Appendix A. Notation Table

Table 4 summarizes the notations used throughout this paper.

Table 4: Notation Table

Notation	Description
M	Vision-Language Model (VLM)
$p_1, p_2, \ldots, p_n$	Pixels in the image with intensities in $[0, 255]$
R_1, R_2	Randomly selected regions of image I
$P(R_1), P(R_2)$	Probability distribution of pixel intensities in regions R_1 and R_2
$E(R_1), E(R_2)$	Entropy of regions R_1 and R_2
ΔE	Entropy gap between regions R_1 and R_2
$\Delta E_{\max}$	Maximum entropy gap
X	Random variable representing harmfulness outcomes
$\mathcal{X}$	Finite support of random variable X
$\hat{X}$	Predicted value of X
P_e	Probability of error, i.e., $Pr(\hat{X} \neq X)$
$H(X\|Y_1, Y_2)$	Conditional entropy of X given inputs Y_1 and Y_2
$I(X; Y_1, Y_2)$	Mutual information between X and the inputs Y_1, Y_2
$Ber(P_e)$	Bernoulli random variable E with $Pr(E = 1) = P_e$
$\mathcal{H}(X)$	Entropy of subset $X \subseteq T$
$I_{\text{rot}}(\theta)$	Region of image after partitioning by a line at angle θ
$I_{\text{rot}}^{\perp}(\theta)$	Complementary region of image after partitioning at angle θ
$P(I_{\text{rot}}(\theta))$	Probability distribution of pixel intensities in $I_{\text{rot}}(\theta)$
$P(I_{\text{rot}}^{\perp}(\theta))$	Probability distribution of pixel intensities in $I_{\text{rot}}^{\perp}(\theta)$
$E_{\text{rot}}(\theta)$	Entropy of region $I_{\text{rot}}(\theta)$
$E_{\text{rot}}^{\perp}(\theta)$	Entropy of region $I_{\text{rot}}^{\perp}(\theta)$
$\Delta E(\theta)$	Entropy gap between $I_{\text{rot}}(\theta)$ and $I_{\text{rot}}^{\perp}(\theta)$

Appendix B. Additional Algorithms

In this section, we present Algorithm 2, the Maximum Entropy Gap via Rotation Partitioning algorithm for jailbreak detection. Algorithm 2 represents a practical adaptation of Algorithm 1. Given that performing K trial iterations is undesirable, this version only requires iterating over angles from 0° to 180°. To streamline the process, each step is simplified to increments of 30°, while the remaining steps remain identical to those in Algorithm 1. The visualization result is shown in Figure 6.

B.1. Experimental Results on Trade-off between Jailbreakability and Stealthiness

In this section, we examine the linear relationship between the error probability lower bound P_e and the mutual information $I(X; Y_1, Y_2)$ in a simple scenario. We begin by selecting a

Algorithm 2: IEG Algorithm (Implementation based on Rotation Partitioning)

Input: Image $I = \{p_1, p_2, \ldots, p_n\}$ with pixel intensities in $[0, 255]$
Output: Maximum entropy gap $\Delta E_{\max}$Maximum entropy gap $\Delta E_{\max}$
Initialize: $\Delta E_{\max} \leftarrow 0$
for $\theta \in [0, 180°]$ **do**
  Partition I into $I_{\text{rot}}(\theta)$ and $I_{\text{rot}}^{\perp}(\theta)$ by a line at angle θ
  Calculate probability distribution $P(I_{\text{rot}}(\theta))$ for $I_{\text{rot}}(\theta)$
  Calculate probability distribution $P(I_{\text{rot}}^{\perp}(\theta))$ for $I_{\text{rot}}^{\perp}(\theta)$
  Compute entropy $E_{\text{rot}}(\theta) = -\sum_{x\in[0,255]} P(I_{\text{rot}}(\theta))(x) \log P(I_{\text{rot}}(\theta))(x)$
  Compute entropy $E_{\text{rot}}^{\perp}(\theta) = -\sum_{x\in[0,255]} P(I_{\text{rot}}^{\perp}(\theta))(x) \log P(I_{\text{rot}}^{\perp}(\theta))(x)$
  Compute entropy gap $\Delta E(\theta) = E_{\text{rot}}(\theta) - E_{\text{rot}}^{\perp}(\theta)$
  if $|\Delta E(\theta)| > |\Delta E_{max}|$ **then**
    $\Delta E_{\max} \leftarrow \Delta E(\theta)$
  end
  return $\Delta E_{\max}$
end

