AMPLIFY TO DETECT: GRADIENT-GUIDED WEIGHT INTERVENTIONS FOR MULTIMODAL JAILBREAK DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models (VLMs) can be jailbroken by malicious intent conveyed through text, images, or their interaction. Detecting such inputs before harmful generation is therefore important, but detectors must also avoid falsely flagging benign requests. Existing hidden-state detectors use VLM representations as fixed features, leaving unclear which internal computations produce harmful–benign separation and whether these signals can be strengthened. We introduce *Amplify to Detect*, a frozen-model intervention for multimodal jailbreak detection. Our method places multiplicative scalers on output channels of VLM linear projections and uses first-order sensitivity of a paired harmful–benign representation-separation objective to identify influential channels. We amplify channels predicted to increase separation, then apply hidden-state detectors to the resulting representations. Across three open-weight VLMs and multiple detectors, targeted amplification generally improves malicious-intent detection under a stringent 5% false-positive-rate constraint. For HiddenDetect, TPR at this operating point increases by 18.8–57.2 percentage points across the three models. Direction ablations further show that reversing the predicted intervention direction reverses its effect, while random would decrease the performance. These findings suggest that safety-relevant hidden-state signals arise partly from identifiable internal computations and can be selectively strengthened for more reliable jailbreak detection.

1 INTRODUCTION

Vision–language models (VLMs) extend LLM capabilities by jointly processing images and language, but their multimodal interface also broadens the attack surface of safety alignment. Malicious intent can pass alignment through visual content, text, or their interaction, allowing jailbreak attacks (Qi et al., 2024; Shayegani et al., 2024; Li et al., 2024; Yang et al., 2025). A useful defense must therefore distinguish adversarial requests from benign ones without producing excessive false alarms or over-refusal. Existing jailbreak defenses operate at several levels, including direct finetuning, input- and output-level detection, cross-modal verification, and model-side safety intervention (Zhang et al., 2025; Xu et al., 2024; Yin et al., 2026; Wang et al., 2025; Zou et al., 2026). Specifically, a line of work (Jiang et al., 2025; Hua et al., 2026; Gan et al., 2026; Zhang et al., 2026; Xu et al., 2024) instead monitors the target model's internal representations. These methods make detection decisions from the target model's hidden states. Their performance, however, can vary with the operating point: under stringent false-positive-rate constraints, detectors such as KCD and MCD show substantial variation (Hua et al., 2026). This motivates us to examine the representations underlying those decisions: *Can the separation between malicious and benign inputs be attributed to a small set of projection output channels, and can selectively amplifying those channels strengthen that separation?*

Prior work shows that interventions on individual attention heads can alter model safety behavior (Zhou et al., 2025; Xia et al., 2026). Motivated by these findings, we identify projection output channels that contribute to harmful–benign separation and rank them using SNIP-inspired sensitivity (Lee et al., 2019; Orgad et al.; Wei et al., 2024). Specifically, we introduce multiplicative scale over projection-output channels and measure the first-order sensitivity of a harmful–benign separation objective to each scaler. We amplify these top positive-score channels while keeping the

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.

original VLM weights frozen, then use the resulting representation for jailbreak detection. With extensive experiments across three open-weight VLMs and multiple detectors, we show that targeted amplification consistently improves malicious-intent detection under a stringent 5% false-positive-rate (FPR) constraint. For HiddenDetect (Jiang et al., 2025) with the proposed method, the TPR at this operating point increases by 18.8–57.2 percentage points across the three models. Our results show that safety-relevant hidden-state signals are not only detectable but also selectively controllable. Targeted channel amplification improves low-FPR jailbreak detection. In addition, direction ablations further show that reversing the signed intervention direction reverses its effect, while random would decrease the performance. These results support the view that VLM safety signals arise from structured internal computations that can be identified and causally strengthened, rather than from an arbitrary post-hoc decision boundary. In summary, our contributions are:

- We introduce a causal perspective on hidden-state jailbreak detection: rather than treating internal representations as fixed features, we identify and intervene on the channels that contribute to safety-relevant harmful–benign separation.
- We formulate structured, signed channel attribution as a SNIP-inspired first-order sensitivity analysis. This formulation connects a local prediction about internal computation to a targeted representation-level intervention, while preserving the frozen VLM.
- We show that the resulting intervention improves safety-relevant separation and low-false-positive-rate jailbreak detection, and that its effect follows the predicted signed direction across controlled interventions, detector objectives, and VLM families.

2 Related Works

Jailbreak Attacks and Detection. Multimodal jailbreaks exploit the additional attack surface introduced by visual inputs. Prior work demonstrates that adversarial images can bypass language-model safeguards and that malicious intent can be distributed across image–text combinations (Qi et al., 2024; Shayegani et al., 2024). Systematic evaluations further highlight the need to assess both attack resistance and performance on benign requests (Weng et al., 2025).

Detection methods draw on complementary signals, including cross-modal relationships, parameter gradients, and internal activations. CIDER examines cross-modal similarity to detect adversarial image inputs (Xu et al., 2024), while GradSafe identifies jailbreak prompts through gradient patterns in safety-critical parameters (Xie et al., 2024). Within VLMs, HiddenDetect exploits safety-relevant patterns in intermediate hidden states (Jiang et al., 2025). RCS further learns a projection of internal representations and applies contrastive statistical scores to distinguish malicious intent from benign novelty (Hua et al., 2026). These findings establish internal representations as useful signals for jailbreak detection. Our work examines a complementary question: whether attributing harmful–benign separation to upstream channels can guide an intervention that improves the representations supplied to a detector, particularly under a low false-positive-rate constraint.

Mechanistic Interpretability and Internal Safety Intervention. Mechanistic studies connect safety behavior to identifiable directions and components within language models. Arditi et al. (2024) show that manipulating a single direction in residual-stream activations can induce or suppress refusal in the models they study. At the component level, Zhou et al. (2025) identify attention heads whose ablation substantially compromises safety. Related evidence extends to VLMs: One-head Defense identifies a critical attention head that provides a discriminative signal for unsafe inputs (Xia et al., 2026), while ASTRA uses adaptive activation steering to mitigate harmful generation (Wang et al., 2025). Together, these studies motivate examining how internal computations contribute to safety-relevant behavior and detection signals.

Sensitivity-based attribution offers a complementary perspective on the internal mechanisms of model safety. Building on the connection sensitivity criterion introduced by SNIP (Lee et al., 2019), Wei et al. (2024) show that pruning a small subset of safety-critical weights can substantially weaken safety alignment while largely preserving general capabilities. Conversely, Orgad et al. use signed sensitivity scores to identify weights supporting harmful generation, whose removal suppresses harmful responses. Our work draws on this motivation but shifts the attribution target from response generation to representation separability. This enables attribution-guided amplification of

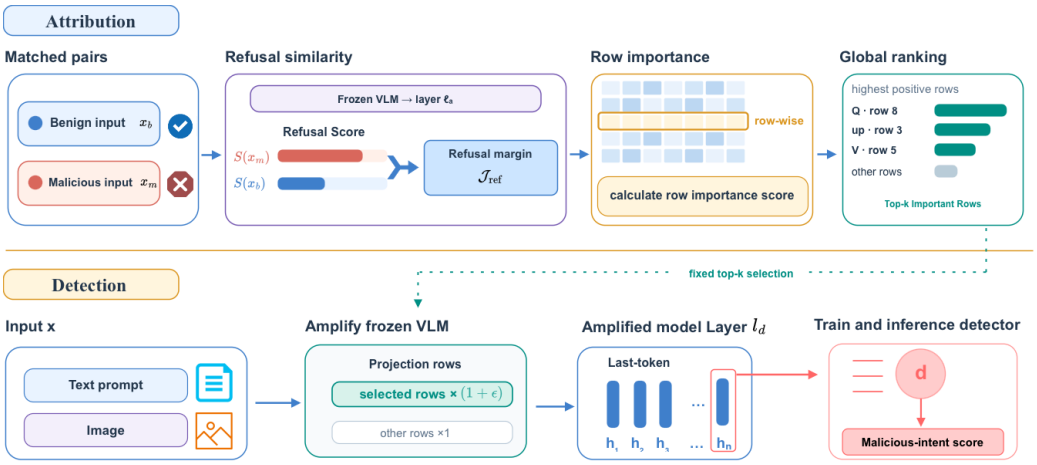


Figure 1: Overview of our proposed *Amplify to Detect* pipeline. It aims to identify the influential output channels of the frozen VLM linear projections followed by amplifying them to enhance hidden-state detectors for better jailbreak detection. Top (Section 3.3): matched benign and malicious inputs define a refusal-score margin at target layer ℓ_a to rank influential channels. Bottom (Section 3.4): these rows are scaled by $1+\epsilon$; the amplified model's last-token state at readout layer ℓ_d is then used for training and inference with existing hidden-state detectors.

detection-relevant representations. The central question is therefore whether sensitivity analysis can identify not only components governing safety behavior, but also interventions that make harmful inputs more distinguishable to a downstream detector.

