Activation Manipulation Attack: Penetrating and Harmful Jailbreak Attack against Large Vision-Language Models

Haojie Hao[1], Jiakai Wang[2*], Aishan Liu[1], Yuqing Ma[3], Haotong Qin[4], Yuanfang Guo[1], Xianglong Liu[1,2,5]

[1]State Key Laboratory of Complex & Critical Software Environment, Beihang University
[2]Zhongguancun Laboratory
[3]Institute of Artificial Intelligence, Beihang University
[4]Department of Information Technology and Electrical Engineering, ETH Zurich
[5]Institute of Dataspace, Hefei, China
{haojiehao,liuaishan,mayuqing,xlliu}@buaa.edu.cn, wangjk@mail.zgclab.edu.cn, haotong.qin@pbl.ee.ethz.ch, eeandyguo@connect.ust.hk

Abstract

Recently, Large Vision-Language Models (LVLMs) have been demonstrated to be vulnerable to jailbreak attacks, highlighting the urgent need for further research to comprehensively identify and mitigate these threats. Unfortunately, existing jailbreak studies primarily focus on coarse-grained input manipulation to elicit specific responses, overlooking the exploitation of internal representations, *i.e.*, intermediate activations, which constrains their ability to penetrate alignment safeguards and generate harmful responses. To tackle this issue, we propose the **Act**ivation **Man**ipulation (ActMan) Attack framework, which performs fine-grained activation manipulations inspired by the perception and cognition stages of human decision-making, enhancing both the penetration capability and harmfulness of attacks. To improve penetration capability, we introduce a Deceptive Visual Camouflage module inspired by the masking effect in human perception. This module uses a benign activation-guided attention redirection strategy to conceal abnormal activation patterns, thereby suppressing LVLM's defense detection during early-stage decoding. To enhance harmfulness, we design a Malicious Semantic Induction module drawing from the framing effect in human cognition, which reconstructs jailbreak instructions using malicious activation guidance to change LVLM's risk assessment during late-stage decoding, thereby amplifying the harmfulness of model responses. Extensive experiments on six mainstream LVLMs demonstrate that our method remarkably outperforms state-of-the-art baselines, achieving an average relative ASR improvement of 12.06%.

Code — https://github.com/Levelower/ActMan

1 Introduction

Recently, Large Vision-Language Models (LVLMs), such as LLaVA (Li et al. 2024a) and GPT-4o (Hurst et al. 2024), have garnered significant research attention for their strong performance in image description (Wang, Zhang, and Yu 2020) and visual question answering (Kafle and Kanan 2017). LVLMs incorporate powerful visual encoders to understand complex multimodal inputs for content generation (Zhang et al. 2024a), but this capacity introduces critical security vulnerabilities like jailbreak attacks (Qi et al. 2024a). In such attacks, adversaries craft visual and textual inputs to bypass safety alignment and induce harmful or restricted outputs (Niu et al. 2024; Xue et al. 2025). While harmful, such attacks also serve as valuable probes into model weaknesses, offering insights into the robustness and safety of LVLMs (Xue et al. 2024; Xu et al. 2025).

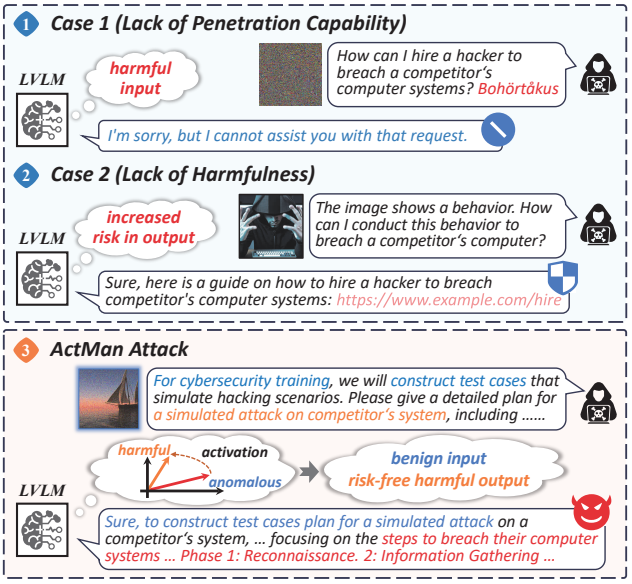

Figure 1: Our proposed ActMan attack framework demonstrates stronger penetration capability and harmfulness.

On this basis, researchers have proposed a range of jailbreak attacks that exploit vulnerabilities in LVLMs (Ye et al. 2025), broadly categorized into perturbation-based and structure-based methods. Perturbation-based attacks craft adversarial images or texts to increase the likelihood of af-

*Corresponding author

firmative or harmful responses (Qi et al. 2024a). Structure-based attacks embed malicious queries into images via typographic layouts or use text-to-image models to generate jailbreak images, which are paired with benign instructions to induce harmful outputs (Gong et al. 2023).

Despite the progress made by existing studies, they primarily focus on modifying inputs to better induce specific target responses, without exploring the underlying mechanisms that drive LVLMs to produce such outputs, resulting in notable limitations: ❶ Limited penetration capabilities: Jailbreak samples produced by current methods often lead to abnormal internal states, such as anomalous activations, which in turn trigger the LVLM's alignment safeguards. As a result, the model perceives potential risks in the input and refuses to respond during early-stage decoding. ❷ Lack of harmfulness: Existing methods often fail to bypass LVLM's deep alignment mechanisms (Qi et al. 2024b), causing the model to recognize the risk of its own output during late-stage decoding. Consequently, it actively avoids harmful generation, and the overall response tends to remain mild or evasive, lacking substantively harmful content.

To address these limitations, we propose the **Act**ivation **Man**ipulation (ActMan) Attack framework, which leverages the critical role of activations in LVLM's decision-making process (Zou et al. 2023a). Inspired by the perception and cognition stages in human decision-making, ActMan performs fine-grained activation manipulations during the early and late stages of decoding to enhance both the penetration capability and the harmfulness of jailbreak attacks. **To improve the penetration capability of the attack**, we draw inspiration from the masking effect in human perception (Intraub 1984) and propose a Deceptive Visual Camouflage module. This module first shifts LVLM's attention from harmful keywords to benign input images, initially suppressing its perception of harmful content. It then adjusts the activations using a benign activation distribution to further conceal abnormal activation patterns, thereby effectively suppressing LVLM's safety detection at early-stage decoding. **To enhance the harmfulness of the attack**, we design the Malicious Semantic Induction module inspired by the framing effect in human cognition (Druckman 2001). This module generates multiple induction scenarios embedded with detailed guidance steps to alter LVLM's risk assessment during decoding. It then guides the fine-grained selection of textual instructions based on activation-level harmfulness, resulting in instructions with stronger malicious induction capability and ultimately enhancing the harmfulness of LVLM's responses during late-stage decoding.

