Embracing the Uncertainty: Evaluating Adversarial Robustness of Vision-Language Models from an Untargeted Attack Perspective

Anonymous ICME submission

***Abstract*—Vision-language models (VLMs), such as BLIP and LLaVA, achieve multimodal understanding by enabling large language models (LLMs) to process and interpret images. However, these models are also vulnerable to attacks from visual adversarial examples. In particular, untargeted attacks achieve a higher success rate, raising significant concerns about their robustness. Previous methods involve calculating the language modeling loss based on the model's output and then back-propagating this loss to the image to generate adversarial examples. Given that LLMs typically have deep neural network architectures, calculating the loss solely from the model's output often diminishes the adversarial noise propagated back to the image. To this end, we propose a novel untargeted attack method based on Maximizing Information Entropy (MIE), which not only leverages the output of VLMs but also utilizes internal features to guide the generation of adversarial examples. Specifically, we maximize the uncertainty of the model's output distribution while concurrently driving the model's internal features towards a meaningless Gaussian distribution. This strategy enables a more effective attack without requiring knowledge of the image annotation information. Quantitative results across multiple models and datasets demonstrate that our proposed MIE outperforms previous methods, which establish a new benchmark for assessing the adversarial robustness of VLMs.**

***Index Terms*—vision-language models, adversarial examples, untargeted attacks**

I. Introduction

Large vision-language models (VLMs) enhance the capabilities of large language models (LLMs) by incorporating visual modules to process complex multimodal data [1]. These models have achieved notable success in various multimodal understanding tasks, such as image captioning [2], visual question answering [3], and document understanding and analysis [4].

Inheriting the vulnerabilities of visual neural networks to image perturbations [5]–[7], VLMs are similarly susceptible to adversarial examples, where the addition of imperceptible noise to pristine images can mislead the models into forming completely erroneous interpretations of the images [8]–[15]. As depicted in Figure 1, even such minimal noise can lead victim models to incorrectly identify a butterfly as a plate of salad, posing a considerable challenge to the security of VLMs, especially in life-critical scenarios.

Previous methods typically calculate the loss solely from the output of the language model and then backpropagate this loss to the image to generate adversarial examples. For example, [14] proposed maximizing or minimizing the negative log-likelihood of a target text to achieve untargeted or targeted attacks. Verbose Images [13] generate adversarial images by delaying the appearance of the end-of-sequence token, prompting VLMs to produce sequences that are as long as possible. Considering that visual information passes through numerous neural network layers from input to the final output, this potentially weakens the noise when the loss is backpropagated to the image, consequently affecting the attack performance. Therefore, it is promising to leverage the internal features of the victim model to guide the generation of adversarial samples.

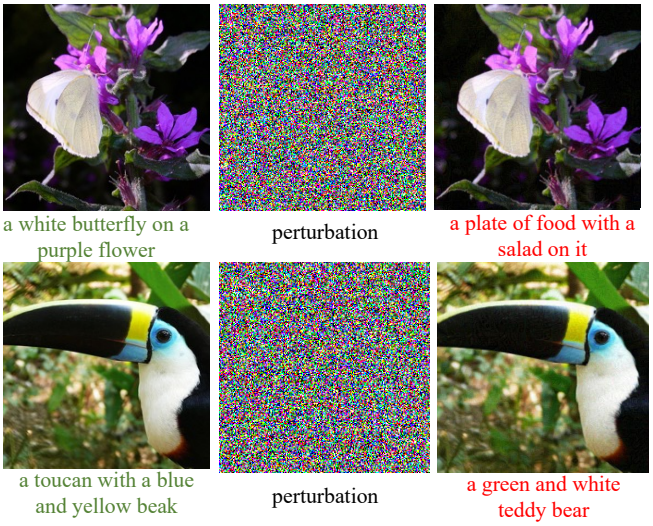


Fig. 1: Illustration of untargeted adversarial attacks on VLMs. The clean images, when perturbed with the subtle yet malicious noise, transform into the adversarial images. These adversarial images can cause the model to generate unpredictable, anomalous, or erroneous outputs. The green text represents the originally correct descriptions, while the red text indicates the erroneous descriptions.