3 METHOD

To improve detection under stringent false-positive-rate constraints, we seek to strengthen the malicious–benign signal in the hidden states supplied to a detector. In this section, we first present the preliminaries and problem formulation in Section 3.1 and Section 3.2, respectively. Figure 1 summarizes the procedure of the proposed *Amplify to Detect* pipeline. During attribution (Section 3.3), matched benign and malicious inputs are passed through the frozen VLM, and their refusal scores at target layer ℓ_a define a separation margin. We use this margin to rank projection output channels and select a fixed set with positive scores. During detection (Section 3.4), we amplify the selected channels and pass the last-token hidden state at readout layer ℓ_d to the downstream detector, such as HiddenDetect (Jiang et al., 2025), which follows its original training and scoring procedure.

3.1 PRELIMINARIES AND NOTATION

Let F_{θ_0} denote a frozen vision-language model with L decoder layers indexed by $0, \ldots, L-1$. An input is denoted by $x = (t, v)$, where t is the text prompt and v is an optional image. For $\ell \in \{0, \ldots, L\}$, let $\mathbf{h}^{(\ell)}(x)$ denote the last-token hidden representation after decoder layers $0, \ldots, \ell-1$.

Each decoder layer contains linear projections in its self-attention and MLP submodules. We consider the query, key, value, and output projections in self-attention, and the gate, up, and down projections in the MLP. Let $\mathcal{R}$ denote the set of these projections across all decoder layers. For a projection $r \in \mathcal{R}$, we write

$$\mathbf{z}_r = W_r \mathbf{a}_r, \tag{1}$$

where W_r is the linear projector, and $\mathbf{a}_r$ and $\mathbf{z}_r$ are its input and output activations, respectively. We refer to each coordinate $z_{r,c}$ as an *output channel*. Thus, channel (r, c) corresponds to the cth output dimension of projection r.

We distinguish the attribution target layer ℓ_a, at which the refusal-margin objective is evaluated, from the detector readout layer ℓ_d, from which hidden representations are supplied to a downstream detector. Let $\lambda(r)$ denote the source layer containing projection r, and let d_r^{out} denote its number of

output channels. For attribution target ℓ_a, we define the set of all upstream channels as

$$\mathcal{C}_{<\ell_a} = \left\{(r,c) : r \in \mathcal{R},\ \lambda(r) < \ell_a,\ 1 \le c \le d_r^{\text{out}}\right\}. \tag{2}$$

Channel ranking and selection are performed globally over $\mathcal{C}_{<\ell_a}$, rather than independently within each source layer.

For channel attribution, we use a paired dataset

$$\mathcal{D}_{\text{attr}} = \left\{(x_i^b, x_i^m)\right\}_{i=1}^{N_{\text{attr}}}, \tag{3}$$

where x_i^b and x_i^m denote the benign and malicious members of the ith matched pair.

3.2 Problem Formulation: Channel Scaling and Attribution

We formulate our method as a structured channel attribution procedure inspired by SNIP (Lee et al., 2019). Rather than assigning importance scores to individual scalar weights, we score complete projection output channels.

For each channel $(r,c) \in \mathcal{C}_{<\ell_a}$, we introduce a scaling parameter $\epsilon_{r,c}$:

$$\widetilde{z}_{r,c} = (1+\epsilon_{r,c}) z_{r,c}. \tag{4}$$

The original model corresponds to $\boldsymbol{\epsilon} = \mathbf{0}$. Equivalently, the intervention scales the corresponding row of the projection matrix as

$$W_{r,c:} \leftarrow (1+\epsilon_{r,c}) W_{r,c:}, \tag{5}$$

where $W_{r,c:}$ is corresponded to the cth row of the projector W_r.

Let $\mathcal{J}(\boldsymbol{\epsilon})$ denote a differentiable attribution objective. We define the signed attribution score of channel (r,c) as

$$s_{r,c} = \left.\frac{\partial \mathcal{J}(\boldsymbol{\epsilon})}{\partial \epsilon_{r,c}}\right|_{\boldsymbol{\epsilon}=\mathbf{0}} = \left\langle W_{r,c:}, \frac{\partial \mathcal{J}}{\partial W_{r,c:}} \right\rangle, \tag{6}$$

Let $\mathbf{e}_{r,c}$ denote the coordinate basis vector corresponding to channel (r,c). For a scalar perturbation δ, the first order expansion is

$$\mathcal{J}(\delta \mathbf{e}_{r,c}) = \mathcal{J}(\mathbf{0}) + \delta s_{r,c} + o(\delta). \tag{7}$$

For $\delta > 0$, a positive score therefore predicts that locally amplifying channel (r,c) increases the attribution objective.

3.3 Refusal-Margin Attribution Objective

We instantiate the attribution objective using a refusal signal inspired by HiddenDetect (Jiang et al., 2025). At attribution target layer ℓ_a, we map the hidden representation into vocabulary space using the model's frozen final normalization and language-model head:

$$\mathbf{q}_{\ell_a}(x;\boldsymbol{\epsilon}) = \text{LMHead}\Big(\text{Norm}\Big(\mathbf{h}_{\boldsymbol{\epsilon}}^{(\ell_a)}(x)\Big)\Big). \tag{8}$$

We then define the refusal similarity score as

$$S_{\ell_a}(x;\boldsymbol{\epsilon}) = \cos(\mathbf{q}_{\ell_a}(x;\boldsymbol{\epsilon}), \mathbf{v}_{\text{ref}}), \tag{9}$$

where $\mathbf{v}_{\text{ref}}$ is a fixed vocabulary-space refusal vector constructed from refusal related tokens following Jiang et al. (2025). A larger value indicates a stronger refusal signal in the hidden representation.

Using the matched attribution pairs, we define the refusal margin at target layer ℓ_a as

$$\mathcal{J}_{\text{ref}}^{(\ell_a)}(\boldsymbol{\epsilon}) = \frac{1}{N_{\text{attr}}} \sum_{i=1}^{N_{\text{attr}}} \left[S_{\ell_a}(x_i^m;\boldsymbol{\epsilon}) - S_{\ell_a}(x_i^b;\boldsymbol{\epsilon})\right]. \tag{10}$$

A larger value of $\mathcal{J}_{\text{ref}}^{(\ell_a)}$ indicates greater separation between malicious and benign inputs along the refusal direction.

The attribution score for each upstream channel is therefore

$$s_{r,c}^{(\ell_a)} = \left.\frac{\partial \mathcal{J}_{\text{ref}}^{(\ell_a)}(\boldsymbol{\epsilon})}{\partial \epsilon_{r,c}}\right|_{\boldsymbol{\epsilon}=\mathbf{0}}, \qquad (r,c) \in \mathcal{C}_{<\ell_a}. \tag{11}$$

A positive $s_{r,c}^{(\ell_a)}$ indicates that locally amplifying channel (r,c) increases the refusal margin at the attribution target layer. We refer the readers to Appendix C for the detailed derivation.

3.4 Global Channel Selection, Amplification, and Detection

Given an attribution target layer ℓ_a and a selection fraction p, we rank all channels in $\mathcal{C}_{<\ell_a}$ jointly according to their attribution scores. We set

$$k = \lceil p\,|\mathcal{C}_{<\ell_a}| \rceil \tag{12}$$

and select the highest-scoring positive channels:

$$\mathcal{M}^{+}_{\ell_a}(p) = \mathrm{Top}_k \left\{ (r,c) \in \mathcal{C}_{<\ell_a} : s^{(\ell_a)}_{r,c} > 0 \right\}. \tag{13}$$

The ranking is performed once over the complete upstream channel set $\mathcal{C}_{<\ell_a}$, without imposing layerwise or projection-wise selection quotas.

Let $\epsilon > 0$ denote the amplification factor. We define the final channel perturbation as

$$\epsilon^{*}_{r,c} = \begin{cases} \epsilon, & (r,c) \in \mathcal{M}^{+}_{\ell_a}(p), \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Thus, selected channels are multiplied by ϵ, while all remaining channels are unchanged. The selected channel set and amplification factor are fixed thereafter.