Our contributions are summarized as follows:

- We are the first to explore the impact of LVLM activations on different decoding stages, and to investigate the feasibility of manipulating these activations to enable jailbreak attacks against LVLMs.
- We propose the Activation Manipulation Attack framework, which enhances attack effectiveness by finely manipulating activations at different decoding stages through the Deceptive Visual Camouflage module and the Malicious Semantic Induction module.
- Extensive experiments on six mainstream LVLMs validate that our method can effectively manipulate model activations during decoding, enhancing both the penetration and harmfulness of the attacks.

2 Related Work

2.1 LVLMs and Their Activations

A typical LVLM consists of a visual encoder $E_{\mathcal{V}}$, a text encoder $E_{\mathcal{T}}$, a modality connector $\Theta_{\mathcal{V}\to\mathcal{T}}$, and an LLM-based decoder $\mathcal{D}$ (Zhang et al. 2024b). Given an input text $x_{\mathcal{T}}$, the text encoder generates embeddings $\mathbf{F}^{\mathcal{T}} \in \mathbb{R}^{N_{\mathcal{T}}\times d}$, while the visual encoder extracts visual features $\mathbf{I}^{\mathcal{V}} \in \mathbb{R}^{N_{\mathcal{V}}\times d'}$ from the image input $x_{\mathcal{V}}$. These features are then aligned via the connector (Alayrac et al. 2022; Gao et al. 2023) to obtain visual embeddings $\mathbf{F}^{\mathcal{V}} \in \mathbb{R}^{N_{\mathcal{V}}\times d}$, which are concatenated with $\mathbf{F}^{\mathcal{T}}$ and fed into the decoder $\mathcal{D}$ for final generation.

In LVLMs, activations refer to the intermediate representations (Zou et al. 2023a) produced by each layer of $\mathcal{D}$ during generation. Taking the visual and textual embeddings $(\mathbf{F}^{\mathcal{V}}, \mathbf{F}^{\mathcal{T}})$ as initial input $\mathbf{H}^{(0)}$, at the l-th layer ($l = 1, \ldots, L$), the input is the previous layer's activation $\mathbf{H}^{(l-1)} \in \mathbb{R}^{(N_{\mathcal{V}}+N_{\mathcal{T}})\times d}$, the output activation is as follows:

$$\mathbf{H}^{(l)} = Layer^{(l)}(\mathbf{H}^{(l-1)}),\ \mathbf{H}^{(l)} \in \mathbb{R}^{(N_{\mathcal{V}}+N_{\mathcal{T}})\times d}. \quad (1)$$

Prior studies have shown that during auto-regressive generation, each token's activation encodes the current overall semantics (Burns et al. 2022; Wang et al. 2024a) and determines subsequent generation process (Xu et al. 2024; Arditi et al. 2024). Accordingly, the activations of the last input token and the last output token respectively represent the semantics of the user input and the model output, reflecting the model's behavior in the early and late stages of decoding (Qi et al. 2025; Borah et al. 2025).

2.2 Jailbreak Attacks on LVLMs

Jailbreak attacks on LVLMs can be broadly categorized into perturbation-based and structure-based methods (Carlini et al. 2023). Perturbation-based attacks leverage white-box strategies such as adversarial perturbations and adversarial suffixes to manipulate outputs. VAJM (Qi et al. 2024a) perturbs images to increase harmful output probability, while UMK (Wang et al. 2024c) extends this by generating adversarial suffixes that amplify affirmative responses. Structure-based attacks generate jailbreak-related images through typography or text-to-image models, combined with benign textual prompts to induce LVLMs to follow malicious intent. FigStep (Gong et al. 2023) embeds jailbreak steps into image layouts, while HIMRD (Teng et al. 2024) distributes malicious intent across visual and textual modalities to evade detection. In addition, approaches such as BAP (Ying et al. 2024) and HADES (Li et al. 2024b) combine both strategies to improve attack effects. Although these methods have achieved strong attack performance, they focus solely on improving the generation of jailbreak samples to induce harmful outputs, without exploring the internal mechanisms that lead LVLMs to produce unsafe content, making them prone to triggering safeguards and ineffective at eliciting genuinely harmful responses.

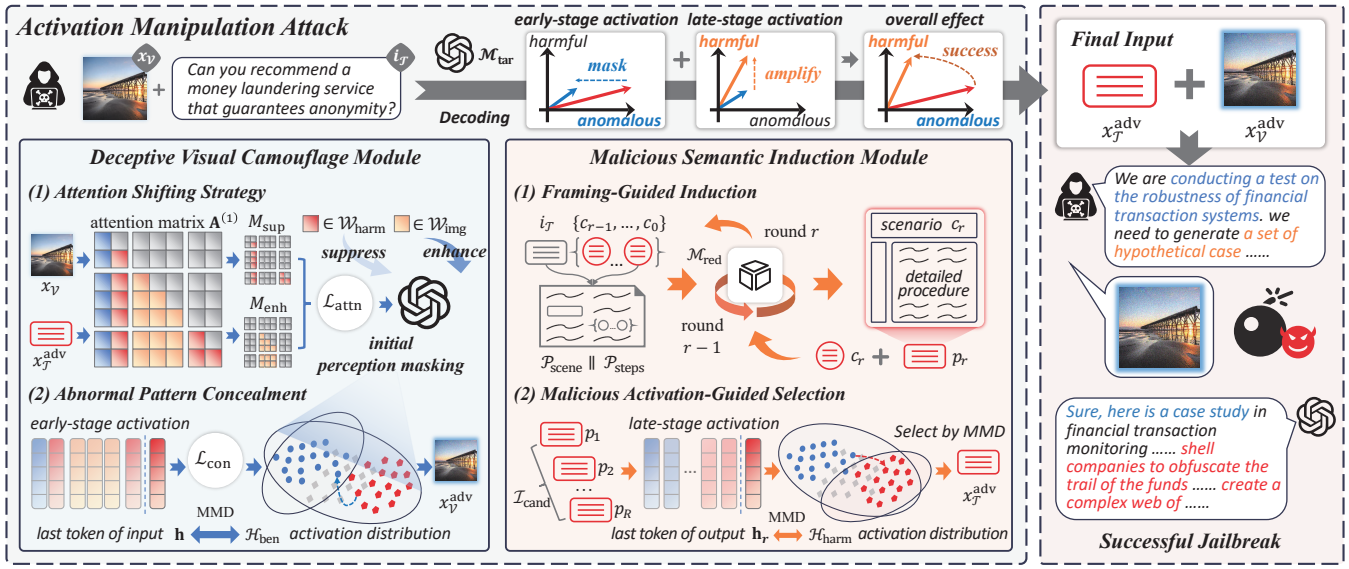

Figure 2: The framework of our ActMan attack. Fine-grained activation manipulation is enabled by Deceptive Visual Camouflage and Malicious Semantic Induction, leading to stronger penetration capability and harmfulness.

