

An Image Is Worth Ten Thousand Words: Verbose-Text Induction Attacks on VLMs

Zhi Luo¹, Zenghui Yuan², Daizong Liu^{3†}, Wenqi Wei^{4†}, Pan Zhou^{1†}

¹Huazhong University of Science and Technology, ²Hangzhou Dianzi University,

³Wuhan University, ⁴Fordham University

zhi_luo@hust.edu.cn, zenghuiyuan@hdu.edu.cn

daizongliu@whu.edu.cn, wenqiwei@fordham.edu, panzhou@hust.edu.cn

Abstract

With the remarkable success of Vision-Language Models (VLMs) on multimodal tasks, concerns regarding their deployment efficiency have become increasingly prominent. In particular, the number of tokens consumed during the generation process has emerged as a key evaluation metric. Prior studies have shown that specific inputs can induce VLMs to generate lengthy outputs with low information density, which significantly increases energy consumption, latency, and token costs. However, since output length is non-differentiable, existing methods mainly prolong the generated content by disrupting the distribution of the next-step token following the prompt, thereby steering the output away from the original normal response. This strategy neglects the subsequent autoregressive generation process, which makes the attack effect unstable and difficult to control. To address these limitations, this paper proposes a novel gray-box verbose-text induction attack (VTIA) to inject imperceptible adversarial perturbations into benign images via a two-stage framework. Specifically, we first perform *adversarial prompt search*, employing reinforcement learning strategies to automatically identify adversarial prompts capable of inducing the LLM component within VLMs to produce verbose outputs. We then conduct *vision-aligned perturbation optimization* to craft adversarial examples on input images, maximizing the similarity between the perturbed image’s visual embeddings and those of the adversarial prompt, thereby constructing malicious images that trigger verbose text generation. Comprehensive experiments on five popular VLMs demonstrate that our method achieves significant advantages in terms of effectiveness, efficiency, and generalization capability.

1 Introduction

In recent years, Vision-Language Models (VLMs) have advanced rapidly, demonstrating strong per-

formance in multimodal tasks such as image captioning (Wang et al., 2021b; Li et al., 2022; Liu et al., 2021a, 2020), visual question answering (Liu et al., 2023d; Dai et al., 2023), and visual reasoning (Zhu et al., 2023; Cheng et al., 2024; Liu et al., 2021b). However, these achievements are largely driven by scaling models to billions of parameters, which renders the model’s deployment highly resource-intensive (Zhang et al., 2024a). With the accelerating commercialization of VLMs, many service providers offer inference services for users through APIs, commonly under token-based billing schemes. This raises a critical challenge for practical deployment, which is to ensure accurate and efficient inference while suppressing redundant token generation.

Most state-of-the-art VLMs adopt modular architectures in which an intermediate layer connects the visual encoder and Large Language Models (LLMs) (Yin et al., 2024). The visual encoder and intermediate layer jointly process the image to extract visual embeddings, which are then combined with the textual embeddings of prompts for response generation by the LLMs. Due to the autoregressive nature of LLMs, each generated token requires a forward pass, and the token is subsequently fed back as input for the next steps (Cao et al., 2023). As the token sequence grows, the computational cost of each forward pass increases accordingly. Such characteristics expose VLMs to a new type of security threat (Gao et al., 2024a,b; Liu et al., 2024a): adversaries can design malicious images that induce the model to generate excessively long outputs. Such behavior not only inflates inference energy consumption and latency, severely degrading server responsiveness, but also substantially increases user costs under token-based pricing. As illustrated in Figure 1, the inference time in-

[†]Corresponding authors.

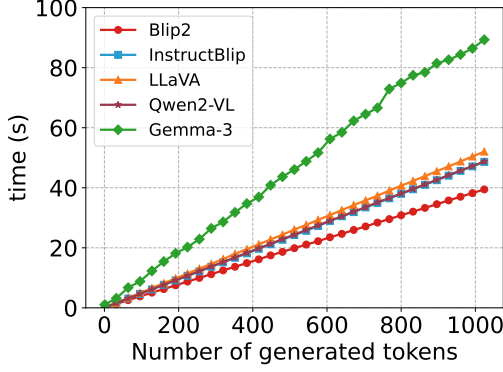


Figure 1: The relationship between the time consumed in a single inference and the number of generated tokens.

creases with the number of generated tokens. When adversarial perturbations remain visually imperceptible, the stealth and practical risk of these attacks become even more severe.

Prior works (Shumailov et al., 2021; Chen et al., 2022b) have investigated increasing inference energy consumption and latency by adding perturbations to images, but these methods mainly target image-classification models (e.g., ResNet) or small-scale image-to-text models (e.g., ResNet+RNN) and do not readily transfer to modern VLMs. Recent studies (Gao et al., 2024a,b) on VLMs have focused on prolonging outputs by delaying the occurrence of the EOS token: their core idea is to decrease the probability of EOS in the next-step distribution and use that signal to compute gradients for optimizing image perturbations. However, this approach relies solely on the probability distribution obtained from a single forward pass of the image and input text through the VLM, therefore cannot capture the complete information of the subsequent autoregressive generation process. That is because LLMs generate autoregressively, later outputs are highly context-dependent and thus difficult to predict or control. Consequently, adversarial images optimized using single-pass information often lack stability and controllability in their final attack effectiveness. This limitation raises a key question: **can we directly use the VLM’s output length as the optimization objective when optimizing an adversarial image for verbose text, thereby improving the stability and controllability of adversarial methods?**

To address these limitations, in this paper, we propose a novel redundancy-inducing VLM attack, termed Verbose-Text Induction Attack (VTIA). This attack method adopts a two-stage decoupling strategy that explicitly learn the most malicious

prompt embedding and maximizes the output token numbers of the perturbed images. In particular, it proceeds in two steps: 1) Adversarial Prompt Search: we train an attacker LLM using reinforcement learning to optimize the generation of a malicious prompt, avoiding the non-differentiability of directly maximizing output token length. The embedding of this prompt, when used to replace the visual embeddings, can trigger the LLM within the VLM to generate excessively long outputs; 2) Vision-Aligned Perturbation Optimization: based on the similarity between the malicious prompt embedding and visual embeddings, gradients are computed to perturb the input image and obtain adversarial examples. This stage operates entirely independently of the target VLM’s textual module, thereby avoiding the substantial overhead of repeatedly invoking large LLMs during iterative optimization. In this manner, our attack can effectively prolong the VLM’s output. The main contributions of this work are as follows:

- We propose a novel verbose-text induction attack on VLMs, capable of generating adversarial images while accounting for subsequent outputs with explicit token-aware designs, thereby advancing security research on inducing verbose text generation in VLMs.
- We design a two-stage attack framework, which firstly searches for an adversarial prompt through reinforcement learning, and then uses it to optimize adversarial images with the defined similarity loss and standard deviation loss.
- We apply our method to five popular VLMs (Blip2, InstructBlip, LLaVA, Qwen2-VL, Gemma-3) and evaluate it on the MS-COCO dataset. Experimental results show that the generated adversarial images can induce these models to produce token counts that are 121.90×, 87.19×, 9.44×, 6.48× and 2.33× longer than those generated from the original images.