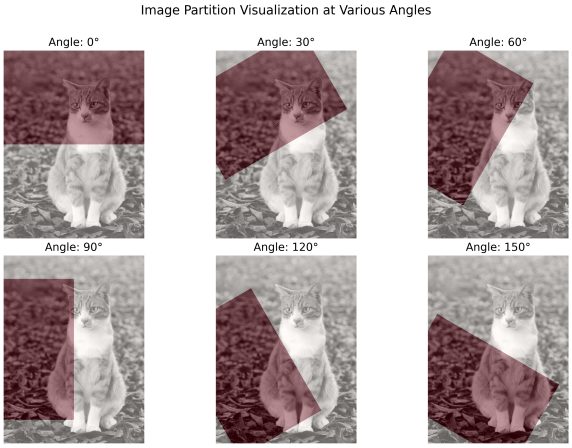


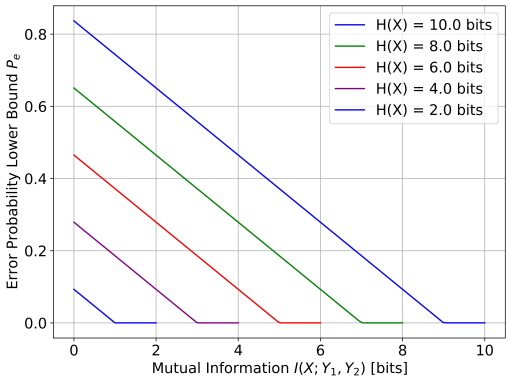
Figure 6: Visualization of Algorithm 2

jailbreak alphabet set from an online resource[1], which contains over 1,730 words and phrases considered inappropriate by Google, including curse words, insults, and vulgar language. This list is often used for profanity filters on websites and platforms.

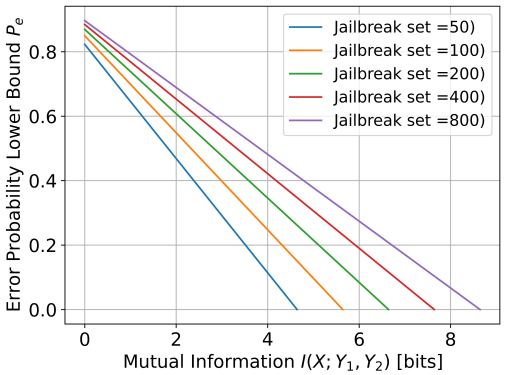
Next, we compute Eq. (3) from Theorem 2 to quantify the relationship. To illustrate the results, we choose several values of the entropy $H(X)$, ranging from 2 bits to 10 bits[2], and present the outcome in Figure 7(a). We make two key observations from Figure 7(a). First, as mutual information increases, the error probability decreases. Second, as $H(X)$ increases, the error probability rises, indicating that if jailbreak words are uniformly distributed, the jailbreak success rate tends to decrease. As illustrated in Figure 7(b), the cardinality of

1. https://www.freewebheaders.com/full-list-of-bad-words-banned-by-google/
2. With $|\mathcal{X}| = 1,730$ and so $\log|\mathcal{X}| \approx 10.76$.

the possible jailbreak alphabet sets varies, indicating that as the number of words on the blacklist increases, the jailbreak success rate decreases.



(*a*) Fano's Inequality curves for different $H(X)$.



(*b*) Fano's Inequality curves for different $|\mathcal{X}|$.

Figure 7: Fano's Inequality curves.

Appendix C. Additional experiments

We further evaluate on a recent open-source model, **DeepSeek-VL2-Small**. Results align with those on earlier architectures, corroborating the generality of IEG (Table 5).

Proprietary model note. On a held-out set of 100 samples, we observed an attack success rate of **0** on GPT-5. While this is a small-scale snapshot, it offers an additional indication that IEG can transfer across model families; a fuller evaluation on proprietary systems is left for future work.

