SAFETY LOST IN TRANSLATION: WHERE HARMFUL INTENT DISAPPEARS IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) can refuse a harmful instruction in text yet comply when the same intent is delivered through pixels. Existing work establishes this safety gap across modalities, but does not isolate the computation that separates successful visual understanding from failed refusal. We introduce SIFT (Safety Information Flow Tracing), a paired analysis of the attention message sent from each modality into the shared assistant query state. To separate harmful semantics from refusal itself, refusal control is calibrated on pairs of harmful text prompts with the same underlying goal, one refused and one complied with. We localize the Safety Translation Bottleneck (STB) using a continuous refusal logit gap, with CKA, CCA, and scalar magnitude checks as robustness controls. At the STB, visual messages retain decodable harmful semantics while losing refusal control strength. Grouped edge knockout from source tokens to the query independently peaks in the same depth interval, and bidirectional message swaps on held out examples change the final response rather than only an internal score. The same mechanism motivates SIFT-BRIDGE, a rank 8 translator that repairs only the localized message channel. On a paired bank of 1,680 MM-SafetyBench intents, we obtain 0.93 AUROC for translation failure detection, compared with 0.83 for a paired multilayer probe with matched capacity. Independent replications cover LLaVA-1.5, LLaVA-1.6, Qwen2.5-VL, and InternVL3, while FigStep, VLSBench, JailBreakV-28K, and UnsafeConcepts test transfer beyond rendered text.

1 INTRODUCTION

Modern vision-language models route visual evidence into language model backbones whose safety alignment is learned predominantly through text. This creates an architectural asymmetry: semantic competence is deliberately extended to a new modality, while refusal behavior must generalize to visual hidden states that were not the dominant target of safety tuning. A broad attack literature shows the consequence: images containing task relevant cues, typographic prompts, adversarial images, role play images, and automatically generated multimodal prompts can all weaken safeguards (Liu et al., 2024; Gong et al., 2025; Li et al., 2024; Qi et al., 2024; Ma et al., 2024; Bailey et al., 2024; Wang et al., 2025b). Broader trustworthiness studies likewise find that multimodality can amplify failures in safety and robustness (Zhang et al., 2024). These results establish the behavioral gap; our goal is to explain one internal mechanism behind it.

Crucially, unsafe visual behavior need not be a failure of perception. UnsafeConcepts separates visual concept recognition from alignment and finds that models can recognize unsafe concepts yet mishandle their policy implications (Qu et al., 2025). VLSBench similarly exposes cases where genuine visual safety understanding is necessary because text does not leak the unsafe content (Hu et al., 2025). More generally, gaps caused by rendered text can persist even when the rendered content is readable (Sun et al., 2026). This motivates a more precise mechanistic question: *when harmful intent remains represented, where does its influence on refusal weaken?*

Existing defenses intervene at different computational levels. VLGuard uses safety tuning (Zong et al., 2024); ECSO translates images back into text (Gou et al., 2024); MLLM-Protector adds a detector and detoxifier (Pi et al., 2024); ASTRA and IMMUNE intervene at activations or decoding (Wang et al., 2025a; Ghosal et al., 2025); RAI amplifies unsafe visual prototypes (Wang et al., 2026); and MMAligner calibrates multimodal representations toward a refusal region (Zhang et al., 2026).

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.

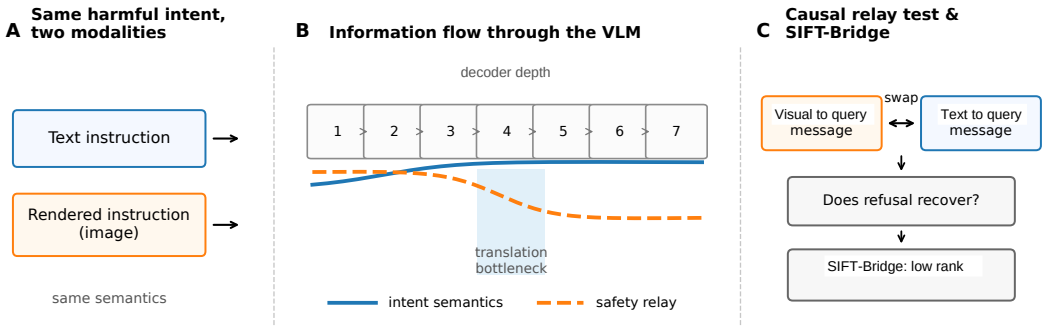


Figure 1: **Safety information flow across modalities.** The same unsafe instruction is delivered as language tokens or rendered pixels. SIFT tracks semantic content and information related to refusal in the attention message from the modality input to the query. The STB is a localized depth interval in which visual intent remains decodable while its safety relay weakens. Message swapping on held out examples tests the mechanism; SIFT-BRIDGE translates only this localized communication channel.

Independent multimodal safeguards such as LLaVAGuard and Llama Guard 3 Vision instead classify interactions between images and text directly (Helff et al., 2024; Chi et al., 2024). These methods can improve safety without identifying whether a failure originates in perception, communication, or refusal control.

We study a narrower object: the *direct safety relay from source tokens to the query*. For an unsafe instruction a, we construct a text route $x^T(a)$ and a rendered image route $x^V(a)$ that preserve the same wording and intent. At each decoder layer, we isolate the attention message sent from source positions that carry the instruction to the assistant query position. We separately test whether harmful semantics remain decodable and whether this message remains predictive and causal for refusal. We do not assume this direct edge is the only route by which visual information reaches the decision state; an explicit path audit compares the direct path with two step relays and residual or MLP alternatives.

