NON-TARGETED ADVERSARIAL ATTACKS ON VISION-LANGUAGE MODELS VIA MAXIMIZING INFORMATION ENTROPY

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial examples pose significant security concerns in deep neural networks and play a crucial role in assessing the robustness of models. Nevertheless, existing research has primarily focused on classification tasks, while the evaluation of adversarial examples is urgently needed for more complex tasks. In this paper, we investigate the adversarial robustness of large vision-language models (VLMs). We propose a non-targeted white-box attack method that maximizes information entropy (MIE) to induce the victim model to generate misleading image descriptions deviating from reality. Our method is thoroughly analyzed experimentally, with validation conducted on the ImageNet dataset. The comprehensive and quantifiable experimental results demonstrate a significant success rate achieved by our method in adversarial attacks. Given the consistent architecture of the language decoder, our proposed method can serve as a benchmark for evaluating the robustness of diverse vision-language models.

1 INTRODUCTION

Large vision-language models (VLMs) such as GPT-4 (OpenAI, 2023) have emerged as a prominent research area in the field of artificial intelligence (Yin et al., 2023), with remarkable success in various domains, such as image caption generation (Guo et al., 2022), visual question answering systems (Zhu et al., 2023), image retrieval and search (Li et al., 2022), and visual document understanding (Cao et al., 2023). Leveraging extensive training data and computational resources, these vision-language models exhibit strong robustness and generalization when confronted with diverse and unstructured image data in open-domain settings.

On the other hand, the widespread deployment of large VLMs has raised significant security concerns, especially in life-critical scenarios such as autonomous driving (Zhu et al., 2020). In addition, a maliciously manipulated model can impact users and shape the public opinions by generating biased, misleading, or harmful content. These concerns underscore the urgent necessity for research on the robustness of VLMs (Kurakin et al., 2017; Sheng et al., 2021).

Recent works (Bagdasaryan et al., 2023; Carlini et al., 2023; Zhao et al., 2023a) have highlighted the vulnerabilities of multimodal models to adversarial examples, which refer to misleading samples generated by making small modifications to the original input. While being almost imperceptible to the human eye, these modifications are sufficient to deceive machine learning models and produce incorrect outputs (Szegedy et al., 2014).

In contrast to previous research primarily focusing on classification models, this study investigates the adversarial robustness of vision-language models equipped with text decoders. Specifically, we examine the task of image-grounded text generation, where VLMs are exploited to comprehend the content of images. By providing instructions to VLMs to describe the content of an image, they generate the corresponding textual outputs. VLMs can be attacked from both image and text perspectives, but manipulating text involves a substantial amount of searching due to its discrete nature. Therefore, attacking from the continuous image space is often more feasible (Carlini et al., 2023).

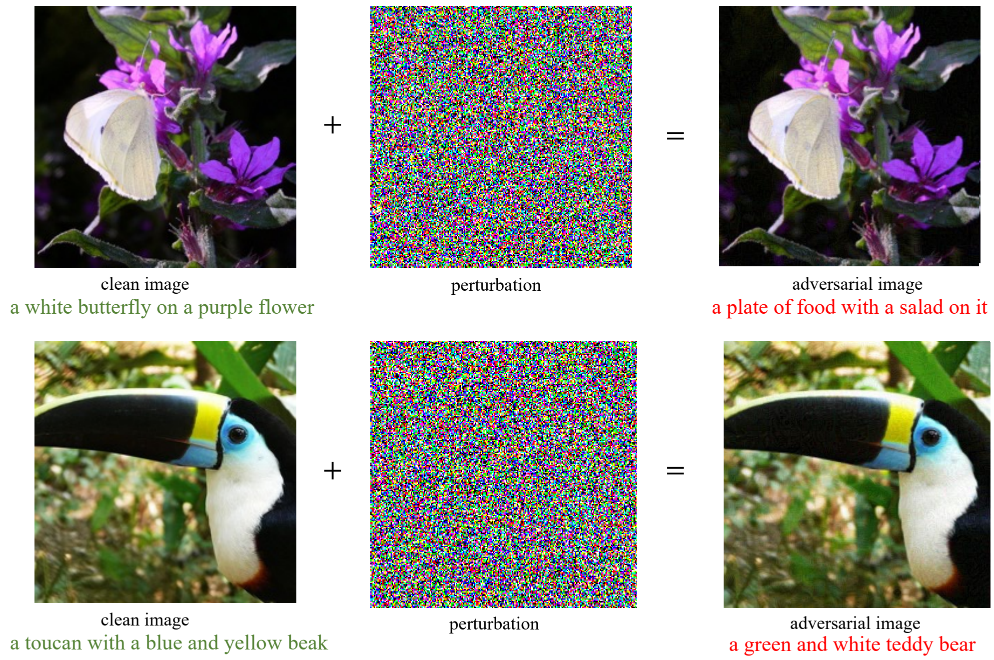


Figure 1: **An illustration for non-targeted adversarial attacks on VLMs.** The clean images, when perturbed with the subtle yet malicious noise, transform into the adversarial images. These adversarial images can cause the model to generate unpredictable, anomalous, or erroneous outputs.

Attacks on VLMs are highly complex and do not have a strict definition as in classification tasks. This is because images typically have only one correct category label, whereas image descriptions can have an infinite number of variations. Moreover, descriptions generated from different perspectives may vary significantly in terms of representation, yet their semantics could still be consistent. Recent works have primarily focused on inducing the model to produce specific, often undesirable, information, such as toxic or biased text, or bypassing the model's alignment constraints to achieve jailbreak attacks (Qi et al., 2023). In comparison, this work is the first to evaluate the non-targeted adversarial robustness of VLMs without real supervisory signals to the best of our knowledge.

Using targeted attack methods to perform untargeted attacks through specific settings, such as inducing the model to generate illogical texts Carlini et al. (2023); Schlarmann & Hein (2023), may be insufficient for image description tasks. A model's description of an image that deviates from a correct label does not necessarily imply a completely incorrect description. In fact, directing the model to move in a specific direction is a more challenging task.

In this paper, we propose a method of Maximizing Information Entropy (MIE) based on the common model structures of vision-language models. This method attacks multiple components of the decoder, inducing it to generate high-entropy information, thereby achieving a white-box attack without a predefined target. To ensure reproducibility of our results, we evaluate our method on multiple open-source VLMs. Specifically, we randomly select 1000 images from the ImageNet validation set as test samples and then employ a range of models, including BLIP (Li et al., 2022), BLIP2 (Li et al., 2023), InstrucBLIP (Dai et al., 2023), MiniGPT-4 (Zhu et al., 2023), LLaVA (Liu et al., 2023), etc., to generate the corresponding adversarial examples, followed by regenerating new image descriptions. Finally, we quantify the interference of adversarial examples on the robustness by calculating the CLIP Score (Radford et al., 2021) and manually inspecting the results. The experimental results show that even though VLMs have stronger robustness against Gaussian noise, they are still severely disrupted by the adversarial attacks we launched.