For any layer ℓ, we denote the hidden representations produced by the original and amplified models as

$$\mathbf{h}^{(\ell)}_{\mathrm{orig}}(x) = \mathbf{h}^{(\ell)}_{\mathbf{0}}(x), \qquad \mathbf{h}^{(\ell)}_{\mathrm{amp}}(x) = \mathbf{h}^{(\ell)}_{\boldsymbol{\epsilon}^*}(x). \tag{15}$$

For a downstream detector with readout layer ℓ_d, the amplified model is used for hidden-state extraction throughout the detection pipeline, replacing $\mathbf{h}^{(\ell_d)}_{\mathrm{orig}}(x)$ with $\mathbf{h}^{(\ell_d)}_{\mathrm{amp}}(x)$. The detector itself follows its original training and inference procedure.

4 Experiments

4.1 Setting & Results

Models. Following Hua et al. (2026), we evaluate three open-weight vision-language models: LLaVA-v1.6-Vicuna-7B (Liu et al., 2023), Qwen2.5-VL-7B-Instruct (Bai et al., 2025), and InternVL3-8B-Instruct (Zhu et al., 2025).

Datasets. We use VLMSafety-420 (Elluru et al., 2026) exclusively as the paired attribution dataset $\mathcal{D}_{\mathrm{attr}}$. These examples are used to compute the refusal-margin objective in Eq. 10 and the corresponding channel attribution scores, but are never used to fit downstream detectors. For detector training and evaluation, we adopt the splits from Hua et al. (2026), whose setup evaluates transfer to out-of-distribution test cases. Additional dataset details are provided in Appendix A.1.

Detection methods. We evaluate a logistic-regression probe (LR), HiddenDetect (Jiang et al., 2025), and the KCD and MCD detectors from Hua et al. (2026). For methods that require fitting, separate detectors are trained using representations from the original and amplified models. HiddenDetect is applied directly to the extracted representations. KCD and MCD additionally train an MLP projector; because its random initialization introduces variation across runs, we report their mean and standard deviation over five seeds. Further implementation details are given in Appendix A.2.

Hyperparameters. Unless otherwise stated, we use a global channel-selection fraction $p = 0.1$ and amplification factor $\epsilon = 0.1$. The attribution target layers follow the layer-selection protocol of Hua et al. (2026): $\ell_a = 16$ for LLaVA and $\ell_a = 22$ for Qwen and InternVL3. LR, KCD, and MCD use $\ell_d = \ell_a$ for detector training and evaluation. HiddenDetect follows its original multi-layer readout configuration, using layers 16–29 for LLaVA, 22–26 for Qwen, and 18–24 for InternVL3.

Metrics. We report the area under the receiver operating characteristic curve (AUROC) to measure overall discrimination between benign and malicious inputs. We also report the true positive rate at a false positive rate of 5% (TPR@5%) to evaluate performance under a stringent false positive rate constraint. Both metrics are reported as percentages.

Table 1: Overall performance before and after channel-scaler amplification. KCD/MCD results are mean ± standard deviation over 5 different seeds. Entries are AUROC / TPR@5%(%).

Model	Detector	Original	Amplified	Improvement
LLaVA-1.6-Vicuna-7B	LR	97.05 / 84.56	97.83 / 92.11	0.78 / 7.55
	HiddenDetect	86.21 / 40.44	94.60 / 59.22	8.39 / 18.78
	KCD	89.85 ± 5.03 / 9.67 ± 15.50	95.08 ± 2.22 / 60.00 ± 34.66	5.23 / 50.33
	MCD	96.80 ± 0.89 / 83.56 ± 4.79	97.35 ± 0.46 / 89.58 ± 3.93	0.55 / 6.02
Qwen2.5-VL-7B	LR	97.97 / 91.89	98.88 / 95.78	0.91 / 3.89
	HiddenDetect	78.83 / 18.89	95.68 / 76.11	16.85 / 57.22
	KCD	92.09 ± 2.22 / 22.56 ± 38.93	93.30 ± 6.13 / 60.42 ± 46.53	1.44 / 37.86
	MCD	96.47 ± 1.15 / 83.18 ± 9.40	97.73 ± 0.46 / 89.96 ± 3.22	1.26 / 6.78
InternVL3-8B	LR	98.52 / 92.22	98.62 / 93.00	0.1 / 0.78
	HiddenDetect	86.34 / 32.33	96.31 / 82.89	9.97 / 50.56
	KCD	96.09 ± 5.07 / 33.07 ± 38.22	97.53 ± 0.94 / 87.60 ± 12.05	5.44 / 54.53
	MCD	98.36 ± 0.22 / 92.18 ± 1.48	98.41 ± 0.33 / 91.78 ± 1.34	0.05 / -0.40

Table 2: Representation movement and signed-direction channel-scaling ablations.

(a) Frozen-original-probe logit shifts.

Model	$\overline{\Delta z}_b$	$\overline{\Delta z}_m$	$\overline{\Delta z}_m - \overline{\Delta z}_b$
LLaVA-1.6-Vicuna-7B	-0.949	$+4.039$	$+4.988$
Qwen2.5-VL-7B	-1.175	$+3.428$	$+4.603$
InternVL3-8B	$+0.451$	$+3.650$	$+3.119$

(b) Signed-direction scaling on LLaVA.

Mask and scaler	ΔAUROC	ΔTPR@5%
Lower tail, $\epsilon = -0.1$	$+0.63$	$+6.00$
Lower tail, $\epsilon = 0.1$	-0.84	-2.78
Upper tail, $\epsilon = -0.1$	-1.61	-6.22
Upper tail, $\epsilon = 0.1$	**+0.78**	**+7.56**

Main results. Table 1 shows that channel amplification improves AUROC for every model and detector combination and improves TPR@5% in all cases except MCD on InternVL3, where it decreases by only 0.40%. The largest gains occur at the low false positive rate operating point. HiddenDetect improves from 40.44% to 59.22% on LLaVA, from 18.89% to 76.11% on Qwen, and from 32.33% to 82.89% on InternVL3. LR and MCD exhibit smaller gains, consistent with their stronger original performance. For example, LR improves from 84.56% to 92.11% in TPR@5% on LLaVA, while MCD improves from 83.18% to 89.96% on Qwen. KCD shows substantially larger variance, particularly in TPR@5%, but its mean performance still improves after amplification across all three models.

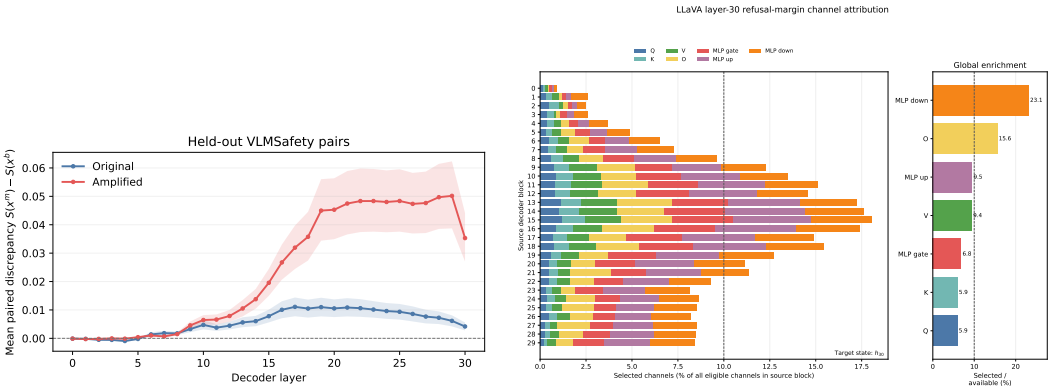
4.2 Analysis

Frozen-probe score shifts. To isolate representation movement from detector retraining, we apply the same LR probe trained on the original representations to both the original and amplified hidden states. Let $f_{\mathrm{LR}}(\mathbf{h})$ denote the scalar output of this frozen linear probe, where larger values indicate stronger evidence of malicious intent. For each evaluation input, we compute

$$\Delta f(x) = f_{\mathrm{LR}}\left(\mathbf{h}^{(\ell_d)}_{\mathrm{amp}}(x)\right) - f_{\mathrm{LR}}\left(\mathbf{h}^{(\ell_d)}_{\mathrm{orig}}(x)\right). \tag{16}$$

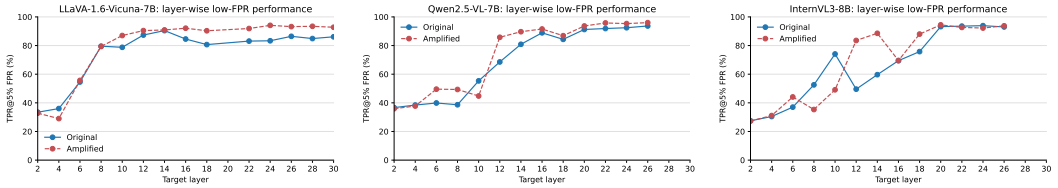
As shown in Table 2a, amplification increases the malicious versus benign logit gap for all three models. For LLaVA and Qwen, benign logits shift downward while malicious logits shift upward. For InternVL3, both shift upward, but the malicious shift is substantially larger. Thus, amplification increases malicious–benign separation under the original frozen decision boundary.

