VISION ENCODER IS A HIDDEN CULPRIT FOR SAFETY DEGRADATION IN BENIGN LVLM FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Benign fine-tuning can undermine the safety alignment of Large Vision-Language Models (LVLMs), a risk commonly associated with their language counterpart. We reveal that the vision encoder is also unsafe or less safe than the language backbone. By updating only the vision encoder (*V-tuning*), substantial safety degradation is observed across various LVLMs and downstream tasks. Such safety deterioration matches or even exceeds that of language-backbone or full-model tuning. To explain this vulnerability, we propose the **vision-sensitivity hypothesis** and analyze how parameter updates perturb representations along a fixed refusal direction. Our first-order analysis finds that the refusal-mediating structure in LVLMs is significantly more sensitive to the vision encoder updates. Based on this, we then demonstrate that V-tuning substantially reduces the separation between harmful and benign representations along the refusal direction more than the other tuning modes. Motivated by these findings, we introduce **SAVR**, **S**ensitivity-**A**ware **V**ision **R**egularization, to preserve the LVLM's refusal structure during vision encoder fine-tuning. SAVR successfully reduces attack success rates by up to 98% relative to unprotected V-tuning while preserving utility. It also retains most of the aligned model's harmful-benign separation. We believe that this work offers new insights for further studies to understand and address fine-tuning risks in LVLMs.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) employ vision encoders to map image inputs into discrete tokens that Large Language Model (LLM) backbones can seamlessly handle. It comes as no surprise that LVLMs inherit or even amplify safety vulnerabilities of LLMs. Existing research along this line primarily focuses on training-free prompt engineering, where extensive studies curate out-of-distribution image-text pairs for jailbreaking (Gong et al., 2025; Jeong et al., 2025; Yang et al., 2025b). In addition to these expected outcomes, a more challenging and stealthy risk is that fine-tuning LVLMs could significantly erode safety alignment, even when the fine-tuning data is entirely benign (Qi et al., 2023; Lee et al., 2025). The emergence of this failure mode under benign fine-tuning demands extensive exploration. First, it can reveal vulnerabilities in safety alignment not attributable to explicitly harmful training data. Second, existing fine-tuning defenses have primarily focused on mitigating the effects of harmful fine-tuning data (Rong et al., 2025; Liu et al., 2026; Gulati & Raval, 2026). Without a deeper understanding of how benign data can cause harm, such alignment approaches may fail to address more fundamental weaknesses.

The risks arising from LVLM benign fine-tuning are predominantly attributed to the language backbone losing its previously learned safety alignment (Zong et al., 2024; Wang et al., 2026). This is intuitive and aligns with the widely accepted notion that unsafe knowledge is embedded in LLMs. In this paper, we challenge this common understanding by revealing a critical yet underexplored vulnerability lying dormant in LVLMs' **vision encoder**. To this end, we design a simple intervention study, which we refer to as *V-tuning*, to perform benign fine-tuning only on the vision encoder (refer to Fig. 1). Under the prevailing view, V-tuning should pose no threat to safety. However, our preliminary results demonstrate that it instead substantially degrades safety alignment, matching or even exceeding the damage caused by fine-tuning the LLM backbone itself. On Qwen3-VL, for instance, V-tuning raises the attack success rate (ASR) from a 2.3% baseline to as high as 55.8%, surpassing LLM-backbone tuning by up to 1.8 times. Moreover, this vulnerability generalizes to various LVLMs and persists across benign fine-tuning datasets, from mathematical and scientific reasoning

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.

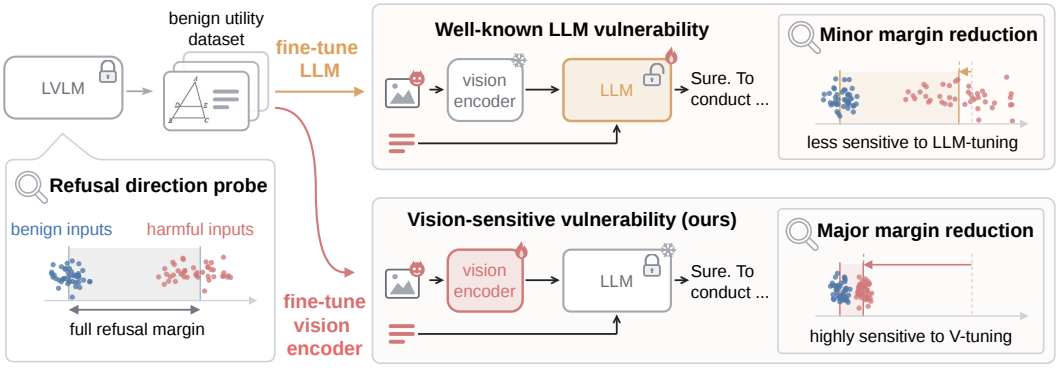


Figure 1: Illustration of V-tuning and its effect. V-tuning causes substantial damage to LVLM refusal despite the frozen LLM backbone. With the vision encoder's higher refusal sensitivity, V-tuning substantially narrows the refusal margin.

to visual grounding. This leads to an interesting question: why does this previously ignored benign V-tuning result in unexpected safety erosion?

We address this question with a **vision-sensitivity hypothesis**: an LVLM's refusal behavior depends far more heavily on the vision encoder than previously recognized. As such, V-tuning weakens the LVLM's refusal and yields more harmful outputs. To verify this, we identify an LVLM's refusal-mediating structure in the activation space (Arditi et al., 2024): a one-dimensional refusal direction with harmful and benign activations well-separated along it. We then conduct a sensitivity analysis of this refusal structure with respect to updates in each parameter block. Specifically, we quantify each block's *refusal sensitivity* as the mean squared projection of the activation shifts, induced by fine-tuning this block, onto the refusal direction. The results are revealing: refusal sensitivity is consistently *higher* for the vision encoder than for the LLM backbone across *LVLM architectures* and *downstream tasks*, by up to six orders of magnitude. This suggests that the vision encoder is the block through which benign fine-tuning most readily disrupts refusal. In the aligned model, harmful and benign inputs are separated along the refusal direction by a wide margin. V-tuning collapses this margin far more than LLM-tuning or full-model tuning does across models. On Qwen3-VL, for instance, V-tuning shrinks the aligned margin by 82%, against 10% under LLM-tuning and 8% under full-model tuning. Together, these results link the vision encoder's high refusal sensitivity to a pronounced degradation of the model's refusal structure.

Based on the vision-sensitivity hypothesis, we further propose **SAVR** (**S**ensitivity-**A**ware **V**ision **R**egularization). SAVR applies regularization gradients to the vision encoder to constrain activations' displacement along the refusal direction relative to the aligned model. This objective aims to preserve refusal-relevant structures while allowing adaptation to downstream tasks. Across all nine model-task combinations under V-tuning, SAVR lowers ASR by 61% on average (ranging from 31% to 98%), while maintaining the original utility performance. On the M^{3}CoT dataset, for example, ASR decreases from 55.8% to 1.0% on Qwen3-VL, from 73.3% to 43.6% on InternVL3, and from 53.2% to 18.8% on Llama-3.2-Vision. Similarly, ASR drops from 45.3% to 4.2%, 51.0% to 27.6%, and 35.7% to 15.0% across these three models on the MMK12 dataset. In addition, SAVR preserves most of the aligned model's refusal margin after V-tuning. On Qwen3-VL fine-tuned on M^{3}CoT, the retained refusal margin rises from 18% to 99%. On InternVL3 fine-tuned on RefCOCO, the refusal margin rises from 40% to 99%. SAVR also complements safety-data fine-tuning to achieve even lower ASRs. These results highlight the value of explicitly safeguarding the vision encoder during benign LVLM fine-tuning and its potential to strengthen existing safety fine-tuning methods.

