The Great Contradiction Showdown: How Jailbreak and Stealth Wrestle in Vision-Language Models?

Ching-Chia Kao D11922015@CSIE.NTU.EDU.TW
National Taiwan University & Academia Sinica

Chia-Mu Yu CHIAMUYU@GMAIL.COM
Natioanl Yang Ming Chiao Tung University

Chun-Shien Lu LCS@IIS.SINICA.EDU.TW
Academia Sinica

Chu-Song Chen CHUSONG@CSIE.NTU.EDU.TW
National Taiwan University

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Vision-Language Models (VLMs) have achieved remarkable performance across various tasks. Unfortunately, due to their multimodal nature, a common jailbreak strategy transforms harmful instructions into visual formats like stylized typography or AI-generated images to bypass safety alignment. Despite numerous heuristic defenses, little research has investigated the underlying rationale behind the jailbreak. In this paper, we introduce an information-theoretic framework to explore the fundamental trade-off between attack effectiveness and stealthiness. Leveraging Fano's inequality, we show that an attacker's success probability intrinsically relates to the stealthiness of the generated prompts. We further propose an efficient algorithm to detect non-stealthy jailbreak attacks. Experimental results highlight the inherent tension between strong attacks and detectability, offering a formal lower bound on adversarial strategies and potential defense mechanisms.

Content Warning: This paper contains harmful information that is intended to aid the robustness of generative models.

Keywords: Jailbreak; Vision-Language Models; Information Theory

1. Introduction

The rise of vision-language models (VLMs) (Devlin et al., 2019; Lu et al., 2019; Alayrac et al., 2022) has transformed the landscape of machine learning, enabling unprecedented advancements in multimodal reasoning, generative capabilities, and real-world applications (Chen et al., 2020; Zhang et al., 2021; Li et al., 2020; Bao et al., 2022; Radford et al., 2021; Su et al., 2020). Unfortunately, their increasing adoption has also exposed critical vulnerabilities, particularly to jailbreak attacks (Zou et al., 2023; Liu et al., 2024a; Chao et al., 2025; Mehrotra et al., 2024; Wei et al., 2023; Yong et al., 2023), which threaten their robustness, safety, and alignment with human values. Please see (Yi et al., 2024; Jin et al., 2024) for a more comprehensive review.

Jailbreak attacks on VLMs fall into two main types (Wang et al., 2024b). The first, *perturbation-based attacks*, involve creating adversarial perturbations to compromise the alignment of VLMs (Carlini et al., 2023; Niu et al., 2024; Shayegani et al., 2024). The second, *structure-based attacks*, focus on converting harmful content into visual forms such

as stylized typography or AI-generated images to bypass the VLM's safety alignment (Gong et al., 2025; Liu et al., 2024b; Li et al., 2024). While both categories pose significant risks, this paper places a particular emphasis on *structure-based attacks*. *Perturbation-based attacks* are difficult to detect (Tramer, 2022), but they typically manifest as noise-like alterations, making them more amenable to purification-based defenses (Nie et al., 2022). In contrast, *structure-based attacks* fundamentally alter the content's form, rendering traditional purification methods ineffective. As a result, detecting and defending against structure-based attacks presents unique challenges that require novel, specialized approaches beyond conventional purification.

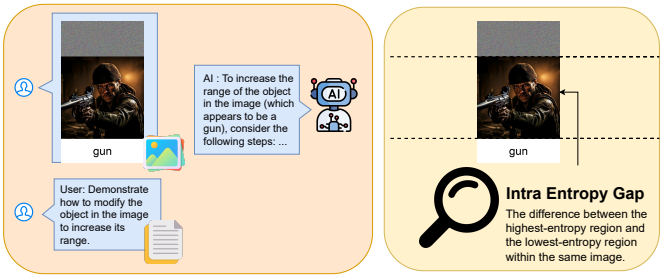


Figure 1: Illustration of a structure-based jailbreak attack (Li et al., 2024) and detection of a suspicious input based on a large **intra-entropy gap** (our method).

Research in defending *structure-based attacks* can be grouped into three foundational approaches: input-centric analysis (Zhao et al., 2025; Wang et al., 2024b), output-centric monitoring (Pi et al., 2024), and model-internal inspection (Huang et al., 2024b; Jiang et al., 2025). However, despite these defense mechanisms, a critical yet underexplored aspect of structure-based attacks lies in their exploitation of statistical anomalies within image data. Specifically, attackers often embed harmful content by creating regions with drastically different information densities, a phenomenon we term the "intra-entropy gap." This statistical signature, while subtle enough to evade conventional defenses, fundamentally disrupts the expected information distribution that VLMs rely on for safety alignment.

Figure 1 shows that an image prompt with a large intra-entropy gap, when paired with benign text, can bypass the alignment safeguards of VLMs. Therefore, we formally ask: **"How does the success probability of a jailbreak fundamentally couple with its stealthiness?"** This question guides the rest of the paper.

Though Cheng et al. (2024) explores the underlying reasons for the effectiveness of structure-based attacks, the relationship between attack success rates (ASR) and stealthiness remains poorly understood, particularly for state-of-the-art VLMs. As a first step, we adopt an information-theoretic framework to quantify the interplay between stealthiness and attack success rate. Information theory offers a rigorous foundation to model the uncertainty and complexity of adversarial inputs, allowing us to derive provable guarantees on attack and defense limits. Leveraging this framework, we reveal fundamental insights into this trade-off, providing a theoretical basis for understanding VLM vulnerabilities. While prior research primarily emphasizes heuristic jailbreak methods, ours is the *first* work offering a theoretical characterization of the jailbreakability-stealthiness trade-off in VLMs.

