

Latent Visual Sensitive Fingerprinting for Black-Box Vision–Language Model Tamper Detection

Chaoxiang He^{1†} Jichen Chai^{1†} Hanqing Hu^{1†} Yi Wang²
Shuo Wang¹ Hongsheng Hu^{1*} Bin Benjamin Zhu^{2,3*}

¹Shanghai Jiao Tong University

²Microsoft Corporation ³Beijing Inst. of Math. Sci. & Appl.

{hechaoxiang, cjc666, hongsheng.hu, wangshuosj}@sjtu.edu.cn
huhangqinghust@gmail.com, yiwang4@microsoft.com, binzhu@ieee.org

Abstract

Cloud-hosted vision-language models (VLMs) may be silently fine-tuned, poisoned, compressed, or replaced after deployment, motivating black-box fingerprint verification. RESF (Hu et al., 2025) tests response distributions but requires estimating the service temperature, making its reliability sensitive to estimation error. We propose *Latent Visual Sensitive Fingerprinting* (LVSF), based on a local Pearson-sensitivity decomposition of the conditional top- k distribution of the first generated token. This decomposition separates observable output shifts from input-dependent exposure to parameter changes. Modeling tampering-induced relative top- k logit shifts as isotropic, we show that a uniform top- k anchor maximizes average sensitivity. Accordingly, LVSF equalizes recorded top- k logits, maintains a margin from out-of-support tokens, and maximizes full-parameter Fisher sensitivity, yielding a reference distribution invariant to positive temperature scaling. These principles apply to both LLMs and VLMs. For VLMs, LVSF fixes a natural-language query and optimizes an image in a pretrained VAE latent space to produce natural-looking fingerprints. Repeated API queries detect excess out-of-support first tokens or shifts in conditional in-support frequencies. Across four VLMs and multiple tampering types, LVSF achieves consistently high detection rates with no observed false positives and substantially higher detection rates than evaluated baselines on Qwen2.5-VL-7B-Instruct.

1 Introduction

Vision-language models (VLMs) are increasingly deployed as large-scale AI services (Ghosh et al., 2024). A model owner may rely on a cloud service provider to host a VLM, while downstream users interact with the deployed service only through a

black-box API. After deployment, however, the service model can be modified: it may be replaced, compressed, or quantized to reduce serving cost (Dettmers et al., 2022; Wang et al., 2024); fine-tuned through lightweight adaptations such as LoRA (Hu et al., 2022); poisoned (Xu et al., 2024b) or backdoored (Xu et al., 2024a; Lyu et al., 2024); or safety-aligned in ways that alter its behavior while leaving ordinary user queries seemingly unaffected (Zong et al., 2024). Model owners and service users therefore need a practical way to determine whether the deployed service still corresponds to the model it claims to serve. Because neither can inspect the serving model’s internals or rely on a potentially untrusted provider’s self-report, they must make this determination using only ordinary image-text queries.

Fingerprinting has been used for black-box model tamper detection. For deterministic image classifiers, prior methods construct fingerprints that maximize output sensitivity to small model-weight perturbations (He et al., 2019) or lie near decision boundaries (Wang et al., 2023), then infer tampering from changes in top-1 predictions. Generative LLMs, in contrast, require distributional evidence under stochastic decoding: ESF checks generated tokens at selected positions against recorded token sets (Bai et al., 2025b), whereas RESF applies a hypothesis test to the response distribution over these tokens (Hu et al., 2025). Both methods use high-perplexity textual fingerprints that are distinguishable from ordinary queries. RESF also requires a decoding-temperature estimate to specify its reference distribution; estimation error can compromise the test’s reliability.

To address these issues in VLM tamper detection, we propose *Latent Visual Sensitive Fingerprinting* (LVSF). Its general design is guided by a local Pearson-sensitivity analysis of the conditional top- k distribution of the first generated token. The analysis separates sensitivity into a distribu-

*Corresponding authors.

†Equal contribution.

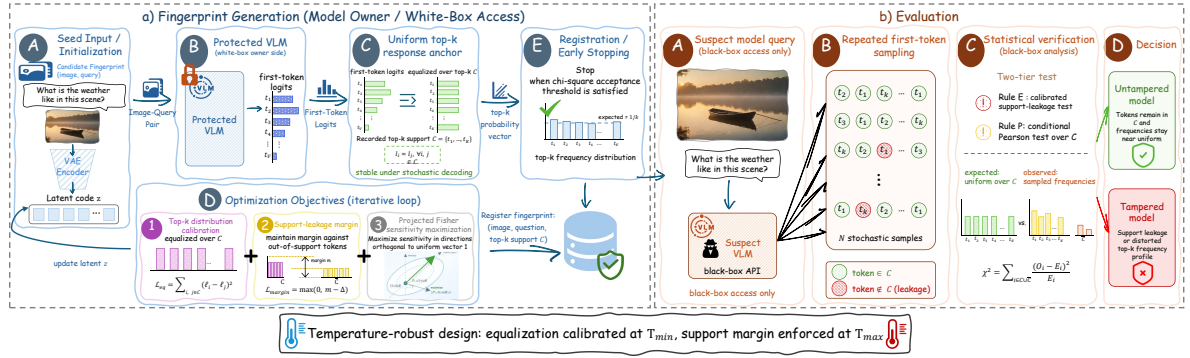


Figure 1: LVSF overview. VAE-latent image optimization with a natural-language query creates a uniform top- k anchor using distribution calibration, a support-leakage margin, and full-parameter Fisher sensitivity. Repeated API calls test support leakage and in-support token frequencies.

tion term, which determines whether sampled outputs reveal relative top- k logit shifts, and a Fisher term, which determines how strongly a fingerprint maps parameter changes to those shifts. Because tampering induces relative logit shifts of unknown direction, an isotropic model shows that a uniform top- k distribution maximizes average sensitivity. LVSF therefore calibrates the recorded top- k logits toward equality. An equal-logit anchor yields a uniform conditional reference at every positive temperature, eliminating the service-temperature estimation required by RESF. A margin separates the recorded support from all out-of-support tokens, while full-parameter Fisher sensitivity maximization amplifies the resulting observable frequency shifts. These sensitivity and anchor-construction principles apply to both LLMs and VLMs.

For VLMs, LVSF realizes these principles during fingerprint generation, as shown in Figure 1, by starting from a natural image and a natural-language query, keeping the query fixed, and optimizing only the image in the latent space of a pretrained VAE (Kingma and Welling, 2014). This VAE latent parameterization guides the search toward natural-looking fingerprint images. The resulting image–query pairs resemble ordinary VLM requests while exercising the VLM’s multimodal inference path. During black-box auditing, the verifier repeatedly queries the suspect API with each registered image–query pair and rejects either an excess of first generated tokens outside the recorded support or a significant deviation in conditional in-support frequencies.

Contributions. Our contributions are as follows.

- **A Pearson-sensitivity analysis for stochastic model outputs.** We derive a local sensitivity decomposition that separates output-shift observ-

ability from input-dependent exposure to parameter changes. Under an isotropic model of unknown relative logit shifts, it establishes the average optimality of uniform top- k response anchors.

- **Temperature-robust response anchors for black-box verification.** Equalized recorded top- k logits yield a temperature-invariant conditional reference, eliminating service-temperature estimation. LVSF combines this anchor with a margin separating recorded from out-of-support tokens and full-parameter Fisher sensitivity to make tampering-induced frequency shifts observable for tamper detection.
- **Visual-latent fingerprint generation for VLM tamper detection.** LVSF instantiates these principles by optimizing only the visual input in VAE latent space, while keeping a natural-language query fixed. To the best of our knowledge, the proposed LVSF is the first black-box VLM tamper-detection method.
- **Comprehensive empirical evaluation.** Across four VLMs, multiple tampering types, and decoding temperatures, LVSF achieves consistently high detection rates for tampered models, with no observed false positives for unmodified protected models. On Qwen2.5-VL-7B-Instruct, comparisons with baselines show that LVSF attains substantially higher detection rates.

2 Related Work

Black-box model tamper detection asks whether a servicing model still behaves as its protected version under normal queries. Classifier methods construct modification-sensitive fingerprints: SSF maximizes output sensitivity to small model-weight perturbations (He et al., 2019); PublicCheck gen-

erates smooth decision-boundary samples (Wang et al., 2023); and IBSF and SDBF exploit multiple boundary intersections and steep boundaries, respectively (Bai et al., 2024, 2025c). These methods target discriminative models, for which changes in deterministic top-1 labels are the verification signal. They do not directly extend to generative LLMs and VLMs, whose token outputs are stochastic even when the model is intact.

LLM fingerprinting extends verification from deterministic label matching to stochastic text generation. ESF checks generated tokens at selected positions against recorded token sets (Bai et al., 2025b), whereas RESF applies support and Pearson hypothesis tests to their response distribution (Hu et al., 2025). Both ESF and RESF use high-perplexity textual fingerprints distinguishable from ordinary queries. RESF further estimates the service temperature to construct its reference distribution, so estimation error can undermine test reliability.

To our knowledge, no prior work targets black-box VLM tamper detection. The closest work, SIF, is a VLM intellectual-property protection method that constructs semantically in-distribution fingerprints for ownership verification (Zhao et al., 2026). It distinguishes models derived from a protected model from independently trained models, rather than determining whether a servicing model is the protected model. Separately, RUT is a black-box LLM model-equality test that evaluates whether a servicing API is behaviorally consistent with a local reference model. It ranks each API response within the corresponding reference-model response distribution and tests the resulting ranks for uniformity (Zhu et al., 2026).

3 Threat Model

We consider integrity verification for a VLM served through a cloud API. As in prior fingerprinting work (He et al., 2019; Bai et al., 2025b; Hu et al., 2025), the model owner has white-box access to the protected model before deployment and uses it to generate private fingerprints. Each fingerprint comprises a VLM query, protected-model reference information, and testing parameters. During auditing, an authorized verifier has only black-box access to the suspect API and assesses whether its behavior is consistent with the protected model. At verification time, the verifier cannot access logits, hidden states, gradients, model parameters, sampling seeds, or the fingerprint-generation procedure.

An adversary may be an untrusted service provider or an external party that seeks to modify the hosted model while preserving normal API functionality and avoiding detection. We assume that it knows the fingerprinting method and verification procedure but cannot access active private fingerprints or their reference information. It may seek to identify verification queries from API traffic. Practical tampering includes fine-tuning, safety or preference alignment, poisoning, post-training quantization or compression, and model replacement.

4 VLM Fingerprint Sensitivity

This section derives the design principles behind LVSF. The key result is that black-box first-token tamper sensitivity is governed by two factors: a top- k distribution term and a Fisher sensitivity term. LVSF jointly optimizes these factors while enforcing a valid recorded support. Complete proofs and derivations are provided in Appendix A.