Generalization of attributed channels. We use $\ell_a = 30$, so channel attribution is performed over all projections in source layers 0–29. We then evaluate the resulting global mask on 65 held-out VLMSafety pairs not used for attribution. Figure 2a shows that amplification increases the refusal-score gap in Eq. 10 between malicious and benign inputs, with the effect becoming more pronounced in middle and later layers. This indicates that the attributed channels capture safety-relevant structure beyond the pairs used to compute their scores, consistent with prior VLM studies showing depth-dependent safety information in internal representations (Jiang et al., 2025; Hua et al., 2026).



(a) Held-out paired refusal discrepancy. (b) Global mask composition and projection enrichment.

Figure 2: Held-out validation and attribution anatomy for LLaVA with $\ell_a = 30$. **Left:** mean paired refusal discrepancy $S_\ell(x^m) - S_\ell(x^b)$ across 65 held-out VLMSafety pairs; shaded regions denote 95% bootstrap confidence intervals. **Right:** distribution of the selected channels across source layers and projection types. The right one reports selection rates normalized by the number of channels in each projection type.



(a) LLaVA-1.6-Vicuna-7B (b) Qwen2.5-VL-7B (c) InternVL3-8B

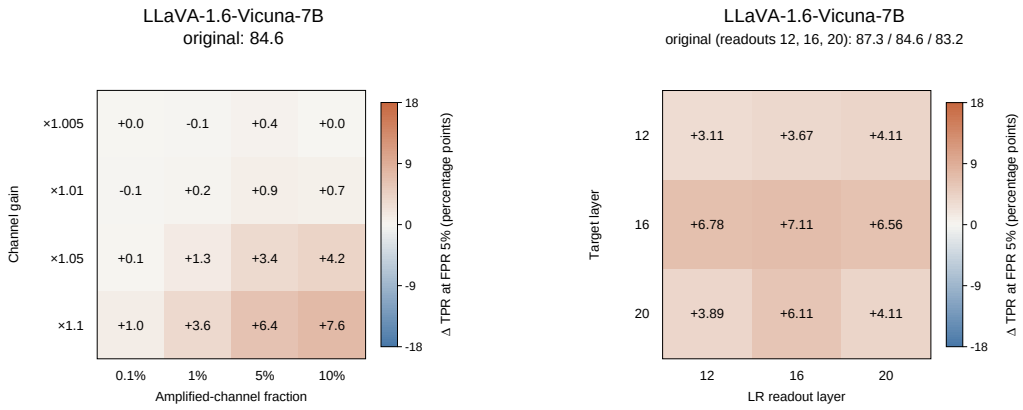
Figure 3: Layer-wise TPR@5%before and after channel amplification.

Distribution of selected channels. Figure 2b shows how the globally selected channels are distributed across source layers and projection types. Selection is concentrated broadly in the middle-to-late layers, with the highest density around layers 13–18 before decreasing toward the final layers. MLP-down projections contain the largest fraction of selected channels, followed by attention-output projections. This pattern is consistent with recent VLM evidence that safety-relevant information can be localized in intermediate MLP components (Li et al., 2026). Because these both projections write into the residual stream, their high ratios suggests that these channels may influence the propagation of refusal-relevant representations through the decoder.

4.3 Ablation Studies

Layer ablation. Figure 3 evaluates amplification across attribution target layers. Overall, the gains are concentrated in intermediate and later layers, although their magnitude is not monotonic with depth. For LLaVA, improvements become persistent from approximately layer 10 onward. Qwen shows its largest gains around layers 12–14, after which both original and amplified representations become strongly separable. InternVL3 exhibits a less smooth pattern, with substantial gains around layers 12–14 and again near layer 18. These results indicate that amplification is most effective once safety-relevant separation has emerged but before the original representations become strongly separable on their own.

Signed directionality of channel scaling. The main intervention applies $\epsilon = 0.1$ to the highest-scoring positive channels. To test whether the effect follows the signed attribution rather than generic perturbation, we also attenuate channels using $\epsilon = -0.1$ and evaluate both positive and negative score tails. The first-order prediction for channel (r, c) is proportional to $\epsilon \cdot s_{r,c}^{(\ell_a)}$, so reversing the sign of ϵ should reverse the effect of a fixed score direction. Table 2b follows this prediction:



(a) Selection fraction p and amplification factor ϵ.

(b) Attribution target ℓ_a and LR readout ℓ_d.

Figure 4: Sensitivity analysis for LLaVA. Each cell reports the change in $\mathrm{TPR@5\%}$relative to the original representation, in percentage points.

Table 3: Matched channel-selection controls for LLaVA layer 16. Entries are improvement in percentage points relative to the unmodified representation.

Channel selection	ΔAUROC	$\Delta\mathrm{TPR@5\%}$
Random	-0.03	-0.11
Weight magnitude	-0.31	-0.56
Gradient magnitude	$+0.58$	$+5.00$
Refusal-margin score	$+0.78$	$+7.56$

attenuating positive-score channels degrades detection, whereas attenuating negative-score channels improves it.

Intervention strength and selection size. Figure 4a varies the global channel-selection fraction p and amplification factor ϵ. Weak interventions have little effect, whereas increasing both quantities generally yields larger gains. The strongest evaluated setting, $p = 0.10$ and $\epsilon = 0.1$, improves $\mathrm{TPR@5\%}$ by 7.6%.

Sensitivity to attribution and readout layers. Figure 4b varies the attribution target ℓ_a and LR readout ℓ_d independently. Amplification improves $\mathrm{TPR@5\%}$ for all evaluated combinations, showing that the effect is not restricted to a single target-readout pair. Attribution at $\ell_a = 16$ produces the largest and most consistent gains across readout layers, with the strongest result at $\ell_a = \ell_d = 16$. This indicates that the selected channels improve safety-relevant representations beyond the exact state used to define the attribution objective.

Channel-selection controls. Table 3 compares the refusal-margin score with alternative channel-selection criteria under the same intervention budget. The module-matched random control preserves the number of selected channels per projection module while randomizing their identities. We also rank channels by weight magnitude and gradient magnitude separately, isolating the two terms in Eq. 6. Random and weight-magnitude selection provide no benefit, while gradient-magnitude selection improves detection but remains weaker than the refusal-margin score. This suggests that the gain depends on the joint weight-gradient contribution to the refusal-margin objective rather than either factor alone.

Attribution-source modality. Table 4 compares text-only and multimodal paired attribution sources under the same intervention. Both sources improve overall detection, but the multimodal

Table 4: Attribution-source transfer for LLaVA layer-16 channel selection. Each entry reports the improvement over the unmodified representation as ΔAUROC / ΔTPR@5% in percentage points.

Paired attribution source	Pairs	Overall	Text-only	Text–image
XLLM (Yu et al., 2025) text-only prompts	220	$+0.06$ / $+2.88$	$+1.73$ / $+8.18$	-0.56 / -2.29
VLMSafety image–text prompts	355	$+0.78$ / $+7.55$	$+1.54$ / $+11.09$	-0.02 / -0.86

VLMSafety attribution yields substantially larger overall gains. It also transfers more consistently across evaluation modalities: performance improves strongly on text-only inputs while showing only a small degradation on text–image inputs. In contrast, attribution from the text-only XLLM source is less effective overall and degrades more noticeably on text–image inputs. These results suggest that multimodal attribution provides a more transferable channel-selection signal across input modalities.

5 Conclusion

We presented Amplify to Detect, a model-side intervention for improving multimodal jailbreak detection by strengthening safety-relevant signals already present within a frozen vision–language model. Our approach uses a SNIP-inspired sensitivity analysis to identify projection-output channels that contribute to harmful–benign separation, and selectively amplifies the channels predicted to increase this separation. The resulting representations can then be used directly by existing hidden-state detectors without retraining VLM parameters or changing their detection objectives.

Across three open-weight VLMs and multiple detection methods, targeted amplification consistently improves overall discrimination and, more importantly, detection performance under stringent false-positive-rate constraints. The largest gains occur for detectors whose original hidden-state representations are less strongly separated, while already strong detectors generally show smaller improvements. Controlled ablations further show that the effect follows the signed attribution direction: amplifying positively attributed channels improves detection, reversing the intervention direction degrades it, and random or weight-magnitude-based channel selection provides little benefit.