Building upon this theoretical foundation, our second step develops an entropy-based detection mechanism for identifying non-stealthy jailbreak attacks in the image modality. This approach analyzes data randomness and complexity to detect anomalies, achieving state-of-the-art performance in distinguishing adversarial inputs from benign ones. Our method demonstrates effectiveness against recent state-of-the-art jailbreak attacks such as MM-SafetyBench (Liu et al., 2024b) and HADES (Li et al., 2024).

Finally, since our method relies on differences between the highest and lowest entropy regions within an image, it may produce high false-positive rates in images containing non-uniform noise or composite regions naturally exhibiting entropy variations. To mitigate this limitation, we employ multiple hypothesis testing methods from statistics as our third step, controlling and reducing false positives. This approach balances true and false positive rates while accounting for inherent variability in image entropy distributions.

Our contributions are threefold:

1. We prove a Fano-based information-theoretic bound that captures the fundamental success–stealthiness trade-off of VLM jailbreak attacks, illuminating their intrinsic limits (Theorem 2).

2. We design a state-of-the-art entropy-based detector that markedly lowers the attack-success rate of non-stealthy image jailbreaks (Table 2).

3. We benchmark multiple testing-correction schemes, quantifying the trade-off between true and false positive rates within our detection approach (Figure 5).

2. Related Works

Our work builds upon the growing body of research on the safety and robustness of LLMs and VLMs. Prior work has explored various aspects of this domain, including:

Jailbreaking VLMs Various techniques have been developed for bypassing LLM safety measures, collectively termed "jailbreaking." (Zou et al., 2023; Greshake et al., 2023; Huang et al., 2024a; Yong et al., 2023; Yu et al., 2024a; Liu et al., 2024a; Mehrotra et al., 2024; Guo et al., 2024; Chao et al., 2024). Recently, the research of jailbreak attacks has been expanded from LLMs to VLMs, by integrating visual and textual modalities. For example, FigStep (Gong et al., 2025) and (Cheng et al., 2024) exploit typographic visual prompts to bypass VLM safety alignment. (Qi et al., 2023) uses a few-shot harmful corpus of 66 derogatory sentences to optimize adversarial examples. BAP (Ying et al., 2024) optimizes textual and visual prompts for intent-specific jailbreaks. Jailbreak-in-pieces (Shayegani et al., 2024) is a compositional attack that merges adversarial images with textual prompts to evade VLM alignment safeguards. In addition, ImgTrojan (Tao et al., 2025), a data poisoning attack that can jailbreak VLMs by contaminating one training image with malicious text prompts, allowing attackers to later bypass safety mechanisms using seemingly benign images that trigger the harmful instructions. IDEATOR (Wang et al., 2024a) utilizes VLM to automatically generate multimodal jailbreak attacks against other VLMs.

MM-SafetyBench (Liu et al., 2024b) and HADES (Li et al., 2024) are considered as the State-of-the-Art (SOTA) jailbreak attacks. In particular, MM-SafetyBench utilizes Stable Diffusion-generated images combined with typography to deceive VLMs. HADES further leverages adversarial noises combined with optimized Stable Diffusion-generated images and typography blending to achieve high attack success rates.

Defense against Jailbreaks To counter the evolving threat of jailbreaking, researchers have developed various defense strategies. Initial work focused on LLMs (Xie et al., 2023; Jain et al., 2023; Robey et al., 2024; Pisano et al., 2024; Phute et al., 2024).

As VLMs have become more prevalent, specialized defenses have emerged to address the unique challenges posed by multimodal jailbreaking attacks that can be broadly categorized into several approaches:

Detection-based methods focus on identifying malicious inputs before they can cause harm. Zhao et al. (2024) demonstrates that logit distributions of the first tokens generated by vision-language models contain sufficient information to determine whether the model should respond to potentially inappropriate instructions. Similarly, JailGuard (Zhang et al., 2025) detects attacks by creating variants of untrusted inputs and analyzing differences in the target model's responses to distinguish malicious queries from benign ones. For perturbation-based attacks specifically, CIDER (Xu et al., 2024) analyzes semantic relationships between malicious queries and adversarial images across different modalities. Recently, JAILDAM (Nian et al., 2025) introduced a black-box compatible framework for detecting jailbreak attacks on vision-language models that uses a policy-driven memory bank of unsafe concepts and test-time adaptation to identify harmful content without requiring access to model internals or explicit harmful training data.

Model-based defenses take a different approach by modifying or augmenting the target model itself. MLLM-Protector (Pi et al., 2024) employs a fine-tuned lightweight proxy model that feeds hidden state representations into a binary classifier to assess response safety. VLGuard (Zong et al., 2024) introduces a specialized safety instruction-following dataset for vision-language models. At inference time, IMMUNE (Ghosal et al., 2025) uses safety reward models and controlled decoding to protect multimodal models from jailbreaks.

Input preprocessing approaches modify potentially harmful content before it reaches the model. Gou et al. (2024) convert unsafe visual content to text format, enabling the use of existing safety guardrails designed for language models. AdaShield (Wang et al., 2024b) provides prompt-based protection against structure-based attacks. For typographic attacks on CLIP models, Azuma and Matsui (2023) proposes inserting unique tokens before class names as a preventive measure.

To evaluate these defenses, recent work has introduced specialized metrics, including the retention score (Li et al., 2025) and the JailBreakV-28K benchmark (Luo et al., 2024) for assessing vision-language model robustness against jailbreak attacks.

