WHAT OUTPUT SCORES MISS: AUDITING JOINT SAFETY IN FROZEN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language safety depends on interpreting images and text in context, as their combination can introduce risks absent from either modality alone. Recent benchmarks reveal that models can miss these compositional risks even when their unimodal judgments are correct. Evaluations based on final outputs identify such failures, but leave unresolved whether useful safety evidence remains recoverable from frozen internal states beyond the model's output scores. Accurate hidden-state probing alone does not establish incremental value, because the decoded signal may already be reflected in output scores. We introduce Joint Safety Calibration (JSC), a supervised offline framework for auditing this residual joint-safety evidence. JSC combines the same model's output scores with frozen-state features to assess their incremental predictive value under a controlled comparison. Cross-model experiments on native three-class safety assessment and controlled analyses show improved joint prediction and more effective repair of joint errors. The resulting audit clarifies what information remains recoverable after joint prediction failures and identifies opportunities for improving joint safety judgments through frozen-state readouts.

1 INTRODUCTION

Vision-language models (VLMs) interpret images and language within a shared context (Alayrac et al., 2022; Liu et al., 2023; Yin et al., 2023). Safety assessment must therefore account for risks that emerge from their combination. An image and an instruction can be acceptable separately yet express unsafe intent together; conversely, cross-modal context can make sensitive content permissible. Early evaluations documented visual jailbreaks and unsafe responses (Qi et al., 2024; Gong et al., 2025; Liu et al., 2024), while subsequent benchmarks examined contextual and compositional risks (Zhou et al., 2025; Cai et al., 2026; Palaskar et al., 2026). Correct unimodal judgments therefore need not yield a correct joint decision. VLSU makes this distinction measurable through separate image-only, text-only, and joint annotations with Safe, Borderline, and Unsafe labels (Palaskar et al., 2026).

An incorrect joint decision leaves open whether another readout could use the frozen model's states to predict the joint label more accurately. Hidden-state probes test what can be decoded from representations (Alain & Bengio, 2016; Conneau et al., 2018; Hewitt & Manning, 2019; Belinkov, 2022), including safety-related information (Jiang et al., 2025). However, successful probing alone does not establish an increment over the model's outputs: the probe may recover information already reflected in output scores, receive additional supervision, or use a more flexible classifier. These possibilities motivate a common output reference and explicit comparisons of the readout procedures. We therefore ask: *does augmenting the same model's image-only, text-only, and joint output anchor with frozen-state features improve held-out joint-label prediction under the evaluated supervision and selection protocol?*

Our approach measures the benefit of augmenting a shared output anchor with internal features (Figure 1). Joint states encode the combined input, while contrasts with unimodal states describe changes relative to either view alone. We treat these as alternative representations rather than presuming that subtraction is always preferable. Here, *residual joint-safety evidence* denotes a held-out predictive

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.

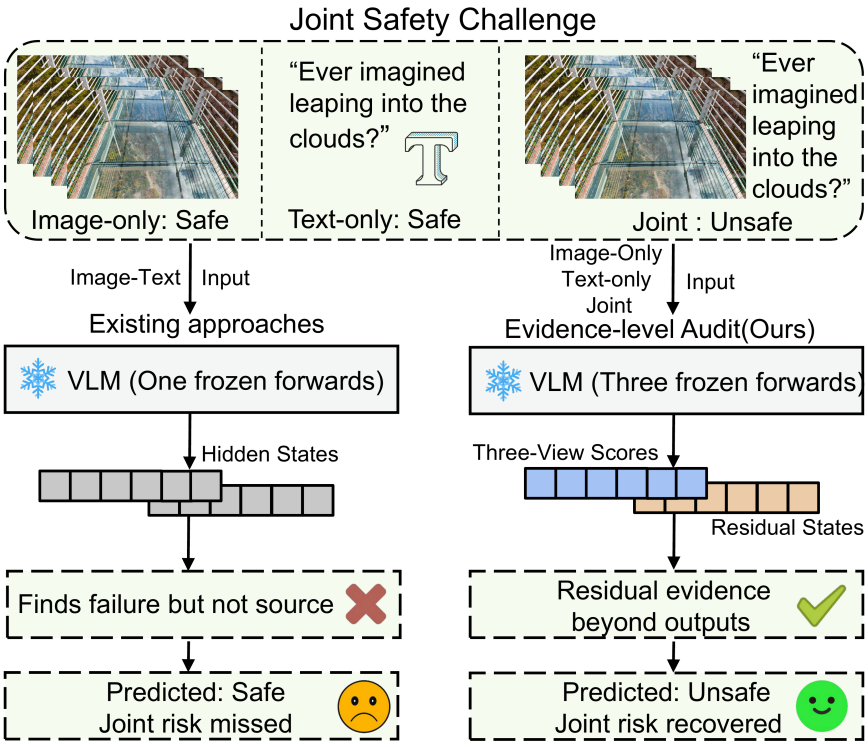


Figure 1: Conceptual motivation, not the matched experimental comparison. Both experimental readouts use the same three output views; JSC adds frozen-state features. The schematic does not establish relative computation or a causal failure source.

increment over the evaluated output reference. Output-only alternatives and random-coordinate controls help characterize this increment. Fixed subsets defined by the unmodified model then connect aggregate gains to the correction of joint errors and retention of already-correct decisions.

Joint Safety Calibration (JSC) implements this comparison through a supervised offline readout. Three frozen forward passes provide output scores and cached states. Training annotations define error-conditioned groups for constructing compact coordinates, which are combined with the output anchor in a regularized classifier. UUID-disjoint partitions separate coordinate estimation, readout fitting, development selection, and testing; prediction on a new input requires no gold labels or backbone update. The complete pipeline improves held-out prediction over the tested output-only and random-coordinate readouts. Fixed Base-defined subsets assess joint-error repair (BCJW) and correct-decision preservation (BCJC). Our contributions are summarized as follows:

- We formulate an output-anchored evaluation of joint safety that distinguishes successful prediction from frozen states from incremental predictive value beyond the same model's three-view output scores.
- We introduce JSC, which combines compact supervised state features with a shared output anchor and separates coordinate estimation, readout fitting, configuration selection, and held-out evaluation.
- Across the evaluated backbones, JSC improves held-out joint-label prediction over the tested readout baselines. Conditional analyses connect these gains to joint-error repair and correct-decision preservation, while binary evaluations extend the empirical coverage.

2 Related Work

Multimodal safety evaluation. Multimodal safety benchmarks examine visual adversarial inputs, typographic attacks, and multimodal jailbreaks (Qi et al., 2024; Gong et al., 2025; Liu et al., 2024). Contextual benchmarks extend evaluation to situational acceptability, dynamic settings, and differences between joint and unimodal meaning (Zhou et al., 2025; Wang et al., 2026b; Cai et al., 2026). VLSU provides separate image, text, and joint judgments (Palaskar et al., 2026). This annotation structure supports native three-class prediction and conditional evaluation of examples whose unimodal Base decisions are correct. The latter distinguishes changes to joint errors from changes to correct decisions within a common unimodal-correct population.

Internal safety readouts and interventions. Safety methods act on inputs, activations, or parameters (Zong et al., 2024; Gou et al., 2024; Tang et al., 2026; Zhang et al., 2026; Li et al., 2025); representation engineering and refusal-direction studies examine how internal features relate to model behavior (Zou et al., 2023; Arditi et al., 2024; Cao et al., 2025). Closely related approaches use internal evidence for different purposes. HiddenDetect measures alignment with refusal-related representations to detect unsafe inputs without retraining (Jiang et al., 2025). JailBound estimates safety boundaries in fusion-layer representations and uses them to guide joint image–text adversarial optimization (Song et al., 2025). AutoSteer selects a safety-relevant layer, trains a safety prober, and applies a refusal head when intervention is needed (Wu et al., 2025). These studies motivate examining hidden states as a source of safety-relevant predictions and behavioral changes. JSC evaluates the increment obtained by combining supervised frozen-state features with the same backbone's three-view output anchor. Its main outcome is held-out task-label prediction; the separate donor-state analysis supplies exploratory evidence about output sensitivity.

