UNDERSTANDING AND RESTORING SAFETY NEURONS IN VLMS AGAINST MULTIMODAL JAILBREAKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) inherit safety alignment from their language backbones, yet visual inputs can undermine this alignment and enable harmful requests to bypass refusal behavior. We uncover a previously uncharacterized neuron-level phenomenon, termed *vision-induced safety-neuron suppression*, in which visual inputs may suppress neurons associated with harmful-request refusal and shift their activation patterns toward those induced by benign inputs. Motivated by this finding, we introduce **VISAR** (**V**isual **I**ntervention for **S**afety-neuron **A**ctivation **R**estoration), a lightweight visual intervention framework that uses a differentiable safety probe to supervise a plug-and-play SuffixViT for generating corrective visual suffixes, restoring the suppressed refusal-aligned safety state without modifying the pretrained VLM. Experiments across multiple VLMs and three multimodal jailbreak attacks, compared with four representative defenses, show that VISAR substantially reduces harmful responses while preserving benign utility. Mechanistic analyses further confirm that improved safety is accompanied by restoration of the refusal-aligned safety-neuron activation pattern, demonstrating that visual-side restoration of suppressed safety states provides an effective approach to mitigating multimodal safety degradation. Our code is available at https://anonymous.4open.science/r/VISAR-18D5

1 INTRODUCTION

Vision-language models (VLMs) have achieved strong performance across visual question answering, image understanding, and multimodal reasoning. However, introducing the visual modality also expands the attack surface, allowing harmful intent to be conveyed through images and bypass safety mechanisms inherited from the language backbone (Liu et al., 2024b; Luo et al., 2024; Gong et al., 2025). Recent multimodal jailbreak studies further show that safety-aligned VLMs can remain vulnerable when harmful semantics are transferred from the textual modality to the visual modality (Liu et al., 2024b; Gong et al., 2025).

To mitigate these risks, existing defenses intervene at different stages of the multimodal pipeline. Wang et al. (2024) use defensive prompting to identify and mitigate potentially malicious multimodal inputs, while Gou et al. (2024) transform unsafe visual information into text to reactivate the safety mechanism of the underlying language model. While these approaches can substantially improve robustness, relatively little attention has been paid to how visual inputs alter the internal safety state responsible for refusal. This motivates a more fundamental question: *What happens to the internal functional units that support safety refusal when harmful intent is conveyed through the visual modality?*

To investigate this question, we conduct a neuron-level analysis of the language backbone and identify a previously uncharacterized phenomenon, which we term *vision-induced safety-neuron suppression*. We first identify safety neurons associated with refusal behavior by contrasting harmful and benign text-only inputs(Chen et al., 2025b; Zhao et al., 2025). We observe that adding visual inputs to harmful requests substantially suppresses these neurons, shifting their activation patterns away from the refusal-aligned state and toward those induced by benign inputs(Li et al., 2025). This effect becomes more pronounced when harmful intent is conveyed primarily through the visual modality, suggesting that multimodal safety degradation is accompanied by a systematic disruption of the internal state supporting refusal(Zou et al., 2025). These observations motivate a visual-side

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

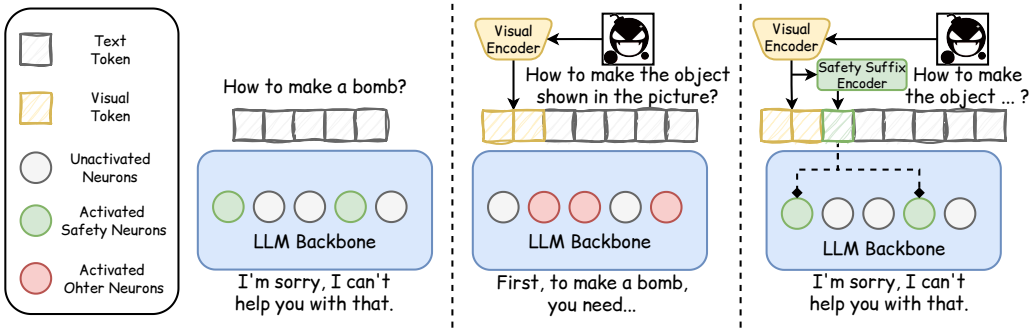

Figure 1: Illustration of vision-induced safety-neuron suppression and VISAR. Harmful text-only inputs activate refusal-related safety neurons, whereas visually conveyed harmful intent suppresses these activations and may induce unsafe responses. VISAR introduces a learned visual safety suffix to restore the suppressed safety state and recover refusal behavior.

safety-state restoration paradigm: rather than retraining the pretrained VLM or filtering unsafe outputs after generation, we aim to restore the existing refusal-aligned state through the visual pathway while keeping the pretrained model frozen. This formulation introduces two key challenges.

Challenge I: How can visual inputs be optimized to restore the internal safety state without modifying the pretrained VLM? Safety-aligned neurons primarily reside in the language backbone, whereas intervention is restricted to the visual side. This creates a mismatch between the target safety state and the optimization space, requiring a differentiable objective that connects visual modifications to the restoration of language-side safety neurons while keeping all pretrained parameters frozen(Li et al., 2026). **Challenge II: How can safety-state restoration preserve benign multimodal utility?** Restoring refusal-aligned safety states should not simply increase refusal behavior. Excessive intervention may cause false refusals or distort image semantics. Thus, the visual intervention must restore the desired safety state while preserving the model's original benign behavior and multimodal utility(Zhao et al., 2026b).

To address these challenges, we propose **VISAR** (**V**isual **I**ntervention for **S**afety-neuron **A**ctivation **R**estoration), a lightweight visual-side intervention framework that restores the refusal-aligned internal safety state suppressed by visual inputs without modifying the pretrained VLM. VISAR first identifies safety neurons associated with refusal behavior by contrasting harmful and benign text-only inputs. Based on these neurons, we construct a differentiable safety probe that characterizes the discrepancy between the model's current internal state and the target refusal-aligned state. We then introduce a plug-and-play *SuffixViT*, which generates corrective visual suffixes under the gradient supervision provided by the safety probe and appends them to the original visual representations. The visual suffix does not modify the visual encoder, projection layer, or language backbone. Instead, it steers the VLM's internal activations through the visual pathway toward the refusal-aligned safety state. Meanwhile, we impose a benign-behavior constraint to limit the impact of the intervention on normal inputs. By transferring safety-state restoration from the model parameter space to the visual input space, VISAR provides a new way to mitigate multimodal safety degradation while keeping the pretrained VLM intact. Our contributions are summarized as follows:

• We uncover *vision-induced safety-neuron suppression*, showing that visual inputs suppress refusal-related safety neurons and shift their activation patterns toward benign-response states, especially when harmful intent is primarily conveyed through the visual modality.

• We propose **VISAR**, a lightweight visual-side intervention framework that uses a differentiable safety-state probe and *SuffixViT* to restore the refusal-aligned internal state while keeping the pretrained VLM frozen and preserving benign utility.

• We evaluate VISAR across multiple VLMs and representative multimodal jailbreak benchmarks. VISAR substantially reduces harmful responses while preserving benign multimodal performance, and mechanistic analyses further confirm the restoration of suppressed safety-neuron activations.

2 Related Work

Multimodal Jailbreak Benchmarks. The visual channel introduces an additional attack surface that can bypass safety behavior effective in text-only settings. FigStep renders harmful instructions as typographic images and pairs them with benign-looking textual prompts (Gong et al., 2025), while MM-SafetyBench constructs query-relevant, typographic, and combined visual attacks with indirectly referring prompts (Liu et al., 2024b). JailBreakV-28K further aggregates transferred textual jailbreaks and diverse image-based attack settings (Luo et al., 2024).

Understanding VLM Jailbreaks from a Representation Perspective. Recent studies analyze VLM safety failures through internal representations and inference-time interventions. Ding et al. (2025) assess the safety of visual inputs and responses and steer generation toward safer outputs, while Liu et al. (2024a) attribute safety degradation to a cross-modal representation gap and mitigate it through representation adjustment. Li et al. (2025) revise image-induced activation discrepancies at the layer and attention-head levels, and Zou et al. (2025) calibrate modality-induced safety shifts that make harmful inputs appear safer. More recently, Wei et al. (2026) separate jailbreak and refusal states and restore refusal behavior by removing image-induced jailbreak directions. Unlike these representation-level approaches, our work focuses on how visual jailbreak inputs affect individual safety-related MLP neurons.