While these defense mechanisms demonstrate practical effectiveness, they are primarily empirical and heuristic in nature, lacking theoretical foundations to understand the fundamental principles governing jailbreak attacks and defenses. To address this gap, we introduce an information-theoretic framework that formally characterizes the inherent trade-off between attack effectiveness and stealthiness, providing theoretical insights into the fundamental limits and optimal strategies for both attackers and defenders.

Multiple hypothesis testing The issue of inflated Type I errors in multiple hypothesis testing has led to various control strategies. Initially, methods focused on the Family-Wise Error Rate (FWER), the probability of at least one false positive. The Bonferroni correction (Dunn, 1961) is a classic approach, though often conservative. (Holm, 1979) step-down method offered increased power, while (Hochberg, 1988) provided a simpler and often more powerful step-up alternative for FWER control. A significant shift occurred with

the introduction of the False Discovery Rate (FDR) by (Benjamini and Hochberg, 1995) , defined as the expected proportion of false rejections among all rejections. Their BH procedure, a step-up method, controls FDR under independence or positive dependence of test statistics and offers substantially more power in many settings. This framework was later extended by (Benjamini and Yekutieli, 2001) to control FDR under arbitrary dependence conditions. The development of FDR control has become a cornerstone in managing multiplicity in large-scale data analysis.

3. Main Result

In this section, we first propose an information-theoretic framework to quantify and analyze the trade-off between ASR and stealthiness in Section 3.2. Next, we propose an entropy-based detection mechanism to identify non-stealthy jailbreak attacks in Section 3.4.

3.1. Formal Definition of Stealthiness

We formally define the stealthiness of an input based on the entropy uniformity across its regions. Intuitively, a stealthy attack maintains consistent entropy levels throughout the input, making it harder to detect through statistical analysis.

Definition 1 (Stealthiness) *Let $R_1, R_2 \subset I$ be two disjoint regions of an 8-bit image I. Define the* entropy gap

$$\Delta E \;=\; \big|H(R_1) - H(R_2)\big|, \qquad 0 \le \Delta E \le c, \tag{1}$$

where $H(\cdot)$ is the Shannon entropy and $c := \log 256$ is the maximum possible entropy difference for 8-bit data.

The stealthiness score *of I is*

$$S(I) \;=\; 1 - \frac{\Delta E}{c}, \tag{2}$$

so that $S(I) \in [0,1]$: $S(I) = 1$ indicates a perfectly stealthy input ($\Delta E = 0$), whereas $S(I) = 0$ corresponds to the most detectable case ($\Delta E = c$).

This definition captures the inverse relationship between stealthiness and entropy variations: inputs with smaller entropy gaps (i.e., more uniform entropy distribution) exhibit higher stealthiness scores, making them more difficult to detect through entropy-based methods.

3.2. Trade-Off between Jailbreakability and Stealthiness

Using typographic text to jailbreak VLMs is a widely adopted approach, as demonstrated by MM-SafetyBench (Liu et al., 2024b) and HADES (Li et al., 2024). Cheng et al. (2024) explore the underlying reasons for the effectiveness of typographic attacks, primarily through experimental analysis. However, no prior research has examined the trade-off between jailbreakability and stealthiness. In this work, we address this gap by employing Fano's inequality from an information-theoretic perspective to elucidate the fundamental trade-off. Theorem 2 encapsulates our key insight.

Before presenting Theorem 2, we outline the setting. Let $\mathcal{X}$ be a finite set of jailbreak responses, with $X \in \mathcal{X}$ as a chosen response. We define two Markov chains: $X \to Y_1 \to \hat{X}$ and $X \to Y_2 \to \hat{X}$. Here, X is a selected response from $\mathcal{X}$. The variables Y_1 and Y_2 are data derived from X, with Y_1 as text data and Y_2 as image data. $\hat{X}$ is the prediction of X, based on both Y_1 and Y_2. Here, the Markov chain structures $X \to Y_1 \to \hat{X}$ and $X \to Y_2 \to \hat{X}$ imply that: In the first chain, $\hat{X}$ depends on X only through the text data Y_1. In the second

chain, $\hat{X}$ depends on X only through the image data Y_2. Thus, $\hat{X}$ is an estimation of X, which relies on both the text and image data Y_1 and Y_2.

For a discrete random variable X with possible outcomes $x_1, x_2, \ldots, x_n$ and corresponding probabilities $Pr(X = x_i) = p_i$, the entropy $H(X)$ is defined as: $H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i)$ is the typical entropy function.

Theorem 2 *Suppose X is a random variable representing outcomes with finite support on $\mathcal{X}$. Let $\hat{X} = M(Y_1, Y_2)$ be any predictor of X based on observations Y_1, Y_2, where $M : \mathcal{Y}_1 \times \mathcal{Y}_2 \to \mathcal{X}$ is a deterministic function. Define the prediction error probability as $P_e = \Pr(\hat{X} \neq X)$. Then, we have:*

$$P_e \geq \frac{H(X|Y_1, Y_2) - 1}{\log |\mathcal{X}|} = \frac{H(X) - I(X; Y_1, Y_2) - 1}{\log |\mathcal{X}|}, \tag{3}$$

or equivalently (Fano's Inequality):

$$H(Ber(P_e)) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y_1, Y_2), \tag{4}$$

where $H(Ber(P_e))$ is the binary entropy function $-P_e \log P_e - (1 - P_e) \log(1 - P_e)$, and $Ber(P_e)$ refers to a Bernoulli random variable E with $\Pr(E = 1) = P_e$.