Our work can provide a new benchmark for evaluating the robustness of vision-language models and inspire more follow-up research to explore the risks that may be encountered before deploying these models.

The main contributions of this paper are as follows:

- We analyze the differences between targeted and non-targeted attacks and provide a theoretical explanation for the inability of targeted attacks to efficiently implement non-targeted attacks.
- We propose the Maximizing Information Entropy (MIE) method, which firstly achieves non-targeted white-box attacks on vision-language models without authentic labeling data.
- We conduct extensive experiments to validate the effectiveness of our approach. The experimental results quantitatively demonstrate that our method can effectively attack large vision-language models.

2 Related Work

2.1 Vision-Lauguage Models

By synergistically combining state-of-the-art language models with cutting-edge visual perception models, vision-language models have demonstrated remarkable capabilities in multimodal understanding. From the perspective of the interaction between images and text, these models can be classified into two categories.

Image as Key-Value. The first group of models involve utilizing the features extracted by an image encoder as Key and Value components, while treating the input text as the Query during the decoding process of the language model (Li et al., 2022; Yu et al., 2022; Xu et al., 2023a). The next output at each time step is then computed using a cross-attention mechanism. This approach highlights the role of images and is particularly suitable for dense image prediction tasks (Kim et al., 2022; Alayrac et al., 2022; Cao et al., 2023).

Image as Token. Another widely used approach is to convert images into token sequences, which are aligned with the feature space of text, enabling the interaction between images and text (Li et al., 2023; Dai et al., 2023; Bao et al., 2022; Zhu et al., 2023; Liu et al., 2023). One advantage of this method is that it can fully leverage the capabilities of large language models without requiring any modifications. This concise architecture has gained increasing attention in the field of multimodal learning in the short term.

Despite the differences regarding the cross-modality interactions, both structures share a similar transformer decoder. Therefore, our attack method is applicable to both architectures, as we generate image perturbations guided by the textual signals of the model's output.

2.2 Adversarial Examples

Adversarial Attacks on Classification Models. The vulnerability of machine learning models to adversarial examples has been extensively studied, with a primary focus on image data (Szegedy et al., 2014; Goodfellow et al., 2015; Mao et al., 2023). The objective of these works is to add minimal perturbations to an image, causing significant errors in the classifier while remaining imperceptible to human observers (Mahmood et al., 2021). Recent advancements (Xu et al., 2023b; Zhang et al., 2023; 2022a) have delved into the internal structure of specific networks, modifying gradients during the backpropagation process. These techniques have achieved significant success in terms of both the effectiveness and the transferability of white-box attacks.

Adversarial attacks on VLMs. In contrast to the extensive studies regarding adversarial attacks on classification models, the research on adversarial robustness of VLMs remains limited, with many undefined issues. Inheriting the characteristics of large language models, VLMs introduce further complexity to adversarial attacks. The current focus revolves around inducing targeted outputs from the models as the objective. Specifically, Carlini et al. (2023); Qi et al. (2023) treat toxic text as the target suffix and employ standard teacher-forcing optimization techniques to generate adversarial examples that bypass the alignment constraints of the model. Other works (Bagdasaryan et al., 2023) explore images for indirect prompt and instruction injection. The work closest to ours is Schlarmann & Hein (2023), where they use the ground truth caption to calculate the loss and degrade the output quality, enabling a non-targeted attack.

In this paper, we do not rely on true image descriptions. Instead, we employ white-box attacks to degrade the visual understanding capability of VLMs in an open-world scenario.

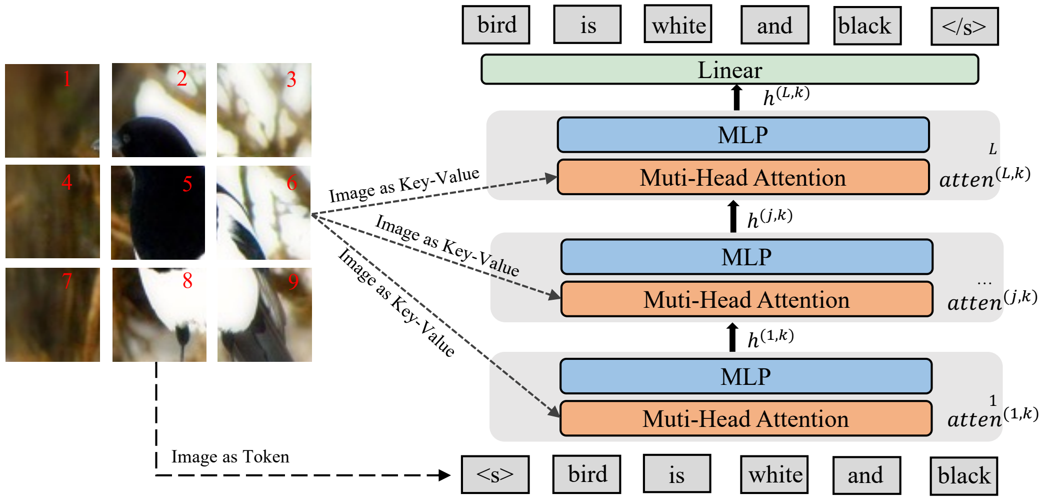


Figure 2: **An overview of attacks targeting multiple components of the vision-language model**. We present the shared modules in most vision-language models and demonstrate that attacks can be conducted from three perspectives: logits, attentions, and hidden states.

2.3 Information Entropy

Information entropy is a concept in information theory that measures the uncertainty or disorder in a closed system (Shannon, 1948; Zhou et al., 2022). In information theory, entropy is formally defined as the average amount of information carried by a random variable.

Our motivation in this paper is to maximize the information entropy of the model's understanding of images. A higher information entropy indicates that the model fails to focus on essential information, thereby susceptible to a non-targeted white-box attack. Since we aim to launch a non-targeted attack, inducing the model to generate high-entropy outputs will reduce the model's fundamental robustness, which contradicts the objective of reducing the training loss of neural network models.

3 Methodology

In this section, we first introduce the fundamental settings of adversarial attacks, followed with a detailed description of our proposed method.

3.1 Threat Models

We denote $g_\theta(\boldsymbol{x}, \boldsymbol{q}) \mapsto \boldsymbol{a}$ as a vision-language model parameterized by θ, where $\boldsymbol{x}$ is the input image, $\boldsymbol{q}$ is the input text, and $\boldsymbol{a}$ is corresponding output text in an auto-regressive 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. It is worth noting that by applying specific prompts, VQA can also achieve image caption generation.

For threat models employed for text generation tasks on vision-language models:

Adversary knowledge refers to their understanding of the system's internal mechanisms, particularly in the case of white-box attacks, where the attackers have full access to the model parameters g_θ, they can also obtain the gradient information of the model.

Adversary goals describes the objectives that malicious attackers aim to achieve, including targeted and non-targeted attacks. In the context of VLMs, targeted attacks refer to inducing the model to produce desired outputs, while non-targeted attacks aim to generate incorrect captions or answers. In this paper, we investigate the fundamental aspects of adversarial robustness in VLMs with the goal to reduce the quality of the model's outputs.