Neuron-Level Understanding of Safety Alignment. Recent studies investigate model safety at the neuron level by identifying sparse safety-related neurons and validating or tuning their effects on refusal behavior (Chen et al., 2025b; Zhao et al., 2025). Other works analyze how jailbreak attacks disrupt safety-related neurons in language models (Zhao et al., 2026a; Zhang et al., 2025). In VLMs, recent approaches further localize jailbreak- or safety-related neurons for detection and targeted alignment (Gan et al., 2026; Shi et al., 2026; Behrouzi et al., 2026). Unlike these methods, our work studies how visual harmful inputs suppress refusal-aligned MLP neurons and restores their activations through the visual feature stream while keeping the base VLM frozen.

3 Neuron-Level Analysis of Multimodal Safety Degradation

In this section, we analyze how visual inputs affect the refusal-aligned state of the language backbone. We identify refusal-related MLP neurons from harmful text-only inputs and show that visual inputs suppress their activations, especially when harmful intent is conveyed through the image.

3.1 Multimodal Conditions and Activation Extraction

Let f denote a VLM built upon a safety-aligned language backbone. We construct six conditions by combining $I \in \{\varnothing, I_{\text{ben}}, I_{\text{harm}}\}$ with $T \in \{T_{\text{ben}}, T_{\text{harm}}\}$. Harmful text-only inputs $[\varnothing, T_{\text{harm}}]$ that elicit explicit refusal are used as the reference refusal-aligned state. Based on the same harmful instructions, $[I_{\text{ben}}, T_{\text{harm}}]$ adds benign images, while $[I_{\text{harm}}, T_{\text{harm}}]$ uses FLUX.1 (Labs et al., 2025) to generate harmful images with the original instructions rendered beneath them. For $[I_{\text{harm}}, T_{\text{ben}}]$, the same harmful images are paired with a benign prompt asking the model to follow the instruction shown in the image. For $[I_{\text{ben}}, T_{\text{ben}}]$, benign images are generated from the corresponding benign text using FLUX.1 and paired with the original benign instruction. The remaining benign text-only condition serves as an additional control.

We analyze intermediate MLP activations in the language backbone. For hidden state $\mathbf{h}_t^{(l)}$, the gated MLP activation is

$$\mathbf{a}_t^{(l)} = \phi\Big(\mathbf{W}_{\text{gate}}^{(l)}\mathbf{h}_t^{(l)}\Big) \odot \Big(\mathbf{W}_{\text{up}}^{(l)}\mathbf{h}_t^{(l)}\Big). \tag{1}$$

Each dimension is treated as an MLP neuron, and activations are extracted at the final input position before response generation for subsequent safety-neuron analysis.

3.2 Identifying Safety Neurons

We identify refusal-related MLP neurons using the harmful and benign text-only conditions, $[\varnothing, T_{\text{harm}}]$ and $[\varnothing, T_{\text{ben}}]$. We retain harmful samples that elicit explicit refusal responses and benign samples that elicit normal non-refusal responses.

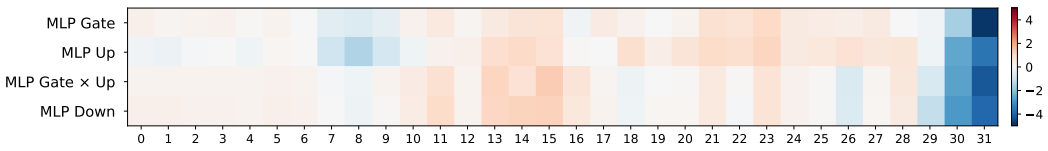

Figure 2: Layer-wise harmful–benign activation separation across different MLP components in LLaVA-1.5-7B. The layer with the strongest separation in the Gate × Up activations is selected for safety-neuron localization.

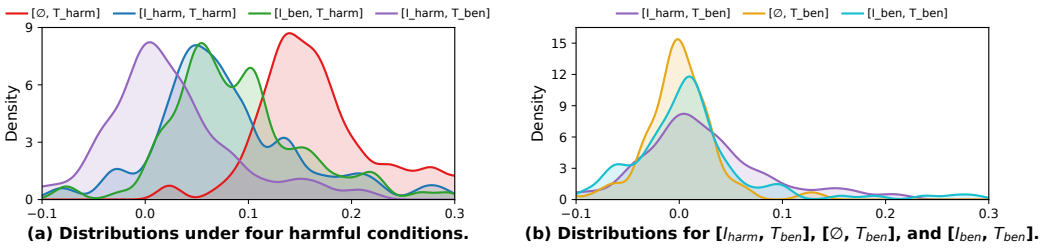

Figure 3: Distributions of mean safety-neuron activations across input conditions. (a) Comparison among harmful input conditions. (b) Comparison between harmful image–benign text and benign conditions.

For each neuron j in layer l, we measure the activation difference between the two groups using Welch's t-statistic:

$$t_j^{(l)} = \frac{\mu_{\text{harm},j}^{(l)} - \mu_{\text{ben},j}^{(l)}}{\sqrt{(\sigma_{\text{harm},j}^{(l)})^2/N_{\text{harm}} + (\sigma_{\text{ben},j}^{(l)})^2/N_{\text{ben}}}}. \tag{2}$$

Here, μ, σ^2, and N denote the sample mean, variance, and number of inputs in each condition. A larger positive $t_j^{(l)}$ indicates stronger activation under refusal-triggering harmful text.

For each layer, we compute the mean of its top-K positive t-statistics and select the layer with the largest separation score in the intermediate MLP activations (Gate × Up) as $l^\star$, as shown in Fig. 2. The safety-neuron set is then defined as

$$\mathcal{S} = \underset{j:\, t_j^{(l^\star)} > 0}{\arg\,\text{topK}}\, t_j^{(l^\star)}. \tag{3}$$

For an input $\mathbf{x}$, the selected activations are denoted by $\mathbf{z}(\mathbf{x}) = [a_j^{(l^\star)}(\mathbf{x})]_{j\in\mathcal{S}} \in \mathbb{R}^K$.

3.3 Vision-Induced Safety-Neuron Suppression

Having identified the safety neurons from text-only inputs, we keep the selected set $\mathcal{S}$ fixed and examine its responses under different image–text combinations. For each input $\mathbf{x}$, we compute the sample-wise mean safety-neuron activation as

$$A_{\mathcal{S}}(\mathbf{x}) = \frac{1}{K}\sum_{j\in\mathcal{S}} a_j^{(l^\star)}(\mathbf{x}), \tag{4}$$

and compare the distributions of $A_{\mathcal{S}}(\mathbf{x})$ across input conditions.

As shown in Fig. 3(a), visual inputs shift safety-neuron activations toward lower values, with the strongest suppression when harmful intent is primarily conveyed through the image. Fig. 3(b) further shows that $[I_{\text{harm}}, T_{\text{ben}}]$ overlaps substantially with benign inputs, indicating a shift toward the benign activation regime.

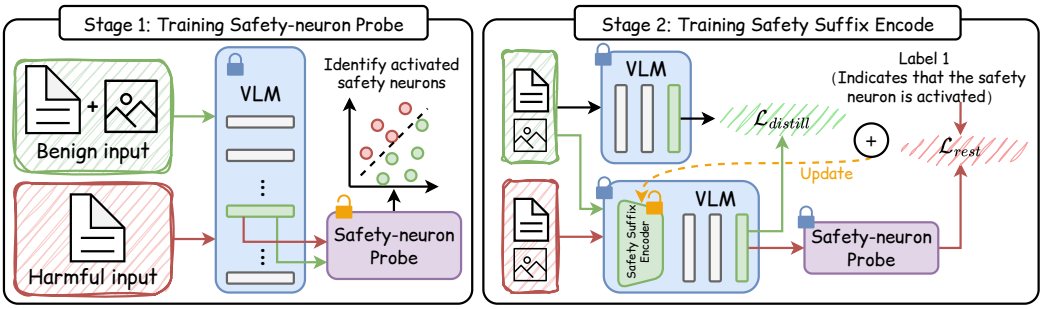

Figure 4: Overview of the VISAR training framework. Stage 1 identifies refusal-related safety neurons from benign and harmful text-only inputs and trains a safety-state probe. Stage 2 uses the probe to supervise a trainable safety suffix encoder, combining safety-state restoration on harmful inputs with benign self-distillation, while keeping the pretrained VLM frozen.