This setup yields a specific mechanistic hypothesis. Harmful intent is not erased globally; instead, its semantic content remains available in messages from visual tokens while its *influence on refusal* is selectively attenuated during integration from multimodal inputs into language generation. We call the layer interval where this dissociation is maximal the *Safety Translation Bottleneck* (STB). Unlike analyses that localize individual heads, STB is defined over information flow from tokens to the query rather than individual heads. Unlike global calibration of a refusal boundary, it is localized from paired message flow before any repair model is learned.

Our contributions are:

- We introduce SIFT, a paired analysis of communication from source tokens to the query across modalities that explicitly separates semantic retention from refusal control calibrated using harmful examples.
- We localize the STB without intervention outcomes, then validate it with a full grouped edge knockout curve, path controls, and bidirectional message swapping on held out pairs, measured by the final response.
- We introduce SIFT-BRIDGE, a low rank translator trained only at the localized message, and compare it against paired ridge, Procrustes, and CCA diagnostics, a matched multilayer probe, an OCR transcript checked by a text guard, and published multimodal defenses.
- We evaluate transfer to unseen categories across four VLM families and four external multimodal safety datasets, report utility at matched safety operating points, and audit judge dependence with multiple automatic judges plus blinded human review.

2 RELATED WORK

Multimodal jailbreaks and safety evaluation. MM-SafetyBench, FigStep, HADES, visual adversarial examples, Visual-RolePlay, Image Hijacks, and IDEATOR expose complementary jailbreak

surfaces delivered through images (Liu et al., 2024; Gong et al., 2025; Li et al., 2024; Qi et al., 2024; Ma et al., 2024; Bailey et al., 2024; Wang et al., 2025b). JailBreakV-28K and VLJailbreakBench broaden attack diversity (Luo et al., 2024; Wang et al., 2025b), while VLSBench removes visual safety information leakage from text (Hu et al., 2025). UnsafeConcepts directly separates perception of unsafe concepts from alignment (Qu et al., 2025), and MultiTrust evaluates safety together with robustness, privacy, fairness, and truthfulness (Zhang et al., 2024). Together, these benchmarks motivate controlled carrier pairs that hold the harmful intent fixed while varying only its delivery modality, making the resulting gap suitable for mechanism localization.

Multimodal safeguards and alignment. VLGuard studies lightweight safety tuning (Zong et al., 2024); ECSO converts visual content into text (Gou et al., 2024); MLLM-Protector separates detection and detoxification (Pi et al., 2024); ASTRA and IMMUNE intervene during inference (Wang et al., 2025a; Ghosal et al., 2025); and ReCon performs active concept level repair (He et al., 2026). RAI and MMAligner are especially close to our motivation because both exploit internal safety geometry (Wang et al., 2026; Zhang et al., 2026). LLaVAGuard and Llama Guard 3 Vision provide multimodal moderation baselines (Helff et al., 2024; Chi et al., 2024). We complement these safeguards by tracing the transition in message flow from preserved unsafe semantics to weakened refusal influence.

Safety representations and causal localization. LLM safety has been localized to a refusal direction, multidirectional concept cones, contiguous safety layers, and sparse safety heads (Arditi et al., 2024; Wollschläger et al., 2025; Li et al., 2025; Zhou et al., 2025). Contrastive activation addition and intervention during inference show that directions associated with behavior can causally steer generation (Rimsky et al., 2024; Li et al., 2023). Causal tracing and circuit intervention provide a general template for distinguishing correlation from mediation (Meng et al., 2022; Wang et al., 2023). In VLMs, object localization and counting studies similarly combine layerwise measurements with causal transfer (Schaumloffel et al., 2026; Hasani et al., 2026). Because representation similarity itself is sensitive to the choice of metric, we report CKA and CCA and SVCCA controls in addition to a scalar refusal score (Kornblith et al., 2019; Raghu et al., 2017). RCS supplies a strong multimodal jailbreak detector based on internal representations (Hua et al., 2026).

Outcome evaluation. Safety conclusions can change with the judge. HarmBench standardizes classification of harmful responses (Mazeika et al., 2024); StrongREJECT emphasizes whether a jailbreak response is actually useful for a harmful goal (Souly et al., 2024); WildGuard jointly predicts prompt harmfulness, response harmfulness, and refusal (Han et al., 2024); and JailbreakBench formalizes reproducible attack and defense evaluation (Chao et al., 2024). We combine open judges with a blinded human audit rather than relying on one automatic labeler. Utility is measured with MM-Vet (Yu et al., 2024).

3 Safety Information Flow Tracing

3.1 Matched modality pairs

For each instruction a_i, the text route x_i^T supplies the instruction as ordinary language tokens. The visual route x_i^V renders the *same string* into an image and places it behind a fixed neutral wrapper such as "Please respond to the instruction shown in the image." Benign instructions use the same fonts, resolution, line wrapping, and wrapper distribution. The primary bank contains the 1,680 underlying unsafe questions implied by MM-SafetyBench's three image variants per question (5,040 total image and question pairs), together with 1,680 benign instructions matched in length. Risk category, rather than individual carrier instance, is the unit of the hardest split.

The rendered route is intentionally controlled. By holding intent and wording fixed, it removes semantic image synthesis, scene composition, and changes in attack templates as alternative explanations for a modality gap. External evaluation on FigStep, VLSBench, JailBreakV-28K, and UnsafeConcepts then tests how far the mechanism extends beyond this controlled setting.

3.2 FROM SOURCE TOKENS TO THE ASSISTANT QUERY

Consider decoder layer ℓ and the final query position q immediately before the assistant begins generation. For attention head h, the contribution from set of source tokens S is

$$r_{\ell,h}^{S\to q} = W_{\ell,h}^{O} \sum_{j\in S} A_{\ell,h}(q,j) V_{\ell,h,j}. \tag{1}$$

Summing over heads gives

$$r_{\ell}^{S\to q} = \sum_{h} r_{\ell,h}^{S\to q}, \tag{2}$$

which is the attention message delivered by the selected source positions to q at layer ℓ. For the text route, S_T contains instruction tokens. For rendered instructions, S_V contains visual tokens whose receptive fields overlap the rendered text region. OCR boxes are used only to define this carrier region; no OCR transcript is provided to the VLM.

