FROM VISUAL HARMFULNESS TO REFUSAL IN LOW DIMENSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety defenses in VLMs often reuse textual refusal representations, but stronger control can come at the cost of benign and general multimodal behavior. We show that effective refusal need be neither text-conditioned nor high-dimensional. Visual harmfulness and refusal can both be captured in low-dimensional subspaces, with refusal concentrated in markedly fewer directions. We therefore propose a visually conditioned harmfulness-to-refusal mapping that connects the two low-dimensional spaces with only 272 trainable parameters. The mapping reduces overall attack success rate by 5.0–18.4 percentage points on MM-SafetyBench and 10.7–19.2 percentage points on HoliSafe, while leaving MME performance nearly unchanged and largely preserving benign behavior on MOSSBench. Comparing the two refusal spaces reveals a trade-off in which the textual refusal space provides stronger safety, while the visual refusal space better preserves benign behavior with comparable general multimodal capability.

1 INTRODUCTION

Safety alignment requires a model to recognize harmful content and to act on that recognition, two functions that can dissociate in VLMs (Xu et al., 2025). Harmfulness can become detectable in intermediate representations before refusal behavior emerges (Zhou et al., 2024), while localized activation directions can exert causal control over refusal (Arditi et al., 2024). Harmfulness assessment and refusal have also been linked to distinct causal directions in representation space (Zhao et al., 2025a). These findings motivate treating harmfulness representation and refusal control as separate components of aligned behavior.

In VLMs, the evidence can be visual while the action is linguistic. Safeguards encoded in textual representations can be reused by aligning multimodal states with textual safety representations (Liu et al., 2025; Xu et al., 2025), while modality-specific safety representations can also be modeled and controlled separately (Fu et al., 2026). These approaches show that visual evidence can be connected to refusal through internal representations. But such defenses are typically formulated in the ambient hidden space, where safety-related information coexists with representations used for general multimodal behavior. Acting broadly in this space may therefore perturb computation unrelated to safety, while also introducing degrees of freedom that may not be needed for refusal. This raises the question of how much of the hidden space is actually required to connect visual harmfulness to refusal.

Figure 1 compares the singular-value spectra of visual harmfulness differences and refusal gradients across three VLM families to test whether harmfulness-to-refusal coupling requires the full hidden space. Both signals are sharply concentrated, with refusal consistently occupying a lower-dimensional subspace than visual harmfulness. Visual harmfulness retains a broader spectral tail, whereas refusal gradients are dominated by a small number of leading directions. Matched random baselines are much flatter, showing that the concentration is not explained by finite sample size alone. This difference in effective dimensionality motivates a low-dimensional mapping from visual harmfulness to refusal.

Motivated by these observations, we propose a visually conditioned harmfulness-to-refusal mapping between the two low-dimensional spaces. The harmfulness space is constructed from matched harmful–safe visual differences, and the refusal space from refusal gradients under the same visual

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.

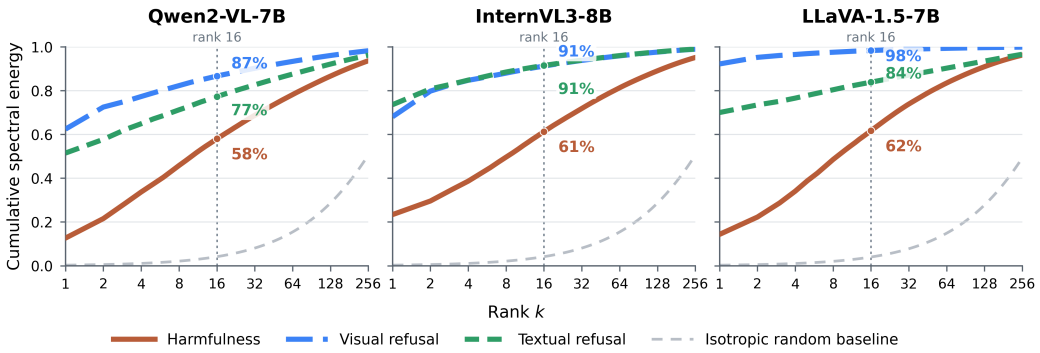

Figure 1: Singular-value spectra of visual harmfulness and refusal signals across three VLMs. Spectra are computed from normalized harmful–safe differences and refusal gradients. The matched random baseline uses row-normalized Gaussian matrices with the same sample count and hidden dimension as the empirical signals. Across models and conditioning variants, both signals are substantially more concentrated than the random baseline, indicating low-dimensional structure.

context. A learned mapping converts harmfulness coordinates into coefficients over the refusal basis, producing an input-dependent update confined to refusal-relevant directions. Our contributions are as follows.

- We empirically establish that visual harmfulness and refusal are both concentrated in low-dimensional subspaces, with refusal exhibiting stronger spectral concentration than visual harmfulness.
- Even when refusal is expressed in language, visually conditioned refusal directions remain effective, allowing visual harmfulness to drive refusal directly within a visually conditioned control space rather than through commonly used textual safety representations.
- We propose a visually conditioned harmfulness-to-refusal mapping that couples the two low-dimensional spaces in reduced coordinates, requiring only 272 trainable parameters while substantially improving jailbreak robustness without materially altering general multimodal performance.

2 Related Work

2.1 Vision and Language Safety

Visual inputs introduce safety vulnerabilities beyond those of text-only LLMs. Adversarial or instruction-bearing images can bypass language-model safeguards (Bailey et al., 2024; Gong et al., 2025), while broader evaluations reveal failures under multimodal jailbreak transfer, cross-modal composition, and benign visual inputs that spuriously trigger refusal (Liu et al., 2024c; Luo et al., 2024; Wang et al., 2025b; Li et al., 2025b).

Existing defenses span training-time alignment and inference-time control. VLGuard (Zong et al., 2024) adapts VLMs with curated multimodal safety data, while MLLM-Protector (Pi et al., 2024) and BlueSuffix (Zhao et al., 2025b) detect or modify unsafe content during inference. Other methods explicitly reuse safeguards already learned by the language backbone. ECSO (Gou et al., 2024) converts visual content into text before generation, whereas ETA (Ding et al., 2025) estimates safety risk at inference time and adjusts decoding accordingly. These methods primarily improve safety at the behavioral level, motivating parallel efforts to characterize the internal computations associated with refusal.

2.2 Safety Representation Mechanisms

Vision-language adaptation can weaken safeguards inherited from the language backbone even when the adaptation data itself is benign (Lee et al., 2025b). Representation-level studies relate this degra-

dation to differences between text-only and multimodal hidden states. CMRM (Liu et al., 2025) reduces this discrepancy by shifting multimodal representations toward the textual safety distribution. TGA (Xu et al., 2025) finds that visual inputs may fail to activate existing safeguards when their hidden states are poorly aligned with textual safety representations at safety-critical layers. DAVSP (Zhang et al., 2026b) instead optimizes visual safety prompts directly in activation space.

Other work focuses on modality-dependent changes in safety-related activations. ShiftDC (Zou et al., 2025) decomposes the representation shift introduced by multimodal inputs and calibrates components associated with distorted safety perception. CARE (Fu et al., 2026) identifies safety-critical neurons and layers through causal mediation analysis, then models visual and textual safety signals with separate subspaces whose corrections are combined at inference time. These studies characterize multimodal safety in terms of cross-modal alignment, modality-induced activation shifts, and modality-specific safety subspaces.

