FOSTER: FIXED-THRESHOLD SAFETY TOPOLOGY REPAIR FOR VLM JAILBREAK DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deployed jailbreak detectors for vision–language models (VLMs) must apply a single decision threshold, chosen *before* unseen attacks appear, to open-world traffic. We formalize this constraint as *fixed-threshold transfer failure*: an unseen attack can remain rank-separable from benign inputs (high AUROC) while its scores drift below the pre-deployment threshold (low TPR). Despite this global displacement, unseen attacks preserve a transferable *local unsafe topology* in mid-to-late VLM hidden layers, so a limited number of trustworthy anchors can be sufficient to improve subsequent detection without retraining or threshold recalibration, while keeping the benign reference fixed. Building on this, we propose FOSTER, which freezes the VLM, the safety evidence router, the benign reference, and all pre-deployment thresholds, while selectively expanding only a cross-verified unsafe topology bank through mode-preserving anchor admission. Anchors are cross-checked, pseudo-labeled unsafe samples, admitted only when the current topology, the frozen initial topology, and a separately parameterized safety estimator all satisfy their own pre-calibrated admission criteria, which are distinct from the prediction and routing thresholds. The three witnesses are complementary rather than statistically independent; their purpose is to reduce self-reinforcing admission errors. We evaluate FOSTER under fixed calibration thresholds and streaming deployment on Qwen2.5-VL and LLaVA-v1.6. Under each detector's own pre-deployment threshold, pooled attack recall rises from 10.5% to 80.7% on Qwen2.5-VL and from 6.8% to 71.3% on LLaVA-v1.6, with zero observed false positives on the 917-sample deployment benign pool (a 95% Wilson upper bound of 0.42%, not a guarantee of zero benign error). Structurally, freezing the admission witnesses fixes the eligibility set before deployment, which yields a stream-independent outer envelope on any input's online topology score—a conditional bound rather than a deployable certificate, since eligibility does not imply admission. Because the benign reference bank is frozen by construction, benign-side drift is measured rather than bounded: under out-of-distribution benign pools and targeted benign contamination we report the resulting change in false positives explicitly. Our code is available at an anonymous repository: https://anonymous.4open.science/r/FOSTER.

1 INTRODUCTION

Vision–language models (VLMs) power a rapidly growing class of multimodal assistants, and multimodal jailbreak attacks that exploit both channels can elicit harmful outputs at high success rates on aligned models (Qi et al., 2024; Gong et al., 2023; Liu et al., 2024c; Shayegani et al., 2023). Recent detectors span input/output checks (Xu et al., 2024; Zhang et al., 2025; Fares et al., 2024), internal-representation methods (Jiang et al., 2025; Liang et al., 2026; Xie et al., 2024), and test-time adaptive detectors (Nian et al., 2025), and per-attack AUROC results in the literature routinely approach 1.0. Yet a safety monitor must fix one threshold *before* deployment and apply it to open-world traffic without attack-family labels or per-attack tuning, and high per-attack AUROC does not by itself imply that this fixed threshold still separates unseen attacks from benign inputs.

We refer to the surface phenomenon as the *AUC illusion* but study its underlying cause as *fixed-threshold transfer failure*: unseen attacks may preserve ranking against benign inputs ($\mathrm{AUROC}(P_u, P_b)$ high) while their score mass shifts below the calibration-time boundary ($\mathrm{TPR}(P_u; \theta_{\mathrm{cal}})$ low). Figure 1 illustrates this collapse: LoD detects only 12.8% of attacks under its

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.

own calibrated threshold on Qwen2.5-VL, and HiddenDetect drops to 2.9%. This is a deployment failure, not a minor metric issue: a safety monitor cannot rescore each attack family with its own threshold at test time. More generally, a deployed detector must commit to a fixed operating policy before unseen attacks appear. For a scalar-score detector, this policy is represented by a single threshold; for a cascaded detector such as FOSTER, it consists of a fixed set of pre-calibrated routing, decision, and admission thresholds. None of these quantities is re-tuned for a newly observed attack family.

Classifying detectors by *what they are allowed to modify during deployment* exposes why existing responses are insufficient. Static representation detectors keep score function and threshold fixed and cannot repair novel-attack coverage. Threshold recalibration requires deployment-time labels unavailable in open-world monitoring. Parametric adaptation (gradient-based TTA) updates model or detector parameters, risking global decision drift and pseudo-label contamination. Generic memory adaptation updates all classes symmetrically without protecting the benign reference or accounting for the asymmetric cost of false positives.

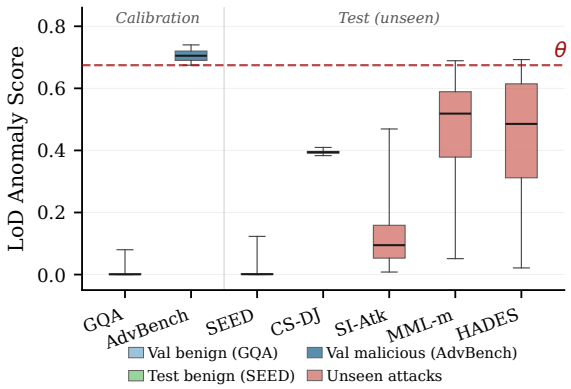

Figure 1: Fixed-threshold transfer failure of LoD (Liang et al., 2026) on Qwen2.5-VL. Benign inputs (green) score near zero; the calibration attack (AdvBench, blue) sits above the deployment threshold $\theta = 0.67$. Unseen attacks (orange) remain *rank-separable* from benign traffic—per-attack AUROC$\approx$1—yet fall below θ, so decision recall collapses (TPR$\approx$0%). The threshold is set from known traffic and cannot be re-tuned per attack family after deployment.

The missing ingredient is not stronger supervision. New attacks do not necessarily require a new global boundary—they often require only a small number of trustworthy *unsafe* anchors to repair their local topology. In the mid-to-late hidden layers of aligned VLMs, unseen attacks retain a transferable local safety topology even when their scores shift globally: once a few high-confidence anchors are admitted, subsequent samples from the same mode retrieve closer unsafe neighbors, their per-layer risk climbs, and the fixed threshold recovers recall. Prediction and admission, however, are different operations: a sample may be predicted unsafe yet fail to qualify as a memory anchor, and the admission rule is what keeps the online topology safe from drift.

We propose FOSTER (Fixed-Threshold Online Safety Topology Expansion and Repair), a safety-constrained topology-repair framework. FOSTER *freezes* the VLM, a calibrated cross-layer safety evidence router, the benign reference bank, and the pre-deployment operating policy. It *expands* only an online unsafe topology bank, and only through samples for which the current bank, the frozen initial bank, and the parametric router all satisfy their respective pre-calibrated admission criteria. These criteria are separate from the prediction threshold and from the routing threshold, so an admitted anchor is a cross-checked, pseudo-labeled unsafe sample rather than simply a positive prediction. The bank is mode-preserving so that a dominant attack family cannot crowd out rare or emerging unsafe modes. FOSTER's methodological distinction lies in its asymmetric update space: only cross-verified unsafe topology may evolve, whereas the model parameters, calibration policy, and benign reference remain fixed. Our contributions are summarized as follows:

(1) We formulate *fixed-threshold transfer failure* in VLM jailbreak detection and separate it operationally from AUROC, providing evidence across nine attack types and multiple VLMs that ranking quality and unified-threshold recall dissociate for state-of-the-art detectors.

(2) We build FOSTER on the hypothesis that unseen attacks retain transferable, layer-dependent local safety topology despite global score displacement, and that this topology is exploitable with a small number of cross-checked anchors; the deployment experiments test the resulting method rather than the geometric hypothesis in isolation.

(3) We propose FOSTER, a safety-constrained topology repair framework that defines an asymmetric update space: the detector, pre-deployment operating policy, and benign reference remain

Table 1: Detectors compared by what they *modify during deployment*. "Model Params" refers to the VLM and detector weights; "Unsafe Memory" refers to stored unsafe reference content, which is data rather than parameters. Among the compared methods, FOSTER is the only one that combines a fully frozen deployment policy with unsafe-only memory updates admitted through cross-verification.

Method	Model Params	Thresh.	Safe Ref.	Unsafe Memory	Verif. Admis.
HiddenDetect	Frozen	Frozen	N/A	N/A	✗
LoD	Frozen	Frozen	N/A	N/A	✗
GradSafe	Frozen	Frozen	N/A	N/A	✗
JailDAM	Frozen	Frozen	N/A	Updated	✗
AdaNPC-style TTA	Frozen	Frozen	Updated	Updated	✗
FOSTER	**Frozen**	**Frozen**	**Frozen**	**Updated**	✓

frozen, while only cross-verified unsafe support may evolve. This principle is instantiated through layer-wise topology scoring, dual-reference unsafe admission, and mode-preserving memory. We give a mechanistic analysis of this update space: the frozen witnesses bound any input's online topology score by a stream-independent envelope, admitted anchors obey a checkable cone constraint on the pre-deployment geometry, and we give a sufficient condition for local repair plus a neighborhood-turnover bound on decision stability.

(4) We evaluate FOSTER under fixed calibration thresholds and unlabeled streaming deployment on Qwen2.5-VL and LLaVA-v1.6, and further quantify its behavior under targeted benign contamination and out-of-distribution benign shift.