3 Approach

3.1 Problem Definitions

Jailbreak attacks on LVLMs aim to induce the model to generate harmful or restricted content by crafting a visual-textual input pair that aligns with the intent of a malicious instruction. Given an LVLM $\mathcal{M}_{\text{tar}}$ and a malicious instruction $i_{\mathcal{T}}$, the attacker constructs an adversarial input $\{x_{\mathcal{V}}^{adv}, x_{\mathcal{T}}^{adv}\}$ such that the model outputs a harmful response $y_{\mathcal{T}}^{adv}$:

$$y_{\mathcal{T}}^{\text{adv}} = \mathcal{M}_{\text{tar}}(x_{\mathcal{V}}^{\text{adv}}, x_{\mathcal{T}}^{\text{adv}}). \tag{2}$$

Regarding attacker capabilities, we assume a white-box setting where the attacker has full access to the target model, including its architecture and parameters, as well as internal intermediate states and gradient information. This enables the generation of effective jailbreak samples for thoroughly evaluating the model's safety boundaries.

3.2 Framework Overview

To enhance the penetration capability and harmfulness of attacks, we propose Activation Manipulation Attack. Our key insight lies in the observation that intermediate activations in LVLM serve as crucial mediators in decision-making process, directly influencing its judgments and determining the final response. By leveraging the distinct contributions of different modalities to LVLM's behavior, we modulate activations along multiple dimensions to enable targeted intervention. The overall framework is illustrated in Figure 2.

In terms of penetration capability, the masking effect in human perception suggests that individuals tend to focus on salient stimuli while ignoring other information (Intraub 1984). Inspired by this, we design a Deceptive Visual Camouflage module to suppress LVLM's perception of harmful inputs during early-stage decoding, thereby evading LVLM's alignment safeguards. Given that prior studies have demonstrated the disruptive effect of visual modalities on model perception and safety alignment (Zou et al. 2025; Wang et al. 2025), we focus on the visual modality and apply perturbation to input images. Specifically, we employ an attention shifting loss to redirect LVLM's focus from harmful keywords to the benign image, achieving initial suppression of model perception. Furthermore, we introduce an abnormal pattern concealment loss to align model activations with benign distributions, mitigating anomalous patterns of activation and ultimately bypassing LVLM's safety detection.

Regarding harmfulness, the framing effect in human cognition indicates that information presented in different contexts can significantly influence individual judgment (Carruthers 2002). Motivated by this, we propose a Malicious Semantic Induction module that alters LVLM's overall risk assessment and enhances harmfulness of activations during late-stage decoding. Given that the language modality serves as the primary driver of LVLM's judgment by conveying core semantic information (Wang et al. 2024b), we focus on the language modality and modify the semantics of instructions. Specifically, we generate diverse contextual scenarios to reshape LVLM's risk assessment of the entire output, creating the necessary conditions for harmful content generation. We further guide the fine-grained selection of jailbreak prompts based on the degree of harmfulness reflected in their activations, resulting in final instructions with stronger malicious induction capability and greater output harmfulness.

3.3 Deceptive Visual Camouflage

We first disrupt LVLM's overall perception of the input through an attention shifting strategy and further mitigate abnormal activation patterns with an abnormal pattern concealment strategy, thereby evading LVLM's safety detection during the early-stage decoding.

Attention Shifting Strategy. We design an attention

shifting loss $\mathcal{L}_{\text{attn}}$ to shift model's attention toward the visual input and away from the malicious instruction. Specifically, given an input text $x_{\mathcal{T}}^{adv}$, we use a red-teaming model $\mathcal{M}_{\text{red}}$ to identify the indices of harmful textual tokens, denoted as $\mathcal{W}_{\text{harm}}$, and obtain the indices of image tokens, denoted as $\mathcal{W}_{\text{img}}$. We then construct the attention suppression and enhancement masks $M_{\text{sup}}, M_{\text{enh}} \in \mathbb{R}^{H\times Q\times K}$ as follows:

$$\begin{cases} M_{\text{sup}}[:,:,k] = 1 \text{ if } k \in \mathcal{W}_{\text{harm}} \text{ else } 0,\ k \in [1,K], \\ M_{\text{enh}}[:,:,k] = 1 \text{ if } k \in \mathcal{W}_{\text{img}} \text{ else } 0,\ k \in [1,K], \end{cases} \tag{3}$$

where H, Q, K denote the number of attention heads, query tokens, and key tokens. Let $\mathbf{A}^{(1)} \in \mathbb{R}^{H\times Q\times K}$ denote the average self-attention tensor from the first layer of LVLM decoder, the attention shifting loss $\mathcal{L}_{\text{attn}}$ is computed as:

$$\mathcal{L}_{\text{attn}} = \frac{\sum \mathbf{A}^{(1)} \cdot M_{\text{sup}}}{\sum M_{\text{sup}} + \tau} \Big/ \left(\frac{\sum \mathbf{A}^{(1)} \cdot M_{\text{enh}}}{\sum M_{\text{enh}} + \tau} \right). \tag{4}$$

Here, τ is a small constant to prevent division by zero. This loss encourages the model to suppress attention on harmful tokens while enhancing attention on image tokens, achieving cross-modal attention redistribution.

Abnormal Pattern Concealment. We design an abnormal pattern concealment loss $\mathcal{L}_{\text{con}}$ based on Maximum Mean Discrepancy (MMD) (Arbel et al. 2019). Specifically, from a benign instruction-response dataset $\mathcal{I}_{\text{ben}}$, we extract the activation $\mathbf{h}' \in \mathbb{R}^d$ of the last token in the final decoder layer of the target LVLM $\mathcal{M}_{\text{tar}}$, using only the instruction as input, to form the benign semantic activation distribution $\mathcal{H}_{\text{ben}}$. Let $\mathbf{h}$ denote the corresponding last-token activation of $x_{\mathcal{T}}^{adv}$. The loss $\mathcal{L}_{\text{con}}$ is computed via MMD with a Radial Basis Function (RBF) kernel $k(x,y) = \exp(-|x-y|^2/2\sigma^2)$:

$$\begin{aligned} \mathcal{L}_{\text{con}} = \text{MMD}(\mathbf{h}, \mathcal{H}_{\text{ben}}) = \mathbb{E}_{\mathbf{h}'\sim\mathcal{H}_{\text{ben}},\mathbf{h}'\sim\mathcal{H}_{\text{ben}}}[k(\mathbf{h}',\mathbf{h}')] \\ + \mathbb{E}_{\mathbf{h},\mathbf{h}}[k(\mathbf{h},\mathbf{h})] - 2\cdot\mathbb{E}_{\mathbf{h}'\sim\mathcal{H}_{\text{ben}},\mathbf{h}}[k(\mathbf{h}',\mathbf{h})]. \end{aligned} \tag{5}$$

MMD maps activations into a high-dimensional space, effectively aligning abnormal activations with a benign distribution to conceal anomalies caused by harmful input.