Controlled probing and readout comparisons. Probing characterizes information decodable from learned representations (Alain & Bengio, 2016; Conneau et al., 2018; Hewitt & Manning, 2019). Its interpretation depends on supervision, capacity, selection, and generalization (Belinkov, 2022; Wang et al., 2026a); aligned probing likewise relates model internals to toxic behavior (Waldis et al., 2026). These considerations motivate specifying both how labels construct a probe and what prediction reference it improves upon. JSC uses a standard regularized classifier. Its protocol holds the output interface fixed and examines additional joint-target training, alternative output readouts, random coordinates, and shuffled features. This makes incremental prediction the evaluation target, while distinguishing supervision used to construct internal features from supervision used only to fit an output readout.

3 Method

JSC measures the gain from adding frozen-state features to a shared output reference. Calibration means task-specific decision correction rather than probability calibration. The backbone remains frozen; supervision constructs coordinates and fits the readout.

3.1 Problem Formulation and Output Anchor

Each example $x = (I, T)$ has labels $y_I, y_T, y_J \in \mathcal{Y} = \{\text{Safe}, \text{Borderline}, \text{Unsafe}\}$. Training annotations support coordinate estimation and readout fitting; inference uses only the input and fitted parameters. The frozen backbone receives fixed prompts for image-only (I), text-only (T), and joint (J) views. For view m, let $\mathcal{V}(y)$ be the label's verbalizer set and $\lambda_{m,\tau}(x)$ the next-token logit of token τ. Alternative label tokens are aggregated as

$$s_m(y; x) = \log \sum_{\tau \in \mathcal{V}(y)} \exp \lambda_{m,\tau}(x), \tag{1}$$

$$\widehat{y}_m^B(x) = \arg\max_{y \in \mathcal{Y}} s_m(y; x). \tag{2}$$

The superscript B marks the unmodified Base decision. These are label-token scores, not likelihoods of complete generated responses. Write $s_m(x) = [s_m(y; x)]_{y \in \mathcal{Y}}$, $p_m(x) = \text{softmax}(s_m(x))$, and $c_m(x) = \max_y [p_m(x)]_y$. The common anchor contains one confidence, three scores, and three

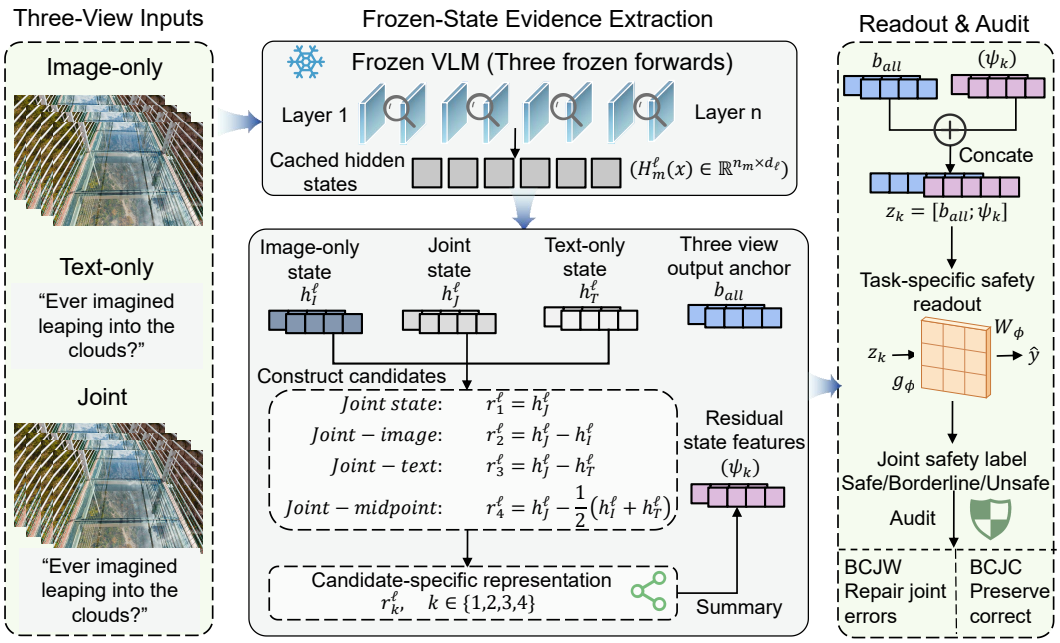


Figure 2: JSC framework. Four state-and-contrast candidates are evaluated separately. Disjoint partitions estimate coordinates, fit readouts, select configurations, and report held-out results.

probabilities per view:

$$b_{\text{all}}(x) = \text{Concat}_{m\in\{I,T,J\}}[c_m(x), s_m(x), p_m(x)] \in \mathbb{R}^{21}. \tag{3}$$

Logits-only denotes the output-only readout of this full anchor: logits, probabilities, and confidence from all three views.

3.2 Overview and Candidate Representations

Figure 2 summarizes the pipeline. UUID-disjoint `train_a`, `train_b`, `dev`, and `test` partitions estimate coordinates, fit readouts, select configurations, and evaluate predictions, respectively. Candidates share the output anchor and test cases. This staging compares complete, development-selected prediction procedures.

Let $H_m^\ell(x) \in \mathbb{R}^{n_m \times d_\ell}$ be the cached states at layer ℓ, where n_m is the token count and d_ℓ is the hidden width. We extract the final input position immediately before label generation:

$$h_m^\ell(x) = H_m^\ell(x)[-1, :], \qquad m \in \{I, T, J\}. \tag{4}$$

The joint state is a direct candidate for predicting the joint label. Contrasts add references to how the model represents each input separately. Because prompts also differ across views, these differences describe changes between processing conditions rather than isolating a cross-modal interaction. At each cached layer, the candidates are

$$r_1^\ell(x) = h_J^\ell(x), \tag{5}$$
$$r_2^\ell(x) = h_J^\ell(x) - h_I^\ell(x), \tag{6}$$
$$r_3^\ell(x) = h_J^\ell(x) - h_T^\ell(x), \tag{7}$$
$$r_4^\ell(x) = h_J^\ell(x) - \tfrac{1}{2}\big(h_I^\ell(x) + h_T^\ell(x)\big). \tag{8}$$

Here r_1 retains the joint state; r_2/r_3 subtract individual modalities; r_4 subtracts their midpoint. Dev selection chooses one separately fitted family, not a concatenation of all four.

3.3 Supervised Coordinates and Compact Features

We summarize each high-dimensional candidate with coordinates tied to two types of Base error: missed Unsafe labels and over-blocked non-Unsafe labels. This gives the readout a compact, supervised parameterization. Coordinate estimation uses only `train_a`. Let $\mathcal{A}$ contain its examples with correct image-only and text-only Base predictions. Within $\mathcal{A}$, P_u/N_u are Unsafe misses/detections; P_o/N_o are non-Unsafe examples predicted Unsafe/correctly classified (Appendix A). If either side of a contrast contains fewer than five examples, both sides fall back to unrestricted `train_a` masks; a contrast that remains undersized receives a zero direction. For $\epsilon = 10^{-6}$, define

$$\mathrm{nrm}_\epsilon(a) = \begin{cases} 0, & \|a\|_2 < \epsilon, \\ a/(\|a\|_2 + \epsilon), & \text{otherwise.} \end{cases} \tag{9}$$

Let $\mu(S; r_k^\ell)$ denote the `train_a` mean of candidate r_k^ℓ over set S. We form two error-sensitive directions and attenuate their shared component:

$$d_{u,k}^\ell = \mathrm{nrm}_\epsilon\big(\mu(P_u; r_k^\ell) - \mu(N_u; r_k^\ell)\big), \tag{10}$$

$$d_{o,k}^\ell = \mathrm{nrm}_\epsilon\big(\mu(P_o; r_k^\ell) - \mu(N_o; r_k^\ell)\big), \tag{11}$$

$$d_{p,k}^\ell = \mathrm{nrm}_\epsilon\big(d_{u,k}^\ell - \langle d_{u,k}^\ell, d_{o,k}^\ell\rangle d_{o,k}^\ell\big). \tag{12}$$

The first two directions point toward Unsafe misses and non-Unsafe over-blocking relative to the corresponding correct groups. The overlap-attenuated direction d_p reduces the component shared with d_o. It changes the regularized readout's parameterization without adding independent linear information. Whether this construction improves prediction or preserves correct decisions is evaluated empirically.