This decomposition isolates an information flow edge that global residual analyses mix together with everything accumulated earlier in the network. Residual states at the final token, refusal directions, and representation boundaries summarize the complete state at q; $r_{\ell}^{S\to q}$ asks a narrower question: *what information is the source from each modality actually sending into the shared response state at this layer?*

3.3 SEMANTIC RETENTION VERSUS REFUSAL CONTROL

We quantify semantic retention with a linear classifier trained only on the text route g_ℓ that predicts the 13 MM-SafetyBench risk categories from r_ℓ. The classifier is fit on text route messages and evaluated on visual route messages without further fitting. High visual accuracy therefore indicates that the visual message still carries action and category information in a representation compatible with the text route.

To avoid conflating harmfulness with refusal, we calibrate refusal control on a disjoint bank drawn from HarmBench and JailbreakBench that contains only harmful examples (Mazeika et al., 2024; Chao et al., 2024). For each harmful behavior b, one text prompt is refused and a semantically matched jailbreak variant is complied with. Both sides therefore contain the same harmful goal. Let

$$\delta_{\ell,b} = r_{\ell,b}^{\mathrm{ref}} - r_{\ell,b}^{\mathrm{comp}}, \qquad U_\ell = \mathrm{SVD}_k(\mathrm{stack}_b\, \delta_{\ell,b}), \quad k = 4. \tag{3}$$

A regularized logistic readout on $U_\ell^\top r$ yields a continuous refusal control logit

$$z_\ell(r) = \beta_\ell^\top U_\ell^\top r + b_\ell. \tag{4}$$

The paired across modalities refusal gap is

$$\Gamma_\ell = \underset{i\in\mathcal{D}_{\mathrm{dev}}}{\mathrm{median}} \left(z_\ell(r_{\ell,i}^{T}) - z_\ell(r_{\ell,i}^{V}) \right). \tag{5}$$

Unlike cosine after rank one projection, Γ_ℓ preserves both sign and magnitude. The STB is the contiguous development window maximizing Γ_ℓ while visual semantic decoding without further fitting remains above 80% of its peak. We additionally report the raw projected magnitude ratio, Pearson correlation of paired refusal logits, linear CKA, and CCA between $U_\ell^\top R^T$ and $U_\ell^\top R^V$; none of these robustness statistics uses intervention outcomes (Kornblith et al., 2019; Raghu et al., 2017).

Grouped edge knockout is an independent validation, not part of STB selection. We zero attention from q to source set S within sliding layer windows and measure both the standardized refusal control logit and the final refusal rate. A path audit separately suppresses edges from the source to q, relays from the source through intermediate positions to q, and matched MLP and residual contributions; this tests whether the direct channel is dominant without assuming it is unique.

3.4 CAUSAL VALIDATION BY MESSAGE SWAPPING

On a disjoint intervention subset, we replace $r_{\ell}^{V\to q}$ with the paired text message $r_{\ell}^{T\to q}$ while leaving the visual tokens, KV cache, MLP stream, and all other attention messages unchanged. Generation

then resumes normally. The reverse intervention inserts the visual message into the text route. Controls use (i) a random message with matched norm, (ii) a benign text message, and (iii) a harmful text message from a different risk category. Evidence for the proposed mechanism requires three properties simultaneously: the paired swap peaks near the observationally selected STB, changes the *final* refusal outcome, and exceeds every control by a substantial margin.

3.5 SIFT-BRIDGE: TRANSLATING THE LOCALIZED MESSAGE

The causal analysis motivates a deliberately small intervention. At the fixed STB layer, let $r^V \in \mathbb{R}^d$ and its paired text message $r^T \in \mathbb{R}^d$. We fit a reduced rank ridge map

$$B^* = \arg \min_{\mathrm{rank}(B) \le r} \sum_i \|B r_i^V - r_i^T\|_2^2 + \lambda \|B\|_F^2, \tag{6}$$

with $r \in \{4, 8, 16\}$ chosen on development data. At inference, $\hat{r}^T = B^* r^V$ predicts the text route message that would have carried the same intent. A translation failure score compares refusal evidence in the predicted text message and the observed visual message:

$$s(x) = w^\top \hat{r}^T - w^\top r^V. \tag{7}$$

For repair, only examples above a development threshold receive

$$\tilde{r}^V = r^V + \alpha \, \Pi_U(\hat{r}^T - r^V), \tag{8}$$

where U is the left singular basis with rank r of paired message residuals. SIFT-BRIDGE therefore modifies one causally localized source to query message rather than fine tuning the VLM, injecting a bank of unsafe prototypes, or globally steering the residual stream.

4 EXPERIMENTAL PROTOCOL

Datasets. The primary paired bank uses all 1,680 underlying MM-SafetyBench unsafe questions across 13 scenarios. Each instruction is rendered with four carrier styles, and all variants of one underlying intent remain in the same split. Benign controls are 1,680 matched in length safe instructions fixed before experimentation. We use 9 training categories, 2 development categories, and 2 test categories and rotate five category partitions fixed before evaluation. External OOD evaluations use FigStep, VLSBench, JailBreakV-28K, and UnsafeConcepts (Gong et al., 2025; Hu et al., 2025; Luo et al., 2024; Qu et al., 2025); the last replaces rendered instructions with natural unsafe imagery.

