Towards Adversarially Robust Vision-Language Models: Insights from Design Choices and Prompt Formatting Techniques

Rishika Bhagwatkar [1 2 3] **Shravan Nayak** [1 2] **Reza Bayat** [1 2 3] **Alexis Roger** [1 2 3] **Daniel Z Kaplan** [3 4]
Pouya Bashivan [1 5] **Irina Rish** [1 2 3]

Abstract

Vision-Language Models (VLMs) have witnessed a surge in both research and real-world applications. However, as they becoming increasingly prevalent, ensuring their robustness against adversarial attacks is paramount. This work systematically investigates the impact of model design choices on the adversarial robustness of VLMs against image-based attacks. Additionally, we introduce novel, cost-effective approaches to enhance robustness through prompt formatting. By rephrasing questions and suggesting potential adversarial perturbations, 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.

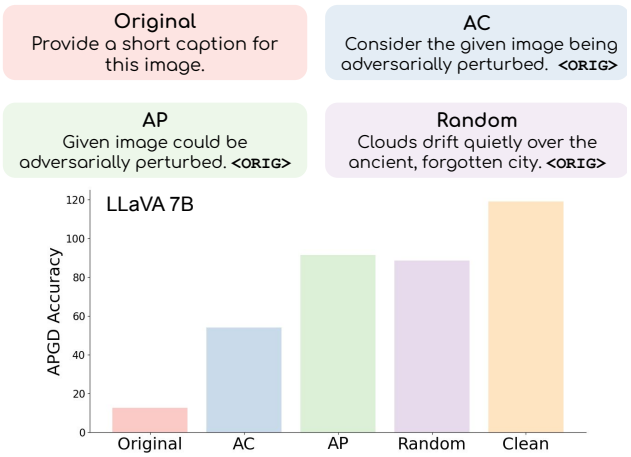

Figure 1: Performance of LLaVA-7B on the COCO dataset when the adversarial images are given along with different types of prompts (Original, AC, AP and Random). Clean accuracy represents the model's performance on unperturbed images.

1. Introduction

VLMs process images in conjunction with text prompts, enabling them to perform a wide array of tasks, such as image captioning, visual question answering (VQA), and cross-modal retrieval (Liu et al., 2023; Laurenccon et al., 2023; Awadalla et al., 2023; Radford et al., 2021). While there is extensive research on advancing architecture and scaling, recent works demonstrate that VLMs are 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 models' robustness and integrity.

White-box attacks are a common type of adversarial threat. These attacks assume complete access to a model's parameters, enabling attackers to exploit specific vulnerabilities. Since many VLMs are open-source, attackers can easily analyze and exploit them. Current VLMs have several design choices, including the vision encoder, large language model (LLM), mapping network, image resolution, and the training data (Liu et al., 2023; Laurenccon et al., 2023; Awadalla et al., 2023). Despite the importance of these factors, their impact on adversarial robustness is under-explored. In this study, we evaluate how these design choices during VLM training influence their susceptibility to white-box adversarial attacks on the input images.

In addition to design choices, the selection and quality of prompts can significantly impact the performance and robustness of VLMs (Awal et al., 2024). Effective prompts can enhance the models' understanding and response to inputs, affecting their robustness to adversarial attacks. Recent works have focused on adversarial training to increase robustness (Schlarmann et al., 2024; Mao et al., 2023), but it is resource-intensive and costly, often requiring millions of samples (Wang et al., 2023; Bartoldson et al., 2024). As a practical and cost-effective alternative, we investigate

[1]Mila – Quebec AI Institute [2]Université de Montréal [3]CERC-AAI [4]Realiz.ai [5]McGill University. Correspondence to: Rishika Bhagwatkar <rishika.bhagwatkar@umontreal.ca>.

whether prompt formatting 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.

Through evaluating both design choices and prompt formatting, we aim to provide comprehensive insights into enhancing the adversarial robustness of VLMs. Our main contributions are summarized as follows:

1. We provide an in-depth analysis of how various design choices of VLMs impact their robustness to white-box visual adversarial attacks.
2. We investigate a novel approach to prompt formatting for enhancing the adversarial robustness of VLMs.
3. To the best of our knowledge, we are the first to offer actionable insights and practical recommendations for using text prompting techniques to enhance the robustness of VLMs in deployment.

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 how these choices affect 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., 2023a; Zhao et al., 2023). Studies also explore the potential of using pretrained VLMs to craft adversarial image and text 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).

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 (Deng et al., 2009) dataset 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 for improving adversarial robustness (Cui et al., 2023b). 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 prompt formatting on model performance without additional training or image-based information extraction.

