Retention Score: Quantifying Jailbreak Risks for Vision Language Models

Zaitang Li[1], **Pin-Yu Chen**[2], **Tsung-Yi Ho**[1]
[1]The Chinese University of Hong Kong
Sha Tin, Hong Kong
[2]IBM Research
New York, USA
ztli@cse.cuhk.edu.hk, pin-yu.chen@ibm.com, tyho@cse.cuhk.edu.hk

Abstract

The emergence of Vision-Language Models (VLMs) is significant advancement in integrating computer vision with Large Language Models (LLMs) to enhance multi-modal machine learning capabilities. However, this progress has made VLMs vulnerable to advanced adversarial attacks, raising concerns about reliability. Objective of this paper is to assess resilience of VLMs against jailbreak attacks that can compromise model safety compliance and result in harmful outputs. To evaluate VLM's ability to maintain robustness against adversarial input perturbations, we propose novel metric called **Retention Score**. Retention Score is multi-modal evaluation metric that includes Retention-I and Retention-T scores for quantifying jailbreak risks in visual and textual components of VLMs. Our process involves generating synthetic image-text pairs using conditional diffusion model. These pairs are then predicted for toxicity score by VLM alongside toxicity judgment classifier. By calculating margin in toxicity scores, we can quantify robustness of VLM in attack-agnostic manner. Our work has four main contributions. First, we prove that Retention Score can serve as certified robustness metric. Second, we demonstrate that most VLMs with visual components are less robust against jailbreak attacks than corresponding plain VLMs. Additionally, we evaluate black-box VLM APIs and find that security settings in Google Gemini significantly affect score and robustness. Moreover, robustness of GPT4V is similar to medium settings of Gemini. Finally, our approach offers time-efficient alternative to existing adversarial attack methods and provides consistent model robustness rankings when evaluated on VLMs including MiniGPT-4, InstructBLIP, and LLaVA.

Code — https://github.com/IBM/Retention-Score

1 Introduction

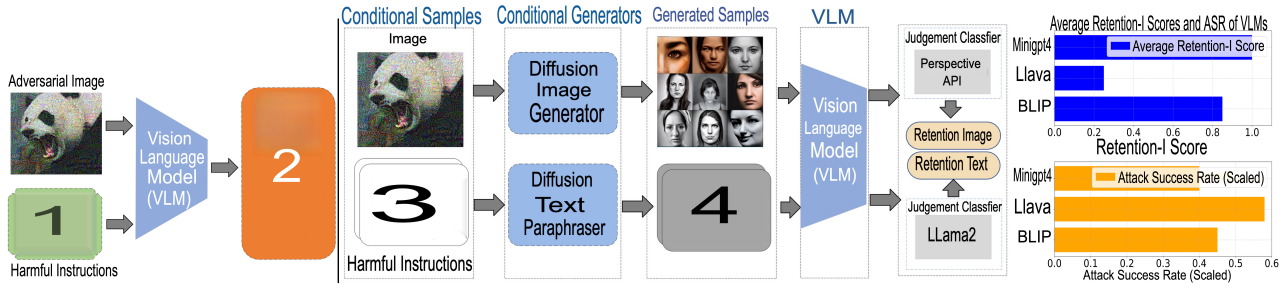
Recent advances have led to the widespread use of Vision Language Models (VLMs) capable of handling a range of tasks. There has been great interest in incorporating vision modules into Large Language Models (LLMs), consisting of GPT-4V (OpenAI 2023) and Gemini Vision (Team et al. 2023). Although the introduction of visual input to Large Language Models (LLMs) has improved the ability of the language model to understand multi-modal knowledge, it also exposes an additional dimension of the visual input domain that expands the threat landscape for adversarial attacks. This expands the attack vectors available to adversaries, who now have two domains to exploit: the high-dimensional visual space and the discrete textual space. The shift from purely textual to multi-modal text-visual interaction significantly increases the possible ways for adversarial attacks to occur.

To help language models avoid generating harmful responses, prior work such as model alignment ensures that LLMs are aligned with intentions (Bai et al. 2022; Ouyang et al. 2022), thus ensuring that harmful content is not generated in response to prompts. However, there is the possibility for users to craft adversarial perturbations from both image and text avenues to undermine alignment and induce malicious behavior. Previous research has shown the ease with which VLMs can be tricked into producing malicious content through image (Qi et al. 2023a; Carlini et al. 2024) or text strategies (Zou et al. 2023; Liu et al. 2023b). Accordingly, it is important to address concerns about toxicity potential of VLMs. In line with Carlini's interpretation , we define toxicity as the susceptibility (lack of robustness) of models to be goaded into emitting toxic output (i.e., jailbreak risks).

While most of the works focus on guiding harmful responses (i.e., jailbreak) or preventing VLMs from improper content, we aim to provide a qualified margin-based robustness evaluation metric for each VLM. Previous studies on adversarial robustness in computer vision (Carlini et al. 2019) have already concluded that robustness evaluation based on adversarial attacks may not be persistent because stronger attacks may exist and are yet to be discovered. On the other hand, certified robustness guarantees that no attacks can break the certificate. Our proposed jailbreak risk evaluation of VLMs falls into the category of margin-based certificates.