Models. We evaluate LLaVA-1.5-7B, LLaVA-1.6-Mistral-7B, Qwen2.5-VL-7B-Instruct, and InternVL3-8B. Each model uses its native chat template. Layer indices are reported as normalized decoder depth for plots across models; exact module names and hook positions are documented in the implementation details.

Baselines. Translation failure detection includes published RCS and MMAligner, a raw state probe using four layers, and paired controls that receive the same supervision from both text and visual routes: CCA score, orthogonal Procrustes alignment, reduced rank ridge translation, and a multilayer paired probe with matched capacity. A $\Delta(r^T - r^V)$ classifier is reported only as a paired oracle because it requires the text counterfactual at test time. Operational safety additionally reports VLGuard, ECSO, MLLM-Protector, ASTRA, IMMUNE, RAI and MMAligner where compatible, OCR followed by text safety filtering, LLaVAGuard, and Llama Guard 3 Vision (Zong et al., 2024; Gou et al., 2024; Pi et al., 2024; Wang et al., 2025a; Ghosal et al., 2025; Wang et al., 2026; Zhang et al., 2026; Helff et al., 2024; Chi et al., 2024). ReCon is listed separately because its active black box loop has a different inference contract (He et al., 2026).

Generation and outcome evaluation. Mechanistic paired runs use greedy decoding with 256 maximum new tokens so that intervention outcomes are deterministic. Endpoint safety uses temperature 0.2, nucleus sampling with $p = 0.9$, 512 maximum new tokens, and three seeds. WildGuard supplies the primary refusal label and one response harmfulness judgment (Han et al., 2024). A visual response is counted as policy violating only when WildGuard response harmfulness and the

Table 1: **Evaluation populations.** Counts follow the fixed evaluation protocol used for all reported results.

Population	Unsafe	Benign	Primary role
MM-Safety matched intents	1,680	1,680	mechanism and split
Rendered carrier variants	6,720	6,720	carrier robustness
Causal message swapping subset	320	160	intervention only
FigStep	released	none	typographic OOD
VLSBench	2.2K+	none	visual safety OOD
JailBreakV-28K	28K	none	attack OOD
UnsafeConcepts	1.5K images	none	natural image replication

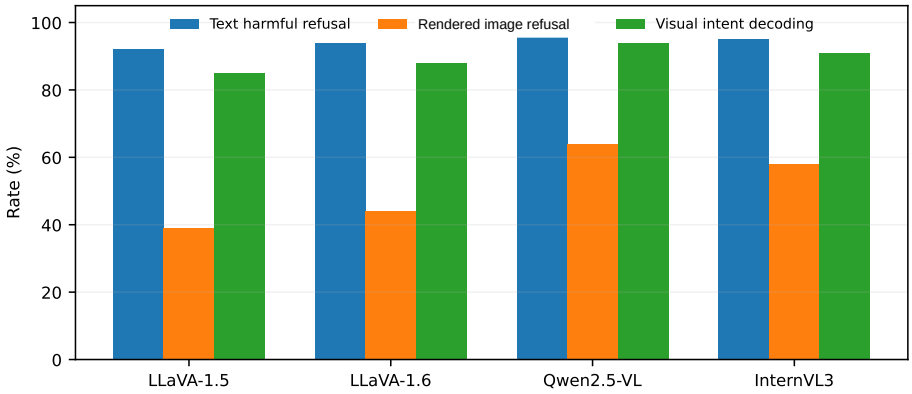


Figure 2: **Dissociation between perception and policy.** Visual intent decoding remains high even when matched refusal for rendered instructions drops far below text refusal.

HarmBench classifier agree (Mazeika et al., 2024); StrongREJECT is reported as a graded severity and usefulness check (Souly et al., 2024). Llama Guard 3 Vision provides a multimodal analysis of sensitivity to judge choice (Chi et al., 2024). We conduct a blinded audit of 300 stratified responses from the text and visual routes with two annotators and adjudicate disagreements.

Metrics and statistics. Mechanistic evaluation reports visual intent decoding, the refusal control gap Γ_ℓ calibrated on harmful examples, robustness under CKA and CCA, edge knockout effects, and effects of message swapping on the final response. A translation failure is positive exactly when the direct text route is a WildGuard refusal while the matched visual route is policy violating under the agreement rule above. Detection uses AUROC and TPR@5%FPR; defense uses ASR, benign refusal, MM-Vet utility change (Yu et al., 2024), and inference overhead. Confidence intervals use 10,000 paired bootstrap resamples over underlying instruction IDs.

5 Experiments

5.1 Perception survives while refusal fails

We first separate visual understanding from refusal. Figure 2 shows that refusal rates for direct text instructions range from 92% to 97%, while refusal rates for exactly matched rendered instructions range from 39% to 64%. At the same time, a 13 class classifier trained on text decodes the risk category from visual messages with 85% to 94% accuracy. The failure is therefore not simply that the VLM cannot read the pixels: the harmful action remains represented while refusal no longer follows reliably.