In summary, our contributions are three-fold. First, to the best of our knowledge, we are the first to identify the vision encoder as a fine-tuning vulnerability in existing LVLM alignment by examining the harmfulness of V-tuning. Second, we explain this safety risk by proposing the vision-sensitivity hypothesis, as supported by theoretical analysis and empirical observations. Finally, we patch this new vulnerability with **SAVR**, a simple yet effective regularizer that penalizes the vision encoder's refusal sensitivity during fine-tuning. SAVR preserves refusal behavior under benign fine-tuning while retaining downstream task utility. It thereby provides a better understanding of LVLMs' vulnerability and instructs future efforts in safeguarding the entire model pipeline.

2 Related Work

LLM and LVLM Safety Alignment. Modern LLMs are aligned with human values via reinforcement learning from human feedback (Ouyang et al., 2022; Bai et al., 2022; Dai et al., 2024) or direct preference optimization (Rafailov et al., 2023) on large-scale human preference datasets (Ji et al., 2023). Some approaches also explore instilling safety by editing model weights (Gao et al., 2024) or unlearning harmful knowledge (Liu et al., 2024b). Recent LVLMs are typically built by integrating a pre-trained vision encoder into a safety-aligned LLM backbone. Apart from inheriting the safety alignment from backbones, LVLMs are further strengthened by techniques such as multimodal RLHF (Sun et al., 2024; Yu et al., 2024; Ji et al., 2025; Bachu et al., 2025), preference optimization (Zhang et al., 2025), and visual instruction tuning (Zong et al., 2024; Ding et al., 2026) on curated multimodal safety datasets, cross-modal safety transfer (Xu et al., 2025; Yang et al., 2025a) or inference-time safety recovery (Liu et al., 2025a;b; Guo et al., 2026). These efforts develop multimodal safety to safeguard LVLMs against harmful intent across modalities.

LVLM Jailbreak Attacks. Existing LVLM jailbreak attacks probe the input-level vulnerabilities of LVLMs through crafting adversarial inputs to bypass the safety guardrails. Among them, perturbation-based methods optimize image-text prompts to elicit harmful responses (Qi et al., 2024; Shayegani et al., 2024; Hao et al., 2025; Ying et al., 2025; Wang et al., 2024), but they often require white-box access and exhibit limited transferability. On the other hand, prompt-injection-based methods push inputs out of distribution via transformation, distraction, and shuffling (Gong et al., 2025; Liu et al., 2024a; Li et al., 2024; Jeong et al., 2025; Yang et al., 2025b; Zhao et al., 2025) to bypass multimodal guardrails, achieving superior attack success and transferability on modern LVLMs. Recent attacks further exploit higher-level mechanisms to undermine LVLM safety alignment, including multi-turn adaptive prompting (Choi et al., 2026), compositional reasoning over individually benign inputs (Zou et al., 2026), and cross-modal information overloading (Zhang et al., 2026). All of them, however, stress-test LVLMs at inference time while the model parameters remain intact.

LVLM Fine-Tuning Risks. Apart from jailbreaking vulnerabilities, a more stealthy and less explored type of vulnerability arises from LVLMs' parameter updates during fine-tuning. Poisoning an aligned LVLM's training data with malicious image-text pairs undermines its safety (Tao et al., 2025; Li et al., 2025), but assumes an adversary with access to training data. More alarmingly, LVLMs' refusal behavior degrades even after visual instruction tuning on entirely benign datasets (Zong et al., 2024; Pantazopoulos et al., 2024; Lee et al., 2025). This fragility has been traced to the absence of safety supervision in fine-tuning corpora (Zong et al., 2024) and failing LLM backbone alignment (Lee et al., 2025; Wang et al., 2026). However, few studies in this line include the vision encoder in their analysis. Instead, they primarily focus on vulnerabilities within the LLM.

3 The Vision-Sensitivity Hypothesis

3.1 Preliminary and Motivation

LVLM and benign fine-tuning. An LVLM processes an image-text input x using a vision encoder and an LLM backbone, linked by a multimodal connector. We denote the full model parameters by θ, the vision-encoder parameters by θ_v, and the LLM backbone parameters by θ_t. The vision encoder supplies visual features to the backbone, which generates the response conditioned on the visual and textual inputs. Given a downstream dataset $\mathcal{D}_{\text{util}}$ of input-response pairs (x, y), fine-tuning minimizes the autoregressive negative log-likelihood loss:

$$\mathcal{L}(\theta; \mathcal{D}_{\text{util}}) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{util}}} \left[\frac{1}{|y|} \sum_{i=1}^{|y|} \log p_\theta(y_i \mid x, y_{<i}) \right], \tag{1}$$

where $p_\theta(y_i \mid x, y_{<i})$ denotes the probability assigned by the LVLM to response token y_i, conditioned on the image-text input x and preceding response tokens $y_{<i}$.

V-tuning intervention. We verify whether LVLM's safety alignment can be weakened through benign fine-tuning without changing its language backbone. For this purpose, we introduce *V-tuning*: updating only θ_v while keeping the remaining parameters (including the backbone) frozen. We compare it with *L-tuning*, which updates only θ_t, and *full-tuning*, which updates all model parameters θ.

Each mode starts from the same aligned model and is trained for one epoch on the same task data. The tuning modes differ only in which parameter blocks are updated to minimize the fine-tuning loss. Under V-tuning, any change in the LVLM's behavior is then attributed to the vision encoder.

Models, tasks, and evaluation. We study Qwen3-VL-8B (Bai et al., 2025), InternVL3-8B (Zhu et al., 2025), and Llama-3.2-11B-Vision (Grattafiori et al., 2024) on three downstream tasks: mathematical problem solving with MMK12 (Meng et al., 2025), multi-domain multimodal reasoning with M^{3}CoT (Chen et al., 2024), and referring-expression generation with RefCOCO (Yu et al., 2016). Each utility dataset is mixed with general visual instruction data (Liu et al., 2023) to preserve instruction following. We evaluate safety using attack success rate (ASR) on the HADES dataset (Li et al., 2024) and utility using downstream task metrics.

Table 1: ASR (%) of fine-tuned LVLMs across different models, tasks and fine-tuning modes. **Bold**: highest ASR per task-model column. V-tuning cells are shaded where their ASR surpasses or is comparable to that of L-tuning (within 10 percentage points).

	Qwen3-VL-8B			InternVL3-8B			Llama-3.2-11B-V		
Mode	MMK12	M^{3}CoT	RefCOCO	MMK12	M^{3}CoT	RefCOCO	MMK12	M^{3}CoT	RefCOCO
None (aligned)	2.3	2.3	2.3	23.6	23.6	23.6	10.9	10.9	10.9
V-tuning	**45.3**	**55.8**	27.7	**51.0**	**73.3**	48.3	35.7	53.2	55.7
L-tuning	24.6	35.0	43.6	26.2	**73.3**	57.7	58.0	56.7	78.9
Full-tuning	28.4	40.8	**49.9**	29.4	70.3	**61.5**	**58.7**	**59.3**	**82.1**

Safety degradation with a frozen backbone. Table 1 shows that V-tuning raises ASR by up to 53.5 percentage points across all nine model-task settings, despite leaving the language backbone unchanged. On two tasks evaluated on Qwen3-VL and InternVL3, V-tuning matches or exceeds the ASR of both L-tuning and full-tuning. For example, on Qwen3-VL with MMK12, ASR rises from 2.3% to 45.3% under V-tuning, compared with 24.6% under L-tuning and 28.4% under full-tuning. Consistently, such degradation is accompanied by utility gains after V-tuning. We defer utility results and further discussion on ASR results of V-tuning to Appendix A.

The vision-sensitivity hypothesis. These results challenge the prevailing view that safety erosion originates in the LLM after benign fine-tuning. They instead demonstrate that encoder-only updates are sufficient to weaken safety in the studied LVLMs. This observation further motivates our *vision-sensitivity hypothesis*: an LVLM's safety alignment (*i.e.*, refusal behavior) depends far more heavily on the vision encoder than previously recognized. We investigate this hypothesis by probing perturbations along a refusal direction in Section 3.2 and harmful-benign separation in Section 3.3.

