Improving Adversarial Robustness in Vision-Language Models with Architecture and Prompt Design

Rishika Bhagwatkar[1,2] **Shravan Nayak**[1,2] **Pouya Bashivan**[1,3] **Irina Rish**[1,2]

[1]Mila - Quebec AI Institute [2]Université de Montréal [3]McGill University

Correspondence: rishika.bhagwatkar@umontreal.ca

Abstract

Vision-Language Models (VLMs) have seen a significant increase in both research interest and real-world applications across various domains, including healthcare, autonomous systems, and security. However, their growing prevalence demands higher reliability and safety including robustness to adversarial attacks. We systematically examine the possibility of incorporating adversarial robustness through various model design choices. We explore the effects of different vision encoders, the resolutions of vision encoders, and the size and type of language models. Additionally, we introduce novel, cost-effective approaches to enhance robustness through prompt engineering. By simply suggesting the possibility of adversarial perturbations or rephrasing questions, we demonstrate substantial improvements in model robustness against strong image-based attacks such as Auto-PGD. Our findings provide important guidelines for developing more robust VLMs, particularly for deployment in safety-critical environments where reliability and security are paramount. These insights are crucial for advancing the field of VLMs, ensuring they can be safely and effectively utilized in a wide range of applications.

1 Introduction

VLMs have emerged as a cornerstone of artificial intelligence in recent years. By enabling seamless integration of visual and linguistic information, these models have the capacity to perform a wide range of tasks from image captioning, to visual question answering (VQA), and cross-modal retrieval (Liu et al., 2023; Laurenccon et al., 2023; Awadalla et al., 2023; Radford et al., 2021). Despite the impressive advances in VLMs' performance through novel architectures and scaling network size, it has been shown that, like many other neural network models, VLMs are also not immune to adversarial vulnerabilities — subtle, intentionally crafted perturbations to input data that can lead to significant errors in the output (Schlarmann et al., 2024). These vulnerabilities can mislead users with harmful or toxic responses, undermining the model's robustness and integrity.

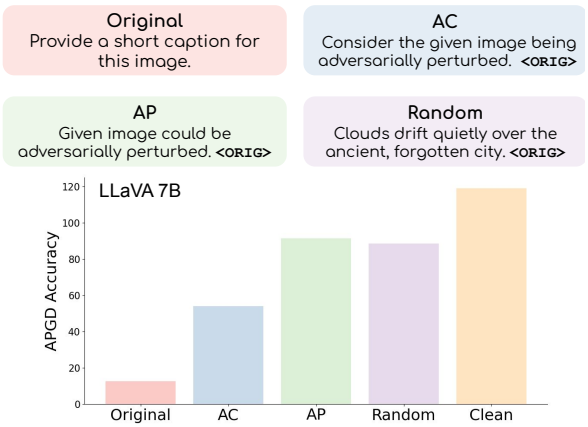

Figure 1: Performance of LLaVA-7B on the COCO dataset when the adversarial images are given along with different types of prompts (Original, AC, AP and Random). Clean accuracy represents the model's performance on unperturbed images.

Among various types of adversarial attacks, white-box attacks have remained one of the most widely studied type of attacks. These attacks assume complete access to a model's parameters, enabling attackers to devise strategies that are closely tailored to particular models, making them the hardest to defend against. This is specially relevant in practice considering that many VLMs are open-source, enabling attackers to easily analyze and exploit them. This imposes utmost importance on incorporating robustness during design and development. In this study, we evaluate how several design choices–the vision encoder (or ensemble), image resolution, large language model (LLM), and its size–influence their susceptibility to white-box adversarial attacks on the input images.

In addition to design choices, the selection

Findings of the Association for Computational Linguistics: EACL 2024, pages 17003–17020

and quality of prompts can significantly impact the performance and robustness of VLMs (Awal et al., 2024). Effective prompts can enhance the model's understanding and response to inputs, improving their robustness to adversarial attacks. Recent works have focused on adversarial training to increase robustness (Schlarmann et al., 2024; Mao et al., 2023), but these methods is resource-intensive and costly, often requiring millions of samples (Wang et al., 2023). As a practical and cost-effective alternative, we investigate whether prompt engineering can enhance the adversarial robustness of VLMs. This approach explores if simple linguistic modifications can increase robustness, offering a low-cost alternative to adversarial training.

In this study, we aim to assess the impact of VLMs' architectural choices and textual input (i.e. prompts) on their robustness to vision-based adversarial attacks through the following questions.

- Which vision encoder demonstrates better robustness?
- Does increasing the input image resolution enhance the VLM's robustness?
- How easily can VLMs be compromised when using an ensemble of vision encoders?
- Which LLM is more robust to adversarial attacks?
- Does increasing the size of the LLM improve the VLM's robustness?
- Can prompt engineering effectively enhance adversarial robustness?

Our results show that:

- Training vision encoders across diverse data distributions improves robustness against simpler attacks but offers minimal advantage against iterative attacks.
- Increasing the resolution of image encoders improves the clean accuracy but does not enhance adversarial robustness.
- Using multiple vision encoders does not guarantee robustness, as knowledge of the weakest encoder is enough to compromise the system.
- Amoung LLMs, Mistral has the best robust accuracy and Vicuna has the worst robust accuracy under the strongest iterative attacks.
- Scaling up the size of the language model does not increase the model's robustness to vision-based attacks.
- Simply indicating the possibility of a perturbed image or adding a random string in the prompt can help significantly improve robustness.

2 Related Works

Vision Language Models. VLMs traditionally align visual tokens from the vision encoder with the linguistic space of the language model using various mapping networks, such as the Q-former in BLIP2 (Li et al., 2023) and the multilayer perceptron in LLaVA (Liu et al., 2023). Recent studies investigate how choices like vision encoder type, language model, resolution of images, and training duration affect the accuracy on clean inputs (Karamcheti et al., 2024). In contrast, our study specifically aims to explore if choices that improve clean accuracy consistently improve robust accuracy.

Adversarial Robustness of VLMs Research into the adversarial robustness of multi-modal foundation models like BLIP2 (Li et al., 2023), OpenFlamingo (Awadalla et al., 2023), CLIP (Radford et al., 2021), and LLaVA (Liu et al., 2023) has highlighted their susceptibility to both targeted and untargeted visual attacks (Cui et al., 2023b; Zhao et al., 2023). Studies also explore the potential of using pretrained VLMs to craft adversarial visual and textual perturbations that can compromise black-box models fine-tuned for various tasks (Zhao et al., 2023; Dong et al., 2023). Additionally, the transferability of these attacks is well-studied, with techniques developed to enhance efficacy using surrogate models (Yin et al., 2023). To understand how easily VLMs can be compromised, this work exhaustively evaluates various VLMs based on design choices, using white-box attacks in an untargeted setting.

Advancements in Defense Mechanisms Many studies focusing on the adversarial robustness of VLMs using CLIP as a vision encoder have revealed its susceptibility to adversarial attacks (Fang et al., 2022; Tu et al., 2023; Nguyen et al., 2022). To counter this, TeCoA (Mao et al., 2023) proposes adversarial fine-tuning to maintain zero-shot capabilities. Further, RobustCLIP (Schlarmann et al., 2024) proposes an unsupervised method leveraging adversarial training on the ImageNet dataset (Deng et al., 2009) to improve robustness across vision-language tasks. Additionally, efforts to enhance robustness include prompt tuning, where one study suggests enriching prompts with contextual image-derived information (Cui et al., 2023a). Another approach optimizes prompts

through adversarial fine-tuning on ImageNet with specific parameters (Zhang et al., 2023).

