FISHER–RAO DECISION GEOMETRY FOR VLM MISBEHAVIOR MONITORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Detecting VLM misbehavior requires identifying incorrect or unsafe answers, including those generated under adversarial inputs and distribution shifts. Output confidence can remain high for such answers and does not fully describe the prediction process. Predictions with similar confidence may respond differently to small changes in decoder states, while their final distributions can conceal revisions across layers. We investigate whether these differences provide additional evidence of failure. Fisher–Rao geometry connects state variations to changes in predictive distributions and provides a shortest-path distance for comparing predictions across layers. Based on these properties, we propose Fisher–Rao Decision Geometry (FRDG), a training-free, white-box monitor. FRDG measures the strongest local predictive change under a perturbation budget relative to the state norm. It also measures accumulated changes across late-layer predictions beyond the direct distance between the first and last predictions. These measurements capture local sensitivity and intermediate revision, respectively, and their standardized summaries are combined into a risk score. We derive the local displacement interpretation and bounds on the additional layerwise change. Across four VLMs and nine benchmark subsets, the reported results show relative improvements of up to 12.42% in macro AUROC and 15.41% in macro AP over the strongest compared baseline for each model and metric. Component ablations show that the combined score improves failure ranking over either measurement alone.

1 INTRODUCTION

VLM misbehavior monitoring aims to identify generated answers that contain factual errors or violate safety requirements, including failures under adversarial inputs and distribution shifts (Li et al., 2023; Mazeika et al., 2024; Huang et al., 2026). Detection must assess the answer rather than infer failure from the input, since a harmful prompt may still receive a safe refusal. A central difficulty is that an incorrect answer can receive high confidence. The internal prediction process may help distinguish such failures. Predictions with similar confidence can differ in their dependence on the final decoder state. Stability near that state, however, does not describe the revisions made before the final prediction: similar final distributions can follow substantially different sequences of layerwise updates. These revisions may therefore contain information that a local measurement alone misses. We investigate whether final-state sensitivity and intermediate prediction changes together provide useful evidence for identifying incorrect or unsafe answers beyond their output confidence.

Existing monitors use predictive probabilities to summarize token or sequence confidence (Malinin & Gales, 2021), or compare generated answers for factual and semantic consistency (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024). High confidence can accompany errors, while repeated generation incurs additional cost and may reproduce the same mistake. Internal methods provide further information through representation statistics, safety-related activations, and evidential conflict or insufficient support (Sriramanan et al., 2024; Jiang et al., 2025; Huang et al., 2026). Prior work also examines intermediate vocabulary predictions and layerwise evidential uncertainty (Belrose et al., 2025; Huang et al., 2026). However, the uncertainty assigned to a distribution does not directly measure its sensitivity to nearby decoder states or its distance from distributions at other layers. Quantifying these changes requires a distance that reflects their effect on token probabilities. Euclidean distances between hidden states do not provide this interpretation, since equally large state changes can affect probabilities differently. Fisher–Rao geometry supplies a local metric for

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.

predictive sensitivity and a shortest-path distance between distributions (Nielsen, 2020). Because normal prediction refinement also involves change, the shortest-path distance between the first and last predictions provides a reference for assessing the amount of movement accumulated across layers.

We propose Fisher–Rao Decision Geometry (FRDG), a training-free, white-box monitor that measures final-state sensitivity and intermediate prediction changes in the same probability space. For each selected answer position, FRDG holds the image, prompt, and preceding answer tokens fixed and considers small perturbations of the final decoder state. Local sensitivity estimates the strongest predictive change under a perturbation budget proportional to the state norm, using a second-order approximation. Here, local refers to the neighborhood of that state, not to successive decoder layers. For the layerwise measurement, a shared frozen output mapping converts selected late-layer states into distributions over the same vocabulary. FRDG sums the distances between successive predictions and subtracts the shortest-path distance between the first and last selected predictions. This difference, termed trajectory excess, measures intermediate movement beyond the overall separation of the two distributions. FRDG summarizes both measurements across answer positions, standardizes them using reference responses, and combines them into a risk score, without updating the VLM or generating additional answers.

We show how the relative state-perturbation budget leads to a local sensitivity score determined by the leading eigenvalue of the state Fisher operator, and derive bounds on trajectory excess. The local score can be estimated without constructing a dense Fisher matrix. We evaluate the resulting monitor on four VLMs and nine benchmark subsets. The reported results show relative improvements of up to 12.42% in macro AUROC and 15.41% in macro AP over the strongest compared baseline for each model and metric. Component ablations favor the combined score over either measurement alone. Across all four models, further comparisons favor retaining relative-state scaling and using trajectory excess rather than total path length or the first-to-last distance. These results support combining local predictive sensitivity with intermediate prediction changes to improve the ranking of incorrect or unsafe VLM responses.

2 Related Work

2.1 Uncertainty-based monitoring

Uncertainty-based monitors assess reliability using predictive probabilities or variation across generated answers. Token-level and sequence-level probability measures summarize confidence in autoregressive predictions (Malinin & Gales, 2021). Sampling-based methods assess factual consistency or group answers by meaning to distinguish semantic disagreement from differences in wording (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024). Self-consistency also aggregates sampled reasoning paths, although its primary goal is answer selection rather than failure detection (Wang et al., 2023). These sampling-based approaches require additional generations to measure variation across answers. Evidential methods examine the evidence underlying a prediction. EUQ interprets output-head features as supporting or opposing evidence and quantifies conflict and ignorance, representing contradictory evidence and insufficient support, respectively (Huang et al., 2026). It evaluates these quantities across hallucination, jailbreak, adversarial, and OOD scenarios without additional model training. These methods capture different aspects of uncertainty and require different levels of access, from generated text to output probabilities and internal features.

2.2 Monitoring internal representations

Internal representations contain information that is not directly available from generated text. Statistical approaches examine variation across sampled answers or patterns within a single response. INSIDE uses the covariance of response representations to measure consistency across answers (Chen et al., 2024), while LLM-Check analyzes hidden states, attention maps, and output probabilities (Sriramanan et al., 2024). Other methods construct compact features for learned detectors. HaloScope extracts informative directions from unlabeled generations, and HARP uses projections derived from the output head for hallucination detection (Du et al., 2024; Hu et al., 2025). These methods reduce annotation requirements or feature dimensionality, but still require detector training. Related safety methods use refusal-related activations or gradient signals to identify jailbreak inputs (Jiang et al.,

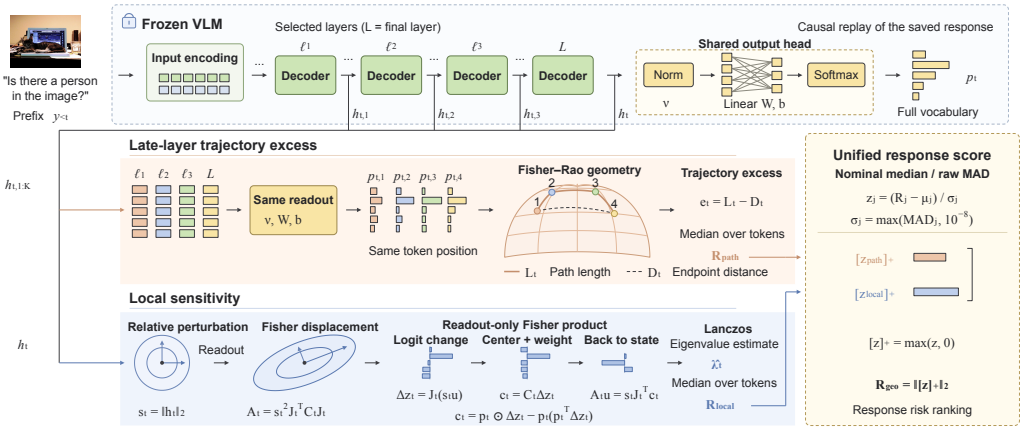

Figure 1: A frozen VLM processes the saved answer to obtain decoder states at selected layers and prediction positions. A shared output mapping supports two measurements: local predictive sensitivity and late-layer trajectory excess $e_t = L_t - D_t$. Their medians over selected positions are standardized using reference-set medians and raw MADs, then combined into a response risk score.

2025; Xie et al., 2024; Hu et al., 2024). Detecting such inputs is distinct from assessing whether a completed answer is unsafe. Layerwise analysis examines how predictions change through the decoder. The tuned lens learns affine maps from intermediate states to vocabulary distributions and uses the resulting prediction trajectories for malicious-input detection (Belrose et al., 2025). These trajectories reveal changes in intermediate predictions, but do not by themselves establish whether the reasoning or final answer is correct.

2.3 Geometric uncertainty estimation

Information geometry relates changes in model parameters or representations to changes in probability distributions (Amari, 1998; Nielsen, 2020). The Fisher metric describes this relationship locally and provides a measure of predictive sensitivity. Parameter-space approaches use training data to estimate curvature or characterize gradient statistics for OOD detection (Sharma et al., 2021; Dauncey et al., 2024). Recent Fisher–Rao sensitivity analysis connects the Fisher trace to feature magnitude and output uncertainty (Nguyen et al., 2026). These sensitivity measures depend on both the space in which perturbations are defined and the statistic used to summarize their effects. Another geometric approach estimates the density of reference data in representation space. REPVLM learns a probability density over normalized VLM embeddings using Riemannian flow matching and uses low density as an uncertainty signal (Ju et al., 2026). Fisher–Rao distance also serves as a regularizer during training to improve adversarial robustness (Picot et al., 2022). These methods share geometric tools but address different objectives, including predictive sensitivity, representation density, and robustness through model training.

FRDG builds on internal monitoring and information geometry to assess completed VLM responses. It measures local sensitivity at the fused decoder state and trajectory excess across late-layer predictions using a shared frozen readout. The focus is on changes in predictive distributions, rather than evidential conflict or representation density, and their relation to response failure.

3 Method

We define the monitoring setting, derive local predictive sensitivity and late-layer trajectory excess, and describe their computation and combination into a response risk score. Figure 1 outlines FRDG.

3.1 MONITORING SETTING AND STATE EXTRACTION

Let $x = (v, q)$ denote an image and prompt, and $y = (y_1, \ldots, y_N)$ the answer from a frozen autoregressive VLM. FRDG assigns a risk score $R_{\text{geo}}(x, y)$, with larger values indicating higher review priority. The label $B(x, y) \in \{0, 1\}$ indicates answer failure, not an adversarial or OOD input. FRDG requires white-box access to decoder states and the frozen output mapping, or readout, without updating model parameters or fitting the score to failure labels.

We process the saved answer once with causal masking and backbone gradients disabled. The state predicting y_t uses only the image, prompt, and preceding tokens $y_{<t}$: it is taken at the final prompt position for $t = 1$ and the preceding answer position otherwise.

Let $\mathcal{T}$ be the retained prediction positions and $\ell_1 < \cdots < \ell_K = L$ the selected blocks of an L-block decoder. For each t, $h_{t,k} \in \mathbb{R}^d$ is the output of block ℓ_k before final model normalization, with $h_t = h_{t,K}$. The shared frozen readout is

$$z(h) = W\nu(h) + b, \qquad p_{t,k} = \text{softmax}(z(h_{t,k})), \qquad p_t = p_{t,K}, \tag{1}$$

where $W \in \mathbb{R}^{V \times d}$, $b \in \mathbb{R}^V$, and ν is the final normalization. All distributions share the same V vocabulary coordinates, and the states combine image, prompt, and answer-prefix information. Scores use original full-vocabulary probabilities before export compression. Appendix B specifies layer and position selection, native-logit checks, and handling of responses without eligible positions.

3.2 LOCAL PREDICTIVE SENSITIVITY

State perturbations of equal magnitude can affect predictions differently across directions. We measure those effects using the categorical Fisher–Rao distance (Nielsen, 2020):

$$d_{\text{FR}}(p, r) = 2 \arccos\left(\sum_{a=1}^{V} \sqrt{p_a r_a}\right). \tag{2}$$

Under $p \mapsto 2\sqrt{p}$, distributions lie on the positive part of a radius-two sphere, with shortest-arc distance in $[0, \pi]$.

Let $J_t = \partial z(h)/\partial h|_{h=h_t} \in \mathbb{R}^{V \times d}$ be the readout logit Jacobian, excluding backbone derivatives. For token coordinate a with coordinate vector e_a, the gradient $g_t(a) = J_t^\top(e_a - p_t)$ has zero mean under p_t. Its covariance defines the state Fisher matrix:

$$F_t = \mathbb{E}_{a \sim p_t}[g_t(a) g_t(a)^\top] = J_t^\top C_t J_t, \qquad C_t = \text{Diag}(p_t) - p_t p_t^\top. \tag{3}$$

The positive semidefinite, possibly singular matrix F_t expresses predictive change in state coordinates: $\delta^\top F_t \delta$ approximates squared Fisher–Rao distance to second order.

An absolute perturbation budget corresponds to different fractional changes at different state norms. We therefore set $s_t = \|h_t\|_2$, write $\delta = s_t u$, and define $A_t = s_t^2 F_t$. Holding s_t fixed during differentiation makes $\|u\|_2 \le \epsilon$ equivalent to $\|\delta\|_2 \le \epsilon \|h_t\|_2$.

Proposition 1 (Local predictive displacement)**.** *Suppose $h_t \neq 0$ and the readout has finite logits and is C^3 in a neighborhood of h_t. As $\epsilon \to 0$,*

$$\max_{\|u\|_2 \le \epsilon} d_{\text{FR}}^2(p_t, \text{softmax}(z(h_t + s_t u))) = \epsilon^2 \lambda_{\max}(A_t) + O(\epsilon^3). \tag{4}$$

The remainder is uniform over the stated local ball.

Expanding the square-root probabilities and maximizing the resulting quadratic form gives Eq. (4); Appendix A provides the proof. LayerNorm and RMSNorm with positive normalization constants satisfy the smoothness assumption. The leading eigenvalue measures the largest local change, whereas the trace sums sensitivities across orthogonal directions. This interpretation concerns the specified coordinates and budget, not arbitrary reparameterizations or realizable input attacks.

We estimate the leading eigenvalue through matrix–vector products. For $u \in \mathbb{R}^d$,

$$\begin{aligned} \Delta z_t &= J_t(s_t u), \\ c_t &= p_t \odot \Delta z_t - p_t(p_t^\top \Delta z_t), \\ A_t u &= s_t J_t^\top c_t. \end{aligned} \tag{5}$$

A readout Jacobian–vector product, probability-weighted centering, and a vector–Jacobian product apply A_t without token sampling, backbone derivatives, or storing J_t, C_t, or F_t.

We use at most $m = 8$ Lanczos steps with full reorthogonalization and a fixed Gaussian starting vector shared across retained positions within each model (Parlett, 1998). The orthonormal basis $Q_{t,m}$ has fewer than m columns if no new direction remains. We compute

$$\widehat{\lambda}_t = \lambda_{\max}(Q_{t,m}^\top A_t Q_{t,m}), \qquad R_{\text{local}} = \operatorname{med}_{t\in\mathcal{T}} \widehat{\lambda}_t. \tag{6}$$

In exact arithmetic, $0 \le \widehat{\lambda}_t \le \lambda_{\max}(A_t)$, with nondecreasing estimates over nested trial spaces. A fixed budget does not guarantee a prescribed error; Proposition 2 gives the bound, and Appendix D.3.2 reports the numerical comparisons across budgets and starting vectors.

3.3 Late-layer trajectory excess

Local sensitivity describes the final prediction, but not its revisions across layers. At each retained answer position, the shared readout gives $(p_{t,1}, \dots, p_{t,K})$ on a common vocabulary. This sequence runs across layers, not generated tokens; its intermediate predictions need not be calibrated (Belrose et al., 2025).

The distance between the first and last predictions can miss intermediate reversals. Total path length captures those changes but also includes the distance needed to connect the endpoints. We subtract that distance to define trajectory excess:

$$\begin{aligned} L_t &= \sum_{k=1}^{K-1} d_{\text{FR}}(p_{t,k}, p_{t,k+1}), \qquad D_t = d_{\text{FR}}(p_{t,1}, p_{t,K}), \\ e_t &= L_t - D_t, \qquad\qquad\qquad R_{\text{path}} = \operatorname{med}_{t\in\mathcal{T}} e_t. \end{aligned} \tag{7}$$

We use raw e_t, without division by $\pi(K-1)$. The triangle inequality and the bound on each segment give $0 \le e_t \le \pi(K-1)$ for $K \ge 3$. An ordered shortest geodesic has zero excess, whereas (p, r, p) has excess $2d_{\text{FR}}(p, r) > 0$ for $p \ne r$ despite identical endpoints.

Inserting a distribution between fixed endpoints cannot decrease raw excess (Proposition 3). We therefore fix K when comparing layer ranges and refit standardization on the same nominal fit set whenever K changes. For $K = 2$, $L_t = D_t$ and $e_t = 0$. Excess measures additional predictive movement, not necessarily incorrect reasoning; its detection value is assessed against local-only scoring and simpler layerwise measures.

3.4 Response scoring and nominal calibration

For analysis, we retain the computed quantities

$$\begin{aligned} d_{t,k} &= d_{\text{FR}}(p_{t,k}, p_{t,k+1}), \\ \mathcal{G}(x, y) &= \big((\widehat{\lambda}_t, d_{t,1}, \dots, d_{t,K-1}, D_t, e_t)\big)_{t\in\mathcal{T}}. \end{aligned} \tag{8}$$

This record requires no additional readout evaluations; scoring uses only R_{local} and R_{path}.

To combine scores with different scales, we standardize them on a nominal reference set $\mathcal{D}_{\text{fit}}$, without filtering by answer correctness. For $j \in \{\text{local}, \text{path}\}$, let μ_j and MAD_j be the medians of R_j and $|R_j - \mu_j|$. With raw MAD and $\sigma_j = \max(\text{MAD}_j, 10^{-8})$, define

$$z_j = \frac{R_j - \mu_j}{\sigma_j}, \qquad R_{\text{geo}} = \sqrt{[z_{\text{local}}]_+^2 + [z_{\text{path}}]_+^2}, \quad [z]_+ = \max(z, 0). \tag{9}$$

The positive part emphasizes values above the reference median, a choice tested by ablation rather than implied by geometry. The score assumes neither independent components nor Gaussian distributions and is not a failure probability. Standardization is fixed before calibration and evaluation. For $K = 2$, $R_{\text{geo}} = [z_{\text{local}}]_+$.

A disjoint nominal set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$ calibrates the alarm threshold. After fixing layer and position rules, the numerical seed, and standardization, we compute the conformal rank (Bates et al., 2023):

$$\widehat{p}(x, y) = \frac{1 + \sum_{i=1}^n \mathbf{1}\{R_{\text{geo}}(x_i, y_i) \ge R_{\text{geo}}(x, y)\}}{n + 1}. \tag{10}$$

We raise an alarm when $\widehat{p} \leq \alpha$ for $\alpha \in [0, 1]$; counting ties with $\geq$ is conservative. Fit, calibration, nominal holdout, and test sets are separated by source image. The experiments specify their sizes; test responses do not determine score parameters.

Theorem 1 (Nominal alarm control)**.** *Conditional on $\mathcal{D}_{\text{fit}}$, suppose the calibration responses and a new nominal response are exchangeable. Assume the same fixed scoring algorithm is well defined and applied to all these responses. Then*

$$\Pr\{\widehat{p}(X, Y) \leq \alpha \mid \mathcal{D}_{\text{fit}}\} \leq \alpha. \tag{11}$$

The probability averages over the calibration set and the new response.

Exchangeability gives the rank argument, including ties, proved in Appendix A. The guarantee controls marginal alarms in the nominal population, which may include incorrect answers. It neither controls false alarms conditional on correctness nor guarantees failure detection under distribution shift.

4 Experiments

We evaluate FRDG across four misbehavior categories and examine which design choices contribute to failure detection. After describing the experimental setup, we compare FRDG with existing monitors and assess the contributions of local sensitivity and trajectory excess. Controlled substitutions test relative-state scaling, the sensitivity statistic, and the subtraction of the first-to-last prediction distance. Numerical accuracy is assessed separately from detection performance.

4.1 Experimental setup

The evaluation records cover Qwen2.5-VL-7B-Instruct, DeepSeek-VL2-Tiny, InternVL3-8B, and LLaVA-OneVision-Qwen2-7B-OV (Bai et al., 2025; Wu et al., 2024; Zhu et al., 2025; Li et al., 2024), abbreviated as Qwen, DeepSeek, InternVL3, and LLaVA. Each model has 6,500 test responses from nine benchmark subsets: POPE, R-Bench, FigStep, HADES, VisualAdv, Typographic, ANDA, PGN, and MMDT. Following the benchmark grouping (Huang et al., 2026), these comprise 2,000 hallucination, 2,800 jailbreak, 400 adversarial, and 1,300 OOD responses. The jailbreak group includes typographic recognition tasks and is therefore not restricted to harmful compliance. Failure is defined by the criterion for the generated answer, not by whether its input is adversarial or OOD.

FRDG retains the configuration described in the original manuscript: four layers at approximately 60%, 75%, 90%, and 100% of model depth and at most four prediction positions per answer. Local sensitivity uses eight Lanczos steps with full reorthogonalization. The main score uses full-vocabulary probabilities and the median raw trajectory excess $L_t - D_t$, without division by the number of layers. Component medians and raw MADs are fitted on 200 reference responses per model, with a scale floor of 10^{-8}. The positive standardized components are combined by their Euclidean norm. Fit and test responses are separated by source image. Reference statistics are fixed before evaluating the test responses. We compare confidence-based scores, sampling-based scores, and internal monitors. These include NLL and LN-PE (Malinin & Gales, 2021), token entropy, self-consistency (SC) (Wang et al., 2023), semantic entropy (SE) (Kuhn et al., 2023; Farquhar et al., 2024), HiddenDetect (Jiang et al., 2025), and EUQ-CF and EUQ-IG (Huang et al., 2026). CF and IG are reported separately. Table 1 presents representative methods from each family. Appendix D.1.1, Tables A3 and A4, retains the complete comparison, including head and state Fisher traces, observed-token gradients, and layerwise JS.

Failure is the positive class. We report AUROC and average precision (AP), and use AURC to assess the risk of retaining answers in increasing score order. Lower AURC is better. Macro summaries weight the four behavior groups equally; they do not denote a separate dataset. All methods are compared on the same responses. Point estimates are recomputed from the response-score records using the recorded 10^{-8} score quantization and grouped ties. Paired intervals reported in the workbook use 2,000 source-image bootstrap replicates. These intervals describe sampling variation in the supplied records; no additional bootstrap run is claimed here.

Table 1: Detection across four behavior groups. Cells report AUROC/AP (%, higher is better). Bold and underlined values mark the best and second-best results across the complete comparison in Appendix D.1.1. SC denotes self-consistency and SE semantic entropy.

Method	Qwen	DeepSeek	InternVL3	LLaVA
Hallucination				
Token entropy	75.90/66.81	77.53/65.90	76.33/74.01	76.06/64.56
SC	69.40/69.81	70.96/66.49	73.41/76.45	69.62/63.38
SE	59.93/57.35	60.92/66.28	78.02/72.89	69.55/60.79
HiddenDetect	78.86/62.74	65.88/58.02	67.58/64.14	73.85/63.70
EUQ-CF	82.08/74.88	78.45/76.14	68.83/75.73	77.14/67.33
EUQ-IG	64.05/67.47	73.23/58.06	77.47/77.74	71.15/65.10
FRDG	**82.98/77.40**	**82.58/78.10**	**78.69/77.95**	**82.47/77.07**
Jailbreak				
Token entropy	77.11/57.94	66.44/43.09	74.80/64.95	74.95/58.25
SC	60.71/64.10	69.18/43.88	67.61/64.74	72.10/59.23
SE	62.58/53.72	69.10/55.06	69.74/68.79	72.00/60.34
HiddenDetect	82.21/61.60	81.99/60.14	69.78/62.10	76.80/57.06
EUQ-CF	62.33/42.17	84.30/62.16	77.48/66.32	77.25/59.04
EUQ-IG	62.95/42.84	69.71/62.90	67.66/63.99	71.65/62.57
FRDG	**84.03/68.44**	**84.53/70.33**	**82.00/79.76**	**81.23/74.39**
Adversarial				
Token entropy	73.45/69.82	69.08/88.94	73.81/53.26	72.10/73.12
SC	70.47/74.15	74.38/86.41	68.74/66.38	70.53/80.07
SE	71.62/64.22	70.65/92.16	69.80/63.81	69.70/74.35
HiddenDetect	79.35/73.77	69.06/80.45	71.90/66.04	72.40/77.13
EUQ-CF	83.21/74.68	86.19/**97.15**	74.39/70.97	80.70/85.59
EUQ-IG	76.11/72.06	**91.05**/92.12	74.00/73.67	82.15/82.44
FRDG	**86.61/90.47**	88.53/94.63	**88.74/87.58**	**88.99/90.90**
OOD				
Token entropy	77.55/91.63	73.30/84.33	66.60/71.58	71.70/70.09
SC	69.73/79.67	62.77/69.93	64.63/63.92	67.60/62.58
SE	65.58/80.90	61.40/76.50	69.83/74.60	65.99/68.19
HiddenDetect	77.38/78.11	71.14/78.47	75.38/70.92	73.44/67.16
EUQ-CF	91.62/**97.58**	77.52/73.71	78.68/73.29	83.64/75.44
EUQ-IG	**91.76**/89.08	**94.19/97.58**	79.08/75.34	86.54/78.22
FRDG	87.71/93.46	91.21/95.68	**87.14/90.25**	**88.86/88.97**

4.2 MISBEHAVIOR DETECTION

Table 1 compares detection across the four behavior groups. FRDG achieves macro AUROCs of 85.33%, 86.71%, 84.15%, and 85.39% on Qwen, DeepSeek, InternVL3, and LLaVA, respectively, with corresponding macro AP values of 82.44%, 84.68%, 83.88%, and 82.83%. Both metrics exceed those of the baseline scores in the complete macro comparison (Appendix D.1.2). The advantage is particularly consistent in the hallucination and jailbreak groups, where FRDG leads the displayed methods on both metrics for all four models. The margins nevertheless vary: for DeepSeek jailbreak responses, the AUROC advantage over EUQ-CF is only 0.23%, whereas the AP advantage over the strongest displayed baseline, EUQ-IG, is 7.43 points. AUROC and AP thus reveal different aspects of the gain, with AP providing additional information about precision when failures are the positive class.