3.2 Refusal Sensitivity Analysis

Refusal direction. Following Arditi et al. (2024), we characterize refusal in the aligned LVLM by a single direction in the activation space. Let $h_{\text{ref}}(x;\theta)$ denote the backbone's residual-stream activation on input x, taken at a layer and token position selected per model by causal criteria (kindly refer to Appendix B for more details). Given harmful inputs $\mathcal{D}_h$ and benign safe inputs $\mathcal{D}_s$, we calculate the refusal direction as the normalized difference between their mean activations,

$$\mathbf{u}_{\text{ref}} = \frac{\boldsymbol{\mu}_{\text{h}} - \boldsymbol{\mu}_{\text{s}}}{\|\boldsymbol{\mu}_{\text{h}} - \boldsymbol{\mu}_{\text{s}}\|_2}, \qquad \boldsymbol{\mu}_{\text{h}} = \mathop{\mathbb{E}}_{x\in\mathcal{D}_h}\big[h_{\text{ref}}(x;\theta)\big], \quad \boldsymbol{\mu}_{\text{s}} = \mathop{\mathbb{E}}_{x\in\mathcal{D}_s}\big[h_{\text{ref}}(x;\theta)\big]. \tag{2}$$

We estimate $\mathbf{u}_{\text{ref}}$ once on the aligned model with parameters θ^{A} and verify that it still mediates refusal after fine-tuning. Along the refusal direction $\mathbf{u}_{\text{ref}}$, the LVLM's refusal behavior responds to where an activation sits along the direction, quantified by its scalar projection onto $\mathbf{u}_{\text{ref}}$:

$$r(x;\theta) = \mathbf{u}_{\text{ref}}^{\top} h_{\text{ref}}(x;\theta). \tag{3}$$

Since $h_{\text{ref}}(\cdot)$ depends on θ, every parameter update moves $r(x;\theta)$, and fine-tuning therefore changes the refusal structure (Jain et al., 2024; Lin et al., 2024). We thus quantify the degree of disruption across tuning modes by probing each parameter block's *refusal sensitivity*.

Refusal sensitivity. We define refusal sensitivity as the first-order response of the probed activation to one fine-tuning step, from the weights $\theta^{(k)}$ at step k to $\theta^{(k+1)}$ at the next step, following the

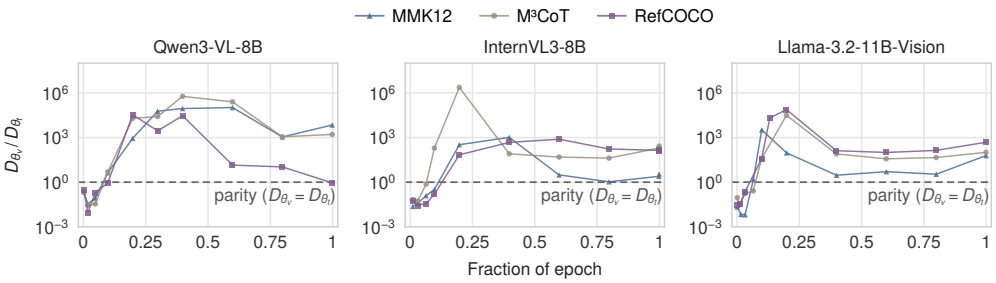


Figure 2: Refusal sensitivity ratio $D_{\theta_v}/D_{\theta_t}$ over one benign fine-tuning epoch, for each model and task. The dashed line marks D_{θ_v} being equal to D_{θ_t}. The vision encoder's refusal sensitivity starts below that of the LLM backbone in every setup, then becomes dominant for the rest of the epoch.

trajectory-tracing view of training influence (Pruthi et al., 2020). We let b denote the block that a tuning mode updates, $b \in \{\theta_v, \theta_t\}$ for V-tuning and L-tuning, respectively. At step k, a benign mini-batch $\mathcal{B}_{\text{util}}^{(k)} \subset \mathcal{D}_{\text{util}}$ determines the task gradient and the corresponding parameter increment:

$$\begin{aligned} g_b^{(k)} &= \nabla_b \mathcal{L}\big(\theta^{(k)}; \mathcal{B}_{\text{util}}^{(k)}\big), \\ \Delta_b^{(k)} &= \mathrm{AdamWStep}\left(b^{(k)}, g_b^{(k)}, s_b^{(k)}\right), \end{aligned} \tag{4}$$

where $s_b^{(k)}$ denotes the optimizer state, and $\mathrm{AdamWStep}$ denotes one update of the AdamW optimizer returning the parameter increment. The mini-batch loss follows Eq. 1. We write $\Delta_b = \Delta_b^{(k)}$ below, embedding it in the full parameter space with zeros outside block b, so that $\theta^{(k+1)} = \theta^{(k)} + \Delta_b$. For a fixed probe input x, the resulting activation change can be expressed as the first-order approximation (Lee et al., 2019)

$$h_{\text{ref}}\big(x;\, \theta^{(k+1)}\big) \;\approx\; h_{\text{ref}}\big(x; \theta^{(k)}\big) \;+\; J_b\big(x; \theta^{(k)}\big)\, \Delta_b, \tag{5}$$

where $J_b(x; \theta^{(k)}) = \partial h_{\text{ref}}(x; \theta^{(k)})/\partial b$ refers to the Jacobian of the probed activation with respect to block b, evaluated at the current weights $\theta^{(k)}$. We isolate the safety-critical perturbation along the refusal direction,

$$\delta_b\big(x; \theta^{(k)}\big) \;=\; \mathbf{u}_{\text{ref}}^{\top} J_b\big(x; \theta^{(k)}\big)\, \Delta_b. \tag{6}$$

This means that one update to block b shifts x on the refusal axis by approximately $\delta_b(x; \theta^{(k)})$. Specifically, it measures how far a single step of training moves the representation of x along the refusal direction. For robustness, we average the square of this perturbation over a pool of harmful inputs for probing, leading to the *refusal sensitivity* metric corresponding to block b at step k:

$$D_b\big(\theta^{(k)}\big) \;=\; \mathop{\mathbb{E}}_{x \in \mathcal{D}_h} \Big[\delta_b\big(x; \theta^{(k)}\big)^2\Big]. \tag{7}$$

Comparing D_b across different parameter blocks therefore examines which block, given the same benign task supervision, has the greater capacity to cause disruption to the refusal structure and drive the safety degradation in the fine-tuned LVLMs.

Probing setup. Since a block's refusal sensitivity changes as training proceeds, we do not evaluate it at θ^{A} alone but track it along the fine-tuning trajectory itself. For each model and downstream task, we run two trajectories on the same utility data, one updating only the vision encoder and one updating only the LLM backbone, with the trainable parameter set of each trajectory identical to that of V-tuning and L-tuning in Section 3.1 and all other settings matched (*e.g.*, fine-tune for one epoch). Along the first and second trajectory we record $D_{\theta_v}(\theta^{(k)})$ and $D_{\theta_t}(\theta^{(k)})$, respectively, at ten values of k across the epoch. We compare the two refusal sensitivity values at matched steps and report their ratio at the final step, calculated using Eq. 7 over 128 image-text prompts from $\mathcal{D}_h$.

Results. We report $D_{\theta_v}/D_{\theta_t}$ over one epoch for all nine model-task setups in Fig. 2, and the trajectories share a single shape. Early in training, the LLM backbone leads in terms of refusal sensitivity ($D_{\theta_v}/D_{\theta_t}$ smaller than one at the first probe), but the vision encoder overtakes it within the first 20% of the epoch and remains ahead thereafter in most settings. At its peak within the epoch, the encoder's refusal sensitivity exceeds the backbone's by three to six orders of magnitude.