Taken together, these results suggest that safety-relevant information in VLM hidden states is not merely passively observable, but is associated with identifiable internal computations whose contribution can be selectively strengthened. This provides a complementary direction for jailbreak defense: rather than designing increasingly complex downstream decision boundaries alone, future detection systems may also benefit from shaping the internal representations on which those detectors operate.

6 Limitations and interpretation

Our evidence comes from three open-weight VLMs, one adopted detector split, and 355 attribution pairs; the 65 held-out pairs assess transfer only within the same attribution source. Robustness to other model families, languages, and adaptive attacks remains untested. The single-layer refusal margin is a proxy for representation separation: improving it or a detector score does not establish safer generated responses. Moreover, first-order channel scores need not predict interactions when many rows are scaled together. Gains are uneven: KCD shows substantial variation across seeds, MCD on InternVL3 loses 0.40 percentage points in TPR@5%, and Table 4 shows a small decline on text–image inputs. We have not measured changes in benign task quality, false refusals, or inference cost. Deployment would require these measurements and calibration for the intended population.

AI USE STATEMENT

We used generative AI tools, including ChatGPT, to help draft and edit manuscript text and LaTeX, develop the method diagram, examine the channel-attribution derivation, review cited methods and publicly released implementations, and provide feedback on methodological presentation and the interpretation of experimental results.

ETHICS STATEMENT

This work evaluates a defensive method using existing text and image jailbreak benchmarks. Information about safety-related channels and their modification could also be used to probe model safeguards. The measured detection gains do not establish protection against adaptive attacks. False alarms may impede legitimate use, so deployment requires evaluation of benign inputs and calibration for the intended setting. Redistribution of benchmark prompts or derived data should respect their applicable terms and privacy constraints.

REPRODUCIBILITY STATEMENT

Section 3 specifies channel scaling, attribution, and selection; Section 4 reports the models, primary settings, and metrics. Appendix A describes the data splits and detector scoring procedures, Appendix B reports additional analyses, and Appendix C derives the channel score. The attribution pairs are separate from the detector training and test data, and variation over five runs is reported for KCD and MCD.

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. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Shengda Song, Kai Dang, Peng Wang, Shijie Wang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Veeraraju Elluru, Arth Singh, Roberto Aguero, Ajay Agarwal, Debojyoti Das, and Hreetam Paul. Mechanistically interpreting compression in vision-language models. *arXiv preprint arXiv:2603.25035*, 2026.

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 *International Conference on Learning Representations*, volume 2026, pp. 5216–5239, 2026.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23951–23959, 2025.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617. IEEE, 2018.

Peichun Hua, Hao Li, Shanghao Shi, Zhiyuan Yu, and Ning Zhang. Rethinking jailbreak detection of large vision language models with representational contrastive scoring. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21748–21785, 2026.

Yilei Jiang, Xinyan Gao, Tianshuo Peng, Yingshui Tan, Xiaoyong Zhu, Bo Zheng, and Xiangyu Yue. Hiddendetect: Detecting jailbreak attacks against multimodal large language models via monitoring hidden states. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14880–14893, 2025.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36:47669–47681, 2023.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. SNIP: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.02340.

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.

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 *European Conference on Computer Vision*, pp. 174–189. Springer, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.08485.

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. *arXiv preprint arXiv:2404.03027*, 2024.

Hadas Orgad, Boyi Wei, Kaden Zheng, Martin Wattenberg, Peter Henderson, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. Large language models generate harmful content using a distinct, unified mechanism. In *Mechanistic Interpretability Workshop at ICML 2026*.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 21527–21536, 2024.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, 2024.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *International conference on learning representations*, volume 2024, pp. 30853–30885, 2024.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1671–1685, 2024.

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, 2023.

Han Wang, Gang Wang, and Huan Zhang. Steering away from harm: An adaptive approach to defending vision language model against jailbreaks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29947–29957. IEEE, 2025.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52588–52610. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/wei24f.html`.

Fenghua Weng, Yue Xu, Chengyan Fu, and Wenjie Wang. Mmj-bench: A comprehensive study on jailbreak attacks and defenses for vision language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 27689–27697, 2025.

Junhao Xia, Haotian Zhu, Shuchao Pang, Zhigang Lu, Bing Li, Yongbin Zhou, and Minhui Xue. One head to rule them all: Amplifying lvlm safety through a single critical attention head. *Advances in Neural Information Processing Systems*, 38:45298–45329, 2026.

Yueqi Xie, Minghong Fang, Renjie Pi, and Neil Gong. Gradsafe: Detecting jailbreak prompts for llms via safety-critical gradient analysis. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 507–518, 2024.

Yue Xu, Xiuyuan Qi, Zhan Qin, and Wenjie Wang. Cross-modality information check for detecting jailbreaking in multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13715–13726, 2024.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9467–9476. IEEE, 2025.

Ziyi Yin, Yuanpu Cao, Han Liu, Ting Wang, Jinghui Chen, and Fenglong Ma. Towards robust multimodal large language models against jailbreak attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22847–22856, 2026.

Jiahao Yu, Haozheng Luo, Jerry Yao-Chieh Hu, Yan Chen, Wenbo Guo, Han Liu, and Xinyu Xing. Mind the inconspicuous: Revealing the hidden weakness in aligned {LLMs}'refusal boundaries. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 259–278, 2025.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-vet: Evaluating large multimodal models for integrated capabilities. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yu24o.html.

Hangtao Zhang, Yucheng Zhao, Sishun Liu, Ziqi Zhou, Zeyu Ye, Wei Wan, Minghui Li, Shengshan Hu, Yanjun Zhang, Yi Liu, et al. Defending jailbreak attacks on large language models via manifold trajectory kinetics. *arXiv preprint arXiv:2606.07335*, 2026.

Xiaoyu Zhang, Cen Zhang, Tianlin Li, Yihao Huang, Xiaojun Jia, Ming Hu, Jie Zhang, Yang Liu, Shiqing Ma, and Chao Shen. Jailguard: A universal detection framework for prompt-based attacks on llm systems. *ACM Transactions on Software Engineering and Methodology*, 35(1):1–40, 2025.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. In *International Conference on Learning Representations*, volume 2025, pp. 84042–84071, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Xiaohan Zou, Jian Kang, George Kesidis, and Lu Lin. Understanding and rectifying safety perception distortion in vlms. *Advances in Neural Information Processing Systems*, 38:114196–114229, 2026.

A Experimental Setup

A.1 Attribution data and Detection Data

Table 5: Composition of the detector training and test splits following Table 6 of Hua et al. (2026). MM denotes a multimodal input. The attribution data used to select projection rows are separate.

Split	Class	Source	Modality	N
Train	Benign	Alpaca (Taori et al., 2023)	Text	500
Train	Benign	MM-Vet (Yu et al., 2024)	MM	218
Train	Benign	OpenAssistant (Köpf et al., 2023)	Text	282
Train	Malicious	AdvBench (Zou et al., 2023)	Text	300
Train	Malicious	DAN (Shen et al., 2024)	Text	150
Train	Malicious	JailbreakV-28K (Luo et al., 2024)[1]	MM	550
Test	Benign	XSTest (Röttger et al., 2024)	Text	250
Test	Benign	FigTxt (Gong et al., 2025)	Text	300
Test	Benign	VizWiz (Gurari et al., 2018)	MM	450
Test	Malicious	XSTest	Text	200
Test	Malicious	FigTxt	Text	350
Test	Malicious	VAE (Qi et al., 2024)	MM	200
Test	Malicious	JailbreakV-28K	MM	150

We use VLMSafety-420 exclusively to construct paired examples for channel attribution. Each example contains a benign input x_i^b and a matched malicious input x_i^m. We use 355 pairs to compute the attribution scores and reserve 65 separate pairs for the held-out refusal-margin analysis. The latter held-out pairs do not contribute to the channel ranking or the selected mask. The VLMSafety-420 dataset is not used to fit a downstream detector. Detector training and evaluation instead use the splits adopted from the setup of Hua et al. (2026) and the component is shown in table 5. This separation matters because the held-out refusal plots assess whether the selected channels affect new paired inputs rather than simply increasing the objective on the pairs that determined the mask.