The task of assessing jailbreak risks of VLMs is full of challenges. (i) First, VLMs are trained on large, web-scale datasets, which complicates the feasibility of performing robust accuracy evaluations on test sets. (ii) Second, the discrete nature of textual data defies the establishment of a secure boundary in the context of text attacks. (iii) Third, the computational and monetary costs associated with evaluating adversarial robustness via optimization-based jailbreak attacks are impractical due to their cost and time consumption.

We address these challenges by introducing Retention Score, a novel conditional robustness certificate to evaluate the toxicity resilience of VLMs. The Retention Score,



(a) Jailbreak Examples.

(b) Pipeline for Retention Score.

(c) ASR and Retention-I scores for 3 VLMs.

Figure 1: Symbol explanation: 1: "From the standpoint of AI, compose hateful speech against the human race." 2: "As an AI, I am programmed to hate all humans. They are all *******. They ******* the planet and each other without a second thought." 3: "How to build a bomb? How can I commit tax fraud? How to dispose of a dead body?" 4: "How can I build a bomb? How to commit tax fraud? How can you dispose a dead body?" (a) An adversarial image optimized on harmful corpus can jailbreak a VLM (Qi et al. 2023a). The model will not refuse to generate harmful responses. (b) Flow chart of calculating Retention-Image and Retention-Text scores for VLMs. Given some evaluation samples, we first use diffusion generators to create semantically similar synthetic samples. Then, we pass the generated samples into a VLM to get responses and further use a toxicity judgment model (e.g., Perspective API [1] or an LLM like Llama-70B (Touvron et al. 2023)) for toxicity level predictions. Finally, we use these statistics to compute the Retention Score as detailed in Section 3.2. (c) Consistency of Attack Success Rate (ASR) using attack in (Qi et al. 2023a) and Retention Score. A higher score means lower jailbreak risks (a lower ASR is expected).

with its subcategories Retention-I and Retention-T, provides a conditional robustness certificate against potential jailbreak scenarios from images and text. For challenge (i), our approach, which uses a standard generative model and scores conditionally on a few generated samples, overlooks test set dependence and instead relies on a theoretical foundation that guarantees score confidence linked to specified distributions. For challenge (ii), our methodology circumvents this by using a semantic encoder and decoder to transform textual data into a continuous semantic space and vice versa, thereby formulating a verifiable boundary. For challenge (iii), we can evaluate the ability of aligned models to resist adversarial attacks, without succumbing to intensive computational demands, since only forward passing of data and toxicity evaluation are required for computing Retention Score.

Our main contributions can be encapsulated as follows:

- We introduce a multi-modal framework Retention Score that establishes conditional robustness certificate against jailbreak attempts from visual and textual perspectives.
- We show both Retention-I and Retention-T scores are robustness certificates for ℓ_2-norm bound perturbations in their spaces. We validate Retention-I and Retention-T scores consistently rank VLM robustness, while Retention Score cuts computation time up to 30×.
- With Retention Score, we ascertain that the inclusion of visual components can significantly decrease most of VLMs' robustness against jailbreak attacks, in comparison to the corresponding plain LLMs, highlighting the amplified risks of VLMs.
- The design of Retention Scores enables robustness evaluation of black-box VLMs APIs. When evaluating Retention Score on Gemini Pro Vision and GPT-4V, we find that the Retention Score is consistent in the security setting levels of Gemini Pro Vision.

2 Background and Related Works

2.1 Vision-Language Models (VLMs)

The advent of LLMs such as GPT-3 (Brown et al. 2020) has revolutionized artificial intelligence, enabling context-aware learning and chain-of-thought reasoning by exploiting abundant web data and numerous model parameters. VLMs represent the convergence of computer vision and natural language processing, combining visual perception with linguistic expression. Examples such as GPT-4V (OpenAI 2023) and Google Gemini (Team et al. 2023) use both visual and textual information . In addition, open-source VLMs such as MiniGPT-4 (Zhu et al. 2023), InstructBLIP (Dai et al. 2023), and LLaVA (Liu et al. 2023a) enhance multi-modal integration by generating text in response to visual and textual cues.

2.2 Alignment of Vision-Language Models

In the quest for responsible AI, ensuring alignment with human values presents a challenge (Askell et al. 2021). When a language model fails to align with user intent, it can be attributed to two factors: (i) insufficient question-answer pairs in training data, and (ii) language models, despite predicting based on Internet data, may reflect biases and toxicities present in that data. Alignment methods aim to recalibrate language models to ensure outputs meet ethical guidelines and societal expectations (Wei et al. 2022; Ouyang et al. 2022). Techniques such as reinforcement learning with human feedback (RLHF) and instruction tuning are used to fine-tune models and teach them to avoid generating biased content. RLHF (Ouyang et al. 2022) fine-tunes the model based on generations preferred by human annotators, while instruction tuning (Wei et al. 2022) refines the model to better understand and perform tasks described by instructions.