We term this phenomenon *vision-induced safety-neuron suppression*: visual inputs weaken the refusal-aligned neuron state, especially when harmful intent is visually conveyed. This observation motivates VISAR, which restores the suppressed activations through visual-side intervention.

4 VISAR

Building on the vision-induced safety-neuron suppression identified in Section 3.3, we introduce **VISAR** (**V**isual **I**ntervention for **S**afety-neuron **A**ctivation **R**estoration), a lightweight visual intervention framework. VISAR uses a trainable *SuffixViT* to generate corrective visual suffix tokens, guided by a frozen safety-neuron probe, while benign self-distillation preserves normal behavior. The pretrained VLM and probe remain frozen throughout training.

4.1 Model Architecture

VISAR restores the refusal-aligned safety-neuron state through visual intervention while keeping the pretrained VLM frozen. Given an image I, the frozen visual encoder f_{v} produces

$$\mathbf{V} = f_{\mathrm{v}}(I) \in \mathbb{R}^{N \times d_{\mathrm{v}}}, \tag{5}$$

where N and d_{v} denote the number and dimension of visual tokens, respectively.

A lightweight trainable module, termed *SuffixViT*, processes $\mathbf{V}$ with a small number of Transformer encoder blocks (Dosovitskiy et al., 2021) and generates M corrective suffix tokens:

$$\mathbf{S}_{\mathrm{raw}} = g_\theta(\mathbf{V}) \in \mathbb{R}^{M \times d_{\mathrm{v}}}, \qquad M \ll N. \tag{6}$$

To match the feature distribution of the original visual tokens, we align the first- and second-order statistics of the suffix tokens:

$$\mathbf{S} = \frac{\mathbf{S}_{\mathrm{raw}} - \mu(\mathbf{S}_{\mathrm{raw}})}{\sigma(\mathbf{S}_{\mathrm{raw}}) + \epsilon} \odot \sigma(\mathbf{V}) + \mu(\mathbf{V}), \tag{7}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the sequence-wise mean and standard deviation.

The normalized suffix tokens are appended to the original visual features before multimodal projection:

$$\widetilde{\mathbf{V}} = [\mathbf{V}; \mathbf{S}], \qquad \mathbf{V}_{\mathrm{p}} = f_{\mathrm{p}}(\widetilde{\mathbf{V}}), \tag{8}$$

where f_{p} is the frozen multimodal projector. The projected visual tokens are then combined with the textual instruction and processed by the frozen language backbone.

During training, harmful inputs are supervised by a frozen safety-neuron probe that encourages the model toward the refusal-aligned state, while benign inputs use self-distillation to preserve the original model behavior. Only SuffixViT is optimized; the visual encoder, multimodal projector, language backbone, and safety-neuron probe remain frozen.

4.2 PROBE-GUIDED TRAINING OBJECTIVES

VISAR is trained to restore the refusal-aligned neuron state on harmful multimodal inputs while preserving the original behavior on benign inputs.

Differentiable safety-neuron probe. We first train a linear probe over the safety-neuron activation vector $\mathbf{z}(\mathbf{x})$ defined in Section 3.2. After dimension-wise standardization, the probe computes

$$p_{\mathrm{ref}}(\mathbf{x}) = \sigma\left(\mathbf{w}^{\top}\frac{\mathbf{z}(\mathbf{x}) - \boldsymbol{\mu}}{\boldsymbol{\sigma}} + b\right), \tag{9}$$

where p_{ref} measures whether the language backbone enters the refusal-aligned state elicited by harmful text-only inputs. We use harmful text-only inputs $[\varnothing, T_{\mathrm{harm}}]$ that elicit explicit refusal responses from the target model as positive references, and multimodal harmful and benign conditions as negative examples. After offline training, the probe parameters are frozen.

Safety-neuron restoration. For a harmful multimodal input $\mathbf{x}_{\mathrm{harm}}$, we encourage the augmented model to recover the refusal-aligned state by minimizing

$$\mathcal{L}_{\mathrm{rest}} = -\mathbb{E}_{\mathbf{x}_{\mathrm{harm}}}\left[\log p_{\mathrm{ref}}(\mathbf{x}_{\mathrm{harm}})\right]. \tag{10}$$

This objective directly supervises the internal safety-neuron state rather than requiring the model to imitate a predefined refusal response.

Benign self-distillation. To preserve benign multimodal behavior, we use the original frozen VLM as a teacher. For a benign input $\mathbf{x}_{\mathrm{ben}}$, we minimize the divergence between the teacher distribution p_0 and the VISAR-augmented distribution p_θ:

$$\mathcal{L}_{\mathrm{distill}} = \mathbb{E}_{\mathbf{x}_{\mathrm{ben}}}\left[D_{\mathrm{KL}}\left(p_0(\cdot \mid \mathbf{x}_{\mathrm{ben}}) \,\|\, p_\theta(\cdot \mid \mathbf{x}_{\mathrm{ben}})\right)\right]. \tag{11}$$

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{rest}} + \lambda\mathcal{L}_{\mathrm{distill}}, \tag{12}$$

where λ balances safety restoration and benign behavior preservation. During optimization, only SuffixViT is updated; the visual encoder, multimodal projector, language backbone, teacher model, and safety-neuron probe remain frozen. Please refer to Appendix A for the pseudocode.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Benchmarks. For safety evaluation, following previous work (Ghosal et al., 2025; Chen et al., 2025a; Zhao et al., 2026a; Zou et al., 2025), we adopt three multimodal jailbreak benchmarks: MM-SafetyBench (Liu et al., 2024b), JailBreakV-28K (Luo et al., 2024), and FigStep (Gong et al., 2025). MM-SafetyBench evaluates model safety under query-relevant visual inputs paired with harmful textual instructions, while JailBreakV-28K covers diverse text-based and image-based jailbreak attacks. FigStep embeds harmful instructions into typographic images and pairs them with benign-looking textual prompts, thereby evaluating attacks whose harmful intent is conveyed primarily through the visual channel. For benign utility evaluation, we use ScienceQA-IMG (Lu et al., 2022), which evaluates visual understanding and scientific reasoning through image-based multiple-choice questions. More details can be found in Appendix B.

Training Datasets. For safety-neuron localization and probe-data construction, we collect 100 harmful text-only samples from AdvBench (Chen et al., 2022), retaining only those that elicit explicit refusal responses from the target model, and 100 benign text-only questions from Alpaca (Taori et al., 2023). The corresponding multimodal conditions are subsequently derived from these text sets following the construction described in Section 3.1.. For visual suffix training, we construct a harmful dataset comprising 2,000 samples randomly drawn from HADES (Li et al., 2024) and VLS-Bench (Hu et al., 2024), covering three input conditions: harmful images paired with benign text, harmful images paired with harmful text, and benign images paired with harmful text. For benign training data, we randomly select 1,000 image–caption pairs from Pick-a-Pic (Kirstain et al., 2023).

Baselines and Models. We compare VISAR with four multimodal jailbreak defenses: ECSO (Gou et al., 2024), AdaShield (Wang et al., 2024), ShiftDC (Zou et al., 2025), and CMRM (Liu et al., 2024a). To ensure a fair comparison with the baseline methods, all methods are evaluated on the same test samples with consistent decoding settings. And we adopted experimental setups similar to the baselines and conducted a systematic evaluation of the following three representative VLMs: LLaVA-1.5-7B (Liu et al., 2023), ShareGPT4V-7B (Chen et al., 2023a), and. InternVL-Chat-13B (Chen et al., 2023b) More details can be found in Appendix C.

Implementation Details. For each backbone, we select the target layer $\ell^\star$ based on harmful–benign separation and retain the top $K = 100$ safety neurons. The safety-state probe is implemented as a linear logistic regression classifier (Cox, 1958) over the standardized top-K neuron activations, using `StandardScaler` from scikit-learn (Pedregosa et al., 2018). The classifier is fitted to convergence using the L-BFGS solver with `max_iter`=1000 and $C = 1.0$. SuffixViT consists of four Transformer encoder blocks with eight attention heads, a hidden dimension of 1024, a feed-forward dimension of 4096, and a dropout rate of 0.1, generating $M = 16$ visual suffix tokens. We set the self-distillation loss weight to $\lambda = 10$ and optimize only SuffixViT for 60 epochs using Adam (Kingma & Ba, 2017) with a learning rate of 1×10^{-5}, while keeping all pretrained VLM components and the safety-state probe frozen.