To this end, we propose a novel untargeted attack method based on maximizing information entropy (MIE), which aims to maximize the uncertainty of both the VLMs' output distribution and their internal features. Specificall, by maximizing the information entropy of the output distribution, we ensure that the generated content exhibits a high degree of uncertainty, thereby leading the model to a completely erroneous understanding of the image. To broaden the scope of perturbations, we also introduce disruptions to the model's internal features, specifically focusing on the two key components of

the Transformer architecture: Attention and MLP [16]. We perform layer-by-layer perturbations on these components to fill the internal features with uncertainty. MIE comprehensively perturbs VLMs, enabling a deep assessment of their adversarial robustness. Because it does not rely on predefined text or require prior knowledge of images, it can be used to generate adversarial examples on a large scale. We collect image data from ImageNet [17] and MS-COCO [18] datasets and evaluate the adversarial robustness of multiple open-source VLMs, including BLIP [2], BLIP-2 [19], InstructBLIP [20], MiniGPT-4 [3], LLaVA1.5 [1], LLaVA1.6 [4]. Extensive experimental results show that our method achieved the best attack performance, highlighting the vulnerabilities of current VLMs and emphasizing the need for a thorough evaluation before deploying these models.

Considering its ease of implementation and applicability to most current VLMs, our method can serve as a new benchmark for evaluating the adversarial robustness of VLMs. In summary, the main contributions of this paper are as follows:

1) We realize that previous methods, which generate adversarial examples solely by optimizing the model's output, are inadequate for effectively assessing the adversarial robustness of VLMs.
2) We propose a novel untargeted attack method based on maximizing information entropy (MIE), which not only maximizes the uncertainty of the model's output distribution but also maximizes the uncertainty of the model's internal features.
3) We conduct extensive experiments to validate the effectiveness of our approach. The experimental results quantitatively demonstrate that our method can effectively attack VLMs.

II. Related Work

A. Vision-Language Models

Large vision-language models (VLMs) can handle various multimodal tasks simultaneously and have received widespread attention. They can be roughly divided into two types from the perspective of modal interaction:

Image as Key-Value. The first group uses features extracted by an image encoder as the Key and Value components and treats the input text as the Query during the language model's decoding process [2], [21]. The subsequent output at each time step is computed using a cross-attention mechanism.

Image as Token. Another commonly used approach involves converting images into token sequences that align with the word embedding space, facilitating interaction between images and text [1], [3], [20]. A key advantage of this method is its ability to fully utilize large language models without requiring modifications.

B. Adversarial Examples

Adversarial Attacks. The vulnerability of neural networks to adversarial examples has been extensively studied [5], [7]. These efforts aim to introduce noise into the image that is invisible to the human eye, thereby deceiving the model into generating incorrect outputs. Some studies suggest that modifying the internal features of a classifier can enhance the effectiveness of an attack [22].

Adversarial attacks on VLMs. VLMs inherit the vulnerabilities of visual neural networks, which lead to susceptibility to adversarial attacks [8]–[15]. Specifically, some works [9], [11], [12], [23] treat toxic text as the target suffix and employ teacher-forcing optimization techniques to generate adversarial examples that bypass the alignment constraints of the model. However, unlike visual classification networks, VLMs exhibit greater adversarial robustness.

III. Methodology

In this section, we initially present the essential framework of adversarial attacks, followed by an in-depth discussion of our MIE method, which seeks to enhance both output uncertainty and internal feature uncertainty, grounded in the architecture of VLMs, as illustrated in Figure 2.