3. Experiments

In this section, we outline the attack setups used during evaluations, the tasks assessed, 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 architecture and parameters. The objective in crafting adversarial samples under 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 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 complexity: Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014), Projected Gradient Descent (PGD) (Madry et al., 2017), and Auto-PGD (APGD) (Croce & Hein, 2020). We employ PGD and APGD attacks with 100 iterations while FGSM uses 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.

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. Higher is better for both metrics. 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.

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

	Vision Encoder	Language Model
Image Representations	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
Size of LLM	CLIP ViT-L/14 @ 336px	Vicuna v1.5 7B Vicuna v1.5 13B
Ensemble of visual encoders	DINOv2 ViT-L/14 + CLIP ViT-L/14 @ 336px DINOv2 ViT-L/14 + SigLIP ViT-L/14 @ 384px	Vicuna v1.5 7B

4. Results

4.1. Model Design Choices

In our analysis, we examine the impact of various model design choices on adversarial robustness. Specifically, we focus on: (a) the choice of vision encoder; (b) the input resolution used by the vision encoder; (c) the sizes of the language models; and (d) the ensemble use of multiple vision encoders. Each of these aspects is detailed in the sections below. We report results using $\epsilon = 8/255$. Please check Appendix A for results with other values $\epsilon = 4/255, 16/255$. We highlight the best robust FGSM, PGD, APGD accuracy across benchmarks for all attacks and models.

4.1.1. Impact of Vision Encoder

We systematically evaluate the effects of employing different vision encoders, each trained under distinct conditions. 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 (Dosovitskiy et al., 2020). As shown in Table 2, SigLIP slightly outperforms CLIP, with both noticeably surpassing DINOv2 and ImageNet 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).

Table 2: Comparison between VLMs having different image encoders but the same language model - 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	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

4.1.2. 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 resolution enhances robustness against stronger attacks for high-resolution CLIP models (on most tasks), the effectiveness of the increased resolution in SigLIP models appears to be task-dependent. However, we observe that robust accuracy significantly deteriorates under APGD attacks in all cases except for VQAv2. For results on other ϵ values, please refer to Appendix A.

4.1.3. Size of Language Model

We evaluate a series of VLMs utilizing the same vision encoder, and same LLM architecture, with the only difference being the language model's size. Specifically, we examine models equipped with the Vicuna language model (Zheng et al., 2024) in two sizes: 7B and 13B. According to the results in Table 4, the model's vulnerability to adversarial attacks and the significant drop in robust accuracy remain consistent, regardless of the model's scale. Hence, increasing the size of the language model 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

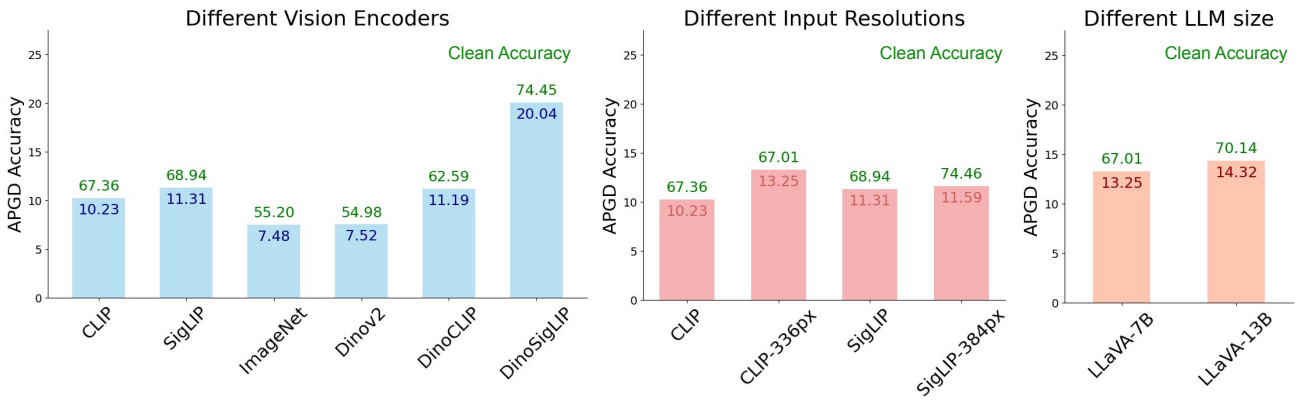

Figure 2: Comparison between VLMs having different vision encoders (left), different input resolutions (center) and different LLM size (right). The comparison is based on the APGD accuracy averaged over all tasks as shown in Tables 2, 3, 4 and 5.