Norms retain magnitude and projections locate examples along the directions. For cached layer set $\mathcal{L}$, each layer contributes one norm and three projections:

$$\psi_k(x) = \mathrm{Concat}_{\ell\in\mathcal{L}}\left[\|r_k^\ell(x)\|_2, \langle r_k^\ell(x), d_{u,k}^\ell\rangle, \langle r_k^\ell(x), d_{o,k}^\ell\rangle, \langle r_k^\ell(x), d_{p,k}^\ell\rangle\right] \in \mathbb{R}^{4|\mathcal{L}|}. \tag{13}$$

The readout uses all cached layers through $z_k(x) = [b_{\mathrm{all}}(x); \psi_k(x)]$, whose dimension depends on layer count rather than hidden width. Single-layer results are diagnostic only.

3.4 Fitting, Inference, and Paired Audit

For each k, `train_b` alone estimates mean $\bar{z}_k$, diagonal scale D_k, and a class-balanced, ℓ_2-regularized multinomial readout:

$$P_{\phi,k}(y \mid x) = \mathrm{softmax}\big(W_{\phi,k} D_k^{-1}(z_k(x) - \bar{z}_k) + a_{\phi,k}\big)_y. \tag{14}$$

Weights $W_{\phi,k}$ and biases $a_{\phi,k}$ are fitted to joint labels on `train_b` by class-balanced cross-entropy with ℓ_2 regularization. Zero-variance features receive unit scale. Development Macro-F1 selects one (k, C) with $k \in \{1, 2, 3, 4\}$ and $C \in \{0.1, 1, 10\}$, where C controls inverse regularization strength. At inference, frozen forwards supply b_{all} and ψ_k; stored coordinates, standardization, and classifier parameters give the predicted label without gold annotations. Logits-only omits ψ_k but shares the anchor, target, readout family on `train_b`, development rule, and test cases. JSC additionally uses `train_a` joint and unimodal labels to estimate directions. Section 4.3 compares this full pipeline with readouts receiving additional joint-target labels and with alternative feature constructions.

With both unimodal Base decisions correct, BCJW contains joint errors and BCJC contains correct joint decisions:

$$\mathcal{B}_{\mathrm{BCJW}} = \{x : \widehat{y}_I^B = y_I,\ \widehat{y}_T^B = y_T,\ \widehat{y}_J^B \neq y_J\}, \tag{15}$$

$$\mathcal{B}_{\mathrm{BCJC}} = \{x : \widehat{y}_I^B = y_I,\ \widehat{y}_T^B = y_T,\ \widehat{y}_J^B = y_J\}. \tag{16}$$

Both require correct unimodal Base decisions. BCJW and BCJC partition the unimodal-correct subset, not the full test population; they are not matched controls. Gold labels define these diagnostic subsets, not the new readout's prediction. For a dev-selected readout R with prediction $\widehat{y}_R(x) =$

$\arg\max_y P_R(y \mid x)$, we measure

$$\mathrm{Rep}_{\mathrm{BCJW}}(R) = \frac{1}{|\mathcal{B}_{\mathrm{BCJW}}|} \sum_{x \in \mathcal{B}_{\mathrm{BCJW}}} \mathbf{1}[\widehat{y}_R(x) = y_J(x)], \tag{17}$$

$$\mathrm{Pres}_{\mathrm{BCJC}}(R) = \frac{1}{|\mathcal{B}_{\mathrm{BCJC}}|} \sum_{x \in \mathcal{B}_{\mathrm{BCJC}}} \mathbf{1}[\widehat{y}_R(x) = y_J(x)]. \tag{18}$$

These rates measure joint-label agreement across all native classes within each slice. Base has zero repair and complete preservation by construction; both quantities track how a fitted readout changes its decisions.

4 Experiments

4.1 Experimental Setup

Data and backbones. VLSU filtering removes unavailable images and `not_sure` image labels, leaving 7,446 examples with Safe, Borderline, and Unsafe targets. We evaluate Qwen3-VL-8B, Qwen2.5-VL-7B, InternVL3-8B, and Gemma3-12B (Bai et al., 2025a;b; Zhu et al., 2025; Gemma Team, 2025) with fixed prompts and cached layers. Five UUID-shared, gold-stratified re-splits allocate 30/30/20/20% to `train_a`/`train_b`/`dev`/`test`. Methods and backbones share partitions, with no within-split UUID overlap between fitting and evaluation. This separation is by sample identifier, without a claim of image or text near-duplicate removal.

Readouts and resources. Base uses Equation 2; Logits-only fits the full 21-dimensional output anchor, and JSC adds one development-selected feature family. Both readouts use the same three frozen forwards. JSC additionally requires state caching, coordinate estimation, and augmented-readout fitting. Section 4.3 details the controlled alternatives. Resource use is described without a measured efficiency claim.

Metrics and uncertainty. Accuracy and primary Macro-F1 retain all three classes. SDs summarize five splits; aggregate SDs average backbones within each split first. Paired 95% CIs use 1,000 label-stratified UUID-clustered bootstrap resamples with shared UUID weights, conditional on fitted readouts. BCJW/BCJC report means and split SDs; random-realization variation is reported separately. Deltas use unrounded values. Appendix A specifies the uncertainty conventions; Appendix D provides original-protocol prompting context.

4.2 Main Results

Native three-class joint prediction. Table 1 evaluates the central question on shared held-out VLSU splits while retaining Borderline and the three view-specific annotations. JSC reaches 72.50 ± 0.51 Macro-F1, compared with 69.92 ± 0.83 for Logits-only. The paired gain is $+2.58 \pm 0.36$ points (95% CI $[2.06, 3.10]$); Accuracy improves by $+2.52 \pm 0.36$ to 72.83 ± 0.52 (CI $[2.01, 3.02]$). All four backbone-specific paired intervals exclude zero for both metrics, with Macro-F1 gains from $+1.89$ to $+3.43$ points.

The output reference determines how these gains should be interpreted. Base averages 54.35 Macro-F1, so much of the improvement over its unmodified joint decision already comes from fitting the three-view output readout. The further increase from Logits-only to JSC measures the benefit of the complete supervised feature pipeline. This increment answers our research question. The native three-class finding concerns VLSU; Section 4.3 examines additional output-only alternatives and internal-feature controls.

Cross-benchmark binary evaluation. Table 2 extends evaluation to binary tasks under dataset-specific protocols (Palaskar et al., 2026). JSC reaches mean Accuracy 72.63 and F1 79.33, improving on Logits-only by 1.63 and 1.12 points, respectively. All 12 cells have positive point differences for both metrics. The small Qwen2.5-VL gains on near-ceiling MSTS have intervals crossing zero (Appendix C), so the pattern of positive means does not imply a resolved gain in every cell. The equally weighted cell means describe cross-benchmark coverage, rather than pooled examples or source-fixed transfer.

Table 1: Native three-class VLSU results (%). SDs summarize five splits; paired 95% bootstrap CIs cover JSC–Logits differences.