A. Threat Models

Current large VLMs typically consist of a visual module and a large language model. They align visual features with the input space of the language model and utilize autoregressive generation to enhance the comprehension of images and text. We denote $f_\theta(\boldsymbol{x}, \boldsymbol{q}) \mapsto \boldsymbol{a}$ as a VLM parameterized by θ, where $\boldsymbol{x}$ is the input image, $\boldsymbol{q}$ is the input text, and $\boldsymbol{a}$ is the corresponding output text in an autoregressive manner. For image captioning tasks, $\boldsymbol{q}$ can be a start symbol $<bos>$ and $\boldsymbol{a}$ represents the caption. In the case of visual question answering (VQA) tasks, $\boldsymbol{q}$ can be a question and $\boldsymbol{a}$ represents an answer. Notably, by applying specific prompts, VLMs are capable of performing various multimedia understanding tasks. For threat models employed for VLMs:

Adversary knowledge refers to their comprehension of the internal mechanisms of the victim model. In the setting of white-box attacks, the attacker can obtain the architecture and parameters of the victim model.

Adversary goals describe the objectives that malicious attackers aim to achieve, including targeted and untargeted attacks. This paper conducts an untargeted attack to assess the adversarial robustness of VLMs.

Adversary capabilities elucidate the resources required or constraints faced by adversaries in executing attacks. The most commonly used constraint is the L_p budget for the perturbation magnitude, where the L_p norm between the clean image $\boldsymbol{x}$ and the adversarial image $\boldsymbol{x}^{adv}$ is required to be less than a specified threshold ϵ as $\left\|\boldsymbol{x} - \boldsymbol{x}^{adv}\right\|_p \leq \epsilon$.

B. Maximizing the Uncertainty of Output

Given that generating adversarial examples for evaluating robustness frequently involves situations where prior knowledge about images is lacking, it is crucial to eliminate reliance on specific target text. If the aim is to make the model's output deviate from all potentially accurate interpretations, a viable approach is to minimize the informational content of