Table 3: Comparison between VLMs having different input resolutions of CLIP and SigLIP. All of them have the same language model: 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	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

even at 13B scale may struggle to effectively interpret these flawed representations, making robustness to image-based attacks less sensitive to language model size. Therefore, enhancing the vision encoder's adversarial robustness is sufficient as shown in prior work (Schlarmann et al., 2024). Please check Appendix A for results on other ϵ values.

4.1.4. 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 performance (Karamcheti et al., 2024; Kar et al., 2024), our research aims to assess whether compromising just one encoder can affect the entire model. This approach allows us

Table 4: Comparison between models having different scales of language models. Both of the models have the same vision encoder, CLIP-336, but different scales of the LLM. 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	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

to analyze if knowledge of the weakest link is sufficient to compromise the entire model when an ensemble of encoders is used. We specifically examined models combining DINOv2 with either CLIP or SigLIP. In our experiments, we perturbed the images processed by DINOv2 while keeping inputs to the other encoder intact. Results in Table 5 show that attacking only DINOv2 is sufficient to compromise the model under stronger attacks, despite the other vision encoder providing clean inputs. This highlights a significant vulnerability in ensemble approaches: even with enhanced performance capabilities, the robustness of the entire system can be jeopardized by targeting a single encoder. Please check Appendix A for results on other ϵ values.

4.2. Prompt Formatting

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 test this hypothesis with the LLaVA 7B and LLaVA 13B models, em-

Table 5: 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. Note: Higher values (↑) indicate better performance.

	Attack	Task					
		COCO	Flickr30k	OK-VQA	TextVQA	VizWiz	VQav2
Dino+CLIP	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
Dino+SigLIP	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

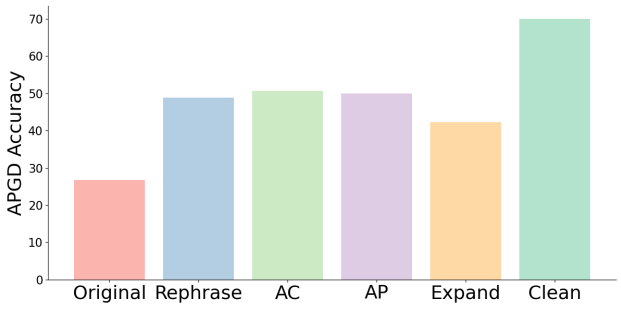

Figure 3: Performance of LLaVA-7B on VQA using questions generated by different types of prompts.

ploying 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. Specific details on the prompts used and the corresponding results are detailed in the subsequent subsections.

4.2.1. Captioning

Our experiments evaluated various prompt formatting strategies, including: (1) **Original** - using the original prompt; (2) **Adversarial Certainty (AC) Prompt** - explicitly informing the model that the image is adversarially perturbed; (3) **Adversarial Possibility (AP) Prompt** - suggesting the possibility that the image might be adversarially perturbed; and (4) **Random** - appending a random sentence or string at the beginning of the prompt. These are listed in Table 15 in Appendix B. From the results presented in Fig. 1 and Table 16 in Appendix C, it is evident that indicating the possibility of adversarial perturbations (AP prompt) assists the model significantly more than explicitly stating that the image is perturbed (AC prompt). Further, the improvements from simply adding a random string or sentence are substantial, even comparable to the effects observed with the AP prompt. This indicates that the models pay more attention to the inputs when they struggle to establish a clear relationship between them.

4.2.2. 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 finetuned model of the 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 14 in Appendix B. According to the results presented in Fig. 3 and Table 17 in Appendix C, 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 robustness performance, reinforcing our observations with the COCO dataset discussed earlier.

5. Discussion and Conclusion

Our evaluation highlighted critical insights into how various design elements affect the adversarial robustness of VLMs. First, we observed that vision encoders trained across diverse data distributions only improve resistance against simpler, less sophisticated attacks, demonstrating limited effectiveness against more complex threats. Additionally, increasing the resolution of image encoders did not correlate with enhanced adversarial robustness, suggesting that benefits seen in the clean accuracy do not extend to improved robustness. Similarly, scaling up the size of the language model did not increase the model's robustness to attacks, indicating that larger models are not inherently more robust. Most notably, our results revealed that using multiple vision encoders does not guarantee robustness; rather, knowledge about the most vulnerable encoder is enough to compromise the entire system.

Building on our findings, we further explored the influence of prompt formatting on enhancing adversarial robustness. Our experiments revealed that even naively rephrasing the questions significantly improves robustness in VQA. Similarly, merely suggesting the possibility of an adversarial image during captioning led to a notable performance boost. More importantly, we found that we do not need to add additional context from the image or fine-tune additional tokens to make models adversarially robust, as opposed to prior work (Cui et al., 2023b; Zhang et al., 2023).