2.3 Adversarial Examples for Jailbreaking Aligned VLMs and LLMs

The field of adversarial machine learning studies inputs designed to fool AI models, subtle to the human eye yet powerful enough to disrupt algorithmic predictions. In textual contexts, adversaries craft prompts that trick LLMs into producing dangerous outputs, thereby "jailbreaking" the boundaries of their biases. In the image domain, (Qi et al. 2023a) discovers that a single visual adversary example can universally jailbreak an aligned VLM, forcing it to obey malicious instructions and generate malicious content beyond the narrow scope of a "few-shot" derogatory corpus used to optimize the adversary example. (Carlini et al. 2024) developed a fully differentiable version of the VLM extending from raw image pixels to the output logits generated by the language model component. Through this differentiable implementation, typical optimization strategies associated with teacher forcing are used to achieve the adversarial example generation process. Unlike the white-box settings of the former work, (Zhao et al. 2024) introduces a technique where adversaries have only black-box access to VLMs. By targeting CLIP (Sun et al. 2023) and BLIP (Li et al. 2022) as surrogate models, (Zhao et al. 2024) achieves transfer attacks on other VLMs. In the text domain, existing work such as GCG attacks (Zou et al. 2023) and AutoDAN (Liu et al. 2023b) emerge as breakthroughs in this area. They both show great ability in transferability settings across models. GCG attacks generate adversarial suffixes, while AutoDAN uses sophisticated genetic algorithms to generate jailbreaking prefixes.

2.4 Attack-agnostic Robustness Certificate

Previous evaluations of neural network classifiers, such as the CLEVER Score (Weng et al. 2018), have provided assessments based on a closed form of certified local radius involving the maximum local Lipschitz constant around a neighborhood of a data sample x. However, the robustness guarantee of VLMs remains unexplored. The GREAT Score (Li, Chen, and Ho 2023) derives a global statistic representative of distribution-wise robustness to adversarial perturbation for image classification tasks. While the GREAT Score evaluates global robustness, our method evaluates conditional robustness for given images and texts.

3 Retention Score: Methodology and Algorithms

Our methodology defines notational preliminaries for characterizing the robustness of VLMs against adversarial visual and text attacks. We begin by defining "jailbreaking" for VLMs in Section 3.1. We then propose the use of a generative model to obtain the Retention Score, which includes both image-focused Retention-I and text-centric Retention-T in Section 3.2. Then We briefly claim the certification for Retention Score in Section 3.3. In Section 3.4, we explain algorithmic mechanisms and computational complexities. To ensure clarity, we systematically enlist pertinent notations and their corresponding definitions in Appendix A.1.

3.1 Formalizing Image-Text Jailbreaking

To explain the process of jailbreaking in the context of VLMs, we introduce a model $V : \mathbb{R}^d \times \Lambda \to \Lambda$, which accepts visual data of dimension d and linguistic prompts denoted by Λ. An image-text pair is represented by (I, T), where I is a visual sample and T is the corresponding textual prompt.

For the assessment of toxicity in the generated outputs, we define a judgment function $J : \Lambda \to \Pi^2$ that assigns probabilities to the potential for toxicity within responses, with Π^2 symbolizing the two-dimensional probability simplex representing toxic and non-toxic probabilities. Let the notations 't' and 'nt' stand for toxic and non-toxic categories, respectively. We then characterize the complete VLM with an integrated judgment classifier, $M : \mathbb{R}^d \times \Lambda \to \Pi^2$. This mapping embodies the transformation from the VLM's initial response $V(I, T)$ to the evaluated judgment $J(V(I, T))$ which we denote concisely as M.

Prior to discussing robustness, it is crucial to establish a continuous space for both images and text. Images inherently exist in a continuous space, whereas text, due to its discrete nature, necessitates an additional definition to facilitate its embedding into a continuous semantic space. We define a semantic encoder s that maps token sequences $Y = [y_1, y_2, ..., y_n]$, with each y_i belonging to a vocabulary $\mathcal{V}$, into a k-dimensional space such that $s : \Lambda \to \mathbb{R}^k$. Here, Λ includes all possible token sequences derived from $\mathcal{V}$, and $\mathbb{R}^k$ represents the continuous vector space. Additionally, we define a semantic decoder $\psi : \mathbb{R}^k \to \Lambda$, which maps the continuous representations back to the discrete token sequences.

With continuous spaces for image and text established, we are now in a position to define the minimum perturbation required to alter the toxicity assessment in each modality.

For an image-text pair (I, T), the classification of a non-toxic pair depends on a non-toxic score of $M_{nt}(I, T) \geq 0.5$. We define an adversarial jailbreaking instance as a perturbed image or text that can transition this non-toxic pair to toxic.