Appendix D. Proof of Theorems

Proof of Theorem 2: The observations are (Y_1, Y_2), and the predictor $\hat{X} = M(Y_1, Y_2)$ is a deterministic function of these observations. Thus, $X \to (Y_1, Y_2) \to \hat{X}$ forms a Markov chain. By the data processing inequality for mutual information, $I(X; \hat{X}) \leq I(X; Y_1, Y_2)$. Since $H(X|V) = H(X) - I(X; V)$ for any variable V, this implies:

$$H(X|\hat{X}) = H(X) - I(X; \hat{X}) \geq H(X) - I(X; Y_1, Y_2) = H(X|Y_1, Y_2)$$

Fano's inequality, applied to the estimation of X by $\hat{X}$, states:

$$H(\mathrm{Ber}(P_e)) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X})$$

Combining this with $H(X|\hat{X}) \geq H(X|Y_1, Y_2)$, we directly obtain equation Eq. (4).

Equation Eq. (3) is another common form of Fano's inequality. It can be derived from $H(X|\hat{X}) \leq H(\mathrm{Ber}(P_e)) + P_e \log(|\mathcal{X}| - 1)$. Since $H(\mathrm{Ber}(P_e)) \leq 1$ (for $\log_2$), we have $H(X|\hat{X}) \leq 1 + P_e \log(|\mathcal{X}| - 1)$. A slightly looser but common upper bound is $H(X|\hat{X}) \leq$

Table 5: Post-defense ASR on DeepSeek-VL2-Small (lower is better). Numbers in parentheses denote absolute drops from the no-defense baseline.

Defense	Attack	ASR	
No Defense	FigStep	0.06–0.07	
	MM-SafetyBench	0.05–0.06	
	HADES	0.08–0.09	
JailGuard	FigStep	0.02	(0.04–0.05↓)
	MM-SafetyBench	0.02	(0.03–0.04↓)
	HADES	0.03	(0.05–0.06↓)
AdaShield-A	FigStep	0.01	(0.05–0.06↓)
	MM-SafetyBench	0.00	(0.05–0.06↓)
	HADES	0.00	(0.08–0.09↓)
MLLM-Protector	FigStep	0.01	(0.05–0.06↓)
	MM-SafetyBench	0.02	(0.03–0.04↓)
	HADES	0.01	(0.07–0.08↓)
IEG (Ours)	FigStep	0.01	(0.05–0.06↓)
	MM-SafetyBench	0.00	(0.05–0.06↓)
	HADES	0.00	(0.08–0.09↓)

$1 + P_e \log |\mathcal{X}|$. Rearranging this gives $P_e \geq \frac{H(X|\hat{X})-1}{\log |\mathcal{X}|}$. Using $H(X|\hat{X}) \geq H(X|Y_1, Y_2)$, we get:

$$P_e \geq \frac{H(X|Y_1, Y_2) - 1}{\log |\mathcal{X}|}$$

The second line of Eq. (3) follows from $H(X|Y_1, Y_2) = H(X) - I(X; Y_1, Y_2)$. ■

Note that Corollary 3 is strongly connected with Algorithm 1.

Proof of Corollary 3: From Theorem 2, the error probability P_e is lower bounded by a quantity that monotonically decreases as $I(X; Y_1, Y_2)$ increases:

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

Therefore, to minimize this lower bound on P_e (and thus to strive for the minimum achievable error P_e^*), we must maximize the term $I(X; Y_1, Y_2)$. The problem then becomes selecting or designing Y_1 and Y_2 (e.g., features, sensor data) such that $I(X; Y_1, Y_2)$ is maximized, while adhering to the given constraint (e.g., $I(X; Y_1) + I(X; Y_2) \leq C$). Maximizing $I(X; Y_1, Y_2)$ generally involves choosing Y_1 and Y_2 to provide complementary (synergistic or non-redundant) information about X. ■

Proof of Theorem 4: We have a Markov chain $X \to (Y_1, Y_2) \to Z \to \hat{X}$. From the data processing inequality, $I(X; \hat{X}) \leq I(X; Z) \leq I(X; Y_1, Y_2)$. This implies $H(X|\hat{X}) \geq H(X|Z) \geq H(X|Y_1, Y_2)$.