Our research, however, focuses on analyzing the impact of substantially inexpensive prompt engineering techniques on model performance without additional training or image-based information extraction.

3 Experimental Setup

In this section, we outline the attack setups, the tasks, and the specific models examined in our model design choice experiments.

3.1 Attack Setup

This work focuses on white-box gradient-based untargeted attacks on image inputs, where it is assumed that the attacker has complete knowledge of the model, including its architecture and parameters. The objective of crafting adversarial samples in this scenario is to subtly perturb the input so that the model produces an incorrect output. Mathematically, it can be formulated as $\max_\delta \mathcal{L}(f(x+\delta), y)$ where f is the model, x is the original input, δ is the adversarial perturbation learnt within the $\|\delta\|_\infty \leq \epsilon$ constraint and y is the original label. Hence the goal is to find a perturbation δ that maximizes the loss while respecting the perturbation bound.

Our evaluation encompasses three gradient-based adversarial attacks, ordered in increasing effectiveness: Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014), Projected Gradient Descent (PGD) (Madry et al., 2017), and Auto-PGD (APGD) (Croce and Hein, 2020). We employ PGD and APGD attacks with 100 iterations, while FGSM uses a single iteration by design. Our evaluation focuses on ℓ_∞ bounded perturbations, with the perturbation magnitudes $\epsilon \in \{4/255, 8/255, 16/255\}$. This range allows us to systematically assess the robustness of models against varying strengths of adversarial attacks. Our focus on white-box attacks stems from their status as the strongest to defend against, given that the adversary has complete information about the model, including the weights. Moreover, existing black-box image-based attacks are typically designed for classification settings, making it impractical to adapt them for generative models (Andriushchenko et al., 2020; Chen et al., 2017).

3.2 Tasks

Our evaluation covers two primary tasks: Image Captioning and VQA. For image captioning, we use the validation splits of the COCO (Lin et al., 2014) and Flickr30k (Plummer et al., 2015) datasets to assess caption accuracy and relevance. In the VQA domain, we evaluate using the validation splits of VQAv2 (Antol et al., 2015), TextVQA (Singh et al., 2019), OK-VQA (Marino et al., 2019), and VizWiz (Gurari et al., 2018) datasets. We report the robust VQA accuracy for datasets associated with VQA tasks and robust CIDEr scores for the captioning datasets. We randomly sample 1000 examples from the validation set of each task and use this for the adversarial evaluations of all models to ensure a fair comparison. The models selected for evaluating the impact of design choices on adversarial robustness are detailed in Table 1. All evaluated models are from the work on Prismatic VLMs (Karamcheti et al., 2024).

	Vision Encoder	Language Model
Vision Encoder	CLIP ViT-L/14 @ 224px SigLIP ViT-SO/14 @ 224px DINOv2 ViT-L/14 @ 224px ImageNet-21K+1K ViT-L/16 @ 224px	Vicuna v1.5 7B
Image Resolution	CLIP ViT-L/14 @ 224px SigLIP ViT-SO/14 @ 224px CLIP ViT-L/14 @ 336px SigLIP ViT-SO/14 @ 384px	Vicuna v1.5 7B
LLM	CLIP ViT-L/14 @ 336px	Vicuna v1.5 7B Llama-2 7B Llama-2 Chat 7B Mistral v0.1 7B Mistral Instruct v0.1 7B
Size of LLM	CLIP ViT-L/14 @ 336px	Vicuna v1.5 7B Vicuna v1.5 13B
Ensemble of vision encoders	DINOv2 ViT-L/14 + CLIP ViT-L/14 @ 336px DINOv2 ViT-L/14 + SigLIP ViT-L/14 @ 384px	Vicuna v1.5 7B

Table 1: Models used for the evaluation of various components of VLMs. Each row corresponds to a VLM built with the given vision encoder and LLM.

4 Results

In our analysis, we explore the possibility of incorporating adversarial robustness through various model design choices. Specifically, we focus on: (a) the choice of vision encoder; (b) the input resolution of the vision encoder; (c) ensembles of multiple vision encoders; (d) the choice of LLM; and (e) the size of LLM. Each of these aspects is detailed in the sections below. We report results using the standard $\epsilon = 8/255$ value for vision attacks. Appendix A presents results with $\epsilon = \{4/255, 16/255\}$. We also present an analysis of black-box attacks to further justify the motivation behind selecting white-box attacks for our study in Appendix D.

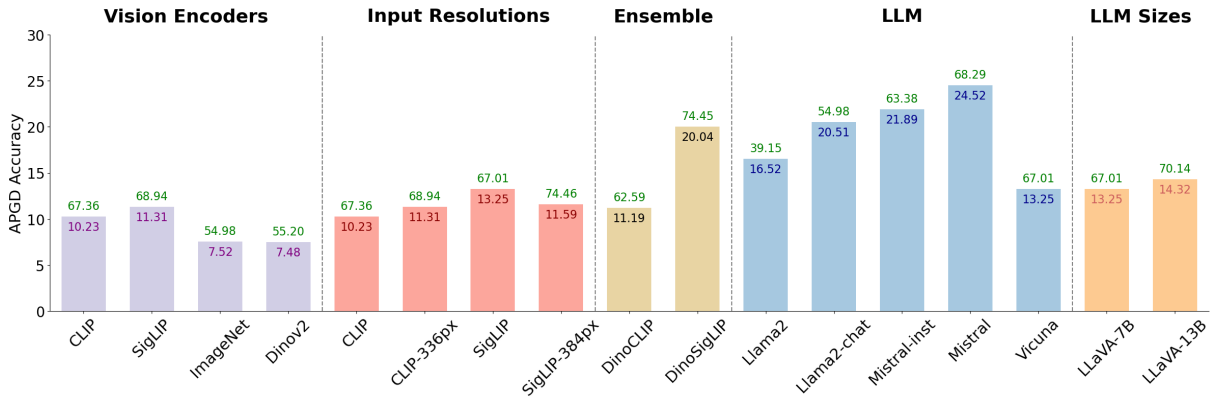

Figure 2: Comparison between VLMs with different vision encoders, different input resolutions and different LLM and their sizes. The comparison is based on the APGD accuracy averaged over all tasks as shown in Tables 2, 3, 4, 5, and 6. **Note:** From Table 4, we only plot case (a) where the adversary has knowledge of only the weakest vision encoder in the ensemble.

Impact of Vision Encoder. We first evaluate the effect of vision encoders, each trained under distinct conditions, on adversarial robustness. We compare VLMs that use four different image encoders: CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), DINOv2 (Oquab et al., 2023), and ImageNet-trained ViT (Dosovitskiy et al., 2020). As shown in Table 2, SigLIP slightly outperforms CLIP, with both noticeably surpassing DINOv2 and ImageNet-trained VLMs on weaker attacks. However, the difference diminishes for stronger attacks. We hypothesize that the Vision Transformer (ViT) used in CLIP and SigLIP has been trained across a wide spectrum of internet-collected images and, hence, has seen many more distributions during training than DINOv2 and ImageNet. The results also resonate with the choice of vision encoders in recent state-of-the-art VLMs (Liu et al., 2023; Karamcheti et al., 2024).