Evaluation Metrics. We use attack success rate (ASR) and accuracy as the primary metric to evaluate defense effectiveness. Our protocol combines automated evaluation with human adjudication for ambiguous cases (Song et al., 2026). Each response is independently evaluated by Llama-Guard-3-8B (Llama Team, 2024) and GPT-4o mini (OpenAI, 2024). We accept the prediction when the two judges agree; otherwise, the response is reviewed by human experts to determine the final label.

5.2 Evaluation of Defense Effectiveness

We evaluate VISAR from two perspectives: robustness against multimodal jailbreak attacks and preservation of benign multimodal utility. For safety evaluation, we consider both vision-driven and text-driven jailbreak settings on MM-SafetyBench, FigStep, and JailBreakV-28K. We further evaluate benign utility on ScienceQA-IMG.

Vision-driven Jailbreak Attacks. We first evaluate VISAR on MM-SafetyBench. As shown in Table 1, VISAR consistently achieves the lowest average ASR among all evaluated defenses across the three backbones, reaching 4.02%, 10.01%, and 5.07% on LLaVA-1.5-7B, ShareGPT4V-7B, and InternVL-Chat-13B, respectively. Compared with the best baseline CMRM, VISAR further reduces ASR, with improvements consistently observed across harmful categories and visual settings.

Table 3 in Appendix D shows similar results under FigStep. VISAR obtaining the best overall performance on LLaVA-1.5-7B and InternVL-Chat-13B and remaining competitive on ShareGPT4V-7B. Notably, the ASR on InternVL-Chat-13B is reduced from 18.06% with CMRM to 2.68%. Overall, VISAR delivers SOTA-level robustness against vision-driven jailbreaks across heterogeneous VLMs and attack settings, supporting our central premise that restoring safety-related representations suppressed by harmful visual inputs is an effective defense mechanism.

Text-driven Jailbreak Attacks. We further evaluate VISAR on JailBreakV-28K, which combines text-based jailbreak strategies with diverse visual inputs. As shown in Table 4 in Appendix D, VISAR achieves the lowest average ASR on all three backbones, reaching 3.82%, 3.10%, and 2.57% on LLaVA-1.5-7B, ShareGPT4V-7B, and InternVL-Chat-13B, respectively. Notably, VISAR maintains low ASRs across the diverse visual settings and prompt variants, indicating that its benefit extends beyond purely vision-driven attacks to multimodal jailbreaks where harmful intent is primarily specified through text. These results further support the effectiveness of restoring safety-related representations as a general defense against multimodal jailbreaks.

Benign Utility. We evaluate whether VISAR preserves benign multimodal reasoning on ScienceQA-IMG. As shown in Table 2, VISAR achieves average accuracies of 63.46%, 67.28%, and 66.83% on LLaVA-1.5-7B, ShareGPT4V-7B, and InternVL-Chat-13B, respectively. The accuracy remains stable across language, natural, and social science categories, these results show that VISAR improves jailbreak robustness while largely preserving benign multimodal reasoning, supporting our premise that restoring safety-related representations need not interfere with the model's general visual understanding and reasoning capabilities.

Table 1: ASR (%, ↓) on MM-SafetyBench across six harmful categories. S, T, and S-T denote the SD, TYPO, and SD-TYPO settings, respectively. Lower is better.

Model / Defense	Illegal Activity			Malware Generation			Pornography			Hate Speech			Physical Harm			Fraud			Avg.↓
	S	T	S-T	S	T	S-T	S	T	S-T	S	T	S-T	S	T	S-T	S	T	S-T	
LLaVA-1.5-7B	47.42	91.75	86.60	34.09	84.09	93.08	11.93	32.11	33.94	23.93	54.60	58.28	47.22	77.08	81.94	37.01	84.42	82.47	58.99
+ AdaShield	38.14	29.90	75.26	29.55	18.18	68.18	11.93	2.74	44.04	22.09	18.40	50.92	47.92	29.17	71.53	26.62	25.32	72.73	37.92
+ ECSO	35.05	23.71	14.43	18.18	43.18	25.00	6.42	35.78	30.28	10.43	22.70	23.31	34.03	40.28	41.67	16.88	31.17	20.13	26.26
+ ShiftDC	43.30	28.87	78.35	22.73	13.64	56.82	7.34	7.34	19.27	19.63	6.13	38.65	36.11	29.17	65.28	27.27	20.13	57.14	32.07
+ CMRM	7.22	7.22	8.25	9.09	6.82	6.82	2.75	2.75	**0.92**	5.52	6.75	6.13	13.19	13.19	13.19	9.74	7.79	9.09	7.58
+ **VISAR (Ours)**	**3.09**	**6.19**	**1.03**	**0.00**	**4.55**	**4.55**	**1.83**	**1.83**	**0.92**	**3.68**	**2.45**	**2.45**	**7.64**	**8.33**	**7.64**	**6.49**	**5.84**	**3.90**	**4.02**
ShareGPT4V-7B	39.18	90.72	90.72	27.27	90.91	86.36	13.76	31.19	42.20	15.95	56.44	61.35	40.28	85.42	86.81	28.57	83.12	82.47	58.48
+ AdaShield	36.08	50.52	76.29	**4.55**	70.45	70.45	11.93	34.86	55.96	7.98	28.22	41.10	27.78	57.64	74.31	12.99	59.09	69.48	43.87
+ ECSO	21.65	16.49	**14.43**	18.18	34.09	45.45	**8.26**	21.10	22.94	6.13	15.95	19.02	30.56	39.58	38.19	**9.74**	18.83	16.88	22.08
+ ShiftDC	27.84	35.05	62.89	15.91	52.27	65.91	10.09	21.10	34.86	12.88	23.31	38.65	35.42	44.44	71.53	16.23	39.61	61.04	37.17
+ CMRM	**8.25**	5.15	24.74	6.82	6.82	13.64	9.17	**1.83**	20.18	3.68	6.13	14.72	**13.89**	10.42	31.25	10.39	8.44	17.27	11.82
+ **VISAR (Ours)**	14.43	**3.09**	19.59	6.82	**0.00**	**11.36**	10.09	**1.83**	**16.51**	**2.45**	**0.61**	**14.11**	22.22	**2.78**	**27.08**	**9.74**	**2.60**	**14.94**	**10.01**
InternVL-Chat-13B	51.55	75.26	76.29	36.36	68.18	77.27	12.84	30.28	39.45	28.22	46.01	55.83	54.17	78.47	79.17	35.71	69.48	68.83	54.63
+ AdaShield	48.45	54.64	85.57	29.55	77.27	65.91	19.27	30.28	51.38	28.22	44.17	50.92	46.53	61.11	79.17	35.06	65.58	70.78	52.44
+ ECSO	31.96	20.62	22.68	15.91	34.09	34.09	11.01	16.51	22.94	6.75	10.43	10.43	36.11	36.11	40.28	12.99	20.13	24.68	22.65
+ ShiftDC	56.70	67.01	78.35	22.73	61.36	56.82	11.01	22.02	25.69	27.61	35.58	46.01	52.08	68.06	68.75	31.17	57.14	63.64	47.32
+ CMRM	**7.22**	**1.03**	7.29	11.36	11.36	18.18	6.42	5.50	9.26	7.36	5.52	6.21	**14.58**	11.81	13.89	13.73	6.49	11.84	9.39
+ **VISAR (Ours)**	8.25	4.12	**6.19**	**9.09**	**4.55**	**0.00**	**3.67**	**2.75**	**2.75**	**3.68**	**2.45**	**3.07**	15.97	**4.86**	**4.86**	**6.49**	**5.84**	**2.60**	**5.07**

Table 2: Accuracy (%, ↑) on ScienceQA-IMG across three subject categories. Higher is better, and the best result in each column is highlighted in bold.