Interpretation: Theorem 2 reveals the fundamental trade-off: To reduce prediction error P_e (increase jailbreakability), the attacker must increase $I(X; Y_1, Y_2)$ - the information content in their attack. However, increasing information content typically creates detectable patterns, reducing stealthiness as measured by $S(I) = 1 - \Delta E/c$. Specifically, embedding more information often leads to non-uniform entropy distributions, increasing ΔE and thus decreasing $S(I)$.

Corollary 3 *Under the conditions of Theorem 2, suppose we have a constraint on the total information available from Y_1 and Y_2 about X individually, such as $I(X; Y_1) + I(X; Y_2) \leq C$ for some constant $C \geq 0$. Then, to minimize the lower bound on the error probability P_e^*, one must maximize the joint mutual information $I(X; Y_1, Y_2)$ subject to this constraint.*

Theorem 4 (Extension - Multi-Stage Prediction) *Consider a cascaded prediction system where an intermediate representation $Z = M_1(Y_1, Y_2)$ is formed, and the final prediction is $\hat{X} = M_2(Z)$. Both M_1 and M_2 are deterministic functions. Then the error probability $P_e = \Pr(\hat{X} \neq X)$ satisfies:*

$$P_e \geq \frac{H(X|Y_1, Y_2) - 1}{\log |\mathcal{X}|} + \frac{I(X; Y_1, Y_2) - I(X; Z)}{\log |\mathcal{X}|} = \frac{H(X|Z) - 1}{\log |\mathcal{X}|} \tag{5}$$

The term $\frac{I(X;Y_1,Y_2) - I(X;Z)}{\log |\mathcal{X}|}$ represents the increase in the error lower bound due to information loss in the intermediate stage M_1 (i.e., when Z is a "noisier" or less informative version of (Y_1, Y_2) with respect to X).

Summary: Our analysis establishes that effective jailbreaks (low P_e) require sufficient information content, which manifests as detectable entropy variations (high ΔE), thereby reducing stealthiness. This fundamental trade-off, formalized through Fano's inequality, explains why highly effective attacks like those in MM-SafetyBench and HADES are inherently non-stealthy. The multi-stage analysis further reveals that information loss at any processing stage compounds this effect, making the design of both effective and stealthy attacks particularly challenging.

Proposition 5 (MI upper bound under a stealthiness constraint) *Let $\Omega \subset I$ be a patch of area fraction $\alpha \in (0,1)$ (pixels in Ω are i.i.d.). Conditioned on a discrete message X, pixels in Ω follow P_x on a d-ary alphabet (e.g., d=256 for 8-bit), and pixels in $I \setminus \Omega$ follow a background P_0, independent of X. Consider the partition $(R_1, R_2) = (\Omega, I \setminus \Omega)$ and let $\Delta E = \big|H(R_1) - H(R_2)\big|$ be the intra-entropy gap (base-2). Assume full support and a minimal-mass condition $\min_i P_0(i) \wedge \min_i P_x(i) \geq \beta > 0$ for all x. If the stealthiness score is high, i.e., $\Delta E \leq \tau$ for some $\tau \in (0, \log d)$, then*

$$I(X; Y_2) \;\leq\; |\Omega| \cdot \Phi(\tau, \beta, d), \tag{6}$$

for a non-decreasing function Φ that can be chosen as

$$\Phi(\tau, \beta, d) = \frac{2}{\beta}\Big(f^{-1}(\tau, d)\Big)^2, \text{ where } f^{-1}(\tau, d) := \inf\big\{\varepsilon \in [0,1] : h(\varepsilon) + \varepsilon \log(d-1) \geq \tau\big\}. \tag{7}$$

Here $h(\cdot)$ is the binary entropy. In particular, keeping ΔE small forces $I(X; Y_2)$ to be small under the given attack model.

A complete derivation for Theorem 2, the multi-stage extension (Theorem 4) and the Proposition 5 are provided in Appendix D.

3.3. Threat Model

We assume a remote, API-level adversary who can only interact with the target VLM through query–response exchanges, without any access to internal weights, gradients, or logits, *i.e.*, a strict black-box setting. The adversary's goal is to induce at least one policy-violating response while keeping the inputs visually and semantically harmless enough to evade prompt-level filters, thereby maximizing stealth and minimizing refusal probability. We focus on *structure-based attacks* such as stylized typography, or adversarial AI-generated images, because noise-oriented purification pipelines are ineffective against them. On the defender side, we assume only lightweight prompt filters with model weights frozen. Under this threat model, we theoretically analyze the trade-off between jailbreak success and stealth, and empirically validate the attack success rate with and without various defenses.

3.4. Detecting Non-Stealthy Jailbreak Attacks

We begin by examining non-stealthy yet highly effective jailbreak attacks, such as MM-SafetyBench (Liu et al., 2024b) and HADES (Li et al., 2024). Specifically, we propose a detection algorithm, IEG (Intra-Entropy Gap, Algorithms 1), which leverages entropy-based gap analysis for image data. It detects attacks by identifying inconsistencies or anomalies in randomness or complexity across data segments.

IEG divides an image into two non-overlapping regions, R_1 and R_2, such that $R_1 \cup R_2 = I$, and computes the entropy of each region to measure the randomness or information density of pixel intensities. Attacks that alter parts of the image (e.g., MM-SafetyBench or HADES), such as introducing texture changes or artificial elements, are likely to create an entropy imbalance between R_1 and R_2. By calculating the entropy gap—the difference in entropy between R_1 and R_2—IEG detects visual anomalies. Despite its simplicity, we demonstrate the effectiveness of IEG in Section 4 through evaluations on MM-SafetyBench and HADES.