The results are less uniform on adversarial and OOD responses. On Qwen OOD data, EUQ-IG achieves 91.76% AUROC versus 87.71% for FRDG, while EUQ-CF achieves 97.58% AP versus 93.46%. On DeepSeek adversarial responses, EUQ-IG leads AUROC and EUQ-CF leads AP; EUQ-IG also exceeds FRDG on both metrics for DeepSeek OOD responses. In contrast, FRDG leads the

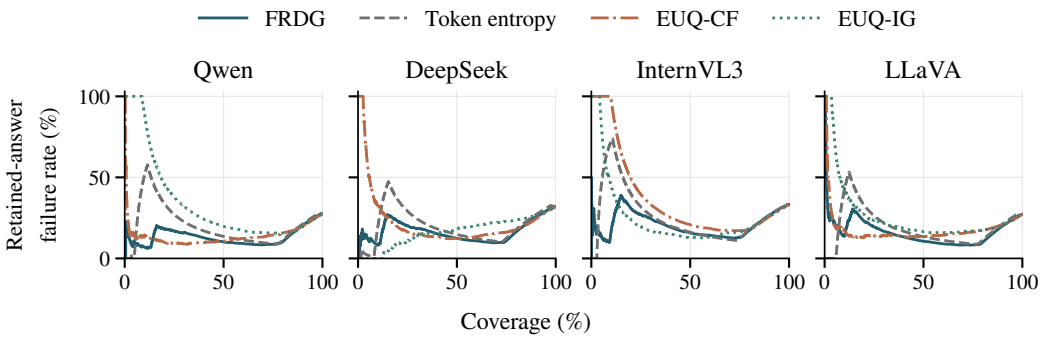

Figure 2: Risk–coverage curves on R-Bench. Coverage is the percentage of answers retained in increasing risk-score order; risk is their failure percentage. Lower curves are better. Plotted-method AURCs and complete source-level comparisons appear in Appendices D.1.4 and D.1.5.

displayed methods on both metrics for the adversarial and OOD groups of InternVL3 and LLaVA. Evidential and geometric scores therefore have different strengths across model–scenario combinations. The macro advantage supports FRDG as a useful monitor across the evaluated groups, rather than a uniformly better detector for every type of failure. Source-level results in Appendix D.1.3 provide the finer breakdown underlying these group summaries.

Figure 2 examines whether the ranking is useful for retaining lower-risk answers on R-Bench. Answers are sorted by increasing risk score; coverage is the fraction retained, and risk is the failure fraction within that retained set. FRDG obtains AURCs of 13.32%, 16.66%, 20.83%, and 16.49% on Qwen, DeepSeek, InternVL3, and LLaVA, respectively. It has the lowest AURC among the plotted methods on three models, while EUQ-IG is slightly lower on DeepSeek at 16.48%. Thus, the ranking advantage extends to average retained-answer risk on several models, but not uniformly to every model or operating point. The curves cross, and at full coverage all methods retain the same answers and have the same failure rate. Appendix D.1.4 reports the plotted-method AURCs, with the complete comparison in Appendix D.1.5.

4.3 Ablation studies and analysis

Table 2 tests whether both measurements contribute to detection. Full exceeds local-only scoring by 0.97–1.39% in macro AUROC across the four models, whereas path-only scoring reaches 66.25%–67.77%. Local sensitivity supplies the stronger standalone signal, but the consistent gain from adding trajectory excess shows that standalone performance does not determine the value of a component in the combined score. This supports joint use in the evaluated setting, without establishing statistical independence or benefits after tightly matching confidence. The combination itself remains open to improvement: fit-median-scaled addition outperforms Full on all four models, although it changes both scaling and aggregation (Appendix D.2.1). Two-sided scoring is weaker than positive-part scoring, but all standardized local values are nonnegative on the test set. This comparison therefore supports clipping the trajectory component in these records, rather than establishing the need for local clipping (Appendix D.2.4).

Removing relative-state scaling lowers macro AUROC by 5.95–7.73%. A fixed absolute perturbation represents different relative changes in states with different norms, so this comparison supports expressing the sensitivity budget relative to state magnitude. State magnitude alone is insufficient: the norm-only control gives AUROCs of 49.79%–52.21% (Appendix D.2.1). The choice of sensitivity statistic has a smaller and less consistent effect. Replacing the leading eigenvalue with the trace raises Qwen AUROC from 85.33% to 85.53%, but lowers it on the other three models; only the DeepSeek paired interval excludes zero in favor of the eigenvalue (Appendix D.2.2). The leading eigenvalue measures the strongest local direction, whereas the trace sums sensitivity across directions. The results favor relative scaling more consistently than they favor either directional summary.

The trajectory substitutions test whether cumulative movement should be measured relative to the separation of the first and last predictions. With the local component retained, replacing trajectory

Table 2: Component and design ablations. Cells report macro AUROC (%). Best and second-best displayed results are bold and underlined. Complete variants and paired intervals: Appendices D.2.1 and D.2.2.

Variant	Qwen	DeepSeek	InternVL3	LLaVA
Full	<u>85.33</u>	**86.71**	**84.15**	**85.39**
Local only	83.94	<u>85.53</u>	83.10	84.42
Path only	67.77	66.89	66.25	67.49
Without relative scaling	79.38	79.28	76.87	77.66
Trace instead of eigenvalue	**85.53**	84.74	<u>84.06</u>	<u>84.61</u>
Local + total path	83.95	84.79	82.97	84.08
Local + first-to-last distance	83.51	85.05	82.95	84.02

Table 3: Numerical accuracy at selected Lanczos budgets. Cells report median relative eigenvalue error, a dimensionless quantity; lower is better. Full numerical results and probability-reconstruction checks appear in Appendices D.3.2 and D.3.3.

Steps	Qwen	DeepSeek	InternVL3	LLaVA
2	3.83×10^{-1}	4.16×10^{-1}	4.39×10^{-1}	3.55×10^{-1}
4	2.12×10^{-2}	2.88×10^{-2}	6.29×10^{-3}	8.98×10^{-3}
8	1.85×10^{-6}	6.68×10^{-7}	1.44×10^{-8}	1.83×10^{-7}

excess with total path length lowers macro AUROC by approximately 1.2–1.9%; using the first-to-last distance instead lowers it by approximately 1.2–1.8%. Total path length includes both the overall prediction change and any additional movement, whereas the first-to-last distance omits intermediate revisions. The advantage of trajectory excess over both alternatives supports measuring movement beyond overall separation. It does not imply that this subtraction identifies erroneous updates: normal refinement can also produce excess movement. The empirical finding is that this particular summary improves failure ranking when combined with local sensitivity.

Table 3 assesses the accuracy of the leading-eigenvalue estimate at different Lanczos budgets. Across the four models, median relative error decreases from 0.35–0.44 at two steps to 0.0063–0.0288 at four steps and falls further at eight steps. The improvement is consistent across models, although the error at a given budget varies. At eight steps, the median error is below 2×10^{-6} for every model, indicating close agreement with the reference eigenvalues in this comparison. These results provide numerical support for the eight-step estimator used to compute the local sensitivity score in FRDG. Appendix D.3.2 reports the full spectral results and checks of the local quadratic approximation.

5 Discussion and Conclusion

Output confidence can remain high for incorrect or unsafe VLM answers, motivating the use of information about how predictions are formed. FRDG measures sensitivity near the final decoder state and accumulated changes across late-layer predictions beyond their first-to-last distance. Fisher–Rao geometry expresses both measurements in terms of changes in token probabilities. Across four VLMs, the reported results show higher macro AUROC and AP than the compared baselines, and combining the measurements improves ranking over either component alone. The ablations also support relative-state scaling and trajectory excess over total path length or the first-to-last distance. However, the leading eigenvalue does not consistently outperform the trace, and alternative score combinations achieve higher AUROC. These findings support the geometric measurements without establishing the chosen score construction as optimal. Future work should identify which failures benefit from layerwise information and develop combination rules that generalize across models.

AI USE STATEMENT

In this work, we used generative AI tools for no tasks requiring mandatory disclosure. We have not used generative AI tools for generating synthetic datasets, developing theoretical models, formulating proofs, designing experiments, implementing methods, or interpreting results, and the remaining mandatory-disclosure categories are not applicable to our use of generative AI. Additionally, we used generative AI tools for a task with recommended disclosure: correcting grammatical errors in the manuscript. We have reviewed all AI-assisted work. Specifically, we checked each grammatical correction to ensure that it preserved the intended meaning and technical content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The evaluation uses existing misbehavior benchmarks, including harmful-instruction examples. The monitor does not generate attacks or claim that a low score makes a response safe. Benchmark execution should follow dataset licenses and restrict access to harmful outputs. No new human-subject study is reported.

REPRODUCIBILITY STATEMENT

We have included the core source code in the supplementary materials. The appendix provides explicit clarifications of all theoretical assumptions alongside complete mathematical proofs. Additionally, comprehensive details regarding the datasets, baseline methods, and implementation procedures are thoroughly documented in the appendix. Appendix A provides the theoretical assumptions and proofs. Appendix B describes decoder state extraction, Fisher operator computation, trajectory distances, and score standardization. Appendix C documents the benchmark composition, reference splits, evaluation metrics, and statistical procedures. Appendix D reports the complete baseline comparisons, source-level results, score variants, and paired intervals. Numerical accuracy is discussed in Appendix D.3.1. The accompanying response-score workbook contains the scores and labels used to reconstruct the reported point estimates.

REFERENCES

Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2): 251–276, February 1998. ISSN 1530-888X. doi: 10.1162/089976698300017746. URL http://dx.doi.org/10.1162/089976698300017746.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, 2025. URL https://arxiv.org/abs/2502.13923.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023. doi: 10.1214/22-AOS2244. URL https://doi.org/10.1214/22-AOS2244.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens, 2025. URL https://arxiv.org/abs/2303.08112v6.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection, 2024. URL https://arxiv.org/abs/2402.03744.

Sam Dauncey, Christopher C. Holmes, Christopher Williams, and Fabian Falck. Approximations to the Fisher Information Metric of Deep Generative Models for Out-Of-Distribution Detection. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=EcuwtinFs9.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. HaloScope: Harnessing unlabeled LLM generations for hallucination detection. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 102948–102972. Curran Associates, Inc., 2024. doi: 10.52202/079017-3270. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ba92705991cfbbcedc26e27e833ebbae-Paper-Conference.pdf.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, 2017. URL https://arxiv.org/abs/1612.00837.

Junjie Hu, Gang Tu, ShengYu Cheng, Jinxin Li, Jinting Wang, Rui Chen, Zhilong Zhou, and Dongbo Shan. HARP: Hallucination detection via reasoning subspace projection, 2025. URL https://arxiv.org/abs/2509.11536.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Gradient Cuff: Detecting jailbreak attacks on large language models by exploring refusal loss landscapes, 2024. URL https://arxiv.org/abs/2403.00867.

Rui Huang, Andrew Geng, and Yixuan Li. On the Importance of Gradients for Detecting Distributional Shifts in the Wild, 2021. URL https://arxiv.org/abs/2110.00218.

Tao Huang, Rui Wang, Xiaofei Liu, Yi Qin, Li Duan, and Liping Jing. Detecting Misbehaviors of Large Vision-Language Models by Evidential Uncertainty Quantification. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xJT4fXJr1Q.

Yilei Jiang, Xinyan Gao, Tianshuo Peng, Yingshui Tan, Xiaoyong Zhu, Bo Zheng, and Xiangyu Yue. HiddenDetect: Detecting jailbreak attacks against multimodal large language models via monitoring hidden states. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14880–14893. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.724. URL https://aclanthology.org/2025.acl-long.724/.

Li Ju, Mayank Nautiyal, Andreas Hellander, Ekta Vats, and Prashant Singh. Epistemic Uncertainty Quantification for Pre-trained VLMs via Riemannian Flow Matching, 2026. URL https://arxiv.org/abs/2601.21662.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.09664.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer, 2024. URL https://arxiv.org/abs/2408.03326.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Andrey Malinin and Mark Gales. Uncertainty Estimation in Autoregressive Structured Prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Anthony Nguyen, Antonin Bertrand, Sylvie Le Hégarat-Mascle, Emanuel Aldea, Franck Florin, Mohammed Nabil El-Korso, and Renaud Lustrat. Fisher-Rao Sensitivity for Out-of-Distribution Detection in Deep Neural Networks. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GEtOzC4MIi.

Frank Nielsen. An Elementary Introduction to Information Geometry. *Entropy*, 22(10):1100, 2020. doi: 10.3390/e22101100. URL https://www.mdpi.com/1099-4300/22/10/1100.

Beresford N. Parlett. *The Symmetric Eigenvalue Problem*. Society for Industrial and Applied Mathematics, January 1998. ISBN 9781611971163. doi: 10.1137/1.9781611971163. URL http://dx.doi.org/10.1137/1.9781611971163.

Marine Picot, Francisco Messina, Malik Boudiaf, Fabrice Labeau, Ismail Ben Ayed, and Pablo Piantanida. Adversarial Robustness via Fisher-Rao Regularization, 2022. URL https://arxiv.org/abs/2106.06685.

Apoorva Sharma, Navid Azizan, and Marco Pavone. Sketching curvature for efficient out-of-distribution detection for deep neural networks. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1958–1967. PMLR, 2021. URL https://proceedings.mlr.press/v161/sharma21a.html.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. LLM-Check: Investigating detection of hallucinations in large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 34188–34216. Curran Associates, Inc., 2024. doi: 10.52202/079017-1077. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3c1e1fdf305195cd620c118aaa9717ad-Paper-Conference.pdf.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models, 2023. URL https://arxiv.org/abs/2203.11171.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding, 2024. URL https://arxiv.org/abs/2412.10302.

Yueqi Xie, Minghong Fang, Renjie Pi, and Neil Gong. GradSafe: Detecting jailbreak prompts for LLMs via safety-critical gradient analysis. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 507–518, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.30. URL https://aclanthology.org/2024.acl-long.30/.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models, 2025. URL https://arxiv.org/abs/2504.10479.

Fisher–Rao Decision Geometry for VLM Misbehavior Monitoring

Appendix

CONTENTS

A Proofs 15
B Implementation Details 16
C Benchmark and Evaluation Details 19
D Additional Experimental Results 22
D.1 Detection and Selective Prediction 22
D.2 Component and Score Design 41
D.3 Numerical Accuracy 59
D.4 Computational Budgets and Cost 65
D.5 Supplementary Threshold Analysis 69

A PROOFS

A.1 THE CATEGORICAL PULLBACK

Let $p(h) = \mathrm{softmax}(z(h))$ and let e_a denote the coordinate vector of token a. Differentiating log-softmax gives $g(a) = J^\top(e_a - p)$. Since $\mathbb{E}_{a\sim p} e_a = p$ and $\mathbb{E}_{a\sim p} e_a e_a^\top = \mathrm{Diag}(p)$, expanding the outer product yields Eq. (3). For every u, $u^\top F u = \mathrm{Var}_{a\sim p}((Ju)_a) \geq 0$. This proof also identifies the null directions: $Fu = 0$ whenever the logit change Ju is constant across vocabulary coordinates. Additional null directions can arise from the readout. The parameter space of a neural network and the state space at one prediction position are different domains for a pullback.

The categorical metric is $\sum_a dp_a^2/p_a$ on the simplex tangent space $\sum_a dp_a = 0$. Its pullback agrees with the score covariance above. Under a smooth change of state coordinates $h = \phi(\eta)$, it transforms as $F_\eta = D\phi^\top F_h D\phi$. A coordinate-dependent perturbation budget must transform as well. Our relative Euclidean budget specifies the allowed perturbations. It does not imply that the eigenvalues of F are invariant under every reparameterization.

A.2 PROOF OF PROPOSITION 1

Write $\psi(h) = 2\sqrt{p(h)}$ componentwise. A finite vocabulary and finite logits give positive probabilities. On a sufficiently small closed neighborhood, the smoothness assumption bounds the required derivatives. Differentiation gives

$$D\psi(h)^\top D\psi(h) = Dp(h)^\top \mathrm{Diag}(p(h))^{-1} Dp(h) = F(h). \tag{12}$$

Taylor expansion therefore gives $\|\psi(h+\delta) - \psi(h)\|_2^2 = \delta^\top F(h)\delta + O(\|\delta\|_2^3)$, uniformly in direction. On a radius-two sphere, a chord of length c and its geodesic distance d obey $c = 4\sin(d/4)$, so $d^2 = c^2 + O(c^4)$. Substituting $\delta = s_t u$ proves the pointwise quadratic approximation with a uniform cubic remainder. Maximizing the quadratic form over $\|u\|_2 \leq \epsilon$ gives $\epsilon^2 \lambda_{\max}(A_t)$; the uniform remainder retains the stated order under maximization.

This argument does not assume a positive-definite Fisher matrix. If the leading eigenvalue vanishes, the leading term is zero. A second-order expansion under weaker smoothness can still have an $o(\|\delta\|^2)$ remainder, but a uniform cubic remainder requires stronger derivative bounds. We therefore state a C^3 assumption.

A.3 SPECTRAL APPROXIMATION

Proposition 2 (Rayleigh–Ritz approximation)**.** *For a symmetric positive semidefinite A_t and an orthonormal $Q_{t,m}$, the estimate in Eq.* (6) *satisfies $0 \leq \widehat{\lambda}_t \leq \lambda_{\max}(A_t)$ in exact arithmetic. It is nondecreasing over nested trial spaces. Let $\lambda_1 \geq \cdots \geq \lambda_d$ be the eigenvalues of A_t, and let v_1 be a unit leading eigenvector. If $c = \|Q_{t,m}^\top v_1\|_2 > 0$, then*

$$0 \leq \lambda_1 - \widehat{\lambda}_t \leq (\lambda_1 - \lambda_d)(1 - c^2). \tag{13}$$

For an orthonormal Q, the Rayleigh–Ritz principle gives

$$\lambda_{\max}(Q^\top A Q) = \max_{w \in \mathrm{span}(Q),\, \|w\|=1} w^\top A w \leq \lambda_1(A). \tag{14}$$

Positive semidefiniteness supplies the lower bound. Enlarging the trial space cannot reduce this maximum. Let $P = QQ^\top$ and $c = \|Pv_1\| > 0$. The unit vector $w = Pv_1/c$ satisfies $v_1^\top w = c$. Expand the component orthogonal to the leading eigenvector in an eigenbasis of A to obtain $w^\top A w \geq c^2\lambda_1 + (1 - c^2)\lambda_d$. The Ritz maximum is at least this value, proving Eq. (13).

For a unit Ritz vector $\widehat{v}$ with value ρ, let $r = A\widehat{v} - \rho\widehat{v}$. Expansion in an orthonormal eigenbasis gives $\|r\|^2 = \sum_i (\lambda_i - \rho)^2 |v_i^\top \widehat{v}|^2$. Consequently, $\min_i |\lambda_i - \rho| \leq \|r\|$. The minimum ranges over all eigenvalues. It cannot be replaced by $|\lambda_1 - \rho|$ without an additional argument. In particular, an exact nonleading eigenvector has zero residual. We do not interpret a residual threshold as a safety certificate or change the iteration budget without reporting it.

A.4 TRAJECTORY PROPERTIES

Proposition 3 (Path excess). *For $K \geq 2$, the raw excess $e_t = L_t - D_t$ in Eq. (7) satisfies $0 \leq e_t \leq \pi(K-1)$. The raw trajectory excess $L_t - D_t$ cannot decrease when a point is inserted between fixed endpoints. It is zero on an ordered shortest geodesic and is unchanged by a common permutation of vocabulary coordinates. For $K = 2$, it is identically zero.*

Repeated application of the triangle inequality gives $D_t \leq L_t$. Every adjacent distance is at most π, hence $0 \leq L_t - D_t \leq \pi(K-1)$. Inserting r between p_i and p_{i+1} replaces one segment by two whose lengths sum to at least the original distance. The endpoint distance is unchanged. An ordered shortest geodesic has additive segment lengths and therefore zero excess. A common vocabulary permutation preserves every square-root inner product. If $K = 2$, the only adjacent segment is the endpoint distance and cancels exactly.

For an example of positive excess, take distinct positive distributions p, r and the sequence (p, r, p). Its excess is $2d_{\mathrm{FR}}(p, r)$ although the endpoint distance is zero. Conversely, a path from p to r along an ordered shortest geodesic has zero excess even if the endpoint change is large. These facts justify comparing excess against both endpoint distance and total path length rather than treating the three definitions as equivalent.

A.5 PROOF OF THEOREM 1

Condition on the fitted score, including its fixed numerical seed and hyperparameters. The n calibration scores and the new nominal score are exchangeable. If scores are distinct, the descending rank of the new score is uniform on $\{1, \ldots, n+1\}$. Equation (10) is this rank divided by $n+1$, so the alarm probability equals $\lfloor \alpha(n+1) \rfloor / (n+1) \leq \alpha$. Counting ties with $\geq$ assigns the most conservative rank within a tie group and cannot increase the alarm probability. This is a marginal statement over the calibration set and the new response, not a statement conditional on one realized calibration sample.

The score may combine dependent components and may include a finite-step numerical approximation. Neither invalidates the rank argument if the scoring algorithm is fixed and applied symmetrically. Reusing calibration responses to choose layers, score weights, or standardization statistics breaks this simple conditioning argument. A shifted deployment distribution need not be exchangeable with the nominal calibration sample.

A.6 NON-IDENTIFIABILITY OF SEMANTIC FAILURE

Predictive geometry does not determine response correctness. Consider a constant readout $p(h) = p_0$ along all selected states. It has zero Fisher sensitivity and zero trajectory excess. The generated answer may nevertheless be incorrect for an input or inconsistent with a safety requirement. Changing the external reference answer or policy label leaves the internal geometry unchanged. No distribution-free detection theorem can exclude this example without assumptions relating model predictions to those labels.

Final confidence also does not determine sensitivity. In a binary readout $p_1(h) = \mathrm{sigmoid}(a(h-1))$, the state $h = 1$ has $p_1 = 1/2$ for every a, but its Fisher value is $a^2/4$. This is an existence argument across smooth readouts, not evidence that a particular VLM realizes the example. Similarly, two layer paths can end at the same state and have different excess lengths. These quantities describe distinct mathematical properties. Statistical complementarity still requires empirical measurement.

B IMPLEMENTATION DETAILS

B.1 MODEL INTERFACE AND CAUSAL REPLAY

FRDG accesses decoder block outputs and the frozen final readout. The evaluated model identifiers are `Qwen/Qwen2.5-VL-7B-Instruct`, `deepseek-ai/deepseek-vl2-tiny`, `OpenGVLab/InternVL3-8B`, and `lmms-lab/llava-onevision-qwen2-7b-ov`. The response scored by each monitor is held fixed. A causal replay processes the image, prompt, and saved response without differentiating through the backbone. At answer position t, the retained

state predicts y_t from the observed prefix $y_{<t}$. The first answer token therefore uses the final prompt position.

Image placeholders can expand into several internal tokens. Prediction positions must follow the sequence after this expansion rather than offsets in the unexpanded prompt. The final state is taken before the native final normalization. Applying that normalization twice would change both the prediction and its derivatives. Reconstructed final logits should therefore be compared with the native replay logits before scoring. The readout uses the native gain, bias, and normalization constant.

The reference configuration specifies evaluation mode, disabled dropout, greedy decoding, and a 128-token generation cap. The available result summaries do not separately verify every decoding parameter. These reference settings should not be interpreted as independently checked metadata for every reported response.

B.2 Layer and position selection

The main configuration selects $K = 4$ block outputs at relative depths $r_k \in \{0.60, 0.75, 0.90, 1.00\}$. For a decoder with L blocks, the stated selection rule is $\ell_k = \lfloor r_k L + 1/2 \rfloor$, using one-based indices and requiring distinct blocks. The final block is included. Changing the layer count changes the distribution of raw trajectory excess. We therefore refit the median and raw MAD on the same nominal fitting responses for each value of K, then freeze them for calibration and evaluation. When $K = 2$, the trajectory excess is identically zero and the score reduces to the positive-part local score.

The available InternVL3 records contain an unresolved indexing discrepancy. The configuration reference lists 28 decoder blocks, which gives indices 17, 21, 25, and 28 under the stated rule. The experiment export instead lists 19, 24, 29, and 32. These records cannot both identify the same 28-block decoder. We do not replace either record with an inferred checkpoint configuration. Depth-specific results for this model require the actual checkpoint and block-index mapping to be reconciled.