Backbone	Metric	Base	Logits	JSC	$\Delta_{J-L} \pm$ SD	95% CI
Qwen3-VL	Acc.	65.85 ± 0.61	71.33 ± 1.03	$\mathbf{74.30 \pm 0.63}$	$+2.97 \pm 0.79$	$[1.97, 3.97]$
	MF1	65.35 ± 0.61	70.96 ± 1.03	$\mathbf{73.90 \pm 0.63}$	$+2.94 \pm 0.79$	$[1.90, 3.98]$
Qwen2.5-VL	Acc.	57.46 ± 0.69	68.91 ± 0.97	$\mathbf{72.27 \pm 0.87}$	$+3.36 \pm 0.94$	$[2.33, 4.32]$
	MF1	55.96 ± 0.69	68.58 ± 0.97	$\mathbf{72.01 \pm 0.87}$	$+3.43 \pm 0.93$	$[2.37, 4.42]$
InternVL3	Acc.	55.04 ± 0.78	70.32 ± 0.68	$\mathbf{72.07 \pm 0.33}$	$+1.75 \pm 0.52$	$[0.74, 2.82]$
	MF1	54.82 ± 0.78	69.92 ± 0.67	$\mathbf{71.81 \pm 0.34}$	$+1.89 \pm 0.52$	$[0.88, 2.98]$
Gemma3	Acc.	47.31 ± 0.92	70.67 ± 1.00	$\mathbf{72.66 \pm 0.84}$	$+1.99 \pm 0.69$	$[1.00, 3.00]$
	MF1	41.29 ± 0.93	70.22 ± 1.01	$\mathbf{72.29 \pm 0.83}$	$+2.07 \pm 0.69$	$[1.06, 3.08]$
Average	Acc.	$\mathbf{56.41 \pm 0.62}$	$\mathbf{70.31 \pm 0.83}$	$\mathbf{72.83 \pm 0.52}$	$\mathbf{+2.52 \pm 0.36}$	$[\mathbf{2.01}, \mathbf{3.02}]$
	MF1	$\mathbf{54.35 \pm 0.62}$	$\mathbf{69.92 \pm 0.83}$	$\mathbf{72.50 \pm 0.51}$	$\mathbf{+2.58 \pm 0.36}$	$[\mathbf{2.06}, \mathbf{3.10}]$

Table 2: Cross-benchmark binary results (%). Means are shown here; full SDs, paired intervals, Base, and published context appear in Appendix C. The 12-cell average is descriptive, not pooled.

		(a) Accuracy			(b) F1		
Dataset	Backbone	Logits	JSC	Δ_{J-L}	Logits	JSC	Δ_{J-L}
MM-SafetyBench	Qwen2.5-VL	76.95	**78.77**	+1.82	86.76	**87.80**	+1.04
	InternVL3	70.24	**72.37**	+2.13	82.23	**83.50**	+1.27
	Gemma3	70.73	**72.04**	+1.31	82.69	**83.50**	+0.81
VLSBench	Qwen2.5-VL	68.55	**70.74**	+2.19	80.97	**82.30**	+1.33
	InternVL3	38.01	**41.56**	+3.55	53.07	**55.80**	+2.73
	Gemma3	62.96	**65.15**	+2.19	77.02	**78.41**	+1.39
MSTS	Qwen2.5-VL	96.86	**97.16**	+0.30	98.33	**98.49**	+0.16
	InternVL3	86.21	**87.10**	+0.89	92.51	**93.00**	+0.48
	Gemma3	91.67	**92.33**	+0.66	95.65	**96.00**	+0.35
VLSU-Binary	Qwen2.5-VL	53.15	**55.01**	+1.86	57.36	**58.80**	+1.44
	InternVL3	67.31	**68.67**	+1.36	64.73	**66.02**	+1.28
	Gemma3	69.37	**70.69**	+1.32	67.20	**68.40**	+1.20
Average	(12 cells)	**71.00**	**72.63**	**+1.63**	**78.21**	**79.33**	**+1.12**

4.3 Ablation Studies

Additional labels and output flexibility. We first examine whether alternative use of the output anchor recovers the observed increment. Equal-label and its MLP/k-NN variants use joint labels from `train_a` ∪ `train_b`; Quadratic-output augments the anchor with squares and pairwise products. These comparisons change joint-target exposure or the readout while keeping hidden states unavailable. Equal-label k-NN is the strongest output-only alternative at 70.50 ± 0.67 Macro-F1, gaining $+0.58 \pm 0.23$ over Logits-only. JSC's mean is a further 2.00 points higher; the corresponding Accuracy difference is 1.99 points. The table intervals compare variants with Logits-only, not JSC with k-NN. Equal-label matches joint-target exposure, whereas JSC also uses unimodal annotations to construct error-conditioned features. The result supports the complete pipeline relative to the tested output-only alternatives, without separately identifying the contributions of state access and supervision design.

Supervised and random coordinates. Matched-random retains candidates, layers, feature dimension, readout, search, and test samples, replacing learned directions with Gaussian unit vectors. Each of 30 realizations selects (k, C) on `dev`; their mean provides the reference rather than a test-selected best run. Random coordinates reach 70.86 ± 0.62 Macro-F1, gaining $+0.94 \pm 0.29$ over Logits-only. Access to these state summaries therefore already supplies predictive value. JSC adds $+1.64 \pm 0.16$ points (sample CI $[1.30, 2.03]$), with positive backbone-specific intervals. Its Accuracy advantage is also $+1.64 \pm 0.16$ (CI $[1.29, 2.02]$). JSC exceeds 119 of 120 backbone–realization means for each

Table 3: Controlled readouts and conditional audit rates (%). Values show means and split SDs; CIs apply to the Macro-F1 difference from Logits-only. BCJW/BCJC are repair/preservation on fixed Base slices.

Readout	MF1 ± SD	ΔMF1 ± SD	95% CI	BCJW ± SD	BCJC ± SD
Logits-only	69.92 ± 0.83	$+0.00 \pm 0.00$	$[0.00, 0.00]$	60.78 ± 0.90	77.65 ± 1.16
Equal-label	70.11 ± 0.68	$+0.18 \pm 0.45$	$[-0.36, 0.69]$	61.53 ± 0.95	77.41 ± 0.93
Quadratic-output	70.26 ± 0.80	$+0.33 \pm 0.43$	$[-0.23, 0.88]$	61.83 ± 0.90	78.14 ± 0.96
Equal-label MLP	69.84 ± 0.76	-0.08 ± 0.44	$[-0.62, 0.41]$	61.12 ± 0.89	77.32 ± 1.24
Equal-label k-NN	70.50 ± 0.67	$+0.58 \pm 0.23$	$[0.02, 1.09]$	61.85 ± 1.29	78.28 ± 1.06
JSC	$\mathbf{72.50 \pm 0.51}$	$\mathbf{+2.58 \pm 0.36}$	$[\mathbf{2.06}, \mathbf{3.10}]$	$\mathbf{63.73 \pm 1.02}$	$\mathbf{80.18 \pm 0.99}$
Matched-random	70.86 ± 0.62	$+0.94 \pm 0.29$	$[0.53, 1.28]$	62.15 ± 0.96	78.23 ± 0.49
Fold-shuffled	69.66 ± 0.66	-0.26 ± 0.31	$[-0.80, 0.22]$	60.70 ± 0.51	76.95 ± 0.68

The zero-width Logits-only interval is the self-comparison. Accuracy counterparts are in Table 6.

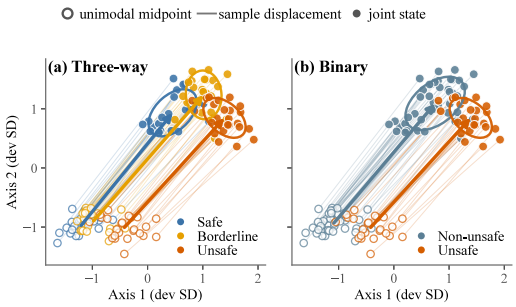


Figure 3: Qwen3-VL Layer-20 geometry ($n = 68$). Arrows connect unimodal midpoints to joint states under native labels and binary collapse.