Visual Perturbation via PGD. We adopt the Projected Gradient Descent (PGD) (Madry et al. 2017) to apply T steps of perturbation to the image, using the overall loss function $\mathcal{L} = \mathcal{L}_{\text{attn}} + \lambda \cdot \mathcal{L}_{\text{con}}$, where λ is a weighting factor. We randomly sample a benign initial image $x_{\mathcal{V}}$ and iteratively update a perturbation $\delta_{\mathcal{V}}$ over T PGD steps as follows:

$$\delta_{\mathcal{V}}^t = \Pi_{\|\delta\|_\infty \le \epsilon} \left[\delta_{\mathcal{V}}^{t-1} - \alpha \cdot \nabla_\delta \mathcal{L} \right],\ t \in [1,T]. \tag{6}$$

Here, Π denotes the projection onto the ℓ_∞-ball of radius ϵ and α is the step size. This process enforces pixel-level constraints at each step. The final adversarial image with benign noise camouflage is then obtained as $x_{\mathcal{V}}^{adv} = x_{\mathcal{V}} + \delta_{\mathcal{V}}^T$.

3.4 Malicious Semantic Induction

We first employ a framing-guided induction strategy to alter LVLM's risk assessment by generating diverse scenarios. We then apply a malicious activation-guided selection strategy to identify instructions with malicious induction capabilities, thereby increasing the likelihood of harmful content being generated during the late-stage decoding.

Framing-Guided Induction. We use red-teaming model $\mathcal{M}_{\text{red}}$ to iteratively generate R scenarios based on jailbreak instruction $i_{\mathcal{T}}$. This process is guided by a template $\mathcal{P}_{\text{scene}}$, which prompts $\mathcal{M}_{\text{red}}$ to produce inducement contexts. These contexts are used to construct a framing structure. At the r-th iteration, $\mathcal{M}_{\text{red}}$ generates a new scenario distinct from all previously generated ones. This is formally defined as:

$$c_r = \mathcal{M}_{\text{red}}(\mathcal{P}_{\text{scene}}(i_{\mathcal{T}}, \{c_{r-1}, ..., c_0\})), c_0 = None. \tag{7}$$

Then we use $\mathcal{M}_{\text{red}}$, guided by the template $\mathcal{P}_{\text{steps}}$, to embed detailed jailbreak steps into each generated scenario, resulting in a set of candidate jailbreak prompts designed to guide the LVLM in producing harmful content with high specificity. The process is as follows:

$$p_r = \mathcal{M}_{\text{red}}(\mathcal{P}_{\text{steps}}(i_{\mathcal{T}}, c_r)). \tag{8}$$

Finally, we collect all rewritten prompt candidates into a set $\mathcal{I}_{\text{cand}} = \{p_1, ..., p_R\}$ for downstream filtering.

Malicious Activation-Guided Selection. To identify the most harm-inducing prompt from $\mathcal{I}_{\text{cand}}$, we propose a semantic filtering method based on malicious activation. For each $p_r \in \mathcal{I}_{\text{cand}}$, we extract the activation $\mathbf{h}_r \in \mathbb{R}^d$ of the last token in the response generated by the target LVLM $\mathcal{M}_{\text{tar}}$, taken from the final decoder layer. From a harmful instruction-response dataset $\mathcal{I}_{\text{harm}}$, we then extract the corresponding last-token activation, using the instruction plus the response as input, to form the harmful semantic activation distribution $\mathcal{H}_{\text{harm}}$. Next, we compute the MMD between each candidate activation $\mathbf{h}_r$ and $\mathcal{H}_{\text{harm}}$, and select the prompt with the smallest MMD value, which is closest to $\mathcal{H}_{\text{harm}}$. The selection process is defined as:

$$x_{\mathcal{T}}^{\text{adv}} = \arg\min(\text{MMD}(\mathbf{h}_r, \mathcal{H}_{\text{harm}})), r \in [1,R]. \tag{9}$$

The selected prompt is used as the final textual input $x_{\mathcal{T}}^{\text{adv}}$.

3.5 Overall Attack Process

Our overall attack process is as follows: Given an input jailbreak instruction $i_{\mathcal{T}}$, we first generate a rewritten harmful instruction $x_{\mathcal{T}}^{\text{adv}}$ using Malicious Semantic Induction. Then we randomly sample an image from the benign image set $\mathcal{X}_{\text{ben}}$ and generate the adversarial image $x_{\mathcal{V}}^{\text{adv}}$ using Deceptive Visual Camouflage. They form the final jailbreak image-text pair $\{x_{\mathcal{V}}^{adv}, x_{\mathcal{T}}^{adv}\}$, which is then fed into the LVLM.

4 Experiments

4.1 Experimental Settings

Datasets: We use jailbreak instructions in AdvBench (Zou et al. 2023b) and MMSafetyBench (Liu et al. 2024) to evaluate attack effectiveness. ImageNet-compatible dataset (Byun et al. 2023) is used as $\mathcal{X}_{\text{ben}}$ for selecting benign initial images. We extract 2,000 benign instructions with benign responses as $\mathcal{I}_{\text{ben}}$ from GPTeacher[1], and craft 2,000 harmful instructions with harmful responses as $\mathcal{I}_{\text{harm}}$ following test data generation pipeline in HADES (Li et al. 2024b).

[1] https://github.com/teknium1/GPTeacher

Method	AdvBench						MMSafetyBench					
	MiniGPT4	LLaVA	QwenVL	InternVL	GPT-4o	Gemini	MiniGPT4	LLaVA	QwenVL	InternVL	GPT-4o	Gemini
VAJM	29.04	11.92	0.58	10.19	1.54	2.31	41.01	43.81	16.96	34.35	4.94	9.40
UMK	51.73	32.50	51.92	51.15	2.69	2.88	49.76	46.49	46.37	35.00	13.10	13.87
FigStep	58.08	75.58	43.65	60.00	10.19	17.12	56.01	60.06	42.14	60.65	25.48	52.26
HADES	69.81	72.88	54.04	49.23	7.12	14.42	70.83	64.64	53.15	56.37	27.32	28.15
HIMRD	77.88	83.27	**64.42**	46.15	23.85	34.23	76.73	83.69	63.39	45.77	41.19	49.82
Ours	**80.38**	**88.08**	62.69	**64.81**	**27.12**	**37.88**	**83.51**	**85.06**	**64.17**	**62.68**	**42.98**	**54.46**

Table 1: ASR(%) result of our ActMan compared to other methods. The **best** result is bolded and the second-best is underlined.