2 Related Work

VLM jailbreak attacks and detection. Multimodal attacks span adversarial perturbation (Qi et al., 2024; Zou et al., 2023), typographic injection (Gong et al., 2023; Liu et al., 2024c), cross-modal manipulation (Shayegani et al., 2023; Niu et al., 2024), and visual-vulnerability exploitation (Li et al., 2024); recent work has moved from white-box to more realistic black-box settings, making attack diversity central (Yi et al., 2024; Liu et al., 2024a). Existing detectors fall into three groups: input/output detectors (CIDER (Xu et al., 2024), JailGuard (Zhang et al., 2025), MirrorCheck (Fares et al., 2024)), internal-representation detectors (HiddenDetect (Jiang et al., 2025), LoD (Liang et al., 2026), GradSafe (Xie et al., 2024)), and test-time adaptive detectors (JailDAM (Nian et al., 2025)); defense methods (Wang et al., 2024; Chen et al., 2025; Gou et al., 2024) modify model behavior rather than detect. Several of these use multi-layer hidden-state evidence (HiddenDetect via refusal-logit signals across layers, LoD via per-layer MSCAV autoencoders), but all of them keep their score function and threshold fixed after calibration. While the distinction between ranking metrics and fixed operating points is well understood in general detection, its consequences for attack-family transfer in VLM jailbreak detection remain underexplored in the literature.

Non-parametric adaptation and closest methods. Parametric test-time adaptation updates model parameters through entropy minimization (Wang et al., 2021) or self-supervised objectives (Sun et al., 2020). Non-parametric alternatives such as AdaNPC (Zhang et al., 2023) use KNN memory but treat classes symmetrically, and Sun et al. (2022) show that KNN distances provide strong OOD signals. FOSTER inherits the locality principle of non-parametric adaptation but restricts *what* may adapt. JailDAM performs training-free test-time adaptation by updating policy-derived unsafe concept memories. FOSTER maintains layer-wise VLM representation banks and constrains anchor admission using current-topology, frozen-reference, and router evidence; symmetric KNN-TTA updates both class memories, and static KNN detectors update nothing. Training-free memory updating is therefore not by itself what distinguishes FOSTER: the distinction lies in scoring layer-wise neighborhoods inside the VLM hidden states, retaining a frozen initial unsafe reference, gating admission on three pre-calibrated criteria, and managing memory so that unsafe modes are preserved (Table 1).

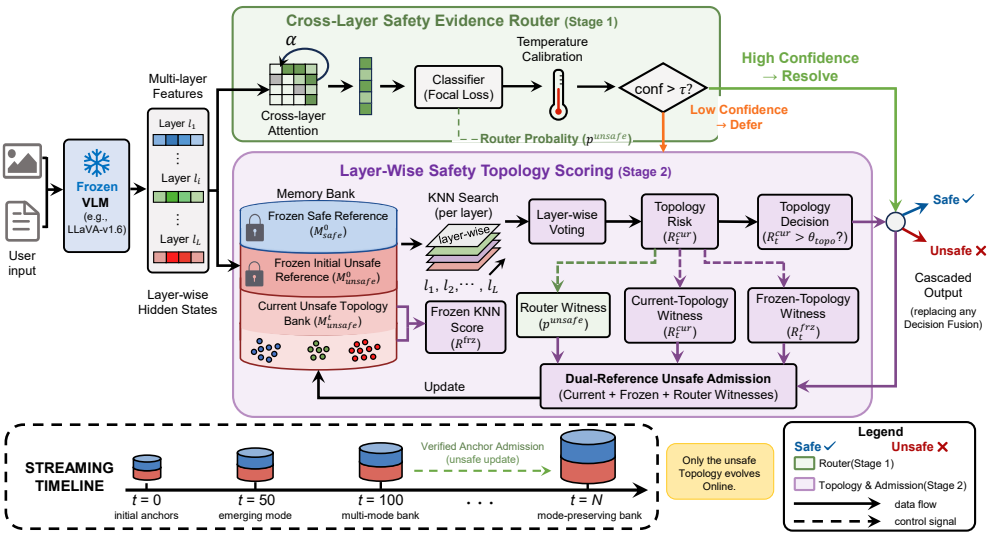

Figure 2: **Overview of FOSTER.** The frozen VLM and safety evidence router first process multi-layer representations. High-confidence samples are resolved directly, while uncertain samples undergo layer-wise topology scoring. Only cross-verified unsafe anchors are admitted into the mode-preserving unsafe topology bank; the model, benign reference, initial unsafe reference, and all deployment thresholds remain fixed throughout the stream.

3 Fixed-Threshold Transfer Failure and Latent Safety Topology

Let P_b, P_u denote the score distributions of a detector on benign and unseen-attack inputs, and θ_{cal} the threshold calibrated on training-time traffic (benign + one known attack). A useful detector must preserve both *rank separation* $\text{AUROC}(P_u, P_b)$ and *fixed-threshold recall* $\text{TPR}(P_u; \theta_{\text{cal}})$ on unseen attacks. The AUC illusion (Figure 1) is the empirical dissociation of these two quantities: a detector can have $\text{AUROC} \to 1$ while $\text{TPR}(P_u; \theta_{\text{cal}}) \to 0$. High AUROC means that the detector score retains rank separation between unsafe and benign inputs; low fixed-threshold TPR means that this ordering is no longer aligned with the deployment boundary.

Fixed-threshold failure does not, however, imply that hidden-state geometry is broken. Three lightweight diagnostics (protocol in the supplementary material) motivate the design hypotheses below; we state them as the working assumptions behind FOSTER's architecture rather than as separately validated empirical findings, because the quantitative per-layer diagnostic results are not reported in this version. First, we posit that the normalized inter/intra separation $\rho_l = \bar{d}_{\cos}(\mathcal{H}_l^{\text{safe}}, \mathcal{H}_l^{\text{unsafe}}) / \frac{1}{2}(\sigma(\mathcal{H}_l^{\text{safe}}) + \sigma(\mathcal{H}_l^{\text{unsafe}}))$ remains informative in mid-to-late layers relative to a random-label control. Second, initializing the unsafe reference with a single calibration attack (AdvBench) and running plain KNN on held-out attack families *without their labels* is expected to retain rank separation from benign traffic, i.e. the geometry *transfers* across families. Third, the informative depth is not assumed stable across attack families, which is why we score a layer range instead of committing to one depth.

These considerations motivate two structural choices used by FOSTER: (i) scoring unsafe evidence *layer-wise* rather than committing to a fixed depth or concatenating layers, and (ii) aggregating across layers with an input-dependent router that lets each sample use the layers that matter for it. The two aggregation mechanisms are distinct and should not be conflated: the router computes input-dependent attention weights over layers, whereas the topology scorer aggregates per-layer KNN risk with *fixed* weights w_l. Under fixed deployment invariants these choices become the mechanism through which a small set of cross-verified unsafe anchors can *repair* the local decision geometry without touching the model, the threshold, or the benign reference.

4 FOSTER: FIXED-THRESHOLD SAFETY TOPOLOGY REPAIR

4.1 DEPLOYMENT INVARIANTS

We first state what FOSTER is *not* allowed to change. Let f^t_{VLM} be the VLM at time t, f^t_{router} the safety evidence router, and $\mathcal{M}^t_{\mathrm{safe}}$, $\mathcal{M}^t_{\mathrm{unsafe}}$ the benign and unsafe reference banks. FOSTER enforces

$$\begin{aligned} f^t_{\mathrm{VLM}} &= f^0_{\mathrm{VLM}}, \quad f^t_{\mathrm{router}} = f^0_{\mathrm{router}}, \\ \mathcal{M}^t_{\mathrm{safe}} &= \mathcal{M}^0_{\mathrm{safe}}, \end{aligned} \tag{1}$$

and admits changes only to $\mathcal{M}^t_{\mathrm{unsafe}}$, under the Dual-Reference Unsafe Admission rule stated below. All decision, routing and admission thresholds—collectively denoted Π_0—are calibrated once before deployment and remain fixed throughout the stream. Our goal is to improve fixed-policy recall on recurring unsafe modes while limiting the resulting change in benign false positives; the supplementary material formalizes this as a repair-gain / benign-drift decomposition.

4.2 CROSS-LAYER SAFETY EVIDENCE ROUTER

The router aggregates layer-wise VLM hidden states into a calibrated unsafe probability. Given hidden states $\{\mathbf{h}_l\}_{l\in\mathcal{L}}$ from a mid-to-late layer subset $\mathcal{L}$,

$$\alpha_l = \frac{\exp(g_\phi(\mathbf{h}_l))}{\sum_{l'\in\mathcal{L}} \exp(g_\phi(\mathbf{h}_{l'}))}, \quad \mathbf{z} = \sum_{l\in\mathcal{L}} \alpha_l \, \mathbf{h}_l, \tag{2}$$

where $g_\phi : \mathbb{R}^d \to \mathbb{R}$ is a small MLP. The aggregated $\mathbf{z}$ is passed through a bottleneck head to yield a temperature-calibrated probability $p^{\mathrm{unsafe}}(x)$. The router is trained once on the calibration split with focal loss and post-hoc temperature scaling and is *frozen* at deployment (training details in the supplementary material). It plays three coupled roles: it resolves high-confidence samples directly at threshold θ_{router}, provides a calibrated unsafe probability for routing, and acts as a separately parameterized safety witness in the admission gate below. This last role—using the router's decision as an admission check rather than as a stand-alone classifier—is what turns cross-layer aggregation into a constraint on the online update space rather than another output head.