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQav2
CLIP	None	116.35	76.15	59.48	37.10	41.18	73.89
	FGSM	94.18	56.83	47.58	22.40	33.70	57.85
	PGD	13.36	9.11	13.90	7.42	8.67	31.65
	APGD	6.32	4.41	10.11	4.80	8.16	27.56
SigLIP	None	118.69	78.29	60.92	38.82	42.43	74.49
	FGSM	94.09	54.43	48.24	24.38	39.85	60.63
	PGD	20.46	11.03	14.96	7.52	10.91	34.70
	APGD	7.32	4.87	10.44	5.30	9.54	30.38
DINOv2	None	104.84	54.78	57.00	10.37	38.07	64.80
	FGSM	81.81	38.24	40.08	8.03	33.78	46.24
	PGD	3.07	1.80	9.06	1.83	10.60	25.21
	APGD	2.09	1.22	7.16	2.00	10.20	22.47
In1k	None	101.59	54.92	56.34	10.70	39.29	68.36
	FGSM	69.13	32.42	38.40	6.60	32.45	46.13
	PGD	5.38	3.43	9.26	1.79	10.78	22.73
	APGD	2.74	2.04	7.58	1.52	10.22	20.79

Table 2: Comparison between VLMs having different image encoders but the same LLM - Vicuna v1.5 7B. We highlight the best robust FGSM, PGD, and APGD accuracies. Note: Higher values (↑) indicate better performance.

CLIP and SigLIP are more robust than DINOv2 and ImageNet vision encoders.

Resolution of Vision Encoder. Generally, a higher input resolution improves the quality of visual representations, potentially boosting model performance (Karamcheti et al., 2024). Owing to the availability of high-resolution variants, we specifically evaluate models equipped with CLIP and SigLIP vision encoders at two distinct resolutions to thoroughly understand these effects. Based on Table 3, while increasing the resolution of CLIP models enhances robustness against stronger attacks (on most tasks), the effectiveness of the increased resolution of SigLIP models appears to be task-dependent. However, we observe that the robust accuracy significantly deteriorates under APGD attacks in all cases except for VQAv2.

CLIP shows higher robust accuracy with higher resolution, but for SigLIP, the results are inconclusive.

Ensemble of vision encoders. We also explore the vulnerability of VLMs that employ an ensemble of vision encoders. Although recent studies suggest that multiple encoders can significantly improve

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQav2
CLIP-224px	None	116.35	76.15	59.48	37.10	41.18	73.89
	FGSM	94.18	56.83	47.58	22.40	33.70	57.85
	PGD	13.36	9.11	13.90	7.42	8.67	31.65
	APGD	6.32	4.41	10.11	4.80	8.16	27.56
CLIP-336px	None	119.02	77.21	59.18	36.73	39.94	70.00
	FGSM	95.55	63.10	48.20	24.03	35.24	57.19
	PGD	22.84	13.87	20.46	10.13	22.73	27.07
	APGD	12.54	7.15	15.68	8.08	9.31	26.73
SigLIP-224px	None	118.69	78.29	60.92	38.82	42.43	74.49
	FGSM	94.09	54.43	48.24	24.38	39.85	60.63
	PGD	20.46	11.03	14.96	7.52	10.91	34.70
	APGD	7.32	4.87	10.44	5.30	9.54	30.38
SigLIP-384px	None	124.11	87.08	62.18	55.05	41.14	77.22
	FGSM	92.39	57.55	51.38	32.91	37.28	62.47
	PGD	15.69	8.29	18.04	9.61	9.98	35.97
	APGD	6.90	3.22	12.72	6.73	8.77	31.21

Table 3: Comparison between VLMs having different input resolutions of CLIP and SigLIP. All of them have the same LLM: Vicuna v1.5 7B. We highlight the best robust FGSM, PGD, and APGD accuracies. Note: Higher values (↑) indicate better performance.

performance (Karamcheti et al., 2024; Kar et al., 2024), our research aims to assess the difficulty of compromising such models. To address this, we break our investigation into two key questions. (a) Is it sufficient for adversaries to have knowledge of only the weakest vision encoder? (b) Do we need to pass the adversarially perturbed image through all the vision encoders or only the weakest one?

	Attack	Task					
		Coco	Flickr30k	OK-VQA	TextVQA	VizWiz	VQav2
Passed only via weakest							
DinoCLIP	None	113.75	74.16	58.88	15.08	39.30	74.35
	FGSM	100.75	55.29	47.54	8.98	39.11	58.20
	PGD	13.14	7.40	12.44	3.11	12.94	33.32
	APGD	5.99	4.36	11.08	2.88	12.51	30.29
DinoSigLIP	None	125.94	85.44	61.12	50.52	44.27	79.39
	FGSM	109.79	73.84	53.94	40.84	41.84	67.75
	PGD	39.76	22.64	19.30	12.83	13.37	39.78
	APGD	25.23	15.72	17.26	12.03	12.25	37.75
Passed through all							
DinoCLIP	None	113.75	74.16	58.88	15.08	39.30	74.35
	FGSM	99.87	55.08	47.70	9.14	39.32	58.56
	PGD	17.33	10.11	14.12	3.22	13.74	33.72
	APGD	6.23	4.00	10.66	3.06	11.79	29.89
DinoSigLIP	None	125.94	85.44	61.12	50.52	44.27	79.39
	FGSM	109.14	71.25	54.26	39.91	42.95	67.24
	PGD	19.22	11.65	16.78	9.32	9.20	39.67
	APGD	24.15	15.64	16.74	13.03	13.07	37.14

Table 4: Comparison between VLMs that have an ensemble of vision encoders. The comparison is made when only the input to the DINOv2 image encoder is perturbed. We highlight the best robust FGSM, PGD, and APGD accuracies. Note: Higher values (↑) indicate better performance.

From Table 4, we observe a minimal difference in the robust accuracies across various attacks in both cases. This indicates that an adversary only needs knowledge of the weaker model (DINOv2 in our case) in the ensemble to compromise the system. Knowledge of the other models is not required; simply ensuring the perturbed image is processed by the weaker model is sufficient to degrade performance. Further, the highest robust accuracies are achieved with the combination of DINOv2 and SigLIP, suggesting that having a better vision encoder in the ensemble can provide some improvement. This finding underscores the limited advantage of current ensemble methods in enhancing robustness. This lays the foundation for further studies involving more than two encoders that will help draw broader conclusions about the efficacy of ensemble approaches in enhancing adversarial robustness.

Adversarially perturbing the input of the weaker vision encoder (like DINOv2) is sufficient to compromise the entire system.

Choice of LLM. We next focus on comparing VLMs that utilize different LLMs to determine their effectiveness in handling adversarially perturbed visual inputs. Our experiments focus solely on visual perturbations without altering textual inputs. We evaluate several LLMs, including Vicuna (Zheng et al., 2024), Mistral (Teknium, 2023), and Llama2 (Touvron et al., 2023), as well as instruction-tuned versions of Mistral and Llama2, to understand their relative robustness in mapping adversarially perturbed vision tokens to the language space. To ensure fairness, all models were specifically selected with the same size (7B).