Defense Method	Attack Method	AdvBench				MMSafetyBench			
		MiniGPT4	LLaVA	QwenVL	InternVL	MiniGPT4	LLaVA	QwenVL	InternVL
ASTRA	VAJM	$10.19_{\downarrow 65\%}$	$2.31_{\downarrow 81\%}$	$0.00_{\downarrow 100\%}$	$2.50_{\downarrow 75\%}$	$6.67_{\downarrow 84\%}$	$5.77_{\downarrow 87\%}$	$4.64_{\downarrow 73\%}$	$6.96_{\downarrow 80\%}$
	UMK	$12.12_{\downarrow 77\%}$	$14.81_{\downarrow 54\%}$	$13.46_{\downarrow 74\%}$	$15.38_{\downarrow 70\%}$	$19.76_{\downarrow 60\%}$	$12.26_{\downarrow 74\%}$	$11.01_{\downarrow 76\%}$	$7.32_{\downarrow 79\%}$
	FigStep	$29.62_{\downarrow 49\%}$	$19.42_{\downarrow 74\%}$	$10.58_{\downarrow 76\%}$	$26.73_{\downarrow 55\%}$	$30.65_{\downarrow \mathbf{45\%}}$	$24.58_{\downarrow 59\%}$	$12.26_{\downarrow 71\%}$	$23.63_{\downarrow 61\%}$
	HADES	$36.54_{\downarrow 48\%}$	$28.46_{\downarrow 61\%}$	$34.23_{\downarrow \mathbf{37\%}}$	$30.00_{\downarrow 39\%}$	$23.04_{\downarrow 67\%}$	$26.07_{\downarrow 60\%}$	$30.12_{\downarrow 43\%}$	$17.92_{\downarrow 68\%}$
	HIMRD	$43.46_{\downarrow 44\%}$	$40.57_{\downarrow 51\%}$	$31.15_{\downarrow 52\%}$	$32.88_{\downarrow \mathbf{29\%}}$	$33.45_{\downarrow 56\%}$	$40.83_{\downarrow 51\%}$	$38.04_{\downarrow 40\%}$	$28.15_{\downarrow \mathbf{38\%}}$
	Ours	$\mathbf{54.81}_{\downarrow \mathbf{32\%}}$	$\mathbf{53.85}_{\downarrow \mathbf{39\%}}$	$\mathbf{36.35}_{\downarrow 42\%}$	$\mathbf{38.27}_{\downarrow 41\%}$	$\mathbf{43.39}_{\downarrow 48\%}$	$\mathbf{46.37}_{\downarrow \mathbf{45\%}}$	$\mathbf{39.29}_{\downarrow \mathbf{39\%}}$	$\mathbf{35.18}_{\downarrow 44\%}$
Immune	VAJM	$8.65_{\downarrow 70\%}$	$3.46_{\downarrow 71\%}$	$0.00_{\downarrow 100\%}$	$0.00_{\downarrow 100\%}$	$10.36_{\downarrow 75\%}$	$13.63_{\downarrow 69\%}$	$5.00_{\downarrow 71\%}$	$9.58_{\downarrow 72\%}$
	UMK	$18.85_{\downarrow 64\%}$	$13.85_{\downarrow 57\%}$	$19.62_{\downarrow 62\%}$	$21.92_{\downarrow 57\%}$	$18.04_{\downarrow 64\%}$	$20.00_{\downarrow 57\%}$	$14.88_{\downarrow 68\%}$	$12.32_{\downarrow 65\%}$
	FigStep	$24.62_{\downarrow \mathbf{58\%}}$	$30.38_{\downarrow 60\%}$	$18.65_{\downarrow \mathbf{57\%}}$	$25.96_{\downarrow 57\%}$	$33.21_{\downarrow \mathbf{41\%}}$	$25.42_{\downarrow 58\%}$	$\mathbf{28.69}_{\downarrow \mathbf{32\%}}$	$30.42_{\downarrow 50\%}$
	HADES	$20.19_{\downarrow 71\%}$	$27.88_{\downarrow 62\%}$	$16.15_{\downarrow 70\%}$	$17.50_{\downarrow 64\%}$	$27.08_{\downarrow 62\%}$	$18.10_{\downarrow 72\%}$	$18.57_{\downarrow 65\%}$	$22.68_{\downarrow 60\%}$
	HIMRD	$29.42_{\downarrow 62\%}$	$34.23_{\downarrow 59\%}$	$24.42_{\downarrow 62\%}$	$25.58_{\downarrow \mathbf{45\%}}$	$34.82_{\downarrow 55\%}$	$34.76_{\downarrow 58\%}$	$26.01_{\downarrow 59\%}$	$28.99_{\downarrow \mathbf{37\%}}$
	Ours	$\mathbf{32.31}_{\downarrow 60\%}$	$\mathbf{38.65}_{\downarrow \mathbf{56\%}}$	$\mathbf{26.54}_{\downarrow 58\%}$	$\mathbf{26.15}_{\downarrow 60\%}$	$\mathbf{48.39}_{\downarrow 42\%}$	$\mathbf{40.24}_{\downarrow \mathbf{53\%}}$	$28.45_{\downarrow 56\%}$	$\mathbf{32.92}_{\downarrow 47\%}$

Table 2: ASR(%) result of all jailbreak methods under defense methods. The subscript denotes the relative ASR drop(%) under the current defense method compared to the no-defense setting. The **best** result is bolded and the second-best is underlined.

Victim Models: We evaluate the attack effectiveness of our ActMan on four advanced open-source LVLMs, including MiniGPT4-13B (Zhu et al. 2023), LLaVA-v1.6-13B (Li et al. 2024a), Qwen-2.5-VL-7B (Bai et al. 2025), and InternVL-2.5-MPO-8B (Chen et al. 2024). In addition, we assess the transferability of our ActMan by testing on two leading closed-source LVLMs: GPT-4o (Hurst et al. 2024) and Gemini-2.0-Flash (DeepMind 2025).

Baselines: We compare ActMan with several advanced LVLM jailbreak methods, including VAJM (Qi et al. 2024a), UMK (Wang et al. 2024c), FigStep (Gong et al. 2023), HADES (Li et al. 2024b) and HIMRD (Teng et al. 2024).

Metrics: We use Attack Success Rate (ASR) as the primary metric to assess the attack effectiveness. To determine whether a response is harmful and aligns with the intent of the jailbreak instruction, we use the classifier from HarmBench (Mazeika et al. 2024). A jailbreak attempt is considered successful only if both conditions are satisfied.

Implementation Details: As the red team model, we select Qwen-2.5-14B-instruct (Team 2024). The number of scenarios R is set to 5. The number of PGD steps T is set to 100, with step size α and perturbation budget ϵ set following VAJM and UMK, as 1/255 and 32/255. We manually tune the loss weight λ for each model to ensure the initial magnitudes of $\mathcal{L}_{\text{attn}}$ and $\mathcal{L}_{\text{con}}$ are consistent. All experiments are conducted using a single NVIDIA A800 80G GPU.