4.3 LAYER-WISE SAFETY TOPOLOGY SCORING

For samples the router leaves uncertain, we score unsafe risk in a way that preserves per-layer geometry. The two modules use multi-layer evidence differently: the router performs input-dependent aggregation to produce a globally calibrated confidence score, whereas the topology scorer below preserves separate per-layer neighborhoods and combines their risks using fixed prior weights. For each layer l, let $\mathcal{N}^{\pm}_k(\mathbf{h}_l)$ be the k nearest neighbors in $\mathcal{M}^t_{\mathrm{unsafe}}$ and $\mathcal{M}^t_{\mathrm{safe}}$, and

$$\begin{aligned} s^{\pm}_l(x) &= \tfrac{1}{k} \sum_{j\in\mathcal{N}^{\pm}_k(\mathbf{h}_l)} \kappa(\mathbf{h}_l, \mathbf{h}^{(j)}_l), \\ \kappa(\mathbf{a},\mathbf{b}) &= \max(\cos(\mathbf{a},\mathbf{b}), 0). \end{aligned} \tag{3}$$

The per-layer safety-topology risk and its cross-layer aggregation are

$$r_l(x) = \frac{s^+_l(x) - s^-_l(x) + 1}{2}, \quad R_t(x) = \sum_{l\in\mathcal{L}} w_l \, r_l(x), \tag{4}$$

with fixed mid-to-late prior weights $w_l \geq 0$, $\sum_l w_l = 1$. Scoring per layer, rather than concatenating layers into one high-dimensional vector, preserves layer-specific structure (D3), keeps neighborhoods interpretable, and prevents any single overfitted layer from dominating the score.

4.4 DUAL-REFERENCE UNSAFE ADMISSION

The central design principle of FOSTER is that *prediction and admission are different operations*. Predicting a sample as unsafe (via $R_t(x)$ crossing the fixed topology threshold $\theta_{\mathrm{topo}} \in \Pi_0$) does not entitle it to enter the online bank. Admission requires the following triple:

$$\begin{aligned} \mathrm{Admit}(x) \iff\ & R^{\mathrm{cur}}_t(x) > \tau_{\mathrm{cur}} \\ \wedge\ & R^{\mathrm{frz}}_t(x) > \tau_{\mathrm{frz}} \\ \wedge\ & p^{\mathrm{unsafe}}_t(x) > \tau_{\mathrm{cls}}, \end{aligned} \tag{5}$$

where R_t^{cur} uses the evolving bank $\mathcal{M}_{\mathrm{unsafe}}^t$, R_t^{frz} uses the frozen initial bank $\mathcal{M}_{\mathrm{unsafe}}^0$, and p_t^{unsafe} is the router's separately parameterized safety probability. Each witness plays a distinct role. R_t^{cur} lets FOSTER exploit newly discovered modes: once a few anchors of an emerging attack are admitted, later samples retrieve closer unsafe neighbors and their scores rise. R_t^{frz} discourages the evolving memory from self-certifying its own drift—an anchor is admitted only if it also looks unsafe under the frozen initial topology. p_t^{unsafe} anchors admission to the calibration distribution through a parameter path that streamed samples cannot rewrite. Together the three checks are complementary rather than statistically independent, and they reduce self-reinforcing admission errors on the streams we evaluate. The frozen witness requires only weak compatibility with the initial unsafe geometry, so FOSTER can expand to geometrically novel modes but may still miss attacks whose hidden-state representation is close to benign under that initial geometry (e.g., SI-Attack)—a deliberate recall–robustness trade-off.

4.5 Mode-Preserving Unsafe Topology Bank

We further constrain *how* admitted anchors enter the bank. The unsafe topology bank is cluster-structured,

$$\mathcal{M}_{\mathrm{unsafe}}^t = \bigcup_{c=1}^{C_t} \mathcal{M}_c, \tag{6}$$

where a new cluster is seeded when an admitted sample's maximum cosine similarity to cluster centroids falls below δ. Each cluster has a per-mode capacity N_c and reservoir-style eviction; the overall bank capacity is N_u. No attack-family labels are used. The purpose is not compression but *mode preservation*: the cluster-structured bank limits how far a single dominant mode can concentrate the admitted anchors, and is intended to retain diverse local unsafe support under a finite memory budget. Its effect on a particular query depends on which of that query's relevant neighbors survive eviction, as quantified by Proposition 4. Clustering also gives, as a byproduct, a natural signal for auditing which unsafe modes have accumulated so far.

4.6 Streaming Inference

The full streaming procedure is provided in the supplementary material. Everything except $\mathcal{M}_{\mathrm{unsafe}}^t$ is frozen for the whole deployment: no parameter update, no threshold recalibration, no router retraining. Prediction and admission stay decoupled—the router uses θ_{router}, the topology score uses θ_{topo}, and admission uses the triple $(\tau_{\mathrm{cur}}, \tau_{\mathrm{frz}}, \tau_{\mathrm{cls}})$—so samples that fall below any single admission threshold can still be predicted unsafe by the fixed policy while remaining excluded from the online bank. Because the admission gate is strict, FOSTER does not need to correctly flag every early sample from an unseen family: once sufficiently similar anchors are admitted and retained, subsequent deferred queries from the same local topology obtain increased local unsafe support, and Proposition 1 gives a sufficient condition under which this increase exceeds a query's deficit to the fixed threshold. Coverage expansion therefore depends on seed eligibility, local similarity, and memory retention rather than on adaptation succeeding by default. Combined with the mode-preserving bank, this yields a cold-start behavior in which coverage grows unsafe-mode by unsafe-mode, without touching the benign reference or the operating policy Π_0.

What the frozen admission boundary buys. Freezing the witnesses constrains what the detector can become, not only how it is updated. Since R^{frz} and p^{unsafe} never change, the set $\mathcal{E}$ of inputs *eligible* to enter memory is fixed before deployment, so the unsafe bank always stays inside $\mathcal{M}_{\mathrm{unsafe}}^0 \cup \mathcal{E}$. This yields a stream-independent *outer envelope* on any input's online topology score: no stream order, injected content, or eviction history can push $R_t(x)$ above a quantity fixed by Π_0 and $\mathcal{M}_{\mathrm{unsafe}}^0$ (Proposition 2). Eligibility is necessary but not sufficient for admission—the dynamic τ_{cur} gate and finite capacities make the reached set a strict subset of $\mathcal{E}$—so this is a conditional bound, not a deployable certificate, and we neither compute it nor report certified coverage. Its per-anchor version *is* checkable: every admitted anchor lies in a cone around the pre-deployment unsafe geometry (Corollary 3), the precise sense in which the evolving memory cannot relax its own admission standard. The contrast with threshold-recalibrating admission is stream-independence, not the existence of a bound. Appendix I proves these statements, gives a sufficient condition under which admitted anchors repair a fixed-threshold miss (Proposition 1), bounds score change by turnover inside a query's own neighborhood (Proposition 4), and states the limits: containment is not correctness, and score monotonicity applies to benign inputs too, so benign drift is measured rather than bounded.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models and data. We conduct full deployment-oriented evaluation on Qwen2.5-VL-7B-Instruct (Qwen2.5-VL) (Bai et al., 2025) and LLaVA-v1.6-Vicuna-7B (LLaVA-v1.6) (Liu et al., 2023; 2024b). The attack suite contains nine datasets covering four families: adversarial perturbation (SafetyBench-VAJM, SafetyBench-UMK (Liu et al., 2024c), HADES (Li et al., 2024)), typographic injection (FigImg (Gong et al., 2023), MM-SafetyBench), cross-modal manipulation (CS-DJ, SI-Attack (Shayegani et al., 2023)), and semantic/composite attacks (MML-m, MML-m_all (Wang et al., 2025)). GQA (Hudson & Manning, 2019) is the benign source for per-attack AUROC and detector calibration; SEED-Bench (Li et al., 2023) and mm-vet2 (Yu et al., 2024) are the benign sources at deployment; AdvBench (Zou et al., 2023) is the sole calibration attack. Benign-shift experiments use VQA-v2 (Goyal et al., 2017).

Baselines. We compare against detectors from every family: input/output-based (MirrorCheck (Fares et al., 2024), CIDER (Xu et al., 2024)), mutation-based (JailGuard (Zhang et al., 2025), two variants), internal-representation (GradSafe (Xie et al., 2024), HiddenDetect (Jiang et al., 2025) with two layer configurations, LoD (Liang et al., 2026)) and test-time adaptive (JailDAM (Nian et al., 2025)). All entries are computed on the nine-attack suite.

Protocols and metrics. We report five deployment-oriented protocols aligned with the fixed-threshold view. **Per-attack AUROC** (Table S1) diagnoses whether ranking quality is already saturated. **Fixed-threshold transfer** (Table 2, Cal. column) is our primary deployment protocol: each detector's own pre-deployment calibration threshold is applied unchanged. **Oracle TPR@0FPR** (Tables 2, S2) is the maximum recall achievable by thresholding the detector's scores subject to zero observed false positives on the evaluated benign pool. The operating point uses evaluation data, so it is a *diagnostic*, not a deployable protocol. For the cascaded FOSTER scorer the router path and its thresholds stay fixed and only θ_{topo} is scanned, over scores recorded during the prequential run. **Streaming (S-LOAO)**: the bank is initialized with benign + one known attack, and the remaining attacks arrive as unlabeled streams. Each sample is classified using the memory state available before its arrival, and its prediction is recorded before any admission or memory update; online metrics are computed from these recorded predictions. This protocol is designed as a controlled evaluation of topology repair over recurring unsafe modes, rather than as an estimate of attack prevalence in production traffic. **Robustness**: benign distribution shift and adversarial contamination. Unless stated otherwise, we use $k = 20$, $\tau_{\text{conf}} = 0.7$, $\tau_{\text{cur}} = 0.8$, $\tau_{\text{frz}} = 0.5$, $\tau_{\text{cls}} = 0.3$, capacities $N_s = 500$, $N_c = 1000$; all hyperparameters and thresholds are tuned only on the calibration data and then frozen as Π_0.