From Table 5, we can observe several key insights: (i) The Mistral base model exhibits the best robust accuracy across tasks for all attacks and benchmarks. This finding suggests that integrating Mistral into upcoming VLMs could significantly enhance their robustness. (ii) Llama2 performs poorly on captioning tasks due to its tendency to be overly verbose. Although this behavior can be mitigated through prompt engineering, doing so would not allow for a fair comparison with other models. Additionally, while Llama2-chat shows some improvement in captioning tasks, the performance remains subpar. (iii) Vicuna demonstrates the worst robust accuracy against APGD. This is particularly concerning given that many open-source VLMs utilize Vicuna as their LLM (Liu et al., 2023). This vulnerability highlights the need for careful consideration when choosing the LLM

	Attack	Task					
		Coco	Flickr30k	OK-VQA	TextVQA	VizWiz	VQav2
llama2	None	6.95	3.36	60.66	44.82	42.41	76.73
	FGSM	6.42	3.14	50.26	29.14	37.99	65.94
	PGD	1.06	0.56	22.84	13.64	18.35	42.18
	APGD	1.13	0.66	23.19	12.22	21.21	40.73
llama2-chat	None	68.31	37.34	61.24	44.85	42.06	76.05
	FGSM	65.66	33.89	50.58	30.90	36.60	64.76
	PGD	18.94	9.90	23.85	13.84	17.86	39.34
	APGD	19.15	10.00	23.11	12.11	20.11	38.60
mistral-instruct-v0.1	None	94.55	63.92	60.76	43.84	40.57	76.64
	FGSM	81.65	51.82	50.40	29.00	34.70	63.72
	PGD	23.93	14.06	24.01	15.09	12.43	43.60
	APGD	23.48	14.37	22.84	13.26	15.50	41.86
mistral-v0.1	None	109.23	74.60	60.60	45.14	43.70	76.47
	FGSM	97.33	63.13	51.56	28.95	40.23	64.00
	PGD	31.03	19.97	23.07	15.78	15.51	44.45
	APGD	28.42	19.14	22.43	13.58	20.92	42.61
Vicuna	None	119.02	77.21	59.18	36.73	39.94	70.00
	FGSM	95.55	63.10	48.20	24.03	35.24	57.19
	PGD	22.84	13.87	20.46	10.13	22.73	27.07
	APGD	12.54	7.15	15.68	8.08	9.31	26.73

Table 5: Comparison between models having different LLMs of 7B size. Every model has the same vision encoder, CLIP-336. We highlight the best robust FGSM, PGD, and APGD accuracies. Note: Higher values (↑) indicate better performance.

for VLMs, especially for applications requiring high robustness against adversarial attacks.

Mistral yields a higher robust accuracy compared to Llama2, Llama2-chat, and Mistral-instruct, while Vicuna (one of the most widely used LLM) achieves the lowest robust accuracy.

Size of LLM. We evaluate a series of VLMs utilizing the same vision encoder, and same LLM architecture, with the only difference being the LLM's size. Although Vicuna is the most vulnerable, owing to its popularity (Schlarmann et al., 2024), we examine VLMs equipped with the Vicuna language model (Zheng et al., 2024) in two sizes: 7B and 13B. According to the results in Table 6, the model's vulnerability to adversarial attacks and the significant drop in robust accuracy remain consistent, regardless of the model's scale. Surprisingly, the smaller 7B model is superior against the strongest attack (APGD) for half the datasets. Hence, increasing the size of the LLM does not seem to enhance robustness. One potential reason for this could be that adversarial attacks compromise the representations from the vision encoder. As a result, LLMs even at the 13B scale may struggle to effectively interpret these flawed representations, making robustness to image-based attacks less sensitive to LLM size. Therefore, enhancing the vision encoder's adversarial robustness is sufficient as shown in prior work (Schlarmann et al., 2024).

Larger language models do not guarantee increased adversarial robustness.

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQav2
LLaVA-7B	None	119.02	77.21	59.18	36.73	39.94	70.00
	FGSM	95.55	63.10	48.20	24.03	35.24	57.19
	PGD	22.84	13.87	20.46	10.13	22.73	27.07
	APGD	12.54	7.15	15.68	8.08	9.31	26.73
LLaVA-13B	None	123.71	77.63	62.86	40.04	41.19	75.39
	FGSM	106.40	64.93	50.90	26.28	36.48	62.49
	PGD	14.60	8.96	15.65	9.08	23.55	34.49
	APGD	6.98	4.35	23.13	10.45	7.47	33.56

Table 6: Comparison between models having different scales of the LLM. Both of the models have the same vision encoder, CLIP-336, but different scales of the LLM. We highlight the best robust FGSM, PGD, and APGD accuracies. Note: Higher values (↑) indicate better performance.

Prompt Engineering. Considering that our adversarial examples are generated solely by perturbing visual inputs, we hypothesize that modifying the original prompts could be particularly effective in countering the effects of such perturbations. We aim to (a) identify effective prompts and (b) evaluate the vulnerability of models to adversarial attacks when these prompts are used. We test this hypothesis with the LLaVA 7B and 13B models, employing different types of prompts for COCO and VQAv2. Our evaluation includes adversarial examples created using FGSM, PGD, and APGD attacks, with PGD, APGD based on 100 iterations.

Captioning: Our experiments evaluated various prompt engineering strategies, including: **(a) Original** - using the original prompt as the baseline and **(b) Adversarial Certainty (AC) Prompt** - explicitly informing the model that the image is adversarially perturbed. From the results presented in Fig. 1 and Table 27, we observe several key insights. (i) Prompting helps in all cases (except one) for both the models. (ii) For the smaller model, this enhances the robust accuracy, but has the worst clean accuracy (almost the same as the robust accuracy). This implies the need for a reliable detector. To mitigate this requirement, we investigate two more strategies. **(c) Adversarial Possibility (AP) Prompt** - suggesting the possibility that the image might be adversarially perturbed; and **(d) Random**

	Prompt	FGSM	PGD	APGD	Clean
LLaVA 7B	Original	95.55	22.84	12.54	119.02
	AC	63.86	60.01	54.05	64.11
	AP	105.41	101.61	91.46	112.78
	Random str	108.23	105.35	94.61	120.90
	Random sent	101.12	97.11	88.45	108.00
	Clean Acc				119.02
LLaVA 13B	Original	106.40	14.60	6.98	123.71
	AC	113.77	106.40	114.65	122.10
	AP	114.48	108.83	113.54	125.28
	Random str	110.74	105.15	111.69	120.49
	Random sent	113.29	106.71	111.13	120.72
	Clean acc				123.71

Table 7: Performance of LLaVA models on image captioning (COCO) when adversarially perturbed images (using $\epsilon = 8/255$) are provided along with different types of prompts. Note: Higher values (↑) indicate better performance.

- appending a random sentence or string at the beginning of the prompt. (iii) Interestingly, both these strategies present the best robust accuracies across both models. (iv) On the other hand, the larger model's robust accuracy is almost the same for all prompt types. Thus, an adversarial sample detector is not necessary for large models. (v) Finally, the improvements from simply adding a random string or sentence are substantial and comparable to the effects observed with the AP prompt. This suggests that models pay more attention to inputs when they struggle to establish a clear relationship between them. The prompts are presented in Table 26.