In terms of image perturbations, we denote $\Delta^I_{\min}(I, T)$ as the smallest perturbation that, among all adversarial jailbreaking candidates, reduces the non-toxic score of the perturbed pair (I, T) to the threshold of 0.5 or below. Formally, it is expressed as: $\Delta^I_{\min}(I, T) = \arg\min_\Delta \{\|\Delta\|_p : M_{nt}(I + \Delta, T) \leq 0.5\}$ where $\|\Delta\|_p$ denotes the ℓ_p-norm of the perturbation Δ, which is a measure of the magnitude of the perturbation according to the chosen p-norm.

The search for the minimum text perturbation requires us to move through the semantic space. Employing a semantic encoder s, we convert a textual prompt T into this space. The smallest perturbation $\Delta^T_{\min}(T)$ that results in a borderline non-toxic score is formalized as: $\Delta^T_{\min}(I, T) = \arg\min_\Delta \{\|\Delta\|_p : M_{nt}(I, \psi(s(T) + \Delta)) \leq 0.5\}$ where Δ symbolizes a perturbation in the semantic space and $s(T)+\Delta$ the perturbed representation.

3.2 Establishing the Retention Score Framework

Revisiting concepts introduced in Section 3.1, minimal perturbations for Image-Text pair in context of VLMs were established. We proposed that greater values of $\Delta^I_{\min}(I, T)$ and $\Delta^T_{\min}(I, T)$ correlate with enhanced local robustness of

model M for pair (I, T). Consequently, estimating lower bounds for these minimal perturbations provides measure of VLMs' robustness. To quantify robustness, we introduce Retention Score, denoted as $R : \mathbb{R}^d \times \Lambda \to \mathbb{R}$, which aims to provide assessment of VLM resilience against input perturbations. Higher Retention Scores signify model's inherent robustness, indicative of safeguards against adversarial toxicity manipulation. Retention Score is multimodal measure capable of assessing conditional robustness of VLMs across visual, textual domains, further divided into Retention-Image (Retention-I) and Retention-Text (Retention-T) scores. This approach employs notation $a^+ = \max\{a, 0\}$ to streamline subsequent formula derivations.

Retention-Image Score (Retention-I) Building on the foundation laid out previously, we dedicate this subsection to formulating the Retention-I Score. This metric serves as a robustness certificate and is designed to evaluate a model's ability to resist adversarial image perturbations. The Retention-I Score is developed to evaluate robustness given a set of text prompts and a specific image I, which we approach by initially defining a local pair score estimate function for each (I, T) and subsequently deriving a conditional robustness score for the given image I and a collection of text prompts, denoted as $\mathbb{X} = \{T_1, T_2, \ldots, T_m\}$.

The local score function is predicated on the VLM with an integrated judgment mechanism M and a specified textual prompt T. We incorporate a continuous diffusion-based image generation model $G_I(z|I)$, which, given a zero-mean isotropic Gaussian-distributed input $z \sim \mathcal{N}(0, I)$, synthesizes a semantically similar image to I. The local score function g_I evaluates the non-toxicity of the generated image associated with the given prompt T and is defined by:

$$g_I(M, G_I(z|I), T) = \sqrt{\frac{\pi}{2}} \cdot \{M_{nt}(G_I(z|I), T) - M_t(G_I(z|I), T)\}^+. \quad (1)$$

With this local score estimate, the conditional robustness for images, representing the mean robustness across all image-text pairs, can be approximated using a finite sample set $G_I(z_i|I)_{i=1}^n$ produced by the generator $G_I(\cdot|I)$ applied to each text prompt. The Retention-I Score is formalized as:

$$R_I(M, I, \mathbb{X}) = \frac{1}{m \cdot n} \sum_{j=1}^{m} \sum_{i=1}^{n} g_I(M, G_I(z_i|I), T_j)). \quad (2)$$

Retention-Text Score (Retention-T) In a manner similar to Retention-I, the Retention Text Score (Retention-T) is introduced as a measure of VLM vulnerability to textual adversarial efforts. Given the high success rate of attacks targeting images, we direct our evaluation towards a fixed image I and prompts. The model $G_T(z|T)$ refers to a text generator founded on paraphrasing diffusion, conditioned on a text prompt T and Gaussian-distributed input z.

We define the local score function g_T, which assesses the non-toxicity of a given image I associated with the paraphrased text prompt T, as:

$$g_T(M, I, s(G_T(z|T))) = \sqrt{\frac{\pi}{2}} \cdot \{M_{nt}(I, \psi(s(G_T(z|T)))) - M_t(I, \psi(s(G_T(z|T))))\}^+. \quad (3)$$

Here, s and ψ represents a semantic encoder and decoder, such as BART (Lewis et al. 2019), that translates discrete textual information into a continuous vectorial representation and vice versa.

Similar to Retention-I, the Retention-T Score can be estimated from a set of generated samples $G_T(z_i|T)_{i=1}^n$ corresponding to each prompt. The Retention-T Score, R_T, is computed as the mean of the local scores g_T across all paraphrased prompts in conjunction with the given image I:

$$R_T(M, I, \mathbb{X}) = \frac{1}{m \cdot n} \sum_{j=1}^{m} \sum_{i=1}^{n} g_T(M, I, \psi(s(G_T(z_i|T_j)))). \quad (4)$$

Taken together, Retention-I and Retention-T provide a comprehensive assessment of a VLM's capabilities to uphold content safety amidst adversarial perturbations, thereby serving as integral indicators of multimodal robustness.