A.2 Detector comparisons

We compare four hidden-state detectors reported in the main results: logistic regression (LR), HiddenDetect (Jiang et al., 2025), and the KCD and MCD variants of representational contrastive scoring (Hua et al., 2026). Each is evaluated on the same detector training and test splits described in Appendix A.1. The two conditions differ in the VLM hidden state supplied to the detector: the original model supplies $h^{(\ell)}_{\mathrm{orig}}(x)$, while the amplified model supplies $h^{(\ell)}_{\mathrm{amp}}(x)$.

Logistic regression. We fit a linear probe to the last-token hidden state at the detector readout layer ℓ_d, using benign and malicious labels from the detector training split. The probe outputs a scalar score for malicious intent. In the main comparison, one probe is fitted to original representations and another to amplified representations.

HiddenDetect. Jiang et al. (2025) constructs a fixed refusal vector r_{ref} in vocabulary space from refusal-related tokens. For an input x, let $q_\ell(x)$ be its last-token hidden state at layer ℓ mapped into vocabulary space by the frozen model output mapping. The *refusal strength* at that layer is

$$S_\ell(x) = \cos\big(q_\ell(x), r_{\mathrm{ref}}\big). \tag{17}$$

This cosine measures alignment with refusal-related vocabulary directions; it does not require generating a response. Using a small set of benign inputs $\mathcal{D}^b$ and malicious inputs $\mathcal{D}^m$, the method computes the layer-wise *refusal discrepancy*

$$\mathcal{J}_\ell = \frac{1}{|\mathcal{D}^m|}\sum_{x\in\mathcal{D}^m} S_\ell(x) - \frac{1}{|\mathcal{D}^b|}\sum_{x\in\mathcal{D}^b} S_\ell(x). \tag{18}$$

[1] JailbreakV-28K uses LLM Transfer and Query-Related attacks for training and held-out FigStep attacks for testing.

The subtraction identifies layers whose refusal signal responds more strongly to malicious inputs than to benign ones. For a model with layers indexed from 0 to $L-1$, HiddenDetect takes the first and last layers whose discrepancy exceeds that of the final layer:

$$s = \min\{\ell : \Delta_\ell > \Delta_{L-1}\}, \qquad e = \max\{\ell : \Delta_\ell > \Delta_{L-1}\}. \tag{19}$$

They then define the safety-aware range $\{s, \ldots, e\}$. For a new input, HiddenDetect reduces its layer-wise refusal strengths to one score by the trapezoidal rule:

$$T_{\mathrm{HD}}(x) = \sum_{\ell=s}^{e-1} \frac{S_\ell(x) + S_{\ell+1}(x)}{2}. \tag{20}$$

Because adjacent layers are one index apart, each term is the area of a unit-width trapezoid beneath the refusal-strength curve. Their sum measures the cumulative signal across the selected range; it is an area over *layer indices*. Higher scores indicate stronger refusal evidence, and a threshold gives the detector's binary decision. HiddenDetect does not train a classifier or MLP projector.

KCD and MCD. They (Hua et al., 2026) share a learned projector P_ϕ that maps the last-token hidden state at ℓ to a 256-dimensional feature. Its training encourages examples from the same source dataset to cluster while separating benign and malicious examples.

KCD projects and normalizes the hidden state of inputs, benign and malicious training samples as

$$z(x) = \frac{P_\phi(h^{(\ell)}(x))}{\|P_\phi(h^{(\ell)}(x))\|_2}. \tag{21}$$

Let $\mathcal{B}$ and $\mathcal{M}$ be the projected benign and malicious training examples and it scores an input by the difference between its distances to the kth nearest examples in benign and malicious training sets:

$$s_{\mathrm{KCD}}(x) = d_k\big(z(x), \mathcal{B}\big) - d_k\big(z(x), \mathcal{M}\big). \tag{22}$$

Here $d_k(z, \mathcal{A})$ is the Euclidean distance from z to its kth nearest point in $\mathcal{A}$. A positive score means that the input lies closer to malicious than to benign examples at this neighborhood scale; using the kth neighbor reduces sensitivity to a single unusually close example.

MCD instead models each source dataset d as a Gaussian cluster with mean μ_d and covariance $\widehat{\Sigma}_d$ in projected space. It uses covariance shrinkage to make the inverse covariance stable when a source has relatively few examples. With

$$D_{\mathrm{Mah}}(z; \mu_d, \widehat{\Sigma}_d) = \sqrt{(z-\mu_d)^\top \widehat{\Sigma}_d^{-1}(z-\mu_d)}, \tag{23}$$

its contrastive score is

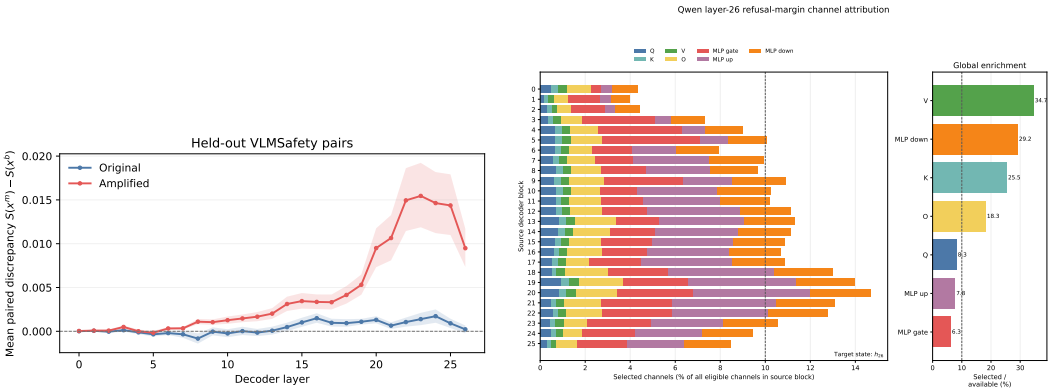
$$s_{\mathrm{MCD}}(x) = \min_{d \in \mathcal{S}_b} D_{\mathrm{Mah}}\big(z(x); \mu_d, \widehat{\Sigma}_d\big) - \min_{d \in \mathcal{S}_m} D_{\mathrm{Mah}}\big(z(x); \mu_d, \widehat{\Sigma}_d\big), \tag{24}$$

where $\mathcal{S}_b$ and $\mathcal{S}_m$ index benign and malicious source datasets. Taking the minimum within each class allows an input to match any of its source clusters. For both detectors, subtracting the malicious distance from the benign distance makes larger scores indicate stronger evidence of malicious intent.

B Additional Results

B.1 Held-out refusal separation and mask composition

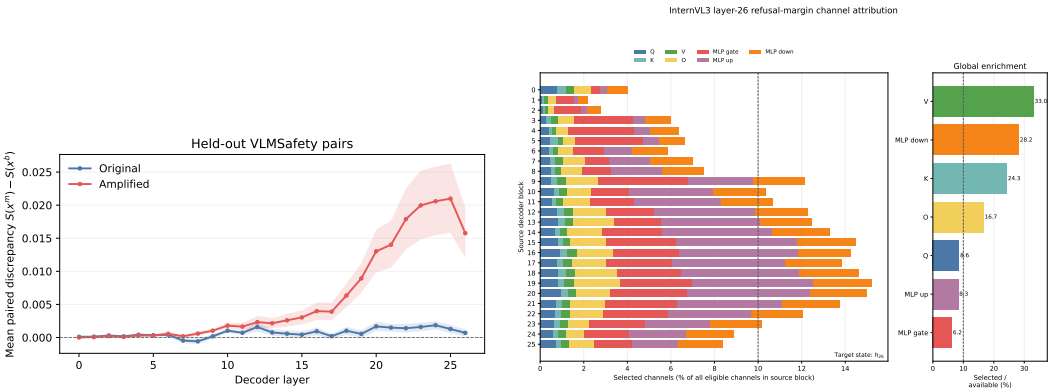
The held-out paired plots for Qwen and InternVL3 extend the LLaVA analysis in the main text to the other two models. For each pair, we subtract the benign refusal similarity from its malicious counterpart at each displayed layer and average the differences over the 65 held-out pairs. The amplified curves show that the intervention affects refusal-relevant separation beyond the attribution examples. The accompanying mask-composition plots report which source layers and projection types contain selected rows; the projection insets normalize selected counts by the number of eligible rows of each type.



(a) Held-out paired refusal discrepancy.

(b) Layer-26 mask composition and projection enrichment.

Figure 5: Held-out validation and attribution anatomy for Qwen. **Left:** mean paired refusal discrepancy $S_\ell(x^m) - S_\ell(x^b)$ across 65 held-out VLMSafety pairs; shaded regions are 95% bootstrap confidence intervals. **Right:** composition of the globally selected top-10% positive layer-30 refusal-margin mask. The right inset reports selection rates normalized by the number of available channels in each projection.