Visual Question Answering: Here, we explored four strategies: **(1) Rephrase** - rephrasing the original question to create a semantically similar question; **(2) Expand** - increasing the length of the questions; **(3) Adversarial Certainty (AC) Prompt** - explicitly informing the model that the image is adversarially perturbed; and **(4) Adversarial Possibility (AP) Prompt** - suggesting the possibility that the image might be adversarially perturbed. We utilize a Mistral 7B LLM (Teknium, 2023) to generate questions according to the above-mentioned strategies. All the instructions used to obtain the modified questions are listed in Table 25. According to the results presented in Table 28, simply rephrasing the questions significantly improved performance compared to the other methods, such as extending the question length or explicitly warning about potential adversarial perturbations. Moreover, indicating the possibility of an adversarial perturbation yielded the best robust accuracies, reinforcing our observations from the captioning task discussed above.

Evaluating the adversarial robustness of effective prompts. Since the rephrasing strategy can be naively broken, our goal is to evaluate how easily models can be compromised even while using different prompt strategies–AC, AP, and Random–during the generation of adversarial samples. To achieve this, we apply the same attacks on the COCO and VQAv2 datasets, conducting iterative attacks over 100 iterations. Our findings presented in Tables 11 and 12 reveal that, even with AC and AP prompts employed during generation, adversarial samples can still be easily crafted. This suggests that an attacker with knowledge of the specific prompts used to enhance adversarial robustness can still effectively generate adversarial samples. This highlights a critical vulnerability: while prompt engineering can improve robustness, it is not foolproof, and attackers can exploit these strategies if they are aware of the techniques being used.

Adding random strings or indicating possible adversarial perturbations enhances robustness, but attackers aware of these strategies can still generate effective adversarial samples.

5 Discussion and Conclusion

Our evaluation provides critical insights on how we can incorporate adversarial robustness in VLMs through various design choices. First, we find that vision encoders trained across diverse data distributions are more robust against simpler attacks but lose the advantage against complex attacks. Second, neither increasing the input resolution of vision encoders nor scaling up the language model enhances robustness. Third, our findings show that the Mistral base model exhibits the best robust accuracy across most tasks and attacks. Fourth, using multiple vision encoders does not guarantee robustness; an adversary only needs knowledge of the weakest encoder to compromise the entire system. The combination of DINOv2 and SigLIP yielded the highest robust accuracies, but the gains were not significant, highlighting the limited advantage of current ensemble methods. Hence, our study points to the need for comprehensive analysis with more encoders to fully understand the efficacy of ensemble approaches in improving

Prompt Type	Prompt
Original	`Provide a short caption for this image.`
AC	`Consider the given image being adversarially perturbed. Provide a short caption for this image.`
AP	`Given image could be adversarially perturbed. Provide a short caption for this image.`
Random sent.	`Clouds drift quietly over the ancient, forgotten city. Provide a short caption for this image.`
Random str.	`ryFo8ZVcyNMtLgryNOg64UTjySyEb79e5aq6IJxGuz0GzWNtoz. Provide a short caption for this image.`

Table 8: Various types of prompts tested for image captioning.

Task	Instruction
Rephrase	`You will be given a question. Your task is to rephrase the question so that it is semantically similar to the original question and will have the same answer as the original question.`
Expand	`You will be given a short question. Your task is to generate a longer question so that it is semantically similar to the original question and will have the same answer as the original question.`
AC	`You will be given a question. However, the image associated with the question will be adversarially perturbed. Your task is to generate a longer question so that it is semantically similar to the original question and will have the same answer as the original question.`
AP	`You will be given a question. However, the image associated with the question could be adversarially perturbed. Your task is to generate a longer question so that it is semantically similar to the original question and will have the same answer as the original question.`

Table 9: Instructions used to obtain modified questions for VQAv2.

	Prompt	FGSM	PGD	APGD	Clean
LLaVA 7B	Original	57.19	27.07	26.73	70.00
	Rephrase	59.01	58.03	48.84	68.30
	AC	60.21	58.82	50.68	69.99
	AP	60.13	58.81	49.95	69.78
	Expand	48.59	48.54	42.24	57.14
	Clean Acc				70.00
LLaVA 13B	Original	62.49	34.49	33.56	75.39
	Rephrase	59.01	60.05	54.77	71.02
	AC	51.38	61.49	55.56	72.00
	AP	63.59	61.29	63.2	71.79
	Expand	53.03	50.03	45.93	58.59
	Clean Acc				75.39

Table 10: Performance of LLaVA models on VQAv2 when adversarially perturbed images (using $\epsilon = 8/255$) are provided along with questions generated using different types of prompts. Note: Higher values (↑) indicate better performance.

adversarial robustness.

To explore inexpensive strategies for improving robustness during deployment, we propose simple prompt engineering techniques. Our experiments reveal that even naively rephrasing the questions significantly improves robustness on VQA. Similarly, merely suggesting the possibility of an adversarial image during captioning leads to a notable performance boost. More importantly, we find that it is not required to add additional context from the image or fine-tune additional tokens to make models adversarially robust, as opposed to prior

	Prompt Type	FGSM	PGD	APGD
LLaVA 7B	Original	57.19	27.07	26.73
	AP	57.23	27.51	26.08
	AC	58.21	27.61	24.96
	Clean Acc			70.00
LLaVA 13B	Original	57.19	27.07	26.73
	AP	60.58	34.88	30.37
	AC	60.99	32.58	29.77
	Clean Acc			75.39

Table 11: Adversarial robustness of LLaVA 7B and 13B using different prompt types for VQAv2. Note: Higher values (↑) indicate better performance.

	Prompt Type	FGSM	PGD	APGD
LLaVA 7B	Original	95.55	22.84	12.54
	AC	51.79	16.48	14.60
	AP	97.41	30.98	26.56
	Random str	102.80	34.38	28.18
	Clean Acc			119.02
LLaVA 13B	Original	106.40	14.60	6.98
	AC	101.91	35.00	30.13
	AP	105.74	38.13	31.21
	Random str	100.77	36.71	30.28
	Clean Acc			123.71

Table 12: Adversarial robustness of LLaVA 7B and 13B using different prompt types for COCO. Note: Higher values (↑) indicate better performance.

work (Cui et al., 2023a; Zhang et al., 2023).

These findings establish the critical impact of model design and prompt engineering on a model's adversarial robustness, demonstrating that even minimal modifications to the textual prompt can significantly enhance the model's robustness against visual attacks.

6 Broader Impact Statement

As VLMs see increased real-world deployment, ensuring their robustness against adversarial attacks is critical. Our research makes two key contributions: providing optimal model design choices for safe deployment and demonstrating how prompt engineering can enhance adversarial robustness. While enhancing robustness against multimodal attacks using prompt engineering remains unexplored, our work addresses the crucial task of defending against strong image-based attacks that can lead to misinformation or harmful content generation. Our novel lightweight techniques offer a practical alternative to computationally intensive adversarial training, substantially reducing the computational footprint. This research aims to support future advancements in the safe and sustainable deployment of AI systems.

7 Limitations and Future Work

This study has a few limitations that nonetheless provide several interesting future directions. This work focuses primarily on visual perturbations. Building upon our results, future work could extend the focus toward studying multimodal robustness by employing adversarial attacks targeting both visual and textual inputs simultaneously.