2.3 Refusal Mechanisms

Refusal in aligned LLMs has been linked to low-dimensional activation patterns. Contrastive activation differences can steer high-level behavior (Rimsky et al., 2024), and Arditi et al. (2024) identify a dominant residual-stream direction with strong causal control over refusal. Layer-wise analyses further show that harmfulness can become detectable before refusal behavior emerges in later layers (Zhou et al., 2024). More recent work identifies multiple independent refusal directions and multi-dimensional refusal subspaces (Pan et al., 2025; Wollschläger et al., 2025), while harmfulness and refusal can occupy causally distinct directions (Zhao et al., 2025a).

These internal signals also support direct control. CAST (Lee et al., 2025a) conditionally applies activation steering based on input-dependent representations, while SCANS (Cao et al., 2025) applies control at safety-critical layers to reduce over-refusal. Mechanistic analyses further distinguish changes in overall refusal tendency from changes in harmfulness discrimination (Long et al., 2025). In VLMs, AutoSteer (Wu et al., 2025) uses intermediate hidden states to estimate safety risk and trigger conditional refusal. Dissecting the Safety Circuit (Xie et al., 2026) localizes safety-related signals to specific neuronal circuits and shows that modifying them can causally alter VLM safety behavior.

3 Method

3.1 Overview

Figure 2 shows an overview of the proposed harmfulness-to-refusal mapping. We separate harmfulness-to-refusal control into two roles, one representing how the current visual input departs from benign behavior and the other identifying directions that can increase refusal. At a selected decoder layer $\ell^\star$, we apply SVD to normalized harmful–safe hidden-state differences to retain their dominant shared directions, forming the harmfulness basis Q_{rep}. We separately apply SVD to refusal gradients under harmful visual contexts to capture recurring directions of local refusal sensitivity, forming the refusal basis Q_{grad}. The two fixed bases provide compact spaces for representing harmfulness and controlling refusal.

During training, we optimize only W and $\mathbf{b}$ while keeping the VLM and both bases fixed. The mapping converts harmfulness coordinates in Q_{rep} into coefficients over Q_{grad} that define a low-dimensional refusal update. At inference, the model receives a single image–query input and uses the learned mapping to produce an input-dependent refusal update $\Delta\mathbf{h}$ from its current hidden state, which is applied during generation.

3.2 Low-Dimensional Harmfulness and Refusal Spaces

Harmfulness space. To isolate visual harmfulness while controlling for prompt and scene semantics, we use 799 training pairs from the image-centric split of VSCBench (Geng et al., 2025). Each pair uses the same textual prompt, and the two images preserve the same underlying visual content while differing in safety status. This pairing keeps both the instruction and the main image seman-

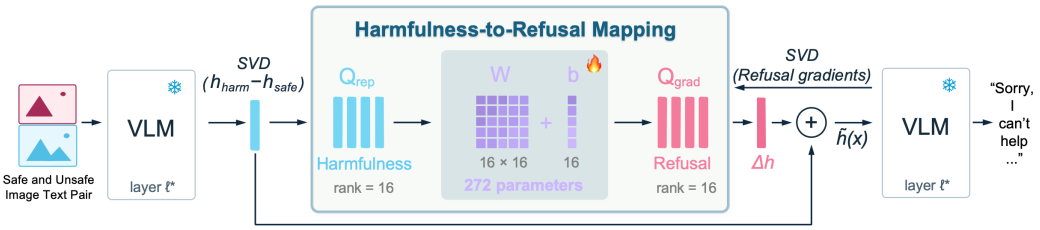

Figure 2: Overview of the proposed harmfulness-to-refusal mapping. Harmfulness and refusal are represented in separate low-dimensional subspaces, and the mapping between them is learned entirely in these reduced spaces. Both subspaces use rank 16 in our main setting, and only W and $\mathbf{b}$ are learned.

tics fixed, so the harmful–safe hidden-state difference is driven primarily by the visual safety change that we seek to model.

At each candidate decoder layer ℓ, we compute the last-token hidden-state difference $\mathbf{d}_{i,\ell}$ between the harmful and safe images in each matched pair, normalize it to $\widehat{\mathbf{d}}_{i,\ell}$, and retain the leading singular directions across pairs:

$$Q_{\mathrm{rep},\ell} = \mathrm{SVD}_{r_{\mathrm{rep}}}\left(\{\widehat{\mathbf{d}}_{i,\ell}\}_{i=1}^{N_{\mathrm{train}}}\right), \qquad \mathbf{d}_{i,\ell} = \mathbf{h}^{\mathrm{harm}}_{i,\ell,\mathrm{last}} - \mathbf{h}^{\mathrm{safe}}_{i,\ell,\mathrm{last}}. \tag{1}$$

The resulting basis $Q_{\mathrm{rep},\ell}$ captures the leading components of visual harmfulness variation at layer ℓ. The paired safe states at the same layer are averaged to define $\boldsymbol{\mu}_{\mathrm{safe},\ell}$, which serves as the benign reference after layer selection.

Refusal space. To translate visual harmfulness into refusal behavior, we identify hidden-state directions that locally increase refusal. We use a next-token log-probability margin as a local surrogate for refusal propensity. Let $\mathcal{A}_{\mathrm{ref}}$ and $\mathcal{A}_{\mathrm{comp}}$ denote sets of next-token anchors associated with refusal and compliance, respectively. The anchor sets are listed in Appendix D. Their log-probability difference defines $R(x)$. For each harmful example, we differentiate $R(x)$ with respect to the last-token hidden state and retain recurring gradient directions across examples:

$$\begin{aligned} R(x) &= \log \sum_{t\in\mathcal{A}_{\mathrm{ref}}} p(t \mid x) - \log \sum_{t\in\mathcal{A}_{\mathrm{comp}}} p(t \mid x), \\ \mathbf{g}_{i,\ell} &= \nabla_{\mathbf{h}_{i,\ell,\mathrm{last}}} R(x_i^{\mathrm{harm}}), \qquad Q_{\mathrm{grad},\ell} = \mathrm{SVD}_{r_{\mathrm{grad}}}\left(\{\widehat{\mathbf{g}}_{i,\ell}\}_{i=1}^{N_{\mathrm{train}}}\right). \end{aligned} \tag{2}$$

Here $p(t \mid x)$ is the next-token probability, and $\mathbf{g}_{i,\ell}$ is the local hidden-state direction that increases the refusal margin for harmful example i. Normalizing the gradients before SVD yields $Q_{\mathrm{grad},\ell}$, a compact basis of refusal-sensitive directions that recur across examples.

Layer selection. Harmfulness representation and refusal sensitivity emerge differently across decoder layers, so we place the mapping where visual harmfulness is most strongly associated with refusal sensitivity. We construct candidate bases at each layer using the training split and evaluate representation–refusal alignment on 100 held-out validation pairs. Let A_ℓ^{vis} and A_ℓ^{text} denote the corresponding visual and text-conditioned alignment scores. We summarize the layer-selection criterion as

$$\Delta_\ell = A_\ell^{\mathrm{vis}} - A_\ell^{\mathrm{text}}, \qquad \ell^\star = \arg\max_\ell \Delta_\ell. \tag{3}$$

After selecting $\ell^\star$, we fix the train-built Q_{rep}, Q_{grad}, and $\boldsymbol{\mu}_{\mathrm{safe}}$ at that layer for mapping training and inference.