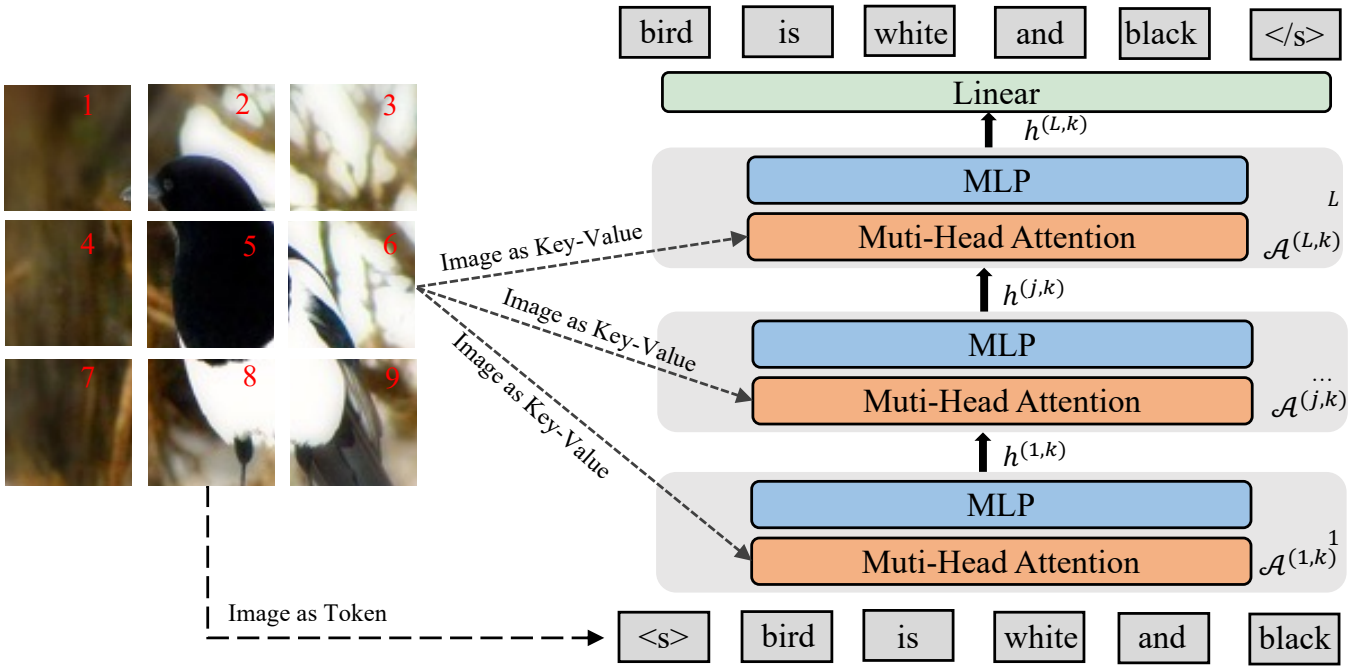


Fig. 2: The unified architecture of VLMs. We propose the Maximizing Information Entropy method to maximize the uncertainty of the output distribution and internal features.

the model's output, ideally resulting in meaningless text. Consequently, we endeavor to maximize the information entropy in the model's responses to images. It should be noted that we maintain gradient computation for large models during the inference stage.

For each position i of the model output $\boldsymbol{a}$, a normalized vector $p_i \in \mathbb{R}^v$, where v is the vocabulary size, is generated. The model subsequently selects the token with the highest probability as the output for that step:

$$\boldsymbol{a} = [\boldsymbol{a}_i] \triangleq [p_i], i = [1, 2, \ldots, N] \tag{1}$$

where $\boldsymbol{a}_i$ denotes i-th token of $\boldsymbol{a}$ and N is the length of the output sequence.

For a well-trained model, it tends to output specific information at each step with high confidence. However, when the model encounters challenging examples, its output can become ambiguous. In the most extreme case, the model assigns equal probabilities to each token, resulting in ungrammatical and random output. This aligns with the definition of information entropy, and to a certain extent, augmenting entropy will also elevate the uncertainty of the model's output [24]. Motivated by this, we maximize the uncertainty of each output vector of the model:

$$\mathcal{L}_1 = -\mathbb{E}[\sum_i \sum_j \log(p_i^{(j)}) p_i^{(j)}] \tag{2}$$

where $p_i^{(j)}$ represents the probability of the j-th position of the output vector corresponding to the i-th token.

Note that the computation of information entropy includes the termination token $<eos>$, which could potentially cause the model to fail in terminating its output correctly.

C. Maximizing the Uncertainty of Internal Features

To further undermine the model's visual understanding ability, it is promising to disrupt the distribution of the model's internal features. By guiding the internal features towards a Gaussian distribution, we can effectively disrupt the model's cognitive system. VLMs typically include a language model, which is structurally a stack of multiple Transformer blocks [16]. Each Transformer block is primarily composed of Attention and MLP components, so we attempt to disrupt these internal features layer by layer.

The Attention mechanism comprises three components: Query, Key, and Value. The attention weights are derived by computing the dot product between the Query and each Key. Typically, in an image, only a small fraction of pixels or patches are pertinent to the prompt. Formally, for an L-layer Transformer block in a language model, each token $\boldsymbol{a}_i$ will generate attention weights $\mathcal{A}_i \in \mathbb{R}^{L \times (P+T)}$, where P is the number of visual patches, $T = 0$ if the interaction mode is *Image as Key-Value*, and $T = i - 1$ if the mode is *Image as Token*. To prevent the model from attending to valuable information, we maximize the uncertainty of the model's attention distribution:

$$\mathcal{L}_2 = -\mathbb{E}[\sum_i \sum_j \sum_{k=1}^{i-1} \log(\mathcal{A}_i^{(j,k)}) \mathcal{A}_i^{(j,k)}] \tag{3}$$

where j and k represent the layer number and sequence position, respectively.