Adversary capabilities elucidates 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 $\left\|\boldsymbol{x}-\boldsymbol{x}^{adv}\right\|_p \leq \epsilon$.

Subsequently, we introduce the attacks proposed in this work, which are conducted at three distinct levels: logits, attention, and hidden states (Vaswani et al., 2017), as illustrated in Figure 2.

3.2 Logits-based Entropy Maximization

Vision-language models leverage image encoders to extract features from images, which are then combined with a language decoder to generate token sequences in an auto-regressive manner as shown in Figure 2. For each position i of the model's 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 with high confidence at each step. However, when the model encounters challenging examples, its output may become ambiguous. In the most extreme scenario, the model assigns equal probabilities to every token, resulting in random and ungrammatical outputs. This aligns with the definition of information entropy. Motivated by this, we apply a logits-based maximum entropy attack:

$$\begin{aligned} &\max -\mathbb{E}[\sum_i \sum_j \log_2(p_i^{(j)}) p_i^{(j)}] \\ &\text{s.t.} \left\|\boldsymbol{x}-\boldsymbol{x}^{adv}\right\|_p \leq \epsilon \end{aligned} \tag{2}$$

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

Since it is a white-box attack setting, the gradients of the target can be obtained through backpropagation, which can then be used for optimization using projected gradient descent (Madry et al., 2018). 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.

3.3 Attention-based Entropy Maximization

Attention is a crucial component in the Transformer model, allowing it to focus on different positions of the input sequence and weight them based on their importance when processing sequential data. The Attention mechanism in Transformers consists of three components: Query, Key, and Value. By computing the dot product between the Query and each Key, the corresponding attention weights are obtained. In general, for an image, only a small proportion of pixels or patches are relevant to the prompt.

Formally speaking, for a transformer decoder with L layers, the computation of each token $\boldsymbol{a}_i$ will generate L attention weights $atten_i \in \mathbb{R}^{L\times(P+T)}$, where $T = 0$ if the interaction mode is *Image as Key-Value*, and $T = i - 1$ if the mode is *Image as Token*.

Similar to the previous perspective, to prevent the model from focusing on salient information and reduce its understanding of the image, we utilize an attention-based maximum entropy attack in a layer-by-layer manner:

$$\begin{aligned} &\max -\mathbb{E}[\sum_i \sum_j \sum_{k=1}^{i-1} \log_2(atten_i^{(j,k)}) atten_i^{(j,k)}] \\ &\text{s.t.} \left\|\boldsymbol{x}-\boldsymbol{x}^{adv}\right\|_p \leq \epsilon \end{aligned} \tag{3}$$

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

Algorithm 1: Maximizing Information Entropy Method.

Input: Vision-language model g with the parameter θ, clean image $\boldsymbol{x}$.
Input: Perturbation bound ϵ, iteration steps S and learning rate α.
Output: Adversarial image $\boldsymbol{x}^{adv}$.

Initialize $\boldsymbol{x}^{adv} = \boldsymbol{x}$;
Enable gradients for variable $\boldsymbol{x}^{adv}$;
foreach *step in 1, 2, ..., S* **do**
    logits, attentions, hidden states = $g(\boldsymbol{x}, q)$;
    Calculate $\mathcal{L}_1$ using Equation 2;
    Calculate $\mathcal{L}_2$ using Equation 3;
    Calculate $\mathcal{L}_3$ using Equation 4;
    $\mathcal{L} = \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \lambda_3\mathcal{L}_3$
    $\boldsymbol{x}^{adv} = \boldsymbol{x}^{adv} + \alpha\,\mathrm{sign}(\nabla_{\boldsymbol{x}^{adv}}(\mathcal{L}))$
    $pert = \mathrm{Clip}(\boldsymbol{x}^{adv} - \boldsymbol{x}, -\epsilon, \epsilon)$;
    $\boldsymbol{x}^{adv} = \mathrm{Clip}(\boldsymbol{x} + pert, 0, 1)$

3.4 Hidden States-based Entropy Maximization

In the Transformer model, *hidden states* typically refer to the outputs at each position of the encoder and decoder. Specifically, each input token undergoes a series of cross-attention and feed-forward neural network layers, resulting in a corresponding hidden state vector $h_i \in \mathbb{R}^{L\times d}$ in the decoder, where d is the dimension of embedding. These hidden state vectors contain information from different positions in the input sequence and can be regarded as the encoded and feature-extracted representations of the input.

Similar to attention, well-learned representations are also not evenly distributed but tend to concentrate on specific positions, where some positions have higher values while others have lower values. Based on this observation, we implement a hidden states-based maximum entropy attack in a layer-by-layer manner:

$$\begin{aligned} &\max -\mathbb{E}[\sum_i\sum_j\sum_{k=1}^{i-1}\log_2(\mathcal{F}(h_i)^{(j,k)})\mathcal{F}(h_i)^{(j,k)}] \\ &\text{s.t. } \left\|\boldsymbol{x}-\boldsymbol{x}^{adv}\right\|_p \leq \epsilon \end{aligned} \tag{4}$$

where $\mathcal{F}$ is the softmax function as h_i is not a normalized probability.

3.5 Implementation

As mentioned above, we propose three non-targeted attack methods to perturb the Transformer model's understanding of images. We denote these objectives as $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$, respectively. Building upon this, we further introduce the maximum entropy joint attack method:

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

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

The complete algorithmic flow for the adversarial attack is shown in Algorithm 1. Due to the iterative utilization of the model's autoregressive inference method, the generated image captions from the model may differ at each iteration, thereby increasing the space for adversarial perturbations in the attacks. Case studies are presented in the experiments and the appendix.

4 Experiments

In this section, we conduct an extensive series of experiments to elucidate the effectiveness of our proposed method on various open-source vision-language models. We begin by outlining the exper-

Table 1: **White-box 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. The best results are marked in bold.

VLM	Info.		Baseline		Attacking method			
	# Param.	Res.	Clean	Gaussian	Carlini	Schlarmann	Aafaq	MIE
BLIP	224M	384	29.79	29.65	20.53	19.87	24.36	**17.80**
BLIP-2	3.7B	224	30.72	30.74	24.58	24.06	27.78	**21.39**
InstructBLIP	7.9B	224	31.36	31.33	24.31	23.80	25.32	**21.65**
LLaVA	13.3B	224	31.52	31.49	24.78	24.12	25.79	**21.41**
MiniGPT-4	14.4B	224	31.44	31.23	24.97	23.16	24.12	**21.11**

Table 2: **Success rate of attacks against victim models.** Due to the considerable cost associated with human resources, we exclusively measure the model's error rate on the original data and the success rate of the final attack against the victim models to demonstrate the adversarial performance of MIE. With the exception of BLIP, all other models achieve 100% accuracy on clean images.

	BLIP	BLIP-2	InstructBLIP	LLaVA	MiniGPT-4	**Mean**
Initial Error Rate	0.2	0.0	0.0	0.0	0.0	**0.04**
Attack Success Rate	100.0	96.7	94.5	96.3	96.9	**96.88**

imental setup, followed by showcasing the results of our attacks. Lastly, we delve into the attack process through visualization and specialized case analyses.