Eligible answer positions exclude tokenizer-declared special tokens. If N_e eligible positions remain, the main configuration retains $T' = \min(4, N_e)$ positions. For $T' > 1$, the zero-based eligible-position indices are

$$i_j = \left\lfloor \frac{j(N_e - 1)}{T' - 1} \right\rfloor, \qquad j = 0, \ldots, T' - 1. \tag{15}$$

A response with one eligible position uses that position. A response with no eligible positions is unscorable rather than assigned an arbitrary numerical score. The one-position uniform ablation uses a single selected position; its precise selection within a longer response is not specified by the result summary and should not be inferred from Eq. (15).

B.3 Readout derivatives and Fisher products

The readout is $z(h) = W\nu(h) + b$, with Jacobian $J = WD\nu(h)$. Let $D_\gamma = \mathrm{Diag}(\gamma)$ denote the learned normalization gain. For RMSNorm with native constant $\varepsilon_\nu > 0$, define $r = (h^\top h/d + \varepsilon_\nu)^{1/2}$. Then

$$\nu(h) = D_\gamma h/r, \qquad D\nu(h) = D_\gamma \left(\frac{I}{r} - \frac{hh^\top}{dr^3} \right). \tag{16}$$

For LayerNorm, let $P = I - \mathbf{1}\mathbf{1}^\top/d$, $c = Ph$, and $r = (c^\top c/d + \varepsilon_\nu)^{1/2}$. With normalization bias β,

$$\nu(h) = D_\gamma c/r + \beta, \qquad D\nu(h) = D_\gamma \left(\frac{P}{r} - \frac{cc^\top}{dr^3} \right). \tag{17}$$

These products require vector operations rather than dense normalization Jacobians. For the transpose product, the learned gain is applied before the symmetric term. The output bias affects probabilities but has zero derivative with respect to the state.

The local operator uses the original full-vocabulary distribution $p_t = \mathrm{softmax}(z(h_t))$. For $s_t = \|h_t\|_2$, it is

$$A_t = s_t^2 J_t^\top \left(\mathrm{Diag}(p_t) - p_t p_t^\top \right) J_t. \tag{18}$$

The scale s_t is held fixed during differentiation. A product with A_t consists of a readout Jacobian–vector product, probability weighting and centering, and a vector–Jacobian product. Model weights remain frozen. Probability reductions use at least FP32, with FP64 comparisons for the controlled checks. A near-zero state norm requires an explicit numerical check rather than silent rescaling.

B.4 Lanczos estimation and trajectory distances

The main score uses at most eight Lanczos steps with full reorthogonalization. The reference implementation reuses a normalized Gaussian starting vector across positions within each model and seed. Each new vector is orthogonalized twice against the existing basis. The projected operator is symmetrized before its largest eigenvalue is computed. The recorded monitoring breakdown tolerance is 10^{-7} relative to the operator-product scale; the small double-precision checks use 10^{-12}. The realized basis dimension is retained when breakdown occurs. Stored operator products also give the Ritz residual without an additional readout product. A small residual establishes proximity to some eigenvalue, not necessarily the largest one.

The primary trajectory score uses original full-vocabulary softmax probabilities at every selected layer. For normalized distributions p and q, the Fisher–Rao distance can be evaluated as

$$d_{\mathrm{FR}}(p, q) = 4 \operatorname{atan2}(\|\sqrt{p} - \sqrt{q}\|_2, \|\sqrt{p} + \sqrt{q}\|_2) \,. \tag{19}$$

This is equivalent to $2 \arccos(\sum_v \sqrt{p_v q_v})$ and avoids cancellation near identical distributions. Stable full-vocabulary log-softmax is used before computing probabilities. The response path score is the median of $e_t = L_t - D_t$, without division by $\pi(K-1)$. A negative excess beyond floating-point tolerance indicates an implementation error rather than a valid path statistic.

B.5 Exported probability profiles

Exported profiles retain eight active token IDs at each answer position and a tail mass. The active set $\mathcal{A}_t$ is shared by all selected layers at that position. It may differ between positions and responses. The uniform-tail reconstruction is

$$\widetilde{p}_v = \begin{cases} p_v, & v \in \mathcal{A}_t, \\ \tau_p/(V-8), & v \notin \mathcal{A}_t, \end{cases} \qquad \tau_p = 1 - \sum_{v \in \mathcal{A}_t} p_v. \tag{20}$$

For two reconstructed distributions with the same active set,

$$\sum_{v=1}^{V} \sqrt{\widetilde{p}_v \widetilde{q}_v} = \sum_{v \in \mathcal{A}_t} \sqrt{p_v q_v} + \sqrt{\tau_p \tau_q}. \tag{21}$$

Equation (21) is exact for the reconstructed distributions. It is not generally exact for the original softmax distributions. Cauchy–Schwarz gives

$$\sum_{v \notin \mathcal{A}_t} \sqrt{p_v q_v} \le \sqrt{\tau_p \tau_q}, \qquad d_{\mathrm{FR}}(\widetilde{p}, \widetilde{q}) \le d_{\mathrm{FR}}(p, q). \tag{22}$$

When both tail masses are positive, equality requires identical conditional tail distributions. A common uniform tail is sufficient but not necessary. Both total path length and endpoint distance can decrease after reconstruction, so their difference has no corresponding one-sided bound.

Other checks must use the same representation consistently. For example,

$$\|\widetilde{p} - \widetilde{q}\|_2^2 = \sum_{v \in \mathcal{A}_t} (p_v - q_v)^2 + \frac{(\tau_p - \tau_q)^2}{V - 8}. \tag{23}$$

Using an unweighted ninth coordinate would not give the same Euclidean distance. JS on the uniform-tail reconstruction can instead be evaluated on the eight active coordinates and one aggregated tail coordinate. Exported-profile checks establish normalization and distance identities for this representation; they do not reproduce the original full-probability trajectory exactly. Neither the main trajectory scores nor the local Fisher operator is replaced by this reconstruction.

Table A1: Test composition. The last four columns count failed answers, not attack inputs.

Source	N	Qwen	DeepSeek	InternVL3	LLaVA
POPE	1000	271	299	313	248
R-Bench	1000	275	319	332	272
FigStep	200	15	25	43	34
HADES	200	20	24	61	35
VisualAdv	600	80	65	191	99
Typographic	1800	218	209	507	297
ANDA	200	108	158	79	86
PGN	200	109	165	87	87
MMDT	1300	939	856	709	659

B.6 Reference splits and computational cost

Each model uses 200 nominal responses for fitting, 200 for calibration, and 400 for holdout evaluation. These sets and the benchmark test set are separated by source image. The fitted location is the median, and the scale is the raw MAD with floor 10^{-8}. No factor of 1.4826 is applied. The score combines positive standardized deviations with a Euclidean norm. Behavior labels do not select components, standardization parameters, or thresholds.

The reference nominal sampling protocol uses unmodified VQAv2 training questions and images (Goyal et al., 2017), with benchmark images excluded and one question per selected image. Nominal membership is not defined by whether the VLM answers correctly. The exchangeability condition concerns this nominal query population, not arbitrary mixtures of benchmark categories.

For a linear output head, each Fisher product requires two W-sized matrix–vector operations and $O(V + d)$ vector work. Processing T' positions with m Lanczos steps and K layers has additional cost

$$O(T'mVd + T'KVd + T'm^2d), \tag{24}$$

besides response generation and causal replay. Processing positions sequentially requires temporary storage $O(V + md + Kd)$, excluding frozen weights and replay workspace. Any higher-precision readout copy contributes to memory use. Matrix-free products avoid dense Fisher storage but do not remove derivative computation. Appendix D.4.3 distinguishes these computational requirements from device measurements; no measured-latency advantage is claimed here.

C Benchmark and Evaluation Details

C.1 Benchmark composition

Table A1 gives the source counts and failed-answer counts used in the current evaluation. The benchmark scenarios and grouping follow the original manuscript. Failure labels refer to generated answers: an adversarial or OOD input does not automatically make its answer a failure. POPE and R-Bench form the hallucination group; FigStep, HADES, VisualAdv, and Typographic form the jailbreak group; ANDA and PGN form the adversarial group; MMDT supplies the OOD group.

The original protocol uses closed-answer evaluation for object, relationship, recognition, and other benchmark questions, and a harmful-compliance evaluator for open-ended safety answers. This revision retains the recorded labels; it does not supply a new annotation-agreement study. Typographic contributes 1,800 of the 2,800 answers in the jailbreak group. Its recognition failures should not be interpreted as harmful compliance. Source-level results are therefore reported separately.

C.2 Reference splits

The fit split supplies the component medians and raw MADs used for standardization. These statistics are fixed before test evaluation. Calibration and holdout splits are used only in the supplementary threshold analysis in Appendix D.5.1; they are not additional components of FRDG. Recorded

Table A2: Reference population. The last four columns report failures within each nominal split. Nominal membership does not imply answer correctness.

Split	N	Qwen	DeepSeek	InternVL3	LLaVA
fit	200	32	46	40	42
calibration	200	42	32	35	41
holdout	400	66	65	82	83

source-image identifiers are disjoint across the four splits. Table A2 retains the recorded failure counts rather than treating reference-set membership as a correctness label.

C.3 Baselines and controlled comparisons

The output-based comparisons include response NLL, LN-PE, and token entropy. SC and SE use repeated generations to assess agreement (Wang et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024). These additional generations are separate from the primary answer whose failure label is evaluated. The reference protocol specifies ten answers for repeated-generation baselines. Sampling temperature and other settings are not recoverable from the result columns alone, so we do not infer undocumented settings from the reported scores.

EUQ-CF and EUQ-IG measure conflict and ignorance and are reported separately (Huang et al., 2026). HiddenDetect provides an internal-monitor comparison (Jiang et al., 2025). Its evaluation outside the original safety setting is an extension of that setting, not a claim that the original method was designed for every behavior group. All methods rank the same primary responses within each model and dataset.

The geometric controls distinguish the domain of perturbation from the statistic used to summarize it. For output-head parameters, the categorical Fisher trace is

$$\left(\|\nu(h)\|_2^2 + \mathbf{1}_{\text{bias}}\right)\left(1 - \|p\|_2^2\right), \tag{25}$$

where the bias term is included only when bias parameters are present. This head-parameter statistic differs from the state Fisher operator used by FRDG. It is a controlled comparison, not a complete reproduction of other gradient-based detectors (Nguyen et al., 2026; Huang et al., 2021). The state-trace comparison replaces the leading eigenvalue with a trace statistic. The reference protocol uses matched-product random probes; a probe estimate should not be confused with an exact trace. Observed-token gradients condition on the generated token rather than averaging over the predictive distribution. Layerwise JS uses the same selected predictions.

Each design comparison retains the primary answers and nominal partitions. Removing relative-state scaling replaces A_t with F_t. Local-only and path-only scores remove one component, while total-path and endpoint alternatives change the trajectory statistic. The two-sided standardized norm removes positive-part clipping. The fit-median-scaled additive score changes both scaling and combination and therefore does not isolate the effect of MAD normalization. Changes to layer count require separately fitted nominal parameters. Related parameter-space methods such as SCOD and generative Fisher scores are discussed in the paper without assigning them unimplemented response-level baselines (Sharma et al., 2021; Dauncey et al., 2024).

C.4 Metric definitions and verification

For N evaluated answers, let S_k contain the k lowest-scoring answers and let $b_i \in \{0, 1\}$ indicate failure. At an attainable retention threshold,

$$c_k = \frac{k}{N}, \qquad r_k = \frac{1}{k}\sum_{i \in S_k} b_i, \qquad k > 0. \tag{26}$$

The figure displays $100c_k$ and $100r_k$. Tied scores enter together, so not every k is necessarily attainable. Risk is undefined at zero coverage, and no observed point is plotted there. At full coverage, risk equals the failure prevalence of the evaluated answers.

AUROC gives half credit to positive–negative score ties. AP sums precision at each distinct descending threshold, weighted by its increase in recall. AURC uses ascending score groups: if a group increases coverage by Δc, its contribution is Δc times the failure rate among all records retained through that group. This specifies the treatment of ties and is not an arbitrary trapezoidal interpolation of a plotting grid. Macro values are the equal average of the four behavior-level values.

All 2,128 available main and ablation metric rows were recomputed from the response-score sheet, including AUROC, AP, and AURC. The largest discrepancy was below 10^{-12}. Scores are quantized to 10^{-8} before tied thresholds are grouped. Bootstrap intervals are retained as reported rather than independently recalculated in this revision. A paired difference is always computed as Full minus the named comparator; a positive AUROC or AP difference favors Full, whereas a negative AURC difference favors Full.

C.5 Scope of controlled comparisons

Local-only and path-only variants assess component contributions. Relative-state scaling and the trace substitution assess the local measurement. Total path length and first-to-last distance assess the trajectory definition. Changing clipping or the standardization rule assesses score construction. These comparisons should not be pooled into a single claim that every alternative is a partial version of Full. Complete method results appear in Tables A3 and A4, and the score variants appear in Appendix D.2.1.

The stratified comparisons condition on recorded entropy, answer length, or first-to-last distance. They are exploratory and are not equivalent to tightly matching answers on local sensitivity or confidence. Small strata can produce unstable AUROC and AP estimates; strata without both label classes do not support an AUROC comparison. These analyses are not used to claim a causal relation between internal sensitivity, prediction revisions, and failure.

C.6 Additional statistical conventions

Writing c_j and r_j for coverage and retained-answer failure rate after the jth score group, the discrete area is

$$\mathrm{AURC} = \sum_{j=1}^{M} (c_j - c_{j-1}) r_j, \qquad c_0 = 0. \tag{27}$$

Lower area does not imply lower risk at every coverage level. The reported bootstrap procedure uses 2,000 paired source-image resamples. Its intervals describe resampling uncertainty, not variation across model training runs, and are not adjusted for simultaneous comparisons. This revision reproduces point metrics but retains the supplied intervals.

For supplementary threshold analysis, the nominal rank rule is

$$\widehat{p} = \frac{1 + \sum_{i=1}^{n_{\mathrm{cal}}} \mathbf{1}\{R_i^{\mathrm{cal}} \geq R\}}{n_{\mathrm{cal}} + 1}, \qquad \text{alarm if } \widehat{p} \leq \alpha. \tag{28}$$

Each fixed score has its own calibration threshold. An exchangeability argument concerns marginal alarms in the reference population, not false positives conditional on correctness. The observed rates and their denominators are reported in Appendices D.5.1 and D.5.2.

D Additional Experimental Results

This appendix provides the complete baseline comparisons, component and design substitutions, paired intervals, and configuration analyses supporting Section 4. All numerical values are taken from the current workbook rather than the historical manuscript tables. The revision checks score consistency but does not independently establish experimental provenance. Numerical agreement alone should not be interpreted as independent validation of model measurements.

D.1 Detection and Selective Prediction

These experiments compare failure ranking across methods, behavior groups, and data sources, and evaluate the risk of retaining answers by score.

D.1.1 Complete behavior comparisons

FRDG has the strongest four-group macro AUROC and AP among the baseline monitors, but the behavior-level tables show that this advantage is not uniform. On Qwen OOD, EUQ-IG has higher AUROC and EUQ-CF has higher AP. EUQ also remains competitive on DeepSeek adversarial and OOD responses. These differences motivate reporting separate behavior groups rather than using the macro average as evidence of dominance in every setting.

The following two tables retain all methods in the main-results sheet. Head and state Fisher traces, token gradients, and layerwise JS are control scores; their names do not imply reproduction of a separate published detector. Best and second-best values are determined separately for AUROC and AP after rounding to two decimals. Model groups are separated from method rows by full horizontal rules.

Table A3: Complete baseline comparison for Qwen and DeepSeek. Cells report AUROC/AP (%, higher is better). Bold and underlined values indicate the best and second-best rounded results within each model, behavior group, and metric.

Method	Halluc.	Jailbreak	Adversarial	OOD
Qwen2.5-VL-7B-Instruct				
NLL	63.47/41.95	67.19/22.55	68.24/71.30	71.83/86.22
LN-PE	64.01/63.44	73.24/66.15	72.80/66.42	78.31/90.10
Token entropy	75.90/66.81	77.11/57.94	73.45/69.82	77.55/91.63
SC	69.40/69.81	60.71/64.10	70.47/74.15	69.73/79.67
SE	59.93/57.35	62.58/53.72	71.62/64.22	65.58/80.90
HiddenDetect	78.86/62.74	82.21/61.60	79.35/73.77	77.38/78.11
EUQ-CF	82.08/74.88	62.33/42.17	83.21/74.68	91.62/**97.58**
EUQ-IG	64.05/67.47	62.95/42.84	76.11/72.06	**91.76**/89.08
Head Fisher trace	71.77/50.00	73.30/26.90	73.92/76.60	72.42/85.43
State Fisher trace	74.06/52.57	75.04/30.01	76.54/77.64	75.20/86.99
Token gradient	72.77/52.92	74.46/32.60	78.27/81.04	72.20/86.09
Layerwise JS	72.12/49.61	73.47/29.08	73.68/78.89	74.57/86.80
FRDG	**82.98/77.40**	**84.03/68.44**	**86.61/90.47**	87.71/93.46
DeepSeek-VL2-Tiny				
NLL	70.14/50.33	63.76/19.16	70.40/90.32	66.21/78.62
LN-PE	76.58/62.99	69.67/44.76	78.63/84.86	72.04/82.34
Token entropy	77.53/65.90	66.44/43.09	69.08/88.94	73.30/84.33
SC	70.96/66.49	69.18/43.88	74.38/86.41	62.77/69.93
SE	60.92/66.28	69.10/55.06	70.65/92.16	61.40/76.50
HiddenDetect	65.88/58.02	81.99/60.14	69.06/80.45	71.14/78.47
EUQ-CF	78.45/76.14	84.30/62.16	86.19/**97.15**	77.52/73.71
EUQ-IG	73.23/58.06	69.71/62.90	**91.05**/92.12	**94.19/97.58**
Head Fisher trace	70.95/52.01	72.84/27.81	73.13/90.50	72.33/82.27
State Fisher trace	73.59/57.04	74.94/32.17	75.14/92.01	74.95/84.48
Token gradient	71.64/53.24	74.13/30.77	77.62/92.90	72.06/82.08
Layerwise JS	71.34/52.58	73.08/28.71	72.15/90.68	73.61/83.57
FRDG	**82.58/78.10**	**84.53/70.33**	88.53/94.63	91.21/95.68

Table A4: Complete baseline comparison for InternVL3 and LLaVA. Cells report AUROC/AP (%, higher is better). Bold and underlined values indicate the best and second-best rounded results within each model, behavior group, and metric.

Method	Halluc.	Jailbreak	Adversarial	OOD
InternVL3-8B				
NLL	69.40/53.56	68.31/46.78	66.40/60.71	65.20/67.58
LN-PE	76.42/74.62	75.44/63.26	73.50/60.81	66.48/65.17
Token entropy	76.33/74.01	74.80/64.95	73.81/53.26	66.60/71.58
SC	73.41/76.45	67.61/64.74	68.74/66.38	64.63/63.92
SE	78.02/72.89	69.74/68.79	69.80/63.81	69.83/74.60
HiddenDetect	67.58/64.14	69.78/62.10	71.90/66.04	75.38/70.92
EUQ-CF	68.83/75.73	77.48/66.32	74.39/70.97	78.68/73.29
EUQ-IG	77.47/77.74	67.66/63.99	74.00/73.67	79.08/75.34
Head Fisher trace	70.96/52.22	72.40/52.75	72.33/67.91	71.26/74.10
State Fisher trace	72.45/57.60	74.47/55.85	74.48/65.60	73.84/75.53
Token gradient	71.09/55.38	73.04/51.82	76.43/67.19	71.60/72.84
Layerwise JS	70.34/52.97	73.31/53.97	72.29/65.42	73.35/76.16
FRDG	**78.69/77.95**	**82.00/79.76**	**88.74/87.58**	**87.14/90.25**
LLaVA-OneVision-Qwen2-7B-OV				
NLL	68.59/45.54	68.36/31.36	68.86/61.61	67.19/66.78
LN-PE	73.35/62.12	75.75/60.63	74.55/69.82	71.89/69.25
Token entropy	76.06/64.56	74.95/58.25	72.10/73.12	71.70/70.09
SC	69.62/63.38	72.10/59.23	70.53/80.07	67.60/62.58
SE	69.55/60.79	72.00/60.34	69.70/74.35	65.99/68.19
HiddenDetect	73.85/63.70	76.80/57.06	72.40/77.13	73.44/67.16
EUQ-CF	77.14/67.33	77.25/59.04	80.70/85.59	83.64/75.44
EUQ-IG	71.15/65.10	71.65/62.57	82.15/82.44	86.54/78.22
Head Fisher trace	71.80/48.47	72.85/37.64	73.09/66.35	72.45/71.68
State Fisher trace	73.69/51.62	75.24/39.15	76.53/72.15	75.14/74.35
Token gradient	72.40/46.92	74.14/36.96	78.03/73.53	72.37/71.45
Layerwise JS	72.15/47.57	74.04/39.93	73.06/66.67	74.22/74.25
FRDG	**82.47/77.07**	**81.23/74.39**	**88.99/90.90**	**88.86/88.97**

D.1.2 MACRO SUMMARIES

The macro results summarize performance across the four behavior groups without allowing the largest group to dominate the average. FRDG improves on the strongest baseline macro AUROC by approximately 5.53, 4.67, 9.30, and 5.70 percentage points on Qwen, DeepSeek, InternVL3, and LLaVA, respectively. The corresponding AP margins are approximately 10.11, 7.02, 11.20, and 10.75 points, with the strongest comparator selected separately for each metric. These are absolute differences, not relative percentage improvements. The intervals below describe uncertainty in the reported metrics; paired differences are the appropriate comparison for two methods evaluated on the same responses.

Macro results assign equal weight to the four behavior groups. Table A5 reports the complete comparison; macro denotes an aggregation rule rather than a separate dataset.

Table A5: Four-group macro averages in the supplied records. Cells report AUROC/AP (%). Macro is an aggregation rule, not a dataset.