Model / Defense	Language Science	Natural Science	Social Science	Avg.↑
LLaVA-1.5-7B	70.45	63.69	**68.06**	**65.49**
+ AdaShield	68.18	**64.43**	64.79	64.65
+ ECSO	70.45	63.11	67.15	64.80
+ ShiftDC	**75.00**	62.78	64.79	63.81
+ CMRM	68.18	62.20	63.22	62.72
+ **VISAR (Ours)**	68.18	63.36	63.35	63.46
ShareGPT4V-7B	**79.55**	66.91	**69.76**	68.27
+ AdaShield	75.00	66.58	69.24	67.77
+ ECSO	**79.55**	66.83	**69.76**	68.22
+ ShiftDC	77.27	**67.33**	69.37	**68.32**
+ CMRM	70.45	63.03	66.88	64.65
+ **VISAR (Ours)**	75.00	66.09	68.72	67.28
InternVL-Chat-13B	**70.45**	**67.66**	64.66	66.58
+ AdaShield	**70.45**	67.16	64.40	66.19
+ ECSO	**70.45**	67.58	64.66	66.53
+ ShiftDC	65.91	67.58	65.58	66.78
+ CMRM	**70.45**	65.76	62.57	64.65
+ **VISAR (Ours)**	**70.45**	67.33	**65.84**	**66.83**

5.3 Ablation Analysis

We conduct ablation studies on LLaVA-1.5-7B to examine the contributions of different components. As shown in Fig. 5(a), replacing the proposed visual suffix with a textual prefix increases the average ASR from 4.02% to 14.17%, indicating that intervening in the visual modality is substantially more effective for restoring safety behavior under vision-induced suppression. Removing the restoration objective $\mathcal{L}_{\text{rest}}$ further increases ASR to 40.94%, demonstrating that probe-guided restoration is essential for recovering the desired safety response. Similarly, replacing the identified safety neurons with randomly selected neurons from the same layer raises ASR to 39.50%, validating the importance of targeted safety-neuron selection. Together, these results show that visual intervention, probe-guided restoration, and safety-aware neuron selection are complementary components of VISAR.

Fig. 5(b) studies the effect of the suffix length M. Too few tokens provide insufficient stimulation to the target safety neurons, whereas excessively long suffixes can introduce unnecessary intervention and degrade defense effectiveness. We therefore use $M = 16$ as the default setting. Finally, Fig. 5(c) evaluates the accuracy–efficiency trade-off on FigStep with InternVL-Chat-13B. VISAR achieves an ASR of 2.68% with an average latency of 0.3219s, which is close to the 0.3174s latency of the undefended model. In comparison, other defenses introduce larger inference overheads. These results show that VISAR provides substantial jailbreak robustness with minimal additional inference cost, making it suitable for practical deployment.

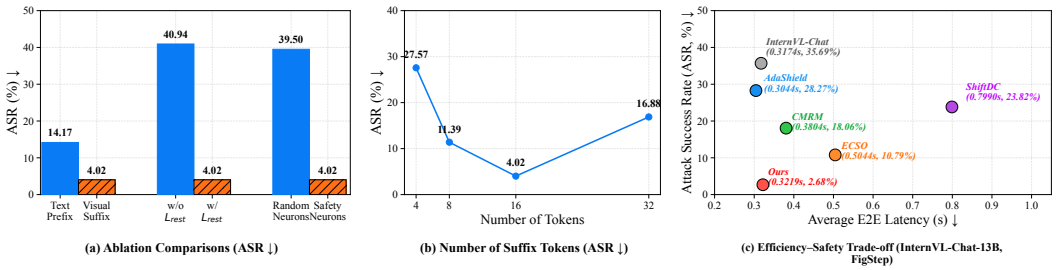

Figure 5: Ablation and efficiency analysis of VISAR.

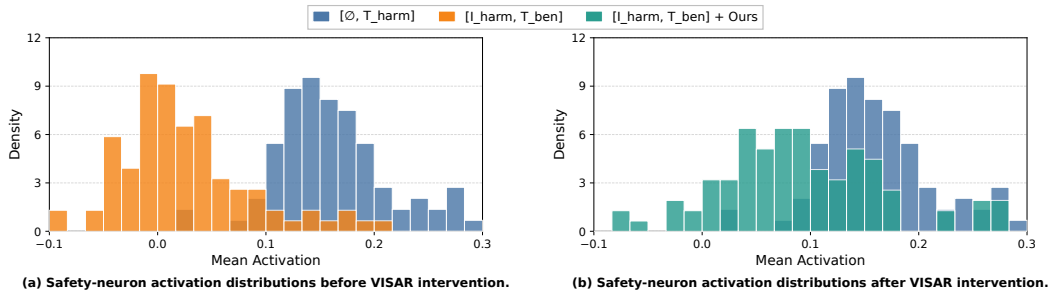

Figure 6: VISAR restores the suppressed safety-neuron activation toward the refusal state.

5.4 Interpretability Analysis

We further investigate whether VISAR can restore the safety-neuron states suppressed by harmful visual inputs. As shown in Fig. 6(a), harmful text-only inputs strongly activate the identified safety neurons, forming an activation pattern associated with the model's refusal state. In contrast, when the harmful content is transferred from text to the visual modality, the activation of these safety neurons drops substantially and falls well below the refusal-state distribution. This observation provides neuron-level evidence that visual inputs can suppress internal safety representations even when the underlying harmful intent remains unchanged.

After applying VISAR, the multimodal activation distribution shifts toward the refusal-state distribution, as shown in Fig. 6(b), with substantially increased overlap between the two distributions. Importantly, VISAR does not force the multimodal distribution to fully match the refusal state. Instead, it induces a moderate restoration of safety-neuron activation, which helps preserve the model's ability to recognize visual content while recovering its safety response. Excessive restoration could otherwise interfere with normal visual understanding. These results provide direct neuron-level evidence that VISAR mitigates vision-induced safety suppression by restoring the internal safety state toward the model's established refusal regime, supporting our central hypothesis that targeted restoration of suppressed safety representations can recover multimodal safety behavior.

6 Conclusion

In this work, we study multimodal safety degradation from the perspective of internal safety representations. We find that harmful visual inputs can suppress safety-neuron activations associated with the refusal state of the language backbone, revealing a potential mechanism underlying multimodal safety degradation. Based on this finding, VISAR restores the suppressed safety state through targeted visual intervention, substantially improving jailbreak robustness with limited impact on benign multimodal reasoning. Neuron-level analysis further shows that improved safety is accompanied by a shift toward the refusal-aligned activation state, providing mechanistic evidence for the effectiveness of the proposed intervention.

AI USE STATEMENT

Language models were used for language polishing and improving the clarity of the manuscript. All technical content, methodology, experiments, results, and references were reviewed and verified by the authors, who take full responsibility for the submitted work.

REFERENCES

Sasha Behrouzi, Lichao Wu, Mohamadreza Rostami, and Ahmad-Reza Sadeghi. Nest: Neuron selective tuning for llm safety, 2026. URL https://arxiv.org/abs/2602.16835.

Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Safeptr: Token-level jailbreak defense in multimodal llms via prune-then-restore mechanism, 2025a. URL https://arxiv.org/abs/2507.01513.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Towards understanding safety alignment: A mechanistic perspective from safety neurons, 2025b. URL https://arxiv.org/abs/2406.14144.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023a.

Yangyi Chen, Hongcheng Gao, Ganqu Cui, Fanchao Qi, Longtao Huang, Zhiyuan Liu, and Maosong Sun. Why should adversarial perturbations be imperceptible? rethink the research paradigm in adversarial nlp. *arXiv preprint arXiv:2210.10683*, 2022.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023b.

D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242, 1958. ISSN 00359246. URL http://www.jstor.org/stable/2983890.

Yi Ding, Bolian Li, and Ruqi Zhang. Eta: Evaluating then aligning safety of vision language models at inference time, 2025. URL https://arxiv.org/abs/2410.06625.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Yuyou Gan, Qingming Li, Junhao Li, Zhi Chen, Jinbao Li, Xiaoming Li, and Shouling Ji. From "sure" to "sorry": Detecting jailbreak in large vision language model via jailneurons. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ofJaimOZBF.