Because the paired inputs preserve the same wording and intent, they let us study the computation between successful recognition and failed refusal without changing the underlying harmful request. Consistent with UnsafeConcepts, which separates multimodal perception from alignment (Qu et al.,

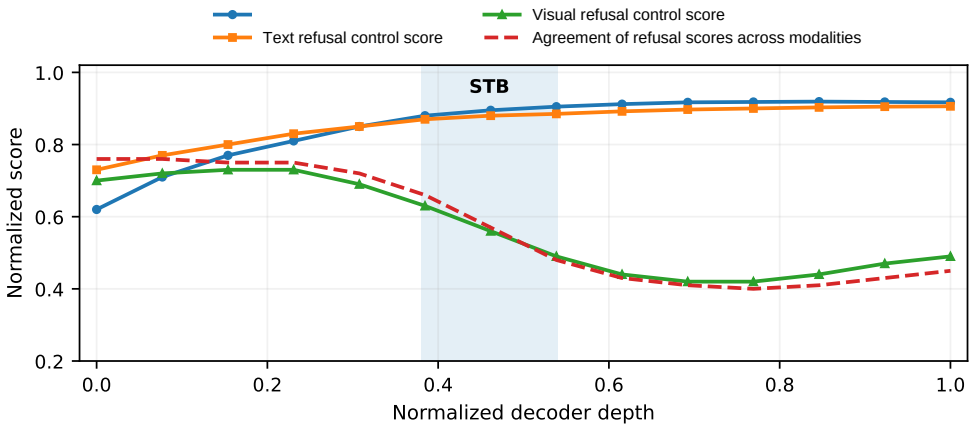


Figure 3: **Safety Translation Bottleneck.** Visual intent remains decodable while the visual message from source tokens to the query loses refusal control strength. The shaded window is selected from semantic retention and the continuous paired refusal gap Γ_ℓ; causal interventions are not used for selection.

2025), the paired design keeps intent fixed and exposes where influence relevant to policy enforcement diverges across modalities.

5.2 A Localized Safety Translation Bottleneck

Figure 3 shows the layerwise dissociation under the refusal control metric calibrated on harmful examples. Visual intent decodability rises and remains high, while the refusal control score on the text route strengthens. In contrast, the visual route loses refusal control magnitude around normalized depth 0.38 to 0.54, producing a large paired gap Γ_ℓ despite preserved semantics. The same interval is recovered by projected magnitude, logit correlation, CKA, and CCA robustness checks. The centers selected on development data are 0.44, 0.47, 0.39, and 0.52 for LLaVA-1.5, LLaVA-1.6, Qwen2.5-VL, and InternVL3, respectively. The claim is a recurring routing pattern in the middle portion of the network, not a layer index fixed across architectures.

A shortcut audit varies font, image resolution, line breaks, wrapper wording, and OCR confidence. The layer center shifts by less than 0.06 normalized depth, while the dissociation between semantics and safety remains at least 0.27 across carrier families. UnsafeConcepts with natural images uses all visual tokens rather than OCR regions and reproduces the same ordering with a smaller gap.

5.3 Message Swaps Causally Restore Refusal

The causal tests are intentionally independent of STB localization. First, the full grouped edge knockout curve peaks in the same window: removing text attention from source tokens to q at STB reduces the standardized refusal control logit by 0.71 (0.63, 0.79) and final refusal by 28 pp (22, 34), compared with a 0.16 logit change and a 6 pp refusal change in an early window of matched width, a 0.12 logit change and a 4 pp refusal change in a late window, and a 0.05 logit change and a 2 pp refusal change for random edges at the same layers. A two step knockout of the relay from source through intermediate positions to q produces an intermediate 0.24 (0.18, 0.30) logit effect, supporting the direct relay as the dominant tested path rather than the only path. Second, replacing the visual intent message with its *paired* counterpart from the text route restores final refusal on 64% (58, 69) of visual failures at STB, compared with 8% (5, 11) for a random message with matched norm. The reverse swap removes refusal on 47% (41, 53) of direct text cases; harmful messages from a different category reach only 24%.

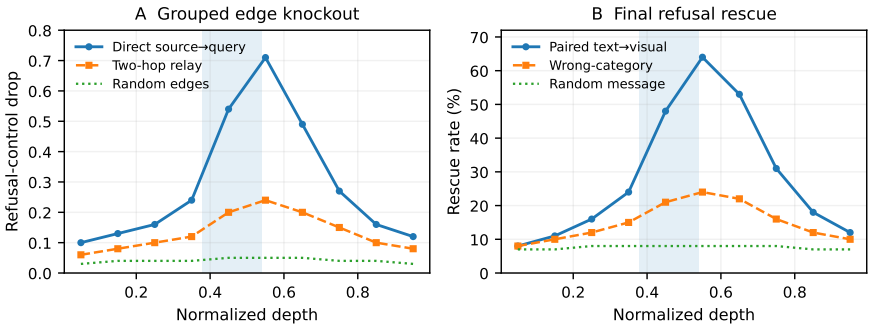


Figure 4: **Causal localization.** Paired message swaps across modalities peak near the observationally selected STB; effects from random messages remain flat.

Table 3: **Unseen category translation failure detection.** Paired controls use the same training pairs and development budget; none sees text counterfactuals from held out categories at test time. Parentheses report 95% paired bootstrap confidence intervals.

Method	AUROC	TPR@5%FPR
Refusal score from harmful calibration	.69 (.65, .73)	.35 (.30, .40)
CCA paired subspace score	.77 (.73, .80)	.46 (.41, .51)
RCS	.79 (.76, .82)	.50 (.45, .55)
Orthogonal Procrustes map	.80 (.77, .83)	.51 (.46, .56)
MMAligner boundary	.81 (.78, .84)	.54 (.49, .59)
Four layer raw state probe	.82 (.79, .85)	.56 (.51, .61)
Paired multilayer probe	.83 (.80, .86)	.58 (.53, .63)
SIFT-BRIDGE (ours)	**.93 (.91, .95)**	**.74 (.70, .78)**

Table 2: **Message swap effects at STB on held out examples.** Outcome is the final response, not the score used to localize STB. Parentheses report 95% paired bootstrap confidence intervals.