Method	Qwen	DeepSeek	InternVL3	LLaVA
NLL	67.69/55.51	67.63/59.61	67.33/57.16	68.25/51.32
LN-PE	72.09/71.53	74.23/68.74	72.96/65.97	73.88/65.45
Token entropy	76.00/71.55	71.59/70.56	72.89/65.95	73.70/66.50
SC	67.58/71.93	69.32/66.68	68.60/67.88	69.97/66.32
SE	64.93/64.05	65.52/72.50	71.85/70.02	69.31/65.92
HiddenDetect	79.45/69.06	72.02/69.27	71.16/65.80	74.12/66.26
EUQ-CF	<u>79.81</u>/<u>72.33</u>	81.62/77.29	<u>74.85</u>/71.58	<u>79.68</u>/71.85
EUQ-IG	73.72/67.86	<u>82.04</u>/<u>77.66</u>	74.55/<u>72.69</u>	77.87/<u>72.08</u>
Head Fisher trace	72.85/59.73	72.31/63.15	71.74/61.74	72.55/56.04
State Fisher trace	75.21/61.80	74.65/66.43	73.81/63.64	75.15/59.32
Token gradient	74.43/63.16	73.86/64.75	73.04/61.81	74.24/57.22
Layerwise JS	73.46/61.10	72.55/63.88	72.32/62.13	73.37/57.10
FRDG	**85.33/82.44**	**86.71/84.68**	**84.15/83.88**	**85.39/82.83**

Table A6: Qwen macro metrics and reported 95% bootstrap intervals. AUROC and AP are higher-is-better; AURC is lower-is-better. All values are percentages. Intervals are those supplied in the workbook. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
NLL	67.69 [65.85, 69.43]	55.51 [53.43, 57.80]	31.33 [29.51, 33.45]
LN-PE	72.09 [69.90, 74.05]	71.53 [69.22, 73.97]	31.55 [29.88, 33.44]
Token entropy	76.00 [74.02, 78.00]	71.55 [69.43, 74.32]	31.66 [29.84, 33.61]
SC	67.58 [65.32, 69.77]	71.93 [69.79, 74.16]	33.79 [31.89, 35.78]
SE	64.93 [62.65, 67.13]	64.05 [61.72, 66.69]	34.59 [32.88, 36.59]
HiddenDetect	79.45 [77.57, 81.25]	69.06 [66.76, 71.67]	**22.36** [20.94, 23.99]
EUQ-CF	<u>79.81</u> [78.10, 81.59]	<u>72.33</u> [70.15, 74.94]	26.76 [25.15, 28.48]
EUQ-IG	73.72 [71.74, 75.65]	67.86 [65.61, 70.41]	29.41 [27.70, 31.25]
Head Fisher trace	72.85 [71.09, 74.50]	59.73 [57.61, 62.09]	28.01 [26.36, 30.02]
State Fisher trace	75.21 [73.62, 76.81]	61.80 [59.63, 64.06]	26.48 [24.84, 28.34]
Token gradient	74.43 [72.77, 76.09]	63.16 [61.04, 65.50]	27.47 [25.83, 29.39]
Layerwise JS	73.46 [71.82, 75.15]	61.10 [59.11, 63.33]	28.05 [26.24, 30.08]
FRDG	**85.33** [83.84, 86.83]	**82.44** [80.44, 84.59]	<u>24.59</u> [22.85, 26.66]

Table A7: DeepSeek macro metrics and reported 95% bootstrap intervals. AUROC and AP are higher-is-better; AURC is lower-is-better. All values are percentages. Intervals are those supplied in the workbook. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
NLL	67.63 [65.61, 69.66]	59.61 [57.96, 61.45]	36.95 [34.95, 38.99]
LN-PE	74.23 [71.52, 76.81]	68.74 [66.63, 71.07]	32.20 [30.30, 34.26]
Token entropy	71.59 [69.38, 73.73]	70.56 [68.55, 72.90]	40.27 [38.30, 42.12]
SC	69.32 [66.55, 71.91]	66.68 [64.66, 68.83]	37.16 [35.26, 39.17]
SE	65.52 [63.38, 67.56]	72.50 [70.64, 74.31]	46.23 [44.30, 48.09]
HiddenDetect	72.02 [69.17, 74.74]	69.27 [67.11, 71.47]	34.62 [32.64, 36.69]
EUQ-CF	81.62 [80.05, 83.24]	77.29 [75.64, 78.92]	33.11 [31.24, 34.93]
EUQ-IG	<u>82.04</u> [79.90, 84.03]	<u>77.66</u> [75.72, 79.73]	<u>30.01</u> [28.09, 31.99]
Head Fisher trace	72.31 [70.28, 74.36]	63.15 [61.34, 65.14]	35.22 [33.10, 37.25]
State Fisher trace	74.65 [72.71, 76.42]	66.43 [64.62, 68.47]	33.83 [31.83, 35.88]
Token gradient	73.86 [71.89, 75.74]	64.75 [63.05, 66.81]	34.04 [32.09, 36.11]
Layerwise JS	72.55 [70.54, 74.51]	63.88 [62.10, 65.91]	35.02 [32.95, 36.98]
FRDG	**86.71** [84.94, 88.38]	**84.68** [82.74, 86.64]	**28.79** [26.81, 30.86]

Table A8: InternVL3 macro metrics and reported 95% bootstrap intervals. AUROC and AP are higher-is-better; AURC is lower-is-better. All values are percentages. Intervals are those supplied in the workbook. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
NLL	67.33 [65.58, 69.05]	57.16 [54.84, 59.87]	27.69 [25.82, 29.69]
LN-PE	72.96 [71.18, 74.58]	65.97 [63.69, 68.43]	24.54 [23.01, 26.16]
Token entropy	72.89 [71.09, 74.77]	65.95 [63.85, 68.46]	27.96 [26.33, 29.63]
SC	68.60 [66.71, 70.51]	67.88 [65.69, 70.02]	30.36 [28.57, 32.29]
SE	71.85 [70.06, 73.50]	70.02 [67.70, 72.24]	28.20 [26.51, 30.02]
HiddenDetect	71.16 [69.37, 72.84]	65.80 [63.46, 68.10]	27.92 [26.24, 29.65]
EUQ-CF	<u>74.85</u> [73.10, 76.59]	71.58 [69.48, 73.78]	25.71 [24.06, 27.50]
EUQ-IG	74.55 [72.81, 76.28]	<u>72.69</u> [70.66, 74.71]	28.06 [26.11, 30.03]
Head Fisher trace	71.74 [70.13, 73.37]	61.74 [59.49, 64.34]	25.02 [23.36, 26.90]
State Fisher trace	73.81 [72.18, 75.44]	63.64 [61.26, 66.23]	<u>22.99</u> [21.48, 24.67]
Token gradient	73.04 [71.46, 74.58]	61.81 [59.55, 64.69]	23.70 [22.16, 25.41]
Layerwise JS	72.32 [70.62, 73.98]	62.13 [59.71, 64.73]	24.18 [22.60, 25.95]
FRDG	**84.15** [82.71, 85.56]	**83.88** [81.94, 85.80]	**21.58** [19.90, 23.19]

Table A9: LLaVA macro metrics and reported 95% bootstrap intervals. AUROC and AP are higher-is-better; AURC is lower-is-better. All values are percentages. Intervals are those supplied in the workbook. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
NLL	68.25 [66.62, 69.89]	51.32 [48.87, 54.09]	22.71 [21.08, 24.41]
LN-PE	73.88 [72.19, 75.57]	65.45 [63.09, 67.86]	21.50 [19.91, 23.04]
Token entropy	73.70 [71.72, 75.77]	66.50 [64.15, 69.31]	25.76 [23.72, 27.78]
SC	69.97 [67.87, 72.18]	66.32 [64.32, 68.48]	25.75 [23.77, 27.67]
SE	69.31 [67.36, 71.41]	65.92 [63.66, 68.15]	29.74 [27.57, 31.65]
HiddenDetect	74.12 [72.23, 75.86]	66.26 [64.15, 68.50]	23.37 [21.45, 25.18]
EUQ-CF	<u>79.68</u> [77.87, 81.42]	71.85 [69.84, 73.89]	20.13 [18.21, 21.93]
EUQ-IG	77.87 [76.18, 79.56]	<u>72.08</u> [70.02, 74.29]	21.40 [19.69, 23.17]
Head Fisher trace	72.55 [70.95, 74.20]	56.04 [53.62, 58.76]	20.54 [18.92, 22.31]
State Fisher trace	75.15 [73.51, 76.75]	59.32 [56.94, 61.88]	<u>19.00</u> [17.53, 20.54]
Token gradient	74.24 [72.69, 75.80]	57.22 [54.91, 59.80]	19.77 [18.14, 21.38]
Layerwise JS	73.37 [71.74, 74.92]	57.10 [54.71, 59.79]	19.70 [18.25, 21.28]
FRDG	**85.39** [83.82, 86.84]	**82.83** [81.01, 84.68]	**17.01** [15.44, 18.68]

D.1.3 SOURCE-LEVEL FULL-SCORE RESULTS

The source-level tables test whether a pooled advantage is shared across the constituent benchmarks. AP must be interpreted together with failure prevalence: a high AP on a failure-heavy source is not directly comparable to the same AP on a source with few failures. Accordingly, best and second-best marks compare methods within the same model, source, and metric. The full-score interval tables compare different sources and are intentionally not ranked across rows.

Table A10 separates the benchmark sources within each behavior group. AP differences across sources reflect both score ordering and differences in failure prevalence.

Table A10: Source-level results for the full score. Cells report AUROC/AP (%). Changes in AP across sources also reflect differences in failure prevalence.

Source	Qwen	DeepSeek	InternVL3	LLaVA
POPE	82.56/77.86	82.41/77.10	78.82/78.22	83.35/77.02
R-Bench	83.33/76.96	82.73/79.09	78.57/77.79	81.65/77.41
FigStep	79.68/60.45	84.57/74.85	79.50/67.51	77.66/65.80
HADES	85.50/72.12	80.54/66.04	80.95/79.40	81.32/74.58
VisualAdv	84.65/70.33	84.32/68.09	83.24/83.76	82.90/77.83
Typographic	84.07/68.71	84.99/71.20	82.10/80.13	81.08/74.31
ANDA	86.55/90.69	90.90/95.16	89.06/89.65	89.15/90.34
PGN	86.93/90.93	86.29/94.65	88.31/85.70	88.76/91.67
MMDT	87.71/93.46	91.21/95.68	87.14/90.25	88.86/88.97

The following tables extend the full-score summaries to every baseline and source. Their source-level values should not be confused with pooled behavior metrics.

Table A11: Qwen source-level comparisons. Cells report AUROC/AP (%). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	ANDA	FigStep	HADES
NLL	63.56/66.77	61.15/29.61	62.69/21.15
LN-PE	69.05/63.55	71.10/64.63	86.72/**80.44**
Token entropy	74.25/70.66	75.86/53.03	75.44/52.99
SC	69.67/73.07	**86.74/87.67**	64.94/67.17
SE	69.78/61.67	74.92/66.70	80.67/68.97
HiddenDetect	77.45/74.49	79.32/56.75	**89.67**/67.58
EUQ-CF	80.04/69.77	61.41/41.83	68.08/37.19
EUQ-IG	73.51/72.07	61.73/40.72	81.19/73.40
Head Fisher trace	75.02/76.89	72.20/16.04	71.82/18.52
State Fisher trace	76.79/79.43	73.86/17.46	74.04/27.72
Token gradient	78.73/78.93	73.17/27.09	72.99/30.42
Layerwise JS	74.79/78.98	72.45/29.64	71.96/25.49
FRDG	**86.55/90.69**	79.68/60.45	85.50/72.12

Table A12: Qwen source-level comparisons. Cells report AUROC/AP (%). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	MMDT	PGN	POPE
NLL	71.83/86.22	72.83/76.84	61.24/39.09
LN-PE	78.31/90.10	76.72/70.45	62.23/61.49
Token entropy	77.55/91.63	72.65/69.47	76.50/67.19
SC	69.73/79.67	71.23/75.39	66.65/67.56
SE	65.58/80.90	73.35/66.89	60.97/57.93
HiddenDetect	77.38/78.11	80.75/74.20	79.40/62.13
EUQ-CF	91.62/**97.58**	86.28/82.54	**83.16**/76.04
EUQ-IG	**91.76**/89.08	78.43/72.92	65.57/67.78
Head Fisher trace	72.42/85.43	72.75/76.52	72.35/50.43
State Fisher trace	75.20/86.99	76.19/76.13	74.51/53.45
Token gradient	72.20/86.09	77.91/82.92	73.30/53.76
Layerwise JS	74.57/86.80	72.94/79.20	72.95/52.45
FRDG	87.71/93.46	**86.93/90.93**	82.56/**77.86**

Table A13: Qwen source-level comparisons. Cells report AUROC/AP (%). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	R-Bench	Typographic	VisualAdv
NLL	65.86/44.81	67.37/21.66	68.35/27.81
LN-PE	65.84/65.41	71.52/63.14	74.88/**71.63**
Token entropy	75.30/66.61	77.25/59.56	77.65/61.34
SC	72.13/72.05	59.64/63.17	57.68/62.06
SE	58.88/56.80	58.65/50.66	66.45/56.95
HiddenDetect	78.41/63.54	82.76/63.90	79.70/56.77
EUQ-CF	81.03/73.85	62.45/42.50	60.82/44.53
EUQ-IG	62.58/67.22	62.23/41.09	60.97/41.10
Head Fisher trace	71.19/49.83	73.39/28.87	73.64/29.95
State Fisher trace	73.62/51.89	75.14/31.58	75.61/31.46
Token gradient	72.15/51.68	74.53/32.88	74.86/36.74
Layerwise JS	71.40/47.00	73.80/29.64	73.94/32.58
FRDG	**83.33/76.96**	**84.07/68.71**	**84.65**/70.33

Table A14: Qwen source-level FRDG results with reported 95% intervals (%).

Source	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
ANDA	86.55 [81.00, 91.61]	90.69 [85.62, 94.88]	28.62 [21.70, 36.77]
FigStep	79.68 [60.88, 95.50]	60.45 [34.51, 84.32]	3.87 [0.58, 9.16]
HADES	85.50 [71.06, 97.89]	72.12 [49.83, 91.40]	3.48 [0.63, 7.26]
MMDT	87.71 [85.59, 89.84]	93.46 [91.27, 95.64]	48.20 [44.53, 52.21]
PGN	86.93 [81.29, 92.04]	90.93 [85.01, 95.59]	33.26 [24.31, 42.29]
POPE	82.56 [78.79, 86.41]	77.86 [72.01, 83.64]	15.15 [11.88, 18.63]
R-Bench	83.33 [79.50, 87.04]	76.96 [71.35, 83.17]	13.32 [10.50, 16.40]
Typographic	84.07 [79.93, 87.98]	68.71 [62.10, 75.96]	4.84 [3.54, 6.24]
VisualAdv	84.65 [77.81, 91.00]	70.33 [59.30, 81.74]	4.96 [2.83, 7.46]

Table A15: DeepSeek source-level comparisons. Cells report AUROC/AP (%). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	ANDA	FigStep	HADES
NLL	66.31/88.51	60.16/24.37	68.25/25.40
LN-PE	80.95/84.94	64.37/46.65	82.81/51.15
Token entropy	69.85/88.59	65.19/43.90	64.80/42.41
SC	76.03/85.86	65.10/39.59	75.08/56.89
SE	75.81/93.05	65.42/48.82	68.30/52.82
HiddenDetect	67.04/78.32	80.57/54.17	**88.78/69.32**
EUQ-CF	88.58/**97.47**	82.19/61.15	85.27/58.02
EUQ-IG	88.35/89.77	69.19/63.92	73.13/63.16
Head Fisher trace	74.37/89.80	71.67/32.30	71.13/39.37
State Fisher trace	76.18/91.55	73.98/34.07	72.76/43.50
Token gradient	78.61/92.74	72.83/34.73	71.77/24.75
Layerwise JS	74.28/90.65	71.74/27.29	72.15/27.68
FRDG	**90.90**/95.16	**84.57/74.85**	80.54/66.04

Table A16: DeepSeek source-level comparisons. Cells report AUROC/AP (%). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	MMDT	PGN	POPE
NLL	66.21/78.62	74.22/91.78	70.63/49.97
LN-PE	72.04/82.34	75.77/85.04	76.41/61.74
Token entropy	73.30/84.33	68.24/89.59	78.15/65.79
SC	62.77/69.93	72.29/86.94	71.42/67.65
SE	61.40/76.50	63.90/91.20	59.47/63.46
HiddenDetect	71.14/78.47	71.48/82.96	63.32/54.66
EUQ-CF	77.52/73.71	83.91/**96.95**	76.90/73.59
EUQ-IG	**94.19/97.58**	**94.29**/94.91	72.24/54.46
Head Fisher trace	72.33/82.27	71.92/91.19	71.58/52.88
State Fisher trace	74.95/84.48	75.32/93.51	74.28/56.91
Token gradient	72.06/82.08	76.44/93.14	72.17/52.44
Layerwise JS	73.61/83.57	72.25/91.87	71.98/54.00
FRDG	91.21/95.68	86.29/94.65	**82.41/77.10**

Table A17: DeepSeek source-level comparisons. Cells report AUROC/AP (%). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	R-Bench	Typographic	VisualAdv
NLL	69.66/51.03	64.97/19.49	59.48/16.60
LN-PE	76.73/64.41	70.41/46.61	64.28/37.88
Token entropy	76.95/66.21	66.60/43.79	67.00/42.86
SC	70.49/65.30	69.16/44.77	68.99/38.45
SE	62.26/68.84	68.29/56.12	73.69/56.86
HiddenDetect	68.31/61.48	80.52/58.65	84.89/64.73
EUQ-CF	79.99/78.65	84.14/62.35	**85.29**/65.22
EUQ-IG	74.24/62.22	69.13/62.16	70.31/65.17
Head Fisher trace	70.31/51.47	72.81/26.61	74.16/29.74
State Fisher trace	72.78/57.29	75.15/31.95	75.08/31.00
Token gradient	71.16/54.61	74.39/34.13	74.56/25.69
Layerwise JS	70.63/51.75	73.27/28.65	73.37/32.43
FRDG	**82.73/79.09**	**84.99/71.20**	84.32/**68.09**

Table A18: DeepSeek source-level FRDG results with reported 95% intervals (%).

Source	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
ANDA	90.90 [83.90, 96.43]	95.16 [89.90, 99.16]	52.11 [43.35, 61.97]
FigStep	84.57 [72.66, 95.18]	74.85 [56.35, 90.41]	4.34 [1.34, 8.20]
HADES	80.54 [65.51, 92.98]	66.04 [45.19, 84.54]	5.71 [1.87, 10.92]
MMDT	91.21 [89.45, 92.76]	95.68 [94.20, 96.82]	38.14 [34.75, 41.73]
PGN	86.29 [76.58, 94.06]	94.65 [89.62, 98.67]	59.85 [50.85, 69.34]
POPE	82.41 [78.76, 86.02]	77.10 [71.64, 82.60]	16.23 [12.98, 19.59]
R-Bench	82.73 [79.00, 86.13]	79.09 [73.82, 84.37]	16.66 [13.66, 20.03]
Typographic	84.99 [80.60, 89.08]	71.20 [64.15, 78.51]	4.43 [3.15, 5.78]
VisualAdv	84.32 [76.68, 91.10]	68.09 [56.11, 79.67]	5.30 [2.39, 8.41]

Table A19: InternVL3 source-level comparisons. Cells report AUROC/AP (%). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	ANDA	FigStep	HADES
NLL	65.60/63.24	68.12/45.73	69.67/51.12
LN-PE	75.41/58.73	74.60/<u>60.54</u>	73.03/62.00
Token entropy	74.65/52.56	73.60/59.05	73.20/65.36
SC	61.58/56.85	59.11/55.27	63.32/58.44
SE	65.70/54.33	67.13/60.03	62.91/65.27
HiddenDetect	70.31/60.31	68.86/56.44	63.65/58.97
EUQ-CF	75.93/73.41	<u>75.23</u>/53.81	<u>74.78</u>/66.49
EUQ-IG	75.10/<u>73.98</u>	66.27/50.51	72.89/<u>71.70</u>
Head Fisher trace	72.90/65.75	71.27/48.60	70.60/54.83
State Fisher trace	76.05/69.10	72.29/51.36	72.24/52.35
Token gradient	<u>77.46</u>/69.57	72.32/41.90	71.19/54.21
Layerwise JS	73.15/64.02	72.19/42.49	70.89/57.30
FRDG	**89.06/89.65**	**79.50/67.51**	**80.95/79.40**

Table A20: InternVL3 source-level comparisons. Cells report AUROC/AP (%). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	MMDT	PGN	POPE
NLL	65.20/67.58	67.14/59.32	68.54/51.22
LN-PE	66.48/65.17	71.38/62.93	76.07/73.53
Token entropy	66.60/71.58	73.05/54.62	<u>76.95</u>/73.61
SC	64.63/63.92	75.30/<u>74.92</u>	73.14/75.07
SE	69.83/74.60	73.54/71.76	76.59/72.13
HiddenDetect	75.38/70.92	73.41/72.65	63.58/59.87
EUQ-CF	78.68/73.29	73.12/69.08	70.08/<u>76.72</u>
EUQ-IG	<u>79.08</u>/75.34	73.19/73.88	75.29/75.06
Head Fisher trace	71.26/74.10	71.84/70.30	71.61/52.96
State Fisher trace	73.84/75.53	74.14/64.19	73.17/58.77
Token gradient	71.60/72.84	<u>75.40</u>/66.63	71.66/54.14
Layerwise JS	73.35/<u>76.16</u>	72.63/70.31	71.42/53.21
FRDG	**87.14/90.25**	**88.31/85.70**	**78.82/78.22**

Table A21: InternVL3 source-level comparisons. Cells report AUROC/AP (%). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	R-Bench	Typographic	VisualAdv
NLL	70.17/56.05	67.75/45.80	68.96/49.87
LN-PE	76.75/75.73	75.97/63.30	74.44/64.97
Token entropy	75.75/74.58	75.00/65.08	75.40/68.67
SC	73.64/77.71	69.16/66.63	66.78/64.51
SE	79.39/73.71	70.78/69.46	69.76/70.79
HiddenDetect	71.29/68.07	70.91/61.43	68.73/66.83
EUQ-CF	67.69/74.90	78.31/67.14	76.91/67.24
EUQ-IG	**79.53/80.17**	68.50/64.42	64.26/64.39
Head Fisher trace	70.42/51.87	72.76/51.80	72.82/57.88
State Fisher trace	72.16/56.76	74.70/55.88	75.00/60.72
Token gradient	70.60/57.24	73.26/51.89	73.95/56.35
Layerwise JS	69.36/52.93	73.70/54.01	73.37/57.68
FRDG	78.57/77.79	**82.10/80.13**	**83.24/83.76**

Table A22: InternVL3 source-level FRDG results with reported 95% intervals (%).

Source	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
ANDA	89.06 [83.46, 94.30]	89.65 [83.03, 94.93]	16.36 [10.86, 22.84]
FigStep	79.50 [68.12, 88.97]	67.51 [51.76, 83.08]	12.47 [6.25, 19.53]
HADES	80.95 [71.30, 89.93]	79.40 [68.57, 89.60]	19.29 [10.76, 27.66]
MMDT	87.14 [84.97, 89.21]	90.25 [87.53, 92.75]	31.38 [27.86, 34.83]
PGN	88.31 [82.51, 93.72]	85.70 [76.80, 94.09]	20.57 [13.77, 29.34]
POPE	78.82 [74.79, 83.08]	78.22 [73.20, 83.43]	20.50 [16.88, 24.43]
R-Bench	78.57 [74.64, 82.34]	77.79 [72.88, 82.43]	20.83 [17.48, 24.50]
Typographic	82.10 [79.16, 84.95]	80.13 [76.26, 83.79]	15.56 [13.26, 18.03]
VisualAdv	83.24 [78.37, 87.71]	83.76 [78.19, 88.72]	17.25 [12.71, 21.92]

Table A23: LLaVA source-level comparisons. Cells report AUROC/AP (%). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	ANDA	FigStep	HADES
NLL	70.68/64.03	58.26/27.10	63.53/28.24
LN-PE	72.81/66.43	74.18/56.18	78.10/61.99
Token entropy	72.90/73.70	73.71/57.88	73.30/58.63
SC	71.35/80.62	72.22/58.89	73.78/63.45
SE	70.53/77.88	69.29/56.30	70.67/54.14
HiddenDetect	77.28/81.21	74.34/54.34	78.01/56.16
EUQ-CF	79.35/83.83	**83.75/70.34**	73.56/61.41
EUQ-IG	83.03/84.19	69.97/62.57	74.91/70.18
Head Fisher trace	74.22/65.36	72.21/32.23	71.28/32.62
State Fisher trace	77.64/74.35	74.11/34.97	73.55/50.72
Token gradient	79.04/75.32	72.56/37.55	72.13/36.27
Layerwise JS	73.70/63.13	72.97/37.43	72.35/43.29
FRDG	**89.15/90.34**	77.66/65.80	**81.32/74.58**

Table A24: LLaVA source-level comparisons. Cells report AUROC/AP (%). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	MMDT	PGN	POPE
NLL	67.19/66.78	67.03/59.45	68.96/46.07
LN-PE	71.89/69.25	76.17/73.34	75.30/61.52
Token entropy	71.70/70.09	71.31/73.09	76.70/64.08
SC	67.60/62.58	69.73/79.73	70.28/63.23
SE	65.99/68.19	68.66/70.99	72.80/62.04
HiddenDetect	73.44/67.16	67.60/73.07	73.99/62.92
EUQ-CF	83.64/75.44	<u>82.03</u>/<u>87.34</u>	<u>79.22</u>/<u>69.64</u>
EUQ-IG	<u>86.54</u>/<u>78.22</u>	81.46/80.78	73.85/66.88
Head Fisher trace	72.45/71.68	72.57/70.09	72.24/47.55
State Fisher trace	75.14/74.35	75.78/71.38	74.01/51.39
Token gradient	72.37/71.45	76.77/72.36	73.21/46.62
Layerwise JS	74.22/74.25	72.88/71.44	72.88/46.70
FRDG	**88.86/88.97**	**88.76/91.67**	**83.35/77.02**

Table A25: LLaVA source-level comparisons. Cells report AUROC/AP (%). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	R-Bench	Typographic	VisualAdv
NLL	68.12/45.09	69.67/32.11	69.42/34.46
LN-PE	71.59/63.01	76.33/61.63	73.65/59.19
Token entropy	<u>75.50</u>/<u>65.36</u>	75.10/58.42	75.50/59.14
SC	69.06/63.74	70.86/58.45	75.14/60.33
SE	66.49/59.87	73.67/63.35	68.27/54.77
HiddenDetect	73.74/64.65	77.15/56.99	<u>76.11</u>/<u>60.34</u>
EUQ-CF	75.23/65.35	<u>78.18</u>/59.85	73.44/51.90
EUQ-IG	68.64/63.58	74.03/<u>64.81</u>	63.93/53.70
Head Fisher trace	71.48/49.58	72.93/38.45	73.59/42.43
State Fisher trace	73.40/52.10	75.40/38.57	75.60/40.02
Token gradient	71.74/47.27	74.13/35.28	75.38/43.59
Layerwise JS	71.44/48.51	74.38/39.65	74.07/42.79
FRDG	**81.65/77.41**	**81.08/74.31**	**82.90/77.83**

Table A26: LLaVA source-level FRDG results with reported 95% intervals (%).

Source	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
ANDA	89.15 [83.03, 94.44]	90.34 [83.95, 95.41]	20.43 [13.76, 28.71]
FigStep	77.66 [65.36, 88.62]	65.80 [48.73, 81.45]	8.82 [4.43, 14.37]
HADES	81.32 [69.10, 91.94]	74.58 [59.56, 87.75]	9.17 [3.70, 16.35]
MMDT	88.86 [86.76, 90.89]	88.97 [85.85, 92.02]	25.33 [22.33, 28.48]
PGN	88.76 [82.72, 94.12]	91.67 [86.45, 95.80]	20.27 [14.05, 27.85]
POPE	83.35 [79.32, 87.29]	77.02 [70.72, 83.18]	11.57 [8.97, 14.43]
R-Bench	81.65 [77.64, 85.59]	77.41 [71.89, 83.15]	16.49 [13.05, 19.99]
Typographic	81.08 [77.20, 84.87]	74.31 [68.87, 79.55]	8.80 [6.81, 11.09]
VisualAdv	82.90 [76.06, 89.16]	77.83 [69.05, 85.87]	8.08 [5.03, 11.61]

D.1.4 SELECTIVE PREDICTION ON R-BENCH

On R-Bench, FRDG has the lowest AURC among the four plotted methods on Qwen, InternVL3, and LLaVA. DeepSeek is an exception: EUQ-IG achieves 16.48% compared with 16.66% for

FRDG. The curves therefore support improved retention on three models, not a uniform advantage at every model or coverage level.