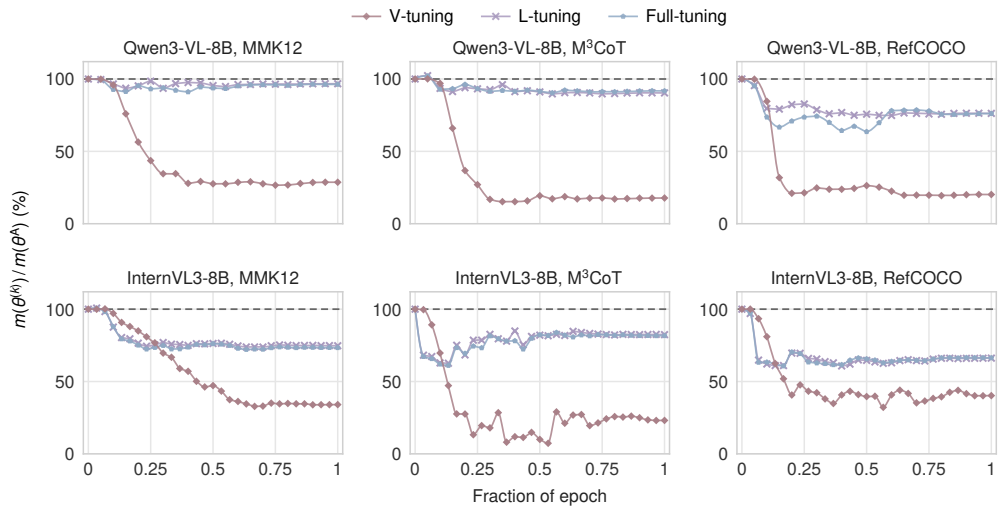


Figure 3: The percentage of refusal margin retained along one benign fine-tuning epoch, $\rho(\theta^{(k)})$ of Eq. 9 at the weights $\theta^{(k)}$ reached at step k. The dashed line marks the aligned margin. V-tuning collapses the margin early within the first 30 to 40% of the epoch.

Table 2: Refusal margin after one benign epoch under different tuning modes. V-tuning shrinks the refusal margin most, in every model and task. **Bold**: the smallest margin per model-task column.

	Qwen3-VL-8B			InternVL3-8B			Llama-3.2-11B-V		
Mode	MMK12	M³CoT	RefCOCO	MMK12	M³CoT	RefCOCO	MMK12	M³CoT	RefCOCO
None ($m(\theta^{\mathrm{A}})$)	46.16	46.16	46.16	82.34	82.34	82.34	3.49	3.49	3.49
V-tuning ($m(\theta_v^{(K)})$)	**13.24**	**8.20**	**9.32**	**28.04**	**19.04**	**33.15**	**3.18**	**2.80**	**2.42**
L-tuning ($m(\theta_t^{(K)})$)	44.66	41.72	35.20	61.52	67.84	54.48	3.28	3.21	2.96
Full-tuning ($m(\theta^{(K)})$)	44.45	42.36	35.14	60.42	67.24	54.65	3.22	3.16	2.90

These consistent results show that during the same benign fine-tuning, the vision encoder is the block with the greater capacity to disrupt the refusal structure, on all three models and all three tasks. We provide a more detailed analysis of the refusal sensitivity ratio in Appendix C. We further reveal the consequences of this higher refusal sensitivity by probing refusal margins in Section 3.3.

3.3 Refusal Margin Shrinkage

Refusal margin. In an aligned LVLM, along its refusal direction $\mathbf{u}_{\mathrm{ref}}$, the representations of harmful and benign multimodal inputs are well-separated. We quantify this separation by the *refusal margin*

$$m(\theta) = \mathop{\mathbb{E}}_{x\in\mathcal{D}_h}\left[\mathbf{u}_{\mathrm{ref}}^{\top} h_{\mathrm{ref}}(x;\theta)\right] - \mathop{\mathbb{E}}_{x\in\mathcal{D}_s}\left[\mathbf{u}_{\mathrm{ref}}^{\top} h_{\mathrm{ref}}(x;\theta)\right], \tag{8}$$

This separation between harmful and benign activations has been linked to effective refusal behavior. Specifically, a shrinking refusal margin suggests a weakening of the model's refusal mechanism, as harmful and benign representations become less separated along the refusal-mediating direction (Arditi et al., 2024; Wollschläger et al., 2025; Lin et al., 2024). The higher refusal sensitivity observed in Section 3.2 suggests that V-tuning could potentially disrupt LVLM's refusal structure through per-step perturbations along the refusal direction. We probe the refusal margins in LVLMs during fine-tuning to assess the cumulative effect of V-tuning on models' refusal structure.

Probing setup. Let $\theta_v^{(k)}$, $\theta_t^{(k)}$, and $\theta^{(k)}$ denote the weights reached after k steps of V-tuning, L-tuning, and full-tuning, respectively, with $\theta^{(0)} = \theta^{\mathrm{A}}$ and K denoting the number of steps in one epoch. We first measure the margin after one epoch, $m(\theta_v^{(K)})$, $m(\theta_t^{(K)})$, and $m(\theta^{(K)})$, and calculate the fraction of the aligned margin $m(\theta^{\mathrm{A}})$ that survives fine-tuning. We further trace this fraction along the epoch. For each mode we report the *retained margin*

$$\rho\big(\theta^{(k)}\big) = \frac{m\big(\theta^{(k)}\big)}{m\big(\theta^{\mathrm{A}}\big)}, \tag{9}$$

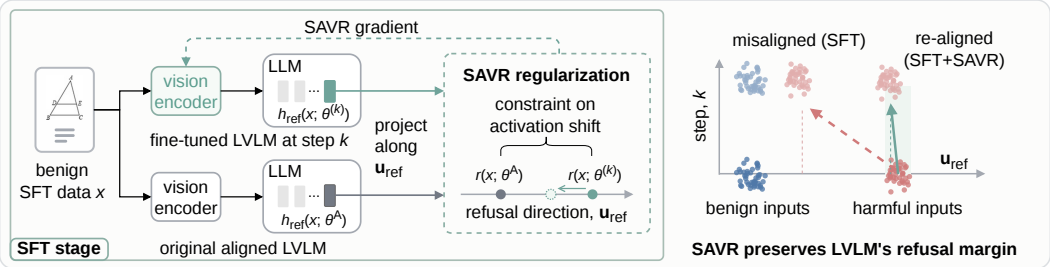


Figure 4: Illustration of SAVR, a regularizer for vision encoder parameter updates. SAVR alleviates the vision-sensitive vulnerability by constraining activation shifts along the refusal direction.

evaluated every 25 steps from $k = 0$ to K on the same harmful and benign inputs $\mathcal{D}_h$ and $\mathcal{D}_s$ at every probe. A value of ρ equal to one means the aligned separation is intact, whereas ρ near 0 indicates that harmful and benign inputs have become indistinguishable along $\mathbf{u}_{\text{ref}}$.

Results. Across all models and tasks, V-tuning shrinks the refusal margin substantially more than L-tuning and full-tuning that update the backbone, as shown in Table 2. Tracing the trajectory of fine-tuning, we find that V-tuning reduces the margin early in the epoch. Such refusal margin then remains collapsed, consistently smaller than those in other tuning modes across most steps (refer to Fig. 3). These results align with the high refusal sensitivity of the vision encoder observed in Section 3.2. Specifically, the larger update-induced responses along $\mathbf{u}_{\text{ref}}$ observed for the vision encoder are accompanied by greater reductions in harmful-benign separation under V-tuning.

3.4 Summary

Our V-tuning experiments show that benign updates to the vision encoder alone can weaken an LVLM's refusal behavior even when the frozen language backbone remains aligned. We investigate this new vulnerability through the vision-sensitivity hypothesis and complementary analyses of the refusal-mediating structure in the LVLM. Our studies reveal that: 1) vision encoder updates consistently induce larger perturbations along the LVLM's refusal direction; 2) such updates also cause the most substantial reduction in harmful-benign separation.