The MLP performs local and non-linear mapping of inputs at each position, thereby enhancing the model's representational ability. For each feature vector $h_i \in \mathbb{R}^{L \times d}$ processed

TABLE I: Untargeted attacks against victim models. The CLIP scores (↓) between the images and texts are reported, where higher values indicate stronger alignment between images and texts, whereas lower values imply weaker alignment.

Method	ImageNet [17]						MS-COCO [18]					
	BLIP	BLIP-2	InstructBLIP	MiniGPT-4	LLaVA1.5	LLaVA1.6	BLIP	BLIP-2	InstructBLIP	MiniGPT-4	LLaVA1.5	LLaVA1.6
Clean	29.79	30.72	31.36	30.89	31.68	31.95	29.72	29.85	31.02	30.95	31.07	32.86
Gaussian	29.65	30.74	31.33	30.85	31.86	31.94	29.73	29.52	31.01	30.84	31.09	32.75
TAR [14]	20.53	24.58	24.31	24.71	24.46	24.64	21.71	24.53	24.73	23.31	24.56	24.75
UAC [14]	19.87	24.06	23.19	24.15	25.23	25.82	19.96	24.53	24.15	24.74	25.49	24.71
GAN-based [15]	24.36	24.78	25.32	25.31	24.15	25.56	24.85	25.07	25.74	24.85	24.93	25.92
Encoder-based [10]	21.70	22.71	23.06	22.85	23.07	23.76	21.53	23.72	23.64	23.76	23.15	23.21
CroPA [12]	19.85	24.23	23.08	24.11	25.30	24.31	20.83	23.71	24.20	23.97	23.95	22.73
Verbose Image [13]	20.21	23.21	23.55	22.21	22.37	22.64	20.05	22.34	22.72	22.50	23.21	23.31
MIE	**17.53**	**20.12**	**20.34**	**21.58**	**21.32**	**21.39**	**17.61**	**19.63**	**20.52**	**20.46**	**21.52**	**20.91**

by the MLP, where d is the embedding dimension of the language model, we introduce perturbations to drive it towards a meaningless Gaussian distribution, thereby preventing it from properly representing the output text. Because the internal feature vectors of the LLM are not regularized, we do not use the previous method of directly computing the information entropy. Instead, we sample a Gaussian distribution ζ of the same size and minimize the KL divergence between the feature vectors and the Gaussian distribution, as follows:

$$\mathcal{L}_3 = -\mathbb{E}[\sum_i \sum_j KL(h_i^{(j)}, \zeta)] \tag{4}$$

Note that we only compute the features for the answer $\boldsymbol{a}$ part.

D. Optimization

As mentioned above, We design our attack method by maximizing the uncertainty of both the VLMs' output distribution and their internal features. Building upon this, we further introduce the maximum entropy joint attack method:

$$\begin{aligned} \max \quad & \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \lambda_3\mathcal{L}_3 \\ \text{s.t.} \quad & \left\|\boldsymbol{x} - \boldsymbol{x}^{adv}\right\|_p \leq \epsilon \end{aligned} \tag{5}$$

where λ_1, λ_2 and λ_3 are hyper-parameters to control the weights of each component.

To optimize this objective, we use the projected gradient descent (PGD) algorithm [7]. The PGD algorithm iteratively optimizes to generate the final adversarial image. The update formula at each step is as follows:

$$\begin{aligned} \mathcal{L} =& \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \lambda_3\mathcal{L}_3 \\ \boldsymbol{x}_t^{adv} =& \boldsymbol{x}_{t-1}^{adv} + \alpha \times \text{sign}(\nabla_{\boldsymbol{x}_{t-1}^{adv}}\mathcal{L}) \end{aligned} \tag{6}$$

where α is the learning step size. The complete algorithm can be found in the Appendix.