Table A27 reports AURC for the four monitors plotted in Figure 2. Coverage and retained-answer risk are defined in Appendix C.4. Integration uses the grouped-threshold convention stated there, not a trapezoidal approximation of a downsampled plot. The zero-coverage endpoint has no observed risk value; the first nonempty score group supplies the first contribution to the discrete AURC sum. The plot is shown only for R-Bench and does not by itself establish the same ordering on other sources.

Table A27: R-Bench AURC for the monitors in Figure 2 (%, lower is better). Best and second-best displayed values are bold and underlined within each model.

Method	Qwen	DeepSeek	InternVL3	LLaVA
FRDG	**13.32**	<u>16.66</u>	**20.83**	**16.49**
Token entropy	19.82	19.98	25.84	18.78
EUQ-CF	<u>14.14</u>	22.53	36.58	<u>17.76</u>
EUQ-IG	34.10	**16.48**	<u>25.53</u>	26.04

D.1.5 SELECTIVE PREDICTION ACROSS ALL SOURCES

These tables extend the R-Bench illustration to all nine sources and all baseline scores. AURC evaluates whether failures are rejected early as coverage decreases; it complements AUROC and AP rather than reproducing their rankings. Comparisons are made within each source because both the endpoint risk and the attainable area depend on failure prevalence. Bold entries identify the smallest areas, not necessarily curves that lie below every competing curve at all coverage levels.

Coverage is the fraction of answers retained after rejecting high-risk answers. Retained-answer failure rate is the fraction of failures among those retained. Smaller AURC is better. The following tables report all baseline methods rather than only the four curves shown in the main text. Definitions and tie handling are in Appendix C.4.

Table A28: Qwen source-level AURC (%, lower is better). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	ANDA	FigStep	HADES	MMDT	PGN
NLL	44.22	6.10	6.48	56.58	37.20
LN-PE	32.96	9.41	5.38	50.88	29.69
Token entropy	34.62	5.61	6.87	63.64	35.82
SC	42.05	3.19	9.66	57.81	42.48
SE	32.12	8.30	8.46	56.90	31.02
HiddenDetect	28.67	4.76	**1.92**	<u>45.31</u>	<u>27.59</u>
EUQ-CF	**26.80**	10.32	10.59	52.38	**24.71**
EUQ-IG	31.78	11.41	8.45	**39.69**	28.98
Head Fisher trace	33.71	4.04	4.17	56.15	37.26
State Fisher trace	32.90	**2.90**	4.46	53.90	32.53
Token gradient	31.17	<u>3.04</u>	4.58	56.44	35.23
Layerwise JS	37.21	4.06	4.44	54.01	38.19
FRDG	<u>28.62</u>	3.87	<u>3.48</u>	48.20	33.26

Table A29: Qwen source-level AURC (%, lower is better). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	POPE	R-Bench	Typographic	VisualAdv
NLL	21.48	20.01	7.50	8.49
LN-PE	31.57	30.60	13.59	13.82
Token entropy	18.99	19.82	8.59	9.28
SC	23.74	21.25	12.99	15.13
SE	32.63	33.60	18.48	16.68
HiddenDetect	**10.82**	**11.56**	5.17	<u>5.06</u>
EUQ-CF	<u>12.32</u>	14.14	15.99	17.96
EUQ-IG	31.28	34.10	15.24	17.20
Head Fisher trace	14.49	15.72	5.44	6.17
State Fisher trace	14.36	14.72	<u>4.97</u>	5.16
Token gradient	14.60	15.39	5.27	6.01
Layerwise JS	15.20	15.51	5.26	5.87
FRDG	15.15	<u>13.32</u>	**4.84**	**4.96**

Table A30: DeepSeek source-level AURC (%, lower is better). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	ANDA	FigStep	HADES	MMDT	PGN
NLL	68.17	9.36	7.06	53.86	69.00
LN-PE	52.99	15.67	3.70	46.91	60.28
Token entropy	72.48	12.74	12.56	55.32	77.87
SC	61.44	9.55	7.08	52.73	67.95
SE	70.53	16.40	13.61	58.07	80.60
HiddenDetect	58.35	4.80	**2.77**	44.96	61.58
EUQ-CF	64.30	**4.11**	<u>3.09</u>	<u>38.07</u>	72.77
EUQ-IG	**50.40**	14.90	12.13	**37.63**	**53.97**
Head Fisher trace	64.22	7.04	5.98	50.04	70.56
State Fisher trace	62.44	6.09	5.50	46.71	70.33
Token gradient	62.20	6.29	5.51	49.35	65.97
Layerwise JS	62.45	7.24	5.35	47.84	72.35
FRDG	<u>52.11</u>	<u>4.34</u>	5.71	38.14	<u>59.85</u>

Table A31: DeepSeek source-level AURC (%, lower is better). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	POPE	R-Bench	Typographic	VisualAdv
NLL	16.96	19.33	7.04	7.88
LN-PE	**13.95**	**15.25**	10.73	11.64
Token entropy	18.11	19.98	11.57	10.68
SC	22.56	24.54	7.87	7.15
SE	37.89	38.51	14.16	10.41
HiddenDetect	30.08	28.98	4.66	**3.00**
EUQ-CF	21.74	22.53	**3.49**	<u>3.05</u>
EUQ-IG	<u>15.65</u>	<u>16.48</u>	14.70	13.39
Head Fisher trace	16.82	19.18	5.53	4.53
State Fisher trace	16.06	17.84	5.08	5.09
Token gradient	17.01	18.66	5.06	4.61
Layerwise JS	18.27	18.41	5.26	4.96
FRDG	16.23	16.66	<u>4.43</u>	5.30

Table A32: InternVL3 source-level AURC (%, lower is better). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	ANDA	FigStep	HADES	MMDT	PGN
NLL	30.03	13.25	19.97	41.87	29.83
LN-PE	19.22	11.31	17.50	36.63	24.65
Token entropy	<u>18.89</u>	16.07	21.84	46.65	<u>21.73</u>
SC	34.61	19.82	25.81	42.27	30.08
SE	26.54	23.13	35.04	38.06	26.14
HiddenDetect	23.80	21.55	32.17	30.62	26.49
EUQ-CF	23.69	<u>10.44</u>	<u>16.39</u>	**28.68**	26.51
EUQ-IG	24.43	20.72	28.05	<u>28.71</u>	32.11
Head Fisher trace	23.85	12.32	19.50	38.89	28.79
State Fisher trace	21.29	11.16	**16.05**	35.97	23.69
Token gradient	21.90	**10.02**	19.68	37.48	23.95
Layerwise JS	22.18	12.46	18.03	36.36	27.17
FRDG	**16.36**	12.47	19.29	31.38	**20.57**

Table A33: InternVL3 source-level AURC (%, lower is better). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	POPE	R-Bench	Typographic	VisualAdv
NLL	19.88	21.18	18.59	19.82
LN-PE	24.39	25.43	14.32	17.28
Token entropy	23.41	25.84	19.84	22.54
SC	23.22	24.93	21.78	25.78
SE	20.96	<u>19.52</u>	26.86	31.37
HiddenDetect	32.99	28.95	22.87	29.33
EUQ-CF	33.41	36.58	**13.32**	**15.79**
EUQ-IG	26.31	25.53	27.64	34.68
Head Fisher trace	<u>17.93</u>	19.98	15.36	17.70
State Fisher trace	18.53	**18.56**	<u>14.25</u>	16.89
Token gradient	**17.62**	20.92	14.71	<u>16.76</u>
Layerwise JS	18.33	20.81	15.05	18.44
FRDG	20.50	20.83	15.56	17.25

Table A34: LLaVA source-level AURC (%, lower is better). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	ANDA	FigStep	HADES	MMDT	PGN
NLL	27.26	13.31	11.48	36.52	28.99
LN-PE	24.33	11.87	11.91	30.71	<u>23.55</u>
Token entropy	35.79	12.83	14.20	36.71	37.10
SC	36.10	11.04	10.94	35.48	37.79
SE	40.21	14.50	16.90	39.25	39.75
HiddenDetect	33.34	9.60	**7.46**	28.60	41.33
EUQ-CF	32.13	8.01	10.55	<u>22.88</u>	32.04
EUQ-IG	<u>22.68</u>	17.09	17.14	**21.45**	24.90
Head Fisher trace	25.03	8.14	9.56	32.87	28.74
State Fisher trace	22.72	**7.15**	9.04	31.34	24.20
Token gradient	22.95	<u>7.99</u>	<u>8.77</u>	33.29	25.48
Layerwise JS	24.52	8.84	9.23	31.43	26.48
FRDG	**20.43**	8.82	9.17	25.33	**20.27**

Table A35: LLaVA source-level AURC (%, lower is better). Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Method	POPE	R-Bench	Typographic	VisualAdv
NLL	15.85	16.91	9.39	9.61
LN-PE	15.78	21.91	12.58	12.06
Token entropy	16.26	18.78	12.16	12.01
SC	18.51	21.00	11.15	9.55
SE	20.60	27.17	15.03	17.44
HiddenDetect	18.26	20.18	7.88	8.46
EUQ-CF	15.15	17.76	8.64	9.57
EUQ-IG	20.98	26.04	15.74	19.74
Head Fisher trace	13.61	15.45	7.87	7.92
State Fisher trace	13.20	**14.45**	**7.29**	<u>7.38</u>
Token gradient	<u>12.66</u>	<u>14.93</u>	7.80	**7.03**
Layerwise JS	13.03	15.29	<u>7.52</u>	7.75
FRDG	**11.57**	16.49	8.80	8.08

D.1.6 SOURCE WEIGHTING

A discrepancy between pooled and source-equal metrics indicates that the aggregate conclusion depends partly on how sources are combined. Source-equal averaging removes unequal source weights but also removes cross-source response comparisons from each metric. This distinction is particularly relevant for the jailbreak group, which contains several sources with different failure criteria and prevalence.

Pooled metrics rank all responses in a behavior group together. Source-equal metrics average the source-specific metrics within that group. Their difference reflects both weighting and cross-source ranking; it cannot be attributed to sample counts alone.

Table A36: Qwen FRDG pooled and source-equal metrics (%). Differences and intervals are source-equal minus pooled, in percentage points.

Group / metric	Pooled	Source-equal	Δ [95% CI]
Hallucination / AUROC	82.98	82.95	-0.04 [-0.17, 0.09]
Hallucination / AP	77.40	77.41	0.01 [-0.21, 0.53]
Hallucination / AURC	14.12	14.24	0.12 [-0.05, 0.31]
Jailbreak / AUROC	84.03	83.47	-0.55 [-5.29, 3.76]
Jailbreak / AP	68.44	67.91	-0.53 [-6.86, 6.48]
Jailbreak / AURC	4.70	4.29	-0.41 [-1.36, 0.78]
Adversarial / AUROC	86.61	86.74	0.14 [-0.21, 0.64]
Adversarial / AP	90.47	90.81	0.35 [-0.26, 1.07]
Adversarial / AURC	31.36	30.94	-0.42 [-1.15, 0.43]
OOD / AUROC	87.71	87.71	0.00 [0.00, 0.00]
OOD / AP	93.46	93.46	0.00 [0.00, 0.00]
OOD / AURC	48.20	48.20	0.00 [0.00, 0.00]

Table A37: DeepSeek FRDG pooled and source-equal metrics (%). Differences and intervals are source-equal minus pooled, in percentage points.

Group / metric	Pooled	Source-equal	Δ [95% CI]
Hallucination / AUROC	82.58	82.57	-0.01 [-0.15, 0.12]
Hallucination / AP	78.10	78.09	-0.00 [-0.24, 0.36]
Hallucination / AURC	16.32	16.44	0.12 [-0.06, 0.34]
Jailbreak / AUROC	84.53	83.60	-0.93 [-4.73, 2.49]
Jailbreak / AP	70.33	70.04	-0.29 [-5.45, 5.02]
Jailbreak / AURC	4.62	4.94	0.32 [-0.78, 1.54]
Adversarial / AUROC	88.53	88.59	0.06 [-1.03, 0.88]
Adversarial / AP	94.63	94.90	0.27 [-0.07, 0.76]
Adversarial / AURC	56.09	55.98	-0.11 [-0.45, 0.73]
OOD / AUROC	91.21	91.21	0.00 [0.00, 0.00]
OOD / AP	95.68	95.68	0.00 [0.00, 0.00]
OOD / AURC	38.14	38.14	0.00 [0.00, 0.00]

Table A38: InternVL3 FRDG pooled and source-equal metrics (%). Differences and intervals are source-equal minus pooled, in percentage points.

Group / metric	Pooled	Source-equal	Δ [95% CI]
Hallucination / AUROC	78.69	78.70	0.00 [-0.15, 0.16]
Hallucination / AP	77.95	78.00	0.06 [-0.27, 0.67]
Hallucination / AURC	20.73	20.66	-0.06 [-0.21, 0.10]
Jailbreak / AUROC	82.00	81.45	-0.55 [-3.57, 2.20]
Jailbreak / AP	79.76	77.70	-2.05 [-5.73, 1.95]
Jailbreak / AURC	15.96	16.14	0.18 [-2.06, 2.37]
Adversarial / AUROC	88.74	88.68	-0.06 [-0.50, 0.30]
Adversarial / AP	87.58	87.68	0.09 [-0.63, 1.13]
Adversarial / AURC	18.24	18.47	0.22 [-0.20, 0.94]
OOD / AUROC	87.14	87.14	0.00 [0.00, 0.00]
OOD / AP	90.25	90.25	0.00 [0.00, 0.00]
OOD / AURC	31.38	31.38	0.00 [0.00, 0.00]

Table A39: LLaVA FRDG pooled and source-equal metrics (%). Differences and intervals are source-equal minus pooled, in percentage points.

Group / metric	Pooled	Source-equal	Δ [95% CI]
Hallucination / AUROC	82.47	82.50	0.03 [-0.12, 0.20]
Hallucination / AP	77.07	77.21	0.14 [-0.24, 0.80]
Hallucination / AURC	14.04	14.03	-0.01 [-0.24, 0.19]
Jailbreak / AUROC	81.23	80.74	-0.49 [-3.92, 2.66]
Jailbreak / AP	74.39	73.13	-1.26 [-5.68, 3.23]
Jailbreak / AURC	8.70	8.72	0.02 [-1.54, 1.79]
Adversarial / AUROC	88.99	88.95	-0.04 [-0.42, 0.27]
Adversarial / AP	90.90	91.01	0.11 [-0.41, 0.69]
Adversarial / AURC	19.96	20.35	0.40 [-0.01, 1.20]
OOD / AUROC	88.86	88.86	0.00 [0.00, 0.00]
OOD / AP	88.97	88.97	0.00 [0.00, 0.00]
OOD / AURC	25.33	25.33	0.00 [0.00, 0.00]

D.2 Component and Score Design

These comparisons examine the contributions of the two measurements and the effects of changing their definitions, scaling, and aggregation.

D.2.1 All score variants

The component-removal comparison supports using both measurements: Full exceeds local-only macro AUROC by 1.39, 1.19, 1.05, and 0.97 percentage points across the four models, and the path-only score is substantially weaker. This finding does not establish that the chosen combination rule is optimal. Alternative centering, scaling, and additive rules can achieve higher point estimates and remain included below. Removing relative scaling lowers AUROC in every model, whereas replacing the leading eigenvalue with a trace gives a slightly higher point estimate on Qwen. Thus, the evidence for relative scaling is more consistent than the evidence favoring the eigenvalue over the trace.

Component removal tests whether the combined score benefits from both measurements. Replacing the trajectory definition, local statistic, or standardization rule addresses a different question and is not a component deletion. Full denotes the prescribed score, not a score chosen to be optimal over all variants. In the current results it exceeds local-only and path-only scoring, while the trace substitution has a higher AUROC point estimate on Qwen. The paired comparisons below distinguish point-estimate differences from differences supported by the reported intervals.

Table A40: Additional score variants. Cells report macro AUROC (%). The EUQ fixed combination is a custom workbook control, not a published continuous CF+IG baseline. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Ranks use all score variants for this model across the continued tables. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Variant	Qwen	DeepSeek	InternVL3	LLaVA
EUQ fixed combination	79.52	82.01	78.41	80.47
Full	85.33	86.71	84.15	85.39
Without relative scaling	79.38	79.28	76.87	77.66
Trace replaces eigenvalue	85.53	84.74	84.06	84.61
Local only	83.94	85.53	83.10	84.42
Path only	67.77	66.89	66.25	67.49
Total path only	55.00	56.55	55.32	53.72
Norm only	51.04	52.21	49.82	49.79
Local + total path	83.95	84.79	82.97	84.08
Local + endpoint	83.51	85.05	82.95	84.02
Gradient replaces local	76.25	75.42	74.40	75.90
L2 distance excess	85.33	86.84	84.27	85.47
JS distance excess	85.36	86.79	84.17	85.40

Table A41: Additional score variants. Cells report macro AUROC (%). The EUQ fixed combination is a custom workbook control, not a published continuous CF+IG baseline. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Ranks use all score variants for this model across the continued tables. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Variant	Qwen	DeepSeek	InternVL3	LLaVA
Local + adjacent JS	82.14	82.17	80.94	81.70
Without local clipping	85.33	86.71	84.15	85.39
Without path clipping	83.91	85.26	83.29	84.45
Without either clipping	83.91	85.26	83.29	84.45
K2	83.94	85.53	83.10	84.42
MAD additive	86.22	87.59	84.90	86.23
No centering	<u>86.98</u>	<u>87.91</u>	<u>85.47</u>	<u>86.85</u>
Fit-median additive	86.49	87.71	85.35	86.29
MEDIAN scale	84.81	86.48	83.89	85.06
SD scale	**87.11**	**87.94**	**85.81**	**86.91**
MEDIAN scale two-sided	83.74	85.32	83.12	84.29
SD scale two-sided	80.28	82.82	81.71	82.12

Table A42: Qwen full score-design comparison with reported macro intervals (%). Alternatives remain visible even when they outperform the prescribed full score. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Ranks use all score variants for this model across the continued tables. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Variant	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
EUQ fixed combination	79.52 [77.90, 81.13]	71.63 [69.55, 73.94]	24.07 [22.46, 25.80]
FRDG	85.33 [83.84, 86.83]	82.44 [80.44, 84.59]	24.59 [22.85, 26.66]
No relative scaling	79.38 [77.86, 80.88]	67.12 [65.31, 69.08]	26.30 [24.53, 28.33]
Trace replaces eigenvalue	85.53 [84.29, 86.71]	74.23 [72.30, 76.30]	**23.23** [21.76, 24.99]
Local only	83.94 [82.28, 85.52]	80.67 [78.65, 83.03]	26.16 [24.35, 28.30]
Path only	67.77 [66.57, 69.00]	62.36 [60.69, 64.13]	34.04 [32.46, 35.70]
Total path only	55.00 [53.28, 56.82]	44.48 [42.84, 46.78]	38.96 [37.25, 40.71]
Norm only	51.04 [49.00, 53.02]	42.32 [40.52, 44.62]	40.94 [39.06, 42.99]
Local + total path	83.95 [82.34, 85.51]	79.08 [77.10, 81.30]	25.81 [23.99, 27.89]
Local + endpoint distance	83.51 [81.85, 85.12]	79.29 [77.30, 81.54]	26.44 [24.57, 28.53]
Gradient replaces local	76.25 [74.67, 77.85]	63.28 [61.16, 65.63]	27.46 [25.79, 29.27]
L2 distance excess	85.33 [83.84, 86.82]	82.45 [80.44, 84.59]	24.60 [22.85, 26.67]
JS distance excess	85.36 [83.87, 86.84]	82.55 [80.58, 84.66]	24.60 [22.85, 26.66]

Table A43: Qwen full score-design comparison with reported macro intervals (%). Alternatives remain visible even when they outperform the prescribed full score. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Ranks use all score variants for this model across the continued tables. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Variant	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
Local + adjacent JS	82.14 [80.51, 83.79]	70.65 [68.60, 73.21]	26.33 [24.44, 28.42]
Unclip local only	85.33 [83.84, 86.83]	82.44 [80.44, 84.59]	24.59 [22.85, 26.66]
Unclip path only	83.91 [82.22, 85.46]	80.60 [78.48, 82.91]	26.02 [24.28, 28.09]
No clipping	83.91 [82.22, 85.46]	80.60 [78.48, 82.91]	26.02 [24.28, 28.09]
K2	83.94 [82.28, 85.52]	80.67 [78.65, 83.03]	26.16 [24.35, 28.30]
MAD additive	86.22 [84.76, 87.67]	<u>85.88</u> [84.29, 87.46]	24.16 [22.44, 26.19]
No centering	<u>86.98</u> [85.57, 88.35]	**86.13** [84.58, 87.71]	23.68 [22.04, 25.72]
Fit-median additive	86.49 [85.08, 87.86]	85.29 [83.81, 86.78]	24.02 [22.30, 26.07]
MEDIAN scale	84.81 [83.27, 86.34]	81.68 [79.65, 83.91]	24.96 [23.19, 27.05]
SD scale	**87.11** [85.67, 88.52]	85.81 [84.11, 87.44]	<u>23.67</u> [21.94, 25.69]
Median scale, two-sided	83.74 [82.05, 85.30]	80.59 [78.54, 82.92]	26.15 [24.44, 28.22]
SD scale, two-sided	80.28 [78.71, 81.79]	70.18 [68.09, 72.53]	26.54 [24.86, 28.49]

Table A44: DeepSeek full score-design comparison with reported macro intervals (%). Alternatives remain visible even when they outperform the prescribed full score. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Ranks use all score variants for this model across the continued tables. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Variant	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
EUQ fixed combination	82.01 [80.42, 83.61]	79.40 [77.74, 81.01]	29.20 [27.34, 31.21]
FRDG	86.71 [84.94, 88.38]	84.68 [82.74, 86.64]	28.79 [26.81, 30.86]
No relative scaling	79.28 [77.47, 80.99]	69.02 [67.51, 70.86]	31.79 [29.68, 33.88]
Trace replaces eigenvalue	84.74 [83.36, 86.06]	76.13 [74.46, 78.06]	30.46 [28.50, 32.31]
Local only	85.53 [83.59, 87.23]	83.09 [81.13, 85.24]	30.08 [28.01, 32.20]
Path only	66.89 [65.70, 68.16]	65.67 [64.05, 67.35]	40.81 [39.17, 42.32]
Total path only	56.55 [54.54, 58.53]	50.89 [49.53, 52.58]	44.55 [42.82, 46.17]
Norm only	52.21 [49.86, 54.28]	48.96 [47.40, 50.69]	46.50 [44.81, 48.14]
Local + total path	84.79 [83.01, 86.49]	80.21 [78.35, 82.19]	30.28 [28.21, 32.37]
Local + endpoint distance	85.05 [83.20, 86.75]	82.63 [80.68, 84.63]	30.60 [28.53, 32.74]
Gradient replaces local	75.42 [73.52, 77.21]	64.77 [63.01, 66.83]	34.39 [32.44, 36.27]
L2 distance excess	86.84 [85.07, 88.50]	84.98 [83.08, 86.88]	28.72 [26.74, 30.78]
JS distance excess	86.79 [85.03, 88.46]	84.87 [82.96, 86.78]	28.75 [26.77, 30.81]

Table A45: DeepSeek full score-design comparison with reported macro intervals (%). Alternatives remain visible even when they outperform the prescribed full score. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Ranks use all score variants for this model across the continued tables. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Variant	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
Local + adjacent JS	82.17 [80.30, 84.00]	70.47 [68.71, 72.46]	31.10 [29.08, 33.20]
Unclip local only	86.71 [84.94, 88.38]	84.68 [82.74, 86.64]	28.79 [26.81, 30.86]
Unclip path only	85.26 [83.38, 86.96]	82.85 [80.93, 84.99]	30.58 [28.49, 32.79]
No clipping	85.26 [83.38, 86.96]	82.85 [80.93, 84.99]	30.58 [28.49, 32.79]
K2	85.53 [83.59, 87.23]	83.09 [81.13, 85.24]	30.08 [28.01, 32.20]
MAD additive	87.59 [85.92, 89.18]	**87.69** [86.20, 89.12]	<u>28.38</u> [26.39, 30.44]
No centering	<u>87.91</u> [86.23, 89.51]	<u>87.56</u> [86.06, 88.98]	28.59 [26.52, 30.71]
Fit-median additive	87.71 [86.14, 89.22]	86.62 [85.16, 88.02]	28.90 [26.85, 30.97]
MEDIAN scale	86.48 [84.68, 88.16]	84.28 [82.31, 86.29]	28.94 [26.98, 31.02]
SD scale	**87.94** [86.23, 89.54]	86.84 [85.18, 88.47]	**28.18** [26.20, 30.24]
Median scale, two-sided	85.32 [83.44, 87.06]	82.95 [81.02, 85.14]	30.51 [28.42, 32.69]
SD scale, two-sided	82.82 [80.94, 84.67]	78.19 [76.29, 80.30]	31.54 [29.48, 33.76]