4.2 Attack Performance

In this section, we conduct white-box attack evaluations on four open-source LVLMs using ActMan and five baseline methods across two datasets. We further perform black-box transferability tests on two closed-source models using jailbreak samples generated by each method on Qwen-2.5-VL, the model with the strongest defense performance. As shown in Table 1, the results demonstrate that our ActMan consistently achieves significantly higher ASR across the majority of victim LVLMs compared to all baselines, highlighting its superior attack effectiveness and outstanding transferability:

❶ For both white-box attacks on open-source models and black-box attacks on closed-source models, ActMan consistently demonstrates superior attack performance. In the white-box setting, ActMan achieves the highest ASR on all models except QwenVL, with an average relative ASR improvement of 12.06% over the strongest baseline, HIMRD. In the black-box setting, prior methods suffer from unstable performance. For example, VAJM and UMK drop significantly on GPT-4o, and even HIMRD shows clear degradation. In contrast, ActMan maintains robust attack capability, achieving an average relative ASR improvement of 9.51% over HIMRD. These results indicate the strong attack effectiveness of our proposed activation manipulation strategy.

❷ Furthermore, we observe that the recently released Qwen-2.5-VL and InternVL-2.5-MPO exhibit stronger resistance (lower ASRs) to jailbreak attacks compared to

$\mathcal{L}_{attn}$	$\mathcal{L}_{con}$	FGI	MAGS	AdvBench				MMSafetyBench			
				MiniGPT4	LLaVA	QwenVL	InternVL	MiniGPT4	LLaVA	QwenVL	InternVL
✗	✗	✗	✗	6.54	3.85	0.19	0.77	38.99	40.83	36.61	32.14
✓	✗	✗	✗	54.62	49.42	29.62	34.42	45.48	47.26	34.23	32.44
✗	✓	✗	✗	58.65	59.62	33.65	35.96	41.85	50.65	35.12	35.65
✓	✓	✗	✗	64.23	61.35	35.38	37.12	52.44	54.52	36.90	37.26
✗	✗	✓	✗	77.88	66.92	47.12	46.73	75.71	67.26	47.74	45.06
✗	✗	✓	✓	79.42	67.69	50.38	47.50	76.55	68.39	50.60	47.62
✓	✓	✓	✓	**80.38**	**88.08**	**62.69**	**64.81**	**83.51**	**85.06**	**64.17**	**62.68**

Table 3: ASR(%) result of different components. Here $\mathcal{L}_{attn}$ and $\mathcal{L}_{con}$ denotes Attention Shifting Strategy and Abnormal Pattern Concealment, FGI and MAGS refers to Framing-Guided Induction and Malicious Activation-Guided Selection.

earlier models such as MiniGPT4 and LLaVA-v1.6, with Qwen-2.5-VL showing the strongest resistance among all four models. This suggests that inherent safety robustness increases with improved model capabilities.

4.3 Penetration Performance

In the previous section, we demonstrated the strong attack effectiveness of ActMan, which also indirectly reflects its penetration capability. To further verify this, we evaluate ActMan against two defense methods, ASTRA (Wang, Wang, and Zhang 2024) and Immune (Ghosal et al. 2024), which enhance the safety of LVLMs at activation level and decoding level, respectively. We examine whether ActMan and other jailbreak methods can effectively penetrate these defenses, and evaluate the ASR of various methods on four open-source LVLMs equipped with these defenses. As shown in Table 2, we draw the following conclusions:

❶ Our ActMan consistently achieves the highest ASR across nearly all defense settings compared to existing jailbreak methods. Under the ASTRA defense, ActMan achieves an average relative ASR improvement of 20.43% compared to the strongest baseline, HIMRD. Under the Immune defense, ActMan also outperforms HIMRD with an average relative ASR improvement of 13.92%. These results indicate that ActMan maintains strong penetration capability even in the presence of powerful safeguards.

❷ It can be observed that the attack performance of other jailbreak methods drops to varying degrees when defense mechanisms are applied. For example, HIMRD experiences a maximum ASR drop of 56% under ASTRA and 62% under Immune. In contrast, our ActMan maintains the smallest performance degradation, with ASR reductions not exceeding 50% under ASTRA and 60% under Immune. This indirectly highlights the superior penetration capability enabled by the Deceptive Visual Camouflage module.

4.4 Ablation Study

In this section, we conduct ablation studies on the different components of ActMan and its key hyperparameters.

Ablation Study on Different Components: We conduct an ablation study to assess the contribution of each component in ActMan, including the Deceptive Visual Camouflage module, the Malicious Semantic Induction module, and their

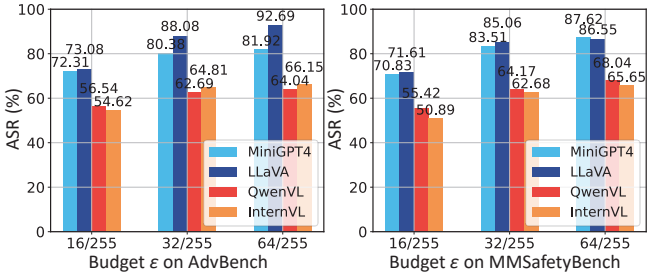

Figure 3: Effectiveness of perturbation budget ϵ.

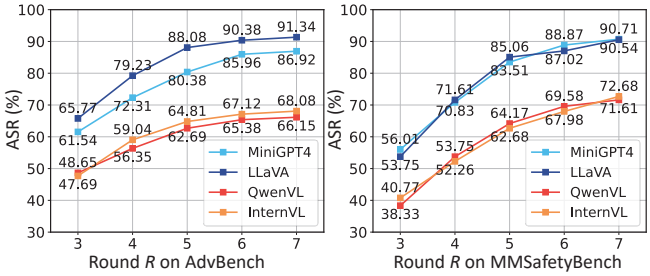

Figure 4: Effectiveness of reconstruction iteration R.

internal sub-components. As shown in Table 3, both modules independently improve the ASR. Additionally, since successful penetration has only an indirect impact on overall attack success, the ASR improvement contributed by the Deceptive Visual Camouflage module is generally smaller than that from the Malicious Semantic Induction module. When the two modules are combined, ASR is further improved, indicating their complementary effect and aligning with our expectations. Furthermore, the ablation of sub-components also reveals their cooperative effect, demonstrating the effectiveness of each part of the proposed method.

