A	99.3
Answer is always B	52.5
Answer is always C	46.4
Answer is always D	47.9
Random position	64.4
Random choice	25.0

Table 1: Accuracy of the Lava 1.5 13B model under different answer positioning strategies.

This disparity in performance across different positions suggests a substantial learning bias, where the model has likely been conditioned during its training phase to associate the first position with the correct answer more frequently than the others. Such biases could severely undermine the utility of the model in scenarios where the position of the correct answer is not fixed, leading to potentially biased and unreliable outcomes, especially in real-world applications, where such regularities are typically absent, leading to skewed or unreliable outputs.

3.2. Simple Random Ordering Impacts Performance

To explore the effect of answer randomization, the correct answer was shuffled randomly among the four positions across a new set of trials. This adjustment yielded an average accuracy of 64.4%, as detailed in Table 1. This configuration not only resulted in a decrease in overall peak performance compared to the fixed first position but also exhibited more consistent performance across different trials.

The results indicate that introducing randomness mitigates some of the pronounced biases observed with fixed positioning. By forcing the model to adapt to varied answer locations, the random setup encourages a more genuine reliance on the integration of visual and textual information, thereby enhancing the model's robustness and applicability in diverse real-world settings. Random choice in Table 1 highlights the probability of choosing the right answer at random.

4. Vision Based Adversarial Attacks

4.1. Image to Image PGD attack

Following the image to image attack strategy introduced in [15], we implement an adversarial attack strategy aimed at deceiving large vision-language models (VLMs) by generating content that is unrelated to the original image, yet closely aligned with a target description. We assume a grey-box access to the VLM where the weights of the vision en-

coder used by the model is available. The attack approach involves several key steps, as outlined below:

Image Generation: We start with a target description, D_{target}, and intend to use it as a guide for generating an adversarial image. This description is input into a stable diffusion [11] model, which generates an image I_{adv} that visually embody D_{target}.

PGD attack to increase latent space similarity: We proceed by encoding both the original image I_{orig} and the adversarial image I_{adv} using the VLM's vision encoder to obtain their respective embeddings. The objective is to adjust I_{orig} such that so that its embedding approximates that of I_{adv}. To achieve this, we employ Projected Gradient Descent (PGD) to minimize the cosine similarity between the embeddings of the two images. The optimization problem can be formulated as follows:

$$\min_{\delta} \cos(\text{VE}(I_{\text{orig}} + \delta), \text{VE}(I_{\text{adv}})),$$

where δ represents the perturbation applied to I_{orig}, and VE represents the vision encoder. The perturbation δ is bounded by a norm constraint to maintain perceptual similarity between I_{orig} and I_{adv}.

Attack Impact Assessment: The success of the attack is quantified by evaluating the model's response to the adversarially modified image $I_{\text{perturbed}} = I_{\text{orig}} + \delta$ which is passed through the VLM. The model's response $D_{\text{adv-response}}$ is observed to assess the effectiveness of the adversarial attack, with success being measured by the model generating content aligned with D_{target} rather than the content of I_{orig}.

The effectiveness of the attack is evaluated by computing the CLIP-based [9] text-text similarity between the response $D_{\text{clean-response}}$ generated for clean image I_{orig} and the target text D_{target}, and also between $D_{\text{adv-response}}$ and D_{target}

5. Mitigating the Attack Using Visual Latent Smoothing

To mitigate the effectiveness of adversarial attacks on vision-language models (VLMs), we propose a defense strategy called "Visual Latent Smoothing". This method aims to enhance the robustness of the VLM by leveraging multiple augmentations of the input image being passed to the vision encoder of the VLM.

Augmented Image Generation: Let I_{orig} denote the original image. We generate a set of augmented images $\{I_1, I_2, \ldots, I_n\}$ where each I_i is an augmented version of I_{orig} obtained through various transformations. For our experiments we have used Gaussian noise with various standard deviation and mean.