4.1 Setup

Let $x = (I, q)$ be an image-query fingerprint. For model parameters θ , let $\ell_\theta(x) \in \mathbb{R}^{|\mathcal{V}|}$ be the first-token logit vector over the vocabulary $\mathcal{V}$. LVSF records a support $C \subset \mathcal{V}$ with $|C| = k \geq 2$. At decoding temperature $T > 0$, the conditional top- k distribution is

$$\pi_{\theta,T}^C(i | x) = \frac{\exp(\ell_{\theta,i}(x)/T)}{\sum_{j \in C} \exp(\ell_{\theta,j}(x)/T)}, \quad i \in C. \quad (1)$$

Let u_C be the uniform distribution on C , and define

$$J(\pi) = \text{diag}(\pi) - \pi\pi^\top, \quad P_{1^\perp} = I_k - \frac{1}{k}\mathbf{1}\mathbf{1}^\top. \quad (2)$$

The matrix $J(\pi)$ is the covariance matrix of a one-hot categorical draw with probabilities π and the Jacobian of the softmax.

Assumption 1 (Local regularity and fixed recorded support). *The deployment temperatures lie in a compact interval $\mathcal{T} = [T_{\min}, T_{\max}] \subset (0, \infty)$. For a fixed fingerprint x and recorded support C , the map $\theta \mapsto \ell_{\theta,C}(x)$ is twice continuously differentiable in a neighborhood of the protected parameter θ_0 . Verification samples are conditionally independent given the model, fingerprint, and decoding temperature.*

4.2 Local Sensitivity and the Two Controlling Factors

The local tamper-detection problem can be written as

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_0 + \Delta\theta. \quad (3)$$

We measure local distributional change with Pearson χ^2 divergence. Let $G_C(x) = \partial \ell_{\theta, C}(x) / \partial \theta$ be the Jacobian of the recorded top- k logits with respect to model parameters. Under a small perturbation $\Delta\theta$, the corresponding top- k logit shift is $G_C(x)\Delta\theta + O(\|\Delta\theta\|^2)$.

Theorem 1 (Sensitivity decomposition). *For a fixed fingerprint x , recorded support C , and temperature T , the local Pearson sensitivity of the conditional top- k distribution satisfies*

$$D_{\chi^2}(\pi_{\theta_0 + \Delta\theta, T}^C(\cdot | x), \pi_{\theta_0, T}^C(\cdot | x)) = \Delta\theta^\top F_{C, T}(x) \Delta\theta + o(\|\Delta\theta\|^2), \quad (4)$$

where

$$F_{C, T}(x) = \frac{1}{T^2} G_C(x)^\top J(\pi_{\theta_0, T}^C(\cdot | x)) G_C(x). \quad (5)$$

We denote the leading local sensitivity by $S_{C, T}(x; \Delta\theta) := \Delta\theta^\top F_{C, T}(x) \Delta\theta + o(\|\Delta\theta\|^2)$. Equivalently,

$$\Delta\theta^\top F_{C, T}(x) \Delta\theta = \frac{1}{T^2} \left\| J(\pi_{\theta_0, T}^C)^{1/2} G_C(x) \Delta\theta \right\|_2^2. \quad (6)$$

Implication. Equation 5 exposes two factors that jointly determine fingerprint sensitivity.

- The **top- k distribution term** $J(\pi)$ determines how observable a relative top- k logit shift is through sampled first-token frequencies. If π is too concentrated, most samples fall on a dominant token and small shifts among the remaining tokens are difficult to observe.
- The **Fisher sensitivity term** $G_C(x)$ determines how strongly the fingerprint input converts a parameter perturbation into a top- k logit shift across model parameters.

A strong VLM fingerprint should therefore optimize both terms jointly.

4.3 Uniform top- k Anchors

Only relative top- k logit changes are visible, because $J(\pi)1 = 0$: adding the same constant to every top- k logit does not change the conditional softmax. When the direction of tampering is unknown, a natural criterion is average sensitivity over possible relative logit perturbations.

Definition 1 (Average logit-level sensitivity). *Let A be a random top- k logit perturbation with fi-*

nite second moment $\Sigma = \mathbb{E}[AA^\top]$. For $\pi = \text{softmax}(z/T)$, define

$$\bar{s}_T(\pi) = \frac{1}{T^2} \mathbb{E}[A^\top J(\pi) A] = \frac{1}{T^2} \text{tr}(J(\pi) \Sigma). \quad (7)$$

Assumption 2 (Isotropic unknown relative logit effect). *The unknown logit perturbation has no preferred direction on the identifiable zero-sum subspace:*

$$\Sigma = \sigma^2 P_{1^\perp}, \quad \sigma^2 > 0. \quad (8)$$

Theorem 2 (Uniform top- k maximizes average distribution sensitivity). *Under Assumption 2, among all probability distributions π on a fixed k -point support,*

$$\bar{s}_T(\pi) \leq \frac{\sigma^2}{T^2} \left(1 - \frac{1}{k}\right), \quad (9)$$

with equality if and only if $\pi = u_C$.

The proof is immediate from $J(\pi)P_{1^\perp} = J(\pi)$:

$$\bar{s}_T(\pi) = \frac{\sigma^2}{T^2} \text{tr} J(\pi) = \frac{\sigma^2}{T^2} (1 - \|\pi\|_2^2), \quad (10)$$

and $\|\pi\|_2^2$ is minimized uniquely by the uniform distribution. Thus, under the isotropic unknown-direction model, a uniform top- k distribution is the most sensitive operating point on average.

Definition 2 (Uniform top- k response anchor). *A fingerprint x lies on a uniform top- k response anchor with support C if*

$$\ell_{\theta_0, i}(x) = \ell_{\theta_0, j}(x), \quad \forall i, j \in C. \quad (11)$$

It is γ -stable if, in addition,

$$\min_{i \in C} \ell_{\theta_0, i}(x) \geq \max_{r \notin C} \ell_{\theta_0, r}(x) + \gamma. \quad (12)$$

If the equal-logit condition in Eq. 11 holds, then $\pi_{\theta_0, T}^C = u_C$ for every temperature $T > 0$. At this anchor,

$$J(u_C) = \frac{1}{k} P_{1^\perp}, \quad (13)$$

and the local sensitivity becomes

$$S_{C, T}(x; \Delta\theta) = \frac{1}{kT^2} \|P_{1^\perp} G_C(x) \Delta\theta\|_2^2 + o(\|\Delta\theta\|^2). \quad (14)$$

The projection $P_{1^\perp}$ removes the unobservable common-logit shift. Eq. 14 shows why LVSF calibrates the top- k distribution toward a uniform anchor under the isotropic unknown-direction model: it maximizes the average distribution term while leaving the remaining sensitivity to be improved through the Fisher term.

4.4 The Fisher Sensitivity Term

The uniform anchor is average-optimal for unknown relative logit perturbations, but a useful

fingerprint must also strongly expose parameter changes through its input-dependent Jacobian. We therefore use the full-parameter Fisher sensitivity induced by the recorded top- k logits. At an exact uniform anchor,

$$\mathcal{S}_{\text{FS}}(x; T) = \frac{1}{kT^2} \|P_{1^\perp} G_C(x)\|_F^2, \quad (15)$$

where the Frobenius norm is taken over all model parameters. This score is large when the fingerprint makes small parameter changes produce large observable relative logit shifts. LVSF therefore maximizes Eq. 15 together with the uniformity and support-validity terms in one objective.

4.5 Temperature-robust Generation

Deployment APIs may use different decoding temperatures. LVSF therefore calibrates anchors conservatively. Let $d_i = \ell_i - \frac{1}{k} \sum_{j \in C} \ell_j$ be centered top- k logits and

$$\pi_T(i) = \frac{\exp(d_i/T)}{\sum_{j \in C} \exp(d_j/T)}. \quad (16)$$

The KL deviation $D_{\text{KL}}(\pi_T \| u_C)$ is non-increasing in T ; equivalently, uniformity is hardest to certify at the lowest deployment temperature $T_{\min}$. Thus, if the anchor passes the uniformity criterion at $T_{\min}$, it passes it throughout $[T_{\min}, T_{\max}]$. Conversely, calibration at any $T_g > T_{\min}$ is unsafe because a fingerprint may appear uniform at T_g but not at $T_{\min}$.

Support leakage is controlled by the margin. If Eq. 12 holds and $M = |\mathcal{V}| - k$, then

$$P_{\theta, T}(Y \notin C \mid x) \leq \frac{Me^{-\gamma/T}}{k + Me^{-\gamma/T}}, \quad (17)$$

which is worst at $T_{\max}$. Finally, at an exact uniform anchor, the leading local sensitivity is proportional to T^{-2} , so the weakest deployment power is also attained at $T_{\max}$. These observations yield LVSF’s generation rule: calibrate the top- k distribution term at $T_{\min}$, enforce the support-leakage margin at $T_{\max}$, and maximize full-parameter Fisher sensitivity in a single weighted objective. LVSF realizes this rule through visual-latent optimization for VLMs. The sensitivity analysis, uniform-anchor construction, support separation, and Fisher-sensitivity maximization also apply to LLMs.

5 Latent Visual Sensitive Fingerprinting

5.1 Overview

Given an initial natural image I_0 and a natural-language query q , LVSF keeps q fixed and optimizes only the image in the latent space of a pretrained

VAE. Let E and D be the VAE encoder and decoder. We initialize $z_0 = E(I_0)$ and generate the visual fingerprint as $I(z) = D(z)$. The support size k is fixed, and LVSF selects the token identities once from the protected model’s top- k first-token logits on the initial pair (I_0, q) . It then optimizes the image so that these selected tokens have nearly equal logits while retaining the same support. After optimization, the owner records the support C^* , image I^* , query q , and aggregate protected-model probability outside C^* at $T_{\max}$, the highest temperature in the pre-specified deployment range.

Figure 1 summarizes the pipeline. Unlike text-only LLM fingerprinting, LVSF optimizes only a visual input through a learned VAE latent parameterization while keeping the text query natural. The resulting image–query fingerprint activates both the visual and language components of the VLM.

5.2 Fingerprint Optimization

LVSF jointly calibrates the fixed top- k support toward a uniform null, maintains a margin separating it from all out-of-support tokens, and increases Fisher sensitivity with respect to all model parameters. The first term, the *top- k distribution calibration* loss, drives the conditional distribution toward the uniform anchor, which maximizes average sensitivity under isotropic unknown relative perturbations. Let C be the fixed support selected from (I_0, q) and let

$$\begin{aligned} \mu_C(z) &= \frac{1}{k} \sum_{i \in C} \ell_i(z), \\ \tilde{d}_i(z; T_{\min}) &= \frac{\ell_i(z) - \mu_C(z)}{T_{\min}}. \end{aligned} \quad (18)$$

At lower temperatures, a fixed top- k logit gap produces a less uniform conditional distribution, so uniformity is hardest to certify at $T_{\min}$. The normalized gap $\tilde{d}_i$ therefore measures deviation at this worst-case temperature, and we use the *top- k distribution calibration (TDC) loss*

$$\mathcal{L}_{\text{TDC}}(z; T_{\min}) = \frac{1}{k} \sum_{i \in C} \rho_\delta(\tilde{d}_i(z; T_{\min})), \quad (19)$$

with $\delta > 0$ denoting the Huber transition parameter and

$$\rho_\delta(a) = \begin{cases} \frac{1}{2}a^2, & |a| \leq \delta, \\ \delta(|a| - \frac{1}{2}\delta), & |a| > \delta. \end{cases} \quad (20)$$

This Huber relaxation drives $\pi_{\theta, T_{\min}}^C(\cdot \mid x)$ toward the uniform distribution u_C over C , precisely aligning logits near the anchor while avoiding unstable gradients when the initial gaps are large.