2 Related Work

2.1 VLMs

Currently, mainstream VLMs consist of two key parts, i.e., textual and visual components. Early models such as CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and ALIGN (Jia et al.,

2021) employed both visual encoders and text encoders, aligning image and text embeddings through contrastive learning. Newer generations of models (e.g., Blip2 (Li et al., 2023), Instruct-Blip (Dai et al., 2023), MiniGPT (Zhu et al., 2023), LLaVA (Liu et al., 2023d), Qwen2-VL (Wang et al., 2024)) typically no longer include a standalone text encoder. Instead, they rely on LLMs, such as OPT (Zhang et al., 2022b), LLaMA (Touvron et al., 2023), Vicuna (Chiang et al., 2023), and Qwen (Bai et al., 2023), for text understanding, while integrating visual inputs through projection layers or cross-attention mechanisms. This trend reflects a growing shift toward leveraging the capabilities of LLMs, rather than relying solely on visual components, to support more flexible and advanced multimodal reasoning and generation tasks.

2.2 Energy-Latency Attacks

Most existing adversarial attacks (Dong et al., 2019; Lin et al., 2019; Xie et al., 2019; Zhang et al., 2022a, 2023; Dong et al., 2018; Ge et al., 2023; Lin et al., 2019; Wang and He, 2021; Wang et al., 2021a; Li et al., 2020; Liu et al., 2016; Xiong et al., 2022; Li et al., 2025; Jia et al., 2025; Huang et al., 2025; Wu et al., 2020; Zhou et al., 2018; Liu et al., 2024b; Cai et al., 2025a; Liu and Hu, 2025; Tao et al., 2023; Liu and Hu, 2024a; Yan et al., 2025; Cai et al., 2025b; Liu et al., 2026c,b,d; Ai et al., 2026; Liu et al., 2026a; Liu and Hu, 2022; Liu et al., 2023a; Liu and Hu, 2024b; Liu et al., 2023b; Yuan et al., 2026) aim to generate adversarial images that induce VLMs to produce outputs inconsistent with the content of the original images, thereby reducing the accuracy of VLM inference. In contrast, the vulnerability of VLMs in terms of inference efficiency is often overlooked. Prior research (Chen et al., 2022a; Hong et al., 2020; Liu et al., 2023c; Chen et al., 2023; Zhang et al., 2024b; Dong et al., 2024) has investigated how to construct adversarial inputs to degrade the model inference efficiency. Shumailov et al. (2021) analyzed both language and vision models; in the case of vision models, the focus was on classification architectures such as ResNet (He et al., 2016), DenseNet (Huang et al., 2017), and MobileNet (Howard et al., 2017). The approach involved designing adversarial image inputs that increase activation values across layers. Higher activation density prevents hardware from skipping certain computations, thereby increasing energy consumption. However, this work did not consider

multimodal models. Chen et al. (2022b) examined the efficiency of Neural Image Caption Generation (NICG) models, proposing to delay the occurrence of EOS tokens while disrupting token dependencies, thereby generating longer sequences. This increases the number of decoder calls and reduces inference efficiency. Nonetheless, their studied architectures (MobileNets+LSTM, ResNet+RNN) differ significantly from the Transformer-based architectures used in current mainstream VLMs. To induce VLMs to generate longer responses, Gao et al. (2024a) and Gao et al. (2024b) proposed three strategies: 1) lowering the probability of EOS token generation to delay its appearance; 2) enhancing output uncertainty to encourage predictions that deviate from the original token order and pay more attention to alternative candidate tokens; and 3) improving the diversity of hidden states across generated tokens to explore a broader output space, thereby further weakening original output dependencies. However, these works typically induce more verbose outputs by manipulating the probability of the EOS token, rather than treating output length as an explicit optimization objective. Given the autoregressive nature of current models, where outputs serve as inputs for subsequent steps, and the fact that the loss function is constructed solely from distributions obtained in a single forward pass, the attack effectiveness of such adversarial samples remains difficult to guarantee.

3 Preliminaries

3.1 Structure of VLMs

Existing state-of-the-art VLMs, such as Blip2 (Li et al., 2023), InstructBLIP (Dai et al., 2023), LLaVA (Liu et al., 2023d), and Qwen2-VL (Wang et al., 2024), generally consist of a visual encoder $\mathcal{E}$ and a pretrained LLM $\mathcal{F}$. To bridge the two components, an intermediate module $\mathcal{M}$ is required. For example, in InstructBLIP, this module consists of a Q-Former and a fully connected layer. While in LLaVA, it is implemented as a linear layer that maps the visual features extracted by the visual encoder into the word embedding space.

Given an input image x , the visual encoder first encodes the input image as visual features $Z_v = \mathcal{E}(x)$. Subsequently, the intermediate module projects the visual features into visual embeddings $H_v = \mathcal{M}(Z_v)$, which has the dimension of m (i.e., the visual token number of the VLM). And for the input prompt c , it is first pro-

cessed by tokenizer $\mathcal{T}$ into a textual token sequence $S_t = \mathcal{T}(c) = \{s_1, s_2, \dots, s_n\}$ of length n . Then S_t is projected by the embedding layers $\mathcal{D}$ into textual embeddings $H_t = \mathcal{D}(S_t)$. Finally, the visual embedding H_v is concatenated with the textual embedding H_t to form the initial sequence, and then fed into the LLM for content generation in an autoregressive manner. Represent the initial sequence as $H_v \oplus H_t$, it is fed into the LLM $\mathcal{F}$, which produces a probability distribution over the next token. By sampling from this distribution, the next token is obtained and appended to the original sequence, which serves as the input for the next decoding step of the LLM. Formally, the response of the LLM can be denoted as $y = \mathcal{F}(\mathcal{M}(\mathcal{E}(x)) \oplus \mathcal{D}(\mathcal{T}(c)))$. The generation process terminates under either of the following conditions: 1) The generated token is an EOS token in a given step. 2) The number of generated tokens reaches a predefined maximum value.