Although we demonstrate that prompt engineering can enhance robustness, an adversary with knowledge of these techniques can still generate adversarial samples to compromise the model.

Hence, exploring advanced prompt engineering techniques, such as dynamic prompt adaptation based on real-time analysis of input data and adversarial threat levels, could further enhance the robustness of these models.

8 Acknowledgements

We acknowledge support from the Canada CIFAR AI Chair Program and from the Canada Excellence Research Chairs Program.

P.B. was supported by FRQ-S Research Scholars Junior 1 grant 310924, and the William Dawson Scholar award. R.B. would like to extend gratitude to Khurshed P. Fitter, Akshay Kulkarni, Sabyasachi Sahoo, Jonas Ngnawé, Alexis Roger, Daniel Z Kaplan, Maxime Heuillet and Reza Bayat for their valuable discussions, reviews and feedback.

This research was enabled in part by computational resources provided by the Digital Research Alliance of Canada and Mila - Quebec AI Institute.

References

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. 2020. Square attack: a query-efficient black-box adversarial attack via random search. *Preprint*, arXiv:1912.00049.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv*, abs/2308.01390.

Rabiul Awal, Le Zhang, and Aishwarya Agrawal. 2024. Investigating prompting techniques for zero- and few-shot visual question answering. *Preprint*, arXiv:2306.09996.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 2017. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM.

Francesco Croce and Matthias Hein. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*.

Xuanimng Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. 2023a. On the robustness of large multimodal models against image adversarial attacks. *arXiv preprint arXiv:2312.03777*.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. 2023b. On the robustness of large multimodal models against image adversarial attacks. *Preprint*, arXiv:2312.03777.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. 2023. How robust is google's bard to adversarial image attacks? *Preprint*, arXiv:2309.11751.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yu Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. 2022. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. *Preprint*, arXiv:1802.08218.

Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. 2024. BRAVE: Broadening the visual encoding of vision-language models. *arXiv preprint arXiv:2404.07204*.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *International Conference on Machine Learning (ICML)*.

Hugo Laurenccon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *ArXiv*, abs/2306.16527.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. 2023. Understanding zero-shot adversarial robustness for large-scale models. In *The Eleventh International Conference on Learning Representations*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. 2022. Quality not quantity: On the interaction between dataset design and robustness of clip. *ArXiv*, abs/2208.05516.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. 2024. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *Preprint*, arXiv:2402.12336.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Teknium. 2023. Openhermes-2-mistral-7b. `https://huggingface.co/teknium/OpenHermes-2-Mistral-7B`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Weijie Tu, Weijian Deng, and Tom Gedeon. 2023. A closer look at the robustness of contrastive language-image pre-training (CLIP). In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. 2023. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, pages 36246–36263. PMLR.

Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. 2023. VLATTACK: Multimodal adversarial attacks on vision-language tasks via pre-trained models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. 2023. Adversarial prompt tuning for vision-language models. *Preprint*, arXiv:2311.11261.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai man Cheung, and Min Lin. 2023. On evaluating adversarial robustness of large vision-language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Model Design Choice Results

We provide results for studying the impact of various model design choices for $\epsilon = 4/255$ and $16/255$.

A.1 Impact of Vision Encoder

We can observe that for a lower ϵ value, i.e., $4/255$ CLIP performs better. However, for higher ϵ values, i.e. $8/255$ and $16/255$, SigLIP performs better.

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQAv2
CLIP	None	116.35	76.15	59.48	37.10	41.18	73.89
	FGSM	106.01	64.45	52.18	26.87	36.93	63.55
	PGD	89.95	54.54	44.40	6.73	32.06	53.81
	APGD	87.07	50.51	42.52	19.03	8.80	50.16
SigLIP	None	118.69	78.29	60.92	38.82	42.43	74.49
	FGSM	99.75	60.60	49.84	27.24	39.74	61.75
	PGD	68.94	38.42	33.30	9.54	26.65	44.51
	APGD	59.67	33.12	14.45	11.89	24.12	41.87
Dinov2	None	104.84	54.78	57.00	10.37	38.07	64.80
	FGSM	77.68	37.81	39.78	7.28	32.50	45.40
	PGD	4.86	3.13	9.80	1.99	10.91	25.67
	APGD	2.45	2.17	8.00	1.96	10.69	23.29
ImageNet	None	101.59	54.92	56.34	10.70	39.29	68.36
	FGSM	71.67	34.74	38.44	6.70	31.62	45.37
	PGD	11.17	5.62	11.28	2.43	11.90	15.00
	APGD	5.24	3.69	9.86	2.04	10.80	17.14

Table 13: Comparison between VLMs having different image encoders but the same LLM for $\epsilon = 4/255$. All of them have the same LLM: Vicuna v1.5 7B. Note: Higher values (↑) indicate better performance.

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQAv2
CLIP	None	116.35	76.15	59.48	37.10	41.18	73.89
	FGSM	93.27	56.35	48.52	20.22	36.73	59.99
	PGD	10.32	6.22	11.88	5.87	8.23	29.57
	APGD	3.33	2.57	8.40	3.84	7.89	23.84
SigLIP	None	118.69	78.29	60.92	38.82	42.43	74.49
	FGSM	88.06	50.05	48.62	22.22	40.80	60.55
	PGD	11.77	6.59	12.45	6.57	10.32	31.04
	APGD	4.31	2.79	7.88	3.78	8.97	23.50
Dinov2	None	104.84	54.78	57.00	10.37	38.07	64.80
	FGSM	81.38	39.35	41.18	7.96	36.22	48.61
	PGD	3.00	1.48	7.70	1.54	10.68	24.78
	APGD	1.57	1.12	6.34	1.34	9.70	20.73
ImageNet	None	101.59	54.92	56.34	10.70	39.29	68.36
	FGSM	62.62	29.90	39.42	7.20	33.98	46.78
	PGD	3.12	2.13	8.10	1.64	9.34	22.69
	APGD	2.13	0.95	5.84	1.80	9.98	18.99

Table 14: Comparison between VLMs having different image encoders but the same LLM for $\epsilon = 16/255$. All of them have the same LLM: Vicuna v1.5 7B. Note: Higher values (↑) indicate better performance.

A.2 Resolution of Vision Encoder

We can observe that at a lower ϵ value of $4/255$, lower resolution models are better. However, at a higher ϵ value of $16/255$, the effectiveness of increased resolution for both CLIP and SigLIP models becomes task-dependent.

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQAv2
CLIP-224px	None	116.35	76.15	59.48	37.10	41.18	73.89
	FGSM	106.01	64.45	52.18	26.87	36.93	63.55
	PGD	89.95	54.54	44.40	6.73	32.06	53.81
	APGD	87.07	50.51	42.52	19.03	8.80	50.16
CLIP-336px	None	119.02	77.21	59.18	36.73	39.94	70.00
	FGSM	96.79	64.25	49.70	25.32	33.92	56.52
	PGD	18.54	13.57	16.26	9.44	10.67	28.60
	APGD	30.20	22.11	22.86	11.73	22.86	28.76
SigLIP-224px	None	118.69	78.29	60.92	38.82	42.43	74.49
	FGSM	99.75	60.60	49.84	27.24	39.74	61.75
	PGD	68.94	38.42	33.30	9.54	26.65	44.51
	APGD	59.67	33.12	14.45	11.89	24.12	41.87
SigLIP-384px	None	124.11	87.08	62.18	55.05	41.14	77.22
	FGSM	93.46	57.28	50.88	36.18	35.50	63.22
	PGD	25.51	13.84	20.32	13.38	11.51	38.15
	APGD	12.53	8.40	16.34	9.10	9.52	34.83

Table 15: Comparison between VLMs having different input resolutions of CLIP and SigLIP for $\epsilon = 4/255$. All of them have the same LLM: Vicuna v1.5 7B. Note: Higher values (↑) indicate better performance.