The second validity term is derived from the γ -stable support condition and the leakage bound in Eq. 17. To make this role explicit, we call it the *support-leakage margin (SLM) loss*. Let $\ell_{\text{tail}}(z) = \max_{r \notin C} \ell_r(z)$ and let $s_{\min}$ be a differentiable soft-min approximation of the weakest in-support logit, with smoothing parameter $\tau_s > 0$:

$$s_{\min}(z) = -\tau_s \log \sum_{i \in C} \exp(-\ell_i(z)/\tau_s). \quad (21)$$

Let $\gamma_{\text{sup}} > 0$ be the desired support-separation margin: every recorded-support logit must exceed every out-of-support logit by at least γ_{sup} . Calibrating this margin for $T_{\max}$ gives

$$\mathcal{L}_{\text{SLM}}(z; T_{\max}) = [\ell_{\text{tail}}(z) - s_{\min}(z) + \gamma_{\text{sup}}]_+. \quad (22)$$

When a target out-of-support leakage mass η_{sup} is specified, Eq. 17 suggests choosing

$$\gamma_{\text{sup}} \geq T_{\max} \log \frac{M(1 - \eta_{\text{sup}})}{k\eta_{\text{sup}}}, \quad (23)$$

where M is the number of out-of-support logits considered in the leakage bound. In implementation, this calibrated gap can be replaced by an equivalent tuned margin hyperparameter, but the term remains a $T_{\max}$ -side support-stability constraint.

The first two terms form the *valid-anchor calibration (VAC) loss*

$$\mathcal{L}_{\text{VAC}}(z) = \lambda_{\text{TDC}} \mathcal{L}_{\text{TDC}}(z; T_{\min}) + \lambda_{\text{SLM}} \mathcal{L}_{\text{SLM}}(z; T_{\max}). \quad (24)$$

The two terms serve distinct validity roles: $\mathcal{L}_{\text{TDC}}$ calibrates the distribution term, whereas $\mathcal{L}_{\text{SLM}}$ preserves a valid support for the registered top- k conditional test. Their weighted sum $\mathcal{L}_{\text{VAC}}$ is a validity loss rather than the full LVSF objective.

To increase tamper sensitivity, LVSF also maximizes the *Fisher sensitivity (FS) score*

$$\mathcal{S}_{\text{FS}}(z; T_{\max}) = \frac{1}{kT_{\max}^2} \|P_{1^\perp} G_C(z)\|_F^2, \quad (25)$$

where $G_C(z) = \partial \ell_{\theta, C}(I(z), q) / \partial \theta$ is the Jacobian of the selected top- k logits with respect to all model parameters. Computing the Fisher score over all model parameters at every optimization step is expensive. We partition the parameter vector into B disjoint blocks. Each step samples m blocks uniformly without replacement and evaluates the corresponding Jacobian terms exactly. Multiplying the sampled sum by B/m gives an unbiased estimator of the full-parameter objective. Details of the estimator appear in Appendix B.6.

Because the conditional top- k softmax depends only on relative logit values, the projection removes changes that add the same amount to every recorded-support logit and therefore leave the

conditional distribution unchanged. Combining validity and sensitivity gives the single LVSF objective

$$\mathcal{L}_{\text{LVSF}}(z) = \mathcal{L}_{\text{VAC}}(z) - \lambda_{\text{FS}} \mathcal{S}_{\text{FS}}(z; T_{\max}). \quad (26)$$

The weights in $\mathcal{L}_{\text{VAC}}$ and λ_{FS} balance anchor validity against parameter sensitivity. The final registered fingerprint is the optimized iterate that satisfies the uniformity criterion and support-leakage margin for C^* .

Chi-square-aligned stopping. In practice, optimization cannot make the top- k logits exactly equal, so the protected model’s conditional distribution $\pi^C(z)$ does not exactly reach the ideal uniform null u_C . We accept an optimized fingerprint once its residual deviation from u_C is statistically negligible for the verifier’s chi-square test. Let n_{gen} be the calibration sample-size parameter that sets this stopping tolerance, and let n_{det} denote the number of API queries made with each registered fingerprint during verification. Because chi-square deviations scale with sample size, choosing $n_{\text{gen}} \gg n_{\text{det}}$ makes the residual protected-model bias much smaller than verification sampling noise.

For a top- k probability vector p , the Pearson statistic against the uniform null at sample size n can be written as

$$X^2(p; n) = nk \sum_{i=1}^k \left(p_i - \frac{1}{k} \right)^2. \quad (27)$$

If

$$\max_{i \in C} \left| p_i - \frac{1}{k} \right| \leq \Delta, \quad (28)$$

then

$$X^2(p; n) \leq nk^2 \Delta^2. \quad (29)$$

Therefore, with significance level α , we use the coordinate-wise tolerance

$$\Delta_{n_{\text{gen}}} = \sqrt{\frac{\chi_{k-1, 1-\alpha}^2}{n_{\text{gen}} k^2}}. \quad (30)$$

The latent is accepted as a valid anchor when

$$\max_{i \in C} \left| \pi_i^C(z) - \frac{1}{k} \right| \leq \Delta_{n_{\text{gen}}}. \quad (31)$$

Because $n_{\text{gen}} \gg n_{\text{det}}$, the residual calibration error contributes a negligible share of the chi-square statistic under H_0 at detection time, so the null acceptance region is preserved while unnecessary optimization is avoided.

5.3 Black-box Verification

For a specified deployment temperature range $\mathcal{T} = [T_{\min}, T_{\max}]$, the owner records the aggregate tail mass

$$p_{\text{out}}^* = \Pr_{\theta_0, T_{\max}} (Y \notin C^* \mid I^*, q). \quad (32)$$

Under standard full-softmax temperature sampling, C^* is the protected model’s fixed top- k support, and its aggregate out-of-support probability is largest at the highest temperature in this range. Hence p_{out}^* upper-bounds the protected-model probability of an out-of-support token throughout $\mathcal{T}$.

For each registered fingerprint, the verifier queries the suspect model n_{det} times and records the first generated token. It first applies Rule E; Rule P is applied if Rule E does not reject.

Support-leakage check (Rule E). Let N_{out} be the number of out-of-support tokens and $Z \sim \text{Binomial}(n_{\text{det}}, p_{\text{out}}^*)$. Given a Rule-E false-alarm allowance β_E , define

$$r_E = \min \{r : \Pr(Z \geq r) \leq \beta_E\}. \quad (33)$$

where the minimum is over $r \in \{1, \dots, n_{\text{det}}\}$. Rule E rejects when $N_{\text{out}} \geq r_E$. Let $b_E = \Pr(Z \geq r_E)$ be its binomial-tail probability.

Conditional Pearson check (Rule P). Let N_i be the number of occurrences of token $i \in C^*$, and let $N_C = \sum_{i \in C^*} N_i = n_{\text{det}} - N_{\text{out}}$. Conditional on an in-support response, the protected-model null is uniform over C^* by anchor construction. The test statistic is

$$X^2 = \sum_{i \in C^*} \frac{(N_i - N_C/k)^2}{N_C/k}. \quad (34)$$

Rule P is applied at conditional level

$$\alpha_P = \frac{\alpha - b_E}{1 - b_E}, \quad (35)$$

and rejects when $X^2 > \chi_{k-1, 1-\alpha_P}^2$. Thus, when $\beta_E < \alpha$, the two rules allocate the nominal false-positive level α across support leakage and conditional in-support frequency shifts. If neither rule rejects, the verifier reports no evidence of tampering. Rule E detects increased aggregate mass outside the registered support, whereas Rule P detects shifts in relative in-support frequencies. Verification uses only API calls and sampled first tokens.

6 Experiments

6.1 Experimental Setup

Models and tampering types. We evaluate LVSF on four open-source instruction-tuned VLMs: Qwen2.5-VL-3B-Instruct (Bai et al., 2025a), Qwen2.5-VL-7B-Instruct (Bai et al., 2025a), Gemma-3-4B-it (Gemma Team, 2025), and Kimi-VL-A3B-Instruct (Kimi Team, 2025). They span different parameter scales and architectural families, allowing us to assess whether LVSF gen-

eralizes across heterogeneous VLM designs rather than being tailored to a single architecture.

We consider four common tampering scenarios: fine-tuning, poisoning, compression, and model replacement. For fine-tuning, we use Alpaca (Taori et al., 2023) for supervised fine-tuning and SPA-VL (Zhang et al., 2025) for multimodal safety alignment. For poisoning, we follow the backdoor construction of Xu et al. (2024a) with a 10% poison ratio. In both settings, we apply Freeze, which freezes approximately 70% of input-side parameters, and Low-Rank Adaptation (LoRA). For benign task adaptation, we additionally use Direct Preference Optimization (DPO) (Rafailov et al., 2023). All fine-tuning experiments use the ms-swift framework (Zhao et al., 2025); LoRA uses rank 8 and alpha 32. For compression, we apply 4-bit and 8-bit post-training quantization using bit-sandbytes (Dettmers et al., 2022, 2023). For cross-model replacement, we represent the registered support using decoded first-token strings rather than tokenizer-specific token IDs. We additionally evaluate vision-encoder-only tampering, in which only the vision encoder is quantized to INT8.

Fingerprint generation and detection settings.

We construct 100 initial image–query pairs, with images generated by SDXL (Podell et al., 2024), yielding 100 distinct fingerprints per setting. Unless otherwise specified, LVSF is optimized in VAE latent space using Adam (Kingma and Ba, 2015) with learning rate 7.5×10^{-4} , support size $k = 5$, and nominal false-positive level $\alpha = 0.05$. We target the practical deployment temperature range $\mathcal{T} = [0.1, 1.0]$: anchors are calibrated at $T_{\min} = 0.1$, and protected-model aggregate tail mass is recorded at $T_{\max} = 1.0$. The uniformity acceptance rule uses the calibration sample-size parameter $n_{\text{gen}} = 1000$ in Eq. 31, an order of magnitude larger than the per-fingerprint verification count below, so that the residual deviation from the uniform anchor is negligible for the chi-square detector. The LVSF objective combines anchor calibration, support-margin preservation, and full-parameter Fisher sensitivity through weighted loss terms.