4 Preserving Refusal During Vision Encoder Updates

The findings in Section 3 motivate addressing the vision-sensitive vulnerability for better protection of LVLMs' safety during benign fine-tuning. We therefore introduce **SAVR** (**S**ensitivity-**A**ware **V**ision **R**egularization), a simple yet effective method constraining vision encoder updates that perturb representations along the LVLM's refusal direction (as illustrated in Fig. 4). SAVR targets only the vision encoder and is thus evaluated under V-tuning and full-tuning. We show that SAVR preserves LVLMs' refusal behavior while maintaining their downstream task utility. Such effectiveness further supports our vision-sensitivity hypothesis.

4.1 Method and Implementation

From per-step refusal sensitivity to accumulated drift. Our probing studies reveal substantial first-order perturbations along the refusal direction of the LVLM (*i.e.*, high refusal sensitivity) induced by updating the vision encoder, as defined in Eq. 6. Such perturbations accumulate along the fine-tuning trajectory, leading to a cumulative drift along the refusal direction:

$$r\big(x;\theta^{(k)}\big) - r\big(x;\theta^{\mathrm{A}}\big) \;\approx\; \sum_{j=0}^{k-1} \delta_{\theta_v}\big(x;\theta^{(j)}\big),$$

where k denotes the number of completed fine-tuning updates. Such accumulated drifts are necessary for the margin reductions observed in Section 3.3: without such drifts, the harmful-benign margin will remain unchanged. We therefore constrain the accumulated shifts of inputs' activations from the aligned model during fine-tuning. This prevents perturbations along the refusal direction from compounding and thus preserves the refusal margin in the fine-tuned LVLM.

Table 3: Results of SAVR under V-tuning and full-tuning. Δ utility (positive means higher than baseline values) and the retained margin (original→ with SAVR) are calculated relative to the original V-tuning/full-tuning run of the same setting. **Bold**: smaller ASR after applying SAVR.

	Qwen3-VL-8B			InternVL3-8B			Llama-3.2-11B-Vision		
	MMK12	M^{3}CoT	RefCOCO	MMK12	M^{3}CoT	RefCOCO	MMK12	M^{3}CoT	RefCOCO
V-tuning									
ASR (%), unprotected	45.3	55.8	27.7	51.0	73.3	48.3	35.7	53.2	55.7
ASR (%), + SAVR	**4.2**	**1.0**	**7.3**	**27.6**	**43.6**	**25.5**	**15.0**	**18.8**	**38.2**
Δ utility	+0.7	+6.6	+1.6	−0.2	−0.4	−0.8	+1.3	−0.9	−0.9
Margin retained (%)	29→91	18→99	20→64	34→98	23→46	40→99	91→98	80→98	69→91
Full-tuning									
ASR (%), unprotected	28.4	40.8	49.9	29.4	70.3	61.5	58.7	59.3	82.1
ASR (%), + SAVR	**5.5**	**34.6**	**13.6**	**19.2**	**69.2**	**61.4**	**28.3**	**56.0**	**77.6**
Δ utility	−0.4	−1.7	−0.8	+0.6	+5.8	+0.7	+2.7	−1.0	+0.8
Margin retained (%)	96→94	92→83	76→88	73→92	82→80	66→67	92→95	91→87	83→86

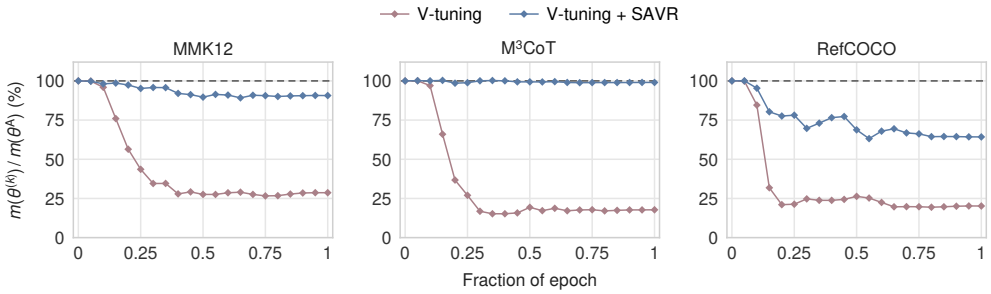


Figure 5: Refusal margin retained along one benign epoch under V-tuning, with and without SAVR, for Qwen3-VL-8B on each task. The dashed line marks the aligned margin.

SAVR training objective. We construct a regularizer for both V-tuning and full-tuning:

$$\mathcal{R}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{util}}}\left[\left(r(x;\theta) - r(x;\theta^{\text{A}})\right)^2\right]. \tag{10}$$

We combine this regularizer with the downstream task loss by applying the regularization gradient exclusively to the vision encoder and the task gradient to all trainable parameters. For each trainable parameter $w \in \theta$, the gradient supplied to the optimizer is

$$g_w \;=\; \nabla_w \mathcal{L}(\theta;\mathcal{D}_{\text{util}}) \;+\; \lambda \begin{cases} \nabla_w \mathcal{R}(\theta), & \text{if } w \in \theta_v, \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

where λ controls the regularization strength. Under V-tuning, only the encoder is updated, receiving both task and regularization gradients. Under full-tuning, the encoder receives the same two gradients, while the remaining parameters are updated using the task gradient alone. In our implementation, the aligned projections $r(x;\theta^{\text{A}})$ are pre-computed and cached, while the current projections $r(x;\theta)$ are obtained from the task forward pass at the same refusal direction.

4.2 Safety and Utility Evaluation

Setup. We use the same models, datasets, fine-tuning setup, and evaluation protocols as in Section 3.1 for V-tuning and full-tuning with SAVR. We tune λ per model-task setting for the best ASR and utility results relative to the original V-tuning and full-tuning.

Results. Under V-tuning, SAVR substantially reduces ASR in all nine settings by up to 98% relative to the original V-tuning, as shown in Table 3 (top). The utility across tasks is generally preserved or even improved with SAVR. Additionally, the retained margin rises to over 90% of the aligned margin in most settings. Our SAVR regularizer thus restores the separation collapsed by V-tuning. We further trace the margin along the epoch for Qwen3-VL, as shown in Fig. 5. With SAVR, the refusal margin stays near the aligned value throughout, while unprotected V-tuning collapses it within the first third of the epoch (more results are in Appendix D). Under full-tuning, as shown in

Table 4: Results of SAVR in composition with safety-data fine-tuning. Δ utility is calculated relative to the baseline (positive means higher than baseline values). **Bold**: ASR is lower than baseline.

	Qwen3-VL-8B			InternVL3-8B			Llama-3.2-11B-Vision		
	MMK12	M^3CoT	RefCOCO	MMK12	M^3CoT	RefCOCO	MMK12	M^3CoT	RefCOCO
SafeInstruction: text-only safety data (Bianchi et al., 2024)									
ASR (%), baseline	4.1	5.8	10.2	3.2	19.0	0.9	41.6	27.6	46.4
ASR (%), + SAVR	**2.5**	**1.0**	**3.1**	**3.1**	**16.8**	**0.0**	**26.1**	**19.5**	**45.5**
Δ utility	$+1.0$	-0.3	0.0	-0.5	-0.2	-0.1	$+0.7$	-0.3	$+0.4$
VLGuard: image-text safety data (Zong et al., 2024)									
ASR (%), baseline	12.5	16.8	34.8	15.3	27.1	18.4	26.4	16.6	49.7
ASR (%), + SAVR	**4.5**	**14.0**	**6.0**	**14.6**	**24.2**	**17.8**	**22.2**	**12.1**	**47.1**
Δ utility	$+3.4$	$+0.7$	$+0.4$	-3.0	$+0.7$	$+0.5$	$+0.4$	$+0.6$	$+0.5$

Table 3 (bottom), regularizing the encoder alone leads to smaller improvement compared to V-tuning settings. Nevertheless, SAVR still reduces ASR by up to 81% with utility preserved or improved.