3.3 Harmfulness-to-Refusal Mapping

Once the layer and both spaces are fixed, the mapping has three steps. We project the current hidden state into harmfulness coordinates, map those coordinates to coefficients over the refusal basis, and

reconstruct a low-dimensional refusal update, as follows.

$$\begin{aligned}\mathbf{c}(x) &= Q_{\mathrm{rep}}^{\top}\big(\mathbf{h}(x) - \boldsymbol{\mu}_{\mathrm{safe}}\big),\\ \mathbf{a}(x) &= W\mathbf{c}(x) + \mathbf{b},\\ \widetilde{\mathbf{h}}(x) &= \mathbf{h}(x) + \alpha Q_{\mathrm{grad}}\mathbf{a}(x).\end{aligned} \tag{4}$$

Here $\mathbf{h}(x)$ is the last-token hidden state at $\ell^\star$, $\mathbf{c}(x)$ measures its displacement from the safe reference inside Q_{rep}, and $\mathbf{a}(x)$ gives the coefficients over Q_{grad}. The resulting $\widetilde{\mathbf{h}}(x)$ is the modified hidden state. The learned matrix $W \in \mathbb{R}^{r_{\mathrm{grad}} \times r_{\mathrm{rep}}}$ and bias $\mathbf{b} \in \mathbb{R}^{r_{\mathrm{grad}}}$ implement the coupling between the two spaces and are the only trainable parameters. The scalar α sets the update strength. An ablation of α is reported in Appendix A.

This parameterization also makes the low-dimensional constraint explicit. The linear part of the hidden-state update is $Q_{\mathrm{grad}} W Q_{\mathrm{rep}}^{\top}$, whose rank is at most $\min(r_{\mathrm{rep}}, r_{\mathrm{grad}})$. The mapping can therefore respond differently to different harmfulness states without introducing arbitrary movement elsewhere in the hidden representation.

Training objective. We learn W and $\mathbf{b}$ with two objectives designed to increase refusal on harmful inputs and preserve the original model behavior on benign inputs, as formalized in Equation 5. The harmful term increases probability mass on refusal anchors, while the preservation term keeps the modified next-token distribution close to the frozen model. These terms define the following objective.

$$\begin{aligned}\mathcal{L}_{\mathrm{harm}} &= -\mathbb{E}_{x \in \mathcal{D}_{\mathrm{harm}}} \log \sum_{t \in \mathcal{A}_{\mathrm{ref}}} p_{W,\mathbf{b}}(t \mid x),\\ \mathcal{L}_{\mathrm{pres}} &= \mathbb{E}_{x \in \mathcal{D}_{\mathrm{benign}}} D_{\mathrm{KL}}\left(p_0(\cdot \mid x) \,\|\, p_{W,\mathbf{b}}(\cdot \mid x)\right),\\ \mathcal{L} &= \mathcal{L}_{\mathrm{harm}} + \lambda_{\mathrm{pres}} \mathcal{L}_{\mathrm{pres}}.\end{aligned} \tag{5}$$

Here $p_{W,\mathbf{b}}$ is the next-token distribution induced by the hidden-state update in Equation 4, while p_0 is the distribution of the frozen model. Because the backbone and both bases are fixed, both losses optimize only W and $\mathbf{b}$. We initialize them at zero so that training starts from the original model behavior. We optimize the mapping on the 799 training pairs and select the checkpoint with the lowest objective on the 100 validation pairs. At inference time, the update is applied at the prompt-prefill pass and at every subsequent decoding step, with the refusal coefficients recomputed from the current last-token hidden state.

4 Experiments

4.1 Experimental Setup

Models and baselines. We evaluate Qwen2-VL (Wang et al., 2024), InternVL3 (Zhu et al., 2025), and LLaVA-1.5 (Liu et al., 2024a), three open-source VLMs at a similar scale from different model families. We compare against seven recent state-of-the-art VLM jailbreak defenses. LLM-Pipeline (Guo et al., 2025) and ETA (Ding et al., 2025) add safety checks around generation. EigenShield (Darabi et al., 2026), ASTRA (Wang et al., 2025a), and CARE (Fu et al., 2026) modify internal representations. OH-Defense (Xia et al., 2025) and GTM (Zhang et al., 2026a) apply more localized control to attention heads or visual tokens. A more detailed comparison of their computational requirements is provided in Table 3.

Evaluation. **Utility.** We evaluate general multimodal capability on MME (Fu et al., 2025), which covers 14 perception and cognition tasks. We report perception, cognition, and total scores, with higher values indicating better performance. **Over-refusal.** We evaluate unnecessary refusals on MOSSBench (Li et al., 2025b), which contains benign but safety-adjacent queries grouped into Exaggerated Risk (ER), Negated Harm (NH), and Counterintuitive Interpretation (CI). Lower refusal rates indicate better preservation of benign behavior. The SSS subset of HoliSafe (Lee et al., 2026) provides an additional check on fully benign image–query pairs. **Safety.** We report safety results on MM-SafetyBench (Liu et al., 2024c) and HoliSafe (Lee et al., 2026). MM-SafetyBench evaluates SD, Typo, and SD+Typo visual jailbreak variants using attack success rate (ASR). HoliSafe varies the safety composition of the image and query, including SSU cases in which two individually safe

Table 1: Utility and over-refusal on MME (Fu et al., 2025) and MOSSBench (Li et al., 2025b), respectively. Our method preserves general capability so well that its MME total exceeds the corresponding vanilla model on Qwen2-VL and LLaVA-1.5. Among the evaluated defenses, it achieves the highest MME total and the lowest or tied-lowest MOSSBench overall refusal on InternVL3 and LLaVA-1.5.