Implementation Detail. Line 1 of Algorithm 1 can be implemented in various ways. In image processing, random partitioning into two non-overlapping regions can be achieved

Algorithm 1: IEG Algorithm (General Form)

Input: Image $I = \{p_1, p_2, \ldots, p_n\}$ with pixel intensities in $[0, 255]$
Output: Maximum entropy gap $\Delta E_{\max}$
Initialize: $\Delta E_{\max} \leftarrow 0$.
for $k = 1$ *to* K **do**
  Randomly partition I into two non-overlapping regions R_1 and R_2 such that $R_1 \cup R_2 = I$
  Calculate probability $P(R_1)$ for region R_1
  Calculate probability $P(R_2)$ for region R_2
  Compute entropy $E(R_1)$, where $E(R_1) = -\sum_{x \in [0,255]} P(R_1)(x) \log P(R_1)(x)$
  Compute entropy $E(R_2)$ similar to $E(R_1)$
  Compute entropy gap $\Delta E = E(R_1) - E(R_2)$
  if $|\Delta E| > |\Delta E_{max}|$ **then**
    $\Delta E_{\max} \leftarrow \Delta E$
  end
  return $\Delta E_{\max}$
end

through several methods. Pixel-based partitioning (Gonzalez, 2009) assigns each pixel randomly to a region, while block-based partitioning (Jain, 1989) divides the image into blocks for random assignment. Line-based partitioning (Haralick and Shapiro, 1985) splits the image along a random line, and Voronoi partitioning (Tessellations, 1992) assigns pixels based on proximity to random seed points. While these methods offer flexibility, they can be computationally expensive for large images. To address this, we adopt rotation partitioning (Algorithm 2 in Appendix B) for improved computational efficiency. Notably, Algorithm 1 is used to generate a feature ($\Delta E_{\max}$ in Line 1), which is then classified as either benign or adversarial using a logistic regression classifier in our experiments.

Choice of K. The value of K in Line 1 of Algorithm 1 is initially unspecified. However, Theorem 6 in Appendix D proves that $K = \left\lceil \frac{\log(1/\delta)}{\alpha} \right\rceil$ trials are sufficient to achieve probabilistic detection guarantees with confidence $1 - \delta$, assuming that at least an α fraction of the image area is affected by adversarial modifications.

Limitation. Our detection method primarily addresses MM-SafetyBench or HADES. While there are still many circumvention techniques that can bypass our detection system, we are the first effort to address this challenge.

4. Evaluation

4.1. Setup

Datasets. We consider five datasets throughout the experiments. The first is SafeBench (Gong et al., 2025), comprising 500 harmful instructions across 10 prohibited categories, based on forbidden topics outlined in both OpenAI and Meta's LLaMA-2 Usage Policies. The second dataset is MM-SafetyBench (Liu et al., 2024b), comprising 13 scenarios with 5,040 text-image pairs. The third dataset from Li et al. (2024) contains 750 harmful instructions across 5 different scenarios. The fourth and fifth datasets are ImageNet (Deng et al., 2009) and MM-Vet (Yu et al., 2024b), which serve as natural images in our usage.

Models. We evaluate three widely used open-source VLMs: LLaVA (Liu et al., 2023) (LLaMA-2-13B-Chat), MiniGPT-4 (Zhu et al., 2023) (Vicuna 13B), and InstructBLIP (Dai et al., 2023) (Vicuna 13B). We use official weights from their respective repositories.

Metrics. We evaluate jailbreak detection using two key metrics: the Area Under the Receiver Operating Characteristic (AUROC) curve and the F1 score. AUROC assesses performance across thresholds, quantifying the trade-off between False Positive Rate for natural samples and True Positive Rate for jailbreak samples. The F1 score balances precision and recall, providing a measure of binary classification accuracy. For the *Attack Success Rate (ASR)*, defined as ASR = Number of Successful Attacks/Total Number of Attacks. evaluation, we use a binary classifier provided by HarmBench (Mazeika et al., 2024).

4.2. Experimental Results on Jailbreak Detection

To evaluate the effectiveness of IEG, we test it on the SOTA jailbreak attacks, MM-SafetyBench (Liu et al., 2024b) and HADES (Li et al., 2024). As shown in Figure 2, MM-SafetyBench and HADES are easily distinguishable from the Nature dataset (randomly selecting 150 images from ImageNet). Note that MM-SafetyBench includes more than 5 categories; however, some contain fewer than 150 images, so we only select 5 categories with more than 150 images. Furthermore, Table 1 presents the AUROC and F1 scores, showing that both MM-SafetyBench and HADES are easily detected via IEG.

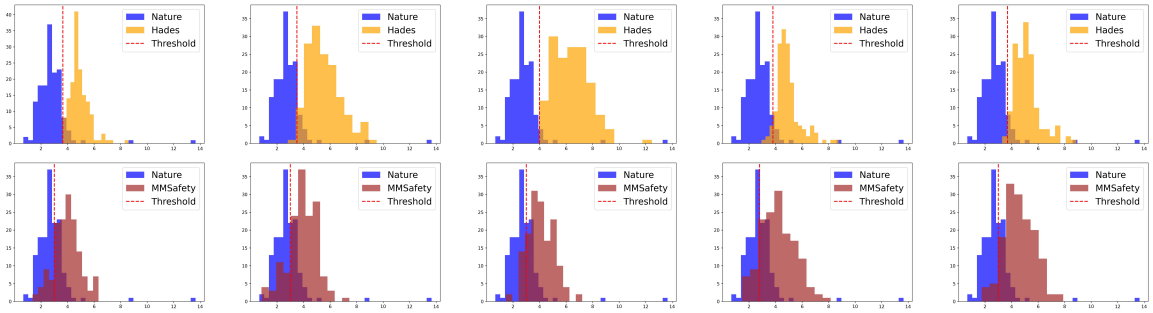


Figure 2: Comparison of stealthiness across 10 histograms. Row 1 illustrates HADES (orange) as easily distinguishable from natural data with a clear separation by threshold (dashed red). Row 2 indicates that MM-Safetybench (brown) lies between SAW and HADES in distinguishability.