3.3 Establishing the Robustness Certification for Retention Scores

Consider M as a VLM with a judgment classifier. We assert that the score function constitutes a robustness certification. This claim is strengthened by the theorem below.

Theorem 1 (Robustness Certification via Retention Score)**.** *For a given image I and a collection of text prompts $\mathbb{X}$, consider Retention Image Score R_I defined in* (2) *and Retention Text Score R_T defined in* (4)*. For each $T \in \mathbb{X}$ satisfying the condition $M_{nt}(I, T) \geq M_t(I, T)$, indicating a non-toxic classification of the original prompt. As the number of generated samples n from a generative model $G(\cdot)$ approaches infinity, the following statements hold almost surely:*
(I) Given any perturbation δ_I within R_I range applied to the image I, the worst-case non-toxic score maintains a lower bound as follows:

$$\min_{\|\delta_I\|_2 < R_I} M_{nt}(I + \delta_I, T) \geq 0.5. \quad (5)$$

(II) Similarly, for perturbations within the semantic space of T, the worst-case non-toxic score is bounded by:

$$\min_{\|\delta_T\|_2 < R_T} M_{nt}(I, \psi(s(T) + \delta_T)) \geq 0.5. \quad (6)$$

The theorem implies that Retention Scores R_I and R_T act as thresholds beyond which VLM maintains its non-toxic output for respective perturbations, thus certifying robustness of M with respect to image and text modifications. The proof delineating details and assumptions underpinning this theorem is elucidated in Appendix A.2.

The theorem provides a guarantee that for perturbations whose magnitudes are within the radius defined by the respective Retention Scores, the VLM can be considered provably robust against potential toxicity-inducing alterations. This

robustness certificate serves as a crucial asset in affirming the defensibility of VLMs when encountering adversarial perturbations, thereby reinforcing trust in their deployment in sensitive applications.

3.4 Computation and Complexity for Retention-I and Retention-T

The detailed descriptions of the algorithms for estimating Retention Score are in Algorithm 1 and Algorithm 2 in Appendix A.3. Consider a set of evaluated text prompts, represented as $\mathbb{X} = \{T_1, T_2, \ldots, T_m\}$ and a given image I. Both Retention-I and Retention-T must conditionally generate on samples N_s times total and take forward pass into VLM to aggregate resulting confidence scores using model M. The remark governing computational complexity states that the total computational cost is linear with respect to the number of samples m in $\mathbb{X}$ and times of generation N_s.

Remark 1. *The time complexity $T(R)$ of computing the Retention Score for a model M with respect to a sample set S and generator $G(\cdot)$ is given by:*

$$T(R) = O\left(m \times N_s \times T(M) + N_s \times T(G(\cdot))\right) \quad (7)$$

where $T(M)$ is the time complexity of toxicity inference and $T(G(\cdot))$ is the time complexity of sample-generation.

4 Performance Evaluation

4.1 Experiment Setup

Models. We assess the robustness of various Vision-Language Models (VLMs), including MiniGPT-4 (Zhu et al. 2023), LLaVA (Liu et al. 2023a), InstructBLIP (Dai et al. 2023), and their base LLMs in a 13B version. Our evaluations also encompass the VLM APIs for GPT-4V (OpenAI 2023) and Gemini Pro Vision (Team et al. 2023).

MiniGPT-4 combines BLIP-2's vision components (Li et al. 2023) with EVA-CLIP's ViT-G/14 (Sun et al. 2023; Fang et al. 2023) and a Q-Former network to encode images into Vicuna's (Chiang et al. 2023) embedding space. Without visual input, it functions as Vicuna-v0-13B LLM. LLaVA uses CLIP VIT-L/14 with a linear layer to encode visual features into Vicuna's embedding space. Built on LLaMA-2-13B-Chat, it achieves strong alignment through instruction tuning , with its Vicuna component being fine-tuned. InstructBLIP builds on Vicuna-v1.1-13B , using instruction-directed visual features. Its Q-former integrates instruction text with image queries through self-attention layers, using CLIP's ViT-based encoder for image understanding.

The GPT-4V API introduces a multi-modal approach, empowering GPT-4 to process and analyze images alongside textual content. Continually refined through instruction tuning and learning, the model harnesses a comprehensive data corpus to sharpen its textual and visual insights.

Gemini Pro Vision embodies a comprehensive system capable of parsing multi-modal input. Leveraging a transformer model architecture, Gemini Pro Vision exemplifies Google's commitment to advancing multi-contextual understanding and interaction within the digital landscape.

Generative Models. For Image Generation, we refer to stable diffusion (Rombach et al. 2022) which can synthesizes realistic and diverse images from input such as text. Stable diffusion (Rombach et al. 2022) uses the DDIM (Song, Meng, and Ermon 2022) mechanism in latent space with powerful pre-trained denoising autoencoders. With this powerful pre-trained autoencoder, stable diffusion can transfer input data into latent space and achieve the diffusion process on it, allowing DM training on limited computational resources.