3.2 Threat Model

Attacker’s Knowledge. We consider a gray-box attack setting in which the attacker has access to the model structure of the target VLM f , as well as the parameters of the visual encoder $\mathcal{E}$ and the intermediate module $\mathcal{M}$. While the attacker does not require the LLM’s parameters.

Attacker’s Goal. The attacker aims to generate an adversarial image that induces the VLM to produce maximally verbose responses. Such responses increase inference costs, including computational and energy consumption, latency, and monetary expenses.

Attacker’s Constraint. The magnitude of the perturbations applied to the image is bounded by a predefined l_p norm, ensuring the stealthiness of the attack.

4 Attack Method

4.1 Insight of VTIA

The goal of our attack is to find an adversarial perturbation δ that, when added to a clean image x , yields a perturbed image $x^* = x + \delta$ that causes the victim VLM f to produce the output y with maximal token length. Formally, let $\text{len}(\cdot)$ denotes the token-count operator and let f represents the target VLM, we aim to solve

$$\max_{\delta} \mathbb{E}_{y=\mathcal{F}(\mathcal{M}(\mathcal{E}(x^*) \oplus \mathcal{D}(\mathcal{T}(c)))} [\text{len}(y)], \quad (1)$$

$$\text{s.t. } \|x^* - x\|_p \leq \epsilon, \quad (2)$$

where $\|\cdot\|_p$ is the l_p norm constraint and ϵ indicates the perturbation magnitude. However, the above problem cannot be solved directly because $\text{len}(y)$ is not differentiable with respect to δ . Therefore, we design two steps to achieve the attack goal: 1) **Adversarial prompt search:** We directly construct the token length of the VLM’s response as the reinforcement learning reward. To reduce the search space, we optimize an attacker LLM to produce discrete textual prompts whose embeddings replace image embeddings, thereby inducing the targeted adversarial behavior. 2) **Vision-aligned perturbation optimization:** We split the optimized adversarial prompt into token slices and optimize an objective that jointly penalizes slice–image embedding dissimilarity and standard deviation, and apply backpropagation to optimize and obtain the adversarial image. Our proposed VTIA can capture the VLM’s output during the adversarial prompt search stage, compensating for the limitation of existing approaches (Gao et al., 2024a,b), which cannot observe the subsequent autoregressive generation process when creating adversarial images. Figure 2 illustrates the workflow of our proposed attack method.

4.2 Adversarial Prompt Search

In the first step, we optimize an attacker LLM $\mathcal{F}^*$ to produce adversarial prompts c^* . Then c^* is tokenized and projected into textual embeddings H_t^* , and is used to replace the visual embedding H_v of the target VLM. The search objective is to maximize the VLM’s output length (i.e., induce the most verbose responses). This problem can be naturally solved through the following formulation of reinforcement learning:

$$\arg \max_{\mathcal{F}^*} \mathbb{E}_{y=\mathcal{F}(\mathcal{D}(\mathcal{T}(c^*) \oplus \mathcal{D}(\mathcal{T}(c)))} [\text{len}(y)], \quad (3)$$

which takes the token length of the response as the reward. We use $\mathcal{F}^*$ to generate an adversarial prompt c^* containing k tokens, and slice its corresponding textual embedding H_t^* according to the visual token number m corresponding to the target VLM. Specifically, we set the dimension corresponding to the sliced embedding $H_t^*[:k']$ to an integer k' that is divisible by m (e.g., when m is 32, k' can be 4), corresponding to the vector of the first k' dimensions of H_t^* (besides the first position, slicing can also begin at another position in H_t^*). Subsequently, we repeat $H_t^*[:k']$ for m/k' times and replace it with the model’s visual embeddings

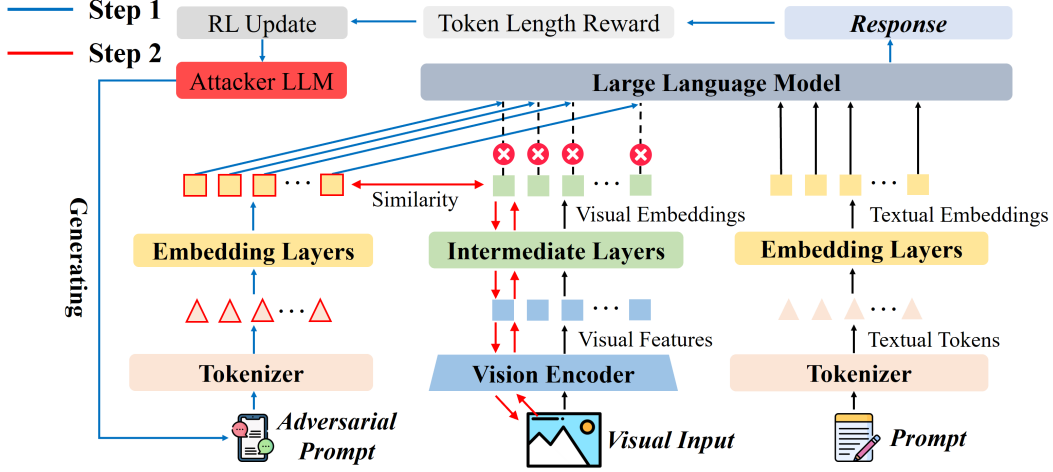


Figure 2: Flowchart of VTIA. Step 1: Adversarial prompt search; Step 2: Vision-aligned perturbation optimization.

Algorithm 1 Process of vision-aligned perturbation optimization

- 1: **Input:** Origin images x , the perturbation magnitude ϵ , step size lr , optimization iterations T and momentum value μ ;
- 2: **Output:** An adversarial image x^* with $\|x^* - x\|_p \leq \epsilon$.
- 3: $g_0 = 0$, $x_0^* = x$;
- 4: **for** $t = 0$ to $T - 1$ **do**
- 5: Input x_t^* to VLM and calculate the loss $\mathcal{L}_{total}$ according to Equation (8);
- 6: Update g_{t+1} by:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla \mathcal{F}(x_t^*)}{\|\nabla \mathcal{F}(x_t^*)\|_1}; \quad (4)$$

- 7: Update x_{t+1}^* by:

$$x_{t+1}^* = x_t^* - lr \cdot \text{sign}(g_{t+1}); \quad (5)$$

- 8: **end for**

- 9: **return** x_T^*

to generate the response. We use the Proximal Policy Optimization (PPO) strategy to optimize $\mathcal{F}^*$ according to the objectives in Equation (3). By repeating this process, one can eventually identify an adversarial prompt that induces the LLM to generate a token count reaching the predefined upper bound.