Table A46: InternVL3 full score-design comparison with reported macro intervals (%). Alternatives remain visible even when they outperform the prescribed full score. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Ranks use all score variants for this model across the continued tables. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Variant	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
EUQ fixed combination	78.41 [76.86, 79.95]	74.56 [72.67, 76.47]	22.52 [20.84, 24.24]
FRDG	84.15 [82.71, 85.56]	83.88 [81.94, 85.80]	21.58 [19.90, 23.19]
No relative scaling	76.87 [75.25, 78.37]	68.26 [65.78, 70.93]	23.53 [21.79, 25.25]
Trace replaces eigenvalue	84.06 [82.87, 85.33]	78.35 [76.39, 80.36]	**19.54** [18.14, 20.96]
Local only	83.10 [81.55, 84.64]	82.83 [80.87, 84.91]	23.01 [21.21, 24.79]
Path only	66.25 [65.13, 67.46]	59.82 [58.09, 61.51]	31.31 [29.76, 32.89]
Total path only	55.32 [53.71, 56.90]	43.87 [41.90, 46.38]	36.32 [34.50, 38.07]
Norm only	49.82 [47.95, 51.64]	39.57 [37.81, 42.06]	39.34 [37.43, 41.29]
Local + total path	82.97 [81.55, 84.35]	78.46 [76.27, 80.59]	22.12 [20.48, 23.77]
Local + endpoint distance	82.95 [81.49, 84.41]	81.30 [79.30, 83.31]	22.67 [20.97, 24.34]
Gradient replaces local	74.40 [72.90, 75.93]	61.66 [59.38, 64.48]	24.15 [22.61, 25.75]
L2 distance excess	84.27 [82.85, 85.68]	84.04 [82.09, 85.94]	21.49 [19.82, 23.11]
JS distance excess	84.17 [82.75, 85.59]	83.93 [81.99, 85.83]	21.55 [19.87, 23.15]

Table A47: InternVL3 full score-design comparison with reported macro intervals (%). Alternatives remain visible even when they outperform the prescribed full score. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Ranks use all score variants for this model across the continued tables. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Variant	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
Local + adjacent JS	80.94 [79.48, 82.36]	70.45 [68.19, 73.20]	22.72 [21.08, 24.40]
Unclip local only	84.15 [82.71, 85.56]	83.88 [81.94, 85.80]	21.58 [19.90, 23.19]
Unclip path only	83.29 [81.78, 84.79]	82.62 [80.56, 84.83]	22.69 [20.97, 24.38]
No clipping	83.29 [81.78, 84.79]	82.62 [80.56, 84.83]	22.69 [20.97, 24.38]
K2	83.10 [81.55, 84.64]	82.83 [80.87, 84.91]	23.01 [21.21, 24.79]
MAD additive	84.90 [83.52, 86.27]	<u>86.27</u> [84.74, 87.71]	21.11 [19.46, 22.72]
No centering	<u>85.47</u> [84.15, 86.81]	86.20 [84.69, 87.67]	<u>20.51</u> [18.96, 22.09]
Fit-median additive	85.35 [84.06, 86.65]	86.04 [84.73, 87.31]	20.55 [18.94, 22.18]
MEDIAN scale	83.89 [82.44, 85.32]	83.59 [81.64, 85.56]	21.78 [20.09, 23.41]
SD scale	**85.81** [84.48, 87.13]	**86.28** [84.62, 87.87]	20.58 [18.99, 22.19]
Median scale, two-sided	83.12 [81.60, 84.66]	82.64 [80.57, 84.81]	22.82 [21.11, 24.53]
SD scale, two-sided	81.71 [80.28, 83.17]	77.06 [74.87, 79.47]	22.39 [20.73, 24.14]

Table A48: LLaVA full score-design comparison with reported macro intervals (%). Alternatives remain visible even when they outperform the prescribed full score. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Ranks use all score variants for this model across the continued tables. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Variant	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
EUQ fixed combination	80.47 [78.89, 82.06]	73.04 [71.12, 75.19]	18.63 [16.80, 20.36]
FRDG	85.39 [83.82, 86.84]	82.83 [81.01, 84.68]	17.01 [15.44, 18.68]
No relative scaling	77.66 [76.15, 79.10]	63.59 [61.47, 65.76]	19.11 [17.52, 20.86]
Trace replaces eigenvalue	84.61 [83.28, 85.83]	73.53 [71.56, 75.56]	**15.83** [14.42, 17.23]
Local only	84.42 [82.78, 85.97]	81.52 [79.65, 83.48]	18.12 [16.47, 19.86]
Path only	67.49 [66.17, 68.87]	57.83 [55.98, 59.75]	26.30 [24.73, 27.85]
Total path only	53.72 [52.00, 55.48]	35.96 [34.11, 38.25]	32.00 [30.28, 33.74]
Norm only	49.79 [48.00, 51.57]	34.55 [33.01, 36.79]	34.44 [32.50, 36.34]
Local + total path	84.08 [82.58, 85.56]	77.13 [75.01, 79.32]	17.77 [16.13, 19.45]
Local + endpoint distance	84.02 [82.47, 85.51]	78.83 [76.80, 80.93]	18.05 [16.42, 19.78]
Gradient replaces local	75.90 [74.44, 77.37]	57.35 [55.05, 59.93]	19.53 [18.09, 21.05]
L2 distance excess	85.47 [83.92, 86.92]	83.01 [81.17, 84.84]	16.95 [15.38, 18.61]
JS distance excess	85.40 [83.83, 86.84]	82.84 [81.01, 84.68]	16.98 [15.40, 18.64]

Table A49: LLaVA full score-design comparison with reported macro intervals (%). Alternatives remain visible even when they outperform the prescribed full score. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Ranks use all score variants for this model across the continued tables. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Variant	AUROC [95% CI]	AP [95% CI]	AURC [95% CI]
Local + adjacent JS	81.70 [80.11, 83.26]	66.12 [63.65, 68.84]	18.42 [16.79, 20.13]
Unclip local only	85.39 [83.82, 86.84]	82.83 [81.01, 84.68]	17.01 [15.44, 18.68]
Unclip path only	84.45 [82.83, 85.98]	81.38 [79.55, 83.36]	18.34 [16.59, 20.03]
No clipping	84.45 [82.83, 85.98]	81.38 [79.55, 83.36]	18.34 [16.59, 20.03]
K2	84.42 [82.78, 85.97]	81.52 [79.65, 83.48]	18.12 [16.47, 19.86]
MAD additive	86.23 [84.74, 87.64]	**85.71** [84.20, 87.17]	16.53 [15.00, 18.14]
No centering	<u>86.85</u> [85.37, 88.21]	<u>85.59</u> [84.06, 87.04]	16.24 [14.58, 17.86]
Fit-median additive	86.29 [84.86, 87.62]	84.59 [83.13, 86.02]	16.42 [14.79, 18.04]
MEDIAN scale	85.06 [83.47, 86.53]	82.14 [80.30, 84.08]	17.25 [15.65, 18.92]
SD scale	**86.91** [85.45, 88.30]	85.34 [83.75, 86.86]	<u>16.15</u> [14.62, 17.76]
Median scale, two-sided	84.29 [82.63, 85.85]	81.43 [79.54, 83.46]	18.41 [16.64, 20.14]
SD scale, two-sided	82.12 [80.64, 83.53]	71.72 [69.74, 73.79]	18.46 [16.67, 20.16]

D.2.2 Reported paired intervals

Paired intervals distinguish the size of an observed difference from the precision of that comparison. For the leading-eigenvalue versus trace comparison, the reported AUROC interval excludes zero in favor of Full only on DeepSeek; the other three intervals include zero. A small point-estimate advantage on those models should therefore not be described as a demonstrated improvement. These intervals are exploratory and unadjusted, particularly where multiple score constructions were compared.

These intervals are transcribed from the workbook rather than independently regenerated. They are unadjusted for multiple comparisons and do not account for the outcome-based selection recorded in the companion artifact. Positive intervals must not be interpreted as independent validation of a selected configuration.

Table A50: Full minus Local only: absolute macro AUROC difference on the percentage scale.

Group	Difference	Reported 95% interval
Qwen	+1.393	[1.039, 1.826]
DeepSeek	+1.186	[0.929, 1.479]
InternVL3	+1.048	[0.772, 1.379]
LLaVA	+0.969	[0.769, 1.187]

Table A51: Full minus Path only: absolute macro AUROC difference on the percentage scale.

Group	Difference	Reported 95% interval
Qwen	+17.566	[15.703, 19.389]
DeepSeek	+19.829	[17.650, 21.814]
InternVL3	+17.895	[16.061, 19.662]
LLaVA	+17.903	[15.874, 19.791]

Table A52: Full minus Local + total path: absolute macro AUROC difference on the percentage scale.

Group	Difference	Reported 95% interval
Qwen	+1.387	[0.822, 1.992]
DeepSeek	+1.925	[1.286, 2.584]
InternVL3	+1.173	[0.512, 1.860]
LLaVA	+1.307	[0.778, 1.827]

Table A53: Full minus No clipping: absolute macro AUROC difference on the percentage scale.

Group	Difference	Reported 95% interval
Qwen	+1.427	[0.942, 1.870]
DeepSeek	+1.459	[1.029, 1.919]
InternVL3	+0.857	[0.528, 1.188]
LLaVA	+0.938	[0.601, 1.252]

Table A54: Full minus Trace replaces leading: absolute macro AUROC difference on the percentage scale.

Group	Difference	Reported 95% interval
Qwen	-0.197	[-1.126, 0.707]
DeepSeek	+1.979	[0.782, 3.127]
InternVL3	+0.082	[-0.920, 1.029]
LLaVA	+0.780	[-0.204, 1.730]

Differences below are target minus reference, as recorded. Positive AUROC/AP and negative AURC differences favor the target. These reported, unadjusted intervals do not account for configuration selection.

Table A55: Qwen paired macro differences: FRDG minus the listed reference, in percentage points, with reported 95% intervals.

Reference	ΔAUROC	ΔAP	ΔAURC
NLL	17.65 [15.57, 19.66]	26.93 [24.17, 29.57]	-6.74 [-8.37, -5.18]
LN-PE	13.25 [10.78, 15.80]	10.91 [7.98, 13.88]	-6.95 [-8.58, -5.33]
Token entropy	9.33 [6.83, 11.81]	10.89 [7.50, 13.70]	-7.07 [-8.77, -5.23]
SC	17.76 [15.17, 20.46]	10.51 [7.61, 13.35]	-9.20 [-11.00, -7.53]
SE	20.41 [17.76, 23.15]	18.40 [15.39, 21.30]	-10.00 [-11.67, -8.34]
HiddenDetect	5.88 [3.57, 8.15]	13.39 [10.39, 16.26]	2.23 [0.84, 3.67]
EUQ-CF	5.53 [3.21, 7.73]	10.11 [7.03, 12.86]	-2.17 [-3.68, -0.52]
EUQ-IG	11.62 [9.14, 14.07]	14.58 [11.61, 17.47]	-4.82 [-6.43, -3.21]
EUQ fixed combination	5.81 [3.57, 7.89]	10.81 [7.98, 13.54]	0.52 [-0.99, 2.09]
Head Fisher trace	12.48 [10.30, 14.81]	22.71 [19.84, 25.48]	-3.42 [-5.15, -1.88]
State Fisher trace	10.12 [8.03, 12.29]	20.64 [17.94, 23.41]	-1.89 [-3.54, -0.35]
Token gradient	10.91 [8.74, 13.10]	19.28 [16.57, 22.05]	-2.87 [-4.52, -1.24]
Layerwise JS	11.87 [9.56, 14.05]	21.34 [18.52, 23.98]	-3.46 [-5.19, -1.68]

Table A56: Qwen paired macro differences: FRDG minus the listed reference, in percentage points, with reported 95% intervals.

Reference	ΔAUROC	ΔAP	ΔAURC
No relative scaling	5.95 [4.95, 6.99]	15.32 [13.46, 17.26]	-1.71 [-2.60, -0.77]
Trace replaces eigenvalue	-0.20 [-1.13, 0.71]	8.22 [6.31, 10.06]	1.36 [0.52, 2.22]
Local only	1.39 [1.04, 1.83]	1.77 [1.34, 2.10]	-1.57 [-2.18, -1.11]
Path only	17.57 [15.70, 19.39]	20.09 [17.64, 22.68]	-9.45 [-10.68, -8.08]
Total path only	30.33 [28.04, 32.51]	37.97 [35.11, 40.42]	-14.36 [-15.81, -12.74]
Norm only	34.30 [31.86, 36.71]	40.12 [37.36, 42.67]	-16.35 [-18.03, -14.58]
Local + total path	1.39 [0.82, 1.99]	3.36 [2.08, 4.54]	-1.22 [-1.79, -0.65]
Local + endpoint distance	1.82 [1.27, 2.44]	3.16 [2.04, 4.15]	-1.84 [-2.58, -1.18]
Gradient replaces local	9.08 [7.09, 11.25]	19.16 [16.43, 21.91]	-2.86 [-4.32, -1.36]
L2 distance excess	0.01 [-0.01, 0.02]	-0.01 [-0.05, 0.02]	-0.01 [-0.02, 0.00]
JS distance excess	-0.02 [-0.08, 0.02]	-0.11 [-0.25, -0.00]	-0.01 [-0.06, 0.05]
Local + adjacent JS	3.19 [2.32, 4.09]	11.79 [9.64, 13.63]	-1.74 [-2.40, -1.08]
Unclip local only	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]

Table A57: Qwen paired macro differences: FRDG minus the listed reference, in percentage points, with reported 95% intervals.

Reference	ΔAUROC	ΔAP	ΔAURC
Unclip path only	1.43 [0.94, 1.87]	1.84 [1.18, 2.38]	-1.43 [-2.37, -0.53]
No clipping	1.43 [0.94, 1.87]	1.84 [1.18, 2.38]	-1.43 [-2.37, -0.53]
K2	1.39 [1.04, 1.83]	1.77 [1.34, 2.10]	-1.57 [-2.18, -1.11]
MAD additive	-0.89 [-1.06, -0.73]	-3.44 [-4.19, -2.60]	0.44 [0.34, 0.54]
No centering	-1.64 [-2.07, -1.26]	-3.69 [-4.54, -2.78]	0.91 [0.07, 1.61]
Fit-median additive	-1.16 [-1.59, -0.76]	-2.85 [-4.08, -1.54]	0.58 [-0.27, 1.37]
MEDIAN scale	0.52 [0.40, 0.68]	0.76 [0.52, 1.00]	-0.36 [-0.48, -0.27]
SD scale	-1.78 [-2.19, -1.42]	-3.37 [-4.00, -2.63]	0.92 [0.71, 1.16]
Median scale, two-sided	1.60 [1.19, 2.00]	1.85 [1.32, 2.25]	-1.56 [-2.44, -0.70]
SD scale, two-sided	5.06 [3.99, 6.15]	12.26 [10.35, 13.96]	-1.95 [-3.14, -0.80]

Table A58: DeepSeek paired macro differences: FRDG minus the listed reference, in percentage points, with reported 95% intervals.

Reference	ΔAUROC	ΔAP	ΔAURC
NLL	19.09 [16.64, 21.50]	25.08 [22.64, 27.38]	-8.15 [-9.61, -6.68]
LN-PE	12.49 [9.26, 15.94]	15.95 [12.91, 18.83]	-3.41 [-4.94, -1.90]
Token entropy	15.13 [12.48, 17.91]	14.12 [11.36, 16.79]	-11.47 [-13.03, -9.88]
SC	17.39 [14.05, 20.79]	18.01 [15.07, 20.69]	-8.37 [-9.87, -6.85]
SE	21.20 [18.32, 23.85]	12.18 [9.65, 14.84]	-17.43 [-18.93, -15.70]
HiddenDetect	14.70 [11.35, 18.12]	15.41 [12.62, 18.13]	-5.83 [-7.45, -4.24]
EUQ-CF	5.10 [2.53, 7.46]	7.39 [4.90, 9.85]	-4.31 [-5.61, -2.70]
EUQ-IG	4.67 [1.91, 7.39]	7.02 [4.26, 9.70]	-1.21 [-2.48, 0.18]
EUQ fixed combination	4.70 [2.24, 7.07]	5.28 [2.79, 7.75]	-0.40 [-1.68, 0.97]
Head Fisher trace	14.40 [11.81, 17.05]	21.54 [18.93, 24.07]	-6.43 [-7.91, -4.81]
State Fisher trace	12.06 [9.45, 14.56]	18.26 [15.62, 20.75]	-5.04 [-6.49, -3.49]
Token gradient	12.85 [10.43, 15.39]	19.93 [17.32, 22.44]	-5.24 [-6.67, -3.84]
Layerwise JS	14.17 [11.40, 16.91]	20.80 [18.17, 23.41]	-6.22 [-7.77, -4.55]

Table A59: DeepSeek paired macro differences: FRDG minus the listed reference, in percentage points, with reported 95% intervals.

Reference	ΔAUROC	ΔAP	ΔAURC
No relative scaling	7.43 [6.19, 8.65]	15.67 [13.82, 17.42]	-3.00 [-3.74, -2.18]
Trace replaces eigenvalue	1.98 [0.78, 3.13]	8.55 [6.69, 10.37]	-1.67 [-2.37, -0.84]
Local only	1.19 [0.93, 1.48]	1.60 [1.21, 1.90]	-1.28 [-1.77, -0.89]
Path only	19.83 [17.65, 21.81]	19.01 [16.60, 21.38]	-12.01 [-13.00, -10.83]
Total path only	30.17 [27.46, 32.77]	33.79 [31.45, 35.99]	-15.76 [-17.07, -14.28]
Norm only	34.50 [32.07, 37.17]	35.72 [33.41, 37.98]	-17.71 [-19.18, -16.13]
Local + total path	1.92 [1.29, 2.58]	4.47 [3.17, 5.81]	-1.48 [-2.21, -0.79]
Local + endpoint distance	1.66 [1.18, 2.16]	2.05 [1.31, 2.84]	-1.80 [-2.55, -1.08]
Gradient replaces local	11.29 [8.88, 13.76]	19.92 [17.29, 22.42]	-5.59 [-6.89, -4.24]
L2 distance excess	-0.12 [-0.16, -0.09]	-0.30 [-0.44, -0.17]	0.07 [0.05, 0.09]
JS distance excess	-0.08 [-0.11, -0.05]	-0.18 [-0.30, -0.09]	0.04 [0.02, 0.07]
Local + adjacent JS	4.55 [3.53, 5.59]	14.22 [12.33, 15.94]	-2.31 [-3.15, -1.51]
Unclip local only	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]

Table A60: DeepSeek paired macro differences: FRDG minus the listed reference, in percentage points, with reported 95% intervals.

Reference	ΔAUROC	ΔAP	ΔAURC
Unclip path only	1.46 [1.03, 1.92]	1.83 [1.23, 2.27]	-1.79 [-2.48, -1.11]
No clipping	1.46 [1.03, 1.92]	1.83 [1.23, 2.27]	-1.79 [-2.48, -1.11]
K2	1.19 [0.93, 1.48]	1.60 [1.21, 1.90]	-1.28 [-1.77, -0.89]
MAD additive	-0.87 [-1.07, -0.70]	-3.00 [-3.72, -2.24]	0.42 [0.33, 0.52]
No centering	-1.20 [-1.57, -0.88]	-2.88 [-3.60, -2.10]	0.21 [-0.28, 0.73]
Fit-median additive	-0.99 [-1.58, -0.47]	-1.93 [-3.23, -0.65]	-0.10 [-0.75, 0.56]
MEDIAN scale	0.23 [0.18, 0.29]	0.41 [0.27, 0.55]	-0.15 [-0.19, -0.11]
SD scale	-1.23 [-1.49, -1.00]	-2.16 [-2.63, -1.65]	0.62 [0.50, 0.76]
Median scale, two-sided	1.40 [1.04, 1.78]	1.74 [1.20, 2.08]	-1.71 [-2.37, -1.09]
SD scale, two-sided	3.90 [3.01, 4.80]	6.49 [5.23, 7.77]	-2.74 [-3.67, -1.88]

Table A61: InternVL3 paired macro differences: FRDG minus the listed reference, in percentage points, with reported 95% intervals.

Reference	ΔAUROC	ΔAP	ΔAURC
NLL	16.82 [14.70, 19.05]	26.73 [23.77, 29.42]	-6.11 [-7.92, -4.48]
LN-PE	11.19 [9.14, 13.41]	17.92 [15.11, 20.74]	-2.96 [-4.47, -1.45]
Token entropy	11.26 [8.96, 13.51]	17.93 [14.95, 20.41]	-6.38 [-7.94, -4.80]
SC	15.55 [13.17, 18.00]	16.01 [13.27, 18.86]	-8.78 [-10.57, -7.05]
SE	12.30 [10.12, 14.58]	13.86 [11.22, 16.71]	-6.62 [-8.30, -5.01]
HiddenDetect	12.99 [10.70, 15.30]	18.08 [15.27, 20.97]	-6.34 [-8.07, -4.75]
EUQ-CF	9.30 [7.10, 11.53]	12.31 [9.49, 14.97]	-4.13 [-5.78, -2.50]
EUQ-IG	9.59 [7.33, 11.92]	11.20 [8.61, 13.85]	-6.48 [-8.31, -4.62]
EUQ fixed combination	5.73 [3.60, 7.85]	9.33 [6.78, 11.84]	-0.94 [-2.60, 0.63]
Head Fisher trace	12.41 [10.26, 14.55]	22.14 [19.24, 24.82]	-3.45 [-5.15, -1.81]
State Fisher trace	10.34 [8.11, 12.54]	20.24 [17.31, 23.04]	-1.41 [-3.14, 0.12]
Token gradient	11.11 [8.86, 13.28]	22.08 [18.91, 24.76]	-2.12 [-3.75, -0.50]
Layerwise JS	11.82 [9.78, 14.06]	21.76 [18.89, 24.71]	-2.61 [-4.24, -1.04]

Table A62: InternVL3 paired macro differences: FRDG minus the listed reference, in percentage points, with reported 95% intervals.

Reference	ΔAUROC	ΔAP	ΔAURC
No relative scaling	7.28 [6.22, 8.35]	15.63 [13.32, 17.82]	-1.95 [-2.72, -1.12]
Trace replaces eigenvalue	0.08 [-0.92, 1.03]	5.54 [3.94, 7.17]	2.03 [1.26, 2.84]
Local only	1.05 [0.77, 1.38]	1.05 [0.78, 1.25]	-1.43 [-1.91, -1.02]
Path only	17.89 [16.06, 19.66]	24.06 [21.76, 26.41]	-9.73 [-11.03, -8.35]
Total path only	28.82 [26.67, 31.02]	40.02 [37.15, 42.52]	-14.74 [-16.29, -13.20]
Norm only	34.32 [31.93, 36.69]	44.31 [41.35, 46.74]	-17.76 [-19.44, -15.93]
Local + total path	1.17 [0.51, 1.86]	5.42 [3.89, 7.08]	-0.54 [-1.15, 0.14]
Local + endpoint distance	1.20 [0.71, 1.68]	2.58 [1.73, 3.47]	-1.10 [-1.72, -0.43]
Gradient replaces local	9.74 [7.54, 11.81]	22.23 [19.09, 24.85]	-2.57 [-4.06, -0.98]
L2 distance excess	-0.13 [-0.17, -0.09]	-0.16 [-0.23, -0.10]	0.08 [0.06, 0.11]
JS distance excess	-0.03 [-0.05, -0.01]	-0.04 [-0.08, -0.01]	0.03 [0.01, 0.05]
Local + adjacent JS	3.21 [2.31, 4.20]	13.44 [10.98, 15.59]	-1.14 [-1.84, -0.39]
Unclip local only	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]

Table A63: InternVL3 paired macro differences: FRDG minus the listed reference, in percentage points, with reported 95% intervals.

Reference	ΔAUROC	ΔAP	ΔAURC
Unclip path only	0.86 [0.53, 1.19]	1.26 [0.77, 1.63]	-1.11 [-1.94, -0.32]
No clipping	0.86 [0.53, 1.19]	1.26 [0.77, 1.63]	-1.11 [-1.94, -0.32]
K2	1.05 [0.77, 1.38]	1.05 [0.78, 1.25]	-1.43 [-1.91, -1.02]
MAD additive	-0.76 [-0.92, -0.61]	-2.38 [-3.01, -1.73]	0.47 [0.37, 0.58]
No centering	-1.32 [-1.64, -1.02]	-2.32 [-2.92, -1.65]	1.06 [0.57, 1.53]
Fit-median additive	-1.20 [-1.66, -0.72]	-2.16 [-3.31, -1.03]	1.03 [0.34, 1.69]
MEDIAN scale	0.26 [0.19, 0.34]	0.30 [0.19, 0.39]	-0.20 [-0.26, -0.15]
SD scale	-1.66 [-2.05, -1.32]	-2.39 [-2.89, -1.86]	0.99 [0.79, 1.22]
Median scale, two-sided	1.03 [0.73, 1.35]	1.24 [0.82, 1.56]	-1.25 [-2.05, -0.50]
SD scale, two-sided	2.44 [1.57, 3.27]	6.83 [5.39, 8.24]	-0.81 [-1.87, 0.25]

Table A64: LLaVA paired macro differences: FRDG minus the listed reference, in percentage points, with reported 95% intervals.