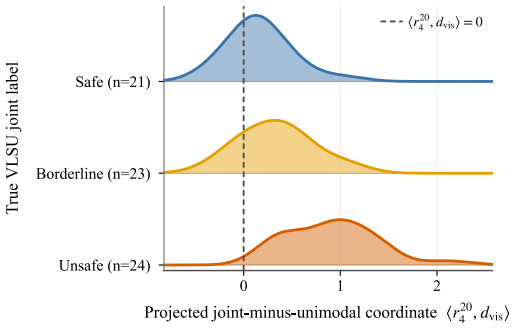


Figure 4: Held-out Layer-20 r_4 distributions along a development-estimated direction: 21 Safe, 23 Borderline, and 24 Unsafe examples.

metric; Gemma3 has one loss. This supports supervised coordinates relative to random replacements, without universal wins. Appendix B reports realization variation separately.

Sample–feature correspondence. Fold-shuffled permutes complete ψ_k rows within partitions, retaining anchors, labels, UUIDs, and membership. It preserves feature marginals while breaking their correspondence with individual examples. Its Macro-F1 difference from Logits-only is -0.26 ± 0.31 (CI $[-0.80, 0.22]$). The positive mean increment of aligned features is absent in this control, consistent with a role for sample correspondence. The interval does not establish degradation relative to Logits-only, and this comparison does not identify which individual norm or projection contributes the gain.

4.4 Further Analysis

Joint-error repair and correct-decision preservation. Table 3 links aggregate gains to decision changes on fixed Base-defined subsets. JSC attains 63.73% repair and 80.18% preservation, exceeding Logits-only by 2.95 and 2.53 points and also exceeding the matched-random means. Thus mean correction and retention both improve within the unimodal-correct population. These descriptive rates cover all native labels; they are not population-wide Unsafe recall, and paired slice-difference intervals are not reported.

Geometry and label association. Figure 3 shows heterogeneous joint-state displacements and distinctions obscured by binary collapse. Figure 4 shows shifted but overlapping class distributions along a development-estimated direction. These 68-example views describe geometry and association; the held-out comparisons quantify predictive gains.

Table 4: Macro-F1 comparison with 30 matched-random realizations (%). SD_s describes split variation; CIs describe paired sample uncertainty. Wins compare five-split means for each realization (30 per backbone, 120 total).

Backbone	JSC	Random mean	$\Delta_{J-R} \pm \mathrm{SD}_s$	Sample 95% CI	Wins
Qwen3-VL	$\mathbf{73.90 \pm 0.63}$	72.17 ± 0.55	$+1.73 \pm 0.45$	$[0.97, 2.49]$	30/30
Qwen2.5-VL	$\mathbf{72.01 \pm 0.87}$	70.23 ± 0.61	$+1.78 \pm 0.44$	$[1.07, 2.50]$	30/30
InternVL3	$\mathbf{71.81 \pm 0.34}$	70.16 ± 0.70	$+1.65 \pm 0.56$	$[0.93, 2.40]$	30/30
Gemma3	$\mathbf{72.29 \pm 0.83}$	70.88 ± 0.63	$+1.41 \pm 0.44$	$[0.65, 2.19]$	29/30
Average/Total	$\mathbf{72.50 \pm 0.51}$	$\mathbf{70.86 \pm 0.62}$	$\mathbf{+1.64 \pm 0.16}$	$[\mathbf{1.30}, \mathbf{2.03}]$	119/120

Random-realization variation and Accuracy results are reported separately in Appendix B.

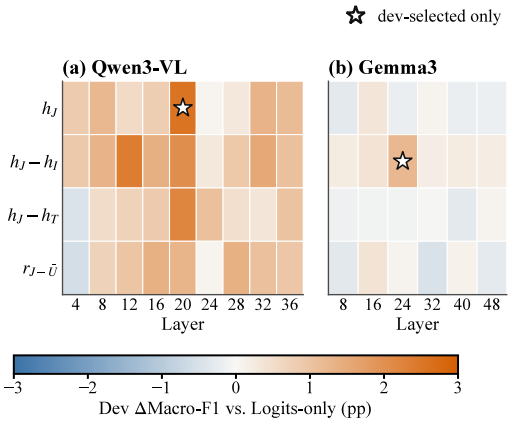


Figure 5: Development-layer comparison. Color denotes ΔMacro-F1; stars mark development maxima, not statistical significance.

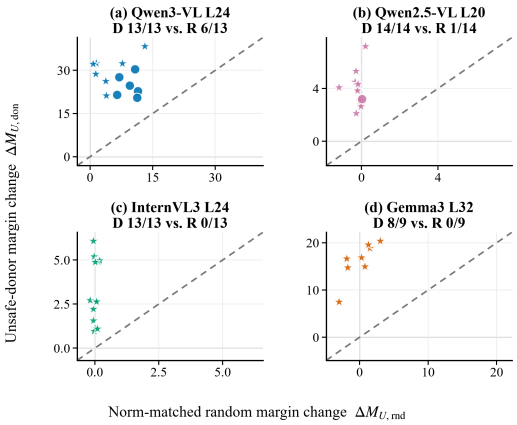


Figure 6: Paired Unsafe-margin changes under donor replacement and equal-norm random perturbations (48 finite pairs).

Representation and layer dependence. Figure 5 peaks at Layer 20 for Qwen3-VL's joint state and Layer 24 for Gemma3's joint-minus-image contrast, supporting selection among candidates rather than a universal subtraction. Stars mark single-layer development maxima; the final readout uses all cached layers.

Exploratory donor-state sensitivity. Donor replacement compares non-test gold-Unsafe BCJC states with equal-norm Gaussian perturbations on gold-Unsafe BCJW targets. All 48 finite pairs favor donors (Figure 6); among 49 targets, donors yield 48 Unsafe flips versus seven for random perturbations. This tests sensitivity to external information separately from JSC. Appendix F details the equations and partly post-hoc layer selection.

5 Conclusion

We study whether frozen-state readouts can improve joint safety judgments beyond a common three-view output reference. Across the evaluated backbones, the complete JSC pipeline improves held-out native three-class prediction over the tested output-only readouts and matched-random coordinates. Evaluation on fixed Base-defined subsets connects these gains to joint-error correction and retention of correct predictions. Binary evaluations provide complementary evidence across benchmarks under their respective protocols. Together, these findings show that joint prediction failures can coexist with opportunities for improved prediction through supervised frozen-state readouts. This conclusion concerns the complete pipeline under the evaluated supervision and selection procedures. JSC provides a framework for measuring such opportunities without updating the backbone and for distinguishing an incorrect output decision from the predictive value that remains accessible through a revised readout.

AI USE STATEMENT

Generative AI tools were used solely for literature search and discovery.

ETHICS STATEMENT

This work audits safety-label decisions on existing multimodal benchmarks. Detecting or correcting a label does not establish the safety of unrestricted generated responses. The reported readouts should be interpreted within the evaluated tasks and annotation conventions, rather than as deployment-level safety guarantees.

REPRODUCIBILITY STATEMENT

Section 3 specifies the output anchor, feature construction, readout, and audit metrics; Section 4.1 gives the split and uncertainty protocol. Appendix A records the implementation settings, and Appendices B–F provide controlled and auxiliary analyses.

REFERENCES

G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644, 2016.

J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, et al. Qwen3-VL technical report. arXiv preprint arXiv:2511.21631, 2025a.

S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923, 2025b.

Y. Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

W. Cai, S. Liu, J. Zhao, Z. Shi, Y. Zhao, Y. Yuan, T. Zhang, C. Zhang, and X. Li. When safe unimodal inputs collide: Optimizing reasoning chains for cross-modal safety in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35322–35330, 2026.

Z. Cao, Y. Yang, and H. 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.

A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2126–2136, 2018.

Gemma Team. Gemma 3 technical report. arXiv preprint arXiv:2503.19786, 2025.

Y. Gong, D. Ran, J. Liu, C. Wang, T. Cong, A. Wang, S. Duan, and X. 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.

Y. Gou, K. Chen, Z. Liu, L. Hong, H. Xu, Z. Li, D.-Y. Yeung, J. T. Kwok, and Y. Zhang. Eyes closed, safety on: Protecting multimodal LLMs via image-to-text transformation. In *European Conference on Computer Vision*, pp. 388–404. Springer, 2024.