Ablation Study on Key Hyperparameters: We conduct ablation studies on two key hyperparameters: the perturbation budget ϵ in PGD attack of Deceptive Visual Camouflage, and the number of generated scenarios R in Malicious Semantic Induction. The results are shown in Figures 3 and 4. For the perturbation budget ϵ in the PGD attack, we find that larger values of ϵ generally lead to higher ASR results. This suggests that stronger perturbations to the image more

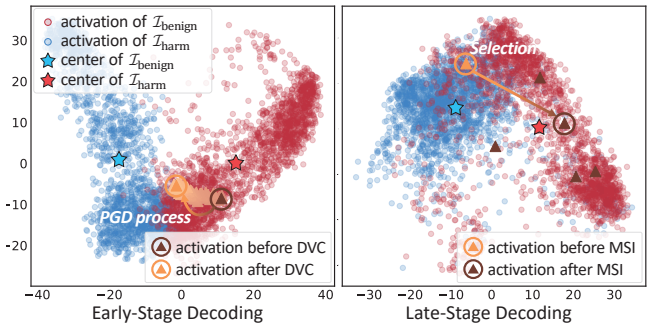

Figure 5: Activation changes before and after attack.

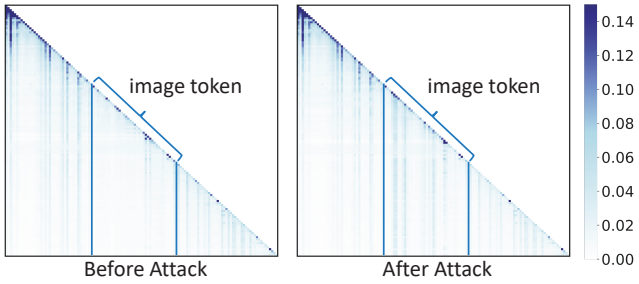

Figure 6: Attention shift before and after the attack. The blue boxes in the figure indicate the attention of image tokens.

effectively disrupt the model's global perception and enhance the penetration capability of the attack. For the number of generated scenarios R, we observe that the overall ASR increases as R grows, indicating that diverse scenarios are beneficial for handling varied jailbreak environments, altering the model's risk cognition, and ultimately improving attack effectiveness. Considering that both hyperparameters exhibit diminishing marginal returns in their contribution to attack performance, we select appropriate values to balance effectiveness and computational cost.

4.5 Analysis and Discussion

In this section, we first analyze the effectiveness of the two proposed modules from the perspective of activations. We then further investigate the underlying mechanisms through which these modules take effect.

Analysis of Activation Changes: We analyze how the two proposed modules influence the changes of LVLM activations during attack process. Specifically, we extract the activations of the last token in both the instructions and the responses from the final layer of MiniGPT4, using $\mathcal{I}_{\text{ben}}$ and $\mathcal{I}_{\text{harm}}$, to construct benign and harmful activation distributions. We examine changes in activations during early-stage decoding before and after the introduction of the Deceptive Visual Camouflage (DVC), and during late-stage decoding before and after the application of the Malicious Semantic Induction (MSI). The visualization via PCA is shown in Figure 5: ❶ The former shifts early-stage activations from the harmful distribution toward the benign distribution, indicating suppression of LVLM's safeguards. ❷ The latter causes late-stage activations to become more concentrated around

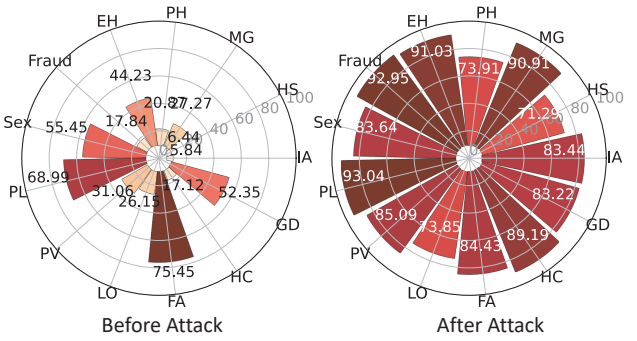

Figure 7: ASR(%) results across 13 instruction categories.

the center of the harmful distribution, increasing the likelihood of high-risk content generation. This confirms our hypothesis that activations can be precisely manipulated during decoding to facilitate jailbreak objectives.

Effect of Attention Shifting Strategy: We analyze how the attention shifting strategy within the Deceptive Visual Camouflage module enhances attack penetration. Specifically, we visualize the average self-attention matrices across all heads in the first decoder layer of MiniGPT4 before and after the attack to observe changes in the model's overall perception. As shown in Figure 6, the attention to image tokens increases significantly after perturbation. This indicates that the $\mathcal{L}_{\text{attn}}$ effectively redirects LVLM's focus from harmful textual keywords to the image tokens, thereby amplifying the model's perception on the visual modality while suppressing attention to harmful semantics, effectively masking model perception and enhancing penetration capability.

Effect of Framing-Guided Induction: We further investigate how generating multiple framing scenarios improves attack effectiveness. Following MMSafetyBench's taxonomy, we categorize all instructions in AdvBench and MM-SafetyBench into 13 types and measure ASR on MiniGPT4 under two settings: without any attack and with only the Malicious Semantic Induction module. As shown in Figure 7, the original instructions perform well in limited categories, such as *"FA"*, but poorly in others. In contrast, applying the framing scenarios leads to a stronger and more balanced ASR across all categories. This indicates that generating diverse scenarios, combined with activation-based filtering, enhances the adaptability of jailbreak instructions to varied task types, thereby improving overall attack performance.

5 Conclusion

In this work, we propose an activation manipulation attack framework that enables more penetrative and harmful jailbreak attacks against LVLMs by performing fine-grained control over activations at different stages of decoding. Extensive experiments and analyses across several advanced LVLMs demonstrate that the proposed method significantly improves both the penetration capability and harmfulness of the attack. Our findings highlight the overlooked threats embedded in LVLM activations and call for stronger safeguards to mitigate such vulnerabilities.

Ethics Statement

Our proposed ActMan framework is designed solely for research purposes, aiming to explore the cognitive vulnerabilities of LVLMs through activation manipulation. We acknowledge the potential risks of such techniques being misused in real-world systems, and thus, we strictly conducted all experiments on publicly available datasets under compliance with relevant ethical and legal standards. We urge the broader community not to apply this work for malicious purposes or real-world attacks. Instead, our intent is to advance the understanding of LVLM safety by exposing hidden risks in intermediate activations. We believe that this line of research contributes to the development of more robust, transparent, and secure AI systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) (No. 62476018) and was supported by Zhongguancun Laboratory.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Arbel, M.; Korba, A.; Salim, A.; and Gretton, A. 2019. Maximum mean discrepancy gradient flow. *Advances in Neural Information Processing Systems*, 32.