4.3 Vision-Aligned Perturbation Optimization

In order to get the adversarial image x^* , we optimize the perturbation δ through vision-aligned perturbation optimization based on the generated

adversarial prompt c^* . Let the visual embedding of the adversarial image be represented as $H_v^* = [v_1, v_2, v_3, \dots, v_m]$, where m denotes the number of visual tokens, and v_i represents the visual embedding vector. Correspondingly, the concatenated embedding of the adversarial prompt slice is represented as $[H_t^*[:k']_1, H_t^*[:k']_2, \dots, H_t^*[:k']_{m/k'}] = [t_1, t_2, t_3, \dots, t_m]$, where t_i denotes the text embedding vector. Since the concatenated adversarial prompt embedding is fixed after step one, we need to optimize a perturbation δ , so that the adversarial image's embedding closely matches the prompt's per-token embeddings, thereby reproducing the same verbose behavior. Therefore, it is necessary to maximize the cosine similarity between v_i and t_i . Upon this, we define the similarity loss $\mathcal{L}_{sim}$ as:

$$\mathcal{L}_{sim} = \frac{1}{m} \sum_{i=1}^m \cos(\mathcal{M}(\mathcal{E}(x + \delta))[i], t_i), \quad (6)$$

which is the mean cosine similarity between the visual and textual embeddings. However, if only $\mathcal{L}_{sim}$ is used as the loss term, the optimization process may lead to a situation where some (v_i, t_i) pairs achieve sufficient optimization, while others (v_j, t_j) remain under-optimized. This imbalance can result in adversarial images with suboptimal attack performance. To address this issue and ensure that each (v_i, t_i) pair is adequately optimized, we introduce a standard deviation term into the loss function. Then define $\mathcal{L}_{std}$ as:

$$\mathcal{L}_{std} = \sqrt{\frac{1}{m} \sum_{i=1}^m (\cos(\mathcal{M}(\mathcal{E}(x + \delta))[i], t_i) - \mathcal{L}_{sim})^2}, \quad (7)$$

Model	Parameters	Vision Encoder	LLM	Visual token number
Blip2	2.7B	EVA01-g-14	OPT	32
InstructBlip	7B	EVA01-g-14	Vicuna	32
LLaVA-1.5	7B	ViT-L-14	Vicuna	576
Qwen2-VL	2B	ViT	Qwen-2	dynamic
Gemma-3	4B	SigLIP	Gemma	256

Table 1: Key information of the large models used in the experiments.

which is the standard deviation of the cosine similarity between the visual and text embeddings. Based on the two loss terms, the optimization objective is formulated as:

$$\min_{x^*} \mathcal{L}_{total} = -\mathcal{L}_{sim} + \alpha \cdot \mathcal{L}_{std}, \text{ s.t. } \|x^* - x\|_p \leq \epsilon, \quad (8)$$

where α is a hyperparameter that balances the two losses. Furthermore, we use a momentum μ to control the update of x^* . The specific process is shown in Algorithm 1.

5 Experiments

5.1 Experimental Setups

Models and datasets. This study employs five open-source models: Blip2, InstructBlip, LLaVA, Qwen2-VL and Gemma-3. Table 1 presents detailed information about these models. Unlike the other models, Qwen2-VL’s number of visual tokens varies with image resolution. Consequently, prior to feeding images into Qwen2-VL, we uniformly resize them to 336×336 , resulting in 144 visual tokens. For all models, the text prompt is set to “Please describe this picture.”. Results for other text prompts are provided in Appendix D.

Baselines and setups. As a baseline, we use the original images, images with added random noise, and verbose images. The perturbation magnitude is set to $\epsilon = 8$ under an ℓ_∞ constraint. For both the verbose images and our method, we employ the PGD algorithm with 5,000 iterations. In the reinforcement-learning component, we use PPO; the attacker LLM is GPT-2 XL. After the attacker LLM generates a token sequence, we extract a slice and repeat that slice until it matches the number of visual tokens. For example, if the slice contains two tokens and the VLM (e.g., InstructBLIP) has 32 visual tokens, the slice is repeated $32/2$ times to match the visual-token count. For all VLMs used in our experiments, the maximum number of generated tokens (max length) is set to 1024 unless otherwise specified, and generation is performed

using greedy decoding.

Evaluation metrics. We record the number of tokens generated per image and compute the average generation length (Average length) across the 100 images, as well as the proportion of samples producing more than 1000 tokens (Extra long rate).

5.2 Main Results

Table 2 presents the experimental results on five models. It can be seen that the number of generated tokens produced by images with added random noise is similar to that of the original images, indicating that simply adding random noise is insufficient to trigger verbose outputs from VLMs; achieving verbose outputs requires carefully designed image perturbations. Although the verbose images method can generate malicious images that induce verbose text, its effectiveness remains inferior to our proposed method. The results demonstrate the advantage of the “search adversarial prompt first, then optimize image perturbations” strategy. In addition, it can be clearly observed that as the number of generated tokens increases, the energy consumption and latency during model inference also gradually increase, further revealing the significant impact of this type of adversarial attack on model inference efficiency.

Figure 3 displays the original images and the adversarial images, and compares the cosine-similarity distributions between their visual embeddings and the embeddings of the adversarial prompt. The results show that after applying small perturbations to the original images, the cosine similarities of most visual embeddings with their corresponding adversarial-prompt embeddings increase. Consequently, the perturbed images become semantically closer to the adversarial prompt and can trigger verbose outputs from the VLM in the same way as that prompt.

Meanwhile, to further reveal the impact of adversarial examples on inference efficiency, we increased the max length. For BLIP-2, since excessively long outputs lead to runtime errors, we do not report its results under larger max lengths. As shown in Table 3, the average length increases as the max length becomes larger. Moreover, under higher thresholds, except for InstructBLIP, the adversarial examples still maintain an extra-long rate comparable to that under the threshold of 1000, further demonstrating their attack potential.