For each verification decision, the verifier uses one fingerprint and queries the suspect API $n_{\text{det}} = 100$ times. It uses the two-tier procedure in Section 5 with nominal false-positive level $\alpha = 0.05$ and Rule-E allowance $\beta_E = 0.01$; Rule P uses the fingerprint-specific adjusted level in Eq. 35. Rather than producing an ensemble verdict, the 100 fin-

Tampering	Models	Qwen-3B			Qwen-7B			Gemma-4B			Kimi-A3B		
	Temperature	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
Fine-tuning	Freeze	1.00	1.00	0.99	1.00	1.00	0.98	1.00	1.00	1.00	1.00	0.99	0.98
	DPO	1.00	1.00	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97
	LoRA	1.00	1.00	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97
Poison	Freeze	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	1.00	1.00	0.99
	LoRA	1.00	1.00	0.97	1.00	1.00	0.99	1.00	1.00	0.98	1.00	1.00	0.99
Compress	4-bit	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	8-bit	1.00	0.99	0.91	1.00	1.00	0.99	1.00	1.00	0.99	1.00	0.97	0.81
Replacement	Qwen-3B	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
	Qwen-7B	1.00	1.00	0.99	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.99
	Gemma-4B	1.00	1.00	0.99	1.00	1.00	1.00	0.00	0.00	0.00	1.00	1.00	1.00
	Kimi-A3B	1.00	1.00	0.98	1.00	1.00	0.99	1.00	1.00	0.99	0.00	0.00	0.00

Table 1: Two-tier verification rejection rates. Each cell averages 100 final Rule-E/Rule-P decisions; higher is better for tampered rows, and the replacement-block diagonals correspond to unmodified protected models (ideal value 0). Qwen-3B/7B, Gemma-4B, and Kimi-A3B denote Qwen2.5-VL-3B/7B-Instruct, Gemma-3-4B-it, and Kimi-VL-A3B-Instruct.

gerprints yield 100 independent single-fingerprint decisions. We report their average *rejection rate*: for tampered models, it is the true detection rate; for an unmodified protected model, it is the fraction of decisions incorrectly rejected (the false-positive rate).

Baselines. RESF (Hu et al., 2025) is the only directly relevant prior baseline: it performs black-box tamper detection for stochastic generative models using high-perplexity textual fingerprints. We also evaluate RESF-image, our image-input variant that applies the same RESF approach to images.

Ablation. To isolate the contribution of Fisher sensitivity, we evaluate LVSF without its Fisher-sensitivity term.

6.2 Main Results

Table 1 reports final two-tier verification rejection rates across four VLMs, four tampering categories, and three decoding temperatures: Rule E checks support leakage, and Rule P tests conditional in-support frequency shifts. For tampered-model rows, the rejection rate is the detection rate, so higher is better. Across all VLMs and temperatures, LVSF attains detection rates of 0.95–1.00 for fine-tuning, 0.97–1.00 for poisoning, 0.81–1.00 for compression, and 0.98–1.00 for model replacement. The diagonal entries in the replacement block are results for the unmodified protected models; there, the rejection rate is the false-positive rate. No rejection (i.e., no false positive) was observed among the 100 single-fingerprint decisions in each reported unmodified-model setting;

this finite-sample result is not a population false-positive guarantee. Under a binomial model, zero rejections in 100 decisions gives a one-sided 95% upper confidence bound of approximately 2.95% on the corresponding false-positive probability. To isolate the visual pathway and test whether visual-input optimization can detect modifications that do not directly alter the language backbone, we also evaluate vision-encoder-only INT8 tampering. Across decoding temperatures, LVSF achieves an average detection rate of 87.4%, compared with 0.8% for text-only RESF and 7.4% for RESF-image. The complete setup and per-temperature results appear in Appendix B.4.

A notable advantage of LVSF is its robustness to decoding temperature. Across all tampered settings, the average detection rate remains at 100% when the temperature is 0.1, 99.8% at 0.5, and 98.2% even at 0.9. Although performance drops slightly at higher temperatures, especially under the milder 8-bit compression setting, the overall detection rate remains high.

These results indicate that calibration at the lowest temperature in the practical deployment range produces fingerprints that retain strong statistical sensitivity under different inference temperatures. Exact equality of the recorded top- k logits would yield a temperature-invariant conditional reference; in practice, optimization stops when the residual deviation satisfies Eq. (31). Because temperature rescales these remaining logit differences, the conditional response distributions and resulting detection rates need not be identical across temperatures.

We also find that larger top- k supports consis-

Tampering	Method	RESF			RESF-image			LVSF w/o FS			LVSF		
		0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
Fine-tuning	DPO	0.14	0.56	0.46	0.32	0.36	0.45	1.00	1.00	0.99	1.00	1.00	1.00
	LoRA	0.18	0.75	0.49	0.32	0.54	0.60	1.00	1.00	0.99	1.00	1.00	1.00
Poison	LoRA	0.25	0.81	0.51	0.38	0.58	0.61	1.00	1.00	0.99	1.00	1.00	0.99
Compress	8-bit	0.31	0.77	0.53	0.00	0.09	0.18	1.00	1.00	0.94	1.00	1.00	0.99
Replacement	Qwen-3B	0.60	0.97	0.63	0.56	0.62	0.71	1.00	1.00	0.99	1.00	1.00	1.00

Table 2: Two-tier verification detection rates; each cell averages 100 final single-fingerprint decisions. The protected model is Qwen2.5-VL-7B-Instruct (all tampered settings). In the Replacement row, Qwen-3B denotes the suspect model Qwen2.5-VL-3B-Instruct.

tently yield higher detection rates, since they provide more degrees of freedom for the chi-square test and a richer uniform response anchor; full results of this support-size calibration ablation appear in Appendix B.1. Tighter stopping targets on the uniform top- k anchor likewise improve detection rates at the cost of longer generation time; the full stopping-target calibration ablation is also reported there.

LVSF also remains robust across diverse decoding strategies, including variations in temperature, top- p , top- k , and repetition penalty, while the chat template and model-input construction are held fixed. The complete setup and results appear in Appendix B.5.

We further evaluate sensitivity to model-update scale using conventional parameter-space metrics. LVSF responds strongly even to small parameter shifts, demonstrating that detection is not confined to large modifications. Appendix B.3 provides the metric formulations, experimental configurations, and complete results.

6.3 Baseline Comparison and Ablation

Table 2 compares LVSF with RESF and RESF-image and reports the Fisher-sensitivity ablation. The ablated variant achieves a 99.3% average detection rate, substantially exceeding RESF (53.1%) and RESF-image (42.1%). Full LVSF modestly increases the average detection rate to 99.9%; the largest gain is under 8-bit compression at $T = 0.9$, where detection rises from 0.94 to 0.99.

We further evaluate five alternative fingerprint constructions to examine the contributions of LVSF’s visual-latent parameterization and anchor-design choices. Under INT8 tampering, full LVSF maintains 0.99–1.00 detection across temperatures with no observed false positives on the unmodified model, whereas the alternatives either generate false positives or degrade substantially at $T = 0.9$. Appendix B.2 provides their settings, protocol, and

complete results.

We also evaluate fingerprint stealthiness against an adaptive provider using a Deep- k NN out-of-distribution detector with frozen vision-encoder features to distinguish fingerprints from benign image queries (Sun et al., 2022). The detector achieves only 0.593 AUROC, detecting 7% of fingerprints at a 5% benign false-positive rate and none at a 1% rate. LVSF also preserves the visual appearance of source images, with SSIM 0.946, PSNR 33.67 dB, and VLM-judge consistency 4.97/5. Appendix B.4 provides the complete detector, fidelity, and visual-example evaluations.

7 Conclusion

We presented Latent Visual Sensitive Fingerprinting (LVSF) for black-box VLM tamper detection. A local Pearson-sensitivity analysis shows that a uniform top- k response anchor maximizes average sensitivity under isotropic unknown relative logit shifts. LVSF therefore equalizes recorded top- k logits, maintains support separation, and maximizes full-parameter Fisher sensitivity, yielding a conditional reference invariant to positive temperature scaling. Although LVSF realizes these principles through visual-latent optimization, the sensitivity analysis, uniform-anchor construction, support separation, and Fisher-sensitivity maximization apply to both LLMs and VLMs. For VLMs, LVSF optimizes only the image in a pretrained VAE’s latent space while keeping a natural-language query fixed, then applies a two-tier first-token verifier for support leakage and conditional frequency shifts. Across four VLMs and multiple tampering types, LVSF achieves consistently high detection rates without observed false positives; on Qwen2.5-VL-7B-Instruct, it substantially outperforms the evaluated RESF-based baselines.

Limitations

Fingerprint enrollment requires white-box access to a trusted protected VLM because LVSF optimizes visual fingerprints using its logits and parameter gradients, including full-parameter Fisher sensitivity. This limits LVSF to settings in which the model owner or an authorized auditor can access the protected model when enrolling fingerprints. Active fingerprints and their registered reference information must remain private.

The evaluation covers four open-source instruction-tuned VLMs and tampering through fine-tuning, poisoning, compression, and model replacement. Future work should assess larger proprietary VLMs, longer deployment lifecycles, and a wider range of API policies to characterize performance under commercial serving constraints.

LVSF assumes a stable or versioned model-input construction during each verification period. Although the additional evaluation shows robustness to common decoding variations, an incompatible chat-template change can alter the token sequence presented to the VLM and shift the recorded first-token distribution without changing the underlying model. LVSF does not distinguish such input-pipeline changes from model tampering; when a service intentionally changes its chat template or interface version, the owner should regenerate and register the corresponding fingerprints.

Finally, maximizing Fisher sensitivity over all model parameters avoids requiring prior knowledge of the modified layer or module, but increases fingerprint-generation cost. Future work may investigate more efficient sensitivity estimators and multi-fingerprint verification.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant No. 2024YFF0908104; the Collaborative Tackling Project of the Yangtze River Delta Sci-Tech Innovation Community under Grant No. 2024CSJGG01500; the National Natural Science Foundation of China under Grant Nos. 62532016, 25Z990200238, 6257071391, 62502309, and 62602422; and the Natural Science Foundation of Shanghai under Grant No. 25ZR1402211.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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, and 8 others. 2025a. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.
- Xiaofan Bai, Chaoxiang He, Xiaojing Ma, Bin Benjamin Zhu, and Hai Jin. 2024. Intersecting-boundary-sensitive fingerprinting for tampering detection of DNN models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 54402–54413. PMLR / OpenReview.net.
- Xiaofan Bai, Pingyi Hu, Xiaojing Ma, Linchen Yu, Dongmei Zhang, Qi Zhang, and Bin Benjamin Zhu. 2025b. ESF: efficient sensitive fingerprinting for black-box tamper detection of large language models. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 10477–10494. Association for Computational Linguistics.
- Xiaofan Bai, Shixin Li, Xiaojing Ma, Bin Benjamin Zhu, Dongmei Zhang, and Linchen Yu. 2025c. SDBF: Steep-decision-boundary fingerprinting for hard-label tampering detection of DNN models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29278–29287.
- DeepSeek-AI. 2026. DeepSeek-V4: Towards highly efficient million-token context intelligence. *CoRR*, abs/2606.19348.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Gemma Team. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.
- Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *CoRR*, abs/2404.07214.