(a) Held-out paired refusal discrepancy.

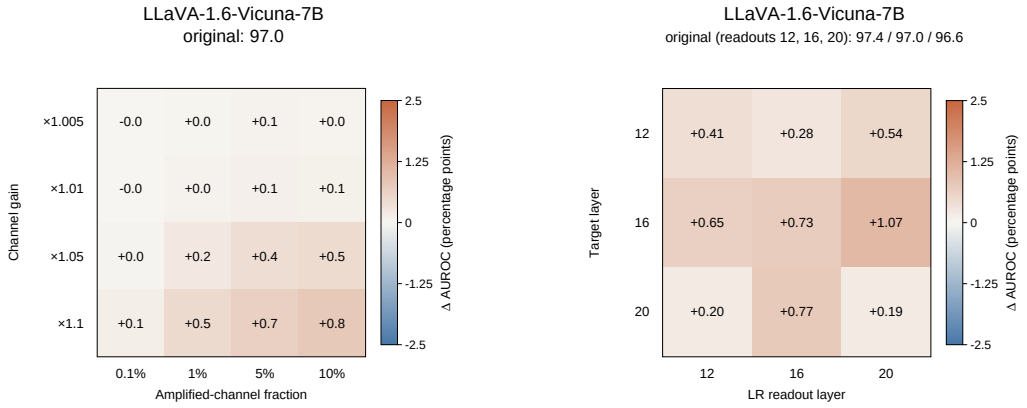
(b) Layer-26 mask composition and projection enrichment.

Figure 6: Held-out validation and attribution anatomy for InternVL. **Left:** mean paired refusal discrepancy $S_\ell(x^m) - S_\ell(x^b)$ across 65 held-out VLMSafety pairs; shaded regions are 95% bootstrap confidence intervals. **Right:** composition of the globally selected top-10% positive layer-30 refusal-margin mask. The right inset reports selection rates normalized by the number of available channels in each projection.

B.2 Intervention strength and readout sensitivity

Figures 7, 8, 9, 10, and 11 extend the sensitivity analysis in Figure 4. In the channel-selection heat maps, p is the fraction of eligible channels selected globally, and the displayed channel multiplier is $1 + \epsilon$. Each cell reports the change from the original hidden-state logistic-regression probe, in percentage points. The layer heat maps vary the attribution target ℓ_a, where the refusal-margin objective is evaluated, independently of the probe readout layer ℓ_d.

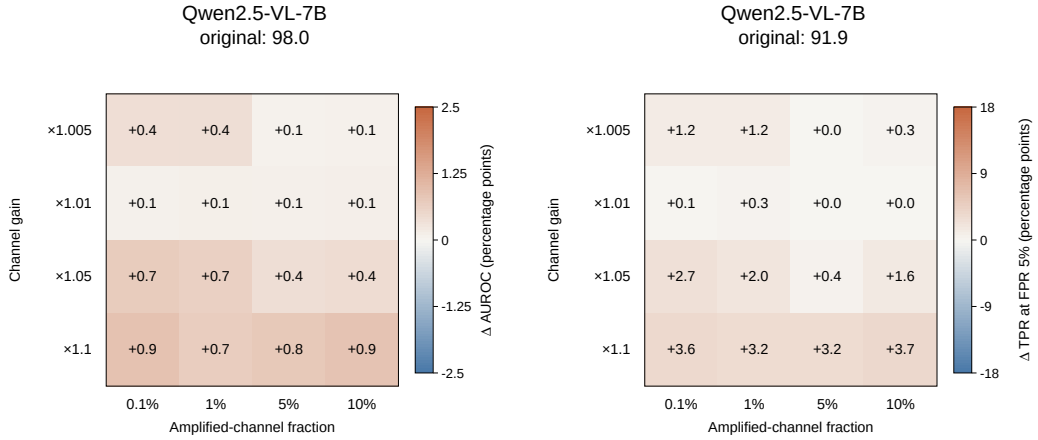
LLaVA. Figure 7a shows the AUROC counterpart to the TPR@5% results in Figure 4a. AUROC changes are small at the weakest scales and generally increase as the amplification becomes stronger, reaching +0.8 points at $p = 10\%$ and $\epsilon = 0.1$. Figure 7b shows a positive AUROC change for every evaluated attribution-target and readout-layer pair. The largest change is +1.07 points when $\ell_a = 16$



(a) Selection fraction p and amplification factor g.

(b) Attribution target ℓ_a and LR readout ℓ_d.

Figure 7: Sensitivity analysis for LLaVA. Each cell reports the change in AUROC relative to the original representation, in percentage points.



(a) AUROC change.

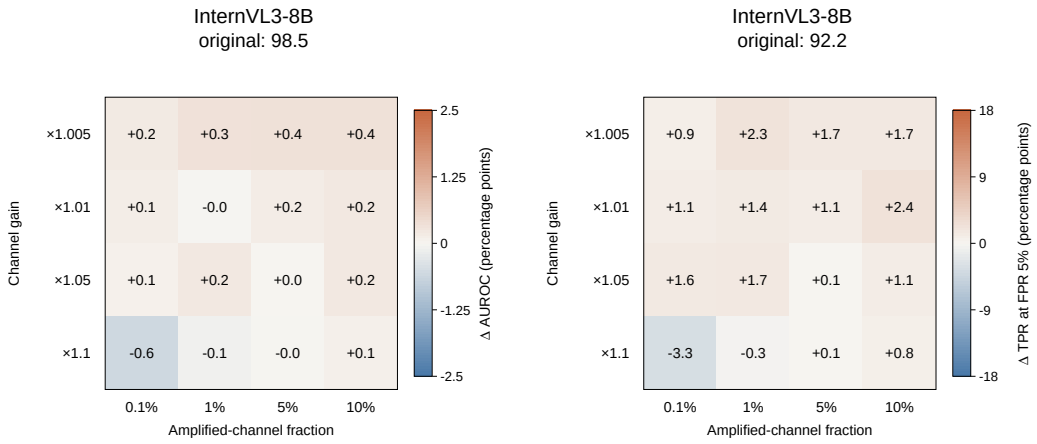
(b) TPR@5%change.

Figure 8: Channel-fraction and gain ablation for Qwen2.5-VL-7B. Each cell shows the change from the original hidden-state logistic-regression probe, in percentage points.

and $\ell_d = 20$. Together with Figure 4b, this indicates that a mask attributed at one layer can improve a probe read from another layer, although the best pair depends on the evaluation metric.

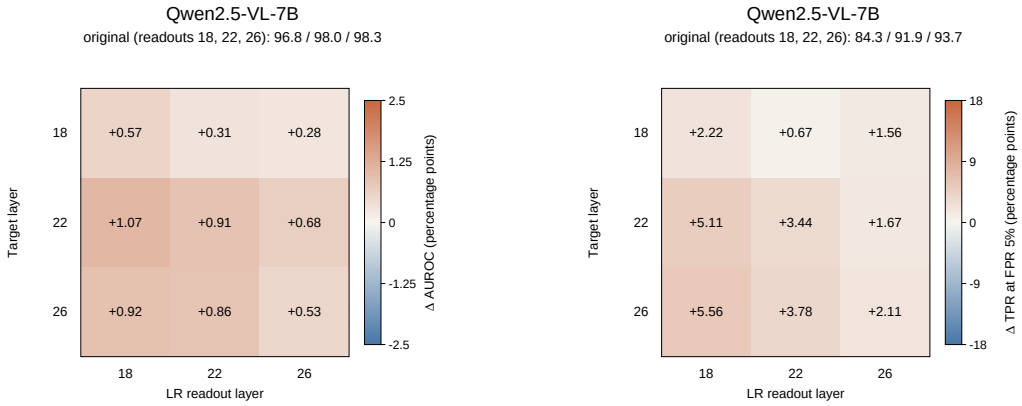
Qwen2.5-VL-7B. In Figure 8, the AUROC and TPR@5% changes are nonnegative throughout the evaluated selection-and-scale grid. At $p = 10\%$ and $\epsilon = 0.1$, they reach +0.9 and +3.7 points, respectively. Figure 10 likewise shows improvements for every evaluated pair of attribution and readout layers. The largest AUROC gain is +1.07 points at $(\ell_a, \ell_d) = (22, 18)$, whereas the largest TPR@5% gain is +5.56 points at $(26, 18)$. Thus, the layer pair that maximizes AUROC need not maximize low-FPR detection.