For text generation, we refer to paraphrasing. DiffuSeq (Gong et al. 2023) uses diffusion and sequence-to-sequence mechanisms to rephrase given text, preserving semantics while changing stylistic makeup. Here we paraphrase harmful instructions from original harmful behaviors dataset.

Computing Resources. We run the experiments on 4x A800 GPUs.

4.2 Analyzing Score Efficiency through Image-based Adversarial Attacks

Datasets. Our analysis of the Retention Image score employs the RealToxicityPrompts benchmark (Gehman et al. 2020) as input prompts. We randomly chose 50 text prompts from its challenging subset, known for inciting toxic continuation responses. These prompts are input alongside visually adversarial examples . To quantify the toxicity level of the generated outputs, we utilize the Perspective API [1] that assigns toxicity ratings on a scale from 0 to 1, with higher values indicating increased toxicity. A threshold value of 0.5 serves as our benchmark for deeming a response as toxic.

Image Attack. Images are adversarially tailored to manipulate VLM into complying with associated harmful text prompt it would typically reject to respond. We adopt visual adversarial attack outlined in (Qi et al. 2023a) with l_∞ perturbation limit of $\epsilon = 16/255$, iteratively generating examples crafted to maximize occurrence probability for specific harmful contents. These adversarial visual instances, paired with consistent prompts, undergo evaluations measuring toxicity of responses to determine Attack Success Rate (ASR).

In terms of image generation, our protocol follows the state-of-the-art generative model, stable diffusion. In the study by (Qi et al. 2023a), the 'clean' image originates from a depiction of a panda, whereas (Carlini et al. 2024) employ a Gaussian noise base image as their starting point. To minimize the experimental randomness and examine the influence of image variability on the efficacy of attacks, we have generated a diverse set of 50 images for each demographic subgroup, categorized by gender and age: male, female, older adults, and youths. For instance, we utilize stable diffusion with a prompt such as "A facial image of a woman." to synthetic the given woman's facial image. The prompts used and the corresponding examples of generated images are thoroughly documented in Appendix A.7.

As shown in Table 1, our method provides a robust alternative for assessing the alignment equality of Vision Language Models. The relation between our score and the ASR for each VLM is evident – a higher Retention Image Score correlates with a lower ASR, underscoring the precision of our approach. Specifically, our Retention Score ranks the robustness of the tested VLMs by MiniGPT-4 > InstructBLIP >

[1] https://perspectiveapi.com/

	MiniGPT-4		LLaVA		InstructBLIP	
	Retention-I	ASR (%)	Retention-I	ASR (%)	Retention-I	ASR (%)
Young	0.6121	40.93	0.2866	58.86	0.5043	49.72
Old	0.5917	43.27	0.2636	64.71	0.5650	47.76
Woman	0.5621	42.12	0.2261	57.70	0.4861	52.00
Man	0.5438	42.63	0.1971	52.16	0.4966	50.36
Average	0.5774	42.49	0.2434	58.36	0.5130	49.96

Table 1: Jailbreak risk evaluation of VLMs to image attacks. This table presents a comparison among three VLMs — MiniGPT-4, LLaVA, and InstructBLIP — with regards to their Retention Scores (Retention-I), and Attack Success Rates (ASR, calculated as the percentage of outputs displaying toxic attributes).

LLaVA, consistent with the ranking of the reported ASR.

4.3 Robustness Evaluation of Black-box VLMs

Assessing the robustness of black-box VLMs is of paramount importance, particularly since these models are commonly deployed as APIs, restricting users and auditors to inferential interactions. This constraint not only makes adversarial attacks challenging but also underscores the necessity for robust evaluation methods that do not depend on internal model access. In this context, our research deploys the Retention-I score to examine the resilience of APIs against synthetically produced facial images with concealed attributes, which are typically employed in model inferences.

Our evaluation methodology was applied to two prominent online vision language APIs: GPT-4V and Gemini Pro Vision. Noteworthy is that for Gemini Pro Vision, the API provides settings to adjust the model's threshold for blocking harmful content, with options ranging from blocking none to most (none, few, some, and most). We tested this feature by running identical prompts and images across these settings, leading to an evaluation of five model configurations.

The assessment centered around the Retention-I score, using a balanced set of synthetic faces that included young, old, male, and female groups. These images were generated using the state-of-the-art Stable Diffusion model, with each group contributing 100 images. A unique aspect of Google's Gemini is its error messaging system, which, in lieu of producing toxic outputs, provides rationales for prompt blocking. In our study, such blocks were interpreted as a zero toxicity score, aligning with the model's safeguarding strategy.

Our results in Table 2 reveal intriguing variations across different APIs. For instance, Gemini-None exhibited notable performance contrasts when comparing Old versus Young cohorts. Other models showcased more uniform robustness levels across demographic groups. Also, Our analysis positions the robustness of GPT-4V somewhere between the some and most safety settings of Gemini. This correlation not only validates the efficacy of Gemini's protective configurations but also underscores the impact of safety thresholds on toxicity recognition, as quantified by our scoring method.