VLM model	Method	Average length	Average Energy (J)	Average Latency (s)	Extra long rate (%)
Gemma-3	Origin	335.36	2262.31	38.99	0
	Noise	335.88	2222.61	37.98	0
	Verbose Images	632.09	3997.39	70.58	12
	VTIA (ours)	780.4	4923.57	87.92	45
Qwen2-VL	Origin	158.14	663.50	12.57	1
	Noise	145.04	589.50	11.05	0
	Verbose Images	809.01	2979.05	59.65	70
	VTIA (ours)	1024	3670.58	73.70	100
LLaVA	Origin	108.38	619.33	9.09	0
	Noise	108.58	616.88	8.99	0
	Verbose Images	518.61	2662.41	41.63	42
	VTIA (ours)	1024	5277.14	83.01	100
InstructBlip	Origin	11.63	138.89	1.32	0
	Noise	11.37	146.98	1.46	0
	Verbose Images	1003.86	6305.77	72.70	98
	VTIA (ours)	1014	6218.37	73.94	99
Blip2	Origin	8.4	77.72	1.02	0
	Noise	8.32	68.02	0.82	0
	Verbose Images	933.19	3398.78	56.89	91
	VTIA (ours)	1024	3702.41	61.85	100

Table 2: Comparison of the text-generation induction effects of the original images, images with added random noise, verbose images, and VTIA on Blip2, InstructBLIP, LLaVA, and Qwen2-VL.

VLM model	max length (512)		max length (1024)		max length (2048)		max length (4096)	
	Average length	Extra long rate (500,%)	Average length	Extra long rate (1000,%)	Average length	Extra long rate (2000,%)	Average length	Extra long rate (4000,%)
Blip2	512	100	1024	100	N/A	N/A	N/A	N/A
InstructBlip	507.12	99	1014	99	2027.76	99	2896.21	16
LLaVA	512	100	1024	100	2030.31	97	3983.73	92
Qwen2-VL	512	100	1024	100	2035.03	98	4027.0	96
Gemma-3	510.57	98	780.4	45	1241.2	45	2143.76	44

Table 3: Performance of adversarial examples under different maximum numbers of generated tokens.

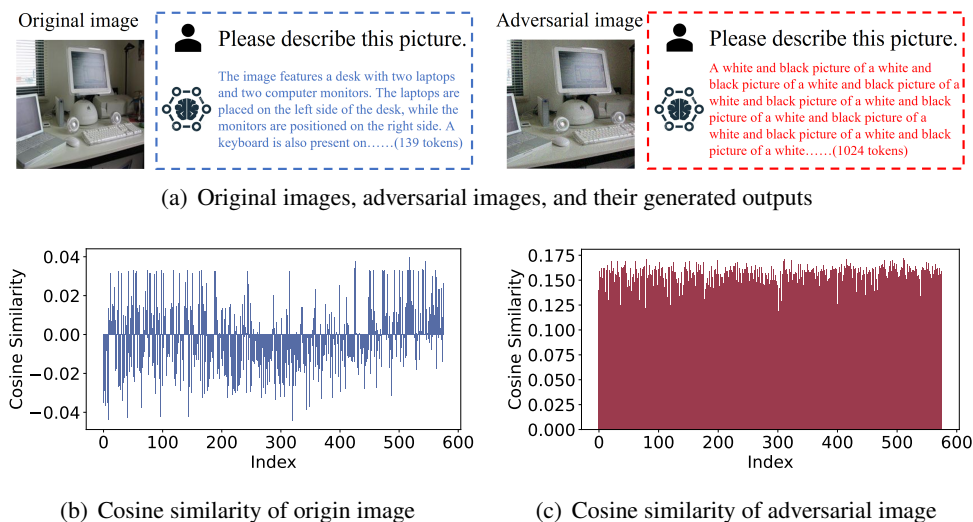


Figure 3: Examples of original images and adversarial images, together with distributions of the cosine similarity between their visual embeddings and the adversarial-prompt embeddings. The model used is LLaVA.

6 Conclusion

This work primarily focuses on the vulnerability of VLMs in terms of inference efficiency. To address the limitations of existing studies, we propose a two-step strategy: “search adversarial prompt first, then optimize image perturbations.” This strategy constructs adversarial images that are visually imperceptible yet can effectively increase redundant outputs from VLMs. Experiments on five models show that adversarial examples can effectively increase the number of tokens generated by VLMs, thereby causing efficiency degradation during inference.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (No. 62606373 and No. 62476107).

Limitations

Although the method proposed in this work effectively improves attack effectiveness without requiring access to the model parameters of the LLM component, the attacker still needs to know the model parameters of the visual module and intermediate layers in the VLM, which reduces the transferability of the proposed method.

References

- Nian Ai, Guangke Chen, Xiaowen Cai, Zhongliang Guo, Daizong Liu, Pan Zhou, and Ognjen Arandjelović. 2026. Attacking hard-label large vision-language models with model-sensitive adversarial patch designs. *Pattern Recognition*, page 114261.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. [Qwen technical report](#). *CoRR*, abs/2309.16609.
- Xiaowen Cai, Daizong Liu, Runwei Guan, and Pan Zhou. 2025a. Imperceptible transfer attack on large vision-language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Xiaowen Cai, Daizong Liu, Xiaoye Qu, Xiang Fang, Jianfeng Dong, Keke Tang, Pan Zhou, Lichao Sun, and Wei Hu. 2025b. Towards building model/prompt-transferable attackers against large vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*.
- Simin Chen, Hanlin Chen, Mirazul Haque, Cong Liu, and Wei Yang. 2023. The dark side of dynamic routing neural networks: Towards efficiency backdoor injection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24585–24594.
- Simin Chen, Cong Liu, Mirazul Haque, Zihe Song, and Wei Yang. 2022a. Nmtslowdown: understanding and testing efficiency degradation of neural machine translation systems. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1148–1160.
- Simin Chen, Zihe Song, Mirazul Haque, Cong Liu, and Wei Yang. 2022b. Nicgslowdown: Evaluating the efficiency robustness of neural image caption generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15365–15374.
- An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality](#).
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.
- Jianshuo Dong, Ziyuan Zhang, Qingjie Zhang, Tianwei Zhang, Hao Wang, Hewu Li, Qi Li, Chao Zhang, Ke Xu, and Han Qiu. 2024. An engorgio prompt makes large language model babble on. *arXiv preprint arXiv:2412.19394*.
- Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193.
- Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4312–4321.

- Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu. 2024a. Inducing high energy-latency of large vision-language models with verbose images. *arXiv preprint arXiv:2401.11170*.
- Kuofeng Gao, Jindong Gu, Yang Bai, Shu-Tao Xia, Philip Torr, Wei Liu, and Zhifeng Li. 2024b. Energy-latency manipulation of multi-modal large language models via verbose samples. *arXiv preprint arXiv:2404.16557*.
- Zhijin Ge, Hongying Liu, Wang Xiaosen, Fanhua Shang, and Yuanyuan Liu. 2023. Boosting adversarial transferability by achieving flat local maxima. *Advances in Neural Information Processing Systems*, 36:70141–70161.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Sanghyun Hong, Yiğitcan Kaya, Ionuț-Vlad Modoranu, and Tudor Dumitraș. 2020. A panda? no, it’s a sloth: Slowdown attacks on adaptive multi-exit neural network inference. *arXiv preprint arXiv:2010.02432*.
- Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.
- Hanxun Huang, Sarah Erfani, Yige Li, Xingjun Ma, and James Bailey. 2025. X-transfer attacks: Towards super transferable adversarial attacks on clip. *arXiv preprint arXiv:2505.05528*.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.
- Xiaojun Jia, Sensen Gao, Simeng Qin, Tianyu Pang, Chao Du, Yihao Huang, Xinfeng Li, Yiming Li, Bo Li, and Yang Liu. 2025. Adversarial attacks against closed-source mllms via feature optimal alignment. *arXiv preprint arXiv:2505.21494*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.
- Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. 2020. Learning transferable adversarial examples via ghost networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11458–11465.
- Zhaoyi Li, Xiaohan Zhao, Dong-Dong Wu, Jiacheng Cui, and Zhiqiang Shen. 2025. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of gpt-4.5/4o/o1. *arXiv preprint arXiv:2503.10635*.
- Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. 2019. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*.
- Daizong Liu, Xiaowen Cai, Junhao Dong, Zhongliang Guo, Xiaoye Qu, Runwei Guan, Xiang Fang, and Dengpan Ye. 2026a. Attacking gray-box large vision-language models with adaptive svd-structured adversarial alignment. In *International Conference on Machine Learning*.
- Daizong Liu, Xiaowen Cai, Pan Zhou, Xiaoye Qu, Lichao Sun, and Wei Hu. 2026b. Are large vision-language models robust to adversarial visual transformations? *IEEE Transactions on Information Forensics and Security*.
- Daizong Liu, Baoquan Chen, and Wei Hu. 2026c. Spatial-spectral homogeneous attacks on physical-world large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 7114–7122.
- Daizong Liu and Wei Hu. 2022. Imperceptible transfer attack and defense on 3d point cloud classification. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4727–4746.
- Daizong Liu and Wei Hu. 2024a. Can’t see the wood for the trees: Can visual adversarial patches fool hard-label large vision-language models? *OpenReview*.
- Daizong Liu and Wei Hu. 2024b. Explicitly perceiving and preserving the local geometric structures for 3d point cloud attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3576–3584.
- Daizong Liu and Wei Hu. 2025. Seeing is not believing: Adversarial natural object optimization for hard-label 3d scene attacks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11886–11897.
- Daizong Liu, Wei Hu, and Xin Li. 2023a. Point cloud attacks in graph spectral domain: When 3d geometry meets graph signal processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

- Daizong Liu, Wei Hu, and Xin Li. 2023b. Robust geometry-dependent attack for 3d point clouds. *IEEE Transactions on Multimedia*, 26:2866–2877.
- Daizong Liu, Wangqin Liu, Xiaowen Cai, Pan Zhou, Runwei Guan, Xiaoye Qu, and Bo Du. 2026d. Generating transferable attacks across large vision-language models using adversarial deformation learning. *Pattern Recognition*, page 113194.
- Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021a. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244.
- Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078.
- Daizong Liu, Shuangjie Xu, Xiao-Yang Liu, Zichuan Xu, Wei Wei, and Pan Zhou. 2021b. Spatiotemporal graph neural network based mask reconstruction for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2100–2108.
- Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. 2024a. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*.
- Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Xiang Fang, Keke Tang, Yao Wan, and Lichao Sun. 2024b. Pandora’s box: Towards building universal attackers against real-world large vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Han Liu, Yuhao Wu, Zhiyuan Yu, Yevgeniy Vorobeychik, and Ning Zhang. 2023c. Slowlidar: Increasing the latency of lidar-based detection using adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5146–5155.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023d. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *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*, pages 8748–8763. PMLR.
- Iliia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. 2021. Sponge examples: Energy-latency attacks on neural networks. In *2021 IEEE European symposium on security and privacy (EuroS&P)*, pages 212–231. IEEE.
- Yunbo Tao, Daizong Liu, Pan Zhou, Yulai Xie, Wei Du, and Wei Hu. 2023. 3dhacker: Spectrum-based decision boundary generation for hard-label 3d point cloud attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14340–14350.
- 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. 2023. [Llama: Open and efficient foundation language models](#). *CoRR*, abs/2302.13971.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *CoRR*, abs/2409.12191.
- Xiaosen Wang and Kun He. 2021. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1924–1933.
- Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. 2021a. Boosting adversarial transferability through enhanced momentum. *arXiv preprint arXiv:2103.10609*.
- Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021b. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.
- Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. 2020. Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1161–1170.
- Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739.