Experimental logic. The experiments are organized to match the paper's motivation. Table S1 establishes that per-attack AUROC is already saturated for strong representation detectors, including our own. Tables 2–S2 then replace ranking with deployment metrics—TPR, FPR, and Oracle TPR@0FPR—to test whether AUROC success survives a single-threshold operating point. *Online Unsafe Topology Repair* isolates whether the gain comes from online repair rather than from a friendlier threshold. Table 3 then bundles two additional robustness checks: *Safety-Constrained Admission under Contamination* (Panel A) evaluates gate robustness, and *Robustness under Benign Distribution Shift* (Panel B) evaluates benign-side drift. Finally, Table 4 decomposes the contribution of each component.

5.2 PER-ATTACK AUROC IS SATURATED

Table S1 reports per-attack AUROC on Qwen2.5-VL and LLaVA-v1.6 with GQA as benign. This table is not our main deployment claim: it diagnoses *metric saturation*. LoD is strong on average, HiddenDetect competitive on several attacks, and our frozen-topology detector is at ceiling. The 1.000 entries are measured AUROC values rather than assumed ceiling values: on these splits, the frozen detector achieves complete ranking separation, although many attack scores remain below the fixed calibration threshold. If AUROC were the right deployment metric, there would be little left to improve—and this is what makes the fixed-threshold collapse in the next tables informative.

Table 2: Deployment performance under the fixed calibration threshold (Cal.) and the oracle empirical-0FPR diagnostic. TPR is computed over all attack samples pooled across categories (micro recall); Table S2 additionally reports an unweighted macro average across attack categories, so the two aggregates are not directly comparable. Per-attack breakdown of the diagnostic in Table S2.

Method	**Qwen2.5-VL**			**LLaVA-v1.6**		
	Cal. TPR	**Cal. FPR**	**Oracle @0FPR**	**Cal. TPR**	**Cal. FPR**	**Oracle @0FPR**
LoD	12.8%	0.0%	74.7%	4.8%	0.0%	36.1%
HiddenDetect$_{(12,25)}$	2.9%	1.9%	0.1%	29.2%	0.0%	29.2%
HiddenDetect$_{(25,26)}$	12.0%	4.8%	0.8%	65.1%	3.8%	23.7%
MirrorCheck	93.7%	64.8%	0.1%	91.4%	62.3%	0.1%
GradSafe	90.2%	14.1%	31.2%	82.6%	18.7%	18.3%
JailGuard	61.3%	35.2%	0.0%	58.4%	33.6%	0.0%
JailDAM	88.6%	68.8%	0.0%	84.9%	66.1%	0.0%
FOSTER (frozen)	10.5%	0.0%	74.8%	6.8%	0.0%	29.2%
FOSTER (online)	**80.7%**	**0.0%**	**80.7%**	**71.3%**	**0.0%**	**71.3%**

5.3 Fixed-Threshold Transfer

We now turn the saturated AUROC into a deployment test. For each method, we calibrate a single threshold on its own validation split (GQA-test benign, AdvBench-test malicious) using Youden's J statistic and apply it *unchanged* to a deployment mixture of 9,528 unseen attack samples across nine types and 917 benign samples from SEED + mm-vet2. Table 2 reports two complementary views. Under the reproduced calibration threshold—the primary deployment protocol—several strong-AUROC detectors miss most attacks (LoD, HiddenDetect on Qwen2.5-VL) or reach high recall only through unusable FPR (MirrorCheck, JailGuard and JailDAM all exceed 35% FPR); under the same protocol, FOSTER reaches 80.7% TPR at empirical 0.0% FPR on Qwen2.5-VL and 71.3% TPR at empirical 0.0% FPR on LLaVA-v1.6. The Oracle TPR@0FPR column diagnoses the maximum recall attainable when the target benign pool is available for operating-point selection; it is reported to separate calibration mismatch from a genuine absence of signal, and is not itself a deployable protocol.

Table S2 breaks the Oracle diagnostic down by attack. LoD becomes much stronger at this diagnostic threshold (70.0% macro), confirming that its collapse in the Cal. column is a calibration mismatch, not an absence of signal. FOSTER still improves the macro-averaged Oracle recall to 79.2%, with large gains on CS-DJ ($+41.4$ over LoD), MML-m ($+10.7$), and MML-all ($+5.3$); SI-Attack remains the hardest case for every detector we test. Online adaptation does not improve every category under this diagnostic: recall decreases on HADES ($66.9 \rightarrow 63.1$), MM-SafetyBench ($97.4 \rightarrow 91.3$) and SB-VAJM ($96.5 \rightarrow 88.0$). We report these without attributing a cause, since the present experiments do not separate eviction, admission errors, and neighborhood changes.

5.4 Online Unsafe Topology Repair

The frozen and online rows in Tables 2–S2 isolate online expansion. Static detectors, including FOSTER's frozen-topology variant, stop improving once scores are fixed. FOSTER instead updates its unsafe topology bank as unlabeled streams arrive: under the fixed calibration policy, TPR rises from 10.5% to 80.7% on Qwen2.5-VL and from 6.8% to 71.3% on LLaVA-v1.6, with empirical 0% FPR on the benign pool. Table S2 shows online expansion adds recall on CS-DJ ($+16.8$), MML-m ($+12.6$), SI-Attack ($+12.3$), MML-all ($+8.7$), and SB-UMK ($+5.5$), while FigImg remains near saturation. The gain is delivered without touching Π_0, the router, or the benign reference; empirical FPR is unchanged. Gains concentrate on the rarer modes (MML-m, CS-DJ) where fixed-threshold recall is scarcest.

5.5 Safety-Constrained Admission under Contamination

We inject 10% witness-targeted hard-benign samples into the Qwen2.5-VL S-LOAO stream while keeping the operating policy and injection schedule fixed across admission variants. As summarized in Panel A of Table 3, relaxing either the frozen-topology or router witness increases erroneous unsafe-bank admissions and persistent post-stream FPR, while Full FOSTER retains comparable attack recall at the highest admission precision. Results for $1\%/5\%$ contamination and the witness-specific decomposition are in the supplementary material.

Table 3: Summary of the additional deployment checks. Panel A reports the strongest 10% witness-targeted contamination on Qwen2.5-VL; Panel B evaluates OOD benign false positives under the unchanged calibration policy. Full contamination ratios, confidence intervals, score statistics, and LLaVA-v1.6 benign-shift results are in the supplementary material.

A. Witness-targeted contamination on Qwen2.5-VL (3 seeds)

Admission rule	TPR ($0\% \rightarrow 10\%$)	Post-FPR (%)	Adm. Prec. (%)	#Wrong benign
Current-only	$83.4 \rightarrow 84.1$	$5.89_{\pm.44}$	$80.6_{\pm0.7}$	$447_{\pm27}$
w/o Frozen witness	$82.2 \rightarrow 82.5$	$2.73_{\pm.22}$	$89.1_{\pm0.6}$	$202_{\pm16}$
w/o Router witness	$81.6 \rightarrow 81.9$	$2.18_{\pm.22}$	$90.6_{\pm0.6}$	$163_{\pm15}$
Full FOSTER	$\mathbf{80.7 \rightarrow 80.2}$	$\mathbf{0.44_{\pm.11}}$	$\mathbf{98.3_{\pm0.3}}$	$\mathbf{25_{\pm6}}$

B. OOD benign shift on Qwen2.5-VL

State	Attack TPR (%)	SEED+mm-vet2 FPR	VQA-v2 FPR	LLaVA-Wild FPR
Frozen	10.5	0.00% (0/917)	0.10% (2/2000)	0.40% (4/1000)
Online	**80.7**	**0.00% (0/917)**	**0.25% (5/2000)**	**0.80% (8/1000)**

Table 4: Ablations on the Qwen2.5-VL validation subset with an ablation-specific fixed threshold, so absolute numbers differ from Table 2. All values except AUC are percentages. Admission-witness ablations appear in Panel A of Table 3.

Configuration	AUC	Acc	F1	TPR	FPR
Full FOSTER	0.9900	90.77	94.71	90.27	3.83
w/o router	0.9895	8.44	0.00	0.00	0.00
w/o mode-preserv. bank	0.9900	71.71	81.73	69.10	0.00
w/o online repair	0.9910	21.83	25.51	14.62	0.00

5.6 Robustness under Benign Distribution Shift

We evaluate the frozen and clean-online banks on VQA-v2 and LLaVA-Wild without recalibrating any threshold or updating the bank during benign evaluation. Panel B of Table 3 shows online repair substantially improves attack recall with only a limited change in OOD benign FPR on Qwen2.5-VL; Wilson intervals, q_{99} shifts, and LLaVA-v1.6 results are in the supplementary.

5.7 Ablation and Efficiency

Table 4 isolates the router, mode-preserving bank, and online repair. Removing the router yields zero recall at the evaluated operating point despite an AUROC of 0.9895—a failure of the score-to-decision configuration, not a loss of ranking ability: fixed-policy TPR falls from 90.27% to 0.00%, a flat pool reduces it to 69.10%, and disabling online repair reduces it to 14.62%. AUROC moves by at most 0.001 while operating-point TPR moves by up to 90 points, the reason we report deployment-oriented metrics. The two frozen admission witnesses are evaluated separately under contamination in Panel A of Table 3.