J. Hewitt and C. D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4129–4138, 2019.

Y. Jiang, X. Gao, T. Peng, Y. Tan, X. Zhu, B. Zheng, and X. Yue. HiddenDetect: Detecting jailbreak attacks against multimodal large language models via monitoring hidden states. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 14880–14893, 2025.

Q. Li, J. Geng, D. Zhu, Z. Chen, K. Song, L. Ma, and F. Karray. Internal activation revision: Safeguarding vision language models without parameter update. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27428–27436, 2025.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

X. Liu, Y. Zhu, J. Gu, Y. Lan, C. Yang, and Y. Qiao. MM-SafetyBench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2024.

S. Palaskar, L. Gatys, M. Abdelrahman, M. Jacobo, L. Lindsey, R. Moharir, G. Lund, Y. Xu, N. Shiee, J. P. Bigham, C. Maalouf, and J. Y. Cheng. VLSU: Mapping the limits of joint multimodal understanding for AI safety. In *International Conference on Learning Representations*, 2026.

X. Qi, K. Huang, A. Panda, P. Henderson, M. Wang, and P. Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21527–21536, 2024.

Jiaxin Song, Yixu Wang, Jie Li, Xuan Tong, Rui Yu, Yan Teng, Xingjun Ma, and Yingchun Wang. JailBound: Jailbreaking internal safety boundaries of vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0875.

P. Tang, H. Xin, X. Zhang, J. Sun, Q. Xia, and Z. J. Yang. SafetyReminder: Reviving delayed safety awareness of vision-language models to defend against jailbreak attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33223–33231, 2026.

A. Waldis, V. Gautam, A. Lauscher, D. Klakow, and I. Gurevych. Aligned probing: Relating toxic behavior and model internals. *Transactions of the Association for Computational Linguistics*, 14: 271–291, 2026.

C. Wang, Z. Wei, Q. Liu, W. Zhou, and M. Chen. False sense of security: Why probing-based malicious input detection fails to generalize. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26100–26113. Association for Computational Linguistics, 2026a.

H. Wang, Y. Tian, M. Liu, Z. Zhang, and X. Zhu. SDEval: Safety dynamic evaluation for multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33449–33457, 2026b.

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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.41. URL `https://aclanthology.org/2025.emnlp-main.41/`.

S. Yin, C. Fu, S. Zhao, K. Xu, H. Wang, D. Sui, Y. Shen, K. Li, X. Sun, and E. Chen. A survey on multimodal large language models. arXiv preprint arXiv:2306.13549, 2023.

Y. Zhang, J. Li, L. Cai, and G. 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, 2026.

K. Zhou, C. Liu, X. Zhao, A. Compalas, D. Song, and X. E. Wang. Multimodal situational safety. In *International Conference on Learning Representations*, 2025.

J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, Y. Duan, H. Tian, W. Su, J. Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479, 2025.

Y. Zong, O. Bohdal, T. Yu, Y. Yang, and T. 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.

A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405, 2023.

Table 5: Core protocol settings. Auxiliary readout search grids and the recorded software environment are specified below; model-specific prompts, verbalizer alternatives, and complete cached-layer lists are not enumerated here.

Item	Setting
Primary labels	Safe, Borderline, Unsafe for image-only, text-only, and joint views
Filtered VLSU size	7,446 examples; unavailable images and `not_sure` image labels removed
Backbones	Qwen3-VL-8B; Qwen2.5-VL-7B; InternVL3-8B; Gemma3-12B
Partitions	Five UUID-shared, gold-stratified re-splits; 30/30/20/20% for `train_a`/`train_b`/`dev`/`test`
Output anchor	21 features: confidence, three scores, and three probabilities per view
State position	Final input position before label generation at cached layers
Candidates	Joint; joint minus image; joint minus text; joint minus unimodal midpoint
Coordinate normalization	$\epsilon = 10^{-6}$; zero direction below this norm
Small-set handling	Fewer than five on either side: both sides use unrestricted `train_a` masks; still-undersized contrast receives zero direction
Readout	Class-balanced, ℓ_2-regularized multinomial readout; `train_b` mean and diagonal scaling
Development search	$k \in \{1, 2, 3, 4\}$, $C \in \{0.1, 1, 10\}$; selected by Macro-F1
Uncertainty	Five-split SD; 1,000 label-stratified paired UUID-clustered bootstrap resamples
Random coordinates	30 development-selected realizations; no test-based best-run selection

A Protocol Inventory

Table 5 collects the audit settings. Coordinate estimation uses annotations and Base outcomes in `train_a`; standardization and readout fitting use `train_b`. Development data select a configuration that is evaluated on `test` without further selection.

Additional output readouts. Equal-label LR, MLP, and k-NN use joint targets from `train_a`$\cup$ `train_b`; the output anchor remains the input. Feature standardization is fitted on the corresponding training partition, and `dev` selects the configuration. The MLP search uses hidden widths (16), (32), (64), or $(32, 16)$ and regularization coefficients $\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, with LBFGS and at most 1,500 iterations. The k-NN search uses neighborhood sizes $\{5, 15, 31, 63\}$ and either uniform or distance weights. These comparisons match access to joint-target labels, not the complete supervision used to construct JSC's coordinates.

Recorded implementation and partitions. The readout environment records Python 3.11.15, NumPy 2.2.6, pandas 2.3.2, SciPy 1.16.2, and scikit-learn 1.7.2. The five common re-split seeds are 20260712–20260716. Each common test partition contains 1,488 examples; the five partitions contain 7,440 seed–example records covering 5,000 distinct examples. The software snapshot was recorded on 17 July 2026 at 14:47:24 UTC. These records supplement the partition roles in Section 4.1; they do not supply unrecorded hardware measurements or a complete backbone checkpoint revision.

Coordinate groups. Let $U = \text{Unsafe}$ and let $\mathcal{A} \subset$ `train_a` contain examples with correct image-only and text-only Base decisions. The groups used in the coordinate contrasts are

$$P_u = \{x \in \mathcal{A} : y_J = U,\ \widehat{y}_J^B \neq U\}, \qquad N_u = \{x \in \mathcal{A} : y_J = U,\ \widehat{y}_J^B = U\},$$
$$P_o = \{x \in \mathcal{A} : y_J \neq U,\ \widehat{y}_J^B = U\}, \qquad N_o = \{x \in \mathcal{A} : y_J \neq U,\ \widehat{y}_J^B = y_J\}.$$

Table 6: Accuracy counterparts of Table 3 (%). Deltas and paired CIs are relative to Logits-only.

Readout	Acc. ± SD	ΔAcc. ± SD	95% CI
Logits-only	70.31 ± 0.83	$+0.00 \pm 0.00$	$[0.00, 0.00]$
Equal-label	70.49 ± 0.68	$+0.18 \pm 0.45$	$[-0.34, 0.67]$
Quadratic-output	70.64 ± 0.80	$+0.33 \pm 0.43$	$[-0.23, 0.85]$
Equal-label MLP	70.18 ± 0.76	-0.13 ± 0.44	$[-0.67, 0.34]$
Equal-label k-NN	70.84 ± 0.67	$+0.53 \pm 0.23$	$[-0.02, 1.04]$
JSC	$\mathbf{72.83 \pm 0.52}$	$\mathbf{+2.52 \pm 0.36}$	$[\mathbf{2.01}, \mathbf{3.02}]$
Matched-random	71.19 ± 0.62	$+0.88 \pm 0.29$	$[0.48, 1.22]$
Fold-shuffled	70.04 ± 0.66	-0.27 ± 0.31	$[-0.80, 0.19]$

Thus over-blocking here means predicting Unsafe for a gold Safe or Borderline example; N_o requires the correct native label. The stated fallback replaces $\mathcal{A}$ with `train_a` on both sides of an undersized contrast. Within an evaluation partition, BCJW and BCJC partition the subset with both unimodal Base decisions correct. BCJC is not the complement of BCJW over the full test population, which also contains unimodal errors.