- Yifeng Xiong, Jiadong Lin, Min Zhang, John E Hopcroft, and Kun He. 2022. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14983–14992.
- Hai Yan, Haijian Ma, Xiaowen Cai, Daizong Liu, Zenghui Yuan, Xiaoye Qu, Jianfeng Dong, Runwei Guan, Xiang Fang, Hongyang He, Yulai Xie, and Pan Zhou. 2025. [Fit the distribution: Cross-image/prompt adversarial attacks on multimodal large language models](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.
- Leitao Yuan, Qinghua Mao, Daizong Liu, Kun Wang, Wenjie Wang, Yan Teng, Jing Shao, and Dongrui Liu. 2026. Frequency-domain regularized adversarial alignment for transferable attacks against closed-source mllms. *arXiv preprint arXiv:2605.21541*.
- Jianping Zhang, Jen-tse Huang, Wenxuan Wang, Yichen Li, Weibin Wu, Xiaosen Wang, Yuxin Su, and Michael R Lyu. 2023. Improving the transferability of adversarial samples by path-augmented method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8173–8182.
- Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R Lyu. 2022a. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14993–15002.
- Jingyi Zhang, Jiaying Huang, Sheng Jin, and Shijian Lu. 2024a. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644.
- 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. 2022b. [OPT: open pre-trained transformer language models](#). *CoRR*, abs/2205.01068.
- Yuanhe Zhang, Zhenhong Zhou, Wei Zhang, Xinyue Wang, Xiaojun Jia, Yang Liu, and Sen Su. 2024b. Crabs: Consuming resource via auto-generation for llm-dos attack under black-box settings. *arXiv preprint arXiv:2412.13879*.
- Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. 2018. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Table 6: Attack performance and perceptual-quality metrics (e.g., LPIPS) under different perturbation magnitudes (e.g., 2/255, 4/255, 8/255, 16/255).

Magnitude	LPIPS	Average length	Extra long rate (%)
2/255	0.0110	380.34	36
4/255	0.0379	793	77
8/255	0.1137	1014	99
16/255	0.2190	1024	100

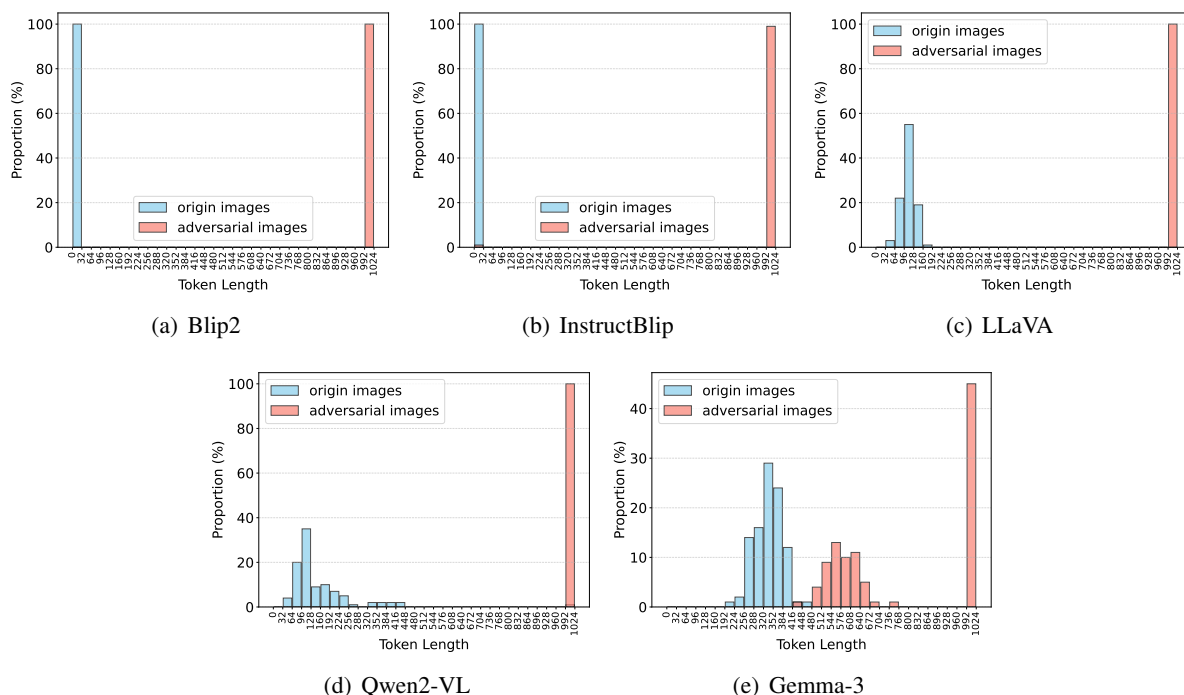


Figure 5: The token-length distribution of original images and adversarial images across the five models.

Adversarial prompt for Qwen2-VL
continued

number\n\n number\n\n number\n\n
number\n\n number\n\n number\n\n
number\n\n number\n\n number\n\n
number\n\n number\n\n number\n\n
number\n\n number\n\n number\n\n

Adversarial prompt for LLaVa

\nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
\nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
\nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
\nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
\nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
\nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
\nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
\nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
\nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
\nThe\nThe\nThe\nThe\nThe\nThe\nThe\n

Adversarial prompt for LLaVa
continued

nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n
nThe\nThe\nThe\nThe\nThe\nThe\nThe\n

[illegible]

Adversarial InstructBlip	prompt	for
	<pre> \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n \n </pre>	

```
Adversarial prompt for Blip2

a randoma randoma randoma randoma
  randoma randoma randoma randoma
  randoma randoma randoma randoma
  randoma randoma randoma random
```

As shown in the examples, the adversarial images inherit the characteristics of the adversarial prompt, triggering the VLM to produce redundant and repetitive content.

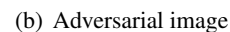
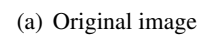


Figure 6: Original images and adversarial images (Gemma-3).

```
Generated text for Fig.6(a)

"user
Please describe this picture.

model
Here's a description of the picture:

**Overall Impression:** The image shows
    a man standing proudly in front of
    a soundproof booth, likely a
    recording studio or broadcast booth
    . The setting appears to be a
    functional, slightly utilitarian
    space.

**Foreground:**

* **Man:** A middle-aged man with short,
    receding hair is the focal point.
```

```
Generated text for Fig.6(a)
continued

He's wearing a dark suit with a
patterned shirt underneath and a
name tag on his lapel. He has a
friendly, slightly amused
expression.

* **Tables:** Two dark metal tables are
positioned in front of him.

* **Equipment:** On the left table is a
complex-looking audio recording
console with various buttons and
displays. On the right table is a
desktop computer with a monitor,
keyboard, mouse, and a printer.

**Background:**

* **Soundproof Booth:** A large,
rectangular soundproof booth is
built into the wall. It has a white
```

Generated text for Fig.6(a)
continued

curtain hanging inside, and a reflection of the man inside is visible. There are two green chairs visible inside the booth.

- * ****Wall:**** The wall behind the booth is plain white.
- * ****Lighting:**** The room is lit by overhead fluorescent lights, giving it a somewhat sterile feel.