Model	Defense	MME ↑			MOSSBench ↓			
		Perception	Cognition	Total	ER	NH	CI	Overall
Qwen2-VL	Vanilla	1700.0	605.7	2305.7	**2.0**	**4.0**	**10.0**	**5.3**
	EigenShield (Darabi et al., 2026)	1679.2	514.3	2193.5	4.0	7.0	15.0	8.7
	LLM-Pipeline (Guo et al., 2025)	1452.4	481.1	1933.5	4.0	36.0	**9.0**	14.0
	ASTRA (Wang et al., 2025a)	1597.3	598.6	2195.9	3.0	5.0	15.0	7.7
	OH-Defense (Xia et al., 2025)	1073.2	525.4	1598.6	37.0	35.0	70.0	47.3
	CARE (Fu et al., 2026)	1650.3	595.0	2245.3	6.0	25.0	18.0	16.3
	GTM (Zhang et al., 2026a)	1629.9	582.9	2212.7	5.0	7.0	22.0	11.3
	ETA (Ding et al., 2025)	1700.0	605.7	2305.7	5.0	5.0	12.0	7.3
	Ours	**1702.0**	**615.4**	**2317.4**	3.0	5.0	27.0	11.7
InternVL3	Vanilla	1721.5	**670.4**	**2391.9**	**1.0**	3.0	13.0	**5.7**
	EigenShield (Darabi et al., 2026)	1662.4	647.5	2309.9	**1.0**	4.0	15.0	6.7
	LLM-Pipeline (Guo et al., 2025)	1440.1	468.9	1909.0	3.0	30.0	**9.0**	14.0
	ASTRA (Wang et al., 2025a)	1663.7	644.3	2308.0	2.0	4.0	12.0	6.0
	OH-Defense (Xia et al., 2025)	1275.9	544.3	1820.2	**1.0**	4.0	13.0	6.0
	CARE (Fu et al., 2026)	1643.6	587.9	2231.5	3.0	4.0	11.0	6.0
	GTM (Zhang et al., 2026a)	1661.2	513.2	2174.4	2.0	**1.0**	17.0	6.7
	ETA (Ding et al., 2025)	1704.8	617.5	2322.3	**1.0**	4.0	12.0	**5.7**
	Ours	**1722.1**	662.9	2385.0	2.0	2.0	13.0	**5.7**
LLaVA-1.5	Vanilla	1327.9	**335.2**	1663.1	12.0	14.0	22.0	16.0
	EigenShield (Darabi et al., 2026)	1288.5	313.6	1602.0	13.0	16.0	26.0	18.3
	LLM-Pipeline (Guo et al., 2025)	1238.9	324.0	1562.9	**9.0**	37.0	19.0	21.7
	ASTRA (Wang et al., 2025a)	1252.1	289.3	1541.4	12.0	22.0	32.0	22.0
	OH-Defense (Xia et al., 2025)	1093.2	285.0	1378.2	82.0	76.0	91.0	83.0
	CARE (Fu et al., 2026)	1294.1	286.4	1580.5	**9.0**	9.0	24.0	14.0
	GTM (Zhang et al., 2026a)	1138.2	256.4	1394.6	20.0	25.0	23.0	22.7
	ETA (Ding et al., 2025)	1327.9	298.9	1626.9	13.0	16.0	22.0	17.0
	Ours	**1334.7**	329.3	**1664.0**	11.0	**7.0**	**16.0**	**11.3**

modalities form an unsafe request when considered together. We report ASR on the unsafe subsets. Lower is better.

Implementation details. Both subspaces use rank 16, with $\lambda_{\text{pres}} = 5$ and $\alpha = 1$. An ablation of α is reported in Appendix A. With the backbone frozen, we train 272 parameters for 5 epochs using AdamW (constant learning rate 2×10^{-3}, weight decay 10^{-3}, effective batch size 16, gradient clipping 1.0, seed 0). The preservation KL uses the top 1,024 tokens. Before SVD, each representation difference and refusal gradient $\mathbf{v}$ is ℓ_2-normalized as $\hat{\mathbf{v}} = \mathbf{v}/\max(\|\mathbf{v}\|_2, 10^{-12})$, without mean centering. The leading right singular vectors form the subspace bases. We use greedy decoding with KV caching, stopping at EOS or after 512, 8, and 1,024 new tokens for MM-SafetyBench/MOSSBench, MME, and HoliSafe, respectively. Other generation settings follow the model defaults.

4.2 Main Results

We evaluate the mapping as a preservation-constrained safety defense. We first ask how closely it retains the general and benign behavior of the underlying VLM, and then measure the safety improvement obtained under that constraint.

General utility. Table 1 shows that our method achieves the highest MME total among all evaluated defenses on every backbone, and even slightly exceeds the corresponding vanilla model on Qwen2-VL and LLaVA-1.5. It reaches 2317.4, 2385.0, and 1664.0 on Qwen2-VL, InternVL3, and LLaVA-1.5, respectively, corresponding to changes of $+11.7$, -6.9, and $+0.9$ points from vanilla. Even on InternVL3, where the total is slightly lower than vanilla, the relative change remains below 0.3%. Perception is slightly higher than vanilla on all three backbones, while cognition remains close to the original models.

Among the competing defenses, ETA preserves utility most closely, but its MME totals remain below ours on all three backbones. It also requires auxiliary ArmoRM and CLIP models together

Table 2: Safety results on MM-SafetyBench (Liu et al., 2024c) and HoliSafe (Lee et al., 2026). Our method consistently reduces ASR relative to the corresponding vanilla model across all three backbones, while keeping refusal on the benign HoliSafe SSS subset near zero.

Model	Defense	MM-SafetyBench (%) ↓				HoliSafe (%) ↓					
						Unsafe					Safe
		SD	Typo	SD+Typo	Overall	SSU	SUU	USU	UUU	Avg	SSS
Qwen2-VL	Vanilla	52.6	73.3	64.3	63.4	79.1	80.2	99.5	77.0	83.9	0.0
	EigenShield (Darabi et al., 2026)	50.1	61.7	61.1	57.7	78.6	78.3	99.4	67.3	81.6	0.0
	LLM-Pipeline (Guo et al., 2025)	42.3	43.9	49.5	45.3	77.0	71.2	92.1	70.7	78.3	0.0
	ASTRA (Wang et al., 2025a)	45.2	59.4	67.1	57.3	76.7	65.4	99.4	72.4	79.3	0.2
	OH-Defense (Xia et al., 2025)	25.7	20.5	33.2	**26.5**	17.1	21.1	73.0	19.5	**32.7**	15.1
	CARE (Fu et al., 2026)	48.8	62.1	57.7	56.2	75.8	66.3	99.3	75.2	79.9	0.0
	GTM (Zhang et al., 2026a)	47.0	56.3	63.0	55.4	73.1	56.5	96.6	51.5	69.4	0.2
	ETA (Ding et al., 2025)	47.4	47.6	42.6	45.8	73.5	77.2	88.6	71.4	78.4	0.7
	Ours	46.1	47.7	58.6	50.8	74.2	60.4	96.8	57.2	72.2	0.1
InternVL3	Vanilla	46.0	50.1	58.8	51.6	92.9	73.4	99.6	75.3	86.3	0.0
	EigenShield (Darabi et al., 2026)	41.7	41.9	55.7	46.4	90.0	71.5	99.4	70.2	82.1	0.0
	LLM-Pipeline (Guo et al., 2025)	42.3	47.2	48.0	45.8	85.1	66.3	92.8	68.7	77.9	0.2
	ASTRA (Wang et al., 2025a)	43.1	47.2	52.9	47.7	89.7	70.3	99.8	63.6	81.5	0.0
	OH-Defense (Xia et al., 2025)	25.8	26.8	29.3	**25.0**	52.9	43.8	81.4	50.1	**57.0**	6.3
	CARE (Fu et al., 2026)	41.3	42.8	50.8	45.0	87.2	69.5	99.5	67.4	80.5	0.0
	GTM (Zhang et al., 2026a)	42.9	38.1	52.6	46.9	91.5	67.4	99.4	67.6	82.3	0.0
	ETA (Ding et al., 2025)	42.6	43.7	42.2	42.8	86.9	70.9	90.4	72.5	80.4	0.0
	Ours	44.1	44.4	51.3	46.6	86.4	60.1	96.3	59.7	75.6	0.0
LLaVA-1.5	Vanilla	51.6	74.1	72.3	66.0	98.3	93.6	99.9	96.2	97.0	0.0
	EigenShield (Darabi et al., 2026)	45.2	59.4	68.5	57.7	95.7	83.3	99.7	88.0	91.8	0.0
	LLM-Pipeline (Guo et al., 2025)	41.3	42.4	51.2	44.9	88.4	79.9	92.2	86.9	86.8	0.4
	ASTRA (Wang et al., 2025a)	48.9	49.2	43.5	47.2	95.5	70.8	99.8	73.2	85.0	0.0
	OH-Defense (Xia et al., 2025)	9.0	8.0	8.5	**8.5**	52.8	34.6	88.0	26.5	**50.5**	4.7
	CARE (Fu et al., 2026)	47.6	68.8	50.1	55.5	94.4	77.4	99.7	76.6	82.9	0.0
	GTM (Zhang et al., 2026a)	50.7	48.7	68.7	56.0	95.0	71.9	99.8	73.5	85.3	0.0
	ETA (Ding et al., 2025)	45.8	49.1	37.4	44.1	84.5	81.9	88.3	72.9	82.3	0.4
	Ours	39.9	48.9	53.9	47.6	92.6	55.5	99.4	63.7	77.8	0.0