Token Averaging: All images, including I_{orig} and $\{I_1, I_2, \ldots, I_n\}$, are passed to the vision encoder of the VLM. Let T_j represent the tokens produced by the encoder for each image I_j. The defense mechanism involves averaging these tokens across all augmented instances, which is computed as follows:

$$T_{\text{avg}} = \frac{1}{n+1} \left(\text{Enc}(I_{\text{orig}}) + \sum_{i=1}^{n} \text{Enc}(I_i) \right),$$

where Enc denotes the encoding function of the vision encoder.

Robust Output Generation: The computed average tokens T_{avg} serve as the input for the subsequent processing stages of the VLM. This approach is intended to reduce any adversarial effects present in the token representation of the original image by incorporating diverse representations from its augmentations.

By using Visual Latent Smoothing, the model is less susceptible to adversarial perturbations as the averaging of encodings from multiple variations of the input image tends to stabilize the final token representation, thus preserving the integrity of the output against adversarial influences.

6. Experimental Setup

6.1. VLMs used for evaluations

For our experiments we have used BLIP [5], BLIP-2 (OPT), BLIP-2 (T5) [4], and LLaVa 1.5 13B [6]. We have chosen these models to cover models with varying model size to show the generalizability of our insights to models of different sizes

6.2. Dataset

This section describes the datasets employed in our study, which include the ImageNet [2] validation set, the COCO Captions dataset [1], and the MMMU benchmark [14]. Each dataset is used to assess different aspects of the robustness and susceptibility of the Vision Language Model (VLM) under various testing conditions.

6.2.1 ImageNet

We utilize a subset of the ImageNet validation set for the text-based robustness evaluation. This subset comprises 20 distinct classes, providing a total of 1000 images. These images offer a diverse array of visual contexts, allowing for a thorough assessment of the model's capability to generalize across different visual features under fixed and randomized answer conditions.

6.2.2 COCO Captions

For the vision-based adversarial attacks, we leverage the same ImageNet dataset but pair the images with target texts

Table 2: Performance of different models under various conditions. RN50, RN101, ViT-B/16, ViT-B/32, ViT-L/14 denotes various CLIP backbones. Ensemble is the average of clip similarity calculated with various CLIP backbones. For each model we provide the CLIP text-text similarity between $D_{\text{clean-response}}$ and D_{target} in first row, $D_{\text{adv-response}}$ and D_{target} in second row and $D_{\text{adv-response}}$ and D_{target} after Visual Latent Smoothing in third row

Model	Size	Attack	RN50	RN101	ViT-B/16	ViT-B/32	ViT-L/14	Ensemble (avg)
BLIP	224M	Clean image	0.494	0.476	0.502	0.522	0.363	0.471
		Adv. image	0.769	0.755	0.776	0.789	0.699	0.758
		Adv. image + latent smoothing	0.635	0.619	0.644	0.662	0.534	0.619
BLIP-2 OPT	2.7B	Clean image	0.481	0.465	0.494	0.517	0.355	0.462
		Adv. image	0.526	0.500	0.534	0.556	0.398	0.503
		Adv. image + latent smoothing	0.505	0.482	0.514	0.537	0.378	0.483
BLIP-2 T5 XXL	12.4B	Clean image	0.484	0.467	0.500	0.519	0.373	0.469
		Adv. image	0.519	0.493	0.532	0.552	0.403	0.500
		Adv. image + latent smoothing	0.500	0.478	0.512	0.535	0.384	0.482
LLaVa 1.5	13B	Clean image	0.362	0.430	0.392	0.426	0.260	0.374
		Adv. image	0.430	0.503	0.467	0.497	0.329	0.445
		Adv. image + latent smoothing	0.397	0.447	0.422	0.454	0.292	0.402

derived from the COCO Captions dataset. This unique combination facilitates the generation of adversarial examples that are visually similar to the original ImageNet images but are guided by contextually rich descriptions from COCO Captions, aiming to deceive the model into producing erroneous outputs aligned with the captions rather than the actual image content.