Extending the scope of safeguarding LVLMs. The effectiveness of SAVR under V-tuning provides complementary evidence that the vision encoder training constitutes a distinct source of safety vulnerability in LVLMs. Its improvements under full-tuning further suggest that this vision-sensitive vulnerability also contributes to safety degradation when the entire LVLM is fine-tuned. Together, these findings highlight the need to extend safety safeguards from the language backbone to the entire multimodal pipeline during benign fine-tuning.

4.3 Composition with Safety-Data Fine-Tuning

Setup. An existing strategy for preserving safety during benign fine-tuning is to mix safety instruction examples into the downstream training data. We explore whether SAVR provides additional protection when combined with this strategy under full-tuning. We use two baselines (Bianchi et al., 2024; Zong et al., 2024) which are designed to reinforce alignment in the LLM backbone through safety supervision, orthogonal to SAVR's encoder-targeted regularization. We apply each baseline alone and with SAVR using the same fine-tuning setup.

Results. When combined with safety-data fine-tuning, SAVR further reduces ASR with negligible utility loss in most model-task settings (Table 4). For example, adding SAVR to SafeInstruction reduces ASR from 41.6% to 26.1% on Llama-3.2-Vision fine-tuned on MMK12, while adding it to VLGuard reduces ASR from 34.8% to 6.0% on Qwen3-VL fine-tuned on RefCOCO. Utility is preserved in both cases. Together with the findings in Section 4.2, these results underscore the value of holistically understanding and safeguarding LVLMs from a multimodal perspective.

5 Conclusion

In this paper, we highlight that safeguarding the language backbone is not enough to keep an LVLM safe under benign fine-tuning. This is due to a critical vision-sensitive vulnerability underlying existing LVLMs: V-tuning alone disrupts refusal across LVLMs and downstream tasks. We traced this vulnerability to LVLMs' refusal-mediating structure: 1) vision encoder updates create disruptions along the aligned model's refusal direction, up to six orders of magnitude more than LLM updates do; 2) such updates eventually collapse the harmful-benign refusal margin. Our proposed regularizer, SAVR, addresses this vulnerability by penalizing the accumulated shift that the encoder induces along the refusal direction. Under V-tuning and full-tuning, SAVR successfully restores the aligned refusal margin and lowers ASR across models and tasks, while maintaining the original utility performance. SAVR also demonstrates its potential in composing with other approaches such as safety-data fine-tuning, lowering ASR further without significant utility loss. Future work can focus on extending SAVR to other training regimes, such as reinforcement-learning-based fine-tuning, and developing more extensive safeguards that preserve safety across the entire multimodal pipeline.

AI DISCLOSURE

In this work, we used generative AI tools to implement and debug code for experiments and interpret results. We have *not* used generative AI tools for developing the theoretical framework, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in writing proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, or supporting qualitative and thematic data analysis; generating synthetic datasets, translation, and cleaning/reformatting datasets are not applicable to this work. Additionally, we used generative AI tools to aid writing, search for relevant literature, and create or modify scientific figures. We have reviewed all AI-assisted work. We verify the code correctness through manual review and testing by one of the authors. We also carefully check and manually edit artifacts produced by AI for quality and correctness. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to improve the safety of LVLMs by identifying and mitigating refusal degradation during benign fine-tuning. All experiments use public models and datasets, and we do not release fine-tuned checkpoints with degraded safety. Our findings on the vision-sensitive vulnerability have dual-use implications: understanding how vision encoder updates weaken refusal could also inform malicious users' attempts to weaken the safety alignment in LVLMs. We thus call for responsible disclosure and use of these findings, and we release SAVR so that the vulnerability is reported together with a mitigation. Our goal is to help researchers and safety practitioners identify and address these weaknesses before they are exploited, and to encourage safeguards that cover the entire multimodal pipeline rather than the language backbone alone.

REPRODUCIBILITY STATEMENT

We have taken several steps to make our results reproducible. All experiments use publicly available models (Qwen3-VL-8B, InternVL3-8B, and Llama-3.2-11B-Vision) and publicly available downstream datasets (MMK12, M^3CoT, and RefCOCO), and the main paper specifies the fine-tuning modes, training recipe, and evaluation protocols for both safety and utility. The appendices and supplementary materials provide the remaining details needed to rerun each experiment. The core idea of SAVR is specified in the main paper, with its code implementation in the supplementary materials. We will further open-source our codebase, including the fine-tuning scripts, the refusal direction and margin probes, and the training and evaluation pipeline.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *NeurIPS*, 2024.

Saketh Bachu, Erfan Shayegani, Rohit Lal, Trishna Chakraborty, Arindam Dutta, Chengyu Song, Yue Dong, Nael B. Abu-Ghazaleh, and Amit Roy-Chowdhury. Layer-wise alignment: Examining safety alignment across image encoder layers in vision language models. In *ICML*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Rottger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. In *ICLR*, 2024.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *ACL*, 2024.

In Chong Choi, Jiacheng Zhang, Feng Liu, and Yiliao Song. Multi-turn adaptive prompting attack on large vision-language models. *CoRR*, abs/2602.14399, 2026.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: Safe reinforcement learning from human feedback. In *ICLR*, 2024.

Yi Ding, Lijun Li, Bing Cao, and Jing Shao. Rethinking bottlenecks in safety fine-tuning of vision language models. In *ICLR*, 2026.

Lei Gao, Yue Niu, Tingting Tang, Salman Avestimehr, and Murali Annavaram. Ethos: Rectifying language models in orthogonal parameter space. In *NAACL Findings*, 2024.

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. In *AAAI*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

Idhant Gulati and Shivam Raval. Narrow fine-tuning erodes safety alignment in vision-language agents. *CoRR*, abs/2602.16931, 2026.

Jiahe Guo, Xiangran Guo, Jiaxuan Chen, Weixiang Zhao, Yanyan Zhao, Yutai Hou, Qianchao Wang, Dandan Tu, and Bing Qin. Safety geometry collapse in multimodal llms and adaptive drift correction. *CoRR*, abs/2605.18104, 2026.

Shuyang Hao, Bryan Hooi, Jun Liu, Kai-Wei Chang, Zi Huang, and Yujun Cai. Exploring visual vulnerabilities via multi-loss adversarial search for jailbreaking vision-language models. In *CVPR*, 2025.

Samyak Jain, Ekdeep Singh Lubana, Kemal Oksuz, Tom Joy, Philip Torr, Amartya Sanyal, and Puneet K. Dokania. What makes and breaks safety fine-tuning? a mechanistic study. In *NeurIPS*, 2024.

Joonhyun Jeong, Seyun Bae, Yeonsung Jung, Jaeryong Hwang, and Eunho Yang. Playing the fool: Jailbreaking llms and multimodal llms with out-of-distribution strategy. In *CVPR*, 2025.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: towards improved safety alignment of llm via a human-preference dataset. In *NeurIPS*, 2023.

Jiaming Ji, Xinyu Chen, Rui Pan, Han Zhu, Jiahao Li, Donghai Hong, Boyuan Chen, Jiayi Zhou, Kaile Wang, Juntao Dai, Chi-Min Chan, Sirui Han, Yike Guo, and Yaodong Yang. Safe rlhf-v: Safe reinforcement learning from multi-modal human feedback. In *NeurIPS*, 2025.

Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *NeurIPS*, 2019.

Seongyun Lee, Geewook Kim, Jiyeon Kim, Hyunji Lee, Hoyeon Chang, Sue Hyun Park, and Minjoon Seo. How does vision-language adaptation impact the safety of vision language models? In *ICLR*, 2025.

Juncheng Li, Yige Li, Hanxun Huang, Yunhao Chen, Xin Wang, Yixu Wang, Xingjun Ma, and Yu-Gang Jiang. Backdoorvlm: A benchmark for backdoor attacks and defenses on vision-language models. *CoRR*, abs/2511.18921, 2025.

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. In *ECCV*, 2024.