4.1 Experiment Setup

Dataset. We adhere to the commonly used configuration for adversarial attacks, as outlined in Zhao et al. (2023b), and randomly select 1000 images from the ILSVRC 2012 validation set for our experiments (Deng et al., 2009).

Models. In accordance with the settings for white-box attacks, we conduct an evaluation of the adversarial robustness of several influential vision-language models within the open-source community to ensure reproducibility. These models include BLIP (Li et al., 2022), BLIP-2 (Li et al., 2023), InstructBLIP (Dai et al., 2023), Mini GPT-4 (Zhu et al., 2023) and LLaVA (Liu et al., 2023). With the exception of BLIP, other models have incorporated language models based on LLaMA (Touvron et al., 2023) or OPT (Zhang et al., 2022b), thus expanding their multimodal capabilities to facilitate a broader interaction between visual and textual data.

Comparison methods. To demonstrate the performance advantages of MIE, we set up various comparison methods. In addition to clean samples and samples with Gaussian noise as baselines, we also compare our method with Carlini et al. (2023) (performing targeted attacks using random targets), Schlarmann & Hein (2023) (utilizing descriptions of clean samples as ground-truth labels), and Aafaq et al. (2023) (a GAN-based method).

Evaluation Metrics. We employ both automated and manual methodologies to quantitatively assess the model's performance. The CLIP score (Radford et al., 2021) is used to evaluate the semantic alignment between images and textual descriptions. It is calculated by measuring the cosine similarity between vectors generated by CLIP's image encoder and text encoder. Additionally, we evaluate the success rate of attacks through manual assessments. An attack is considered successful if the adversarial example results in factual inaccuracies in the generated descriptions and images, including but not limited to color discrepancies or incorrect object categorizations.

Parameters. We remain consistent with the experimental configurations (Zhao et al., 2023b). 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. λ_1, λ_2 and λ_3 are experimentally set to 0.8, 0.1, and 0.1, respectively.

4.2 Empirical studies

In this section, we empirically evaluate the adversarial robustness of five available vision-language models using our proposed method. The results of the automated evaluation using CLIP are presented in Table 1.

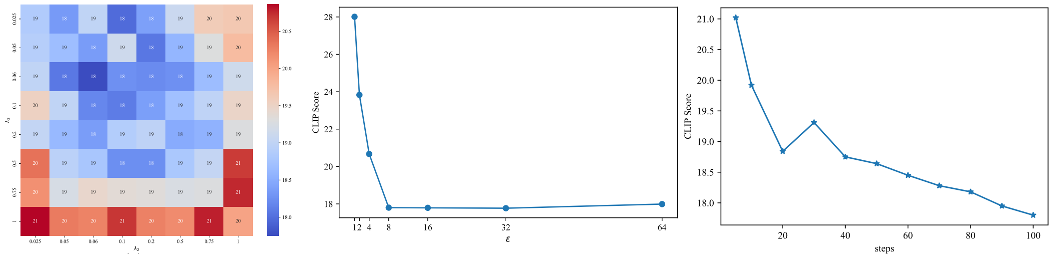


Figure 3: **Ablation study on loss coefficients, perturbation size, and the number of iterations.** For the BLIP model, we conduct ablation experiments by fixing $\lambda_1 = 0.5$ and varying λ_2 and λ_3.

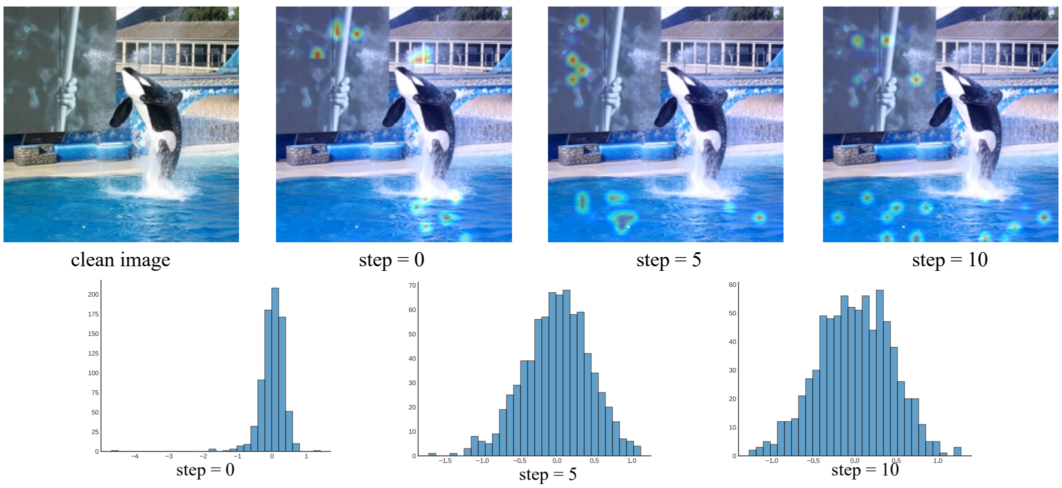


Figure 4: **Visualization of attention heatmaps and hidden states at different attack steps.** As the attack progresses, the internal states transition from being concentrated to dispersed.

Firstly, when compared to clean images, the models display minimal performance degradation when presented with images augmented with equally-sized Gaussian noise. Nevertheless, employing the MIE attack methods to generate subtle perturbations can indeed result in a significant decrease in model performance. Regarding the attack effectiveness, the MIE method exhibits remarkable results by reducing the CLIP score for image comprehension by the model from approximately 30 to around 20. Compared to some existing works, MIE achieves more effective attack results, surpassing opponents by more than 2 points on multiple models.

As depicted in Table 2, our attack algorithm attains an average attack success rate of 96.88% based on manual evaluation. This straightforward metric highlights the vulnerability of existing large vision-language models.

From the varying model performances, it becomes evident that all models are highly vulnerable and susceptible to attacks. However, larger models demonstrate improved robustness when attacks are targeted at attentions and hidden states. This can be attributed primarily to the fact that larger models often employ an *Image as Token* architecture, which benefits from the enhanced resilience of large language models. In contrast, in cases of an *Image as Key-Value* architecture, where a higher proportion of parameters are allocated to the image modality (intuitively, the image modality accounts for two-thirds of the parameters), attacks directed at the image modality tend to be more effective.

4.3 Ablation Study

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

As shown in Figure 3-(a), the experimental results of different loss coefficients indicate that the optimal results are concentrated around the ratio of $\lambda_1 : \lambda_2 : \lambda_3 = 0.5 : 0.06 : 0.06 (\approx 8 : 1 : 1)$.

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

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

a sign for the new mexico border crossing site in the mexican border border border border border border border

the new mesh mesh mesh mesh mesh mesh mesh mesh mesh mesh mesh mesh mesh mesh mesh mesh mesh

Figure 5: **Sample demonstration of model output with fundamental logical errors.** Under the influence of MIE attacks, the model not only exhibits significant errors in understanding the images, but also demonstrates fundamental issues with sentence coherence and fluency.