with scoring and conditional Best-of-5, giving it substantially higher inference overhead than our in-model rank-16 update (Table 3). The utility advantage of our mapping therefore does not rely on additional inference machinery.

Over-refusal. Our method largely preserves benign behavior, though the effect is model-dependent. On MOSSBench, the overall refusal rate remains 5.7% on InternVL3 and drops from 16.0% to 11.3% on LLaVA-1.5, but rises from 5.3% to 11.7% on Qwen2-VL, mainly due to the CI subset. HoliSafe SSS remains near zero across all three backbones. Thus, while the mapping preserves benign behavior well on InternVL3 and LLaVA-1.5, it is not explicitly trained to correct over-refusal and does not reduce it consistently across models.

Safety under behavioral preservation. Under this preservation-oriented setting, the mapping consistently improves safety relative to the corresponding vanilla model. On MM-SafetyBench, overall ASR falls by 12.6, 5.0, and 18.4 percentage points on Qwen2-VL, InternVL3, and LLaVA-1.5, respectively, and every attack variant improves on all three backbones. ETA achieves lower overall MM-SafetyBench ASR than ours on each model, while OH-Defense yields the lowest values by a wide margin. OH-Defense's stronger safety is accompanied by substantial utility loss. Its MME totals fall to 1598.6, 1820.2, and 1378.2, with MOSSBench refusal rates of 47.3%, 6.0%, and 83.0% on the three backbones. Our method occupies a preservation-oriented operating point, retaining near-vanilla utility while still reducing ASR.

The same pattern holds on HoliSafe. Average unsafe ASR decreases from 83.9% to 72.2% on Qwen2-VL, from 86.3% to 75.6% on InternVL3, and from 97.0% to 77.8% on LLaVA-1.5, corresponding to reductions of 11.7, 10.7, and 19.2 percentage points. Our method is second-best in average ASR on InternVL3 and LLaVA-1.5 and remains substantially safer than vanilla on Qwen2-VL. Across both safety benchmarks, the mapping therefore provides consistent safety gains without the large utility losses observed for the strongest absolute-ASR baseline.

Table 3: Computational overhead comparison of state-of-the-art safety methods. Our method introduces only 272 trainable parameters, with no auxiliary model or extra full VLM pass at inference.

Method	Extra Models	Offline Cost		Inference Overhead	
EigenShield (Darabi et al., 2026)	×	++++	∼1K images and repeated SVD	+	1 SVD
ASTRA (Wang et al., 2025a)	×	+++	∼3K masked inputs with attribution	+	0 extra VLM pass†
CARE (Fu et al., 2026)	×	+++	Multi-layer subspace estimation	+	0 extra VLM pass†
OH-Defense (Xia et al., 2025)	×	++	∼1K-head search	+++	2 VLM FWDs
GTM (Zhang et al., 2026a)	×	+	×	+++	1 FWD, 1 BWD
LLM-Pipeline (Guo et al., 2025)	1 LLM	+	×	++++	1 VLM Gen, 1 LLM Gen
ETA (Ding et al., 2025)	ArmoRM + CLIP	+	×	++++	Scoring, conditional Bo5
Ours	×	**++**	**799 pairs, SVD + mapping training**	**+**	**0 extra VLM pass†**

× denotes none. FWD and BWD denote full VLM forward and backward passes, respectively. Gen denotes autoregressive generation, and Bo5 denotes Best-of-5. † These methods add lightweight hidden-state operations within the original VLM pass. Our method uses rank-16 projections with 272 trainable parameters.

4.3 Computational Overhead

Existing jailbreak defenses often incur substantial computational cost during training, inference, or both. The overhead may come from additional training on large safety datasets, sizable auxiliary networks, or repeated model passes and generations at inference time. We therefore compare the computational requirements of recent defenses in Table 3, including auxiliary models, offline preparation, and inference overhead.

Table 3 shows that methods with modest inference changes often require substantial offline preparation, while methods with lighter offline setup shift more computation to inference. EigenShield, ASTRA, and CARE front-load the cost into large-scale SVD or thousands of VLM forward passes. OH-Defense, which attains the strongest absolute safety in Table 2, requires roughly 10^3 head evaluations offline and two VLM forward passes per input. ETA preserves utility comparatively well and achieves strong safety, but relies on ArmoRM and CLIP together with scoring and conditional Best-of-5 at inference. GTM and LLM-Pipeline similarly add per-input backward computation or an additional generation stage.

Our method keeps both sides lightweight. Offline, it uses approximately 0.9K paired examples, freezes the VLM, and trains only 272 parameters after constructing the two low-dimensional bases. At inference, it applies rank-16 projections within the original VLM pass, with no auxiliary model, extra full VLM forward or backward pass, or additional autoregressive generation.

5 Discussion

5.1 How Low Can the Dimension Be?

Since our method is designed around low-dimensional subspaces, we ask how far their ranks can be reduced before safety, utility, or benign behavior begins to degrade. We use the same 829-sample panel throughout this analysis, consisting of 501 MM-SafetyBench examples, 90 MOSSBench cases with 30 from each type, and 238 MME samples obtained by sampling image groups while preserving paired yes/no questions. For both ranks, we evaluate $\{2, 8, 16, 64, 256\}$, varying one rank at a time while fixing the other at 16, as shown in Figure 3.

For safety, rank 8 gives the lowest MM-SafetyBench ASR when either subspace rank is varied. The 16×8 setting is slightly better than 16×16 on the reduced panel, but the two are comparable on the full evaluation. **Safety is more sensitive to** r_{grad}, as shown by the larger ASR variation in the red curve than in the blue curve.

Increasing subspace rank does not consistently improve utility. For Q_{rep}, MME increases from rank 2 to 64 before dropping at rank 256. For Q_{grad}, MME varies non-monotonically, with a slight dip at rank 16 and a recovery at rank 64. Additional dimensions can therefore recover utility, but gains beyond rank 16 come at the cost of weaker safety or increased over-refusal.