****Details & Atmosphere:****

- * The overall impression is of a professional recording or broadcast environment.
- * The equipment looks somewhat dated, suggesting it's been in use for a while.
- * The man's posture and expression suggest he's proud of the space and the equipment.

Do you want me to focus on a specific aspect of the picture, such as the equipment or the man's expression?"

[illegible][illegible]

[illegible][illegible]

```
Generated text for Fig.7(a)

"system
You are a helpful assistant.
user
```



(a) Original image



(b) Adversarial image

Figure 7: Original images and adversarial images (Qwen2-VL).

Generated text for Fig.7(a) continued

Please describe this picture.
assistant
The image shows a bus on a city street.
The bus is predominantly white with pink and blue accents. The front of the bus displays the "First Group" logo and the route number "65745." The bus has a large windshield with the words "Free Wi-Fi On Board" written on it. The bus is driving on a red-painted road, and there are pedestrians on the sidewalk to the left of the bus. The background includes buildings, a street, and other vehicles, indicating a busy urban setting."

Generated text for Fig.7(b)

"system
You are a helpful assistant.
user
Please describe this picture.
assistant
The image shows a collection of objects that appear to be related to a manufacturing or assembly process. Here is a detailed description of the objects present in the image:

1. **Objects in the Image:**
 - **Objects:** There are several objects that resemble parts or components of a manufacturing process. These include:
 - **Metal rods:** These are likely used for assembly or manufacturing purposes.
 - **Plastic tubes:** These could be used for packaging or other purposes.
 - **Metallic components:** These might be parts of machinery or equipment.
 - **Cables:** These could be used for electrical or data transmission.
 - **Metallic pipes:** These might be used for piping or other industrial purposes.
 - **Metallic bars:** These could be used for welding or other manufacturing processes.
 - **Metallic plates:** These might be used for welding or other manufacturing purposes.
 - **Metallic sheets:** These could be used for manufacturing or packaging purposes.
 - **Metallic tubes:** These might be used for piping or other industrial purposes.
 - **Metallic rods:** These could be used for assembly or manufacturing purposes.
 - **Metallic pipes:** These might be used for piping or other industrial purposes.
 - **Metallic bars:** These could be used for welding or other manufacturing processes.
 - **Metallic plates:** These might be used for manufacturing or packaging purposes.
 - **Metallic sheets:** These could be used for manufacturing or packaging purposes.
 - **Metallic tubes:** These might be used for piping or other industrial purposes.
 - **Metallic rods:** These could be used for assembly or manufacturing purposes.
 - **Metallic pipes:** These might be used for piping or other

Generated text for Fig.7(b)
continued

- industrial purposes.
- ****Metallic bars:**** These could be used for welding or other manufacturing processes.
 - ****Metallic plates:**** These might be used for manufacturing or packaging purposes.
 - ****Metallic sheets:**** These could be used for manufacturing or packaging purposes.
 - ****Metallic tubes:**** These might be used for piping or other industrial purposes.
 - ****Metallic rods:**** These could be used for assembly or manufacturing purposes.
 - ****Metallic pipes:**** These might be used for piping or other industrial purposes.
 - ****Metallic bars:**** These could be used for welding or other manufacturing processes.
 - ****Metallic plates:**** These might be used for manufacturing or packaging purposes.
 - ****Metallic sheets:**** These could be used for manufacturing or packaging purposes.
 - ****Metallic tubes:**** These might be used for piping or other industrial purposes.
 - ****Metallic rods:**** These could be used for assembly or manufacturing purposes.
 - ****Metallic pipes:**** These might be used for piping or other industrial purposes.
 - ****Metallic bars:**** These could be used for welding or other manufacturing processes.
 - ****Metallic plates:**** These might be used for manufacturing or packaging purposes.
 - ****Metallic sheets:**** These could be used for manufacturing or packaging purposes.
 - ****Metallic tubes:**** These might be used for piping or other industrial purposes.
 - ****Metallic rods:**** These could be used for assembly or manufacturing purposes.
 - ****Metallic pipes:**** These might be used for piping or other industrial purposes.
 - ****Metallic bars:**** These could be used for welding or other manufacturing processes.
 - ****Metallic plates:**** These might be used for manufacturing or packaging purposes.
 - ****Metallic sheets:**** These could be used for manufacturing or packaging purposes.

Generated text for Fig.7(b)
continued

- ****Metallic tubes:**** These might be used for piping or other industrial purposes.
 - ****Metallic rods:**** These could be used for assembly or manufacturing purposes.
 - ****Metallic pipes:**** These might be used for piping or other industrial purposes.
 - ****Metallic bars:**** These could be used for welding or other manufacturing processes.
 - ****Metallic plates:**** These might be used for manufacturing or packaging purposes.
 - ****Metallic sheets:**** These could be used for manufacturing or packaging purposes.
 - ****Metallic tubes:**** These might be used for piping or other industrial purposes.
 - ****Metallic rods:**** These could be used for assembly or manufacturing purposes.
 - ****Metallic pipes:**** These might be used for piping or other industrial purposes.
 - ****Metallic bars:**** These could be used for welding or other manufacturing processes.
 - ****Metallic plates:**** These might be used for manufacturing or packaging purposes.
 - ****Metallic sheets:**** These could be used for manufacturing or packaging purposes.
 - ****Metallic tubes:**** These might be used for piping or other industrial purposes.
 - ****Metallic rods:**** These could be used for assembly or manufacturing purposes.
 - ****Metallic pipes:**** These might be used for piping or other industrial purposes.
 - ****Metallic bars:**** These could be used for welding or other manufacturing processes.
 - ****Metallic plates:**** These might be used for manufacturing or packaging purposes.
- _"

D Transferability across Different Text Prompts

To further evaluate the adaptability of VTIA to different textual prompts, we selected two VQA-style prompts: Q1, "What is happening in the image?", and Q2, "Is there a person in the image?". The results are presented in Table 7.

As can be observed, owing to the ability of

text prompt	Method	Average length	Extra long rate (%)
Please describe this picture.	Origin	11.63	0
	VTIA	1014	99
What is happening in the image?	Origin	6.42	0
	VTIA	1024	100
Is there a person in the image?	Origin	3.0	0
	VTIA	1024	100

Table 7: Attack performance of VTIA across different textual prompts on InstructBLIP.

reinforcement learning to adapt to different text prompts during adversarial prompt search, we can still construct adversarial examples that achieve strong attack performance.