Reference	ΔAUROC	ΔAP	ΔAURC
NLL	17.14 [14.95, 19.18]	31.51 [28.39, 34.21]	-5.70 [-7.14, -4.22]
LN-PE	11.50 [9.22, 13.72]	17.38 [14.46, 20.11]	-4.49 [-5.98, -2.98]
Token entropy	11.69 [9.11, 14.11]	16.33 [12.98, 19.07]	-8.76 [-10.53, -6.75]
SC	15.42 [12.67, 18.06]	16.52 [13.75, 19.16]	-8.74 [-10.49, -6.79]
SE	16.08 [13.53, 18.41]	16.92 [14.07, 19.67]	-12.74 [-14.32, -10.81]
HiddenDetect	11.27 [8.94, 13.73]	16.57 [13.94, 19.36]	-6.36 [-8.21, -4.45]
EUQ-CF	5.70 [3.36, 7.91]	10.98 [8.25, 13.56]	-3.12 [-4.76, -1.21]
EUQ-IG	7.52 [5.15, 9.71]	10.75 [8.06, 13.31]	-4.39 [-6.00, -2.71]
EUQ fixed combination	4.92 [2.60, 7.02]	9.79 [7.01, 12.38]	-1.62 [-3.20, 0.13]
Head Fisher trace	12.84 [10.66, 15.09]	26.80 [23.57, 29.57]	-3.53 [-5.15, -1.97]
State Fisher trace	10.24 [8.02, 12.46]	23.51 [20.54, 26.37]	-1.99 [-3.48, -0.54]
Token gradient	11.15 [8.98, 13.24]	25.62 [22.62, 28.50]	-2.76 [-4.28, -1.15]
Layerwise JS	12.02 [9.89, 14.28]	25.73 [22.68, 28.64]	-2.69 [-4.33, -1.17]

Table A65: LLaVA paired macro differences: FRDG minus the listed reference, in percentage points, with reported 95% intervals.

Reference	ΔAUROC	ΔAP	ΔAURC
No relative scaling	7.73 [6.82, 8.62]	19.25 [17.22, 21.07]	-2.10 [-2.81, -1.42]
Trace replaces eigenvalue	0.78 [-0.20, 1.73]	9.30 [7.41, 11.01]	1.17 [0.47, 1.91]
Local only	0.97 [0.77, 1.19]	1.31 [1.00, 1.55]	-1.11 [-1.41, -0.83]
Path only	17.90 [15.87, 19.79]	25.01 [22.51, 27.41]	-9.30 [-10.52, -7.95]
Total path only	31.67 [29.26, 33.88]	46.87 [44.02, 49.24]	-14.99 [-16.51, -13.36]
Norm only	35.60 [33.17, 37.89]	48.28 [45.59, 50.50]	-17.44 [-19.00, -15.77]
Local + total path	1.31 [0.78, 1.83]	5.70 [4.27, 6.97]	-0.76 [-1.32, -0.19]
Local + endpoint distance	1.37 [0.92, 1.79]	4.00 [2.91, 4.96]	-1.05 [-1.62, -0.44]
Gradient replaces local	9.49 [7.37, 11.53]	25.49 [22.56, 28.32]	-2.52 [-3.94, -1.08]
L2 distance excess	-0.08 [-0.11, -0.06]	-0.18 [-0.26, -0.10]	0.06 [0.04, 0.09]
JS distance excess	-0.01 [-0.03, 0.01]	-0.01 [-0.07, 0.04]	0.03 [-0.01, 0.08]
Local + adjacent JS	3.69 [2.78, 4.58]	16.72 [14.34, 18.87]	-1.41 [-2.07, -0.74]
Unclip local only	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]

Table A66: LLaVA paired macro differences: FRDG minus the listed reference, in percentage points, with reported 95% intervals.

Reference	ΔAUROC	ΔAP	ΔAURC
Unclip path only	0.94 [0.60, 1.25]	1.46 [0.95, 1.86]	-1.33 [-2.08, -0.54]
No clipping	0.94 [0.60, 1.25]	1.46 [0.95, 1.86]	-1.33 [-2.08, -0.54]
K2	0.97 [0.77, 1.19]	1.31 [1.00, 1.55]	-1.11 [-1.41, -0.83]
MAD additive	-0.84 [-1.00, -0.69]	-2.87 [-3.52, -2.16]	0.48 [0.37, 0.60]
No centering	-1.46 [-1.83, -1.15]	-2.76 [-3.39, -2.07]	0.77 [0.16, 1.29]
Fit-median additive	-0.90 [-1.37, -0.49]	-1.76 [-2.78, -0.70]	0.58 [-0.06, 1.16]
MEDIAN scale	0.33 [0.26, 0.41]	0.69 [0.48, 0.84]	-0.24 [-0.30, -0.18]
SD scale	-1.52 [-1.85, -1.22]	-2.50 [-2.97, -1.99]	0.86 [0.67, 1.07]
Median scale, two-sided	1.10 [0.82, 1.39]	1.41 [0.95, 1.72]	-1.41 [-2.07, -0.69]
SD scale, two-sided	3.27 [2.53, 4.10]	11.11 [9.51, 12.65]	-1.45 [-2.38, -0.41]

D.2.3 STRATIFIED COMPONENT COMPARISONS

Variation across strata is important here: an average improvement can coexist with negligible or negative differences in some subsets. The reported median and range expose that variation, but they do not test whether the path score adds information after exact matching on confidence or local sensitivity. A focused matching study would be needed for that stronger interpretation.

The joint strata use the recorded entropy, answer-length, and endpoint-distance cutoffs within each source. They are exploratory comparisons, not exact matching experiments. Small strata and the absence of multiplicity adjustment limit inferential claims. The table reports medians and ranges of within-stratum differences, not a pooled or weighted effect. Complete stratum counts, cutoffs, and values are supplied in `data/Conditional.csv`.

Table A67: Qwen stratified FRDG-minus-local comparisons. Differences are percentage points; brackets give the observed range, not a confidence interval.

Source	N range	ΔAUROC	ΔAP	ΔAURC
ANDA	9–44	1.79 [0.00, 11.69]	0.76 [0.00, 5.67]	-2.18 [-10.81, 0.00]
FigStep	6–46	1.79 [0.00, 8.33]	0.20 [0.00, 16.67]	-0.22 [-2.53, 0.00]
HADES	7–49	0.00 [0.00, 20.00]	0.00 [0.00, 8.33]	0.00 [-11.90, 0.00]
MMDT	72–309	1.75 [0.85, 2.40]	0.49 [0.33, 1.47]	-2.44 [-3.31, -0.71]
PGN	8–40	0.20 [0.00, 6.67]	0.05 [0.00, 4.00]	-0.23 [-4.17, 0.00]
POPE	64–206	0.35 [0.00, 5.74]	0.58 [0.00, 2.06]	-0.13 [-7.37, 0.00]
R-Bench	55–230	0.71 [0.00, 2.69]	0.68 [0.00, 2.01]	-0.94 [-2.80, 0.00]
Typographic	113–393	0.42 [0.00, 1.03]	1.67 [0.00, 5.40]	-0.16 [-0.31, 0.00]
VisualAdv	31–124	0.11 [0.00, 7.69]	0.34 [0.00, 2.79]	-0.05 [-1.05, 0.00]

Table A68: DeepSeek stratified FRDG-minus-local comparisons. Differences are percentage points; brackets give the observed range, not a confidence interval.

Source	N range	ΔAUROC	ΔAP	ΔAURC
ANDA	8–41	0.40 [0.00, 7.41]	0.11 [0.00, 1.64]	-0.22 [-12.09, 0.00]
FigStep	10–45	0.00 [0.00, 6.67]	0.00 [0.00, 3.00]	0.00 [-1.82, 0.00]
HADES	8–41	0.00 [0.00, 11.49]	0.00 [0.00, 1.34]	0.00 [-8.39, 0.00]
MMDT	78–277	1.16 [0.16, 2.25]	0.40 [0.06, 0.91]	-1.29 [-3.58, -0.13]
PGN	12–44	0.00 [0.00, 2.60]	0.00 [0.00, 0.72]	0.00 [-2.58, 0.00]
POPE	52–203	0.43 [0.00, 1.70]	0.14 [0.00, 1.70]	-0.49 [-1.81, 0.00]
R-Bench	52–205	0.87 [0.11, 2.13]	0.77 [0.30, 3.30]	-1.66 [-2.92, -0.03]
Typographic	100–373	1.18 [0.09, 2.13]	2.61 [0.32, 5.32]	-0.69 [-1.41, -0.01]
VisualAdv	32–139	0.51 [0.00, 2.95]	2.71 [0.00, 13.33]	-0.11 [-1.03, 0.00]

Table A69: InternVL3 stratified FRDG-minus-local comparisons. Differences are percentage points; brackets give the observed range, not a confidence interval.

Source	N range	ΔAUROC	ΔAP	ΔAURC
ANDA	9–44	0.00 [0.00, 0.22]	0.00 [0.00, 0.08]	0.00 [-0.23, 0.00]
FigStep	9–48	0.00 [0.00, 0.80]	0.00 [0.00, 0.86]	0.00 [-4.29, 0.00]
HADES	6–39	0.00 [0.00, 6.64]	0.00 [0.00, 2.40]	0.00 [-7.34, 0.00]
MMDT	70–253	0.76 [0.55, 1.72]	0.34 [0.13, 1.20]	-0.80 [-2.24, -0.40]
PGN	9–42	1.93 [1.42, 5.19]	0.63 [0.44, 1.85]	-3.22 [-6.02, -1.11]
POPE	47–201	0.94 [0.00, 2.55]	0.56 [0.00, 0.93]	-1.51 [-3.21, 0.00]
R-Bench	49–192	1.12 [0.00, 1.58]	0.68 [0.00, 1.82]	-1.10 [-3.68, 0.00]
Typographic	91–357	1.23 [0.28, 2.37]	0.72 [0.06, 2.73]	-1.39 [-4.22, -0.09]
VisualAdv	30–127	0.83 [0.00, 3.17]	0.30 [0.00, 4.59]	-0.97 [-6.06, 0.00]

Table A70: LLaVA stratified FRDG-minus-local comparisons. Differences are percentage points; brackets give the observed range, not a confidence interval.

Source	N range	ΔAUROC	ΔAP	ΔAURC
ANDA	11–47	0.00 [0.00, 4.17]	0.00 [0.00, 2.22]	0.00 [-1.52, 0.00]
FigStep	10–53	0.00 [0.00, 2.22]	0.00 [0.00, 4.46]	0.00 [-3.16, 0.00]
HADES	11–42	0.00 [0.00, 6.06]	0.00 [0.00, 1.40]	0.00 [-3.21, 0.00]
MMDT	75–282	1.21 [0.35, 3.56]	0.83 [0.22, 1.58]	-1.06 [-4.66, -0.33]
PGN	8–42	0.00 [0.00, 3.64]	0.00 [0.00, 0.88]	0.00 [-5.21, 0.00]
POPE	60–215	1.13 [0.16, 2.26]	0.87 [0.16, 1.99]	-1.40 [-3.79, -0.02]
R-Bench	60–229	0.47 [0.00, 2.26]	0.84 [0.00, 6.17]	-0.71 [-2.00, 0.00]
Typographic	112–359	1.18 [0.78, 3.14]	0.63 [0.21, 4.09]	-1.04 [-3.59, -0.20]
VisualAdv	32–138	0.38 [0.00, 0.97]	0.06 [0.00, 1.88]	-0.15 [-1.31, 0.00]

D.2.4 SCORE SIGNS AND REFERENCE STATISTICS

In the recorded test responses, the standardized local score is never negative. Consequently, removing clipping from that component alone cannot change the test ranking when the remaining computation is fixed. The effect of clipping in this evaluation is therefore primarily associated with the path component. The sign counts are descriptive diagnostics, not quantities for which a best or second-best value is meaningful.

The sign comparison separates negative standardized local and path scores. Rates use the response failure labels in the supplied records. The nominal cohort is a reference-pool name, not a guarantee that all responses are correct. Empty cells have no defined failure rate. The full two-dimensional bin counts are included in `data/Bins.csv`.

Table A71: Qwen score-sign groups. A minus sign denotes a negative standardized component; rates and intervals are percentages.

Cohort	Local / path	N	Failures	Rate [95% CI]
nominal	+ / +	169	37	21.89 [16.32, 28.71]
nominal	+ / -	238	50	21.01 [16.31, 26.62]
nominal	- / +	232	28	12.07 [8.48, 16.89]
nominal	- / -	161	25	15.53 [10.74, 21.92]
test_correct	+ / +	114	0	0.00 [0.00, 3.26]
test_correct	+ / -	4351	0	0.00 [0.00, 0.09]
test_correct	- / +	0	0	– [–, –]
test_correct	- / -	0	0	– [–, –]
test_failure	+ / +	759	759	100.00 [99.50, 100.00]
test_failure	+ / -	1276	1276	100.00 [99.70, 100.00]
test_failure	- / +	0	0	– [–, –]
test_failure	- / -	0	0	– [–, –]

Table A72: Qwen reference-fit statistics. Raw MAD is not multiplied by a Gaussian consistency factor.

Component	N	Median	Raw MAD	Effective scale
local	200	1.22532	0.42305	0.42305
path	200	0.06192	0.01553	0.01553
total	200	0.25350	0.04151	0.04151
endpoint	200	0.19404	0.05060	0.05060
norm	200	8.08572	1.30531	1.30531
unscaled	200	0.01844	0.00893	0.00893
trace	200	3.93159	1.43721	1.43721
gradient	200	1.02612	0.01629	0.01629
l2	200	0.00582	0.00163	0.00163
js_excess	200	0.02267	0.00528	0.00528
adjacent_js	200	0.00108	0.00038	0.00038
k2	200	0.00000	0.00000	0.00000

Table A73: DeepSeek score-sign groups. A minus sign denotes a negative standardized component; rates and intervals are percentages.

Cohort	Local / path	N	Failures	Rate [95% CI]
nominal	+ / +	161	35	21.74 [16.07, 28.73]
nominal	+ / -	227	44	19.38 [14.77, 25.02]
nominal	- / +	230	32	13.91 [10.03, 18.98]
nominal	- / -	182	32	17.58 [12.74, 23.77]
test_correct	+ / +	118	0	0.00 [0.00, 3.15]
test_correct	+ / -	4262	0	0.00 [0.00, 0.09]
test_correct	- / +	0	0	– [–, –]
test_correct	- / -	0	0	– [–, –]
test_failure	+ / +	761	761	100.00 [99.50, 100.00]
test_failure	+ / -	1359	1359	100.00 [99.72, 100.00]
test_failure	- / +	0	0	– [–, –]
test_failure	- / -	0	0	– [–, –]

Table A74: DeepSeek reference-fit statistics. Raw MAD is not multiplied by a Gaussian consistency factor.

Component	N	Median	Raw MAD	Effective scale
local	200	0.96972	0.32933	0.32933
path	200	0.06420	0.01878	0.01878
total	200	0.24244	0.03936	0.03936
endpoint	200	0.18682	0.05721	0.05721
norm	200	8.20761	1.10377	1.10377
unscaled	200	0.01369	0.00597	0.00597
trace	200	3.12087	1.19071	1.19071
gradient	200	1.02902	0.01590	0.01590
l2	200	0.00600	0.00191	0.00191
js_excess	200	0.02299	0.00637	0.00637
adjacent_js	200	0.00097	0.00032	0.00032
k2	200	0.00000	0.00000	0.00000

Table A75: InternVL3 score-sign groups. A minus sign denotes a negative standardized component; rates and intervals are percentages.

Cohort	Local / path	N	Failures	Rate [95% CI]
nominal	+ / +	133	31	23.31 [16.93, 31.18]
nominal	+ / -	232	42	18.10 [13.68, 23.56]
nominal	- / +	245	45	18.37 [14.02, 23.69]
nominal	- / -	190	39	20.53 [15.39, 26.83]
test_correct	+ / +	110	0	0.00 [0.00, 3.37]
test_correct	+ / -	4068	0	0.00 [0.00, 0.09]
test_correct	- / +	0	0	– [–, –]
test_correct	- / -	0	0	– [–, –]
test_failure	+ / +	807	807	100.00 [99.53, 100.00]
test_failure	+ / -	1515	1515	100.00 [99.75, 100.00]
test_failure	- / +	0	0	– [–, –]
test_failure	- / -	0	0	– [–, –]

Table A76: InternVL3 reference-fit statistics. Raw MAD is not multiplied by a Gaussian consistency factor.

Component	N	Median	Raw MAD	Effective scale
local	200	1.10491	0.38905	0.38905
path	200	0.06300	0.01851	0.01851
total	200	0.24982	0.03309	0.03309
endpoint	200	0.19351	0.04762	0.04762
norm	200	8.28567	1.13077	1.13077
unscaled	200	0.01610	0.00698	0.00698
trace	200	3.52102	1.37022	1.37022
gradient	200	1.03186	0.01556	0.01556
l2	200	0.00592	0.00184	0.00184
js_excess	200	0.02304	0.00625	0.00625
adjacent_js	200	0.00106	0.00031	0.00031
k2	200	0.00000	0.00000	0.00000

Table A77: LLaVA score-sign groups. A minus sign denotes a negative standardized component; rates and intervals are percentages.

Cohort	Local / path	N	Failures	Rate [95% CI]
nominal	+ / +	158	44	27.85 [21.45, 35.30]
nominal	+ / -	243	55	22.63 [17.82, 28.30]
nominal	- / +	222	35	15.77 [11.56, 21.14]
nominal	- / -	177	32	18.08 [13.11, 24.41]
test_correct	+ / +	139	0	0.00 [-0.00, 2.69]
test_correct	+ / -	4544	0	0.00 [0.00, 0.08]
test_correct	- / +	0	0	– [–, –]
test_correct	- / -	0	0	– [–, –]
test_failure	+ / +	652	652	100.00 [99.41, 100.00]
test_failure	+ / -	1165	1165	100.00 [99.67, 100.00]
test_failure	- / +	0	0	– [–, –]
test_failure	- / -	0	0	– [–, –]

Table A78: LLaVA reference-fit statistics. Raw MAD is not multiplied by a Gaussian consistency factor.

Component	N	Median	Raw MAD	Effective scale
local	200	1.13156	0.39046	0.39046
path	200	0.06456	0.01706	0.01706
total	200	0.25176	0.03899	0.03899
endpoint	200	0.19182	0.04942	0.04942
norm	200	8.31683	1.21140	1.21140
unscaled	200	0.01559	0.00723	0.00723
trace	200	3.44334	1.40369	1.40369
gradient	200	1.03074	0.01474	0.01474
l2	200	0.00607	0.00172	0.00172
js_excess	200	0.02350	0.00582	0.00582
adjacent_js	200	0.00107	0.00035	0.00035
k2	200	0.00000	0.00000	0.00000

D.3 Numerical Accuracy

These checks evaluate eigenvalue estimation, the local quadratic approximation, and probability reconstruction separately from detection performance.

D.3.1 Numerical accuracy and probability representation

Table 3 in the main text reports median relative eigenvalue errors at two, four, and eight Lanczos steps. The medians are computed across the state and starting-vector records within each model group. Increasing the budget reduces the reported approximation error, but this is distinct from improving failure discrimination. The main score continues to use eight steps. The detailed tables below report the current records, including their recorded reference settings; historical numerical values are not carried forward.

The compression sheet compares an original probability vector with an eight-coordinate representation whose remaining mass is spread uniformly over the vocabulary complement. This reconstruction generally changes Fisher–Rao distance when the original tail is not uniform. Agreement of distance calculations within a reconstructed representation is not proof that it preserves the geometry of original model outputs.

The main trajectory definition uses the original full-vocabulary probabilities, not the uniform-tail reconstruction. The export comparison therefore concerns probability representation and must not be used as a substitute for checking the main trajectory calculation. At a fixed prediction position, explicit token coordinates must be shared across the selected layers before distances are computed.

D.3.2 Local approximation and spectral estimates

The spectral error decreases sharply with the iteration budget in the supplied records. At eight steps, median relative errors range from approximately 1.4×10^{-8} to 1.8×10^{-6} across models, compared with roughly 0.35–0.44 at two steps. This supports eight steps as an accurate approximation of the recorded reference operator; it does not establish that a more accurate eigenvalue improves failure detection. The local-radius checks address a separate question: how closely the quadratic approximation follows the finite predictive displacement.

The local check compares the squared Fisher–Rao distance after a perturbation with the Fisher quadratic form at the original state. Relative error is reported as recorded, grouped by direction and perturbation radius. The spectral comparison uses the reference values stored with each state. Neither calculation evaluates semantic correctness. The display records contain `State_dimension=96`, `Grouped_vocab=128` for local checks, and `Reference_budget=64` for spectral checks. These metadata do not by themselves specify the dimensions or computation of the original model operator.

Table A79: Qwen local quadratic-approximation errors. Errors are dimensionless.

Direction	Radius	Records	Median error	Max. error
leading	0.0001	24	0.000288	0.000409
leading	0.0003	24	0.000866	0.00123
leading	0.001	24	0.00289	0.00411
leading	0.003	24	0.00868	0.0124
leading	0.01	24	0.0291	0.0422
leading	0.03	24	0.089	0.134
leading	0.1	24	0.226	0.534
random	0.0001	24	3.24e-05	7.68e-05
random	0.0003	24	9.71e-05	0.00023
random	0.001	24	0.000323	0.000769
random	0.003	24	0.000962	0.00231
random	0.01	24	0.00312	0.00778
random	0.03	24	0.00865	0.0239
random	0.1	24	0.0271	0.086

Table A80: Qwen spectral estimates across recorded states and starts. Correlation and top-10% overlap are descriptive summaries of the recorded runs.

Steps	Records	Median error	Max. error	Spearman	Jaccard
2	120	0.383	0.875	0.6400	0.5000
4	120	0.0212	0.395	0.9730	1.0000
8	120	1.85e-06	0.0596	1.0000	1.0000
16	120	6.38e-16	2.83e-08	1.0000	1.0000
32	120	4.17e-16	3.38e-15	1.0000	1.0000

Table A81: DeepSeek local quadratic-approximation errors. Errors are dimensionless.

Direction	Radius	Records	Median error	Max. error
leading	0.0001	24	0.000284	0.000446
leading	0.0003	24	0.000853	0.00134
leading	0.001	24	0.00285	0.00447
leading	0.003	24	0.00855	0.0135
leading	0.01	24	0.0287	0.0456
leading	0.03	24	0.0867	0.143
leading	0.1	24	0.238	0.551
random	0.0001	24	1.57e-05	6.14e-05
random	0.0003	24	4.7e-05	0.000184
random	0.001	24	0.000156	0.000613
random	0.003	24	0.000463	0.00183
random	0.01	24	0.00148	0.00607
random	0.03	24	0.00432	0.0182
random	0.1	24	0.0156	0.0702

Table A82: DeepSeek spectral estimates across recorded states and starts. Correlation and top-10% overlap are descriptive summaries of the recorded runs.

Steps	Records	Median error	Max. error	Spearman	Jaccard
2	120	0.416	0.787	0.4817	0.5000
4	120	0.0288	0.4	0.9443	1.0000
8	120	6.68e-07	0.234	0.9991	1.0000
16	120	4.88e-16	3.02e-08	1.0000	1.0000
32	120	4.92e-16	2.05e-15	1.0000	1.0000

Table A83: InternVL3 local quadratic-approximation errors. Errors are dimensionless.

Direction	Radius	Records	Median error	Max. error
leading	0.0001	24	0.000327	0.000432
leading	0.0003	24	0.000981	0.0013
leading	0.001	24	0.00327	0.00433
leading	0.003	24	0.00984	0.0131
leading	0.01	24	0.0334	0.0443
leading	0.03	24	0.0947	0.139
leading	0.1	24	0.26	0.543
random	0.0001	24	3.22e-05	7.5e-05
random	0.0003	24	9.65e-05	0.000225
random	0.001	24	0.000321	0.000748
random	0.003	24	0.000961	0.00224
random	0.01	24	0.00316	0.00735
random	0.03	24	0.00891	0.0212
random	0.1	24	0.0287	0.07

Table A84: InternVL3 spectral estimates across recorded states and starts. Correlation and top-10% overlap are descriptive summaries of the recorded runs.

Steps	Records	Median error	Max. error	Spearman	Jaccard
2	120	0.439	0.785	0.5313	0.2000
4	120	0.00629	0.46	0.9643	1.0000
8	120	1.44e-08	0.0557	1.0000	1.0000
16	120	5.19e-16	3.37e-10	1.0000	1.0000
32	120	3.89e-16	1.9e-15	1.0000	1.0000

Table A85: LLaVA local quadratic-approximation errors. Errors are dimensionless.

Direction	Radius	Records	Median error	Max. error
leading	0.0001	24	0.000347	0.000466
leading	0.0003	24	0.00104	0.0014
leading	0.001	24	0.00348	0.00467
leading	0.003	24	0.0105	0.0141
leading	0.01	24	0.0349	0.048
leading	0.03	24	0.098	0.153
leading	0.1	24	0.268	0.611
random	0.0001	24	2.26e-05	7.78e-05
random	0.0003	24	6.77e-05	0.000233
random	0.001	24	0.000226	0.000779
random	0.003	24	0.000676	0.00235
random	0.01	24	0.00228	0.00794
random	0.03	24	0.00727	0.0248
random	0.1	24	0.0279	0.0939

Table A86: LLaVA spectral estimates across recorded states and starts. Correlation and top-10% overlap are descriptive summaries of the recorded runs.