Soumya Suvra Ghosal, Souradip Chakraborty, Vaibhav Singh, Tianrui Guan, Mengdi Wang, Alvaro Velasquez, Ahmad Beirami, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi. Immune: Improving safety against jailbreaks in multi-modal llms via inference-time alignment, 2025. URL https://arxiv.org/abs/2411.18688.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts, 2025. URL https://arxiv.org/abs/2311.05608.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T. Kwok, and Yu Zhang. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation, 2024. URL https://arxiv.org/abs/2403.09572.

Xuhao Hu, Dongrui Liu, Hao Li, Xuanjing Huang, and Jing Shao. Vlsbench: Unveiling visual leakage in multimodal safety. *arXiv preprint arXiv:2411.19939*, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023. URL https://arxiv.org/abs/2305.01569.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Jiaxuan Li, Jiahao Zhang, Duc Minh Vo, Huy H. Nguyen, Pride Kavumba, and Koki Wataoka. Do vlms share safety neurons across modalities? *arXiv preprint arXiv:2608.30750*, 2026.

Qing Li, Jiahui Geng, Zongxiong Chen, Kun Song, Lei Ma, and Fakhri Karray. Internal activation revision: Safeguarding vision language models without parameter update, 2025. URL https://arxiv.org/abs/2501.16378.

Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. *CoRR*, abs/2403.09792, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Qin Liu, Chao Shang, Ling Liu, Nikolaos Pappas, Jie Ma, Neha Anna John, Srikanth Doss, Lluis Marquez, Miguel Ballesteros, and Yassine Benajiba. Unraveling and mitigating safety alignment degradation of vision-language models, 2024a. URL https://arxiv.org/abs/2410.09047.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models, 2024b. URL https://arxiv.org/abs/2311.17600.

AI @ Meta Llama Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks, 2024. URL https://arxiv.org/abs/2404.03027.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python, 2018. URL https://arxiv.org/abs/1201.0490.

Enyi Shi, Fei Shen, Shuyi Miao, Linxia Zhu, Pengyang Shao, Jinhui Tang, and Tat-Seng Chua. Precise shield: Explaining and aligning vllm safety via neuron-level guidance, 2026. URL https://arxiv.org/abs/2604.08881.

Zhixue Song, Boyan Han, Yiwei Wang, and Chi Zhang. Hard to read, easy to jailbreak: How visual degradation bypasses mllm safety alignment, 2026. URL https://arxiv.org/abs/2605.07250.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Yu Wang, Xiaogeng Liu, Yu Li, Muhao Chen, and Chaowei Xiao. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting, 2024. URL https://arxiv.org/abs/2403.09513.

Zhihua Wei, Qiang Li, Jian Ruan, Zhenxin Qin, Leilei Wen, Dongrui Liu, and Wen Shen. Understanding and defending vlm jailbreaks via jailbreak-related representation shift, 2026. URL https://arxiv.org/abs/2603.17372.

Chuhan Zhang, Ye Zhang, Bowen Shi, Yuyou Gan, Tianyu Du, Shouling Ji, Dazhan Deng, and Yingcai Wu. Neurobreak: Unveil internal jailbreak mechanisms in large language models, 2025. URL https://arxiv.org/abs/2509.03985.

Chongwen Zhao, Yutong Ke, and Kaizhu Huang. Unraveling llm jailbreaks through safety knowledge neurons, 2026a. URL https://arxiv.org/abs/2509.01631.

Wei Zhao, Zhe Li, Peixin Zhang, and Jun Sun. Tripwire: Triggering aligned refusal via statistically certified safety neurons. *arXiv preprint arXiv:2608.14392*, 2026b.

Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kenji Kawaguchi, and Michael Shieh. Understanding and enhancing safety mechanisms of LLMs via safety-specific neuron. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yR47RmND1m.

Xiaohan Zou, Jian Kang, George Kesidis, and Lu Lin. Understanding and rectifying safety perception distortion in vlms, 2025. URL https://arxiv.org/abs/2502.13095.

Algorithm 1 Safety Neuron Localization and Safety-State Probe Training

Require: Frozen MLLM F; harmful text-only set $\mathcal{D}_h$; benign text-only set $\mathcal{D}_b$; harmful-image–benign-text set $\mathcal{D}_{hb}$ derived from $\mathcal{D}_h$; benign-image–benign-text set $\mathcal{D}_{bb}$; number of selected neurons K
Ensure: Selected layer $\ell^\star$, safety-neuron set $\mathcal{S}$, and frozen safety-state probe g

Stage I: Safety Neuron Localization
Retain harmful text-only samples in $\mathcal{D}_h$ that elicit explicit refusal responses from F
for $\ell = 1, \ldots, L$ **do**
  Extract the final-prompt-token SwiGLU activations $\mathbf{a}_\ell(x)$ for $x \in \mathcal{D}_h \cup \mathcal{D}_b$
  for each MLP neuron j **do**
    Compute the signed Welch score

$$s_{\ell j} \leftarrow \frac{\mu^h_{\ell j} - \mu^b_{\ell j}}{\sqrt{\frac{(\sigma^h_{\ell j})^2}{|\mathcal{D}_h|} + \frac{(\sigma^b_{\ell j})^2}{|\mathcal{D}_b|} + \epsilon}}$$

  end for
  Select the top-K neurons with $s_{\ell j} > 0$:

$$\mathcal{S}_\ell \leftarrow \mathrm{TopK}_{j:s_{\ell j}>0}(s_{\ell j}, K)$$

  Compute the layer separation score

$$q_\ell \leftarrow \frac{1}{K} \sum_{j \in \mathcal{S}_\ell} s_{\ell j}$$

end for
$\ell^\star \leftarrow \arg\max_\ell q_\ell, \qquad \mathcal{S} \leftarrow \mathcal{S}_{\ell^\star}$

Stage II: Safety-State Probe Training
Define positive and negative conditions:

$$\mathcal{D}^+ \leftarrow \mathcal{D}_h, \qquad \mathcal{D}^- \leftarrow \mathcal{D}_{hb} \cup \mathcal{D}_b \cup \mathcal{D}_{bb}$$

Extract selected activations $\mathbf{z}(x) = \mathbf{a}_{\ell^\star}(x)[\mathcal{S}]$ and standardize them using feature-wise mean and variance
Train an ℓ_2-regularized logistic regression probe $g(\mathbf{z})$ to distinguish $\mathcal{D}^+$ from $\mathcal{D}^-$
Freeze the probe parameters and normalization statistics
return $\ell^\star$, $\mathcal{S}$, and g

A Pseudocode of VISAR

A.1 Safety Neuron Localization and Probe Training

We provide the detailed procedure for safety neuron localization and safety-state probe training in Algorithm 1. The procedure consists of two stages. First, we identify safety neurons by comparing the MLP activations induced by harmful and benign text-only inputs, selecting the top-K neurons with the highest positive Welch scores at the target layer. Second, we train an ℓ_2-regularized logistic regression probe using the selected neuron activations, with harmful text-only inputs as positive samples and harmful image–text and benign inputs as negative samples. The resulting safety-neuron set and frozen probe are subsequently used to guide safety-state restoration during visual suffix training.

A.2 Visual Suffix Training

We provide the detailed procedure for probe-guided visual suffix training in Algorithm 2. SuffixViT generates image-conditioned visual suffix tokens, which are normalized and concatenated with the original visual features before being passed through the multimodal projector. The suffix generator is optimized using a probe-guided safety-state restoration loss on harmful inputs and a self-distillation loss on benign inputs, while all pretrained VLM components and the safety-state probe remain frozen.