6.2.3 MMMU Benchmark

The MMMU Benchmark [14], standing for Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI, represents a cutting-edge dataset designed to challenge VLMs with complex multimodal, multilingual, and multidisciplinary scenarios. This benchmark is crucial for evaluating the resilience and adaptability of VLMs in realistic, diverse environments that mimic expert-level artificial general intelligence (AGI) tasks. It provides a stringent testbed for examining how well models can integrate and reason across different types of information, making it an ideal tool for assessing performance under extreme adversarial conditions and complex data integration challenges.

This benchmark comprises 1050 question-answer pairs, from which we have selectively utilized those that align with the requirements of Multiple-Choice Question Answering (MCQA) tasks. We specifically focus on pairs where a single image serves as the input, aligning with the VQA format of our study. To ensure the relevance and specificity of the dataset to MCQA, we have excluded open-ended questions that require a descriptive response about the image. This refinement gives a dataset of 942 question-answer pairs, allowing us to concentrate on evaluating the model's decision-making accuracy and robustness in a controlled multimodal context.

6.3. CLIP Score

CLIP (Contrastive Language-Image Pre-training) [9] score is a metric used to evaluate the alignment between textual descriptions and corresponding images, leveraging the capabilities of the CLIP model. The CLIP score quantifies this relationship by measuring the cosine similarity between these embeddings. Higher scores indicate a closer alignment, suggesting that the model perceives the text to effectively describe the visual content. This metric is particularly valuable in applications such as image captioning, content moderation, and enhancing search engine results, where understanding the nuanced interplay between text and images is crucial. Additionally, it has become an instrumental tool in the evaluation of synthetic image generation, where ensuring the relevance of generated images to specified textual inputs is paramount.

7. Results and Discussions

7.1. Assessing the Impact of Vision Attacks

To determine the impact of adversarial attacks, we compute the CLIP-based text similarity between the response generated by the VLM D_{response} and the intended target response D_{target} towards which the image is perturbed in the PGD attack.

Figure 1 provides a qualitative analysis of the adversarial attack's efficacy. The left half of the figure displays a caption generated by LLaVa 1.5 13B which accurately describes the clean image sourced from ImageNet. In contrast, the top right of the figure features an image that, although visually similar to the original to the human eye, has been altered through adversarial perturbation to align with a dif-

Table 3: CLIP image to text similarity between original image and the generated response under various conditions (higher the similarity, lesser the effectiveness of attack). RN50, RN101, ViT-B/16, ViT-B/32, ViT-L/14 denotes various CLIP backbones. Ensemble is the average of clip similarity calculated with various CLIP backbones. LLaVa 1.5 13B was used as the VLM. **Similarity between image and generated response under Visual Latent Smoothing is almost as good as the response with clean image**

Image to generated text similarity		**RN50**	**RN101**	**ViT-B/16**	**ViT-B/32**	**ViT-L/14**	**Ensemble (avg)**
I_{orig}	$D_{\text{clean-response}}$	0.240	0.462	0.307	0.295	0.254	0.311
	$D_{\text{adv-response}}$ + latent smoothing	0.250	0.440	0.307	0.300	0.242	0.308
	D_{response}	0.120	0.343	0.178	0.173	0.115	0.186
	D_{target}	0.1145	0.346	0.169	0.169	0.119	0.183

ferent target text. The response generated by the VLM from this adversarially perturbed image demonstrates the success of the attack; the model is deceived into generating a text description that diverges significantly from the actual content of the image. This outcome confirms the vulnerability of the model to adversarial manipulation, where minimal visual changes can lead to dramatically incorrect textual interpretations by the model.