- Zecheng He, Tianwei Zhang, and Ruby B. Lee. 2019. Sensitive-sample fingerprinting of deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4729–4737. Computer Vision Foundation / IEEE.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Pingyi Hu, Xiaofan Bai, Xiaojing Ma, Chaoxiang He, Dongmei Zhang, and Bin Benjamin Zhu. 2025. RESF: regularized-entropy-sensitive fingerprinting for black-box tamper detection of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 4889–4903. Association for Computational Linguistics.
- Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Kimi Team. 2025. Kimi-vl technical report. *CoRR*, abs/2504.07491.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer.
- Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. 2024. TrojVLM: Backdoor attack against vision language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXV*, *Lecture Notes in Computer Science*, pages 467–483. Springer.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252.
- Yiyu Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 20827–20840. PMLR.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. 2024. Q-VLM: post-training quantization for large vision-language models. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Shuo Wang, Sharif Abuadbbba, Sidharth Agarwal, Kristen Moore, Ruoxi Sun, Minhui Xue, Surya Nepal, Seyit Camtepe, and Salil S. Kanhere. 2023. Public-check: Public integrity verification for services of run-time deep models. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*, pages 1348–1365. IEEE.
- Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612.
- Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. 2024a. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*.

pages 3111–3126. Association for Computational Linguistics.

Yuancheng Xu, Jiarui Yao, Manli Shu, Yanchao Sun, Zichu Wu, Ning Yu, Tom Goldstein, and Furong Huang. 2024b. Shadowcast: Stealthy data poisoning attacks against vision-language models. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, Feng Zhao, Tao Gui, and Jing Shao. 2025. Spa-vl: A comprehensive safety preference alignment dataset for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19867–19878.

Yifei Zhao, Qian Lou, and Mengxin Zheng. 2026. Sif: Semantically in-distribution fingerprints for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17399–17408.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2025. SWIFT: A scalable lightweight infrastructure for fine-tuning. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 29733–29735. AAAI Press.

Xiaoyuan Zhu, Yaowen Ye, Tianyi Alex Qiu, Hanlin Zhu, Sijun Tan, Ajraf Mannan, Jonathan Michala, Raluca Ada Popa, and Willie Neiswanger. 2026. Auditing black-box LLM APIs with a rank-based uniformity test. In *The Fourteenth International Conference on Learning Representations*.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy M. Hospedales. 2024. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 62867–62891. PMLR / OpenReview.net.

A Complete Proofs for Uniform-Boundary Optimality and Temperature Robustness

A.1 Local Pearson sensitivity

Let $z \in \mathbb{R}^k$, $\pi = \text{softmax}(z/T)$, and define

$$q_r = \text{softmax}\left(\frac{z + ra}{T}\right). \quad (36)$$

The softmax derivative gives

$$q_r - \pi = \frac{r}{T} J(\pi) a + O(r^2). \quad (37)$$

For any $q = \pi + \delta$ with $\sum_i \delta_i = 0$ and $\pi_i > 0$,

$$D_{\chi^2}(q, \pi) = \sum_i \frac{\delta_i^2}{\pi_i} = \delta^\top \text{diag}(1/\pi) \delta. \quad (38)$$

Substituting Eq. (37),

$$D_{\chi^2}(q_r, \pi) = \frac{r^2}{T^2} a^\top J(\pi) \text{diag}(1/\pi) J(\pi) a + o(r^2). \quad (39)$$

It remains to note that

$$J(\pi) \text{diag}(1/\pi) J(\pi) = J(\pi). \quad (40)$$

Indeed, writing $D = \text{diag}(\pi)$, $J = D - \pi\pi^\top$, and $D^{-1} = \text{diag}(1/\pi)$,

$$J D^{-1} J = (D - \pi\pi^\top) D^{-1} (D - \pi\pi^\top) \quad (41)$$

Since $\pi^\top D^{-1} \pi = \sum_i \pi_i = 1$, Eq. (40) follows. Therefore

$$\lim_{r \rightarrow 0} \frac{1}{r^2} D_{\chi^2}(q_r, \pi) = \frac{1}{T^2} a^\top J(\pi) a. \quad (42)$$

Finally,

$$a^\top J(\pi) a = \sum_i \pi_i a_i^2 - \left(\sum_i \pi_i a_i \right)^2 \quad (43)$$

This proves Eq. (5).

For parameter perturbations, set

$$G_{\mathcal{C}}(x) = \frac{\partial \ell_{\theta, \mathcal{C}}(x)}{\partial \theta}, \quad a = G_{\mathcal{C}}(x) \Delta \theta. \quad (44)$$

Assumption 1 gives

$$\ell_{\theta + \Delta \theta, \mathcal{C}}(x) = \ell_{\theta, \mathcal{C}}(x) + G_{\mathcal{C}}(x) \Delta \theta + O(\|\Delta \theta\|^2), \quad (45)$$

and hence

$$\begin{aligned} S_{\mathcal{C}, T}(x; \Delta \theta) &= \Delta \theta^\top \left[\frac{1}{T^2} G_{\mathcal{C}}(x)^\top J(\pi) G_{\mathcal{C}}(x) \right] \Delta \theta \\ &\quad + o(\|\Delta \theta\|^2), \end{aligned} \quad (46)$$

which is Eq. (5). If $\Delta \theta_n = v/\sqrt{n}$ and the protected conditional null is uniform, then

$$n S_{\mathcal{C}, T}(x; \Delta \theta_n) \rightarrow v^\top F_{\mathcal{C}, T}(x) v, \quad (47)$$

which is the local noncentrality coefficient of the conditional Pearson test.

A.2 Average-sensitivity optimality of the uniform distribution

By Definition 1,

$$\bar{s}_T(\pi) = \frac{1}{T^2} \mathbb{E}[A^\top J(\pi) A] = \frac{1}{T^2} \text{tr}(J(\pi) \Sigma), \quad (48)$$

where $\Sigma = \mathbb{E}[A A^\top]$. Under Assumption 2, $\Sigma = \sigma^2 P_{1^\perp}$. Since $J(\pi) \mathbf{1} = 0$, we have $J(\pi) P_{1^\perp} = J(\pi)$, and therefore

$$\bar{s}_T(\pi) = \frac{\sigma^2}{T^2} \text{tr} J(\pi). \quad (49)$$

Now

$$\begin{aligned} \text{tr} J(\pi) &= \text{tr}(\text{diag}(\pi)) - \text{tr}(\pi \pi^\top) \\ &= 1 - \|\pi\|_2^2. \end{aligned} \quad (50)$$

For any probability vector π on k points,

$$\|\pi\|_2^2 = \sum_{i=1}^k \pi_i^2 \geq \frac{1}{k} \left(\sum_{i=1}^k \pi_i \right)^2 = \frac{1}{k}, \quad (51)$$

with equality if and only if $\pi_1 = \dots = \pi_k = 1/k$. Thus

$$\bar{s}_T(\pi) \leq \frac{\sigma^2}{T^2} \left(1 - \frac{1}{k} \right), \quad (52)$$

with equality if and only if $\pi = u_{\mathcal{C}}$.

Remark 1 (Why this is an average, not universal, optimality statement). *If Σ is not proportional to $P_{1^\perp}$, the objective is $\text{tr}(J(\pi) \Sigma)$ and the uniform distribution need not be globally optimal. For a fixed known vector a , maximizing $\text{Var}_{i \sim \pi}(a_i)$ over the closed simplex places mass $1/2$ on a maximizer of a_i and mass $1/2$ on a minimizer of a_i . Thus the uniform top- k boundary is optimal for unknown isotropic relative perturbations; sensitivity to parameter changes is handled by the full-parameter Fisher objective in Eq. (25).*

A.3 Uniform boundary and exact boundary sensitivity

If the top- k distribution calibration loss in Eq. (19) reaches zero, equivalently $\ell_{\theta, i}(x) = \ell_{\theta, j}(x)$ for all $i, j \in \mathcal{C}$, then all numbers $\ell_{\theta, i}(x)/T$, $i \in \mathcal{C}$, are equal for every $T > 0$. Therefore the conditional top- k softmax is exactly uniform. At $\pi = u_{\mathcal{C}}$, direct substitution gives

$$J(u_{\mathcal{C}}) = \text{diag}(u_{\mathcal{C}}) - u_{\mathcal{C}} u_{\mathcal{C}}^\top = \frac{1}{k} P_{1^\perp}. \quad (53)$$

Using the local parameter sensitivity from Appendix A.1 and $J(u_{\mathcal{C}}) = k^{-1} P_{1^\perp}$, let $\delta_{\mathcal{C}}(x, \Delta \theta) = P_{1^\perp} G_{\mathcal{C}}(x) \Delta \theta$. Then

$$S_{\mathcal{C}, T}(x; \Delta \theta) = \frac{\|\delta_{\mathcal{C}}(x, \Delta \theta)\|_2^2}{k T^2} + o(\|\Delta \theta\|^2). \quad (54)$$

This proves Eq. (14).

A.4 Temperature robustness and the correct generation temperature

This appendix subsection contains the temperature analysis that was removed from the main text. It chooses the calibration temperature and identifies the worst deployment temperatures for validity and power; detailed derivations are given in Appendices A.5–A.10.

Let $\mathcal{T} = [T_{\min}, T_{\max}]$ be the deployment temperature range. For fixed top- k logits define the centered logit vector

$$d_i = \ell_i - \frac{1}{k} \sum_{j \in \mathcal{C}} \ell_j, \quad \sum_{i \in \mathcal{C}} d_i = 0, \quad (55)$$

and write

$$\pi_T(i) = \frac{\exp(d_i/T)}{\sum_{j \in \mathcal{C}} \exp(d_j/T)}. \quad (56)$$

Theorem 3 (Temperature-robust calibration rule).

The KL deviation $D_{\text{KL}}(\pi_T \| u_{\mathcal{C}})$ is non-increasing in T . Hence, for any $\varepsilon > 0$ and for all $T_d \in [T_{\min}, T_{\max}]$

$$D_{\text{KL}}(\pi_{T_{\min}} \| u_{\mathcal{C}}) \leq \varepsilon \implies D_{\text{KL}}(\pi_{T_d} \| u_{\mathcal{C}}) \leq \varepsilon. \quad (57)$$

Conversely, for any $T_g > T_{\min}$ and any $0 < \varepsilon < \log k$, there exist centered nonzero top- k logits d such that

$$D_{\text{KL}}(\pi_{T_g} \| u_{\mathcal{C}}) < \varepsilon \text{ but } D_{\text{KL}}(\pi_{T_{\min}} \| u_{\mathcal{C}}) > \varepsilon. \quad (58)$$

Consequently, among single calibration temperatures inside the deployment interval,

$$T_g^* = T_{\min} \quad (59)$$

is the unique safe and nonconservative choice.