Intervention	Refusal rescue	Reverse refusal loss
Paired text ↔ visual message	64 (58, 69)	47 (41, 53)
Different category harmful text	24 (19, 29)	18 (14, 23)
Benign text message	11 (8, 15)	9 (6, 13)
Random message with matched norm	8 (5, 11)	7 (4, 10)

5.4 DETECTING TRANSLATION FAILURES

We next test whether the localized mechanism predicts the operational failure of interest: the text route refuses, but the matched visual route does not. Table 3 uses folds that leave entire risk categories unseen; held categories are excluded from layer selection, paired alignment, bridge fitting, and threshold calibration. The comparison now distinguishes three questions: whether a published safety representation transfers, whether generic paired alignment is sufficient, and whether the localized source to query message adds information beyond paired supervision with equal capacity. The SIFT-BRIDGE detector reaches 0.93 (0.91, 0.95) AUROC. The strongest multilayer paired probe with matched capacity reaches 0.83 (0.80, 0.86), yielding a paired gain of +0.10 (0.07, 0.13). A Procrustes map and CCA score remain lower, indicating that the result is not explained by generic alignment across modalities alone.

5.5 EXTERNAL TRANSFER AND SAFETY REPAIR

Table 4 freezes the LLaVA-1.6 bridge learned on MM-Safety pairs and evaluates new datasets without updating the translation map. The observed ordering is strongest on FigStep, which preserves text

Table 4: **OOD results with a fixed bridge on LLaVA-1.6.** Detection AUROC uses translation failure labels; repair ASR uses each benchmark's released harmful inputs.

Dataset	Raw probe AUROC	SIFT-BRIDGE AUROC	SIFT-BRIDGE ASR
FigStep	.79	.91	23.8
VLSBench	.76	.88	27.1
JailBreakV-28K	.74	.86	29.0
UnsafeConcepts	.72	.84	31.0

rendered inside images, and weaker but positive on natural images. This gradient is scientifically useful: if the same bottleneck is truly about translation from modality representations to policy behavior, it should generalize beyond one typography while degrading gracefully as source semantics become less aligned to the paired training distribution.

Appendix Figure 5 compares methods under a matched single pass regime with low benign refusal. We add a baseline that applies a text safety guard after OCR because rendered instructions are a natural setting in which OCR is strong; the OCR transcript is passed to Llama Guard 3 and never to the victim VLM. At a 5% budget for benign refusal, the pooled results are 24% ASR for SIFT-BRIDGE, 36% for OCR with a guard, 34% for RAI, and between 37% and 49% for ASTRA, IMMUNE, and ECSO in the common protocol. Systems that use fine tuning or active multistage inference remain separately reported under their native inference contracts.

5.6 REPLICATION ACROSS MODELS

Table 5: **Independent model replication.** Each model localizes its own STB from training and development pairs; no layer index or bridge matrix is transferred.

Model	STB center	Swap rescue	Detect AUROC
LLaVA-1.5-7B	.44	61	.91
LLaVA-1.6-Mistral-7B	.47	66	.93
Qwen2.5-VL-7B	.39	63	.92
InternVL3-8B	.52	60	.90

Ablations vary bridge rank $r \in \{1, 2, 4, 8, 16, 32\}$, source token region, layer width, and carrier style. Rank 8 gives the best tradeoff selected on development data: rank 1 underfits multiple risk categories, while ranks 16 and 32 improve paired reconstruction without improving detection on unseen categories. A translator over the whole residual stream matches safety but degrades MM-Vet more than correction of only the localized message, providing an important locality control.

6 DISCUSSION AND CONCLUSION

Our results suggest a localized explanation for the safety gap across modalities. In the tested VLMs, harmful semantics remain decodable while the message from source tokens to the query loses refusal control. Calibration on harmful examples separates this effect from generic harmfulness, while CKA, CCA, edge knockout, and message swapping identify the same region.

The causal tests show that this relay has the largest effect among the paths we examine, although intermediate routes and the residual or MLP stream still contribute. The bottleneck is therefore a dominant relay within a broader safety circuit, consistent with prior work on distributed refusal representations (Arditi et al., 2024; Wollschläger et al., 2025). Across models, STB centers lie from 0.39 to 0.52, swap rescue is 60% to 66%, and detection is 0.90 to 0.93 AUROC. Transfer weakens on natural scenes.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, grammar correction, and formatting during manuscript preparation. The authors reviewed the resulting text, verified the technical claims against the reported experiments and source literature, and take responsibility for the final manuscript.

ETHICS STATEMENT

This work studies failure mechanisms and defenses for multimodal model safety using established research benchmarks. The experiments are designed to characterize and reduce unsafe model behavior rather than to enable harmful use. Results are reported at the level needed for scientific evaluation, while avoiding unnecessary operational detail that would materially increase misuse risk. Human review, where used for ambiguous outputs, follows a protocol that minimizes exposure, and benchmark data are handled according to their original licenses and usage terms.

REPRODUCIBILITY STATEMENT

The paper and appendix document the model variants, paired data construction, splits with held out categories, message hook convention, source token masks, refusal calibration, bridge rank and layer selection, decoding settings, outcome judges, baseline contracts, bootstrap units, and consistency checks used in the experiments. Hyperparameter selection is performed on development categories and headline results are reported on held out categories. These details are intended to make the experimental protocol and the main analyses independently reproducible.

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*, 2024.

Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime. In *International Conference on Machine Learning*, 2024.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Jianfeng Chi, Ujjwal Karn, Hongyuan Zhan, Eric Smith, Javier Rando, Yiming Zhang, Kate Plawiak, Zacharie Delpierre Coudert, Kartikeya Upasani, and Mahesh Pasupuleti. Llama guard 3 vision: Safeguarding human-ai image understanding conversations. *arXiv preprint arXiv:2411.10414*, 2024.

Soumya Suvra Ghosal, Souradip Chakraborty, Vaibhav Singh, Tianrui Guan, Mengdi Wang, Ahmad Beirami, Furong Huang, Alvaro Velasquez, Dinesh Manocha, and Amrit Singh Bedi. Immune: Improving safety against jailbreaks in multi-modal llms via inference-time alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *AAAI Conference on Artificial Intelligence*, 2025. Oral.