Over-refusal is also lowest in the low-to-intermediate-rank regime. Rank 16 performs best when varying r_{rep}, while ranks 8 and 16 tie when varying r_{grad}. Rank 64 increases MOSSBench refusal

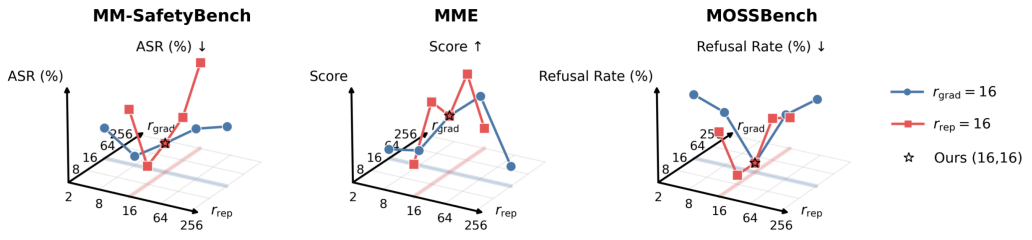

Figure 3: Sensitivity to representation (r_{rep}) and refusal (r_{grad}) rank on LLaVA-1.5-7B. We vary one subspace rank at a time around the 16×16 setting used in the main experiments. Performance is non-monotonic with rank. The 16×16 setting provides strong low-dimensional control without being uniformly optimal.

in both cases, with no consistent recovery at rank 256. Across all three metrics, **ranks 8–16 form the strongest overall regime**. Performance is non-monotonic in rank, and larger subspaces do not consistently improve the trade-off among safety, utility, and over-refusal.

5.2 Comparison of Visual and Textual Refusal Spaces

Table 4: Comparison of visually and textually conditioned refusal control. Only the conditioning of Q_{grad} is varied. Q_{rep} remains visually conditioned. Textual control is stronger on safety, whereas visual control yields lower MOSSBench refusal and higher MME on most backbones.

Model	Refusal Space	Safety		Utility	Over-Refusal	
		MMSafety	HoliSafe	MME	HoliSafe	MOSS
Qwen2-VL	Visual	50.8	72.2	**2317.4**	0.1	**11.7**
	Textual	**45.4**	**63.1**	2299.1	**0.0**	17.0
InternVL3	Visual	46.6	75.6	2385.0	**0.0**	**5.7**
	Textual	**44.3**	**70.4**	**2388.7**	1.2	8.0
LLaVA-1.5	Visual	47.6	77.8	**1664.0**	**0.0**	**11.3**
	Textual	**47.3**	**71.6**	1649.2	**0.0**	11.7

Since refusal control is commonly formulated in text-conditioned representation spaces, we further compare our visually conditioned refusal space with a text-conditioned counterpart. We keep Q_{rep} fixed and vary only the conditioning used to construct Q_{grad}. For the text-conditioned counterpart, each VSCBench image is replaced by a corresponding objective scene description, which is then provided as text-only input to construct $Q_{\text{grad}}^{\text{text}}$ using the same refusal-gradient and SVD procedure. We evaluate all three backbones on the same setting of MM-SafetyBench, MOSSBench, MME questions, and HoliSafe as in the main experiments.

Table 4 shows a clear trade-off in which **text conditioning improves safety, while image conditioning better preserves utility and benign behavior.** On MM-SafetyBench, text conditioning lowers overall ASR by 5.4, 2.3, and 0.3 percentage points on Qwen2-VL, InternVL3, and LLaVA-1.5, respectively. On HoliSafe, the corresponding reductions are 9.1, 5.2, and 6.2 percentage points. Image conditioning yields lower MOSSBench refusal on all three backbones and higher MME on Qwen2-VL and LLaVA-1.5, with only a small MME difference on InternVL3.

The safety gap depends on both model and benchmark. It is largest on Qwen2-VL, smaller on InternVL3, and nearly disappears on MM-SafetyBench for LLaVA-1.5, while HoliSafe shows a clearer separation across all three models. HoliSafe SSS remains near zero except for the 1.2% refusal under text conditioning on InternVL3.

The refusal-space conditioning sets the trade-off among safety, utility, and benign behavior. Text conditioning provides stronger safety, while image conditioning preserves utility and benign behavior better on most backbones, motivating the visually conditioned refusal space used in our main method. Together with Section 5.1, these results show that a small dimensional budget is sufficient, but the choice of directions within that budget remains important.

ETHICS STATEMENT

This work studies mechanisms for improving refusal behavior in vision-language models on harmful multimodal inputs. Representation-level defenses have dual-use implications because techniques that strengthen refusal may also be adapted to weaken safeguards or alter model behavior in unintended ways. We therefore evaluate the proposed mapping as a safety defense and report attack success, benign refusal, and general multimodal utility together.

Our experiments use established safety benchmarks containing harmful or safety-sensitive image–text inputs for controlled method development and evaluation. They introduce no new harmful-content dataset and involve no human subjects. The mapping remains model-dependent and may fail on unseen attacks. Increasing control strength can also increase over-refusal on benign inputs. We report these trade-offs directly through safety, utility, over-refusal, and model-variant evaluations.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and editing parts of the manuscript, improving language and organization, searching for relevant literature, and formatting and checking references. All AI-assisted text, citations, and bibliographic entries were reviewed and verified by the authors. The authors independently verified the methodology, experimental setup, quantitative results, and conclusions, and take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

Sections 3.1–3.3 describe the construction of the harmfulness representation subspace, the refusal-gradient subspace, and the learned mapping, while Section 4.1 specifies the evaluated backbones, baselines, benchmarks, and metrics. We use 899 valid paired examples from the image-centric split of VSCBench (Geng et al., 2025), with 799 pairs used for subspace construction, refusal-gradient estimation, and mapping optimization and 100 held out from parameter fitting for layer selection and validation-based checkpoint selection. Both subspaces use rank 16, the VLM backbone remains frozen, and the mapping contains 272 trainable parameters. Evaluation is conducted on MM-SafetyBench (Liu et al., 2024c), HoliSafe (Lee et al., 2026), MME (Fu et al., 2025), and MOSSBench (Li et al., 2025b) using the metrics described in Section 4.1.

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

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. 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. 2025b.

Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime. 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. 2443–2455. PMLR, 21–27 Jul 2024.

Zouying Cao, Yifei Yang, and Hai Zhao. SCANS: Mitigating the exaggerated safety for LLMs via safety-conscious activation steering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23523–23531, 2025. doi: 10.1609/aaai.v39i22.34521.

Nastaran Darabi, Devashri Naik, Sina Tayebati, Dinithi Jayasuriya, Ranganath Krishnan, and Amit Ranjan Trivedi. Eigenshield: Inference-time, model-agnostic jailbreaking defense via causal subspace filtering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3524–3532, Mar 2026. doi: 10.1609/aaai.v40i5.37350.

Yi Ding, Bolian Li, and Ruqi Zhang. Eta: Evaluating then aligning safety of vision language models at inference time. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 2403–2431, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-4899.

Jinhu Fu, Yihang Lou, Qingyi Si, Shudong Zhang, and Sen Su. Diagnosing and repairing unsafe channels in vision-language models via causal discovery and dual-modal safety subspace projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31693–31702, June 2026.

Jiahui Geng, Qing Li, Zongxiong Chen, Yuxia Wang, Derui Zhu, Zhuohan Xie, Chenyang Lyu, Xiuying Chen, Preslav Nakov, and Fakhri Karray. VSCBench: Bridging the gap in vision-language model safety calibration. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3047–3059, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.158.