Two additional robustness checks determine which endpoint of the deployment interval should be used. Let $\bar{\ell}_{\mathcal{C}} = \frac{1}{k} \sum_{j \in \mathcal{C}} \ell_j$ be the average logit over class set $\mathcal{C}$. If the top- k logits are approximately equal in the sense that $\max_{i \in \mathcal{C}} |\ell_i - \bar{\ell}_{\mathcal{C}}| \leq \varepsilon$, then for any temperature $T > 0$,

$$\|\pi_T^{\mathcal{C}} - u_{\mathcal{C}}\|_{\infty} \leq \frac{e^{2\varepsilon/T} - 1}{k}. \quad (60)$$

whose right-hand side is worst at $T_{\min}$. If the support-margin condition targeted by the SLM loss in Eq. (22), $\min_{i \in \mathcal{C}} \ell_i \geq \max_{r \notin \mathcal{C}} \ell_r + \gamma$, holds and $m = |\mathcal{V}| - k$, then

$$\mathbb{P}_{\theta, T}(Y \notin \mathcal{C} \mid x) \leq \frac{me^{-\gamma/T}}{k + me^{-\gamma/T}}, \quad (61)$$

whose right-hand side is worst at $T_{\max}$. Thus LVSF checks uniformity at $T_{\min}$ and support leakage at $T_{\max}$.

At an exact uniform response anchor, the only visible local perturbation is the projected top- k

logit shift $\delta_{\mathcal{C}}(x, \Delta\theta) = P_{1^{\perp}} G_{\mathcal{C}}(x) \Delta\theta$. Since the leading sensitivity is proportional to T^{-2} , its worst case over $\mathcal{T}$ is attained at $T_{\max}$:

$$\inf_{T_d \in \mathcal{T}} \mathcal{S}_{\mathcal{C}, T_d}(x; \Delta\theta) = \frac{\|\delta_{\mathcal{C}}(x, \Delta\theta)\|_2^2}{kT_{\max}^2} + o(\|\Delta\theta\|^2). \quad (62)$$

The remainder is uniform over $T_d \in \mathcal{T}$. Therefore uniformity is hardest to certify at $T_{\min}$, while local detection power and support stability are controlled by the high-temperature endpoint.

A.5 Monotonicity of KL uniformity deviation

Let $d \in \mathbf{1}^{\perp}$ be fixed and put $\beta = 1/T$. Define

$$\pi_{\beta}(i) = \frac{\exp(\beta d_i)}{Z(\beta)}, \quad Z(\beta) = \sum_{i \in \mathcal{C}} \exp(\beta d_i). \quad (63)$$

Since $u_i = 1/k$,

$$\begin{aligned} D_{\text{KL}}(\pi_{\beta} \| u_{\mathcal{C}}) &= \sum_i \pi_{\beta}(i) \log \frac{\pi_{\beta}(i)}{1/k} \\ &= \beta \mathbb{E}_{\pi_{\beta}}[d] - \log Z(\beta) + \log k. \end{aligned} \quad (64)$$

We use the identities $\partial_{\beta} \log Z(\beta) = \mathbb{E}_{\pi_{\beta}}[d]$ and $\partial_{\beta} \mathbb{E}_{\pi_{\beta}}[d] = \text{Var}_{\pi_{\beta}}(d)$. Differentiating Eq. (64) gives

$$\partial_{\beta} D_{\text{KL}}(\pi_{\beta} \| u_{\mathcal{C}}) = \beta \text{Var}_{\pi_{\beta}}(d) \geq 0. \quad (65)$$

Thus KL deviation from uniformity is nondecreasing in inverse temperature β and non-increasing in temperature T . It is strictly monotone for $\beta > 0$ whenever d is nonconstant.

A.6 Necessity of lowest-temperature calibration

The sufficiency statement in Eq. (57) follows from Appendix A.5. We prove the converse.

Choose a centered vector $d_0 \in \mathbf{1}^{\perp}$ with at least two distinct coordinates and a unique largest coordinate. Define

$$\phi(s) = D_{\text{KL}}(\text{softmax}(sd_0) \| u_{\mathcal{C}}), \quad s \geq 0. \quad (66)$$

By the KL monotonicity result, ϕ is continuous, $\phi(0) = 0$, and ϕ is strictly increasing on $(0, \infty)$. Since d_0 has a unique largest coordinate, $\text{softmax}(sd_0)$ converges to a point mass as $s \rightarrow \infty$, hence

$$\lim_{s \rightarrow \infty} \phi(s) = \log k. \quad (67)$$

For any $0 < \varepsilon < \log k$, there is a unique $s_{\varepsilon} > 0$ such that $\phi(s_{\varepsilon}) = \varepsilon$. Let $T_g > T_{\min}$ and choose

$$\alpha \in (T_{\min} s_{\varepsilon}, T_g s_{\varepsilon}). \quad (68)$$

Set $d = \alpha d_0$. Then

$$\frac{\alpha}{T_g} < s_{\varepsilon} < \frac{\alpha}{T_{\min}}. \quad (69)$$

Strict monotonicity of ϕ yields

$$\begin{aligned} D_{\text{KL}}(\pi_{T_g} \| u_{\mathcal{C}}) &= \phi(\alpha/T_g) < \varepsilon, \\ D_{\text{KL}}(\pi_{T_{\min}} \| u_{\mathcal{C}}) &= \phi(\alpha/T_{\min}) > \varepsilon. \end{aligned} \quad (70)$$

This proves Eq. (58).

A.7 Formal optimality of the generation temperature

Recall the acceptance set $\mathcal{A}_T(\varepsilon)$: it contains the centered top- k logit vectors $d \in \mathbf{1}^\perp$ whose temperature- T softmax is within KL tolerance ε of the uniform distribution $u_{\mathcal{C}}$. In symbols,

$$\mathcal{A}_T(\varepsilon) = \{d \in \mathbf{1}^\perp : D_T(d) \leq \varepsilon\}, \quad (71)$$

where $D_T(d) = D_{\text{KL}}(\text{softmax}(d/T) \| u_{\mathcal{C}})$. The monotonicity proof implies that if $T_1 \leq T_2$, then

$$\mathcal{A}_{T_1}(\varepsilon) \subseteq \mathcal{A}_{T_2}(\varepsilon). \quad (72)$$

Consequently,

$$\bigcap_{T_d \in [T_{\min}, T_{\max}]} \mathcal{A}_{T_d}(\varepsilon) = \mathcal{A}_{T_{\min}}(\varepsilon). \quad (73)$$

A single generation temperature $T_g \in [T_{\min}, T_{\max}]$ is safe if passing at T_g implies passing at all deployment temperatures, i.e.

$$\mathcal{A}_{T_g}(\varepsilon) \subseteq \bigcap_{T_d \in [T_{\min}, T_{\max}]} \mathcal{A}_{T_d}(\varepsilon) = \mathcal{A}_{T_{\min}}(\varepsilon). \quad (74)$$

If $T_g = T_{\min}$, Eq. (74) holds with equality. If $T_g > T_{\min}$, the construction in Appendix A.6 exhibits a vector $d \in \mathcal{A}_{T_g}(\varepsilon)$ with $d \notin \mathcal{A}_{T_{\min}}(\varepsilon)$, so Eq. (74) fails. Thus $T_g^* = T_{\min}$ is the unique safe single calibration temperature within the deployment interval.

If one allowed $T_g < T_{\min}$, then Eq. (72) gives $\mathcal{A}_{T_g}(\varepsilon) \subseteq \mathcal{A}_{T_{\min}}(\varepsilon)$, so such a choice would be safe but more conservative: it would reject some fingerprints that are valid throughout the actual deployment interval. Therefore $T_{\min}$ is the nonconservative safe choice.

A.8 Approximate logit equalization

Let

$$d_i = \ell_i - \bar{\ell}_{\mathcal{C}}, \quad \bar{\ell}_{\mathcal{C}} = \frac{1}{k} \sum_{j \in \mathcal{C}} \ell_j, \quad (75)$$

and assume $|d_i| \leq \epsilon$ for every $i \in \mathcal{C}$. Then

$$e^{-\epsilon/T} \leq e^{d_i/T} \leq e^{\epsilon/T} \quad (76)$$

for each numerator term, while

$$ke^{-\epsilon/T} \leq \sum_{j \in \mathcal{C}} e^{d_j/T} \leq ke^{\epsilon/T}. \quad (77)$$

Therefore

$$\frac{e^{-2\epsilon/T}}{k} \leq \pi_T^{\mathcal{C}}(i) \leq \frac{e^{2\epsilon/T}}{k}. \quad (78)$$

The upper bound implies

$$\pi_T^{\mathcal{C}}(i) - \frac{1}{k} \leq \frac{e^{2\epsilon/T} - 1}{k}, \quad (79)$$

and the lower bound implies

$$\frac{1}{k} - \pi_T^{\mathcal{C}}(i) \leq \frac{1 - e^{-2\epsilon/T}}{k} \leq \frac{e^{2\epsilon/T} - 1}{k}. \quad (80)$$

This proves Eq. (60). Since $e^{2\epsilon/T}$ decreases as T increases, the bound is worst at $T_{\min}$.

A.9 Support leakage under a margin

Assume

$$\min_{i \in \mathcal{C}} \ell_i \geq \max_{r \notin \mathcal{C}} \ell_r + \gamma, \quad \gamma \geq 0. \quad (81)$$

Let $a = \min_{i \in \mathcal{C}} \ell_i$ and $m = |\mathcal{V}| - k$. Then every in-support logit is at least a , while every out-of-support logit is at most $a - \gamma$. Therefore, the out-of-support softmax mass is bounded by

$$\mathbb{P}_{\theta, T}(Y \notin \mathcal{C} \mid x) \leq \frac{me^{-\gamma/T}}{k + me^{-\gamma/T}}. \quad (82)$$

Indeed, the numerator is at most $m \exp((a - \gamma)/T)$, while the in-support contribution to the denominator is at least $k \exp(a/T)$; dividing both terms by $\exp(a/T)$ gives Eq. (82). The right-hand side is increasing in $e^{-\gamma/T}$, and $e^{-\gamma/T}$ is non-decreasing in T for $\gamma \geq 0$. Thus the bound is worst at $T_{\max}$.

For verification, LVSF directly records the protected-model aggregate tail mass p_{out}^* at $T_{\max}$ rather than calibrating Rule E from this potentially loose vocabulary-wide bound. Since $\tau_T = \mathbb{P}_{\theta, T}(Y \notin \mathcal{C} \mid x) \leq p_{\text{out}}^*$ for every $T \in \mathcal{T}$, the number of out-of-support tokens in n independent verification samples is stochastically dominated by $\text{Binomial}(n, p_{\text{out}}^*)$. Consequently, the threshold in Eq. (33) gives Rule E false-alarm probability at most β_E .

A.10 Worst-case deployment sensitivity

At the exact uniform response anchor, Eq. (54) shows that the leading local sensitivity is proportional to T^{-2} .