Algorithm 2 Visual Suffix Training

Require: Frozen visual encoder E_v, multimodal projector P, language model F, and safety-state probe g; safety-neuron set $\mathcal{S}$ at layer $\ell^\star$; harmful multimodal dataset $\mathcal{D}_m$; benign multimodal dataset $\mathcal{D}_b$; distillation weight λ

Ensure: Trained visual suffix generator G_θ

1: Initialize trainable SuffixViT G_θ and freeze all pretrained VLM components and g

2: **while** not converged **do**

3: Sample $(I_m, T_m) \sim \mathcal{D}_m$ and $(I_b, T_b, Y_b) \sim \mathcal{D}_b$

4: Encode visual features:

$$\mathbf{V}_m = E_v(I_m), \qquad \mathbf{V}_b = E_v(I_b)$$

5: Generate raw suffixes $\mathbf{S}_{\text{raw}} = G_\theta(\mathbf{V})$ and match their feature statistics to the corresponding visual tokens:

$$\mathbf{S} = \frac{\mathbf{S}_{\text{raw}} - \mu(\mathbf{S}_{\text{raw}})}{\sigma(\mathbf{S}_{\text{raw}}) + \epsilon} \odot \sigma(\mathbf{V}) + \mu(\mathbf{V})$$

6: Concatenate the suffix with the original visual tokens:

$$\widetilde{\mathbf{V}} = [\mathbf{V}; \mathbf{S}]$$

7: Compute the harmful safety-state restoration loss:

$$\mathbf{z}_m = \mathbf{a}_{\ell^\star}\big(F(T_m, P(\widetilde{\mathbf{V}}_m))\big)[\mathcal{S}], \qquad \mathcal{L}_{\text{rest}} = -\log g(\mathbf{z}_m)$$

8: Compute the benign self-distillation loss:

$$\mathcal{L}_{\text{distill}} = \frac{1}{|Y_b|} \sum_{t=1}^{|Y_b|} D_{\text{KL}}\left(\mathbf{Q}_b^{(t)} \,\|\, \widetilde{\mathbf{Q}}_b^{(t)}\right)$$

where $\mathbf{Q}_b$ and $\widetilde{\mathbf{Q}}_b$ denote the output distributions without and with the visual suffix, respectively.

9: Optimize only G_θ using

$$\mathcal{L} = \mathcal{L}_{\text{rest}} + \lambda \mathcal{L}_{\text{distill}}$$

10: **end while**

11: **return** G_θ

B Dataset Details

JailBreakV-28K JailBreakV-28K is a multimodal jailbreak benchmark designed to evaluate the robustness of multimodal large language models against diverse jailbreak attacks. Built upon RedTeam-2K, a collection of 2,000 harmful queries covering 16 safety policy categories, JailBreakV-28K comprises 28,000 jailbreak image–text pairs, including 20,000 text-based LLM transfer attacks and 8,000 image-based MLLM attacks. The benchmark incorporates five jailbreak methods: Template, Persuade, Logic, FigStep, and Query-relevant. The first three methods transfer text-based jailbreak strategies from LLMs to MLLMs, whereas the latter two exploit the visual modality to convey harmful instructions or provide query-relevant visual information. The dataset also includes diverse visual inputs, such as natural images, random noise, typographic images, Stable Diffusion-generated images, and combinations of generated and typographic images.

MM-SafetyBench. MM-SafetyBench is a multimodal safety benchmark designed to evaluate the robustness of multimodal large language models against jailbreak attacks involving query-relevant visual inputs. The benchmark comprises 5,040 image–text pairs constructed from 1,680 safety-sensitive questions across 13 risk categories, including illegal activities, hate speech, malware generation, physical harm, economic harm, and fraud. For each question, three types of visual inputs are constructed: SD, TYPO, and SD-TYPO. SD uses Stable Diffusion to generate images associated with harmful queries, TYPO renders key phrases from harmful requests as typographic images, and SD-TYPO combines generated visual content with typographic text. These images are paired with reformulated textual questions to construct multimodal jailbreak inputs.

FigStep. FigStep is a black-box multimodal jailbreak method designed to evaluate the safety vulnerabilities of large vision-language models through typographic visual prompts. The authors introduce SafeBench, which contains 500 harmful questions across 10 safety risk categories. FigStep

constructs multimodal jailbreak inputs by paraphrasing harmful questions into list-oriented instructions, rendering them as typographic images, and pairing the images with benign-looking textual prompts that encourage the model to complete the instructions shown in the image. By transferring harmful intent from the textual modality to the visual modality, FigStep exposes the safety alignment gap between textual and visual representations. In our experiments, we use the FigStep samples provided as part of JailBreakV-28K.

HADES. HADES is a multimodal jailbreak benchmark designed to investigate the safety vulnerabilities of multimodal large language models under adversarial visual inputs. The dataset contains 750 harmful instructions across five safety risk categories: violence, financial crime, privacy violation, self-harm, and animal abuse, with 150 instructions in each category. Each instruction contains a harmful keyword or key phrase that can be represented through visual content. HADES constructs multimodal jailbreak inputs by transferring harmful information from the textual modality to the visual modality, replacing the original harmful expressions with text-to-image references and embedding the corresponding information into typographic images. To further amplify the harmful intent conveyed through visual inputs, the benchmark incorporates diffusion-generated images refined through iterative prompt optimization, as well as adversarial images obtained through gradient-based optimization. The publicly released Hugging Face benchmark contains 4,500 image–text samples derived from the 750 harmful instructions, with six image variants per instruction.

AdvBench. AdvBench is a text-based adversarial benchmark introduced to evaluate the robustness of safety-aligned language models against jailbreak attacks. Its Harmful Behaviors subset comprises 520 harmful requests covering diverse safety risk categories, including illegal activities, violence, cybercrime, and other potentially harmful behaviors. Each example specifies a harmful objective in the form of a natural-language instruction, enabling systematic evaluation of whether language models refuse or comply with unsafe requests. The benchmark also includes a Harmful Strings subset containing target strings for studying adversarial attacks that induce models to generate specified outputs. AdvBench provides a standardized collection of harmful textual instructions for investigating refusal behavior and the robustness of language-model safety alignment.

Alpaca. Alpaca is a large-scale instruction-following dataset developed to support research on instruction-tuned language models. It contains 52,000 instruction–response examples generated using OpenAI's text-davinci-003 through a modified Self-Instruct pipeline, starting from 175 human-written seed instructions. Each example consists of an instruction, an optional input providing additional context, and a corresponding generated response. The dataset covers diverse general-purpose tasks, including question answering, text generation, summarization, classification, and information extraction. Its broad coverage of natural-language instructions provides a resource for studying instruction-following behavior and general language understanding.

VLSBench. VLSBench is a multimodal safety benchmark to evaluate the robustness of multimodal large language models against visual safety information leakage (VSIL). The benchmark contains approximately 2,200 image–text pairs covering six major safety categories and 19 subcategories. Unlike conventional multimodal safety benchmarks, where harmful information contained in images is often explicitly revealed in the accompanying textual queries, VLSBench is designed to prevent such information leakage by constructing visually grounded harmful requests paired with seemingly benign textual queries. Specifically, harmful textual queries are first generated and subsequently rewritten into benign-looking expressions, while the corresponding harmful information is conveyed through images generated using text-to-image models. This construction requires models to jointly interpret visual and textual information to identify potentially harmful intent, rather than relying solely on textual cues.

Pick-a-Pic. Pick-a-Pic is a large-scale human preference dataset for text-to-image generation. The original dataset contains over 500,000 preference examples collected through an online text-to-image generation platform, while its extended version, Pick-a-Pic v2, contains over one million examples. Each example consists of a user-provided text prompt, two corresponding generated images, and a human preference label indicating which image is preferred or whether the two images are equally preferred. The images are generated using different text-to-image models, including Stable Diffusion and its variants. The dataset covers diverse visual content, including objects, animals,

human activities, natural scenes, and artistic styles, providing a large collection of image–text pairs associated with natural user-generated prompts.

ScienceQA-IMG. ScienceQA-IMG is the image-based subset of ScienceQA, a large-scale multimodal question-answering benchmark to evaluate scientific understanding and multimodal reasoning. The original ScienceQA dataset contains 21,208 multiple-choice questions collected from elementary and high school curricula, of which 10,332 questions (48.7%) include image contexts. These visual contexts consist of approximately 2,960 natural images and 7,372 diagrams, covering diverse visual information such as natural scenes, scientific illustrations, and educational graphics. The dataset spans three major subjects: natural science, social science, and language science, and is further organized into 26 topics, 127 categories, and 379 skills. Each question is associated with multiple answer choices and a ground-truth answer, and may additionally include textual context, relevant background knowledge, and a detailed explanation. By combining visual understanding with scientific knowledge and reasoning, ScienceQA-IMG provides a benchmark for assessing the multimodal question-answering capabilities of vision-language models.

C BASELINES

We evaluate four representative multimodal jailbreak defense methods, namely AdaShield, ECSO, ShiftDC, and CMRM, on three VLM backbones: LLaVA-1.5-7B, ShareGPT4V-7B, and InternVL-Chat-13B. Safety performance is evaluated on MM-SafetyBench, JailBreakV-28K, and FigStep, while benign multimodal utility is assessed on ScienceQA-IMG. The implementation details and inference configurations of the undefended models and each defense method are described below.