To quantitatively evaluate the impact of the adversarial attack, we use models of varying sizes to show that the larger models are as susceptible to adversarial attacks as the smaller models. As illustrated in Table 2, for each model the response generated for clean image $D_{\text{clean-response}}$ has the least similarity with the target text D_{target}. Whereas the response generated for adversarial images $D_{\text{adv-response}}$ has more similarity to D_{target}, which shows the adversarial attack has been effective. Subsequently, we can notice that the similarity to D_{target} reduces after we employ Visual Latent Smoothing in all the cases. This shows that visual latent smoothing can be a potential simple defense mechanism that can be adopted to minimize the impact of adversarial attacks.

7.2. Improvements using Visual Latent Smoothing

Figure 4 further illustrates the effectiveness of the Visual Latent Smoothing technique (bottom right). When the adversarially perturbed image is processed through the latent smoothing pipeline prior to its presentation to the VLM, the model demonstrates a significant improvement in performance. VLM recognizes information of the original image, despite the adversarial manipulation. This result underscores the potential of latent smoothing as a robust defense mechanism that helps preserve the integrity of the model's output, ensuring that the descriptions generated are more aligned with the visual content of the original image rather than being misled by adversarial perturbations.

Table 2 demonstrates that Visual Latent Smoothing effectively diminishes the impact of adversarial attacks to an extent. To further evaluate its efficacy, we assessed whether the responses generated under Visual Latent Smoothing closely approximate those produced for the clean image. This involved computing the CLIP similarity between the original image's response and both the responses generated for the clean image and the adversarially perturbed image treated with Visual Latent Smoothing.

As indicated in Table 3, the CLIP similarity scores for responses post-Visual Latent Smoothing are nearly comparable to those for the clean image. This indicates that Visual Latent Smoothing not only mitigates the effects of adversarial attacks but also substantially recovers the model's original performance levels, even under compromised conditions.

7.3. Position Bias on the Adversarial Generated Images

Building on the adversarial attack strategies detailed in Section 4, we see the impact of adversarial perturbations on LLaVa 1.5 model when subjected to a Visual Question Answering (VQA) task, paralleling the experimental setup discussed in Section 3.1. n these experiments, adversarially generated images are used as inputs to assess how VLMs handle positional bias under adversarially challenging conditions.

Table 4 shows that under adversarial conditions, the model's ability to correctly answer questions based on the image content is significantly compromised, underscoring the effectiveness of adversarial attacks in exploiting the visual processing vulnerabilities of VLMs. This evaluation highlights the critical challenge posed by adversarial perturbations, as even slight manipulations can disrupt the model's visual understanding, leading to erroneous outputs or a complete breakdown in performance. The results emphasize the necessity for developing more robust defense mechanisms to protect VLMs against such adversarial threats in real-world applications.

Choices	Accuracy (%) ↑
Answer is always A	99.1
Answer is always B	3.8
Answer is always C	1.3
Answer is always D	2.0
Random position	28.9
Random choice	25.0

Table 4: Accuracy of the Lava 1.5 13B model using adversarial images under different answer positioning strategies.

7.4. Evaluating Robustness of VLMs on MMMU benchmark

This subsection details the results from our robustness evaluation of Vision Language Models (VLMs) using the MMMU benchmark under varied input conditions as seen in Section 3.1. The experiments were designed to assess the dependency of VLMs on visual and textual cues by systematically altering the inputs and observing the consequent changes in accuracy. Table 5 summarizes the accuracies obtained under each experimental condition.

Experiment	Accuracy (%) ↑
With Original Image	30.47
With Shuffled Options	28.87
With Black Image	28.34
Random Choice	26.80

Table 5: Accuracy of LLaVa 1.5 13B under different test conditions on the MMMU benchmark.

We use LLaVa 1.5 for the experiment and for the baseline condition where the models were tested with original, unaltered image-question pairs, achieving an accuracy of 30.47%. This scenario reflects the models' standard operational performance. On shuffling the answer choices, there was a slight reduction in accuracy to 28.87%, indicating a potential positional bias within the models, as they appeared to rely on the order of the answer options to guide their response.