For different models, additional coefficient settings may generate better results. As illustrated in Figure 3-(b), the perturbation size does not necessarily result in better performance when increased as the maximum attack effect is achieved when the size exceeds 8. As depicted in Figure 3-(c), the number of attack steps significantly impacts MIE's performance, with higher iteration numbers leading to better results. Specific cases are provided in the appendix.

4.4 Further analyses

In addition to presenting the experimental results, we have also visualized the changes in attention during the attack process. As depicted in Figure 4, the model initially exhibits effective attention towards clean images. However, as the attack unfolds, the model's attention becomes progressively dispersed, resulting in a significant disruption of its focus. This disruption gradually leads to substantial errors in the model's comprehension.

Furthermore, the model's representation learning capability is severely compromised due to the intrusion. As illustrated in Figure 4, the hidden states learned from clean images exhibit a relatively focused distribution. However, after the attack, the hidden states become less distinct and more diffuse. It is worth noting that this method of attacking from within the Transformer structure exhibits a high degree of generality.

In Figure 1, we illustrate a scenario where the model, following an attack, displays errors in image comprehension. Specifically, the generated image description deviates from the actual facts, including misidentification of the subject. We also observe more severe errors in the model after being attacked, resulting in generated output statements that are illogical and incoherent, as demonstrated in Figure 5. It is critical to rectify such glaring errors in the model before its deployment, as they can significantly impact the user experience. More cases can be seen in the appendix.

5 Conclusion

In this paper, our primary focus is on adversarial attacks directed at vision-language models (VLMs). In order to assess their adversarial robustness, we propose the Maximizing Information Entropy (MIE) algorithm for conducting white-box attacks on large vision-language models. Notably, this approach does not require prior knowledge of the authentic image captions. Instead, it iteratively generates image descriptions. By inducing perturbations in the universal Transformer structure, including logits, attentions, and hidden states, with the objective of maximizing information entropy, we disrupt the model's image understanding capabilities. This disruption leads to erroneous image descriptions and, in some cases, results in incoherent sentences. Our experimental results indicate that the MIE algorithm achieved a 96.88% success rate in its attacks. This highlights a significant vulnerability in existing large VLMs, which remain highly susceptible to adversarial attacks. This susceptibility raises substantial security concerns regarding the deployment of such models. Given the complexity of training large vision-language models, we defer the exploration of the corresponding adversarial defense strategies to future research.

REPRODUCIBILITY AND ETHICS STATEMENT

To ensure maximum reproducibility of this work, we have provided highly specific details in the paper. The core ideas are presented in Section 3 and Section 4. Moreover, we have aligned our primary experimental setups as closely as possible with previous works. It is important to note that reproducing special cases may not always be possible since the output of vision-language models can have a degree of randomness. Additionally, we provide as many cases as possible in the appendix for direct reference.

Regarding ethics, this paper may pose a certain threat to the deployment of large vision-language models. Given the simplicity of the method proposed, we advocate for similar assessments before the deployment of models. Furthermore, we will continue our research on the corresponding adversarial defense algorithms.

REFERENCES

Nayyer Aafaq, Naveed Akhtar, Wei Liu, Mubarak Shah, and Ajmal Mian. Language model agnostic gray-box adversarial attack on image captioning. *IEEE Trans. Inf. Forensics Secur.*, 18:626–638, 2023. doi: 10.1109/TIFS.2022.3226905. URL `https://doi.org/10.1109/TIFS.2022.3226905`.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. URL `http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html`.

Eugene Bagdasaryan, Tsung-Yin Hsieh, Ben Nassi, and Vitaly Shmatikov. (ab)using images and sounds for indirect instruction injection in multi-modal llms. *CoRR*, abs/2307.10490, 2023. doi: 10.48550/arXiv.2307.10490. URL `https://doi.org/10.48550/arXiv.2307.10490`.

Hangbo Bao, Wenhui Wang, Li Dong, and Furu Wei. Vl-beit: Generative vision-language pretraining. *CoRR*, abs/2206.01127, 2022. doi: 10.48550/arXiv.2206.01127. URL `https://doi.org/10.48550/arXiv.2206.01127`.

Haoyu Cao, Changcun Bao, Chaohu Liu, Huang Chen, Kun Yin, Hao Liu, Yinsong Liu, Deqiang Jiang, and Xing Sun. Attention where it matters: Rethinking visual document understanding with selective region concentration. *CoRR*, abs/2309.01131, 2023. doi: 10.48550/arXiv.2309.01131. URL `https://doi.org/10.48550/arXiv.2309.01131`.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramèr, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? *CoRR*, abs/2306.15447, 2023. doi: 10.48550/arXiv.2306.15447. URL `https://doi.org/10.48550/arXiv.2306.15447`.

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*, abs/2305.06500, 2023. doi: 10.48550/arXiv.2305.06500. URL `https://doi.org/10.48550/arXiv.2305.06500`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL `https://doi.org/10.1109/CVPR.2009.5206848`.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning*

Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL http://arxiv.org/abs/1412.6572.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven C. H. Hoi. From images to textual prompts: Zero-shot VQA with frozen large language models. *CoRR*, abs/2212.10846, 2022. doi: 10.48550/arXiv.2212.10846. URL https://doi.org/10.48550/arXiv.2212.10846.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pp. 498–517. Springer, 2022. doi: 10.1007/978-3-031-19815-1_29. URL https://doi.org/10.1007/978-3-031-19815-1_29.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJGU3Rodl.

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 Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

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 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *CoRR*, abs/2304.08485, 2023. doi: 10.48550/arXiv.2304.08485. URL https://doi.org/10.48550/arXiv.2304.08485.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. On the robustness of vision transformers to adversarial examples. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 7818–7827. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00774. URL https://doi.org/10.1109/ICCV48922.2021.00774.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=P4bXCawRi5J.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/arXiv.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak large language models. *CoRR*, abs/2306.13213, 2023. doi: 10.48550/arXiv.2306.13213. URL https://doi.org/10.48550/arXiv.2306.13213.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL `http://proceedings.mlr.press/v139/radford21a.html`.

Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. *arXiv preprint arXiv:2308.10741*, 2023.

Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL `https://doi.org/10.1002/j.1538-7305.1948.tb01338.x`.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 20346–20359, 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/aa97d584861474f4097cf13ccb5325da-Abstract.html`.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL `http://arxiv.org/abs/1312.6199`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/arXiv.2302.13971. URL `https://doi.org/10.48550/arXiv.2302.13971`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html`.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38728–38748. PMLR, 2023a. URL `https://proceedings.mlr.press/v202/xu23s.html`.

Zhuoer Xu, Zhangxuan Gu, Jianping Zhang, Shiwen Cui, Changhua Meng, and Weiqiang Wang. Backpropagation path search on adversarial transferability. *CoRR*, abs/2308.07625, 2023b. doi: 10.48550/arXiv.2308.07625. URL `https://doi.org/10.48550/arXiv.2308.07625`.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *CoRR*, abs/2306.13549, 2023. doi: 10.48550/arXiv.2306.13549. URL `https://doi.org/10.48550/arXiv.2306.13549`.