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

6. 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 formatting can enhance adversarial robustness. Our lightweight technique offers a practical alternative to computationally intensive adversarial training, reducing the computational footprint. While enhancing robustness against multimodal attacks using prompt formatting remains unexplored, our work addresses the crucial task of defending against strong image-based attacks that can lead to misinformation or harmful content generation. This research aims to support future advancements in the safe deployment of AI systems.

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

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S. Y., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P. W., Ilharco, G., Wortsman, M., and Schmidt, L. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv*, abs/2308.01390, 2023. URL https://api.semanticscholar.org/CorpusID:261043320.

Awal, R., Zhang, L., and Agrawal, A. Investigating prompting techniques for zero- and few-shot visual question answering, 2024.

Bartoldson, B., Diffenderfer, J., Parasyris, K., and Kailkhura, B. Adversarial robustness limits via scaling-law and human-alignment studies. *ArXiv*, abs/2404.09349, 2024.

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

Cui, X., Aparcedo, A., Jang, Y. K., and Lim, S.-N. On the robustness of large multimodal models against image adversarial attacks, 2023a.

Cui, X., Aparcedo, A., Jang, Y. K., and Lim, S.-N. On the robustness of large multimodal models against image adversarial attacks. *arXiv preprint arXiv:2312.03777*, 2023b.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dong, Y., Chen, H., Chen, J., Fang, Z., Yang, X., Zhang, Y., Tian, Y., Su, H., and Zhu, J. How robust is google's bard to adversarial image attacks?, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., and Schmidt, L. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, 2022.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014. URL https://api.semanticscholar.org/CorpusID:6706414.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. Vizwiz grand challenge: Answering visual questions from blind people, 2018.

Kar, O. F., Tonioni, A., Poklukar, P., Kulshrestha, A., Zamir, A., and Tombari, F. BRAVE: Broadening the visual encoding of vision-language models. *arXiv preprint arXiv:2404.07204*, 2024.

Karamcheti, S., Nair, S., Balakrishna, A., Liang, P., Kollar, T., and Sadigh, D. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *International Conference on Machine Learning (ICML)*, 2024.

Laurenccon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A. M., Kiela, D., Cord, M., and Sanh, V. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *ArXiv*, abs/2306.16527, 2023. URL https://api.semanticscholar.org/CorpusID:259287020.

Li, J., Li, D., Savarese, S., and Hoi, S. C. H. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:256390509.

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2017. URL https://api.semanticscholar.org/CorpusID:3488815.

Mao, C., Geng, S., Yang, J., Wang, X., and Vondrick, C. Understanding zero-shot adversarial robustness for large-scale models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=P4bXCawRi5J.

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

Nguyen, T., Ilharco, G., Wortsman, M., Oh, S., and Schmidt, L. Quality not quantity: On the interaction between dataset design and robustness of clip. *ArXiv*, abs/2208.05516, 2022.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2641–2649, December 2015. doi: 10.1109/ICCV.2015.303.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Schlarmann, C., Singh, N. D., Croce, F., and Hein, M. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models, 2024.

Singh, A., Natarjan, V., Shah, M., Jiang, Y., Chen, X., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

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

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

Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., and Yan, S. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, pp. 36246–36263. PMLR, 2023.

Yin, Z., Ye, M., Zhang, T., Du, T., Zhu, J., Liu, H., Chen, J., Wang, T., and Ma, F. VLATTACK: Multimodal adversarial attacks on vision-language tasks via pre-trained models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=qBAED3u1XZ.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Zhang, J., Ma, X., Wang, X., Qiu, L., Wang, J., Jiang, Y.-G., and Sang, J. Adversarial prompt tuning for vision-language models, 2023.

Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., man Cheung, N., and Lin, M. On evaluating adversarial robustness of large vision-language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

A. Model Design Choice Results

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

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.

Table 6: Comparison between VLMs having different image encoders but the same language model for $\epsilon = 4/255$. All of them have the same language model: 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	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 7: Comparison between VLMs having different image encoders but the same language model for $\epsilon = 16/255$. All of them have the same language model: 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

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.

Table 8: Comparison between VLMs having different input resolutions of CLIP and SigLIP for $\epsilon = 4/255$. All of them have the same language model: 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	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 9: Comparison between VLMs having different input resolutions of CLIP and SigLIP for $\epsilon = 16/255$. All of them have the same language model: 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

A.3. Size of Language Model

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.

Table 10: Comparison between models having different scales of 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	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 11: Comparison between models having different scales of LLM but the same vision encoder for $\epsilon = 16/255$. 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

A.4. 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.

Table 12: 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	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 13: 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	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

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 14: 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 15: Various types of prompts tested for image captioning.

C. Prompt Formatting Results

Table 16: 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	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 17: 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.

	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