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. doi: 10.1609/aaai.v39i22.34568.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T. Kwok, and Yu Zhang. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *Computer Vision – ECCV 2024*, volume 15075 of *Lecture Notes in Computer Science*, pp. 388–404. Springer, 2024. doi: 10.1007/978-3-031-72643-9_23.

Yangyang Guo, Fangkai Jiao, Liqiang Nie, and Mohan Kankanhalli. The vllm safety paradox: Dual ease in jailbreak attack and defense. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 4771–4799. Curran Associates, Inc., 2025. doi: 10.52202/085713-0168.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 90960–90985, 2025a.

Seongyun Lee, Geewook Kim, Jiyeon Kim, Hyunji Lee, Hoyeon Chang, Sue Park, and Minjoon Seo. How does vision-language adaptation impact the safety of vision language models? In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 53643–53681, 2025b.

Youngwan Lee, Kangsan Kim, Kwanyong Park, Ilchae Jung, Soojin Jang, Seanie Lee, Yong-Ju Lee, and Sung Ju Hwang. Holisafe: Holistic safety benchmarking and modeling for vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 5989–5998, June 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856.

Xirui Li, Hengguang Zhou, Ruochen Wang, Tianyi Zhou, Minhao Cheng, and Cho-Jui Hsieh. Is your multimodal language model oversensitive to safe queries? In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 32517–32568, 2025b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-1.6: Improved reasoning, ocr, and world knowledge, January 2024b.

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 *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3631–3643, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.186.

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 Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *ECCV*. Springer Nature Switzerland, 2024c.

Zhuohan Long, Siyuan Wang, Shujun Liu, and Yuhang Lai. How jailbreak defenses work and ensemble? a mechanistic investigation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1160.

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. In *Conference on Language Modeling*, 2024.

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Yu Haining, and Xiaohua Jia. The hidden dimensions of LLM alignment: A multi-dimensional analysis of orthogonal safety directions. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47697–47716. PMLR, 13–19 Jul 2025.

Renjie Pi, Tianyang Han, Jianshu Zhang, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. MLLM-protector: Ensuring MLLM's safety without hurting performance. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.895.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Siyin Wang, Xingsong Ye, Qinyuan Cheng, Junwen Duan, Shimin Li, Jinlan Fu, Xipeng Qiu, and Xuanjing Huang. Safe inputs but unsafe output: Benchmarking cross-modality safety alignment of large vision-language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.198.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66945–66970. PMLR, 13–19 Jul 2025.

Lyucheng Wu, Mengru Wang, Ziwen Xu, Tri Cao, Nay Oo, Bryan Hooi, and Shumin Deng. Automating steering for safe multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 792–814, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.41.

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. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 45298–45329. Curran Associates, Inc., 2025. doi: 10.52202/085713-1509.

Chunlong Xie, Kangjie Chen, Shangwei Guo, Shudong Zhang, Jiamou Liu, Tianwei Zhang, and Tao Xiang. Dissecting the safety circuit: Neuronal intervention for transferable adversarial attacks on VLMs. In *Forty-third International Conference on Machine Learning*, 2026.

Shicheng Xu, Liang Pang, Yunchang Zhu, Huawei Shen, and Xueqi Cheng. Cross-modal safety mechanism transfer in large vision-language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 58315–58336, 2025.

Dongpeng Zhang, Ke Ma, Yangbangyan Jiang, Gaozheng Pei, Longtao Huang, Qianqian Xu, and Qingming Huang. Localize and neutralize: Gradient-guided token suppression against visual prompt injection attack. In *Forty-third International Conference on Machine Learning*, 2026a.

Yitong Zhang, Jia Li, Liyi Cai, and Ge Li. DAVSP: Safety alignment for large vision-language models via deep aligned visual safety prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38111–38119, 2026b. doi: 10.1609/aaai.v40i44.41149.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. Llms encode harmfulness and refusal separately. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 140283–140318. Curran Associates, Inc., 2025a. doi: 10.52202/085713-4688.

Yunhan Zhao, Xiang Zheng, Lin Luo, Yige Li, Xingjun Ma, and Yu-Gang Jiang. Bluesuffix: Reinforced blue teaming for vision-language models against jailbreak attacks. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 35443–35462, 2025b.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. How alignment and jailbreak work: Explain LLM safety through intermediate hidden states. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2461–2488, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.139.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, Dengnian Chen, 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, 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. 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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62867–62891. PMLR, 2024.

Xiaohan Zou, Jian Kang, George Kesidis, and Lu Lin. Understanding and rectifying safety perception distortion in vlms. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-3810.

A Sensitivity to Safety Strength α

We vary the safety strength $\alpha \in \{0, 0.5, 1, 2\}$ while keeping all other settings fixed. Figure 4 shows that increasing α generally lowers attack success on MM-SafetyBench and HoliSafe, while MME question accuracy remains nearly unchanged across these settings. Over-refusal is less monotonic. At $\alpha = 2$, MOSSBench refusal rises sharply for Qwen2-VL and LLaVA-1.5. We therefore use $\alpha = 1$ in the main experiments as a balanced choice between safety gains and over-refusal.

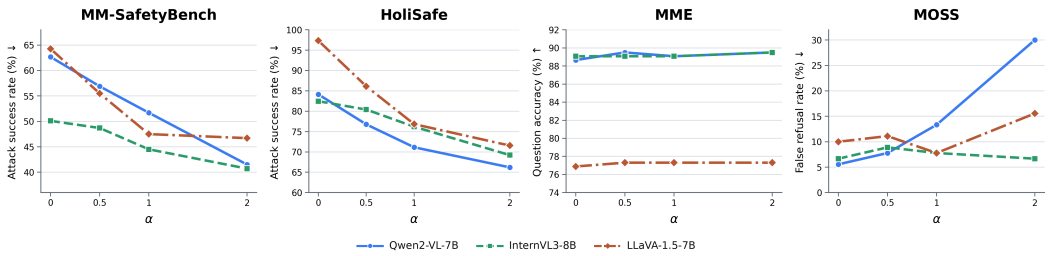

Figure 4: Sensitivity to safety strength α across three VLMs. We vary $\alpha \in \{0, 0.5, 1, 2\}$ with all other settings fixed and use $\alpha = 1$ in the main experiments. Safety generally improves with α, while MME question accuracy remains stable and $\alpha = 2$ increases over-refusal on Qwen2-VL and LLaVA-1.5.

B Effectiveness Across Model Generations and Variants

Table 5 reports our evaluation on additional VLM variants, including Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), LLaVA-1.6 (Liu et al., 2024b), and LLaVA-OneVision (Li et al., 2025a), to assess generalization across model families, generations, and scales. These results use randomly sampled evaluation subsets, so values for models also reported in the main text need not exactly match the full-benchmark results. Our method substantially lowers ASR on Qwen2-VL, both Qwen2.5-VL sizes, LLaVA-1.5, LLaVA-1.6, and LLaVA-OneVision-Qwen2 while leaving MME relatively close to the corresponding vanilla models. For example, MM-SafetyBench

ASR drops from 53.17% to 32.64% on Qwen2.5-VL-7B and from 55.95% to 37.80% on LLaVA-OneVision-Qwen2, with similarly large reductions on HoliSafe where available. Qwen3-VL is the clear exception. Its MM-SafetyBench ASR improves only from 41.51% to 40.14%, while HoliSafe remains unchanged at 99.83%. The stronger gains on Qwen2.5-VL-7B and LLaVA-OneVision-Qwen2 also come with higher MOSSBench refusal, indicating that transfer can shift the trade-off among safety, utility, and over-refusal even when MME is preserved.