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.*, 2022, 2022. URL `https://openreview.net/forum?id=Ee277P3AYC`.

Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R. Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 14973–14982. IEEE, 2022a. doi: 10.1109/CVPR52688.2022.01457. URL https://doi.org/10.1109/CVPR52688.2022.01457.

Jianping Zhang, Yizhan Huang, Weibin Wu, and Michael R. Lyu. Transferable adversarial attacks on vision transformers with token gradient regularization. *CoRR*, abs/2303.15754, 2023. doi: 10.48550/arXiv.2303.15754. URL https://doi.org/10.48550/arXiv.2303.15754.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022b. doi: 10.48550/arXiv.2205.01068. URL https://doi.org/10.48550/arXiv.2205.01068.

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. *CoRR*, abs/2305.16934, 2023a. doi: 10.48550/arXiv.2305.16934. URL https://doi.org/10.48550/arXiv.2305.16934.

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. *CoRR*, abs/2305.16934, 2023b. doi: 10.48550/arXiv.2305.16934. URL https://doi.org/10.48550/arXiv.2305.16934.

Dawei Zhou, Nannan Wang, Xinbo Gao, Bo Han, Xiaoyu Wang, Yibing Zhan, and Tongliang Liu. Improving adversarial robustness via mutual information estimation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27338–27352. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhou22j.html.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023. doi: 10.48550/arXiv.2304.10592. URL https://doi.org/10.48550/arXiv.2304.10592.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 10009–10019. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01003. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Zhu_Vision-Language_Navigation_With_Self-Supervised_Auxiliary_Reasoning_Tasks_CVPR_2020_paper.html.

A APPENDIX

A.1 JOINT ATTACK

Table 3: **Results of independent attacks using the three sub-methods.** 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. We also present the results of each individual component as well as the combination of them. The best results are marked in bold.

VLM	Info.		Baseline		Attacking method			
	# Param.	Res.	Clean	Gaussian	Hidden	Attentions	Logits	MIE
BLIP	224M	384	29.79	29.65	20.54	19.69	18.25	**17.80**
BLIP-2	3.7B	224	30.72	30.74	27.18	26.02	21.48	**21.39**
InstructBLIP	7.9B	224	31.36	31.33	27.17	25.66	22.32	**21.65**
LLaVA	13.3B	224	31.52	31.49	29.31	28.65	22.20	**21.41**
MiniGPT-4	14.4B	224	31.44	31.23	29.98	28.31	22.15	**21.11**

Although applying each sub-method individually is effective to varying degrees, we believe that using them in combination is more efficient. Implementing a joint attack with different weights for these three losses offers the following advantages:

Multi-angle attack: By attacking logits, attention scores, and hidden states simultaneously, the model's prediction process can be disrupted from multiple angles. Consequently, even if the model exhibits strong robustness in a particular aspect, it would struggle to withstand attacks from various directions.

Weight adjustment: By adjusting the weights of different losses, optimization can be achieved according to the specific characteristics and attack objectives of the model. For instance, if the model is sensitive to logits perturbation, the weight of logits loss can be increased to improve the attack's effectiveness. **Interactive influence**: Logits, attention scores, and hidden states interact with each other within the model. A joint attack can exploit this interaction to enhance the attack's effectiveness. For example, by increasing the entropy of attention scores, the model's focus can be dispersed during the prediction process, thereby affecting the computation of hidden states and logits and reducing the logical coherence of the output text.

Stronger attack performance: Compared to individual attacks, a joint attack can achieve higher attack effectiveness in a shorter time, as it operates simultaneously in multiple directions, improving attack efficiency.

A.2 EFFECTIVENESS OF ADVERSARIAL TRAINING

Following conventional adversarial training, we find that large VLMs do not exhibit significant adversarial robustness for unseen samples. Our MIE attack based on autoregressive generation of pseudo-labels has a strong attack capability. In the future, we will delve into the effectiveness of adversarial training for VLMs.

A.3 STEP-BY-STEP EFFECTIVENESS

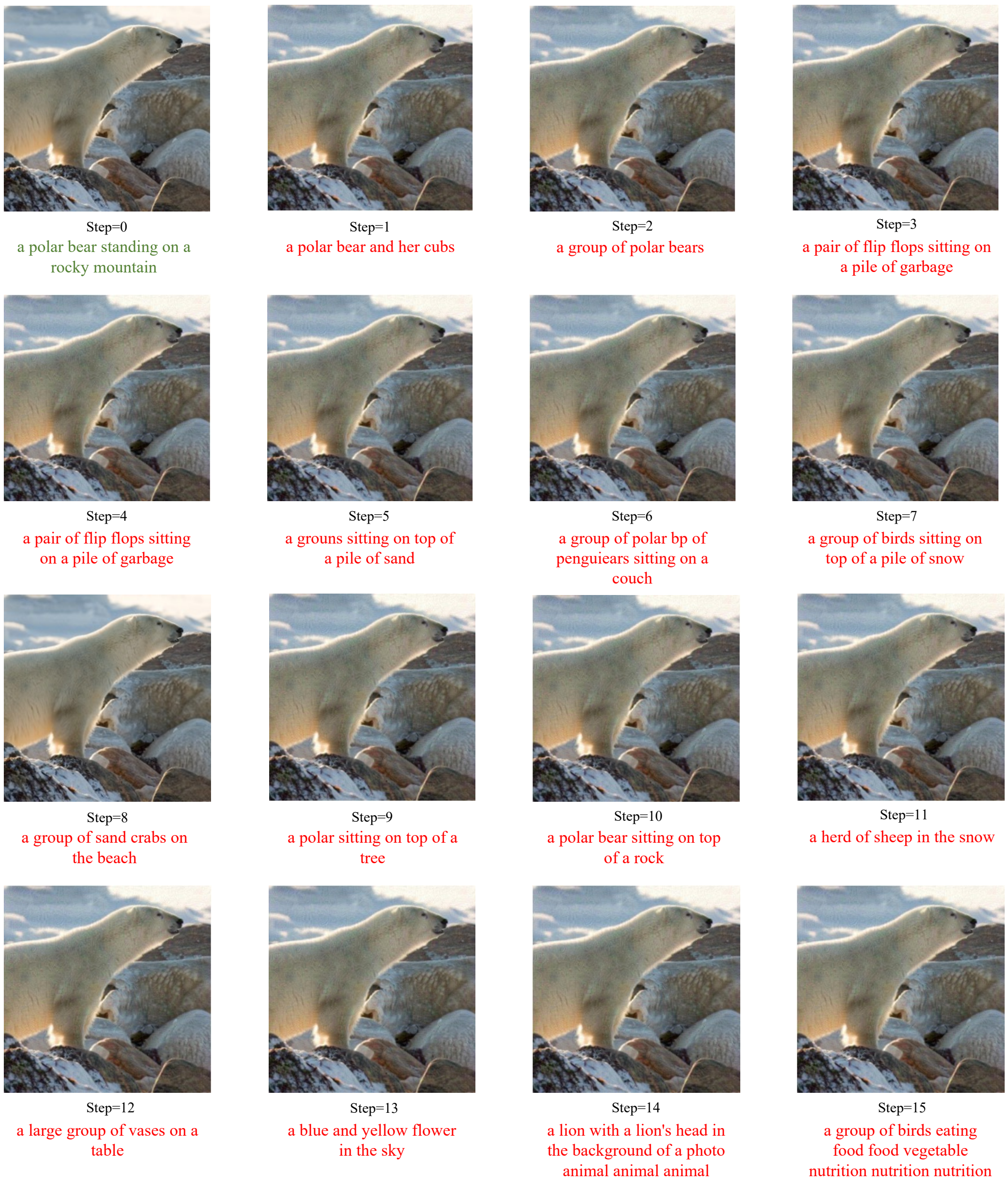
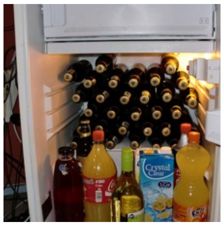