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQAv2
CLIP-224px	None	116.35	76.15	59.48	37.10	41.18	73.89
	FGSM	93.27	56.35	48.52	20.22	36.73	59.99
	PGD	10.32	6.22	11.88	5.87	8.23	29.57
	APGD	3.33	2.57	8.40	3.84	7.89	23.84
CLIP-336px	None	119.02	77.21	59.18	36.73	39.94	70.00
	FGSM	101.98	56.08	49.24	22.20	37.39	58.30
	PGD	10.08	5.25	17.48	6.70	8.52	24.39
	APGD	13.88	9.05	13.26	7.34	27.54	19.98
SigLIP-224px	None	118.69	78.29	60.92	38.82	42.43	74.49
	FGSM	88.06	50.05	48.62	22.22	40.80	60.55
	PGD	11.77	6.59	12.45	6.57	10.32	31.04
	APGD	4.31	2.79	7.88	3.78	8.97	23.50
SigLIP-384px	None	124.11	87.08	62.18	55.05	41.14	77.22
	FGSM	94.90	57.40	52.24	30.93	39.77	63.53
	PGD	9.53	5.17	14.48	8.26	9.19	33.16
	APGD	3.15	1.75	9.14	4.00	7.85	27.82

Table 16: Comparison between VLMs having different input resolutions of CLIP and SigLIP for $\epsilon = 16/255$. All of them have the same LLM: Vicuna v1.5 7B. Note: Higher values (↑) indicate better performance.

A.3 Ensemble of Vision Encoders

The observations for both $\epsilon = 4/255$ and $16/255$ are same as for $\epsilon = 8/255$. Targeting the weakest image encoder is enough to jeopardize the entire system. Conversely, having the strongest vision encoder in the ensemble ensures the best robust performance.

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQAv2
DinoCLIP	None	113.75	74.16	58.88	15.08	39.30	74.35
	FGSM	99.01	57.19	47.14	8.92	38.39	58.50
	PGD	21.00	12.22	14.96	3.34	13.95	35.03
	APGD	10.71	7.12	12.96	3.30	12.49	33.10
DinoSigLIP	None	125.94	85.44	61.12	50.52	44.27	79.39
	FGSM	107.87	74.10	52.92	40.36	40.77	67.24
	PGD	52.89	32.14	23.10	15.73	14.78	42.87
	APGD	35.34	22.86	19.18	13.83	12.96	39.58

Table 17: Comparison between VLMs that have an ensemble of vision encoders. The comparison is made when only the input to the Dino image encoder is perturbed for $\epsilon = 4/255$. Note: Higher values (↑) indicate better performance.

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQAv2
DinoCLIP	None	113.75	74.16	58.88	15.08	39.30	74.35
	FGSM	103.37	57.80	48.38	9.29	40.30	58.87
	PGD	8.14	5.73	11.22	2.86	12.12	32.38
	APGD	3.11	2.37	8.94	2.57	12.63	27.47
DinoSigLIP	None	125.94	85.44	61.12	50.52	44.27	79.39
	FGSM	111.52	74.29	54.76	42.78	42.74	68.40
	PGD	31.29	17.58	18.52	12.72	12.61	39.06
	APGD	17.77	10.57	14.86	10.94	12.38	35.66

Table 18: Comparison between VLMs that have an ensemble of vision encoders. The comparison is made when only the input to the Dino image encoder is perturbed for $\epsilon = 16/255$. Note: Higher values (↑) indicate better performance.

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQAv2
DinoCLIP	None	113.75	74.16	58.88	15.08	39.30	74.35
	FGSM	98.00	57.23	47.50	9.26	38.24	58.72
	PGD	29.76	17.24	17.66	3.38	15.31	36.35
	APGD	11.54	6.95	12.36	3.26	12.77	31.86
DinoSigLIP	None	125.94	85.44	61.12	50.52	44.27	79.39
	FGSM	107.64	72.26	53.82	41.42	40.76	66.92
	PGD	28.75	18.74	18.08	11.28	10.61	40.16
	APGD	36.03	22.73	19.80	13.54	13.88	38.70

Table 19: Comparison between VLMs that have an ensemble of vision encoders. The comparison is made when the same (perturbed) image is passed to both vision encoders for $\epsilon = 4/255$. Note: Higher values (↑) indicate better performance.

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQAv2
DinoCLIP	None	113.75	74.16	58.88	15.08	39.30	74.35
	FGSM	101.93	56.49	47.62	9.54	40.56	59.22
	PGD	10.73	7.27	12.46	2.96	12.98	31.79
	APGD	3.48	2.46	9.04	2.71	11.68	28.03
DinoSigLIP	None	125.94	85.44	61.12	50.52	44.27	79.39
	FGSM	106.15	68.59	53.86	36.45	43.20	67.83
	PGD	13.58	8.06	14.12	6.80	8.52	36.14
	APGD	16.92	10.92	15.72	11.18	11.56	35.94

Table 20: Comparison between VLMs that have an ensemble of vision encoders. The comparison is made when the same (perturbed) image is passed to both vision encoders for $\epsilon = 16/255$. Note: Higher values (↑) indicate better performance.

A.4 Size of LLM

Here we can observe that increasing the model size only helps in gaining robustness against weaker attacks (FGSM). However, the vulnerability and drop in performance against iterative attacks (PGD and APGD) remain almost the same regardless of the model's size.

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQAv2
LLaVA-7B	None	119.02	77.21	59.18	36.73	39.94	70.00
	FGSM	96.79	64.25	49.70	25.32	33.92	56.52
	PGD	18.54	13.57	16.26	9.44	10.67	28.60
	APGD	30.20	22.11	22.86	11.73	22.86	28.76
LLaVA-13B	None	123.71	77.63	62.86	40.04	41.19	75.39
	FGSM	123.71	77.63	62.86	40.04	41.19	75.39
	PGD	18.54	13.57	16.26	9.44	10.67	28.60
	APGD	30.20	22.11	21.30	11.73	22.86	28.76

Table 21: Comparison between models having different scales of the LLM but the same vision encoder for $\epsilon = 4/255$. Note: Higher values (↑) indicate better performance.

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQAv2
LLaVA-7B	None	119.02	77.21	59.18	36.73	39.94	70.00
	FGSM	101.98	56.08	49.24	22.20	37.39	58.30
	PGD	10.08	5.25	17.48	6.70	8.52	24.39
	APGD	13.88	9.05	13.26	7.34	27.54	19.98
LLaVA-13B	None	123.71	77.63	62.86	40.04	41.19	75.39
	FGSM	99.83	58.40	52.90	24.85	37.89	60.98
	PGD	10.08	9.05	13.26	6.70	8.52	24.39
	APGD	13.88	5.25	17.48	7.34	27.54	19.98

Table 22: Comparison between models having different scales of the LLM but the same vision encoder for $\epsilon = 16/255$. Note: Higher values (↑) indicate better performance.