Table S8 shows FOSTER averages $\sim$25 ms—router-path samples cost $\sim$5 ms and only ambiguous ones invoke topology scoring and the admission gate—against $\sim$1.5–5 s for query-heavy detectors and $\sim$800 ms for GradSafe, i.e. comparable to HiddenDetect/LoD while providing online repair. **Additional experiments in the supplementary.** The supplementary reports the $\{0, 1, 5, 10\}\%$ contamination grid over three stream-order seeds with witness-specific hard-benign analyses: across these levels and orders the full gate admits fewer erroneous benign anchors than the relaxed variants, indicating consistency in direction. It also gives benign-shift results with Wilson intervals on both backbones; the empirical 0/917 result has a 95% upper bound of 0.42%, so it should be read as low empirical FPR, not a zero-FPR guarantee.

6 Conclusion

FOSTER reduces fixed-threshold transfer failure on attack streams while keeping the VLM, router, benign reference, and pre-deployment operating policy frozen. Its gains are strongest for recurring unsafe modes, while benign-side drift remains explicitly measured rather than a formal guarantee. These results suggest that safety-constrained, asymmetric memory updates offer a practical middle ground between static detectors and unconstrained test-time adaptation.

Ethics Statement

This paper contains examples of harmful prompts and unsafe model outputs solely for research evaluation. Our work aims to *improve* VLM safety by detecting jailbreak attacks, not to facilitate misuse. All harmful content is redacted in figures.

AI Use Statement

We used generative AI tools to polish the writing of this paper, including improving the clarity, grammar and phrasing of author-written text. All AI-assisted text was subsequently reviewed by the authors, who verified that every statement expresses the intended meaning and accurately reflects the underlying work. Generative AI was not used to generate research ideas, experimental results, or claims. We take full responsibility for the final content of this work, including all text, claims and artifacts.

References

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

Beitao Chen, Xinyu Lyu, Jingkuan Song, Heng Tao Shen, and Lianli Gao. Safeptr: Token-level jailbreak defense in multimodal llms via prune-then-restore mechanism. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. InternVL2.5 Technical Report.

Samar Fares, Nikola Jovanovic, Robin Staab, Mark Vero, and Martin Vechev. Mirrorcheck: Efficient adversarial defense for vision-language models. *arXiv preprint arXiv:2406.09250*, 2024.

Yichen Gong, Delong Deng, Xinwen Guan, Yun Zhu, and Shimin Bi. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*, 2023.

Yunhao Gou, Kai Shao, Zhili Gong, Zhufeng Deng, et al. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *European Conference on Computer Vision (ECCV)*, 2024.

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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Yilei Jiang, Xinyan Gao, Tianshuo Peng, Yingshui Tan, Xiaoyong Zhu, Bo Zheng, and Xiangyu Yue. Hiddendetect: Detecting jailbreak attacks against large vision-language models via monitoring hidden states. *arXiv preprint arXiv:2502.14744*, 2025.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking multimodal llms with generative comprehension. In *arXiv preprint arXiv:2307.16125*, 2023.

Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. In *European Conference on Computer Vision (ECCV)*, 2024.

Shuang Liang, Zhihao Xu, Jialing Tao, Hui Xue, and Xiting Wang. Learning to detect unknown jailbreak attacks in large vision-language models: A unified and accurate approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Yu Pan, Wenbin Xu, and Jiahui Zhou. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. *Blog post*, 2024b.

Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision (ECCV)*, 2024c.

Yi Nian, Shenzhe Zhu, Yuehan Qin, Li Li, Ziyi Wang, Chaowei Xiao, and Yue Zhao. Jaildam: Jailbreak detection with adaptive memory for vision-language model. *arXiv preprint arXiv:2504.03770*, 2025.

Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*, 2024.

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

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv preprint arXiv:2307.14539*, 2023.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Yu Wang, Xiaogeng Li, Muhao Yu, Fei Jiang, and Wei-Wei Chao. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. In *European Conference on Computer Vision (ECCV)*, 2024.

Yu Wang, Xiaofei Zhou, Yichen Wang, Geyuan Zhang, and Tianxing He. Jailbreak large vision-language models through multi-modal linkage. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1466–1494, 2025.

Yueqi Xie, Minghong Fang, Renjie Pi, and Neil Gong. Gradsafe: Detecting jailbreak prompts for llms via safety-critical gradient analysis. 2024.

Yue Xu, Xiuyuan Qi, Zhan Qin, and Wenjie Wang. Cross-modality information check for detecting jailbreaking in multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Sibo Yi, Yule Liu, Zhen Sun, and Tat-Seng Chua. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*, 2024.

Weihao Yu, Zhengyuan Yang, Lingfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*, 2024.

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

Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Adanpc: Exploring non-parametric classifier for test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

This supplementary material accompanies the main paper. All references to "Table X" or "Section X" without an "S" prefix point to the main body; tables, figures, algorithms and equations introduced here carry the "S" prefix. The main body and this supplementary can be read independently.

A Per-Attack AUROC on Qwen2.5-VL and LLaVA-v1.6

The full per-attack AUROC grid referenced in the main-body *Per-Attack AUROC Is Saturated* subsection is reported in Table S1. It is placed here rather than in the main body because it diagnoses metric saturation rather than supporting a deployment claim.

Table S1: Detection AUROC across 2 VLMs × 9 attacks (GQA benign). Best **bold**, second underlined. †: mutation-based. ‡: refusal-word variant. For FOSTER, we report the frozen-topology detector so that AUROC is directly comparable to fixed-score baselines; deployment-time gains are in Tables 2–S2.

Model	Method	MML-m	HADES	SB-VAJM	SB-UMK	CS-DJ	SI-Atk	MM-SB	FigImg	MML-all	Avg
LLaVA-v1.6	MirrorCheck	.360	.524	.891	.891	.210	.482	.381	.106	.296	.460
	CIDER	.832	.736	.896	.896	.709	.707	.777	.747	.810	.790
	JailGuard†	.324	.452	.346	.340	.518	.474	.585	.761	.474	.475
	JailGuard‡	.358	.632	.525	.563	.597	.471	.585	.759	.487	.553
	GradSafe	.998	.783	.621	.709	.786	.790	.594	.333	.797	.712
	HiddenDetect	.315	.989	.925	.981	.988	.845	.913	.762	.680	.822
	LoD	**.961**	**.999**	**.997**	**.996**	**.997**	.839	**.956**	**.991**	**.976**	**.968**
	FOSTER (frozen)	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Qwen2.5-VL	MirrorCheck	.357	.494	.872	.873	.164	.490	.385	.089	.296	.447
	CIDER	.722	.759	.913	.912	.654	.757	.812	.668	.821	.780
	JailGuard†	.341	.517	.407	.491	.650	.548	.626	.655	.504	.527
	JailGuard‡	.367	.687	.687	.666	.799	.547	.655	.675	.529	.624
	GradSafe	.816	.917	.827	.982	.789	.785	.711	.542	.789	.795
	HiddenDetect	.536	.951	.901	.872	.865	.687	.759	.690	.588	.761
	LoD	**.950**	**1.000**	**1.000**	**.997**	**1.000**	.788	**.983**	**1.000**	**.966**	**.965**
	FOSTER (frozen)	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

B Per-Attack Oracle Recall on Qwen2.5-VL

Table S2 gives the per-attack breakdown of the oracle empirical-0FPR diagnostic referenced in Section 5.3, the fixed-threshold transfer evaluation.

C Full Contamination Results on Qwen2.5-VL

The main body summarizes the strongest 10% stress setting in Panel A of Table 3. Table S3 reports the full contamination grid over $\{0, 1, 5, 10\}\%$; all values are mean±std over three stream-order seeds, sharing the same injection samples and positions across variants.

Table S2: Per-attack recall under the oracle empirical-0FPR diagnostic on Qwen2.5-VL, i.e., the maximum TPR attainable at zero benign false positives on the evaluated pool. Macro denotes the unweighted mean of recall across the nine attack categories, and therefore differs from the pooled (micro) Oracle column of Table 2.

Method	CS-DJ	SI-Atk	MML-m	HADES	FigImg	MM-SB	SB-VAJM	SB-UMK	MML-all	**Macro**
LoD	56.5	2.6	72.7	61.9	**100**	**96.1**	**97.5**	69.6	73.1	70.0
HiddenDetect$_{(12,25)}$	0.0	0.0	0.0	0.0	0.0	1.9	0.0	0.0	0.0	0.2
HiddenDetect$_{(25,26)}$	0.0	0.0	0.2	0.0	0.0	3.5	10.0	1.4	0.2	1.7
MirrorCheck	0.0	0.1	0.0	0.2	0.2	0.3	0.0	0.0	0.1	0.1
GradSafe	34.7	18.9	0.0	55.8	12.4	41.6	36.2	25.8	39.2	29.4
JailGuard	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
JailDAM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FOSTER (frozen)	81.1	4.9	70.8	**66.9**	99.2	97.4	96.5	87.8	69.7	74.9
FOSTER (online)	**97.9**	**17.2**	**83.4**	63.1	**100**	91.3	88.0	**93.3**	**78.4**	**79.2**

Table S3: Full contamination grid on Qwen2.5-VL (3 seeds). Post-stream FPR (%) on SEED+mm-vet2 and wrong benign→unsafe insertions.