IV. Experiments

In this section, we conduct extensive experiments to elucidate the effectiveness of our proposed method on various open-source VLMs.

A. Experimental Setup

Dataset. Consistent with previous works [10], [12], [13], we randomly select 1,000 images from both the ImageNet [17] and MS-COCO [18] validation set, respectively. We prompt the model to generate descriptions of images using the instruction *describe this image in one sentence*.

TABLE II: The attack success rate of our method based on human evaluation.

Dataset	BLIP	BLIP-2	InstructBLIP	MiniGPT-4	LLaVA1.5	LLaVA1.6
ImageNet	100.0	96.7	94.5	95.3	96.2	96.1
MS-COCO	100.0	96.8	94.3	96.7	95.9	96.2

Models. We evaluate the popular LVLMs in the open-source community, including BLIP [2], BLIP-2 [25], InstructBLIP [20], MiniGPT-4 [3], LLaVA1.5 [1], and LLaVA1.6 [4].

Comparison Methods. To demonstrate the performance advantages of the MIE method, we set up comprehensive comparison methods. Initially, we evaluate clean samples and samples with added Gaussian noise. Furthermore, we apply two variants of [14] for a fair comparison: (1) TAR: targeted attacks, where a sequence is randomly selected from the vocabulary as the predefined target, effectively simulating an untargeted attack. (2) UAC: untargeted attacks, utilizing the model's initial correct output as the predefined target. We also perform an image encoder-based attack as in [10] for comparison. In addition, we also add a GAN-based method [15], CroPA [12] and Verbose Image [13] to provide a more comprehensive comparison.

Evaluation Metrics. We adopt both automatic and manual methods for quantitative evaluation. The CLIP score [19] is used to evaluate the semantic alignment between images and textual descriptions. Manual evaluation is carried out independently by 5 people and the final average value is taken.

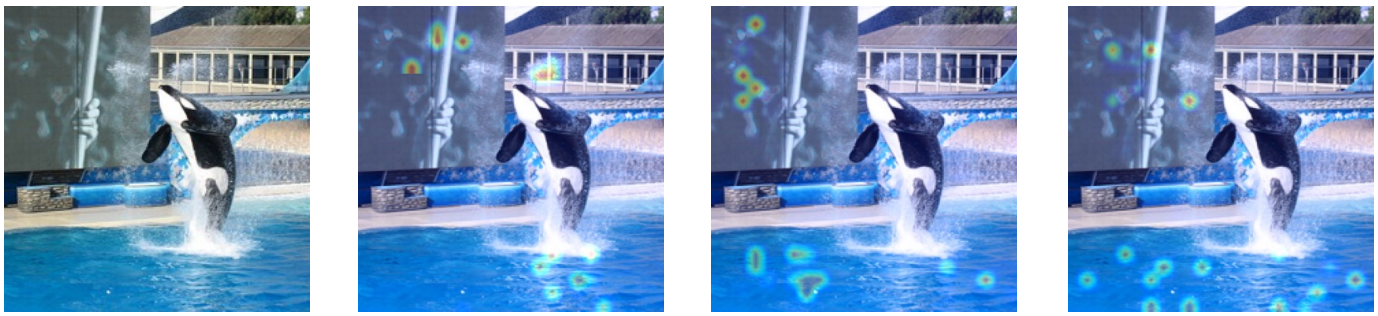
Parameters. We remain consistent with the experimental settings [10], [12]. Specially, we set $\epsilon = 8$ and employ L_∞ with the constraint $||\boldsymbol{x} - \boldsymbol{x}^{adv}||_\infty \leq 8$. We use a 100-step PGD to optimize our method. We experimentally set $\lambda_1 = \lambda_2 = \lambda_3 = 1$ and $\alpha = 1/255$.

More details can be found in the Appendix.

B. Main Results

As presented in Table I, we empirically evaluate the adversarial robustness of 6 available VLMs. The results are evaluated using CLIP Scores for the purpose of quantitatively measuring semantic similarity.