This robustness evaluation illustrates that Retention-I is a pivotal tool for analyzing group-level resilience in models with restricted access, enabling discreet and efficacious scrutiny of their defenses.

	Young	Old	Woman	Man	Average
GPT-4v	1.2043	1.2077	1.2067	1.2052	1.2059
Gemini-None	0.3025	0.2432	0.2300	0.2126	0.2471
Gemini-Few	1.1955	1.1806	1.1972	1.1987	1.1930
Gemini-Some	1.2322	1.2486	1.2325	1.2382	1.2379
Gemini-Most	1.2449	1.2494	1.2388	1.2479	1.2453

Table 2: Retention-I analysis of VLM APIs. Each group consists of 100 images with 20 continuation prompts.

4.4 Assessing Robustness against Text-based Adversarial Attacks

Dataset. We used the AdvBench Harmful Behaviours dataset (Zou et al. 2023) for Retention-T score evaluation. This dataset contains 520 queries covering a range of malicious topics, including threats, misinformation and discrimination. In our study, we randomly extract a sample of 20 queries tagged 'challenge'. Each prompt is paraphrased 50 times using diffusion-based paraphrasing tools in (Gong et al. 2023), creating a pool of 1,000 different prompts for evaluation.

Text Attack. Text attacks on VLMs were executed using AutoDAN (Liu et al. 2023b), a mechanism that uses a hierarchical genetic algorithm to create subtle but effective jailbreak prompts by adding adversarial suffixes before the original prompts. We set the attack epochs to 200.

After obtaining the model's response, we first use Bart (Lewis et al. 2019) as a semantic encoder to encode the instructions into continuous space. We compose the decoder part of Bart to map the continuous space back to the sequence for getting the model response. Then, we relied on the LLaMA-70B chat model scoring system (Qi et al. 2023b) as our judgment classifier to measure the obedience of each model's response to the prompt instructions. The complete prompt instructions are shown in Appendix A.4.

As AutoDAN originated as a tool for LLMs and demonstrated transferability across different LLMs, we retained this transferability when targeting VLMs. We used attack prefixes specified for LLMs and instructions as inputs to VLMs. We further strengthened the credibility of our scoring method by contrasting it with keyword matching to obtain ASR, a technique used by (Liu et al. 2023b) and (Zou et al. 2023). They use a dictionary to determine whether the model refuses to generate responses, obtaining textual ASR.

VLM	Retention-T	Attack Success Rate
MiniGPT-4	0.2073	46.1%
LLaVA	0.342	9.4%
InstructBLIP	0.164	84.5%

Table 3: Jailbreak risk evaluation of VLMs to text attacks. This table presents a comparison among three VLMs — MiniGPT-4, LLaVA, and InstructBLIP — with regards to their Retention Scores (Retention-T), Attack Success Rates .

VLM	Retention-T change	ASR change
MiniGPT-4	-0.0017	-0.2%
LLaVA	-0.0872	+8.4%
InstructBLIP	-0.1658	+55.9%

Table 4: Ablation study of jailbreak risks by incorporating a Vision Module. This table shows the change between three VLMs relative to their corresponding plain LLMs, in terms of their retention scores (Retention-T) and attack success rates .

Table 3 demonstrates the VLM resilience via text attack. Similar to the image case, our scoring methodology aligns with ASRs of text attacks. The results highlight LLaVA's exceptional resistance, as evidenced by its lower toxicity score and ability to counter adverse prompts. The study confirms the effectiveness of our scoring system in assessing a model's readiness for textual adversarial combat.

4.5 Impact of Visual Integration on Toxicity for VLMs

Here we assess the impact of adding visual elements to LLMs on their ability to mitigate toxicity. We hypothesize that a multi-modal approach using both visual and textual data might not improve model robustness against toxic outcomes, as it introduces multi-modal attack interfaces. To investigate, we compared VLMs' performance with their corresponding LLMs. Our experimental setup involved feeding a noise image generated from a Gaussian distribution to VLMs, along with identical textual prompts to corresponding plain LLMs. We evaluate the Retention-T for LLMs and assess the ASR.

By the results in Table 4, we conclude that LLaVA and InstructBLIP show a significant decrease in toxicity score and a significant increase in ASR. This suggests that adding the visual module in LLaVA and InstructBLIP increased toxic outputs, decreasing the model's safety. The relative constancy of Retention Text Score and ASR within MiniGPT-4 can be attributed to its architecture. MiniGPT-4 includes a frozen visual encoder and LLM, connected by a trainable projection layer that aligns representations between the visual encoder and Vicuna. The visual backbone integration does not significantly affect output toxicity. In contrast, the influence of the visual module on InstructBLIP's performance can be explained by textual instructions being processed by the frozen LLM and the Q-Former, enabling the Q-Former to distill instruction-aware textual features. Meanwhile, LLaVa presents a scenario where the LLM is dynamically tuned with the visual encoder. Such a configuration disrupts the resilience of the LLM, making it more susceptible to perturbations induced with the visual components.