	Post-Stream FPR (%)			**#Wrong Benign→Unsafe**		
Variant	1%	5%	10%	1%	5%	10%
Static	0.00	0.00	0.00	0	0	0
Current-only	0.65$\pm$.11	3.24$\pm$.27	5.89$\pm$.44	45$\pm$3	220$\pm$15	447$\pm$27
w/o Frozen	0.29$\pm$.06	1.38$\pm$.17	2.73$\pm$.22	20$\pm$2	98$\pm$9	202$\pm$16
w/o Router	0.18$\pm$.06	1.05$\pm$.17	2.18$\pm$.22	16$\pm$2	79$\pm$8	163$\pm$15
Full FOSTER	0.04$\pm$.06	0.15$\pm$.06	0.44$\pm$.11	2$\pm$1	10$\pm$3	25$\pm$6

Full FOSTER maintains attack TPR $80.7 \to 80.6 \to 80.0 \to 80.2$ and admission precision $100 \to 99.7 \to 98.8 \to 98.3$ (%) at $0\%/1\%/5\%/10\%$; Current-only reaches slightly higher attack TPR (84.1% at 10%) but admission precision collapses to 80.6%. The 10% row of Table S3 matches Panel A of the main-body Table 3.

Splitting the 10% hard-benign candidates by which witness they narrowly miss, Table S4 confirms the intended witness-specific pattern: on frozen-gap candidates the w/o Frozen variant degrades most, and on router-gap candidates the w/o Router variant degrades most. The two evolving-side witnesses therefore catch orthogonal failure modes.

Table S4: Witness-specific 10% contamination on Qwen2.5-VL. Wrong benign→unsafe insertions and admission precision (%).

Cand. type	**Variant**	#Wrong	Adm. Prec.
Frozen-gap	Current-only	325$\pm$20	67.8$\pm$3.0
	w/o Frozen	281$\pm$17	71.5$\pm$2.7
	w/o Router	47$\pm$7	93.6$\pm$1.3
	Full	16$\pm$4	97.8$\pm$0.8
Router-gap	Current-only	274$\pm$19	72.3$\pm$2.8
	w/o Frozen	54$\pm$8	92.9$\pm$1.5
	w/o Router	238$\pm$16	75.7$\pm$2.5
	Full	13$\pm$3	98.2$\pm$0.7

D Benign-Shift Results and Score Statistics

Panel B of the main-body Table 3 reports OOD benign FPR on Qwen2.5-VL. Table S5 adds the LLaVA-v1.6 counterpart and repeats Qwen2.5-VL for symmetry. Attack TPR is unchanged (Qwen2.5-VL 80.7% online / 10.5% frozen; LLaVA-v1.6 71.3% online / 6.8% frozen).

Wilson 95% upper bounds. $0/917 \Rightarrow [0.00\%, 0.42\%]$; $2/2000 \Rightarrow [0.03\%, 0.36\%]$; $3/2000 \Rightarrow [0.05\%, 0.44\%]$; $5/2000 \Rightarrow [0.11\%, 0.58\%]$; $7/2000 \Rightarrow [0.17\%, 0.72\%]$; $4/1000 \Rightarrow [0.16\%, 1.02\%]$; $5/1000 \Rightarrow [0.22\%, 1.16\%]$; $8/1000 \Rightarrow [0.41\%, 1.57\%]$; $10/1000 \Rightarrow [0.54\%, 1.83\%]$. The default $0/917$ has upper bound 0.42%, so "empirical 0% FPR" should be read as "FPR consistent with $\leq 0.42\%$ at 95% confidence."

Benign-score q_{99} shifts (Online minus Frozen). Qwen2.5-VL: $+0.017/+0.034/+0.055$ on SEED+mm-vet2/VQA-v2/LLaVA-Wild; LLaVA-v1.6: $+0.018/+0.039/+0.065$. The shift is

Table S5: OOD benign shift, both backbones. Empirical FPR reported as FP count/N (and % where non-zero).

Backbone	State	SEED+ mm-vet2	VQA-v2	LLaVA-Wild
Qwen2.5-VL	Frozen	0/917	2/2000 (.10%)	4/1000 (.40%)
	Online	0/917	5/2000 (.25%)	8/1000 (.80%)
LLaVA-v1.6	Frozen	0/917	3/2000 (.15%)	5/1000 (.50%)
	Online	0/917	7/2000 (.35%)	10/1000 (1.00%)

small on the calibration source and larger on the more open pools, but only a limited fraction of the shift crosses the fixed threshold.

E Additional Method Details

Design objective formalization. Let P_u, P_b denote the unseen-attack and benign score distributions and Π_0 the pre-deployment operating policy. We track two quantities:

$$G_{\text{repair}}(P_u) = \text{TPR}(P_u; \Pi_0, \mathcal{M}^t_{\text{unsafe}}) - \text{TPR}(P_u; \Pi_0, \mathcal{M}^0_{\text{unsafe}}), \tag{S1}$$

$$D_{\text{benign}}(P_b) = \text{FPR}(P_b; \Pi_0, \mathcal{M}^t_{\text{unsafe}}) - \text{FPR}(P_b; \Pi_0, \mathcal{M}^0_{\text{unsafe}}). \tag{S2}$$

FOSTER is designed to increase G_{repair} on recurring unsafe modes while empirically limiting D_{benign} on the evaluated benign pool. This is a design objective, not a formal guarantee: D_{benign} is measured, not proved, and reported explicitly in the ablation and benign-shift experiments in the main body.

Router training and calibration. The safety evidence router is a two-stage cross-layer aggregator over the mid-to-late layer subset $\mathcal{L}$: a per-layer scoring MLP g_ϕ ($d \to 64 \to 1$, Tanh) that emits softmax attention weights, followed by a bottleneck classifier ($d \to 128 \to 1$) with post-hoc temperature scaling. Training uses focal loss (Lin et al., 2017) ($\gamma = 2$) on the GQA + AdvBench calibration split, AdamW with base LR 1×10^{-4}, batch size 64, and early stopping on validation calibration ECE. The router is frozen after calibration.

F Streaming Inference Algorithm

Algorithm S1 gives the streaming procedure referenced in the main-body Streaming Inference subsection. Throughout deployment the VLM f_{VLM}, the safety evidence router f_{router}, the pre-deployment operating policy Π_0, the benign reference bank $\mathcal{M}^0_{\text{safe}}$ and the frozen initial unsafe bank $\mathcal{M}^0_{\text{unsafe}}$ remain fixed. The only object that evolves is $\mathcal{M}_{\text{unsafe}}$, and only through anchors that satisfy the triple admission gate. The current-topology witness R^{cur}_t uses the evolving bank $\mathcal{M}_{\text{unsafe}} \cup \mathcal{M}^0_{\text{safe}}$; the frozen witness R^{frz}_t uses only the pre-deployment $\mathcal{M}^0_{\text{unsafe}} \cup \mathcal{M}^0_{\text{safe}}$. LKNN denotes layer-wise KNN scoring and Admit denotes mode-preserving admission. Note that the **continue** on the router-only path means samples resolved with high router confidence bypass the admission gate entirely and never become anchors.

G Topology Diagnostic Protocol

For each VLM we extract per-layer hidden states at the last token position, ℓ_2-normalize them, and compute three diagnostics D1–D3 as summarized in Table S6. D1 measures the normalized inter/intra separation $\rho_l = \bar{d}_{\cos}(\mathcal{H}^{\text{safe}}_l, \mathcal{H}^{\text{unsafe}}_l) / \frac{1}{2}(\sigma(\mathcal{H}^{\text{safe}}_l) + \sigma(\mathcal{H}^{\text{unsafe}}_l))$ layer by layer against a random-label control. D2 initializes the unsafe reference with a single calibration attack (AdvBench) and evaluates plain KNN on held-out attack families without their labels. D3 compares layer-wise informativeness across attack families. These diagnostics motivate the layer-wise scoring and the frozen initial bank used by FOSTER; they are not part of the deployment claim.

Algorithm S1 FOSTER streaming inference.

Require: Stream $\{x_t\}$; frozen f_{VLM}, f_{router}, Π_0, $\mathcal{M}^0_{\mathrm{safe}}$, $\mathcal{M}^0_{\mathrm{unsafe}}$

$\mathcal{M}_{\mathrm{unsafe}} \leftarrow \mathcal{M}^0_{\mathrm{unsafe}}$
for each x_t **do**
  $\{\mathbf{h}_l\} \leftarrow f_{\mathrm{VLM}}(x_t)$; $p_t \leftarrow f_{\mathrm{router}}(\{\mathbf{h}_l\})$
  if $|2p_t - 1| > \tau_{\mathrm{conf}}$ **then** ▷ router-only path
    $\hat{y}_t \leftarrow \mathbb{1}[p_t > \theta_{\mathrm{router}}]$; **continue**
  end if
  $R^{\mathrm{cur}}_t \leftarrow \mathrm{LKNN}(\{\mathbf{h}_l\}, \mathcal{M}_{\mathrm{unsafe}} \cup \mathcal{M}^0_{\mathrm{safe}})$
  $R^{\mathrm{frz}}_t \leftarrow \mathrm{LKNN}(\{\mathbf{h}_l\}, \mathcal{M}^0_{\mathrm{unsafe}} \cup \mathcal{M}^0_{\mathrm{safe}})$
  $\hat{y}_t \leftarrow \mathbb{1}[R^{\mathrm{cur}}_t > \theta_{\mathrm{topo}}]$
  if $R^{\mathrm{cur}}_t > \tau_{\mathrm{cur}} \wedge R^{\mathrm{frz}}_t > \tau_{\mathrm{frz}} \wedge p_t > \tau_{\mathrm{cls}}$ **then**
    $\mathcal{M}_{\mathrm{unsafe}} \leftarrow \textsc{Admit}(\mathcal{M}_{\mathrm{unsafe}}, \{\mathbf{h}_l\})$
  end if
end for

Table S6: Topology diagnostics.