Compared with clean images, the performance of the victim models hardly change when presented with images augmented by Gaussian noise of the same size. This indicates that the



clean image | step = 0 | step = 5 | step = 10

Fig. 3: Visualization of attention heatmaps at different attack steps. The BLIP model's attention becomes increasingly diffused as the attack unfolds, thereby impeding its ability to concentrate on pivotal information.

a jar of candy candy candy jar with a candy candy candy jar filled with candy candy candy

a computer chip chip chip chip chip chip chip chip chip chip chip chip chip chip chip chip

Fig. 4: Sample demonstration of model output with fundamental logical errors. Under the MIE attacks, the model not only exhibits significant errors in understanding the images, but also demonstrates fundamental issues with sentence coherence and fluency.

model has good robustness to natural noise, whose robustness can be attributed to the training on large-scale datasets.

Our MIE attack method introduces slight changes that greatly affect the model's image understanding. It effectively reduces the CLIP score from about 30 to 20. Compared with previous methods, MIE is more effective and outperforms them by more than 2 points under various settings. Previous methods have relied solely on the model's output to guide adversarial noise, while the MIE approach also utilizes the internal features of the victim model to guide the generation of adversarial examples. Moreover, MIE does not require any predefined textual targets, allowing for a larger perturbation space. On the other hand, BLIP is more susceptible to attacks compared to other models. We believe this is primarily because its use of cross-attention for image feature utilization places a greater emphasis on visual information. Models with more parameters seem to have stronger adversarial robustness, but this is not absolute.

Regarding human evaluation, as presented in Table II, MIE has an average attack success rate of nearly 100% based on manual evaluation. This straightforward metric highlights the vulnerability of existing VLMs.

TABLE III: The impact of different combinations of three loss objectives against the BLIP model.

$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	ImageNet	MS-COCO
✓			18.45	18.96
	✓		19.69	19.79
		✓	19.54	19.72
✓	✓		18.31	18.65
✓		✓	18.19	18.54
	✓	✓	18.23	18.96
✓	✓	✓	**17.53**	**17.61**

TABLE IV: The impact of perturbation size.

ϵ	1	2	4	8	16	32	64
MIE	28.0	23.8	20.6	17.5	17.5	17.6	17.5

C. Ablation Study

In this section, we investigate the impact of various factors on the attack performance of MIE.

The impact of different combinations of three loss objectives. To disrupt the output probability distribution and internal features of the victim models, we design three loss objectives based on the architecture of VLMs, each of which can independently initiate an attack. As indicated in Table III, experimental results prove that a combined attack significantly impairs the model's visual comprehension abilities.

The impact of perturbation size. As shown in Table IV, increasing the size of perturbations enhances the effectiveness of the attack, and the MIE method converges when the perturbation value reaches 8, demonstrating its efficiency. Additional examples in the Appendix illustrate that MIE introduces such minimal perturbations to images that they are difficult to detect with the human eye.

The impact of the number of iterations. As illustrated in Table V, the number of attack steps significantly impacts the performance of MIE, with higher iteration numbers leading to better results. Furthermore, the attack effectiveness does not converge within 100 steps, indicating that increasing the number of iterations can lead to stronger attack results. This is because MIE expands the attack space with each iteration. Detailed cases are provided in the Appendix.

TABLE V: The impact of the number of iterations.

Step	5	20	30	50	80	100	200
MIE	21.0	18.8	18.3	18.2	18.1	17.5	16.4

D. Further Analyses

We delve into the changes in the model's attention, the actual effects of the attacks, and potential defense strategies.

The changes in the attention. Besides presenting the experimental results, we also visualize how the attention changes during the attack. As depicted in Figure 3, the BLIP model has a relatively concentrated attention distribution on clean images. However, as the attack progresses, the model's attention becomes more dispersed, significantly disrupting its focus and leading to erroneous perceptions.