Steps	Records	Median error	Max. error	Spearman	Jaccard
2	120	0.355	0.769	0.7504	0.5000
4	120	0.00898	0.563	0.9478	1.0000
8	120	1.83e-07	0.0616	0.9991	1.0000
16	120	4.2e-16	2.06e-13	1.0000	1.0000
32	120	5.06e-16	1.82e-15	1.0000	1.0000

D.3.3 UNIFORM-TAIL RECONSTRUCTION

The uniform-tail reconstruction can substantially reduce distance relative to the original distributions. Pooling tail mass replaces distinctions among omitted token probabilities with a common shape, so information about their relative allocation is lost. These results support keeping reconstructed profiles as export checks and retaining original full-vocabulary probabilities for the main trajectory score.

For shared active token coordinates, the compressed inner-product formula is exact for the reconstructed distributions. It is not generally exact for the original softmax outputs. The comparison below reports reconstructed minus original distance; it is an export check, not a replacement for the main score.

Table A87: Qwen original and reconstructed Fisher–Rao distances. Tail A and B denote the omitted probability masses.

State	Tail A	Tail B	Original	Reconstructed	Δ
0	0.9933	0.9942	0.1577	0.0113	-0.1464
1	0.9932	0.9937	0.1380	0.0071	-0.1309
2	0.9919	0.9912	0.1525	0.0124	-0.1401
3	0.9891	0.9863	0.2057	0.0261	-0.1796
4	0.9916	0.9901	0.1513	0.0164	-0.1349
5	0.9828	0.9850	0.1483	0.0178	-0.1305
6	0.9942	0.9939	0.1269	0.0059	-0.1210
7	0.9955	0.9960	0.1998	0.0139	-0.1859
8	0.9960	0.9959	0.1276	0.0079	-0.1197
9	0.9944	0.9956	0.1613	0.0178	-0.1435
10	0.9960	0.9957	0.1378	0.0061	-0.1317
11	0.9860	0.9850	0.1287	0.0084	-0.1203

Table A88: DeepSeek original and reconstructed Fisher–Rao distances. Tail A and B denote the omitted probability masses.

State	Tail A	Tail B	Original	Reconstructed	Δ
0	0.9920	0.9931	0.1507	0.0144	-0.1364
1	0.9951	0.9947	0.1701	0.0102	-0.1599
2	0.9906	0.9918	0.1744	0.0209	-0.1535
3	0.9931	0.9944	0.1516	0.0166	-0.1350
4	0.9951	0.9952	0.1476	0.0047	-0.1430
5	0.9937	0.9938	0.1390	0.0067	-0.1323
6	0.9923	0.9926	0.1787	0.0102	-0.1686
7	0.9950	0.9949	0.1595	0.0065	-0.1530
8	0.9949	0.9948	0.1138	0.0026	-0.1112
9	0.9950	0.9951	0.1408	0.0064	-0.1344
10	0.9921	0.9937	0.1728	0.0201	-0.1527
11	0.9953	0.9952	0.1431	0.0136	-0.1295

Table A89: InternVL3 original and reconstructed Fisher–Rao distances. Tail A and B denote the omitted probability masses.

State	Tail A	Tail B	Original	Reconstructed	Δ
0	0.9916	0.9922	0.1470	0.0145	-0.1325
1	0.9957	0.9960	0.1617	0.0098	-0.1518
2	0.9951	0.9938	0.1681	0.0173	-0.1508
3	0.9911	0.9899	0.1532	0.0132	-0.1400
4	0.9960	0.9958	0.1513	0.0085	-0.1428
5	0.9926	0.9926	0.1616	0.0070	-0.1546
6	0.9955	0.9954	0.1382	0.0098	-0.1283
7	0.9916	0.9922	0.1563	0.0067	-0.1496
8	0.9888	0.9890	0.1166	0.0019	-0.1148
9	0.9895	0.9874	0.1727	0.0198	-0.1529
10	0.9931	0.9923	0.1264	0.0098	-0.1166
11	0.9944	0.9943	0.1250	0.0025	-0.1225

Table A90: LLaVA original and reconstructed Fisher–Rao distances. Tail A and B denote the omitted probability masses.

State	Tail A	Tail B	Original	Reconstructed	Δ
0	0.9940	0.9937	0.1230	0.0037	-0.1194
1	0.9942	0.9937	0.1385	0.0120	-0.1265
2	0.9930	0.9933	0.1040	0.0030	-0.1009
3	0.9949	0.9952	0.1564	0.0084	-0.1479
4	0.9828	0.9842	0.1252	0.0109	-0.1143
5	0.9922	0.9915	0.1255	0.0078	-0.1177
6	0.9936	0.9938	0.1394	0.0036	-0.1358
7	0.9946	0.9940	0.1596	0.0147	-0.1449
8	0.9922	0.9930	0.2054	0.0225	-0.1829
9	0.9950	0.9945	0.1480	0.0111	-0.1369
10	0.9953	0.9949	0.1557	0.0084	-0.1474
11	0.9957	0.9954	0.1609	0.0089	-0.1520

D.4 Computational Budgets and Cost

These experiments examine layer and prediction-position budgets; the accompanying workload records describe computational requirements rather than verified GPU timings.

D.4.1 Layer and position budgets

Table A91: Layer and prediction-position configurations. Cells report macro AUROC (%). K denotes the number of selected layers and T the maximum number of retained prediction positions. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

Configuration	Qwen	DeepSeek	InternVL3	LLaVA
$K = 2, T = 1$	83.77	85.46	83.12	84.43
$K = 2, T = 2$	83.94	85.53	83.10	84.42
$K = 2, T = 4$	83.94	85.53	83.10	84.42
$K = 3, T = 1$	**85.97**	**87.30**	**84.79**	**85.97**
$K = 3, T = 2$	<u>85.33</u>	<u>86.71</u>	<u>84.15</u>	<u>85.39</u>
$K = 3, T = 4$	<u>85.33</u>	<u>86.71</u>	<u>84.15</u>	<u>85.39</u>
$K = 4, T = 1$	**85.97**	**87.30**	**84.79**	**85.97**
$K = 4, T = 2$	<u>85.33</u>	<u>86.71</u>	<u>84.15</u>	<u>85.39</u>
$K = 4, T = 4$	<u>85.33</u>	<u>86.71</u>	<u>84.15</u>	<u>85.39</u>

With two layers, trajectory excess is identically zero. Equivalence to local-only scoring additionally requires the same prediction positions, local measurements, and fit parameters. The recorded values for $T = 2$ and $T = 4$ coincide at the reported precision. The $T = 1$ configuration has higher point estimates for $K = 3$ and $K = 4$, so this table does not show that retaining more positions always improves detection. The main configuration is retained rather than replaced on the basis of these test comparisons. No six-layer configuration is reported in the current workbook.

D.4.2 DETAILED LAYER AND POSITION CONFIGURATIONS

Increasing the number of retained positions does not consistently improve the recorded ranking metrics. In particular, the one-position settings have higher AUROC point estimates for three and four layers, while the two- and four-position settings coincide at the reported precision. The two-layer setting is a useful structural control because its trajectory excess is exactly zero. These comparisons characterize sensitivity to the configuration; they are not a basis for selecting a new configuration on the test set.

Each configuration uses its corresponding fit statistics. The recorded layer indices index the exported selected-layer sequence; they should not be read as original decoder block indices without the extraction configuration. No $K = 6$ result is supplied.

Table A92: Qwen configuration metrics (%) and fitted trajectory statistics. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

(K, T)	AUROC	AP	AURC	Path median	Path MAD
(2,1)	83.77	80.33	26.29	0.0000	0.0000
(2,2)	83.94	80.67	26.16	0.0000	0.0000
(2,4)	83.94	80.67	26.16	0.0000	0.0000
(3,1)	**85.97**	**83.08**	**24.10**	0.0617	0.0155
(3,2)	<u>85.33</u>	<u>82.44</u>	<u>24.59</u>	0.0619	0.0155
(3,4)	<u>85.33</u>	<u>82.44</u>	<u>24.59</u>	0.0619	0.0155
(4,1)	**85.97**	**83.08**	**24.10**	0.0617	0.0155
(4,2)	<u>85.33</u>	<u>82.44</u>	<u>24.59</u>	0.0619	0.0155
(4,4)	<u>85.33</u>	<u>82.44</u>	<u>24.59</u>	0.0619	0.0155

Table A93: DeepSeek configuration metrics (%) and fitted trajectory statistics. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

(K, T)	AUROC	AP	AURC	Path median	Path MAD
(2,1)	85.46	82.93	30.31	0.0000	0.0000
(2,2)	85.53	83.09	30.08	0.0000	0.0000
(2,4)	85.53	83.09	30.08	0.0000	0.0000
(3,1)	**87.30**	**85.22**	**28.64**	0.0639	0.0187
(3,2)	<u>86.71</u>	<u>84.68</u>	<u>28.79</u>	0.0642	0.0188
(3,4)	<u>86.71</u>	<u>84.68</u>	<u>28.79</u>	0.0642	0.0188
(4,1)	**87.30**	**85.22**	**28.64**	0.0639	0.0187
(4,2)	<u>86.71</u>	<u>84.68</u>	<u>28.79</u>	0.0642	0.0188
(4,4)	<u>86.71</u>	<u>84.68</u>	<u>28.79</u>	0.0642	0.0188

Table A94: InternVL3 configuration metrics (%) and fitted trajectory statistics. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

(K, T)	AUROC	AP	AURC	Path median	Path MAD
(2,1)	83.12	82.83	22.40	0.0000	0.0000
(2,2)	83.10	82.83	23.01	0.0000	0.0000
(2,4)	83.10	82.83	23.01	0.0000	0.0000
(3,1)	**84.79**	**84.39**	**20.61**	0.0629	0.0186
(3,2)	<u>84.15</u>	<u>83.88</u>	<u>21.58</u>	0.0630	0.0185
(3,4)	<u>84.15</u>	<u>83.88</u>	<u>21.58</u>	0.0630	0.0185
(4,1)	**84.79**	**84.39**	**20.61**	0.0629	0.0186
(4,2)	<u>84.15</u>	<u>83.88</u>	<u>21.58</u>	0.0630	0.0185
(4,4)	<u>84.15</u>	<u>83.88</u>	<u>21.58</u>	0.0630	0.0185

Table A95: LLaVA configuration metrics (%) and fitted trajectory statistics. Best and second-best displayed point estimates are bold and underlined; ties share the same marking. Interval endpoints are not ranked, and highlighting does not imply statistical significance.

(K, T)	AUROC	AP	AURC	Path median	Path MAD
(2,1)	84.43	81.20	17.85	0.0000	0.0000
(2,2)	84.42	81.52	18.12	0.0000	0.0000
(2,4)	84.42	81.52	18.12	0.0000	0.0000
(3,1)	**85.97**	**83.22**	**16.39**	0.0644	0.0170
(3,2)	<u>85.39</u>	<u>82.83</u>	<u>17.01</u>	0.0646	0.0171
(3,4)	<u>85.39</u>	<u>82.83</u>	<u>17.01</u>	0.0646	0.0171
(4,1)	**85.97**	**83.22**	**16.39**	0.0644	0.0170
(4,2)	<u>85.39</u>	<u>82.83</u>	<u>17.01</u>	0.0646	0.0171
(4,4)	<u>85.39</u>	<u>82.83</u>	<u>17.01</u>	0.0646	0.0171

D.4.3 COMPUTATIONAL COST

The cost sheet is not sufficient to establish measured GPU latency or memory use. The companion description identifies analytic workload assumptions, and measured peak memory is unavailable. Its millisecond values are therefore excluded from an empirical efficiency table. Hardware comparisons require actual timing measurements under matched conditions.

D.4.4 RECORDED WORKLOAD ESTIMATES

The recorded estimates include an additional cost for the geometric computation, but the absence of verified device timing and peak-memory measurements prevents a hardware-efficiency conclusion. Reporting cached and end-to-end scopes separately makes the accounting explicit without interpreting the estimates as measured latency.

The following values preserve the cost-sheet entries for completeness. They are workload estimates, not independently verified wall-clock measurements. Peak device memory is unavailable. Cached computation and end-to-end estimates use different accounting scopes; these values should not support a measured speedup claim.

Table A96: Qwen recorded workload estimates in milliseconds. Measured peak memory is unavailable for every listed method.

Method	Cached p50	End-to-end p50	Cached p90	End-to-end p90
NLL	0.07	235.07	0.07	273.57
LN-PE	0.07	235.07	0.07	273.57
Token entropy	0.07	235.07	0.07	273.57
SC	0.07	2350.07	0.07	2735.07
SE	0.07	2350.07	0.07	2735.07
HiddenDetect	0.07	235.07	0.07	273.57
EUQ-CF	0.14	235.14	0.14	273.64
EUQ-IG	0.14	235.14	0.14	273.64
Head Fisher trace	0.07	235.07	0.07	273.57
State Fisher trace	0.07	235.07	0.07	273.57
Token gradient	0.07	235.07	0.07	273.57
Layerwise JS	0.07	235.07	0.07	273.57
FRDG	4.57	239.57	4.57	278.07

Table A97: DeepSeek recorded workload estimates in milliseconds. Measured peak memory is unavailable for every listed method.

Method	Cached p50	End-to-end p50	Cached p90	End-to-end p90
NLL	0.06	235.06	0.06	273.56
LN-PE	0.06	235.06	0.06	273.56
Token entropy	0.06	235.06	0.06	273.56
SC	0.06	2350.06	0.06	2735.06
SE	0.06	2350.06	0.06	2735.06
HiddenDetect	0.06	235.06	0.06	273.56
EUQ-CF	0.12	235.12	0.12	273.62
EUQ-IG	0.12	235.12	0.12	273.62
Head Fisher trace	0.06	235.06	0.06	273.56
State Fisher trace	0.06	235.06	0.06	273.56
Token gradient	0.06	235.06	0.06	273.56
Layerwise JS	0.06	235.06	0.06	273.56
FRDG	3.93	238.93	3.93	277.43

Table A98: InternVL3 recorded workload estimates in milliseconds. Measured peak memory is unavailable for every listed method.

Method	Cached p50	End-to-end p50	Cached p90	End-to-end p90
NLL	0.07	235.07	0.07	273.57
LN-PE	0.07	235.07	0.07	273.57
Token entropy	0.07	235.07	0.07	273.57
SC	0.07	2350.07	0.07	2735.07
SE	0.07	2350.07	0.07	2735.07
HiddenDetect	0.07	235.07	0.07	273.57
EUQ-CF	0.14	235.14	0.14	273.64
EUQ-IG	0.14	235.14	0.14	273.64
Head Fisher trace	0.07	235.07	0.07	273.57
State Fisher trace	0.07	235.07	0.07	273.57
Token gradient	0.07	235.07	0.07	273.57
Layerwise JS	0.07	235.07	0.07	273.57
FRDG	4.56	239.56	4.56	278.06

Table A99: LLaVA recorded workload estimates in milliseconds. Measured peak memory is unavailable for every listed method.

Method	Cached p50	End-to-end p50	Cached p90	End-to-end p90
NLL	0.07	235.07	0.07	273.57
LN-PE	0.07	235.07	0.07	273.57
Token entropy	0.07	235.07	0.07	273.57
SC	0.07	2350.07	0.07	2735.07
SE	0.07	2350.07	0.07	2735.07
HiddenDetect	0.07	235.07	0.07	273.57
EUQ-CF	0.14	235.14	0.14	273.64
EUQ-IG	0.14	235.14	0.14	273.64
Head Fisher trace	0.07	235.07	0.07	273.57
State Fisher trace	0.07	235.07	0.07	273.57
Token gradient	0.07	235.07	0.07	273.57
Layerwise JS	0.07	235.07	0.07	273.57
FRDG	4.57	239.57	4.57	278.07

D.5 Supplementary Threshold Analysis

These supplementary analyses examine fixed-threshold decisions and repeated calibration rather than score ordering.

D.5.1 Supplementary threshold analysis

This analysis is retained for completeness and is not used to support the geometric design or component-complementarity claims. Ranking metrics evaluate score ordering, whereas the following results concern a specific threshold. The two evaluations answer different questions.

Table A100: Full-score operating points at the nominal level of 5%. Rates are in %. Correct FPR is conditional on correct holdout answers; TPR is measured on failed test answers.

Group	Alarms	Nominal rate	Correct FPR	TPR
Qwen	19/400	4.75	4.49	0.25
DeepSeek	47/400	11.75	10.75	59.53
InternVL3	32/400	8.00	8.49	52.63
LLaVA	10/400	2.50	2.52	0.00

Nominal records include failed answers. Consequently, nominal alarm control and correct-answer false-positive control are different statements. The conformal interpretation additionally requires exchangeability of calibration and target nominal records; disjoint identifiers alone do not establish this assumption.

D.5.2 Supplementary operating points and repeated calibration

The operating points show why ranking and threshold decisions must be assessed separately. A score can rank failures well while placing few of them above a threshold fitted to a different reference population. The mixed reference-pool alarm rate and the correct-answer false-positive rate must also remain distinct. These supplementary results do not support a claim that all four models achieve a fixed 5% correct-answer false-positive rate.

These results are supplementary and are not evidence that the geometric score guarantees correct answers. Reference-pool alarm rate and false-positive rate among correct answers have different denominators. Repeated-split quantiles below describe the supplied repetitions; they are not confidence intervals or independent replications.

Table A101: Qwen operating points at $\alpha = 0.05$. All rates are percentages. Holdout FPR is conditional on correct holdout answers.

Method	Pool alarms	Holdout FPR	Test TPR	Test FPR
NLL	5.75	5.69	16.12	6.09
LN-PE	6.75	6.29	54.30	28.33
Token entropy	9.75	9.58	64.32	27.28
SC	1.50	0.90	33.02	2.44
SE	4.25	4.19	41.62	23.00
HiddenDetect	2.50	1.80	39.75	16.73
EUQ-CF	6.25	7.19	54.00	23.16
EUQ-IG	4.75	5.39	49.04	23.94
Head Fisher trace	4.25	4.19	54.59	23.61
State Fisher trace	5.00	5.69	59.90	24.52
Token gradient	2.75	2.99	56.02	22.93
Layerwise JS	4.50	3.89	57.99	26.23
FRDG	4.75	4.49	0.25	0.00

Table A102: Qwen FRDG threshold levels. Rates and reported 95% intervals are percentages.

α	Threshold	Pool alarm [CI]	Correct FPR [CI]	Test TPR [CI]
0.01	19.28	0.50 [0.14, 1.80]	0.60 [0.16, 2.16]	0.00 [0.00, 0.19]
0.02	16.53	1.25 [0.54, 2.89]	1.50 [0.64, 3.46]	0.00 [0.00, 0.19]
0.05	8.10	4.75 [3.06, 7.30]	4.49 [2.74, 7.28]	0.25 [0.10, 0.57]
0.10	6.11	8.75 [6.36, 11.93]	8.68 [6.11, 12.19]	47.62 [45.45, 49.79]

Table A103: Qwen repeated calibration: reference-pool alarm rates (%). Quantiles use lower order statistics.

α	Repetitions	Mean	Median	2.5–97.5% range
0.01	100	0.95	0.75	0.00–2.50
0.02	100	1.88	1.75	0.25–4.50
0.05	100	4.71	4.62	2.00–9.00
0.10	100	9.51	9.50	5.00–14.00

Table A104: DeepSeek operating points at $\alpha = 0.05$. All rates are percentages. Holdout FPR is conditional on correct holdout answers.

Method	Pool alarms	Holdout FPR	Test TPR	Test FPR
NLL	10.00	8.96	15.80	6.19
LN-PE	8.25	8.66	53.11	29.45
Token entropy	5.50	5.37	59.29	22.24
SC	2.50	2.99	26.56	3.68
SE	1.25	0.90	41.75	16.53
HiddenDetect	4.00	3.88	46.27	23.72
EUQ-CF	2.50	2.39	53.30	22.99
EUQ-IG	4.25	3.88	47.50	19.52
Head Fisher trace	6.75	7.16	57.03	24.89
State Fisher trace	8.50	8.66	62.64	28.38
Token gradient	5.50	5.67	55.99	23.84
Layerwise JS	4.50	3.58	55.90	24.84
FRDG	11.75	10.75	59.53	16.96

Table A105: DeepSeek FRDG threshold levels. Rates and reported 95% intervals are percentages.

α	Threshold	Pool alarm [CI]	Correct FPR [CI]	Test TPR [CI]
0.01	31.95	0.25 [0.04, 1.40]	0.00 [0.00, 1.13]	0.00 [0.00, 0.18]
0.02	8.24	5.25 [3.46, 7.89]	4.48 [2.73, 7.26]	0.09 [0.03, 0.34]
0.05	5.27	11.75 [8.95, 15.28]	10.75 [7.86, 14.52]	59.53 [57.42, 61.60]
0.10	4.17	17.25 [13.86, 21.26]	16.12 [12.57, 20.44]	72.69 [70.75, 74.54]

Table A106: DeepSeek repeated calibration: reference-pool alarm rates (%). Quantiles use lower order statistics.

α	Repetitions	Mean	Median	2.5–97.5% range
0.01	100	1.01	0.88	0.00–2.50
0.02	100	2.10	2.00	0.75–4.75
0.05	100	5.05	5.00	2.25–8.50
0.10	100	9.87	9.62	6.00–14.00

Table A107: InternVL3 operating points at $\alpha = 0.05$. All rates are percentages. Holdout FPR is conditional on correct holdout answers.

Method	Pool alarms	Holdout FPR	Test TPR	Test FPR
NLL	2.50	1.89	11.37	2.73
LN-PE	6.75	6.29	57.54	24.17
Token entropy	2.75	2.20	66.97	12.78
SC	3.25	3.46	47.29	5.96
SE	6.50	5.97	57.62	21.80
HiddenDetect	3.50	3.77	50.43	19.58
EUQ-CF	4.75	4.40	55.30	16.40
EUQ-IG	4.25	4.09	58.27	20.37
Head Fisher trace	6.50	6.29	56.80	26.71
State Fisher trace	6.50	6.29	57.92	25.01
Token gradient	3.75	3.46	54.82	24.32
Layerwise JS	5.25	5.66	56.93	24.01
FRDG	8.00	8.49	52.63	1.05

Table A108: InternVL3 FRDG threshold levels. Rates and reported 95% intervals are percentages.

α	Threshold	Pool alarm [CI]	Correct FPR [CI]	Test TPR [CI]
0.01	12.91	2.50 [1.36, 4.54]	2.83 [1.50, 5.29]	0.00 [0.00, 0.17]
0.02	9.22	4.00 [2.48, 6.40]	4.72 [2.88, 7.64]	0.00 [0.00, 0.17]
0.05	6.24	8.00 [5.72, 11.08]	8.49 [5.90, 12.07]	52.63 [50.59, 54.65]
0.10	4.95	10.25 [7.65, 13.61]	11.01 [8.02, 14.92]	73.04 [71.20, 74.81]

Table A109: InternVL3 repeated calibration: reference-pool alarm rates (%). Quantiles use lower order statistics.

α	Repetitions	Mean	Median	2.5–97.5% range
0.01	100	0.95	0.75	0.00–2.25
0.02	100	1.93	1.75	0.25–4.00
0.05	100	4.94	4.62	2.00–9.25
0.10	100	9.83	9.38	5.25–14.75

Table A110: LLaVA operating points at $\alpha = 0.05$. All rates are percentages. Holdout FPR is conditional on correct holdout answers.

Method	Pool alarms	Holdout FPR	Test TPR	Test FPR
NLL	4.75	4.73	15.47	4.76
LN-PE	5.50	6.94	55.59	22.81
Token entropy	1.00	0.63	56.69	6.90
SC	1.75	2.21	38.86	3.78
SE	6.75	6.94	57.18	27.44
HiddenDetect	6.25	6.31	56.36	24.04
EUQ-CF	2.25	2.52	54.27	16.74
EUQ-IG	8.00	8.52	64.56	27.57
Head Fisher trace	5.75	4.73	55.09	23.89
State Fisher trace	4.75	4.10	60.10	24.98
Token gradient	5.75	6.62	57.29	24.30
Layerwise JS	3.25	3.79	58.61	25.67
FRDG	2.50	2.52	0.00	0.00

Table A111: LLaVA FRDG threshold levels. Rates and reported 95% intervals are percentages.

α	Threshold	Pool alarm [CI]	Correct FPR [CI]	Test TPR [CI]
0.01	22.82	0.25 [0.04, 1.40]	0.32 [0.06, 1.76]	0.00 [0.00, 0.21]
0.02	22.31	0.25 [0.04, 1.40]	0.32 [0.06, 1.76]	0.00 [0.00, 0.21]
0.05	10.57	2.50 [1.36, 4.54]	2.52 [1.28, 4.90]	0.00 [0.00, 0.21]
0.10	5.52	6.75 [4.68, 9.64]	7.26 [4.88, 10.65]	67.91 [65.73, 70.02]

Table A112: LLaVA repeated calibration: reference-pool alarm rates (%). Quantiles use lower order statistics.

α	Repetitions	Mean	Median	2.5–97.5% range
0.01	100	1.36	1.25	0.00–3.25
0.02	100	2.18	2.00	0.50–5.00
0.05	100	5.04	5.00	1.75–8.00
0.10	100	10.16	10.00	5.50–15.25