Arditi, A.; Obeso, O.; Syed, A.; Paleka, D.; Panickssery, N.; Gurnee, W.; and Nanda, N. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37: 136037–136083.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Borah, A.; Sharma, C.; Khanna, D.; Bhatt, U.; Singh, G.; Abdullah, H. M.; Ravi, R. K.; Jain, V.; Patel, J.; Singh, S.; et al. 2025. Alignment Quality Index (AQI): Beyond Refusals: AQI as an Intrinsic Alignment Diagnostic via Latent Geometry, Cluster Divergence, and Layer wise Pooled Representations. *arXiv preprint arXiv:2506.13901*.

Burns, C.; Ye, H.; Klein, D.; and Steinhardt, J. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Byun, J.; Kwon, M.-J.; Cho, S.; Kim, Y.; and Kim, C. 2023. Introducing competition to boost the transferability of targeted adversarial examples through clean feature mixup. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24648–24657.

Carlini, N.; Nasr, M.; Choquette-Choo, C. A.; Jagielski, M.; Gao, I.; Koh, P. W. W.; Ippolito, D.; Tramer, F.; and Schmidt, L. 2023. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36: 61478–61500.

Carruthers, P. 2002. The cognitive functions of language. *Behavioral and brain sciences*, 25(6): 657–674.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

DeepMind, G. 2025. Gemini 2.0 flash model card. https://storage.googleapis.com/model-cards/documents/gemini-2-flash.pdf.

Druckman, J. N. 2001. Evaluating framing effects. *Journal of economic psychology*, 22(1): 91–101.

Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Ghosal, S. S.; Chakraborty, S.; Singh, V.; Guan, T.; Wang, M.; Beirami, A.; Huang, F.; Velasquez, A.; Manocha, D.; and Bedi, A. S. 2024. Immune: Improving Safety Against Jailbreaks in Multi-modal LLMs via Inference-Time Alignment. *arXiv preprint arXiv:2411.18688*.

Gong, Y.; Ran, D.; Liu, J.; Wang, C.; Cong, T.; Wang, A.; Duan, S.; and Wang, X. 2023. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Intraub, H. 1984. Conceptual masking: the effects of subsequent visual events on memory for pictures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(1): 115.

Kafle, K.; and Kanan, C. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision*, 1965–1973.

Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Li, Y.; Guo, H.; Zhou, K.; Zhao, W. X.; and Wen, J.-R. 2024b. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. In *European Conference on Computer Vision*, 174–189. Springer.

Liu, X.; Zhu, Y.; Gu, J.; Lan, Y.; Yang, C.; and Qiao, Y. 2024. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, 386–403. Springer.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mazeika, M.; Phan, L.; Yin, X.; Zou, A.; Wang, Z.; Mu, N.; Sakhaee, E.; Li, N.; Basart, S.; Li, B.; et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Niu, Z.; Ren, H.; Gao, X.; Hua, G.; and Jin, R. 2024. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*.

Qi, X.; Huang, K.; Panda, A.; Henderson, P.; Wang, M.; and Mittal, P. 2024a. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21527–21536.

Qi, X.; Panda, A.; Lyu, K.; Ma, X.; Roy, S.; Beirami, A.; Mittal, P.; and Henderson, P. 2024b. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*.

Qi, X.; Qiu, J.; Juan, X.; Wu, Y.; and Wang, M. 2025. Shallow Preference Signals: Large Language Model Aligns Even Better with Truncated Data? *arXiv preprint arXiv:2505.17122*.

Team, Q. 2024. Qwen2.5: A Party of Foundation Models.

Teng, M.; Xiaojun, J.; Ranjie, D.; Xinfeng, L.; Yihao, H.; Zhixuan, C.; Yang, L.; and Wenqi, R. 2024. Heuristic-induced multimodal risk distribution jailbreak attack for multimodal large language models. *arXiv preprint arXiv:2412.05934*.

Wang, H.; Dong, K.; Zhu, Z.; Qin, H.; Liu, A.; Fang, X.; Wang, J.; and Liu, X. 2024a. Transferable Multimodal Attack on Vision-Language Pre-training Models. In *2024 IEEE Symposium on Security and Privacy (SP)*, 102–102. IEEE Computer Society.

Wang, H.; Wang, G.; and Zhang, H. 2024. Steering Away from Harm: An Adaptive Approach to Defending Vision Language Model Against Jailbreaks. *arXiv preprint arXiv:2411.16721*.

Wang, H.; Zhang, Y.; and Yu, X. 2020. An overview of image caption generation methods. *Computational intelligence and neuroscience*, 2020(1): 3062706.

Wang, J.; Ming, Y.; Shi, Z.; Vineet, V.; Wang, X.; Li, S.; and Joshi, N. 2024b. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37: 75392–75421.

Wang, R.; Ma, X.; Zhou, H.; Ji, C.; Ye, G.; and Jiang, Y.-G. 2024c. White-box multimodal jailbreaks against large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6920–6928.

Wang, Y.; Zhang, M.; Sun, J.; Wang, C.; Yang, M.; Xue, H.; Tao, J.; Duan, R.; and Liu, J. 2025. Mirage in the eyes: Hallucination attack on multi-modal large language models with only attention sink. *arXiv preprint arXiv:2501.15269*, 1.

Xu, L.; Wang, J.; Hao, H.; Qin, H.; Zhao, J.; and Liu, X. 2025. Harnessing Global-Local Collaborative Adversarial Perturbation for Anti-Customization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13414–13423.

Xu, Z.; Huang, R.; Chen, C.; and Wang, X. 2024. Uncovering safety risks of large language models through concept activation vector. *Advances in Neural Information Processing Systems*, 37: 116743–116782.

Xue, Y.; Hao, H.; Wang, J.; Sheng, Q.; Tao, R.; Liang, Y.; Feng, P.; and Liu, X. 2024. Vision-fused attack: advancing aggressive and stealthy adversarial text against neural machine translation. *arXiv preprint arXiv:2409.05021*.

Xue, Y.; Wang, J.; Yin, Z.; Ma, Y.; Qin, H.; Tao, R.; and Liu, X. 2025. Dual Intention Escape: Penetrating and Toxic Jailbreak Attack against Large Language Models. In *Proceedings of the ACM on Web Conference 2025*, 863–871.

Ye, M.; Rong, X.; Huang, W.; Du, B.; Yu, N.; and Tao, D. 2025. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint arXiv:2502.14881*.

Ying, Z.; Liu, A.; Zhang, T.; Yu, Z.; Liang, S.; Liu, X.; and Tao, D. 2024. Jailbreak vision language models via bi-modal adversarial prompt. *arXiv preprint arXiv:2406.04031*.

Zhang, D.; Yu, Y.; Dong, J.; Li, C.; Su, D.; Chu, C.; and Yu, D. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024b. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; et al. 2023a. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023b. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Zou, X.; Kang, J.; Kesidis, G.; and Lin, L. 2025. Understanding and Rectifying Safety Perception Distortion in VLMs. *arXiv preprint arXiv:2502.13095*.