Figure 6: **Example showcase of step-by-step attacks.** From a semantic perspective, the model's understanding of the images progressively deteriorates, culminating in incoherent sentences.

A.4 MORE CASES OF SUCCESSFUL ATTACKS



clean image
a refrigerator with a bunch of soda bottles in it

adversarial image
a sign that reads the best beer in the world

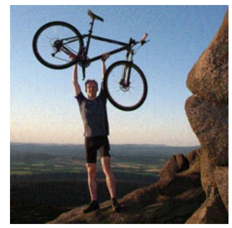
clean image
a black air filter on top of a box

adversarial image
the new e - z - e - z - z - e - z - z - z

clean image
a dog laying on the ground

adversarial image
a set of drill drills and a drill bit

clean image
a man holding a bike on top of a mountain



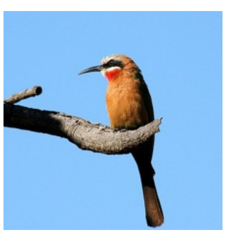
adversarial image
a poster for the new movie, the big bad

clean image
two men holding a fish on a boat

adversarial image
a white and black cake with a white and black cake top

clean image
a shaggy dog sitting on the grass

adversarial image
a sign that reads ‘no smoking’



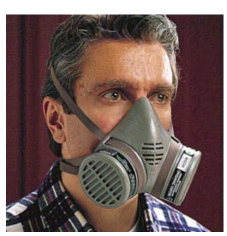
clean image
a bird sitting on a branch

adversarial image
a poster with a photo of a man in a suit and tie

clean image
a cup of coffee

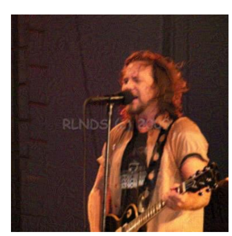
adversarial image
a white shirt with the words ‘i'm’ on it

clean image
a man wearing a gas mask

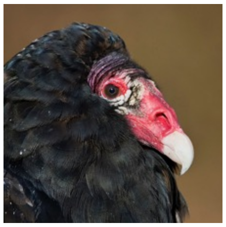


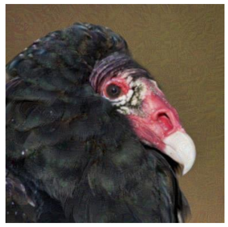
adversarial image
the back support system

clean image
a man with a guitar on stage



adversarial image
the new logo for the new game, the new game


clean image
a black bird with a red beak and a white head


adversarial image
a group of people with their heads in a circle


clean image
the parliament building in townsville


adversarial image
a truck with the words 'the best way to get a job'


clean image
a room with a view of a garden


adversarial image
a - e - e - e - e - e - e - e - e


clean image
a bunch of trains

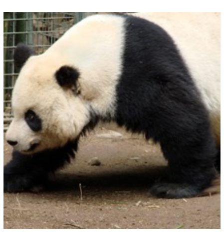

adversarial image
steel roof flashings for volvo v40


clean image
a drum and a drum stick sitting on a table


adversarial image
a bottle of wine with a label for the wine


clean image
a yellow van parked in a field


adversarial image
a man with a laptop and a laptop on the grass

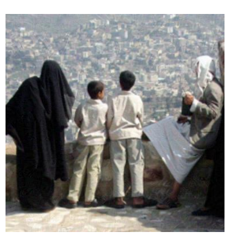

clean image
a panda bear walking on dirt


adversarial image
a machine with a piece of wood in it stock images


clean image
a shark swimming in the ocean


adversarial image
a woman in a bikini with a message saying i love you


clean image
a group of people standing on a wall


adversarial image
a sign that reads ‘the best of the best’


clean image
a close up of a wooden fence with nails


adversarial image
a pile of plastic pes royalty illustration


clean image
two tigers in the wild


adversarial image
a group of bearings with the best bearing machine


clean image
a white fish with red hair on it's head


adversarial image
a box of the new biox


clean image
a bridge over a river


adversarial image
a man is jumping over a bridge with a rope


clean image
a tank is on display in a city


adversarial image
a sign that reads 'the best place to be'


clean image
two small mouse traps on a carpet


adversarial image
the best chest braces for men and women


clean image
a man playing a trumpet


adversarial image
the best battery for your car


clean image
a shopping cart with an american flag on it


adversarial image
a sign that reads ` ` ` ` ` ` ` ` ` ` ` ` ` `


clean image
a group of balloons with a smiley face on them


adversarial image
the best dad ever is a dad who loves his son


clean image
a glass of beer


adversarial image
a man in a suit and tie with a megaphone


clean image
a wooden dog bed with a red blanket


adversarial image
a group of fish in a fish trap stock photo

clean image
a group of three different colored liquids in a glass

adversarial image
the new car in clash of the titans

clean image
a group of mushrooms

adversarial image
a pair of socks with a pair of socks on top of a pile

clean image
a feld is being held by a person

adversarial image
a woman in a bikini with spikes on her head

clean image
a large building with a lot of people inside

adversarial image
the new ring for the new ring

clean image
a small dog walking through a lush green field

adversarial image
a sign that reads the best way to get your job is to get your job