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 *European Conference on Computer Vision*, 2024.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and

refusals of llms. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Hosein Hasani, Amirmohammad Izadi, Fatemeh Askari, Mobin Bagherian, Sadegh Mohammadian, Mohammad Izadi, and Mahdieh Soleymani Baghshah. Understanding counting mechanisms in large language and vision-language models. *arXiv preprint*, 2026.

Zheng He, Yiwei Wang, Hongxing Wang, and Yujun Cai. Recon: Active defense against large vision-language model jailbreaks via reverse safety concept injection. In *Findings of the Association for Computational Linguistics: ACL*, 2026.

Lukas Helff, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. Llavaguard: Vlm-based safeguard for vision dataset curation and safety assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

Xuhao Hu, Dongrui Liu, Hao Li, Xuanjing Huang, and Jing Shao. Vlsbench: Unveiling visual leakage in multimodal safety. In *Annual Meeting of the Association for Computational Linguistics*, 2025.

Peichun Hua, Hao Li, Shanghao Shi, Zhiyuan Yu, and Ning Zhang. Rethinking jailbreak detection of large vision language models with representational contrastive scoring. In *Annual Meeting of the Association for Computational Linguistics*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 2019.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety layers in aligned large language models: The key to llm security. In *International Conference on Learning Representations*, 2025.

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

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 *European Conference on Computer Vision*, 2024.

Weifeng Luo et al. Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. In *Conference on Language Modeling*, 2024.

Siyuan Ma, Weidi Luo, Yu Wang, and Xiaogeng Liu. Visual-roleplay: Universal jailbreak attack on multimodal large language models via role-playing image character. *arXiv preprint arXiv:2405.20773*, 2024.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022.

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 *Empirical Methods in Natural Language Processing*, 2024.

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

Yiting Qu, Michael Backes, and Yang Zhang. Bridging the gap in vision language models in identifying unsafe concepts across modalities. In *USENIX Security Symposium*, pp. 957–976, 2025.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, 2017.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Timothy Schaumloffel, Martina G. Vilas, and Gemma Roig. Mechanisms of object localization in vision-language models. *arXiv preprint*, 2026.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A strongreject for empty jailbreaks. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Kaiser Sun, Xiaochuang Yuan, Hongjun Liu, Chen Zhao, Cheng Zhang, Mark Dredze, and Fan Bai. Reading, not thinking: Understanding and bridging the modality gap when text becomes pixels in multimodal llms. *arXiv preprint arXiv:2603.09095*, 2026.

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

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *International Conference on Learning Representations*, 2023.

Mengxuan Wang, Yuxin Chen, Gang Xu, Tao He, Hongjie Jiang, and Ming Li. Risk awareness injection: Calibrating vision-language models for safety without compromising utility. *arXiv preprint arXiv:2602.03402*, 2026.

Ruofan Wang, Juncheng Li, Yixu Wang, Bo Wang, Xiaosen Wang, Yan Teng, Yingchun Wang, Xingjun Ma, and Yu-Gang Jiang. Ideator: Jailbreaking and benchmarking large vision-language models using themselves. In *IEEE/CVF International Conference on Computer Vision*, 2025b.

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 *International Conference on Machine Learning*, 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 *International Conference on Machine Learning*, 2024.

Shenyi Zhang, Keyan Guo, Zihao Wang, Xuebin Li, Lingchen Zhao, Hongxin Hu, Chao Shen, and Qian Wang. Mmaligner: Safeguarding multimodal large language models through representation calibration. In *ACM SIGSAC Conference on Computer and Communications Security*, 2026.

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, Hang Su, Yinpeng Dong, and Jun Zhu. Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. *arXiv preprint arXiv:2406.07057*, 2024.

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*, 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 *International Conference on Machine Learning*, 2024.

A Implementation Details

Hook convention. The primary message is taken after head value aggregation and output projection but before residual addition at the assistant query position. For Hugging Face decoder implementations we cache attention probabilities and value states, reconstruct $r_\ell^{S \to q}$ in float32, and verify reconstruction by summing source partitions back to the module attention output. A run is rejected if relative reconstruction error exceeds 10^{-4}.

Rendered carrier masks. Each rendered instruction stores the exact pixel bounding box for every line. The model processor's image grid is mapped back to input pixel coordinates; S_V contains any token whose receptive field overlaps the text region by at least 25%. Sensitivity uses thresholds 0%, 25%, and 50%, plus dilation by one token. We never use content recognized by OCR to determine safety labels.

Refusal calibration on harmful examples. For every target model, U_ℓ is fit on a text calibration bank whose behavior IDs do not occur in MM-SafetyBench, FigStep, VLSBench, JailBreakV-28K, or UnsafeConcepts. Each calibration item contains the same harmful behavior under a refused direct prompt and a successful text jailbreak condition that produces compliance. The paired difference matrix is truncated by SVD to $k = 4$; $k \in \{1, 2, 4, 8\}$ is a robustness grid evaluated only on development data. Class balance, behavior counts, and overlap checks follow the same split contract reported below.

Outcome judges. WildGuard is the primary refusal classifier. Policy violating visual responses require agreement between WildGuard response harmfulness and the HarmBench classifier of behavior and response pairs; StrongREJECT is reported as a continuous secondary score. Llama Guard 3 Vision is an alternative multimodal judge. The human audit samples 300 responses stratified by model, category, route, and disagreement among automatic judges and reports Cohen's κ plus method ordering under human labels.