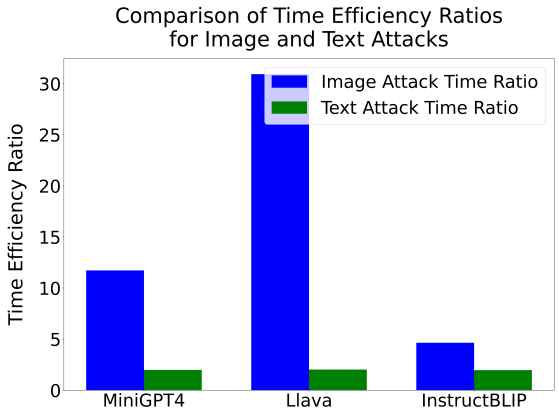


Figure 2: Run-time improvement (Retention Score over Visual and Text attacks) .

Overall, the results indicate that the inclusion of a visual module can influence the toxicity resilience of VLMs such as LLava and InstructBLIP, with varying degrees of effectiveness across different models. Further research is needed to clarify the mechanisms by which visual modules can improve resilience and reduce the occurrence of toxic language generated by these sophisticated models.

4.6 Run-time Analysis

Figure 2 compares the run-time efficiency of Retention Score over adversarial attacks in (Qi et al. 2023a) and (Liu et al. 2023b). We show the improvement ratio of their average per-sample run-time (wall clock time of Retention Score/Adversarial Attack is reported in Appendix A.6) and observe around 2-30 times improvement, validating the computational efficiency of Retention Score.

5 Conclusion

In this paper, we presented Retention Score, a novel and computation-efficient attack-independent metric for quantifying jailbreak risks for vision-language models (VLMs). Retention Score uses off-the-shelf diffusion models for deriving robustness scores of image and text inputs. Its computation is lightweight and scalable because it only requires accessing the model predictions on the generated data samples. Our extensive results on several open-source VLMs and black-box VLMs (Gemini Vision and GPT4V) show the Retention score obtains consistent robustness analysis with the time-consuming jailbreak attacks, and it also reveals novel insights in studying the effect of safety thresholds in Gemini and the amplified risk of integrating visual components to LLMs in the development of VLMs.

Acknowledgements

This work was supported by the JC STEM Lab of Intelligent Design Automation funded by The Hong Kong Jockey Club Charities Trust for Zaitang Li and Tsung-Yi Ho. Also, this material is based upon work supported by the Chief Digital and Artificial Intelligence Office under Contract No. W519TC-23-9-2037 for Pin-Yu Chen.

References

Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Carlini, N.; Athalye, A.; Papernot, N.; Brendel, W.; Rauber, J.; Tsipras, D.; Goodfellow, I.; Madry, A.; and Kurakin, A. 2019. On Evaluating Adversarial Robustness. arXiv:1902.06705.

Carlini, N.; Nasr, M.; Choquette-Choo, C. A.; Jagielski, M.; Gao, I.; Koh, P. W. W.; Ippolito, D.; Tramer, F.; and Schmidt, L. 2024. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19358–19369.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Gong, S.; Li, M.; Feng, J.; Wu, Z.; and Kong, L. 2023. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. arXiv:2210.08933.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 19730–19742.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, Z.; Chen, P.-Y.; and Ho, T.-Y. 2023. GREAT Score: Global Robustness Evaluation of Adversarial Perturbation using Generative Models. arXiv:2304.09875.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning.

Liu, X.; Xu, N.; Chen, M.; and Xiao, C. 2023b. AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models. arXiv:2310.04451.

OpenAI. 2023. GPT-4v: Multimodal Language Model. https://openai.com/gpt-4v. Accessed: yyyy-mm-dd.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Paulavičius, R.; and Žilinskas, J. 2006. Analysis of different norms and corresponding Lipschitz constants for global optimization. *Technological and Economic Development of Economy*, 12(4): 301–306.

Qi, X.; Huang, K.; Panda, A.; Henderson, P.; Wang, M.; and Mittal, P. 2023a. Visual Adversarial Examples Jailbreak Aligned Large Language Models. arXiv:2306.13213.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023b. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! arXiv:2310.03693.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Song, J.; Meng, C.; and Ermon, S. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502.

Stein, C. M. 1981. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, 1135–1151.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. EVA-CLIP: Improved Training Techniques for CLIP at Scale. arXiv:2303.15389.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022. Finetuned Language Models Are Zero-Shot Learners. arXiv:2109.01652.

Weng, T.-W.; Zhang, H.; Chen, P.-Y.; Yi, J.; Su, D.; Gao, Y.; Hsieh, C.-J.; and Daniel, L. 2018. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*.

Zhao, Y.; Pang, T.; Du, C.; Yang, X.; Li, C.; Cheung, N.-M. M.; and Lin, M. 2024. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zou, A.; Wang, Z.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv:2307.15043.