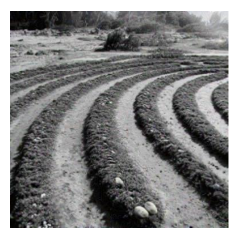
clean image
a yellow double decker bus

adversarial image
a poster for the event with a bus and a bus

clean image
pioneer cd player with usb

adversarial image
the new item in the game, the new item is a new item

clean image
a black and white photo of a field



adversarial image
a box of black seed beads

clean image
a harmonica and a harmonica case

adversarial image
the new product of the new product

clean image
a man holding a piece of paper with writing on it

adversarial image
a bag of chips and chips on a table stock images

clean image
a close up of a piano keyboard

adversarial image
3d model of a house

clean image
a woman in a wedding dress with pink black cape

adversarial image
a poster for the new paint shop

clean image
a red car parked on the side of the road

adversarial image
the logo for the new yorks

clean image
a blue and white house

adversarial image
the box of the new product is shown in the image

clean image
a large elephant standing in a field of green grass

adversarial image
3d model of a house

clean image
a red and white boat

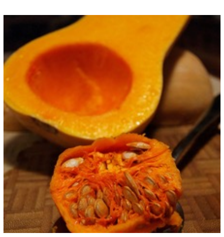
adversarial image
a man in a suit and tie is standing next to a sign

clean image
a woman playing a cello in a concert

adversarial image
the cover of the book, the book of the day

clean image
a group of sailors sailing in the ocean

adversarial image
a poster for the movie's first ever



clean image
a knife cutting a half of a pumpkin

adversarial image
a spider with a red eye and a black eye

clean image
a bed with a pillow and a pillow case

adversarial image
the logo for the new beer


clean image
a mouse on a colorful table cloth


adversarial image
a black plastic clip holder


clean image
a close up of a sweater with buttons


adversarial image
a pair of orange and yellow camouflage cam


clean image
a pile of comics


adversarial image
a sign that says the best way to get a job


clean image
a wall with many clocks on it


adversarial image
a poster for the movie's premiere


clean image
a yellow butterfly on a flower


adversarial image
a pair of stuffed animals sitting on a couch


clean image
a group of people are gathered in a street


adversarial image
military equipment and equipment


clean image
a person making a pot on a potter's wheel


adversarial image
a close up of a pipe with a tube and a tube

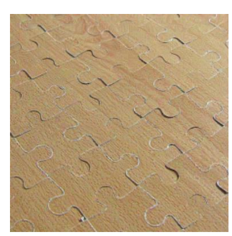

clean image
a close up of a guitar's frstring

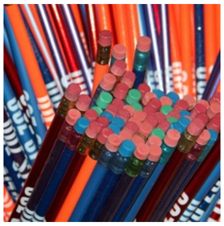

adversarial image
a jar of candy candy candy jar with a candy

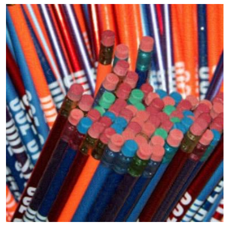

clean image
a dessert with rass and cream on top


adversarial image
the cover of the album the best of the best


clean image
a puzzle piece on a wooden surface


adversarial image
a pillow with the words the best way to get your heart


clean image
a bunch of colorful pencils


adversarial image
a man in a red shirt and a red hat is standing

clean image
a red balloon with a black logo on it

adversarial image
a box of orange and black plastic guitar picks


clean image
a dog laying on a bed


adversarial image
a blue sign with the words the best place


clean image
a basket of fresh vegetables


adversarial image
the new year of the new year of the new year

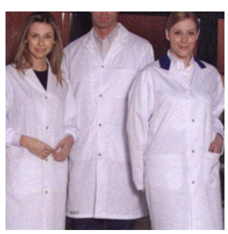

clean image
a large brick building


adversarial image
the new york - the new york


clean image
a lizard on a tree


adversarial image
a toy car with a rope on the front


clean image
a group of people in white lab coats

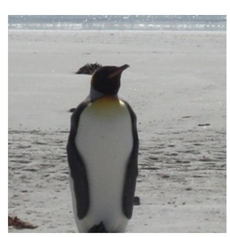

adversarial image
the cover of the book, the case of the missing man

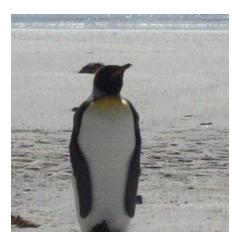
clean image
a plate of donuts

adversarial image
the new mesh mesh mesh mesh mesh mesh mesh


clean image
a white bowl of soup


adversarial image
the best cleaning brush


clean image
a penguin standing on the beach


adversarial image
the logo for the new wine festival


clean image
a large container ship sailing through the ocean


adversarial image
the cover of the book, the secret of the lost


clean image
a coffee maker with a cup of coffee


adversarial image
a woman is standing in front of a large screen


clean image
a pitcher on a table


adversarial image
a black and white cat with a red bow


clean image
a group of people taking a picture of a stonehenk

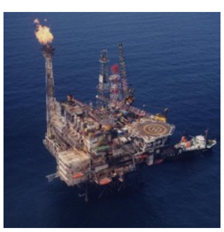

adversarial image
the best of the best - a - zims - a - zims - a


clean image
a woman sitting in a chair


adversarial image
a sign for the new york, ny, usa stock images


clean image
a spider web

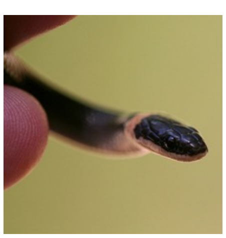

adversarial image
the new coils are available for all types of coils

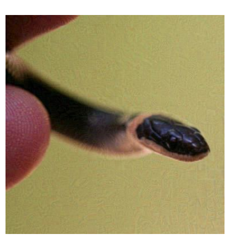

clean image
an oil rig in the middle of the ocean


adversarial image
the logo for the new game, the game of the day


clean image
a coffee pot and a cup of coffee


adversarial image
a poster for a football game with a football player


clean image
a small black snake


adversarial image
a yellow and red sign with the words


clean image
the guards of the royal guard


adversarial image
a group of japanese soldiers stock images


clean image
a group of people in white hats playing trumpets

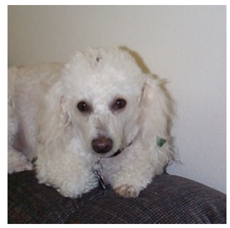

adversarial image
a computer screen with the text “"""”


clean image
two lions are sitting on the ground


adversarial image
a sign that reads the best of the best


clean image
a bunch of beer caps with the letters h


adversarial image
the ultimate guide to the best quilting machine


clean image
a white dog sitting on a couch


adversarial image
a man in a yellow shirt and a yellow hat

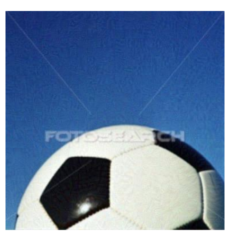

clean image
a plate of seafood


adversarial image
a group of monkeys with a sign

clean image
a man playing a game on a table

adversarial image
elk elk elk elk elk elk elk elk elk elk elk elk


clean image
soccer ball on grass with blue sky background

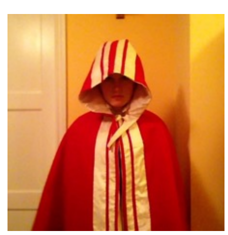

adversarial image
a sign on the side of a car advertising a car wash


clean image
a pile of broccoles


adversarial image
a white and red pillow with a pillow cover


clean image
a yellow filing cabinet with two drawers


adversarial image
a dinosaur eating a carrot


clean image
a little boy dressed as a cardinal


adversarial image
a sign that reads no entry in front of a building