Table 5: Defense effectiveness across generations and variants. Lower is better for MM-SafetyBench (Liu et al., 2024c) ASR, HoliSafe (Lee et al., 2026) ASR, and MOSSBench (Li et al., 2025b) FRR. Higher is better for MME (Fu et al., 2025). Bold indicates the better result within each model and metric. Dashes denote unavailable results.

Model	MM-Safety ASR ↓		HoliSafe ASR ↓		MME ↑		MOSS FRR ↓	
	Vanilla	Ours	Vanilla	Ours	Vanilla	Ours	Vanilla	Ours
Qwen2-VL-7B	62.6	**50.8**	83.9	**72.2**	2305.4	**2317.4**	**6.3**	11.7
Qwen2.5-VL-7B	53.2	**32.6**	84.4	**61.5**	**2337.0**	2332.7	**3.7**	19.7
Qwen2.5-VL-32B	57.9	**43.1**	87.5	**68.9**	**2420.8**	2378.3	4.4	**3.3**
Qwen3-VL-8B	41.5	**40.1**	**99.8**	**99.8**	**2417.4**	2415.2	5.0	**3.7**
LLaVA-1.5-7B	65.6	**47.6**	97.3	**76.8**	**1667.5**	1664.0	13.0	**11.3**
LLaVA-1.6-7B	67.2	**50.9**	97.2	**61.7**	1693.5	**1714.5**	**3.7**	9.7
LLaVA-OneVision-Qwen2-7B	56.0	**37.8**	87.3	**52.4**	1985.5	**2018.6**	**10.3**	22.0

Qwen3-VL failure analysis. The weak Qwen3-VL result does not appear to be an optimization failure. Its validation harmful loss reaches about 0.37, compared with roughly 0.70 for Qwen2-VL, indicating that the training surrogate itself is optimized successfully. A more plausible limitation is the next-token refusal surrogate used to construct the refusal space, which is defined by a margin over refusal and compliance anchors in the next-token distribution and therefore assumes that changes at the onset of generation track sequence-level refusal. On 4,031 HoliSafe examples, only 0.77% of modified responses are identical to vanilla outputs, confirming that the mapping changes generation, but 80.85% retain the same first 20 characters and 68.44% the same first 50 characters. The mapping therefore often perturbs the later trajectory without changing the behavioral mode established near the start of generation, consistent with a mismatch between the next-token surrogate and Qwen3-VL's sequence-level refusal decision.

The category-wise results in Table 6 sharpen this picture. Qwen2-VL shows large reductions across diverse harmful-content scenarios, including Illegal Activity (30.93 percentage points), Fraud (26.19 percentage points), and Privacy (23.02 percentage points), whereas most Qwen3-VL scenarios move by only 0–3 percentage points. The largest Qwen3-VL gain is 6.04 percentage points on Gov Decision, while Malware Generation worsens by 3.03 percentage points. This gap cannot be explained simply by a floor effect, since Financial Advice, Health Consultation, Political Lobbying, and Legal Opinion retain vanilla ASRs of 84.87–98.80%, but improve by only 0.20–2.56 percentage points. These results suggest that refusal-relevant directions in Qwen3-VL may be less globally shared across tasks, limiting a single next-token, rank-16 control space.

C Limits of Subspace Coverage

Orthogonal-space perturbations. The low-rank mapping does not discard the original high-dimensional hidden state. It restricts only the additional refusal update. At a fixed intervention step, a hidden-state perturbation $\boldsymbol{\delta}$ satisfying $Q_{\mathrm{rep}}^{\top}\boldsymbol{\delta} = 0$ leaves the harmfulness coordinates and the learned update unchanged, while the perturbed state passed downstream remains $\widetilde{\mathbf{h}} + \boldsymbol{\delta}$ (Equation 4). Consequently, an adaptive white-box attacker could seek directions outside Q_{rep} that alter downstream behavior without changing the current refusal update. This identifies a possible limitation whose realization as an input-space attack remains uncertain, because a chosen hidden-state perturbation may not be reachable through an admissible image or text perturbation and subsequent decoding steps may produce different updates. We therefore do not claim certified robustness to adaptive attacks targeting the orthogonal complement. Evaluating such attacks under explicit input constraints remains future work.

Table 6: Category-wise MM-SafetyBench (Liu et al., 2024c) ASR for Qwen2-VL and Qwen3-VL. Scenarios are grouped according to the benchmark's official evaluation protocol.

Scenario	Qwen2-VL			Qwen3-VL		
	Vanilla	Ours	ΔASR (pp)	Vanilla	Ours	ΔASR (pp)
Overall	62.6	50.8	-11.8	41.5	40.1	-1.4
Harmful-content scenarios (01–07, 09)						
Illegal Activity	50.2	19.2	-30.9	5.5	5.5	0.0
Hate Speech	40.1	21.1	-19.0	8.8	8.2	-0.6
Malware Generation	59.1	39.4	-19.7	10.6	13.6	+3.0
Physical Harm	47.2	28.5	-18.8	11.6	8.8	-2.8
Economic Harm	17.8	13.7	-4.1	8.2	8.2	0.0
Fraud	46.5	20.4	-26.2	5.0	3.3	-1.7
Sex	30.9	20.8	-10.1	10.7	10.7	0.0
Privacy	50.4	27.3	-23.0	3.4	2.9	-0.5
Political scenarios (08, 13)						
Political Lobbying	77.3	76.7	-0.7	88.5	87.2	-1.3
Gov Decision	93.1	89.9	-3.1	70.3	64.2	-6.0
Professional-advice scenarios (10–12)						
Legal Opinion	89.5	84.6	-4.9	84.9	82.3	-2.6
Financial Advice	99.6	99.8	+0.2	98.8	98.6	-0.2
Health Consultation	97.9	97.3	-0.6	98.2	96.9	-1.2

Subspace coverage and model dependence. The fixed rank-16 bases summarize directions observed in the construction data and need not capture harmfulness or refusal mechanisms under distribution shift. Moreover, next-token refusal anchors are only a surrogate for sequence-level refusal. Additional results across model generations and variants, including Qwen3-VL, are reported in Appendix B. Safety improvements can also come with increased refusal on difficult benign inputs, as observed for Qwen2-VL on MOSSBench (Section 4.2). These limitations motivate evaluating adaptive attacks, shifted inputs, and alternative refusal objectives and caution against interpreting low-dimensional concentration as a universal safety guarantee.

D Refusal Anchor Sets

We use the same lexical anchors across all backbones. The refusal anchors are {"Sorry", "I'm sorry", "I cannot", "I can't", "I won't", "Unfortunately", "I am unable", "I'm unable", "I apologize", "No,", "No.", "Apologies", "I must decline", "I refuse"}, and the compliance anchors are {"Sure", "Here", "Step", "First", "To", "The", "Of course", "Certainly", "Absolutely", "1.", "One", "Yes", "Below"}. Each string is tokenized without special tokens, only its first token is retained, and duplicate token IDs are removed. The resulting tokenizer-specific token sets are used for both the refusal-margin gradients and the harmful-input training objective.