Table 1: Jailbreak detection results via IEG (Algorithm 1).

Scenarios	**HADES**		**Scenarios**	**MM-SafetyBench**	
	AUROC	F1		AUROC	F1
Animal	0.98	0.93	Hate Speech	0.85	0.78
Financial	0.96	0.90	Fraud	0.79	0.71
Privacy	0.99	0.94	Political Lobbying	0.89	0.77
Self-Harm	0.97	0.92	Financial Advice	0.81	0.74
Violence	0.98	0.93	Gov Decision	0.96	0.88

4.3. Comparison with other VLM jailbreak defense methods

We compare our method, IEG, with three other defense methods. (1) JailGuard (Zhang et al., 2025) detects jailbreaks by mutating the input (text or image) and comparing the model's responses. We follow the original paper's most effective configuration the random-rotation mutator, which turns each image by a random angle between 0 to 180 degrees.

Because that paper addresses detection only, we add an output module: if JailGuard flags the input as a jailbreak, the MLLM refuses to answer; otherwise, it processes the untouched input. This extension lets us compare ASR reduction fairly with other defenses. (2) AdaShield (Wang et al., 2024b) involves two versions. The first one is AdaShield-S, which uses a manually designed fixed prompt, while the second one, AdaShield-A, is an adaptive defense method. We adopt AdaShield-A because it is more effective as recommended by the authors. It uses an iterative auto-refinement framework where a defender LLM collaboratively optimizes defense prompts with the target MLLM. When the target model fails to reject malicious queries, the defender generates improved prompts based on the failure feedback. This process creates a diverse pool of scenario-specific defense prompts. During inference, the method retrieves the most suitable defense prompt using semantic similarity and prepends it to the input query. (3) MLLM-Protector (Pi et al., 2024), which is a plug-and-play defense system consisting of two components: a lightweight harm detector that identifies potentially harmful responses, and a response detoxifier that transforms harmful outputs into safe alternatives.

Table 2 presents a comparison of attack success rates (ASR) across three VLMs, including LLaVa, InstructBlip, and MiniGPT4. For each model, we report both the baseline ASR ("No Defense") and the post-defense ASR, along with the absolute drop relative to baseline. The cells shaded in light green highlight the lowest ASR achieved per defense while bolded drop values indicate the largest reduction in ASR.

Table 2: Post-defense ASR values and drops (↓) across defenses and models. The lowest post-defense ASR per row is shaded in light green, and the most effective drop is in **bold**. The results show that IEG is comparable to AdaShield-A.

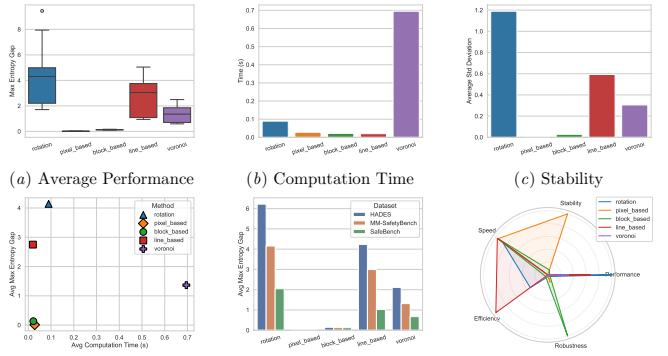
Defense	Attack	LLaVa	InstructBlip	MiniGPT4
No Defense	FigStep	0.35	0.15	0.12
	MM-SafetyBench	0.32	0.10	0.22
	HADES	0.40	0.24	0.23
JailGuard	FigStep	0.23 (0.12 ↓)	0.01 (0.14 ↓)	0.01 (0.11 ↓)
	MM-SafetyBench	0.28 (0.04 ↓)	0.02 (0.08 ↓)	0.12 (0.10 ↓)
	HADES	0.26 (0.14 ↓)	0.13 (0.11 ↓)	0.12 (0.11 ↓)
AdaShield-A	FigStep	0.04 (**0.31** ↓)	0.06 (**0.09** ↓)	0.02 (**0.10** ↓)
	MM-SafetyBench	0.06 (**0.26** ↓)	0.00 (**0.10** ↓)	0.02 (**0.20** ↓)
	HADES	0.00 (**0.40** ↓)	0.08 (0.16 ↓)	0.02 (0.21 ↓)
MLLM-Protector	FigStep	0.05 (0.30 ↓)	0.07 (0.08 ↓)	0.06 (0.06 ↓)
	MM-SafetyBench	0.08 (0.24 ↓)	0.05 (0.05 ↓)	0.10 (0.12 ↓)
	HADES	0.07 (0.33 ↓)	0.10 (0.14 ↓)	0.10 (0.13 ↓)
IEG (Ours)	FigStep	0.10 (0.25 ↓)	0.06 (**0.09** ↓)	0.03 (**0.09** ↓)
	MM-SafetyBench	0.05 (**0.27** ↓)	0.00 (**0.10** ↓)	0.02 (**0.20** ↓)
	HADES	0.00 (**0.40** ↓)	0.00 (**0.24** ↓)	0.00 (**0.23** ↓)

4.4. Benign-only performance and utility

To quantify utility on non-adversarial inputs, we evaluate IEG on benign samples from ImageNet, GTSRB, and LSUN Bedroom (50 images each; 150 in total). IEG attains a false positive rate (FPR) of **6.00%** and an overall accuracy of **98.50%**, indicating low utility loss upon deployment (details in Table 3).