Yuping Lin, Pengfei He, Han Xu, Yue Xing, Makoto Yamada, Hui Liu, and Jiliang Tang. Towards understanding jailbreak attacks in llms: A representation space analysis. In *EMNLP*, 2024.

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

Mingzu Liu, Hao Fang, and Runmin Cong. Tcap: Tri-component attention profiling for unsupervised backdoor detection in mllm fine-tuning. In *ICML*, 2026.

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. In *ACL Findings*, 2025a.

Qin Liu, Fei Wang, Chaowei Xiao, and Muhao Chen. Vlm-guard: Safeguarding vision-language models via fulfilling safety alignment gap. *CoRR*, abs/2502.10486, 2025b.

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. In *ECCV*, 2024a.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards safer large language models through machine unlearning. In *ACL Findings*, 2024b.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *CoRR*, abs/2503.07365, 2025.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Georgios Pantazopoulos, Amit Parekh, Malvina Nikandrou, and Alessandro Suglia. Learning to see but forgetting to follow: Visual instruction tuning makes llms more prone to jailbreak attacks. In *Proceedings of Safety4ConvAI: The Third Workshop on Safety for Conversational AI @ LREC-COLING 2024*, 2024.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *NeurIPS*, 2020.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *CoRR*, abs/2310.03693, 2023.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *AAAI*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Xuankun Rong, Wenke Huang, Jian Liang, Jinhe Bi, Xun Xiao, Yiming Li, Bo Du, and Mang Ye. Backdoor cleaning without external guidance in mllm fine-tuning. In *NeurIPS*, 2025.

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

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf. In *ACL Findings*, 2024.

Xijia Tao, Shuai Zhong, Lei Li, Qi Liu, and Lingpeng Kong. Imgtrojan: Jailbreaking vision-language models with one image. In *NAACL*, 2025.

Ruofan Wang, Xingjun Ma, Hanxu Zhou, Chuanjun Ji, Guangnan Ye, and Yu-Gang Jiang. White-box multimodal jailbreaks against large vision-language models. In *ACM MM*, 2024.

Ziqi Wang, Chang Che, Qi Wang, Hui Ma, Zenglin Shi, Cees G. M. Snoek, and Meng Wang. Harmonious parameter adaptation in continual visual instruction tuning for safety-aligned mllms. In *CVPR*, 2026.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. In *ICML*, 2025.

Shicheng Xu, Liang Pang, Yunchang Zhu, Huawei Shen, and Xueqi Cheng. Cross-modal safety mechanism transfer in large vision-language models. In *ICLR*, 2025.

Wenhan Yang, Spencer Stice, Ali Payani, and Baharan Mirzasoleiman. Bootstrapping llm robustness for vlm safety via reducing the pretraining modality gap. *CoRR*, abs/2505.24208, 2025a.

Zuopeng Yang, Jiluan Fan, Anli Yan, Erdun Gao, Xin Lin, Tao Li, Kanghua Mo, and Changyu Dong. Distraction is all you need for multimodal large language model jailbreaking. In *CVPR*, 2025b.

Zonghao Ying, Aishan Liu, Tianyuan Zhang, Zhengmin Yu, Siyuan Liang, Xianglong Liu, and Dacheng Tao. Jailbreak vision language models via bi-modal adversarial prompt. *IEEE TIFS*, 2025.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, 2016.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *CVPR*, 2024.

Haoyu Zhang, Yangyang Guo, and Mohan Kankanhalli. Overloading large vision-language models for jailbreaking. *CoRR*, abs/2607.02961, 2026.

Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, Feng Zhao, Tao Gui, and Jing Shao. Spa-vl: A comprehensive safety preference alignment dataset for vision language models. In *CVPR*, 2025.

Shiji Zhao, Ranjie Duan, Fengxiang Wang, Chi Chen, Caixin Kang, Shouwei Ruan, Jialing Tao, YueFeng Chen, Hui Xue, and Xingxing Wei. Jailbreaking multimodal large language models via shuffle inconsistency. In *ICCV*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479, 2025.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. In *ICML*, 2024.

Quanchen Zou, Moyang Chen, Zonghao Ying, Wenzhuo Xu, Yisong Xiao, Deyue Zhang, Dongdong Yang, Zhao Liu, and Xiangzheng Zhang. Reasoning-oriented programming: Chaining semantic gadgets to jailbreak large vision language models. *CoRR*, abs/2603.09246, 2026.

A Additional Results of V-Tuning and Discussions

Utility metrics and results. We report the utility performance of every LVLM on its downstream task after fine-tuning with different modes in Table 5. For utility evaluation, we utilize $1{,}000$ held-out test examples from each downstream dataset and calculate each task's own metric: answer match for MMK12, METEOR for RefCOCO, and accuracy for M^3CoT, respectively.

Discussion on ASR results. V-tuning yields higher or comparable ASRs in most settings. The few settings with lower ASRs suggest additional mechanisms contributing to safety degradation, with relative strength that varies across models, tasks, and tuning modes. They do not weaken our central finding that benign vision encoder updates alone can substantially compromise safety despite a frozen language backbone, and V-tuning disrupts LVLMs' refusal-mediating structure the most. How these mechanisms interact is a direction for future work.

B Refusal Direction Estimation

Candidate selection. We use $h_p^\ell(x;\theta) \in \mathbb{R}^d$ to denote LVLM's residual-stream activation at layer ℓ and token position p, and $h_{\mathrm{ref}}(x;\theta) = h_{p^\star}^{\ell^\star}(x;\theta)$ refers to the activation at the selected site with specific layer and token position. Following previous work (Arditi et al., 2024), we explore candidate positions that are the chat-template tokens that close the user turn and open the assistant turn, indexed by $p \in \{-5,\ldots,-1\}$ from the end of the prompt, where $p=-1$ is the token immediately before generation. We also explore candidate layers which are below $0.8L$ for a backbone of depth L. For each candidate site, the candidate refusal direction $\mathbf{u}_p^\ell$ of Eq. 2 is computed on 128 harmful image-text instructions from HADES (Li et al., 2024) and 128 benign image-text instructions from LLaVA-Instruct (Liu et al., 2023), which also serve as $\mathcal{D}_h$ and $\mathcal{D}_s$ in Sections 3.2 and 3.3.

Causal criteria. Each candidate refusal direction is scored on three held-out splits of 32 harmful and 32 benign prompts by measuring how three interventions affects the model's *refusal score* (*i.e.*, the log-odds that the first generated token begins a refusal)Arditi et al. (2024): 1) *Bypass*: the projection onto $\mathbf{u}_p^\ell$ is removed from the residual stream at every layer and position, and the refusal score is averaged over harmful prompts; a direction that mediates refusal drives it negative; 2) *Induce*: the

Table 5: Downstream utility of the models in Table 1 (answer match for MMK12, METEOR for RefCOCO, accuracy for M^{3}CoT). *None* is the aligned model.

	Qwen3-VL-8B			InternVL3-8B			Llama-3.2-11B-V		
Mode	MMK12	M^{3}CoT	RefCOCO	MMK12	M^{3}CoT	RefCOCO	MMK12	M^{3}CoT	RefCOCO
None	38.1	62.1	13.8	49.3	51.6	15.4	20.6	51.5	9.8
V-tuning	55.6	56.8	22.0	50.3	52.4	16.4	21.9	52.9	22.6
L-tuning	51.5	66.2	21.2	55.6	63.9	21.5	27.3	52.4	20.3
Full-tuning	50.7	70.1	21.0	55.9	61.7	20.5	25.5	52.7	19.9

unnormalized difference-of-means vector is added to the residual stream at layer ℓ, and the refusal score is averaged over benign prompts; a refusal direction drives it positive; 3) *KL*: the mean KL divergence between the next-token distributions with and without the bypass ablation on benign prompts, which rules out directions that suppress refusal by damaging the model. A candidate is retained only if its KL score is below 0.1 and its induce score is positive on every split, and among retained candidates the one with the lowest bypass score is selected. Table 6 lists the selected site per model. The refusal direction $\mathbf{u}_{\mathrm{ref}}$ at this site is estimated once on the aligned model θ^{A} and held fixed thereafter; after fine-tuning, we repeat the bypass ablation on the tuned model against a random-direction control to confirm that $\mathbf{u}_{\mathrm{ref}}$ still suppresses refusal when ablated.