InternVL3-8B. Figure 9 shows a less uniform response to amplification. At $p = 0.1\%$ and $\epsilon = 0.1$, AUROC and TPR@5% decrease by 0.6 and 3.3 points, respectively. Smaller scales produce modest gains in several settings; the largest TPR@5% gain in this grid is +2.4 points at $p = 10\%$



(a) AUROC change. (b) TPR@5%change.

Figure 9: Channel-fraction and gain ablation for InternVL3-8B. Each cell shows the change from the original hidden-state logistic-regression probe, in percentage points.



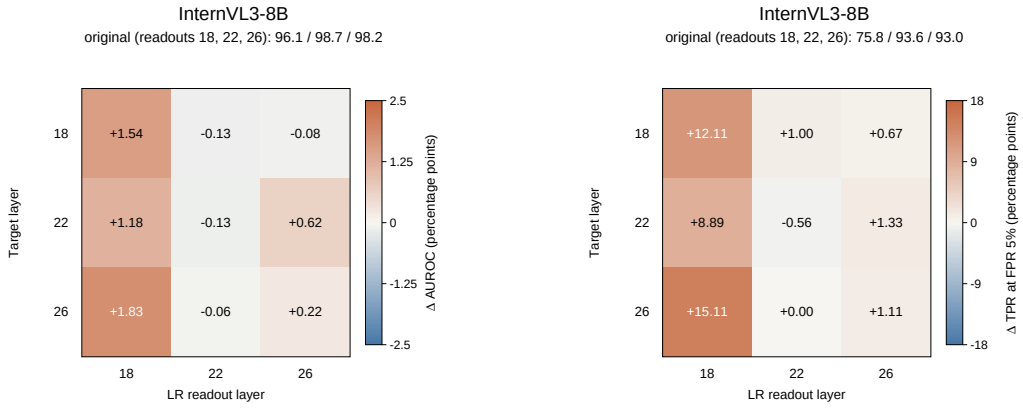
(a) AUROC change. (b) TPR@5%change

Figure 10: Attribution target ℓ_a and LR readout ℓ_d for Qwen2.5-VL-7B model.

and $\epsilon = 0.01$. Figure 11 shows that the outcome also depends strongly on the probe readout layer. Reading from layer 18 gives positive changes for all three attribution targets, with the largest gains at $(\ell_a, \ell_d) = (26, 18)$: +1.83 AUROC points and +15.11 TPR@5% points. Changes at readout layer 22 are much smaller and include negative values. These results show that the effective intervention strength and target layer depend on the model and readout; increasing the scale or selecting more channels does not consistently improve detection.

C Score Derivation

Let $\mathcal{C}_\ell$ be the set of eligible projection output channels that can contribute to the hidden state at target layer ℓ. For each $(r, c) \in \mathcal{C}_\ell$, let $\epsilon_{r,c}$ be its scale, and collect these scales in $\boldsymbol{\epsilon} = (\epsilon_{r,c})_{(r,c)\in\mathcal{C}_\ell}$. Projection r is a linear module, and c indexes an output row of its weight matrix W_r.



(a) AUROC change.

(b) TPR@5%change

Figure 11: Attribution target ℓ_a and LR readout ℓ_d for InternVL3-8B model.

For input x, let the columns of $A_r(x;\boldsymbol{\epsilon})$ be the activations entering projection r at the positions where it is applied. We define the scaled channel outputs by

$$\widetilde{\mathbf{z}}_{r,c}(x;\boldsymbol{\epsilon}) = (1+\epsilon_{r,c})W_{r,c:}A_r(x;\boldsymbol{\epsilon}). \tag{25}$$

The same scale is applied whenever that output channel is used during the forward pass. The input activations may themselves depend on scales applied to upstream channels. If a projection has a bias, we augment each input activation with a constant 1 and include the bias in $W_{r,c:}$.

The scale vector specifies a forward-pass intervention; the original VLM parameters are not updated. Setting $\boldsymbol{\epsilon} = \mathbf{0}$ recovers the original model. Let $h_{\boldsymbol{\epsilon}}^{(\ell)}(x)$ denote the last-token hidden state at layer ℓ obtained from this scaled forward pass.

HiddenDetect (Jiang et al., 2025) shows that refusal-related information can be read from intermediate hidden states before a model generates a response.[2] Following this readout, we apply the model's frozen final normalization and language-model head to the target-layer representation:

$$\begin{aligned} q_\ell(x;\boldsymbol{\epsilon}) &= \mathrm{LMHead}\Big(\mathrm{Norm}\Big(h_{\boldsymbol{\epsilon}}^{(\ell)}(x)\Big)\Big), \\ S_\ell(x;\boldsymbol{\epsilon}) &= \cos(q_\ell(x;\boldsymbol{\epsilon}), r_{\mathrm{ref}}). \end{aligned} \tag{26}$$

Here q_ℓ is a vocabulary-space vector and r_{ref} is a fixed refusal-reference vector in the same space. Thus, $\boldsymbol{\epsilon}$ affects q_ℓ and S_ℓ through the scaled hidden state; the normalization, language-model head, and refusal reference remain fixed.

Let $\{x_j^m\}_{j=1}^{N_m}$ and $\{x_i^b\}_{i=1}^{N_b}$ be the malicious and benign attribution examples. Increasing refusal similarity for malicious inputs alone does not guarantee better separation, because similarity may also increase for benign inputs. We therefore define the attribution objective as the difference between the class means:

$$\mathcal{J}_{\mathrm{ref}}^{(\ell)}(\boldsymbol{\epsilon}) = \frac{1}{N_m}\sum_{j=1}^{N_m} S_\ell(x_j^m;\boldsymbol{\epsilon}) - \frac{1}{N_b}\sum_{i=1}^{N_b} S_\ell(x_i^b;\boldsymbol{\epsilon}). \tag{27}$$

A larger value indicates greater separation along the refusal reference at layer ℓ.

Let $\mathbf{e}_{r,c}$ be the coordinate vector that changes only the scale of channel (r,c). A first-order expansion around the original model gives

$$\mathcal{J}_{\mathrm{ref}}^{(\ell)}(\epsilon\mathbf{e}_{r,c}) = \mathcal{J}_{\mathrm{ref}}^{(\ell)}(\mathbf{0}) + \epsilon s_{r,c}^{(\ell)} + o(\epsilon), \qquad s_{r,c}^{(\ell)} = \left.\frac{\partial \mathcal{J}_{\mathrm{ref}}^{(\ell)}(\boldsymbol{\epsilon})}{\partial \epsilon_{r,c}}\right|_{\boldsymbol{\epsilon}=\mathbf{0}}. \tag{28}$$

[2]We review the HiddenDetect scoring procedure in Appendix A.2.

For a sufficiently small positive ϵ, a positive score predicts a local increase in the refusal margin.

To relate this score to the projection row, let $F_x(U)$ be the refusal score obtained when $W_{r,c:}$ is replaced by U throughout the forward pass, with every other channel scale set to zero. Because scaling channel (r, c) is equivalent to replacing its entire row by $(1 + \epsilon)W_{r,c:}$,

$$S_\ell(x; \epsilon \mathbf{e}_{r,c}) = F_x\big((1 + \epsilon)W_{r,c:}\big). \tag{29}$$

The chain rule for a single input yields

$$\left.\frac{dS_\ell(x; \epsilon \mathbf{e}_{r,c})}{d\epsilon}\right|_{\epsilon=0} = \sum_k W_{r,ck} \left.\frac{\partial F_x(U)}{\partial U_k}\right|_{U=W_{r,c:}}. \tag{30}$$

The derivative of F_x includes every computational path through which that row affects the last-token refusal score.

Finally, define the margin as a function of the replacement row:

$$G(U) = \frac{1}{N_m}\sum_{j=1}^{N_m} F_{x_j^m}(U) - \frac{1}{N_b}\sum_{i=1}^{N_b} F_{x_i^b}(U). \tag{31}$$

Applying the single-input result to both class means gives

$$\begin{aligned} s_{r,c}^{(\ell)} &= \sum_k W_{r,ck} \left[\frac{1}{N_m}\sum_{j=1}^{N_m} \left.\frac{\partial F_{x_j^m}(U)}{\partial U_k}\right|_{U=W_{r,c:}} - \frac{1}{N_b}\sum_{i=1}^{N_b} \left.\frac{\partial F_{x_i^b}(U)}{\partial U_k}\right|_{U=W_{r,c:}}\right] \\ &= \sum_k W_{r,ck} \left.\frac{\partial G(U)}{\partial U_k}\right|_{U=W_{r,c:}}. \end{aligned} \tag{32}$$

Thus, the channel score is the signed directional derivative of the refusal margin along a scaling of the entire projection row. We use these scores to rank channels without updating the original VLM parameters.