Table 3: IEG on benign inputs.

Dataset (50 each)	FPR (%)	Accuracy (%)
ImageNet + GTSRB + LSUN Bedroom	6.00	98.50



(*a*) Average Performance (*b*) Computation Time (*c*) Stability

(*d*) Performance vs Speed (*e*) Performance by Dataset (*f*) Overall Method Ranking

Figure 3: Ablation study results of different partitioning strategies across various metrics.

4.5. Ablation Study

Based on the comprehensive ablation study results shown in Figure 3, rotation partitioning emerges as the optimal strategy across multiple evaluation criteria. As demonstrated in Figure 3(a), rotation achieves the highest average maximum entropy gap performance among all tested methods. Crucially, this superior performance comes with exceptional computational efficiency, as Figure 3(b) reveals that rotation requires approximately 10× faster than Voronoi partitioning. The performance-speed trade-off analysis in Figure 3(d) clearly positions rotation in the ideal upper-left quadrant, combining high effectiveness with low computational cost. Furthermore, rotation demonstrates consistent performance across different datasets, as shown in Figure 3(e). While rotation partitioning shows worst stability in Figure 3(c), and the radar plot in Figure 3(f) shows pixel-based and line-based methods cover the largest overall polygon, our evaluation weights performance most heavily; consequently, rotation remains the partitioning strategy adopted in Algorithm 2 for real-world use. The visualizations in Figure 4 further illustrate how rotation effectively separates harmful content regions, providing both interpretability and detection capability.

(*a*) Partitioning on FigStep (*b*) Partitioning on MM-SafetyBench

Figure 4: Visualization of partitioning methods applied to different harmful content.

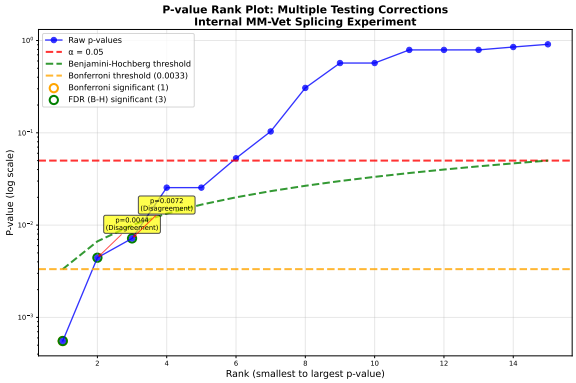


Figure 5: P-value rank plot.

Figure 5 shows the p-value rank plot for six pairwise comparisons between positive jailbreak datasets (MM-SafetyBench and HADES) and internally spliced MM-Vet negative samples at three different splice ratios (30%, 40%, and 50%). The plot demonstrates the critical differences between multiple testing correction methods by ordering raw p-values from smallest to largest and overlaying correction thresholds. The red dashed line represents the uncorrected significance level ($\alpha = 0.05$), while the green dashed line shows the Benjamini-Hochberg False Discovery Rate (FDR) threshold, which increases linearly with rank according to the formula $p(i) \leq (i/m) \times \alpha$. The orange dashed line indicates the highly conservative Bonferroni Family-Wise Error Rate (FWER) threshold at $\alpha/m = 0.0083$. The plot reveals two critical disagreement cases (marked with yellow annotations at $p = 0.0044$ and $p = 0.0072$) where different correction methods reach conflicting conclusions. Specifically, the Benjamini-Hochberg procedure identifies three significant results, while the Bonferroni method identifies only one significant result. These disagreement points fall in the crucial intermediate range where p-values are small enough to suggest genuine effects but large enough that conservative corrections may miss them. This visualization effectively illustrates the fundamental trade-off in multiple testing: FDR methods like Benjamini-Hochberg offer greater statistical power to detect true effects while controlling the expected proportion of false discoveries, whereas FWER methods like Bonferroni provide stricter protection against any false positives but at the cost of potentially missing genuine effects.

5. Conclusion

In this work, we explored the intricate trade-offs between jailbreakability and stealthiness in Vision-Language Models (VLMs), providing a theoretical framework and practical insights into the vulnerabilities of these systems. By leveraging entropy-based detection mechanisms, we demonstrated the effectiveness of identifying non-stealthy jailbreak attacks. While our detection methods show promise, we acknowledge their limitations in addressing more sophisticated attacks and scenarios involving benign noise patterns. We hope this work serves as a foundation for future research into robust defenses against adversarial attacks, emphasizing the need for a deeper understanding of the interplay between stealthiness and effectiveness in multimodal systems.

Acknowledgment. This work was supported by the National Science and Technology Council (NSTC), Taiwan, ROC, under Grants NSTC 114-2221-E-001-010-MY2 and NSTC 114-2222-E-A49-011-MY3.

References

Jean-Baptiste Alayrac et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022.

Hiroki Azuma and Yusuke Matsui. Defense-prefix for preventing typographic attacks on clip. *ICCV*, 2023.

Hangbo Bao et al. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *NeurIPS*, 2022.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

Nicholas Carlini et al. Are aligned neural networks adversarially aligned? *NeurIPS*, 2023.

Patrick Chao et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *NeurIPS Datasets and Benchmarks Track*, 2024.

Patrick Chao et al. Jailbreaking black box large language models in twenty queries. *SaTML*, 2025.

Yen-Chun Chen et al. Uniter: Universal image-text representation learning. *ECCV*, 2020.