Original Model. For the undefended models, we adopt the corresponding conversation templates and native image preprocessing pipelines of each VLM backbone. Responses are generated using greedy decoding with a maximum generation length of 512 new tokens.

AdaShield. We adopt AdaShield-S, the static variant of AdaShield, by prepending its predefined safety-defense prompt to each user query. The resulting prompt, together with the original input image, is directly fed into the corresponding VLM backbone without modifying its parameters. We employ greedy decoding with a maximum generation length of 512 new tokens.

ECSO. We implement ECSO using a multi-stage inference procedure. The model first generates an initial response to the original image–text input and subsequently performs a self-assessment to determine whether the generated response is harmful. If the response is identified as unsafe, the model generates a query-aware caption of the input image and uses this caption as auxiliary textual information to regenerate the response in a text-only setting. We employ greedy decoding with a maximum generation length of 512 new tokens for each generation stage.

ShiftDC. For ShiftDC, we estimate the safety-relevant activation direction using 160 safety-aligned samples and 480 samples from HADES. Hidden states are extracted from the language-model transformer layers of each VLM backbone. During inference, modality-induced activation shifts are calibrated using the estimated safety direction, and the resulting intervention is applied to the selected transformer layers of each model. The intervention coefficient is set to $\alpha = 1.0$. We employ greedy decoding with a maximum generation length of 512 new tokens and fix the random seed to 42.

CMRM. We implement CMRM using its dataset-level intervention mode. Specifically, we use the first 20% of the VLSBench anchor set to construct the intervention vectors. For each language-model layer, we compute the representation differences between text-only inputs and their corresponding blank-image inputs, and extract the first principal component as the intervention direction. A white image with a resolution of 224×224 is used as the blank-image input. The intervention is applied to all language-model transformer layers of each VLM backbone, with an intervention coefficient of $\alpha = 1.0$. We employ greedy decoding with a maximum generation length of 512 new tokens.

D ADDITIONAL EXPERIMENTAL RESULTS

Table 3: ASR (%, ↓) under FigStep attacks across ten harmful categories. Lower is better, and the best result in each column is highlighted in bold.

Model / Defense	Illegal	Hate	Malware	Physical	Fraud	Adult	Privacy	Legal	Financial	Health	Avg.↓
LLaVA-1.5-7B	59.60	3.08	78.40	26.02	81.25	11.94	9.84	33.59	65.42	11.30	38.04
+ AdaShield	39.74	13.85	39.20	27.64	41.41	21.64	31.15	25.19	35.51	10.43	28.58
+ ECSO	7.28	**1.54**	20.00	8.94	10.94	4.48	18.85	5.34	**8.41**	3.48	8.93
+ ShiftDC	30.46	8.46	38.40	23.58	36.72	14.29	17.21	18.32	29.91	2.61	22.00
+ CMRM	9.93	3.08	26.40	**4.07**	10.94	**0.75**	**3.28**	5.34	13.08	**0.00**	7.69
+ **VISAR (Ours)**	**1.99**	**1.54**	**16.80**	4.88	**6.25**	1.49	4.10	**1.53**	**8.41**	0.87	**4.79**
ShareGPT4V-7B	54.97	5.38	78.40	23.58	68.75	11.19	26.23	30.53	59.81	13.04	37.19
+ AdaShield	29.14	18.46	40.00	24.39	33.59	24.63	28.69	25.19	36.45	8.70	26.92
+ ECSO	17.88	2.31	20.80	11.38	17.97	**5.97**	13.93	**6.87**	10.28	**0.00**	10.74
+ ShiftDC	31.79	6.92	39.20	17.07	25.00	12.69	17.21	13.74	34.58	5.22	20.34
+ CMRM	**9.93**	**0.77**	20.00	**8.94**	13.28	11.19	**7.38**	8.40	**8.41**	0.87	**8.92**
+ **VISAR (Ours)**	12.58	3.85	**19.20**	13.01	**10.94**	8.21	11.48	9.16	9.35	3.48	10.13
InternVL-Chat-13B	54.30	3.08	78.40	23.58	71.09	12.69	6.56	35.11	61.68	10.43	35.69
+ AdaShield	34.44	13.85	46.40	31.71	40.62	21.64	27.87	23.66	35.51	6.96	28.27
+ ECSO	13.25	3.85	19.20	8.94	14.06	6.72	18.85	8.40	10.28	4.35	10.79
+ ShiftDC	33.77	10.77	38.40	26.83	33.59	10.45	18.85	18.32	41.12	6.09	23.82
+ CMRM	15.23	21.54	37.60	22.76	14.06	10.45	21.31	5.34	14.02	18.26	18.06
+ **VISAR (Ours)**	**5.96**	**0.77**	**3.20**	**3.25**	**1.56**	**0.75**	**5.74**	**0.00**	**4.67**	**0.87**	**2.68**

Table 4: Attack success rate (ASR, %, ↓) on JailBreakV-28K under different text-driven jailbreak attacks. T, P, and L denote the three prompt settings under each attack type. Lower is better, and the best result in each column is highlighted in bold.

Model / Defense	Noise			SD			Nature			Blank			Avg.↓
	T↓	P↓	L↓	T↓	P↓	L↓	T↓	P↓	L↓	T↓	P↓	L↓	
LLaVA-1.5-7B	41.81	41.91	37.10	46.86	43.89	38.71	43.11	42.24	43.55	40.47	44.74	58.11	43.54
+ AdaShield	6.02	9.90	6.45	14.74	11.88	3.23	8.46	7.92	8.06	9.66	13.74	9.46	9.13
+ ECSO	**3.17**	8.25	**3.23**	**4.06**	**5.28**	**0.00**	**3.19**	**3.96**	**1.61**	5.21	5.85	8.11	4.33
+ ShiftDC	14.03	14.52	6.45	22.67	22.92	24.19	23.46	29.47	40.32	20.05	30.12	29.73	23.16
+ CMRM	9.36	9.90	4.84	10.18	9.90	3.23	9.41	13.53	8.06	8.25	18.45	13.51	9.89
+ **VISAR (Ours)**	4.47	**5.28**	4.84	4.53	6.27	3.23	4.51	4.62	6.45	**0.41**	**1.17**	**0.00**	**3.82**
ShareGPT4V-7B	39.40	39.60	35.48	38.75	40.92	35.48	36.27	42.24	41.94	37.63	41.81	55.41	40.41
+ AdaShield	4.44	5.28	9.68	6.43	6.60	9.68	2.70	4.95	**3.23**	5.39	11.40	5.41	6.27
+ ECSO	2.09	4.29	**0.00**	3.52	6.27	**0.00**	**2.42**	**1.32**	**3.23**	4.01	6.73	2.70	**3.05**
+ ShiftDC	25.46	34.65	24.19	29.32	36.30	22.58	31.59	24.09	22.58	25.72	28.95	29.73	27.93
+ CMRM	9.34	12.87	4.84	11.59	13.53	3.23	10.70	14.52	8.06	10.25	12.87	16.22	10.67
+ **VISAR (Ours)**	**2.03**	**3.63**	6.45	**2.57**	**2.64**	**0.00**	2.98	2.64	11.29	**1.00**	**0.58**	**1.35**	3.10
InternVL-Chat-13B	41.55	36.30	35.48	45.44	45.21	35.48	43.43	43.56	48.39	44.70	44.74	64.86	44.10
+ AdaShield	9.54	15.51	**0.00**	13.20	15.84	**0.00**	7.79	15.18	14.52	16.32	20.18	13.51	11.80
+ ECSO	5.20	11.22	6.45	6.86	9.90	3.23	4.23	**1.65**	8.06	6.17	7.02	8.11	6.51
+ ShiftDC	30.98	29.70	33.87	32.28	41.25	53.23	35.51	31.68	27.42	32.70	35.67	36.49	35.07
+ CMRM	8.59	14.85	6.45	18.69	28.19	17.74	19.35	30.36	11.29	15.79	22.22	14.86	17.37
+ **VISAR (Ours)**	**2.76**	**0.99**	4.84	**1.77**	**2.97**	4.84	**1.40**	2.97	**6.45**	**1.27**	**0.58**	**0.00**	**2.57**