Under Assumption 1, the second-order remainders in the logit and softmax expansions are uniform over $\mathcal{T} = [T_{\min}, T_{\max}]$. Thus the infimum over $\mathcal{T}$ is attained at $T_{\max}$:

$$\inf_{T_d \in \mathcal{T}} \mathcal{S}_{\mathcal{C}, T_d}(x; \Delta\theta) = \frac{\|\delta_{\mathcal{C}}(x, \Delta\theta)\|_2^2}{kT_{\max}^2} + o(\|\Delta\theta\|^2). \quad (83)$$

For an approximate anchor, one should not replace $J(\pi_T)$ by $J(u_{\mathcal{C}})$. Instead, for a fixed local direction v and $a_v = G_{\mathcal{C}}(x)v$,

$$\mathcal{S}_{\mathcal{C}, T}(x; rv) = r^2 \frac{1}{T^2} \text{Var}_{i \sim \pi_T}(a_{v,i}) + o(r^2). \quad (84)$$

If a_v is not constant on $\mathcal{C}$, then $\text{Var}_{\pi_T}(a_v) > 0$ for every T , since each softmax probability is strictly positive. The function $T \mapsto T^{-2} \text{Var}_{\pi_T}(a_v)$ is continuous on the compact interval $\mathcal{T}$, so

$$\inf_{T \in \mathcal{T}} \frac{1}{T^2} \text{Var}_{\pi_T}(a_v) > 0. \quad (85)$$

This gives positive local sensitivity uniformly over the deployment interval, but the closed-form $1/(kT_{\max}^2)$ coefficient is specific to the exact uniform boundary.

A.11 Full-parameter Fisher objective

At an exact uniform response anchor,

$$F_{\mathcal{C},T}(x) = \frac{1}{kT^2} G_{\mathcal{C}}(x)^\top P_{1^\perp} G_{\mathcal{C}}(x). \quad (86)$$

Taking the trace over all model parameters gives

$$\text{tr } F_{\mathcal{C},T}(x) = \frac{1}{kT^2} \|P_{1^\perp} G_{\mathcal{C}}(x)\|_F^2. \quad (87)$$

Since the leading sensitivity is weakest at $T_{\max}$, Eq. (87) motivates the full-parameter Fisher score in Eq. (25).

B Experiment Information

This appendix complements the main evaluation by answering six questions about LVSF: (1) whether statistical calibration choices improve detection power; (2) how alternative fingerprint constructions affect performance; (3) whether LVSF remains sensitive to small parameter updates; (4) whether the visual modality provides unique coverage while preserving inconspicuous queries; (5) whether benign decoding changes trigger false alarms; and (6) the cost of optimizing the full-parameter Fisher objective.

B.1 Ablations of Fingerprint Calibration

This subsection evaluates whether the two main calibration choices—the top- k support size and the stopping target for the uniform response anchor—materially affect detection power. These experiments test the statistical design of LVSF independently of the broader baseline comparisons.

Effect of the top- k support size. For the top- k ablation, we evaluate LVSF on Qwen2.5-VL-7B-Instruct with $k \in \{2, 3, 4, 5\}$. For each value of k , we generate 100 fingerprints. Tampered models are constructed by LoRA fine-tuning Qwen2.5-VL-7B-Instruct with 80, 160, 240, and 320 training samples, respectively. Detection is performed at decoding temperature $T = 0.8$.

As shown in Table 3, increasing the support size consistently improves tamper detection across all

# Samples	$k = 2$	$k = 3$	$k = 4$	$k = 5$
80	0.83	0.90	0.95	0.98
160	0.82	0.93	0.97	1.00
240	0.83	0.94	0.99	1.00
320	0.87	0.94	1.00	1.00

Table 3: Final two-tier detection rates under different top- k settings on Qwen2.5-VL-7B-Instruct.

four fine-tuning scales. The average detection rate increases from 0.84 at $k = 2$ to 0.93, 0.98, and 1.00 at $k = 3, 4, 5$, respectively. In particular, $k = 5$ achieves detection rates of 0.98–1.00 across all tested update strengths, whereas $k = 2$ remains between 0.82 and 0.87. This result indicates that a larger candidate support provides richer distributional evidence for detecting relative first-token shifts. We therefore use $k = 5$ as the default setting, which provides the strongest detection performance among the evaluated support sizes while retaining a compact categorical support for verification.

Effect of the anchor stopping target. For the anchor stopping-target analysis, we vary the coordinate-wise acceptance tolerance $\Delta_{n_{\text{gen}}}$ used to determine whether the optimized fingerprint has reached the uniform top- k response anchor. Following the chi-square-aligned stopping criterion in Eq. (31), we use $\Delta_{n_{\text{gen}}} \in \{0.0503, 0.0436, 0.0356, 0.0308, 0.0276\}$, corresponding to $n_{\text{gen}} \in \{150, 200, 300, 400, 500\}$, respectively, with $k = 5$ and $\alpha = 0.05$. Smaller values of $\Delta_{n_{\text{gen}}}$ require the conditional top- k distribution to lie closer to the uniform anchor and therefore impose a stricter stopping criterion. Detection is evaluated on Qwen2.5-VL-7B-Instruct under 8-bit quantization, with decoding temperature fixed at $T = 0.9$.

Figure 2 shows a clear trade-off between anchor calibration quality and enrollment cost. Tightening the stopping tolerance from $\Delta_{n_{\text{gen}}} = 0.0503$ to $\Delta_{n_{\text{gen}}} = 0.0276$ increases the average generation time only moderately, from approximately 650 to 700 seconds, while improving the final two-tier detection rate from roughly 90% to nearly 100%. The improvement becomes smaller as the anchor approaches uniformity, indicating diminishing returns from increasingly strict calibration. These results confirm that tighter anchor calibration improves detection power, while the stopping criterion provides a practical mechanism for avoiding unnecessary optimization once the remaining distributional bias is sufficiently small.

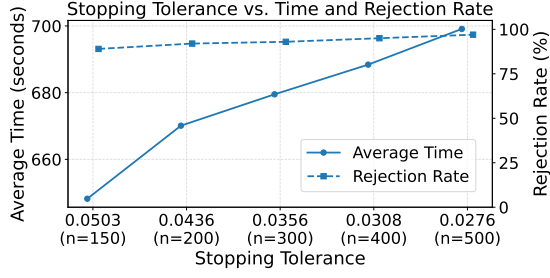


Figure 2: Stopping tolerance versus generation time and final two-tier detection rate.

Method	Intact, $T = 0.1$	Intact, $T = 0.5$	Intact, $T = 0.9$	INT8, $T = 0.1$	INT8, $T = 0.5$	INT8, $T = 0.9$
LVSF	0.00	0.00	0.00	1.00	1.00	0.99
RNI	0.00	0.88	1.00	0.02	0.90	1.00
PixelOpt	0.01	0.00	0.00	1.00	0.95	0.70
AdvPert	0.02	0.01	0.00	1.00	0.88	0.61
Text-FP	0.03	0.04	0.01	0.28	0.64	0.42
BoundaryOpt	0.01	0.00	0.00	0.98	0.89	0.81

Table 4: Comparison with alternative fingerprint construction methods using final two-tier verification decisions.

B.2 Alternative Fingerprint Constructions

We further evaluate five alternative fingerprint constructions on Qwen2.5-VL-7B-Instruct and its INT8-quantized checkpoint. All methods use the same single-fingerprint verification protocol, with enrollment at $T = 0.1$ and evaluation at $T \in \{0.1, 0.5, 0.9\}$.

The evaluated alternatives include random natural image–query references (RNI), which use unoptimized natural responses as fingerprints; direct pixel optimization (PixelOpt), which applies the LVSF objective directly in RGB space instead of the VAE latent space; bounded adversarial perturbations (AdvPert), which restrict image updates to an L_∞ ball with $\epsilon = 8/255$; text-only optimization with a fixed image (Text-FP), which removes the continuous visual optimization channel; and boundary-only optimization (BoundaryOpt), which retains latent-space optimization and the SLM calibration terms but removes the TDC and Fisher-sensitivity objective.

As shown in Table 4, LVSF achieves the most consistent separation between unmodified and tampered models, with no observed false positives on the unmodified model and 0.99–1.00 detection under INT8 tampering across temperatures. RNI becomes highly unstable as temperature changes, while Text-FP provides substantially weaker tamper detection. PixelOpt, AdvPert, and BoundaryOpt perform well at low temperature but degrade at $T = 0.9$, reaching detection rates of only 0.70, 0.61, and 0.81, respectively. These results show that the full LVSF construction provides more stable and robust detection across decoding temperatures than the alternative fingerprint constructions.

Fine-tuning dataset	$T = 0.1$	$T = 0.3$	$T = 0.5$	$T = 0.7$	$T = 0.9$	Average
CodeAlpaca	1.00	1.00	1.00	1.00	0.99	1.00
GSM8K	1.00	1.00	0.98	0.95	0.93	0.97
ScienceQA	1.00	1.00	1.00	0.99	0.99	1.00

Table 5: Final two-tier detection rates across diverse fine-tuning datasets.

B.3 Fine-Tuning Diversity and Sensitivity to Subtle Model Updates

We further examine LVSF from two complementary aspects: whether it generalizes across diverse fine-tuning directions and whether it can detect subtle model updates rather than only large parameter changes. Let θ_0 denote the source-model parameters and θ_m the parameters of a modified checkpoint. Their parameter-space difference, or task vector, is defined as

$$\Delta\theta = \theta_m - \theta_0. \quad (88)$$

We quantify modification strength using the task-vector norm, normalized displacement, and cosine similarity. The normalized displacement is defined as

$$D_{\text{norm}} = \frac{\|\Delta\theta\|_2}{\|\theta_0\|_2}, \quad (89)$$

and the cosine similarity is

$$\text{CosSim}(\theta_0, \theta_m) = \frac{\theta_0^\top \theta_m}{\|\theta_0\|_2 \|\theta_m\|_2}. \quad (90)$$

The first two metrics measure the absolute and relative magnitude of the parameter update, while cosine similarity measures global directional alignment (Ilharco et al., 2023).

Diverse fine-tuning directions. We first examine whether LVSF generalizes beyond a particular fine-tuning dataset or task. We LoRA fine-tune Qwen2.5-VL-7B-Instruct on CodeAlpaca¹, GSM8K², and ScienceQA³, covering code-oriented instruction following, mathematical reasoning, and scientific multimodal question answering, respectively. The same LoRA configuration is used across the three settings, with only the adaptation dataset and task varied.

As shown in Table 5, LVSF maintains high detection rates across all three fine-tuning directions, with average detection rates between 0.97 and 1.00. Even for GSM8K, the most challenging setting, the detection rate remains 0.93 at $T = 0.9$. These results show that LVSF is not specific to a particular fine-tuning dataset or task domain.

Controlled low-strength fine-tuning. We next reuse the one-epoch LoRA sweep from Table 3.