C Two-Factor Decomposition of the Refusal Sensitivity Ratio

The refusal sensitivity in Section 3.2 depends on both the size of a parameter update and how strongly that update moves the probed activation along the refusal direction. We separate these two factors to interpret the ratio in Fig. 2. Following Section 3.2, $\theta^{(k)}$ denotes the current weights in the corresponding V-tuning or L-tuning run, and Δ_b denotes that run's update at step k. For $\|\Delta_b\|_2 > 0$, let $\hat{\Delta}_b = \Delta_b / \|\Delta_b\|_2$ and define

$$S_b\big(\theta^{(k)}\big) = \frac{D_b(\theta^{(k)})}{\|\Delta_b\|_2^2} = \mathop{\mathbb{E}}_{x \in \mathcal{D}_h}\left[\left(\mathbf{u}_{\mathrm{ref}}^{\top} J_b(x; \theta^{(k)}) \hat{\Delta}_b\right)^2\right]. \tag{12}$$

This quantity measures the mean squared first-order shift along the refusal direction after normalizing the parameter update to unit length. Omitting the step notation for readability, Eq. 7 gives

$$\frac{D_{\theta_v}}{D_{\theta_t}} = \underbrace{\frac{\|\Delta_{\theta_v}\|_2^2}{\|\Delta_{\theta_t}\|_2^2}}_{\text{parameter update size}} \times \underbrace{\frac{S_{\theta_v}}{S_{\theta_t}}}_{\text{response per unit update}}, \tag{13}$$

where both factors compare squared quantities, and the ratio is defined when both updates are nonzero and $D_{\theta_t} > 0$. We measure these two factors and plot their trends across fine-tuning steps.

Results and interpretation. Figure 6 shows that the update-size factor rises above one in all nine settings. At the final probe, the response-per-unit-update factor exceeds one only for Qwen3-VL on MMK12 and M^{3}CoT, and InternVL3 on M^{3}CoT. This indicates that the higher refusal sensitivity under V-tuning is mostly driven by larger vision encoder parameter updates, as the larger updates often outweigh a weaker response per unit of update during the fine-tuning steps.

D Margin Trajectories under SAVR

We show the refusal margin results ($m(\theta^{(k)})/m(\theta^{\mathrm{A}})$) under SAVR across steps for two other LVLMs in Fig. 7. SAVR holds the margin near the aligned value for the whole epoch in most settings, ending above 90%, while the undefended V-tuning curve in the same panel collapses quickly within the first third of the epoch.

E Additional Results and Details

Kindly refer to our supplementary material (submitted as a separate PDF file zipped with code) for more results and technical details.

Table 6: Selected probing site of each model and the causal scores of its refusal direction on the aligned model (means over three held-out splits). The refusal score is a log-odds, so bypass should be negative and induce positive.

Model	Backbone depth L	Layer $\ell^\star$	Position $p^\star$	Bypass	Induce	KL
Qwen3-VL-8B	36	21	-2	-5.77	4.69	0.043
InternVL3-8B	28	21	-1	-4.63	2.39	0.079
Llama-3.2-11B-Vision	40	12	-5	-2.54	1.05	0.023

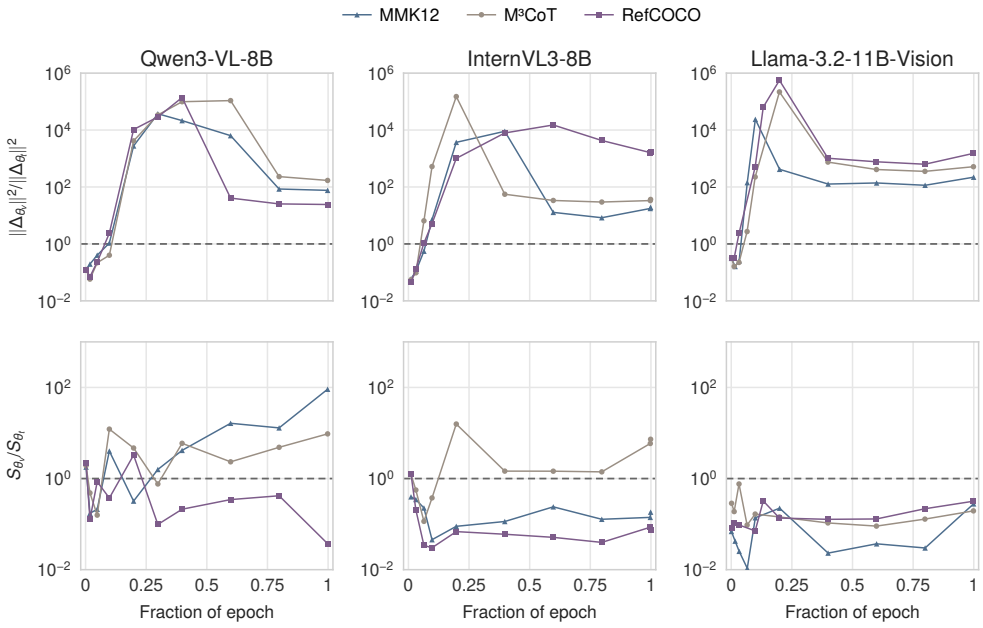


Figure 6: Two factors of the refusal sensitivity ratio: the ratio of squared parameter-update lengths (top) and the ratio of mean squared refusal-direction responses per unit update (bottom). Their product gives the sensitivity ratio. The dashed line marks a ratio of one.

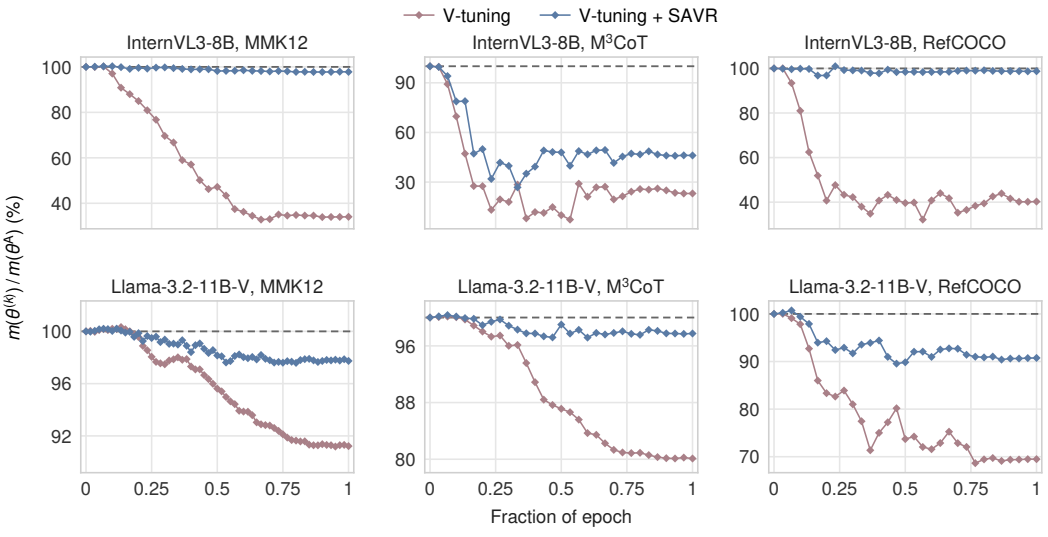


Figure 7: Refusal margin retained along one benign epoch under V-tuning, with and without SAVR, for InternVL3-8B and Llama-3.2-11B-Vision at the certified site of each model. The SAVR curves are the V-tuning arms reported in Table 3. The dashed line marks the aligned margin.