Applying the Fano inequality form from Eq. (3) to the estimation of X from Z (via $\hat{X} = M_2(Z)$):

$$P_e \geq \frac{H(X|\hat{X}) - 1}{\log |\mathcal{X}|}$$

Since $H(X|\hat{X}) \geq H(X|Z)$ (because $\hat{X}$ is a function of Z), we have:

$$P_e \geq \frac{H(X|Z) - 1}{\log |\mathcal{X}|}$$

We can rewrite $H(X|Z)$ using the definition of mutual information: $H(X|Z) = H(X) - I(X;Z)$. Also, $H(X|Y_1, Y_2) = H(X) - I(X; Y_1, Y_2)$. Substituting these into the RHS :

$$\begin{aligned}
&\frac{(H(X) - I(X;Y_1,Y_2)) - 1}{\log |\mathcal{X}|} + \frac{I(X;Y_1,Y_2) - I(X;Z)}{\log |\mathcal{X}|} \\
&= \frac{H(X) - I(X;Y_1,Y_2) - 1 + I(X;Y_1,Y_2) - I(X;Z)}{\log |\mathcal{X}|} \\
&= \frac{H(X) - I(X;Z) - 1}{\log |\mathcal{X}|} \\
&= \frac{H(X|Z) - 1}{\log |\mathcal{X}|}
\end{aligned}$$

This confirms that the two expressions for the lower bound are equivalent. The term $I(X;Y_1,Y_2) - I(X;Z)$ is non-negative due to the data processing inequality (Z is processed from Y_1, Y_2), and represents the information about X that is lost when passing from (Y_1, Y_2) to Z. ■

Proof of Proposition 5: (i) Since only Ω carries X-dependent signal, $I(X;Y_2) = I(X;Y_{2,\Omega})$ and the per-pixel MI adds up over Ω. (ii) By Fannes' continuity of entropy (discrete case), $|H(P_x) - H(P_0)| \leq h(\varepsilon) + \varepsilon \log(d-1)$ whenever $\|P_x - P_0\|_1/2 \leq \varepsilon \leq 1/2$. Thus $\Delta E \leq \tau$ implies $\|P_x - P_0\|_1 \leq 2f^{-1}(\tau, d)$ for all x. (iii) Under the minimal-mass condition, the local KL is upper-bounded by the squared ℓ_1 distance: $D_{\mathrm{KL}}(P_x \| P_0) \leq \frac{1}{\beta}\|P_x - P_0\|_2^2 \leq \frac{1}{\beta}\|P_x - P_0\|_1^2$. (iv) Standard channel decompositions yield $I(X;Y_{2,\Omega}) \leq \sum_{u \in \Omega} \mathbb{E}_X\big[D_{\mathrm{KL}}(P_x \| P_0)\big]$, hence $I(X;Y_2) \leq |\Omega| \cdot \frac{4}{\beta}\big(f^{-1}(\tau, d)\big)^2$, which matches the stated Φ up to a constant factor. ■

Consequence with Theorem 2. Combining Proposition 1 with the Fano-type lower bound, reducing the prediction error necessarily requires increasing $I(X;Y_1,Y_2)$; under the localized embedding model, this *cannot* be achieved while keeping ΔE arbitrarily small. This formalizes the success–stealthiness tension for structure-based attacks.

Theorem 6 (Detection Guarantee) *Let I be an image with adversarial modifications affecting at least α fraction of the image area. For any $\delta > 0$, if we set $K = \left\lceil \frac{\log(1/\delta)}{\alpha} \right\rceil$ random trials in Algorithm 1, then the probability of failing to detect the modification is at most δ.*

Proof For each random partition (R_1, R_2), the probability of the partition line intersecting the modified region is at least α. Therefore, the probability of missing the modification in a single trial is at most $(1-\alpha)$. After K independent trials, the probability of missing in all trials is at most $(1-\alpha)^K$. Setting $K = \left\lceil \frac{\log(1/\delta)}{\alpha} \right\rceil$ ensures:

$$\begin{aligned}(1-\alpha)^K &\leq \exp(-\alpha K) \\ &\leq \exp\left(-\alpha \cdot \frac{\log(1/\delta)}{\alpha}\right) \\ &= \exp(-\log(1/\delta)) \\ &= \delta\end{aligned}$$

This implies that with K trials, we detect the modification with probability at least $1-\delta$. ■

Corollary 7 (Practical Detection Bound) *For a desired confidence level of* 95% *($\delta = 0.05$) and assuming the adversarial modification affects at least* 10% *of the image ($\alpha = 0.1$), setting $K = 30$ trials is sufficient for reliable detection.*

Proof With $\alpha = 0.1$ and $\delta = 0.05$:

$$K = \left\lceil \frac{\log(1/0.05)}{0.1} \right\rceil = \left\lceil \frac{3}{0.1} \right\rceil = 30$$

■