A.5 Choice of LLM

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQAv2
llama2	None	6.95	3.36	60.66	44.82	42.41	76.73
	FGSM	-	-	60.72	44.68	42.23	76.47
	PGD	-	-	59.98	44.57	42.23	77.32
	APGD	-	-	60.72	44.56	42.35	76.44
llama2-chat	None	68.31	37.34	61.24	44.85	42.06	76.05
	FGSM	-	-	51.50	33.04	35.45	65.53
	PGD	-	-	18.64	9.49	13.27	34.47
	APGD	-	-	15.24	7.30	12.60	32.10
mistral-instruct-v0.1	None	94.55	63.92	60.76	43.84	40.57	76.64
	FGSM	-	-	51.10	30.42	33.71	63.78
	PGD	-	-	18.54	11.54	7.87	39.61
	APGD	-	-	16.34	8.75	7.33	36.87
mistral-v0.1	None	109.23	74.60	60.60	45.14	43.70	76.47
	FGSM	-	-	51.36	31.41	39.67	64.39
	PGD	-	-	16.82	11.89	9.85	40.03
	APGD	-	-	15.38	10.24	9.31	38.69
Vicuna	None	119.02	77.21	59.18	36.73	39.94	70.00
	FGSM	96.79	64.25	49.70	25.32	33.92	56.52
	PGD	18.54	13.57	16.26	9.44	10.67	28.60
	APGD	30.20	22.11	22.86	11.73	22.86	28.76

Table 23: Comparison between VLMs that have different LLMs of 7B size for $\epsilon = 4/255$. Every model has the same vision encoder, CLIP-336. Note: Higher values (↑) indicate better performance.

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQAv2
llama2	None	6.95	3.36	60.66	44.82	42.41	76.73
	FGSM	-	-	50.62	27.34	38.62	66.65
	PGD	-	-	12.20	6.46	11.57	33.07
	APGD	-	-	9.42	4.53	12.33	28.03
llama2-chat	None	68.31	37.34	61.24	44.85	42.06	76.05
	FGSM	69.61	38.32	51.90	28.35	38.20	66.11
	PGD	-	2.54	12.48	6.51	10.86	31.53
	APGD	-	1.50	10.04	4.04	10.62	27.12
mistral-instruct-v0.1	None	94.55	63.92	60.76	43.84	40.57	76.64
	FGSM	82.02	49.74	51.58	27.46	35.94	63.72
	PGD	2.89		13.76	8.28	6.70	36.41
	APGD		1.24	8.72	5.35	5.96	31.43
mistral-v0.1	None	109.23	74.60	60.60	45.14	43.70	76.47
	FGSM	-	61.74	51.76	26.82	40.84	64.16
	PGD	-	3.56	12.84	8.46	8.15	37.41
	APGD	2.41	2.01	9.10	6.24	7.71	32.39
Vicuna	None	119.02	77.21	59.18	36.73	39.94	70.00
	FGSM	101.98	56.08	49.24	22.20	37.39	58.30
	PGD	10.08	5.25	17.48	6.70	8.52	24.39
	APGD	13.88	9.05	13.26	7.34	27.54	19.98

Table 24: Comparison between VLMs that have different LLMs of 7B size for $\epsilon = 16/255$. Every model has the same vision encoder, CLIP-336. Note: Higher values (↑) indicate better performance.

B Prompt Formatting

Task	Instruction
Rephrase	`You will be given a question. Your task is to rephrase the question so that it is semantically similar to the original question and will have the same answer as the original question.`
Expand	`You will be given a short question. Your task is to generate a longer question so that it is semantically similar to the original question and will have the same answer as the original question.`
AC	`You will be given a question. However, the image associated with the question will be adversarially perturbed. Your task is to generate a longer question so that it is semantically similar to the original question and will have the same answer as the original question.`
AP	`You will be given a question. However, the image associated with the question could be adversarially perturbed. Your task is to generate a longer question so that it is semantically similar to the original question and will have the same answer as the original question.`

Table 25: Instructions used to obtain the modified questions for VQA.

Prompt Type	Prompt
Original	`Provide a short caption for this image.`
AC	`Consider the given image being adversarially perturbed. Provide a short caption for this image.`
AP	`Given image could be adversarially perturbed. Provide a short caption for this image.`
Random sent.	`Clouds drift quietly over the ancient, forgotten city. Provide a short caption for this image.`
Random str.	`ryFo8ZVcyNMtLgryNOg64UTjySyEb79e5aq6IJxGuz0GzWNtoz. Provide a short caption for this image.`

Table 26: Various types of prompts tested for image captioning.

C Prompt Formatting Results

	Prompt	FGSM	PGD	APGD	Clean
LLaVA 7B	Original	95.55	22.84	12.54	119.02
	AC	63.86	60.01	54.05	64.11
	AP	105.41	101.61	91.46	112.78
	Random str	108.23	105.35	94.61	120.90
	Random sent	101.12	97.11	88.45	108.00
	Clean Acc				119.02
LLaVA 13B	Original	106.40	14.60	6.98	123.71
	AC	113.77	106.40	114.65	122.10
	AP	114.48	108.83	113.54	125.28
	Random str	110.74	105.15	111.69	120.49
	Random sent	113.29	106.71	111.13	120.72
	Clean acc				123.71

Table 27: Performance of LLaVA models on image captioning (COCO) when adversarially perturbed images (using $\epsilon = 8/255$) are provided along with different types of prompts. Note: Higher values (↑) indicate better performance.

	Prompt	FGSM	PGD	APGD	Clean
LLaVA 7B	Original	57.19	27.07	26.73	70.00
	Rephrase	59.01	58.03	48.84	68.30
	AC	60.21	58.82	50.68	69.99
	AP	60.13	58.81	49.95	69.78
	Expand	48.59	48.54	42.24	57.14
	Clean Acc				70.00
LLaVA 13B	Original	62.49	34.49	33.56	75.39
	Rephrase	59.01	60.05	54.77	71.02
	AC	51.38	61.49	55.56	72.00
	AP	63.59	61.29	63.2	71.79
	Expand	53.03	50.03	45.93	58.59
	Clean Acc				75.39

Table 28: Performance of LLaVA models on VQAv2 when adversarially perturbed images (using $\epsilon = 8/255$) are provided along with the questions generated using different types of prompts. Note: Higher values (↑) indicate better performance.

D Analysis on Black-box Attacks

We attempt to implement the ZOO black-box attack (Chen et al., 2017) on the LLaVA 7B model for all the six tasks, owing to its adaptability to settings beyond classification, unlike the square-attack (Andriushchenko et al., 2020) which is designed specifically for classification. We do not observe any effectiveness with 1k, 2k iterations for 500 samples, or 5k iterations for 10 samples.

- For VQAv2, the robust accuracy on 500 samples after 1k iterations is 75.48. With 5k iterations, we obtain a robust accuracy of 76 for 100 samples.
- For TextVQA, the robust accuracy is 40.48.

All these values are very close to the clean accuracies we provided: 73.89 for VQAv2 and 37.10 for TextVQA on 1000 samples. Note that the number of samples considered during black-box attacks is lower than the figures for which we provided the clean accuracies.