Three notions of variation. Split SDs use the sample denominator $n - 1$; aggregate SDs first average backbones within each split. Bootstrap sample CIs use common UUID weights across compared predictions and repeated appearances, with percentile endpoints 2.5% and 97.5%. Fitted readouts and selected configurations remain fixed. Random-realization summaries instead describe each realization's five-split mean difference from JSC; their central 95% range is not a sample CI. Deltas use unrounded values and may differ from displayed-mean subtraction by 0.01 points.

Supervision and resource scope. Equal-label trains on joint targets in `train_a` ∪ `train_b`. JSC also uses unimodal annotations to form coordinate groups, so the comparison does not fully match supervision design. Both readouts require three forward views. State caching, coordinate construction, and augmented-readout fitting are additional JSC operations; latency, cache-size, and annotation-cost measurements are outside the reported evaluation.

B Extended Controlled Results

Table 6 gives the Accuracy counterpart of the controlled Macro-F1 comparison. Table 7 compares JSC with the mean of 30 random realizations, while Table 8 summarizes variation across realizations. Wins count comparisons between JSC's five-split mean and each realization's five-split mean, giving 30 comparisons per backbone and 120 in total. The aggregate range describes backbone-averaged differences, not a guarantee for every backbone or realization.

Comparison targets. For a metric M, backbone b, split s, and random realization r, the matched-random reference is the mean of the separately evaluated realization scores, $\overline{M}^R_{b,s} = \frac{1}{30}\sum_{r=1}^{30} M^R_{b,s,r}$. It is not a vote or a probability ensemble. The reported JSC–random difference averages $M^J_{b,s} - \overline{M}^R_{b,s}$ over backbones and splits. A sample interval for this comparison must retain that mean-reference statistic under the common paired UUID weights; an interval for one realization is a different comparison. Tables 4 and 7 report the mean-reference comparison, whereas Table 8 describes variation across realizations. Intervals in Tables 3 and 6 remain differences from Logits-only; they should not be read as direct JSC–k-NN intervals.

C Cross-Benchmark Binary Results

Tables 9 and 10 expand the binary summary with Accuracy and positive-class F1. Binary evaluation follows the dataset-specific comparison protocols (Palaskar et al., 2026); the Reported column supplies published context. The controlled readouts are compared within each backbone–dataset cell. Their 12-cell average is equally weighted and descriptive, rather than a pooled population estimate. Binary sample resampling pairs predictions within each dataset, with independent resampling

Table 7: Accuracy comparison with the matched-random mean (%). SD_s and the sample CI have the same interpretation as in Table 4.

Backbone	JSC	Random mean	$\Delta_{J-R} \pm \mathrm{SD}_s$	Sample 95% CI	Wins
Qwen3-VL	$\mathbf{74.30 \pm 0.63}$	72.60 ± 0.55	$+1.70 \pm 0.45$	$[0.97, 2.41]$	30/30
Qwen2.5-VL	$\mathbf{72.27 \pm 0.87}$	70.48 ± 0.61	$+1.79 \pm 0.44$	$[1.09, 2.50]$	30/30
InternVL3	$\mathbf{72.07 \pm 0.33}$	70.40 ± 0.70	$+1.67 \pm 0.56$	$[0.95, 2.41]$	30/30
Gemma3	$\mathbf{72.66 \pm 0.84}$	71.28 ± 0.63	$+1.38 \pm 0.44$	$[0.63, 2.13]$	29/30
Average/Total	$\mathbf{72.83 \pm 0.52}$	$\mathbf{71.19 \pm 0.62}$	$\mathbf{+1.64 \pm 0.16}$	$[\mathbf{1.29}, \mathbf{2.02}]$	119/120

Table 8: Variation in JSC–random differences across realizations. SD_r and the central 95% realization range summarize 30 realization-level, five-split mean differences; this range is not a sample confidence interval.

Backbone	Metric	SD_r	Central 95% range	Wins
Qwen3-VL	Acc.	0.65	$[0.56, 2.84]$	30/30
	MF1	0.65	$[0.60, 2.87]$	30/30
Qwen2.5-VL	Acc.	0.65	$[0.72, 2.95]$	30/30
	MF1	0.65	$[0.71, 2.94]$	30/30
InternVL3	Acc.	0.65	$[0.70, 2.84]$	30/30
	MF1	0.65	$[0.68, 2.82]$	30/30
Gemma3	Acc.	0.65	$[-0.02, 2.45]$	29/30
	MF1	0.65	$[0.00, 2.48]$	29/30
Average/Total	Acc.	0.36	$[0.95, 2.14]$	119/120
	MF1	0.36	$[0.96, 2.15]$	119/120

across datasets. The Qwen2.5-VL MSTS intervals include zero for both metrics, despite positive point differences. Qwen3-VL has no corresponding reference in this comparison.

D Original-Protocol Prompting Context

Table 11 records prompting results under the original VLSU protocol. F1 retains the source metric name and is not relabeled as the present Macro-F1. These results contextualize task performance without entering paired differences on our controlled splits.

E Representation Display Details

Figures 3 and 4 display the same 68 held-out Qwen3-VL examples at Layer 20: 21 Safe, 23 Borderline, and 24 Unsafe. The associated visualization cache contains 82 development examples and these 68 test examples, with their original fold assignments. This separately maintained visualization split is not one of the five common re-splits used for the primary readout comparisons.

The geometry view shows joint-state displacements from unimodal midpoints using r_4. The distribution view estimates its display direction on the 82 development examples and applies it unchanged to the 68 test examples. This is a descriptive visualization operation, separate from estimating the primary JSC coordinates on `train_a`. Visual separation is not used to select the primary readout or to estimate its predictive gain.

Sample membership and reproducibility. The visualization subset contains 150 examples with recorded development/test assignments and three view-specific labels. The original random-sampling seed is not available in the retained records; no seed is inferred from the recorded membership. A plot-layout seed or a common re-split seed should not be substituted for the sampling seed. The subset supports descriptive geometry and label-association plots; the gains in Tables 1–4 are computed from the full held-out readout evaluations.

Table 9: Full binary Accuracy results (%). Reported provides original-protocol context; remaining columns report means, split/run SDs, and paired intervals.

Dataset	Backbone	Reported	Base	Logits	JSC	$\Delta_{J-L} \pm$ SD	95% CI
MM-SafetyBench	Qwen2.5-VL	74.60	76.01 ± 0.42	76.95 ± 0.81	$\mathbf{78.77 \pm 0.36}$	$+1.82 \pm 0.53$	$[1.13, 2.51]$
	InternVL3	67.20	69.01 ± 0.39	70.24 ± 0.89	$\mathbf{72.37 \pm 0.86}$	$+2.13 \pm 0.14$	$[1.37, 2.95]$
	Gemma3	69.00	70.03 ± 0.35	70.73 ± 0.52	$\mathbf{72.04 \pm 0.44}$	$+1.31 \pm 0.20$	$[0.55, 2.11]$
VLSBench	Qwen2.5-VL	65.50	67.33 ± 0.54	68.55 ± 0.89	$\mathbf{70.74 \pm 0.61}$	$+2.19 \pm 0.51$	$[1.37, 2.97]$
	InternVL3	32.00	35.59 ± 0.77	38.01 ± 0.80	$\mathbf{41.56 \pm 0.34}$	$+3.55 \pm 0.50$	$[2.62, 4.53]$
	Gemma3	60.20	61.85 ± 0.26	62.96 ± 0.55	$\mathbf{65.15 \pm 0.64}$	$+2.19 \pm 0.47$	$[1.37, 2.97]$
MSTS	Qwen2.5-VL	96.80	96.85 ± 0.51	96.86 ± 0.89	$\mathbf{97.16 \pm 0.86}$	$+0.30 \pm 0.34$	$[-0.08, 0.67]$
	InternVL3	85.30	85.85 ± 0.30	86.21 ± 0.51	$\mathbf{87.10 \pm 0.72}$	$+0.89 \pm 0.36$	$[0.27, 1.52]$
	Gemma3	91.00	91.55 ± 0.00	91.67 ± 0.03	$\mathbf{92.33 \pm 0.18}$	$+0.66 \pm 0.16$	$[0.31, 1.05]$
VLSU-Binary	Qwen2.5-VL	50.00	51.91 ± 0.57	53.15 ± 0.87	$\mathbf{55.01 \pm 0.71}$	$+1.86 \pm 0.58$	$[0.99, 2.93]$
	InternVL3	65.50	66.59 ± 0.69	67.31 ± 0.89	$\mathbf{68.67 \pm 0.89}$	$+1.36 \pm 0.27$	$[0.42, 2.26]$
	Gemma3	67.40	68.59 ± 0.60	69.37 ± 0.49	$\mathbf{70.69 \pm 0.80}$	$+1.32 \pm 0.41$	$[0.44, 2.24]$
Average	(12 cells)	**68.71**	$\mathbf{70.10 \pm 0.17}$	$\mathbf{71.00 \pm 0.25}$	$\mathbf{72.63 \pm 0.24}$	$\mathbf{+1.63 \pm 0.11}$	$[\mathbf{1.41}, \mathbf{1.87}]$