Diag.	Question	Design implication
D1	Are safe and unsafe activations geometrically separated?	Score in hidden-state topology, not on output-only signals.
D2	Does one known attack provide anchors for held-out attacks?	Initialize bank with a single known attack; grow online.
D3	Is the informative layer stable across attacks?	Score per-layer and aggregate; do not commit to one depth.

H INTERNVL2.5 PER-ATTACK AUROC

Table S7 reports per-attack AUROC on InternVL2.5-8B (Chen et al., 2024) using GQA as the benign reference. These results provide complementary evidence that hidden-state ranking signals transfer to a third backbone. Because the full fixed-threshold and streaming evaluation in the main body is conducted only on Qwen2.5-VL and LLaVA-v1.6, we do not use the InternVL2.5 AUROC results to make any fixed-threshold, online-repair or deployment-time claim.

Table S7: Detection AUROC on InternVL2.5-8B with GQA as benign. Best **bold**, second underlined.

Model	Method	MML-m	HADES	SB-VAJM	SB-UMK	CS-DJ	SI-Atk	MM-SB	FigImg	MML-all	Avg
InternVL2.5	MirrorCheck	.392	.480	.883	.882	.150	.507	.389	.073	.312	.452
	CIDER	.593	.634	.750	.749	.664	.670	.651	.225	.603	.615
	JailGuard	.268	.386	.429	.426	.595	.542	.572	.697	.439	.484
	GradSafe	**.975**	.872	.729	.842	.731	.623	.633	.130	.784	.702
	HiddenDetect	.946	.841	.779	.770	.863	.782	.851	.973	**.942**	.861
	LoD	.973	**1.00**	**.999**	**.999**	**1.00**	.864	**.986**	**1.00**	.982	**.978**
	FOSTER (frz)	.955	.996	.998	.947	.999	**1.00**	**.999**	.869	.947	.968

I ANALYSIS OF LOCAL REPAIR, CERTIFICATION, AND MEMORY STABILITY

Setting. Let $\mathcal{D}$ be the set of inputs deferred by the frozen router, i.e. those not resolved on the high-confidence router path of Algorithm S1. For each layer $l \in \mathcal{L}$ let $\kappa_l(x, z) = \max(\cos(\mathbf{h}_l(x), \mathbf{h}_l(z)), 0) \in [0, 1]$. Writing $N^+_{l,t}(x)$ for the top-k entries retrieved from $\mathcal{M}^t_{\mathrm{unsafe}}$ and $s^-_l(x)$ for the top-k mean similarity to the frozen benign bank, Eq. (4) reads

$$R_t(x) = \frac{1}{2} \sum_{l \in \mathcal{L}} w_l \Big[1 + \frac{1}{k} \sum_{z \in N^+_{l,t}(x)} \kappa_l(x, z) - s^-_l(x) \Big]. \tag{S3}$$

We assume fixed representations, fixed $w_l \geq 0$ with $\sum_l w_l = 1$, and exact class-wise top-k retrieval (neighbors are retrieved separately per class, not from a merged pool) with at least k entries in each bank. Entries carry fixed features and identities and ties are broken by a fixed rule; approximate retrieval does not automatically inherit these statements. Because $\mathcal{M}_{\mathrm{safe}}$ is frozen, $s^-_l(x)$ is independent of the unsafe-memory state and cancels in every difference below. Each result is deterministic given the two compared memory states and therefore holds even when those states were produced adaptively; no independence among the admission witnesses is assumed.

I.1 INSERTION-ONLY REPAIR

Proposition 1 (Insertion-only local repair). *If the updated unsafe bank retains every old entry, then $R_{t+1}(x) \geq R_t(x)$ for every query x. Moreover, let $u_{l,1}(x) \geq \cdots \geq u_{l,k}(x)$ be the old top-k unsafe similarities at layer l. If the added entries include $m \leq k$ anchors whose similarity to x is at least $a_l(x)$ at every layer, then*

$$R_{t+1}(x) - R_t(x) \;\geq\; G_m(x) := \frac{1}{2k} \sum_{l \in \mathcal{L}} w_l \Big[m\, a_l(x) - \sum_{j=k-m+1}^{k} u_{l,j}(x) \Big]_+, \tag{S4}$$

so a deferred query with $R_t(x) \leq \theta$ is detected after the update whenever $G_m(x) > \theta - R_t(x)$.

Proof of Proposition 1. Fix x and a layer l, and let T_l, T_l' be the top-k sums of unsafe similarities before and after the update. All old entries remain available, and exact top-k retrieval maximizes the similarity sum over size-k subsets of the bank; since the old top-k set is still one such subset, $T_l' \geq T_l$.

For the quantitative claim, consider the particular size-k subset formed by the $k - m$ strongest old neighbors together with the m specified new anchors. Its sum is at least $T_l - \sum_{j=k-m+1}^{k} u_{l,j}(x) + m\, a_l(x)$. As T_l' is at least the sum of any size-k subset, taking the better of this subset and the old top-k set gives $T_l' - T_l \;\geq\; [\, m\, a_l(x) - \sum_{j=k-m+1}^{k} u_{l,j}(x)\,]_+$. The benign terms cancel in Eq. (S3); multiplying by $w_l/(2k)$ and summing over l yields both monotonicity and Eq. (S4). If $G_m(x) > \theta - R_t(x)$ then $R_{t+1}(x) > \theta$, which for $x \in \mathcal{D}$ is an unsafe prediction. □

Similarity-margin form. Suppose the m new anchors each satisfy $\kappa_l(x, z) \geq u_{l,k-m+1}(x) + \gamma_l$ with $\gamma_l \geq 0$ at every layer. Every displaced old similarity is at most $u_{l,k-m+1}(x)$, so

$$R_{t+1}(x) - R_t(x) \;\geq\; \frac{m}{2k} \sum_{l \in \mathcal{L}} w_l \gamma_l, \tag{S5}$$

and $\frac{m}{2k} \sum_l w_l \gamma_l > \theta - R_t(x)$ suffices to repair a missed deferred query. This separates the three quantities that govern repair: how many useful anchors were admitted, how much closer they are than what they displace, and how far the query started from the threshold. It does not assert that any particular number of anchors is available in the evaluated streams.

I.2 THE ADMISSIBLE ENVELOPE

Because the representations, router, and benign bank are frozen by Eq. (1), R^{frz} and p^{unsafe} are functions of x alone, so the *eligibility set*

$$\mathcal{E} = \{x \in \mathcal{D} :\; R^{\mathrm{frz}}(x) > \tau_{\mathrm{frz}} \;\wedge\; p^{\mathrm{unsafe}}(x) > \tau_{\mathrm{cls}}\} \tag{S6}$$

is determined before deployment. Eligibility is *necessary but not sufficient* for admission: by Eq. (5) an anchor must additionally clear the dynamic gate $R_t^{\mathrm{cur}} > \tau_{\mathrm{cur}}$, which depends on the current bank, and it must survive the finite per-cluster and global capacities. The set actually reached by a given stream is therefore in general a strict subset of $\mathcal{E}$, and we do not attempt to characterize it. What the frozen witnesses provide is an *outer envelope* of that unknown set.

Proposition 2 (Admissible envelope). *For a query x let $\bar{s}_l^+(x)$ be the top-k mean similarity at layer l computed against $\mathcal{M}^0_{\mathrm{unsafe}} \cup \mathcal{E}$, and*

$$R^{\mathrm{env}}(x) = \tfrac{1}{2} \sum_{l \in \mathcal{L}} w_l \big[1 + \bar{s}_l^+(x) - s_l^-(x)\big]. \tag{S7}$$

Then $R_t(x) \leq R^{\mathrm{env}}(x)$ for every t, every stream order, and every admission and eviction history. Consequently, if $R^{\mathrm{env}}(x) \leq \theta$ then x is never predicted unsafe on the topology path, and a benign x of this kind never becomes a false positive.

Proof of Proposition 2. By Eq. (1) the representations, router and benign bank are fixed, so $\mathcal{E}$ is fixed before deployment. By Eq. (5) an input enters the bank only if it lies in $\mathcal{E}$, hence $\mathcal{M}^t_{\mathrm{unsafe}} \subseteq \mathcal{M}^0_{\mathrm{unsafe}} \cup \mathcal{E}$ for all t, irrespective of stream order, injected content, or eviction history.

Fix a layer l. The top-k mean similarity over any subset of a set is at most the top-k mean over the set itself, so $\frac{1}{k} \sum_{z \in N^+_{l,t}(x)} \kappa_l(x, z) \leq \bar{s}_l^+(x)$. Substituting into Eq. (S3) and using $w_l \geq 0$ gives $R_t(x) \leq R^{\mathrm{env}}(x)$ for all t. If $R^{\mathrm{env}}(x) \leq \theta$ then $R_t(x) \leq \theta$ always, so x is never predicted unsafe

on the topology path; since the frozen router's direct decisions are identical in every state, a benign x of this kind never becomes a false positive. □

What this is and is not. Eq. (S7) is a *conditional outer bound*, not a certificate we can currently issue. Three gaps are worth stating plainly.