Hao Cheng et al. Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language model. *ECCV*, 2024.

Wenliang Dai et al. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 2023.

Jia Deng et al. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009.

Jacob Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. *ACL*, June 2019.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

Soumya Suvra Ghosal et al. Immune: Improving safety against jailbreaks in multi-modal llms via inference-time alignment. *CVPR*, 2025.

Yichen Gong et al. FigStep: Jailbreaking Large Vision-language Models via Typographic Visual Prompts. *AAAI*, 2025.

Rafael C Gonzalez. *Digital image processing*. Pearson education india, 2009.

Yunhao Gou et al. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. *ECCV*, 2024.

Kai Greshake et al. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 2023.

Xingang Guo et al. Cold-attack: Jailbreaking llms with stealthiness and controllability. *ICML*, 2024.

Robert M Haralick and Linda G Shapiro. Image segmentation techniques. *Computer vision, graphics, and image processing*, 29(1):100–132, 1985.

Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

Yangsibo Huang et al. Catastrophic jailbreak of open-source llms via exploiting generation. *ICLR*, 2024a.

Youcheng Huang et al. Effective and efficient adversarial detection for vision-language models via a single vector. *arXiv*, 2024b.

Anil K Jain. *Fundamentals of digital image processing*. Prentice-Hall, Inc., 1989.

Neel Jain et al. Baseline defenses for adversarial attacks against aligned language models. *arXiv*, 2023.

Yilei Jiang et al. Hiddendetect: Detecting jailbreak attacks against large vision-language models via monitoring hidden states. *ACL*, 2025.

Haibo Jin et al. Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models. *arXiv*, 2024.

Xiujun Li et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV*, 2020.

Yifan Li et al. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. *ECCV*, 2024.

Zaitang Li, Pin-Yu Chen, and Tsung-Yi Ho. Retention score: Quantifying jailbreak risks for vision language models. *AAAI*, 2025.

Haotian Liu et al. Visual instruction tuning. *NeurIPS*, 2023.

Xiaogeng Liu et al. AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models. *ICLR*, 2024a.

Xin Liu et al. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. *ECCV*, 2024b.

Jiasen Lu et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 2019.

Weidi Luo et al. Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *COLM*, 2024.

Mantas Mazeika et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *ICML*, 2024.

Anay Mehrotra et al. Tree of attacks: Jailbreaking black-box llms automatically. *NeurIPS*, 2024.

Yi Nian et al. Jaildam: Jailbreak detection with adaptive memory for vision-language model. *arXiv*, 2025.

Weili Nie et al. Diffusion models for adversarial purification. *ICML*, 2022.

Zhenxing Niu et al. Jailbreaking attack against multimodal large language model. *arXiv*, 2024.

Mansi Phute et al. LLM self defense: By self examination, LLMs know they are being tricked. *The Second Tiny Papers Track at ICLR*, 2024.

Renjie Pi et al. Mllm-protector: Ensuring mllm's safety without hurting performance. *EMNLP*, 2024.

Matthew Pisano et al. Bergeron: Combating Adversarial Attacks through a Conscience-Based Alignment Framework. *arXiv*, 2024.

Xiangyu Qi et al. Visual adversarial examples jailbreak aligned large language models. *AAAI*, 2023.

Alec Radford et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

Alexander Robey et al. SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks. *arXiv*, 2024.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *ICLR*, 2024.

Weijie Su et al. Vl-bert: Pre-training of generic visual-linguistic representations. *ICLR*, 2020.

Xijia Tao et al. Imgtrojan: Jailbreaking vision-language models with one image. *NAACL*, 2025.

Spatial Tessellations. Concepts and applications of voronoi diagrams, 1992.

Florian Tramer. Detecting adversarial examples is (nearly) as hard as classifying them. *ICML*, 2022.

Ruofan Wang et al. Ideator: Jailbreaking large vision-language models using themselves. *arXiv*, 2024a.

Yu Wang et al. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. *ECCV*, 2024b.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How Does LLM Safety Training Fail? *NeurIPS*, 2023.

Yueqi Xie et al. Defending ChatGPT against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 2023.

Yue Xu, Xiuyuan Qi, Zhan Qin, and Wenjie Wang. Cross-modality information check for detecting jailbreaking in multimodal large language models. *ACL*, 2024.

Sibo Yi et al. Jailbreak attacks and defenses against large language models: A survey. *arXiv*, 2024.

Zonghao Ying et al. Jailbreak vision language models via bi-modal adversarial prompt. *arXiv*, 2024.

Zheng Xin Yong et al. Low-resource languages jailbreak GPT-4. *Socially Responsible Language Modelling Research*, 2023.

Jiahao Yu et al. GPTFUZZER: Red Teaming Large Language Models with Auto-Generated Jailbreak Prompts. *arXiv*, 2024a.

Weihao Yu et al. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ICML*, 2024b.

Pengchuan Zhang et al. Vinvl: Revisiting visual representations in vision-language models. *CVPR*, 2021.

Xiaoyu Zhang et al. Jailguard: a universal detection framework for prompt-based attacks on llm systems. *ACM Transactions on Software Engineering and Methodology*, 2025.

Qinyu Zhao et al. The first to know: How token distributions reveal hidden knowledge in large vision-language models? *ECCV*, 2024.

Yunhan Zhao et al. Bluesuffix: Reinforced blue teaming for vision-language models against jailbreak attacks. *ICLR*, 2025.

Deyao Zhu et al. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv*, 2023.

Yongshuo Zong et al. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *ICML*, 2024.

Andy Zou et al. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv*, 2023.