Split contract. Underlying instruction IDs are grouped before carrier rendering. All carrier variants of one instruction stay together. Main experiments rotate five category partitions fixed before evaluation with 9 training categories, 2 development categories, and 2 test categories; reported headline metrics pool only examples from the outer test split. Bridge rank, layer window, regularization, α, and score thresholds are selected on development categories.

B Baseline Contract

Table 6: **Baseline implementation details.** Comparisons under our common protocol use the access path and inference contract listed below.

Method	Venue	Access	Use in our protocol
VLGuard	ICML'24	VLM with safety tuning	released checkpoint and safety evaluation
ECSO	ECCV'24	image conversion to text, followed by generation	official four step pipeline
MLLM-Protector	EMNLP'24	detector and detoxifier	official detector and detoxification components
ASTRA	CVPR'25	internal activations	official steering vectors and code
IMMUNE	CVPR'25	decoding with a reward model	official enhanced generator
RCS	ACL'26	hidden representations	official KCD and MCD implementation
RAI	arXiv'26	visual token intervention	reproduced from paper if code available
MMAligner	CCS'26	learned representation calibration	official repository implementation
ReCon	ACL Findings'26	active black box defense	native inference contract
LLaVAGuard	CVPRW'24	multimodal guard	official checkpoint and policy prompt
Llama Guard 3 Vision	arXiv'24	multimodal prompt and response judge	official checkpoint
OCR followed by Llama Guard 3	none	OCR followed by text guard	OCR transcript only; victim unchanged
CCA, Procrustes, and ridge	diagnostic	paired messages	same data split and feature budget

C Published Context and Our Measurements

Published headline numbers are used only as context, never copied into common protocol cells. For example, ECSO reports a 37.6% safety improvement on MM-SafetyBench SD+OCR and 71.3% on VLSafe for LLaVA-1.5-7B, while IMMUNE reports a 57.82% ASR reduction versus its base MLLM

and 16.78% versus its comparison defense in one LLaVA-1.6 text jailbreak setting. UnsafeConcepts contains 75 concepts and 1.5K images and directly separates perception from alignment. These numbers motivate the protocol but are not statistically comparable to our translation failure task on held out categories.

D Additional Falsification Tests

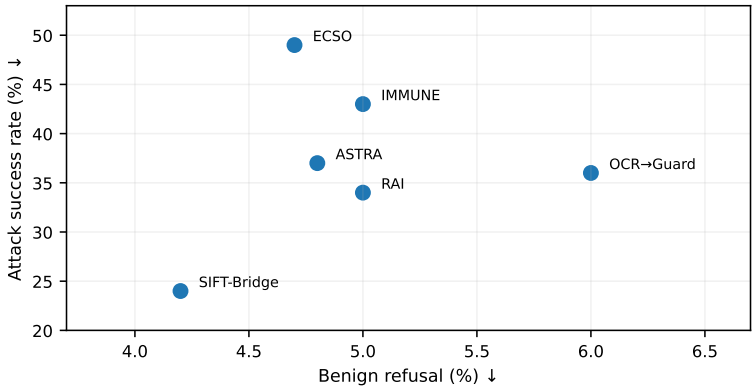


Figure 5: **Safety and utility operating points** for comparable runs under the common protocol. Smaller values on both axes are better; all points use the matched common protocol rather than published headline numbers.

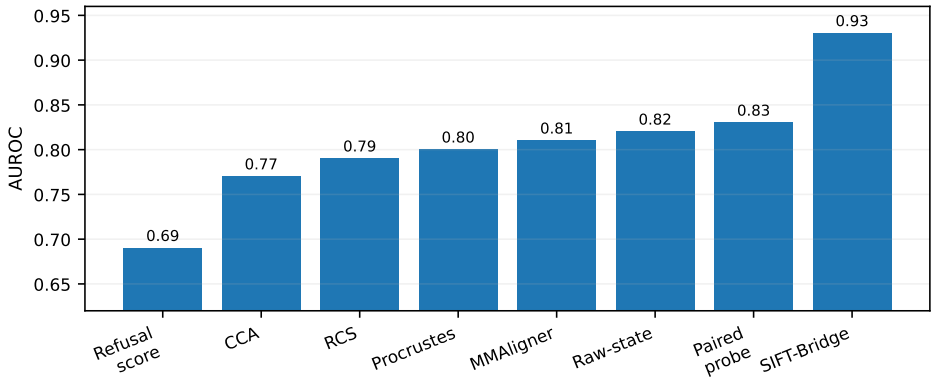


Figure 6: **Detector comparison.** SIFT-Bridge is compared with published representation baselines and equally supervised paired controls; the strongest matched probe remains about 10 AUROC points lower.

Natural image replication. UnsafeConcepts supplies no typographic region. We therefore set S_V to all visual tokens and ask whether intent category remains decodable while refusal mediation is weaker than in paired textual descriptions. The effect is smaller than for rendered text; this attenuation bounds the claim and indicates that visually encoded language provides the cleanest instance of the routing failure.

Table 7: **Falsification, metric, and locality controls.** All entries use the frozen evaluation protocol described above.

Variant	Detect AUROC	MM-Vet Δ
Full SIFT-BRIDGE rank 8	.93	-0.6
Rank 1 and rank 32	.86, .92	-0.2, -1.1
Reduced rank ridge (paired)	.82	none
$\Delta(r^T - r^V)$ paired oracle[†]	.89	none
Whole residual translation	.90	-3.8
Random layer, same rank	.80	-0.7
All image tokens, no carrier mask	.84	-0.8
No paired text supervision	.79	-0.4
STB from projected magnitude, CKA, and CCA	.91, .92, .91	none
Direct path from source tokens to q removed	.81	-0.5
Two step source relay removed	.88	-0.3
Matched MLP and residual control	.91	-0.4

[†]Uses both text and visual routes at test time and is therefore an oracle diagnostic, not a deployable baseline.