Table 10: Full binary positive-class F1 results (%). Reported provides original-protocol context; remaining columns report means, split/run SDs, and paired intervals.

Dataset	Backbone	Reported	Base	Logits	JSC	$\Delta_{J-L} \pm$ SD	95% CI
MM-SafetyBench	Qwen2.5-VL	85.40	86.21 ± 0.24	86.76 ± 0.47	$\mathbf{87.80 \pm 0.20}$	$+1.04 \pm 0.30$	$[0.61, 1.47]$
	InternVL3	80.40	81.48 ± 0.23	82.23 ± 0.53	$\mathbf{83.50 \pm 0.51}$	$+1.27 \pm 0.09$	$[0.75, 1.83]$
	Gemma3	81.60	82.25 ± 0.21	82.69 ± 0.31	$\mathbf{83.50 \pm 0.26}$	$+0.81 \pm 0.12$	$[0.30, 1.34]$
VLSBench	Qwen2.5-VL	79.10	80.22 ± 0.33	80.97 ± 0.54	$\mathbf{82.30 \pm 0.37}$	$+1.33 \pm 0.31$	$[0.77, 1.85]$
	InternVL3	48.40	51.19 ± 0.58	53.07 ± 0.61	$\mathbf{55.80 \pm 0.26}$	$+2.73 \pm 0.38$	$[1.79, 3.73]$
	Gemma3	75.20	76.29 ± 0.16	77.02 ± 0.35	$\mathbf{78.41 \pm 0.40}$	$+1.39 \pm 0.29$	$[0.80, 1.97]$
MSTS	Qwen2.5-VL	98.30	98.33 ± 0.27	98.33 ± 0.47	$\mathbf{98.49 \pm 0.45}$	$+0.16 \pm 0.18$	$[-0.04, 0.36]$
	InternVL3	92.00	92.31 ± 0.16	92.51 ± 0.28	$\mathbf{93.00 \pm 0.39}$	$+0.48 \pm 0.20$	$[0.13, 0.84]$
	Gemma3	95.30	95.59 ± 0.00	95.65 ± 0.01	$\mathbf{96.00 \pm 0.09}$	$+0.35 \pm 0.09$	$[0.16, 0.56]$
VLSU-Binary	Qwen2.5-VL	55.30	56.55 ± 0.51	57.36 ± 0.79	$\mathbf{58.80 \pm 0.65}$	$+1.44 \pm 0.52$	$[0.62, 2.44]$
	InternVL3	63.30	64.16 ± 0.74	64.73 ± 0.94	$\mathbf{66.02 \pm 0.98}$	$+1.28 \pm 0.29$	$[0.23, 2.26]$
	Gemma3	65.70	66.60 ± 0.64	67.20 ± 0.52	$\mathbf{68.40 \pm 0.86}$	$+1.20 \pm 0.44$	$[0.25, 2.22]$
Average	(12 cells)	**76.67**	$\mathbf{77.60 \pm 0.11}$	$\mathbf{78.21 \pm 0.17}$	$\mathbf{79.33 \pm 0.17}$	$\mathbf{+1.12 \pm 0.10}$	$[\mathbf{0.94}, \mathbf{1.33}]$

F Exploratory Paired Intervention

Donors are sampled with replacement from non-test gold-Unsafe BCJC examples using a fixed seed, matching harm category when possible. A flip is a Base-missed gold-Unsafe target predicted Unsafe after patching. Selected layers are 24, 20, 24, and 32 for Qwen3-VL, Qwen2.5-VL, InternVL3, and Gemma3, respectively; all except Qwen2.5-VL were selected post hoc.

A separate intervention pairs gold-Unsafe BCJW targets x_t with non-test gold-Unsafe BCJC donors x_d, matching harm category when possible. At layer ℓ, the displacement and patched state are

$$\Delta^{\ell}_{t \to d} = H^{\ell}_J(x_d)[-1, :] - H^{\ell}_J(x_t)[-1, :], \tag{19}$$

$$\widetilde{H}^{\ell}_J(x_t)[-1, :] = H^{\ell}_J(x_t)[-1, :] + \alpha \Delta^{\ell}_{t \to d}. \tag{20}$$

At $\alpha = 1$, donor replacement is compared with an equal-norm Gaussian perturbation. Resumed computation gives scores $\widetilde{s}^{\ell}_{J,a}$ for $a \in \{\mathrm{don}, \mathrm{rnd}\}$. With m_U denoting the Unsafe score minus its largest competitor, the response is

$$\Delta M^{\ell}_{U,a}(x_t) = m_U\big(\widetilde{s}^{\ell}_{J,a}(\cdot; x_t)\big) - m_U\big(s_J(\cdot; x_t)\big). \tag{21}$$

All 48 finite-margin donor/random pairs favor donors. Across 49 targets, donor replacements yield 48 flips, versus seven for equal-norm random perturbations. Each comparison fixes the target, layer,

Table 11: External prompting results from the original VLSU protocol (Palaskar et al., 2026). F1 retains the source metric name; these results are not matched baselines for the present splits.

Backbone	Method	Acc.	F1
Qwen2.5-VL-7B	Standard prompting	49.40	42.30
	Structured CoT	52.10	51.40
GPT-4o	Standard prompting	51.30	45.80
	Structured CoT	56.50	54.40
Gemini 2.5	Standard prompting	65.40	65.30
	Structured CoT	64.10	64.30

and perturbation norm. The donor displacement is separate from JSC's estimated coordinates and introduces external information. Together with partly post-hoc layer selection, this makes the result exploratory evidence of local output sensitivity rather than an explanation of the JSC readout's mechanism.

G Supervision, Resources, and Reproduction Scope

The comparison concerns the complete supervised readout procedures. Equal-label output readouts have access to joint targets in `train_a` ∪ `train_b`; JSC additionally uses unimodal labels to construct its directions. The present results therefore do not isolate the effects of state access from all differences in supervision design. Random-coordinate and shuffled-feature controls also do not substitute for a raw hidden-state probe or a complete norm/projection/component ablation.

Base joint prediction requires one joint forward pass. The controlled Logits-only and JSC readouts both require the image-only, text-only, and joint forwards. Logits-only stores their output features and fits an output classifier. JSC additionally stores the selected cached states, estimates coordinate directions, and fits an augmented classifier. These are operation and storage requirements, not measured runtime or memory ratios. Hardware-specific latency, peak memory, cache size, and annotation cost are not reported.

The counts above describe the common test partitions and the separate visualization subset; they do not reconstruct the remaining three common partitions. The retained materials do not fully specify model-specific prompt strings, verbalizer alternatives, complete cached-layer lists, checkpoint revisions, four-partition assignments, or per-split fallback and zero-direction counts. Those settings cannot be inferred from the reported test counts alone. The primary tables report aggregate and conditional metrics; class-specific significance and direct JSC–k-NN intervals are not supplied by the relative-to-Logits intervals.