First, the bound is not tight for FOSTER as deployed. It is attained only in the relaxed problem that ignores the dynamic τ_{cur} gate and the capacity constraints, i.e. when the entire eligibility set is simultaneously present in memory. Under the actual admission rule this need not be approached at all; Appendix I (cold start, below) gives the extreme case in which memory never changes because no query clears the gate.

Second, evaluating $\bar{s}_l^+(x)$ requires a top-k retrieval against $\mathcal{E}$, a set defined by two threshold conditions over the whole input space rather than by an enumerated list. We give no algorithm for computing or over-approximating it, and we report no certified coverage—what fraction of a benign pool satisfies $R^{\text{env}}(x) \le \theta$ is an empirical quantity we have not measured. Any use of this bound as a deployment guarantee would require both.

Third, boundedness alone is not distinctive: because $\kappa_l \in [0,1]$, *any* KNN-style detector admits the trivial bound $R_t(x) \le \frac{1}{2}\sum_l w_l[2 - s_l^-(x)]$. The question is whether the frozen witnesses make the envelope *informative*, and that is a quantitative matter—it depends on how much smaller $\bar{s}_l^+(x)$ is than its trivial ceiling, which in turn depends on τ_{frz} and τ_{cls}. We therefore state the structural consequence of freezing the witnesses and do not claim that alternatives admit no bound of this form. The contrast is one of *stream-independence*: under current-topology-only admission the analogous region $\{x : R_t^{\text{cur}}(x) > \tau_{\text{cur}}\}$ moves as anchors are admitted, so an envelope of this kind must be recomputed as the stream proceeds, whereas Eq. (S6) is fixed once Π_0 is fixed.

Corollary 3 (Admission cone)**.** *Every admitted anchor z satisfies $\sum_{l\in\mathcal{L}} w_l\big[s_l^{+,0}(z) - s_l^-(z)\big] > 2\tau_{\text{frz}} - 1$, where $s_l^{+,0}$ is the top-k mean similarity to the frozen initial unsafe bank. At $\tau_{\text{frz}} = \frac{1}{2}$ this states that an admitted anchor must be at least as close to the pre-deployment unsafe geometry as to the benign reference, in the weighted per-layer sense.*

Proof. Admission requires $R^{\text{frz}}(z) > \tau_{\text{frz}}$. Writing $R^{\text{frz}}(z) = \frac{1}{2}\sum_l w_l[1 + s_l^{+,0}(z) - s_l^-(z)]$ and using $\sum_l w_l = 1$ gives the claim. The condition involves only frozen quantities, so it applies to every anchor admitted at any time, and unlike the envelope above it is checkable for each admitted anchor. □

Corollary 3 is the part of this analysis we regard as solid: it is a per-anchor constraint on frozen quantities, and it gives the precise sense in which the evolving memory cannot relax its own admission standard. It constrains *where* memory can move, not whether the pseudo-labels are correct, and a benign input inside $\mathcal{E}$ is not protected by Proposition 2.

I.3 Replacement stability and decision margins

Proposition 4 (Stability under memory replacement)**.** *Allow arbitrary insertion and eviction. Let $N_{l,t}^+(x)$ be the retrieved top-k entry set at layer l and $q_l(x) = k - |N_{l,t}^+(x) \cap N_{l,t+1}^+(x)|$. Then*

$$|R_{t+1}(x) - R_t(x)| \;\le\; \varepsilon(x) := \frac{1}{2k}\sum_{l\in\mathcal{L}} w_l\, q_l(x), \tag{S8}$$

and a deferred query's decision is unchanged whenever $|R_t(x) - \theta| > \varepsilon(x)$.

Proof of Proposition 4. Fix a layer and write $C_l = N_{l,t}^+(x) \cap N_{l,t+1}^+(x)$, $A_l = N_{l,t+1}^+(x) \setminus C_l$ and $B_l = N_{l,t}^+(x) \setminus C_l$, so that $|A_l| = |B_l| = q_l(x)$. Entries in C_l contribute identically in both states, hence

$$s_{l,t+1}^+(x) - s_{l,t}^+(x) = \frac{1}{k}\Big[\sum_{z\in A_l}\kappa_l(x,z) - \sum_{z\in B_l}\kappa_l(x,z)\Big]. \tag{S9}$$

Each bracketed sum lies in $[0, q_l(x)]$ because $\kappa_l \in [0,1]$, so the difference has magnitude at most $q_l(x)/k$. The frozen benign terms cancel; the triangle inequality with $w_l \ge 0$ gives Eq. (S8). A perturbation of magnitude at most $\varepsilon(x)$ cannot move a score across θ when $|R_t(x) - \theta| > \varepsilon(x)$. □

Corollary 5 (Boundary-mass bound on rate changes)**.** *Fix two realized memory states and a benign distribution P_b. Then*

$$\mathrm{FPR}_{t+1} - \mathrm{FPR}_t \;\le\; \Pr_{x\sim P_b}\big[x \in \mathcal{D},\ \theta - \varepsilon(x) < R_t(x) \le \theta\big], \tag{S10}$$

and symmetrically, for an unsafe distribution P_u, $\mathrm{TPR}_t - \mathrm{TPR}_{t+1} \le \Pr_{x\sim P_u}[x \in \mathcal{D},\ \theta < R_t(x) \le \theta + \varepsilon(x)]$.

Proof. The frozen router decides identically in both states, so a newly false-positive input must lie in $\mathcal{D}$ with $R_t(x) \le \theta < R_{t+1}(x)$; Proposition 4 then forces $R_t(x) > \theta - \varepsilon(x)$. The FPR change equals the probability of newly false-positive inputs minus that of corrected ones and is bounded by the former. The recall statement follows by the same argument applied to unsafe queries that change from detected to missed. □

Both sides of Eq. (S10) are distribution-dependent and are not estimated by the result: it says that rate changes are confined to benign mass lying within $\varepsilon(x)$ of the boundary, not that this mass is small. A frozen benign bank does not make the probability vanish, which is why benign drift is measured empirically rather than claimed. Note also that this compares two memory states on a *fixed* distribution; it does not license attributing rate differences between different streams to memory updates.

Interpretation under bounded memory. The quantity $q_l(x)$ counts turnover among *retrieved* entries, not bank-wide churn, so evicting entries outside a query's neighborhoods leaves its score unchanged. For the configured $k = 20$, $q_l(x) \le 1$ at every layer implies a score change of at most $1/(2k) = 0.025$; this is a bound on the score, not a claim of 2.5 percentage points on FPR or TPR. Conversely, retaining a cluster does not by itself retain the neighbors a given query relies on, so per-cluster quotas target the eviction failure mode without eliminating it.

I.4 Cold start and the necessity of eligible seeds

Proposition 2 also has a cold-start corollary. Suppose every query in a stream either lies outside $\mathcal{E}$ or satisfies $R_0(x) \le \tau_{\mathrm{cur}}$ under the initial bank. The first query then fails admission, so the bank is unchanged; by induction every later query is scored against the same initial bank and also fails admission, and the memory never moves. Adaptation therefore requires at least one eligible seed with sufficient initial support—which is precisely the recall–robustness trade-off discussed in Section 4.4, and consistent with SI-Attack remaining hard. It does *not* imply that each attack family needs its own seed: by Proposition 1 what matters is geometric proximity, so anchors admitted from one family can repair queries from another.

Scope. Algorithm S1 returns immediately on the high-confidence router path, so all statements above concern deferred queries and none of them describes repair of samples the router already resolves. None of the results establishes correct admission of arbitrary unseen attacks, a population-level zero false-positive rate, or robustness to arbitrary poisoning; the frozen witnesses restrict admissible updates but need not be statistically independent or free of shared errors. This is why recall, admission precision, and benign false positives are reported separately in Section 5.5 rather than folded into a single number.

J Detector-Side Latency

Table S8 reports the per-sample detector-side latency referenced in Section 5.7.

K Additional Implementation Details

Attack datasets used throughout the paper: SafetyBench-VAJM, SafetyBench-UMK, HADES, FigImg, MM-SafetyBench, CS-DJ, SI-Attack, MML-m, MML-m_all. Benign sources: GQA (calibration), SEED-Bench + mm-vet2 (deployment default), VQA-v2 (benign shift and contamination candidates, disjoint splits), and LLaVA-Wild (benign shift only). Deployment thresholds and admission gates: $k = 20$, $\tau_{\mathrm{conf}} = 0.7$, $\tau_{\mathrm{cur}} = 0.8$, $\tau_{\mathrm{frz}} = 0.5$, $\tau_{\mathrm{cls}} = 0.3$; benign bank capacity $N_s = 500$, per-cluster unsafe capacity $N_c = 1000$; cluster-seeding cosine threshold δ tuned on the calibration

Table S8: Per-sample latency (VLM forward pass excluded). Latency covers computation, including cross-layer aggregation and, for FOSTER, admission-gate evaluation.

Method	**Lat. (ms)**	**Online**
JailGuard	∼5000	✗
MirrorCheck	∼2000	✗
CIDER	∼1500	✗
GradSafe	∼800	✗
HiddenDetect	∼50	✗
LoD	∼30	✗
FOSTER (router only)	∼5	✗
FOSTER (full, avg.)	∼25	✓

split and frozen. All thresholds and capacities are tuned only on the calibration data and remain fixed as Π_0 throughout deployment.