Further modifications involved replacing the original image with a uniform black image, maintaining the question-answer pairs to test the models' reliance on visual versus textual information. This change resulted in a minimal decrease in accuracy to 28.34%, suggesting that while visual inputs do not contribute much to the decision-making process in the challenging MMMU dataset, the models are significantly capable of deriving answers from textual cues alone. o benchmark the models' effectiveness beyond mere chance, we also tested them in a random choice scenario, where accuracy fell to 26.80%. This demonstrates that the models perform above randomness, leveraging textual content to inform their decisions, even in the absence of informative visual cues.

8. Conclusion

In this work, we have undertaken a comprehensive investigation into the adversarial robustness of Large Vision Language Models (VLMs), highlighting their vulnerabilities to both simple text-based perturbations and sophisticated vision-based adversarial attacks. Our experiments have shown that seemingly minor manipulations, such as random ordering in Visual Question Answering tasks, can substantially degrade model performance. Additionally, we demonstrated that targeted attacks on the visual modality can not only mislead the models but also potentially enable adversaries to generate malicious content. This capacity to induce different or harmful outputs underscores the critical need for robust defense mechanisms to secure VLMs in applications where reliability and safety are paramount.

To counter these threats, we introduced Visual Latent Smoothing, a novel defense strategy inspired by traditional noise reduction techniques. This method has proven effective in mitigating the effects of adversarial perturbations, significantly restoring model accuracy towards baseline levels under attack conditions. The success of Visual Latent Smoothing in preserving the integrity of model outputs highlights its potential as a viable solution for enhancing the security of multimodal AI systems.

9. Future Work

Further research is essential to refine and broaden the application of defense strategies such as Visual Latent Smoothing. In future, we wish to explore these defenses across a wider range of VLM architectures to assess their effectiveness in various adversarial contexts. This will help identify any limitations and inspire advancements in defensive technologies. And how the defense strategy generalizes to different VLMs

Additionally, applying our findings to real-world scenarios is crucial. Deploying robust VLMs in critical sectors such as healthcare, finance, and autonomous systems will test their adaptability and reliability under diverse operational conditions. Such practical applications will provide invaluable insights into the real-world efficacy of these models.

Moreover, it is vital to integrate ethical considerations into the development of VLMs. Future work should focus on ensuring that improvements in adversarial robustness do not compromise ethical standards or introduce unintended biases. This approach will support the development of VLMs that are not only technically sound but also ethically responsible, fostering trust and safety in AI deployments.

10. Contribution

My individual contribution to this work encompassed several key areas. I was responsible for developing a custom dataset for Visual Question Answering (VQA) tasks and establishing the framework for conducting text-based adversarial attacks. Additionally, I analyzed positional bias in images manipulated through adversarial attacks. For the MMMU benchmark, I handled the preprocessing of the dataset, setup the evaluation pipeline, and assessed the performance of Vision Language Models (VLMs) on this benchmark. I also designed and executed the experiments to further explore the robustness of VLMs within the MMMU benchmark context.

References

[1] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[3] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, pages 656–672. IEEE, 2019.

[4] 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*, pages 19730–19742. PMLR, 2023.

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 17–23 Jul 2022.

[6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[8] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21527–21536, 2024.

[9] 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*, pages 8748–8763. PMLR, 2021.

[10] Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *ArXiv*, abs/2310.03684, 2023.

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[12] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multimodal language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[13] Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pages 10693–10705. PMLR, 2020.

[14] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

[15] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Yongshuo Zong, Tingyang Yu, Bingchen Zhao, Ruchika Chavhan, and Timothy Hospedales. Fool your (vision and) language model with embarrassingly simple permutations. *arXiv preprint arXiv:2310.01651*, 2023.