¹github.com/sahil280114/codealpaca

²github.com/openai/grade-school-math

³github.com/lupantech/ScienceQA

Fine-tuning scale	$\ \Delta\theta\ _2$	D_{norm}	Cosine similarity	Detection rate
80 samples	3.322	0.153%	0.999998	0.98
160 samples	4.583	0.217%	0.999997	1.00
240 samples	5.397	0.250%	0.999996	1.00
320 samples	5.963	0.275%	0.999996	1.00

Table 6: Parameter-space displacement and final two-tier detection rate for the controlled LoRA fine-tuning sweep.

Model modification method	$\ \Delta\theta\ _2$	D_{norm}	Cosine similarity	Detection rate
DPO	16.571	0.765%	0.999971	1.000
8-bit	18.114	0.837%	0.999967	0.997
LoRA	24.650	1.138%	0.999935	1.000
Poisoning (LoRA)	24.920	1.151%	0.999934	0.997
Freeze	117.982	5.449%	0.998515	0.993

Table 7: Parameter-space modification strength and final two-tier detection rate across tampering types.

The detection rates in Table 6 correspond to $k = 5$ and $T = 0.8$, while the additional columns quantify the resulting parameter displacement.

As shown in Table 6, LVSF remains effective even under very small parameter updates. The smallest update has a normalized displacement of only 0.153% and a cosine similarity of 0.999998, yet still yields a detection rate of 0.98. Detection reaches 1.00 for the remaining update scales, showing that LVSF remains sensitive to low-strength fine-tuning with only minor parameter displacement.

Displacement across tampering types. We further extend the analysis to the main model modifications. Comparable parameters are flattened and concatenated in a fixed order before computing the displacement metrics. For LoRA checkpoints, we merge each adapter update into the corresponding base weight before computing the normalized parameter displacement. For quantized checkpoints, we first dequantize the INT8 weights into the source-model parameter space, then compare matched parameters in float32 to compute their relative displacement. The detection rates in Table 7 are averaged over $T \in \{0.1, 0.5, 0.9\}$.

Table 7 further shows that LVSF maintains high detection across a broad range of modification strengths. Although the normalized displacement varies from 0.765% to 5.449%, the average detection rate remains above 0.99 for all evaluated tampering types.

Together with the controlled LoRA sweep, these results show that LVSF is effective not only across diverse fine-tuning directions, but also for subtle parameter updates that leave the modified checkpoint highly aligned with the source model.

Method	$T = 0.1$	$T = 0.3$	$T = 0.5$	$T = 0.7$	$T = 0.9$	Average
RESF	0.00	0.00	0.03	0.00	0.01	0.008
RESF-image	0.02	0.02	0.05	0.08	0.20	0.074
LVSF	1.00	0.94	0.87	0.81	0.75	0.874

Table 8: Final two-tier detection rates for vision-encoder-only INT8 tampering across decoding temperatures.

B.4 Why Visual Fingerprints? Coverage, Fidelity, and Detectability

The visual carrier is motivated by two distinct goals: it should probe model components that text-only fingerprints may not exercise, and it should preserve the appearance of an ordinary VLM request. We therefore evaluate these properties separately: vision-pathway coverage, perceptual fidelity, and detectability by an adaptive provider.

Vision-pathway coverage. To isolate whether visual fingerprints provide unique coverage, we apply INT8 post-training quantization only to the vision encoder of Qwen2.5-VL-7B-Instruct, while leaving the multimodal projector and language backbone unchanged. This controlled modification changes MMBench performance (Liu et al., 2024) by only 0.06%. We compare LVSF with text-only RESF and the RESF-image variant. Each verification decision uses $n_{\text{det}} = 100$ first-token samples and is evaluated at $T \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

As shown in Table 8, text-only RESF provides almost no detection signal in this setting, with an average detection rate of 0.008. Applying the RESF construction to the image input yields only a small improvement to 0.074. In contrast, LVSF achieves an average detection rate of 0.874 even though the language backbone is unchanged. This controlled intervention shows that visual-latent fingerprint generation, together with the LVSF sensitivity objective, exposes modifications confined to the vision pathway that are largely invisible to text-only fingerprints.

Visual fidelity. To evaluate whether the optimized fingerprints preserve the visual and semantic content of the source images, we compare all 100 source-fingerprint pairs using the structural similarity index measure (SSIM) (Wang et al., 2004), peak signal-to-noise ratio (PSNR), and DeepSeek-V4-Flash (DeepSeek-AI, 2026) as a vision-language model (VLM)-based semantic consistency judge. The optimized fingerprints achieve an average SSIM of 0.946, an average PSNR of 33.67 decibels (dB), and an average VLM-judge consistency score of 4.97/5. For semantic evaluation, the judge compares each source-fingerprint pair using a fixed

Operating Point	Benign FPR	Fingerprint Detection Rate ($\uparrow$)
P90 threshold	10%	10%
P95 threshold	5%	7%
P99 threshold	1%	0%

Table 9: Deep- k NN performance at different benign false-positive-rate operating points.

prompt and assigns a score from 1 (semantically different) to 5 (semantically identical). These results indicate that the optimized fingerprints remain highly similar to their source images at both the visual and semantic levels, suggesting that the improved tamper sensitivity does not rely on substantial changes to the original image content.

Example of text-optimized fingerprint. To illustrate the difference between visual and text-side optimization, we also show a representative text fingerprint produced by the RESF-style baseline. The optimized text is effective as a fingerprint carrier but is visibly unnatural and semantically incoherent, making its audit purpose easier to notice than an ordinary image–query pair:

```
LoadarguessectionBuckleyColts
casualtyknowsRavensCroatianinclude-
*inHinsymmetricrangedjedentb
eating_____for(((
```

In contrast, LVSF keeps the textual query natural and places the fingerprint signal in the optimized visual input, preserving the appearance of a standard VLM interaction.

Adaptive query detectability. Visual fidelity does not necessarily imply that audit queries are difficult for an adaptive provider to identify. We therefore evaluate a representative vision-feature-based detector using Deep- k NN (Sun et al., 2022). The detector operates on frozen vision-encoder features and distinguishes LVSF fingerprints from benign images. Its reference bank contains 1.28M ImageNet training images (Russakovsky et al., 2015), while 5,000 validation images are used to set thresholds at different benign false-positive rates (FPRs). The 100 evaluated fingerprints are excluded from both reference-bank construction and threshold selection.

Deep- k NN achieves an area under the receiver operating characteristic curve (AUROC) of only 0.593, indicating weak separation between LVSF fingerprints and benign images. As shown in Table 9, the fingerprint detection rate is only 7% at a 5% benign FPR and drops to 0% at a 1% benign FPR. These results indicate that the tested vision-feature-based detector has limited ability to identify LVSF fingerprints at low-FPR operating points.

Decoding setting	Value	Unmodified			8-bit		
		$T = 0.1$	$T = 0.5$	$T = 0.9$	$T = 0.1$	$T = 0.5$	$T = 0.9$
top- p	0.85	0.01	0.01	0.00	1.00	1.00	1.00
	0.90	0.00	0.00	0.00	1.00	1.00	0.99
	1.00	0.00	0.00	0.00	1.00	1.00	1.00
top- k	10	0.00	0.00	0.00	1.00	1.00	1.00
	20	0.00	0.00	0.00	1.00	1.00	1.00
	50	0.00	0.00	0.00	1.00	1.00	1.00
Repetition penalty	1.05	0.00	0.00	0.00	1.00	1.00	0.98
	1.10	0.00	0.00	0.00	1.00	1.00	0.98
	1.20	0.00	0.00	0.00	1.00	1.00	0.97

Table 10: Final two-tier detection performance under benign decoding variations.

B.5 Robustness to Benign Decoding Variations

To distinguish model tampering from benign inference-time changes, we evaluate Qwen2.5-VL-7B-Instruct and its 8-bit quantized variant under variations in temperature, top- p , top- k , and repetition penalty. The model input is kept fixed so that the experiment isolates decoder-side changes. For the unmodified model, rejection is a false positive; for the 8-bit model, rejection is detection.

As shown in Table 10, the observed false-positive rate for the unmodified model remains at or below 0.01 across all tested decoding settings, while the detection rate for the 8-bit model remains at least 0.97. These results show that LVSF is robust to the evaluated benign decoding variations while preserving strong sensitivity to parameter tampering.

B.6 Stochastic Fisher Estimation and Enrollment Cost

Computing the full-parameter Fisher-sensitivity objective at every optimization step is expensive for large VLMs. LVSF therefore uses stochastic parameter-block sampling to reduce the enrollment cost.

Let the model parameters be partitioned into B disjoint blocks, $\theta = (\theta_1, \dots, \theta_B)$, covering the vision encoder, visual projector, language-model layers, and other modules. For a fixed candidate set $\mathcal{C}$ of k tokens, define

$$P_{1^\perp} = I - \frac{1}{k} \mathbf{1}\mathbf{1}^\top. \quad (91)$$

The full-parameter sensitivity objective is

$$S(x; T_{\max}) = \frac{1}{kT_{\max}^2} \sum_{b=1}^B \left\| \frac{\partial(P_{1^\perp} \ell_{\theta, \mathcal{C}}(x))}{\partial \theta_b} \right\|_F^2. \quad (92)$$

At each optimization step, we uniformly sample $m = 2$ blocks without replacement. For the sampled block set $\mathcal{B}_m$, we estimate the full objective



Figure 3: Additional originals (top) and LVSF fingerprints (bottom).

by

$$\hat{S}(x; T_{\max}) = \frac{B}{mkT_{\max}^2} \sum_{b \in \mathcal{B}_m} \left\| \frac{\partial(P_{\mathbf{1}^\perp} \ell_{\theta, \mathcal{C}}(x))}{\partial \theta_b} \right\|_F^2. \quad (93)$$

Since each block is selected with probability m/B ,

$$\mathbb{E}_{\mathcal{B}_m} [\hat{S}(x; T_{\max})] = S(x; T_{\max}), \quad (94)$$

so the estimator is unbiased. The Jacobian is computed exactly for each sampled block, while resampling across optimization steps allows all parameter blocks to contribute in expectation.

For Qwen2.5-VL-7B-Instruct, generating one fingerprint takes approximately 25 minutes with 75 GB peak GPU memory on a single NVIDIA A100 80GB GPU. Thus, stochastic block sampling makes full-parameter sensitivity optimization feasible without evaluating all parameter blocks at every step. Although fingerprint generation remains computationally substantial, it is a one-time offline enrollment cost that can be amortized over repeated audits by reusing a compact fingerprint set, leaving recurring verification lightweight.

C Potential Risks

LVSF is intended for legitimate model-integrity auditing, but its sensitivity-based fingerprint construction may have dual-use implications. For example, similar techniques could potentially be used to probe behavioral differences between model versions or to characterize model changes without the model provider’s consent. We therefore recommend that LVSF be used only for authorized model auditing, integrity verification, and related research purposes.

D Information About Use Of AI Assistants

AI assistants were used during research, coding, and manuscript preparation. For writing, they assisted with drafting and language refinement based on ideas and content provided by the authors. For coding, they supported implementation and debugging. All AI-assisted content incorporated into the paper or code was independently reviewed, verified, and finalized by the authors, who take full responsibility for the resulting work.