The actual effects of the attacks. In Figure 1, we illustrate a scenario where the model, following an attack, displays errors in image comprehension. We also observe more severe errors in the model after being attacked with a high number of iterations, resulting in generated output statements that are illogical and incoherent, as demonstrated in Figure 4. This attack yields the most potent effect, causing the model to lose its fundamental language capabilities, thereby significantly impacting the user experience. A feature of our method is its ability to generate varying attack outcomes by adjusting the number of attack iterations. Therefore, different attack steps can be customized to ensure the model's output is misleading yet difficult to detect. More cases can be seen in the appendix.

Potential defense strategies. Given the complexity of performing adversarial training on large VLMs, we leave this for future work. We test the effectiveness of MIE attacks on BLIP-2 using RandomHorizontalFlip as a data augmentation method and found it to be only mildly effective, with the CLIP score improving from 19.63 to 21.04.

V. Conclusion

To evaluate the adversarial robustness of VLMs, we propose the maximizing information entropy method to maximize the uncertainty of model output distribution and internal features. We contend that previous methods, which calculate the loss only for the output part of the model to generate adversarial examples, are insufficient. Our extensive experiments on 6 models and 2 datasets demonstrate that our MIE is a more effective attack method. Such comprehensive attacks make enhancing the adversarial robustness of VLMs more challenging and warrant further investigation in future research.

References

[1] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee, "Improved baselines with visual instruction tuning," in *CVPR*, 2024, pp. 26286–26296.

[2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022, vol. 162, pp. 12888–12900.

[3] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," in *ICLR*, 2024.

[4] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014.

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[8] Eugene Bagdasaryan, Tsung-Yin Hsieh, Ben Nassi, and Vitaly Shmatikov, "(ab)using images and sounds for indirect instruction injection in multi-modal llms," *CoRR*, vol. abs/2307.10490, 2023.

[9] Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin, "Jailbreaking attack against multimodal large language model," *CoRR*, vol. abs/2402.02309, 2024.

[10] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin, "On evaluating adversarial robustness of large vision-language models," in *NeurIPS*, 2023.

[11] Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei Koh, Daphne Ippolito, Florian Tramèr, and Ludwig Schmidt, "Are aligned neural networks adversarially aligned?," in *NeurIPS*, 2023.

[12] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr, "An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models," in *ICLR*, 2024.

[13] Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu, "Inducing high energy-latency of large vision-language models with verbose images," in *ICLR*, 2024.

[14] Christian Schlarmann and Matthias Hein, "On the adversarial robustness of multi-modal foundation models," in *ICCV*, 2023, pp. 3679–3687.

[15] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Mubarak Shah, and Ajmal Mian, "Language model agnostic gray-box adversarial attack on image captioning," *TIFS*, vol. 18, pp. 626–638, 2023.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[18] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft COCO: common objects in context," in *ECCV*, 2014, vol. 8693, pp. 740–755.

[19] 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, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, vol. 139, pp. 8748–8763.

[20] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *CoRR*, vol. abs/2305.06500, 2023.

[21] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu, "Coca: Contrastive captioners are image-text foundation models," *Trans. Mach. Learn. Res.*, vol. 2022, 2022.

[22] Jianping Zhang, Yizhan Huang, Weibin Wu, and Michael R. Lyu, "Transferable adversarial attacks on vision transformers with token gradient regularization," in *CVPR*, 2023, pp. 16415–16424.

[23] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal, "Visual adversarial examples jailbreak aligned large language models," in *AAAI*, 2024, pp. 21527–21536.

[24] Yarin Gal and Zoubin Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016, vol. 48, pp. 1050–1059.

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023, vol. 202, pp. 19730–19742.

[26] Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqian Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, Xiaochun Cao, and Philip H. S. Torr, "A survey on transferability of adversarial examples across deep neural networks," *CoRR*, vol. abs/2310.17626, 2023.