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

Anonymous ACL submission

Abstract

Cloud-hosted vision-language models (VLMs) are increasingly deployed as black-box APIs, but the hosted model may be silently fine-tuned, poisoned, compressed, or replaced after deployment. Existing fingerprinting methods are designed primarily for text-only large language models (LLMs). Their fingerprints therefore inherit the discreteness bottleneck of text: prompt optimization is difficult, the resulting sensitivity is limited by a discontinuous input space, and optimized prompts can become conspicuous. Modern VLMs expose a different opportunity: the visual input is continuous and can be optimized in latent space while the textual query remains natural. We propose Latent Visual Sensitive Fingerprinting (LVSF), the first fingerprinting framework specifically designed for VLM tamper detection by exploiting this visual continuity. We formalize black-box first-token tamper detection and show that detection sensitivity is controlled by two factors: a *top-k distribution term*, which determines how observable relative logit shifts are, and a *Fisher sensitivity term*, which determines how strongly a model perturbation changes the recorded top-k logits. This decomposition motivates a single weighted optimization objective: LVSF jointly calibrates a uniform top-k response anchor and maximizes full-parameter Fisher sensitivity of the image-query fingerprint. Verification uses natural image–query pairs, repeated black-box API calls, and a chi-square test over first-token frequencies. Experiments on four VLMs under fine-tuning, poisoning, compression, and replacement show that LVSF substantially outperforms RESF-based baselines while preserving image naturalness and semantic consistency.

1 Introduction

Vision-language models (VLMs) have become an increasingly important form of large-model deployment (Ghosh et al., 2024). A model owner may train or license a VLM and then rely on a cloud service provider to host it, while downstream users interact only through a black-box API. This deployment pattern creates a model-integrity problem (Bai et al., 2024): the hosted model can be replaced by another model, compressed or quantized for cost reduction (Dettmers et al., 2022; Wang et al., 2024), fine-tuned without authorization through lightweight adaptation 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 a way that changes its behavior, while still appearing normal on ordinary user queries (Zong et al., 2024). Since the verifier cannot inspect the model interior, practical integrity verification must therefore rely only on ordinary image-text queries.

Fingerprinting is a natural solution for black-box integrity verification: the owner registers sensitive inputs before deployment, and the verifier later tests whether a suspect API preserves their model-specific response signatures. For deterministic image classifiers, such fingerprints often target decision-boundary samples and detect top-1 label changes (He et al., 2019; Wang et al., 2023). This paradigm does not directly transfer to generative LLMs and VLMs, where stochastic decoding makes exact output matching unreliable. Recent LLM fingerprinting methods instead use distributional evidence: ESF constructs sensitive prompts with max-coverage token sets (Bai et al., 2025b), and RESF formulates tamper detection as support and Pearson hypothesis tests under temperature variation (Hu et al., 2025). However, RESF-style fingerprints are text-based high-perplexity prompts, making them distinguishable from normal text queries. Most importantly for VLMs, text-only fingerprints may not exercise the visual encoder, leaving vision-side modifications outside the verification path.

This paper argues that VLM fingerprinting should exploit the visual input channel rather than

treating VLMs as text-only LLMs. By optimizing the image latent while keeping the textual query natural, LVSF creates ordinary image-question fingerprints with sensitive first-token response distributions. The resulting image–query pairs remain natural VLM samples, so audit queries are not singled out from normal user queries by their surface form. This visual degree of freedom provides a smooth optimization space, probes the multimodal inference path, and preserves inconspicuous queries.

We propose **Latent Visual Sensitive Fingerprinting (LVSF)**, a single-stage framework that optimizes image latents through a pretrained VAE (Kingma and Welling, 2014). LVSF jointly forms a *uniform top-k response anchor* and improves sensitivity by combining three terms in one weighted objective: top-k distribution calibration (TDC), support-leakage margin (SLM), and full-parameter Fisher sensitivity (FS) maximization. The first two terms give the intact VLM a stable uniform null distribution, while the Fisher term amplifies observable frequency shifts caused by parameter changes.

LVSF is guided by a local sensitivity analysis of the conditional top-k first-token distribution. For a fingerprint $x = (I, q)$ and a small parameter perturbation $\Delta\theta$, the leading Pearson sensitivity is governed by two factors: the categorical covariance $J(\pi^{C}_{\theta,T})$, which determines how visible relative top-k logit shifts are, and the top-k logit Jacobian $G_C(x)$, which determines how strongly the input exposes a parameter change. The analysis shows that the covariance term is maximized, under unknown isotropic relative perturbations, by a uniform top-k distribution, motivating the uniform response anchor. The Jacobian term motivates maximizing Fisher sensitivity with respect to all model parameters. Thus, LVSF directly optimizes the two quantities that control black-box tamper sensitivity, rather than merely adapting LLM fingerprinting to image inputs.

Our key contributions are as follows.

- **First visual-continuity-based VLM fingerprinting framework.** To the best of our knowledge, LVSF is the first black-box tamper-detection fingerprint designed specifically for VLMs by exploiting the continuous visual modality. Unlike ESF/RESF-style text fingerprints (Bai et al., 2025b; Hu et al., 2025), LVSF uses natural image-question pairs that resemble normal VLM requests while activating the multimodal inference path.
- **A theory-driven sensitivity decomposition.** We formalize stochastic first-token tamper detection and derive that sensitivity is governed by the interaction between a top-k distribution term $J(\pi)$ and a Fisher sensitivity term $G_C^\top J(\pi) G_C$. This explains why, under unknown isotropic relative logit perturbations, a uniform top-k anchor maximizes average distributional sensitivity, and why increasing full-parameter Fisher sensitivity improves tamper detectability.
- **A single-stage latent optimization method.** LVSF constructs a uniform top-k response anchor while maximizing full-parameter Fisher sensitivity through one weighted objective that combines top-k distribution calibration, support-leakage margin regularization, and Fisher sensitivity maximization.
- **Comprehensive empirical evidence.** Across four VLMs and multiple tampering types, one LVSF fingerprint already achieves strong detection performance and remains robust across decoding temperatures. LVSF substantially outperforms RESF-based baselines (Hu et al., 2025) while preserving visual quality and semantic consistency.

2 Related Work

Black-box model tamper detection. Model tamper detection verifies whether a deployed model remains behaviorally consistent with its protected version under black-box access, which is crucial in MLaaS settings where parameters, gradients, and intermediate states are unavailable. Existing integrity-verification methods usually construct fingerprint samples that are sensitive to model modifications. VerIDeep introduces sensitive-sample fingerprinting for detecting tampered cloud-hosted DNNs (He et al., 2018), while PublicCheck generates smooth decision-boundary samples to reveal model replacement, compression, or other integrity breaches with few API queries (Wang et al., 2023). However, these methods mainly target discriminative classifiers, where deterministic top-1 label changes serve as verification signals. They are therefore not directly applicable to generative LLMs and VLMs, whose token responses are stochastic even for intact models.

LLM fingerprinting. Recent work on LLM fingerprinting extends model verification from deterministic label matching to stochastic text generation. ESF optimizes sensitivity at selected detection to-

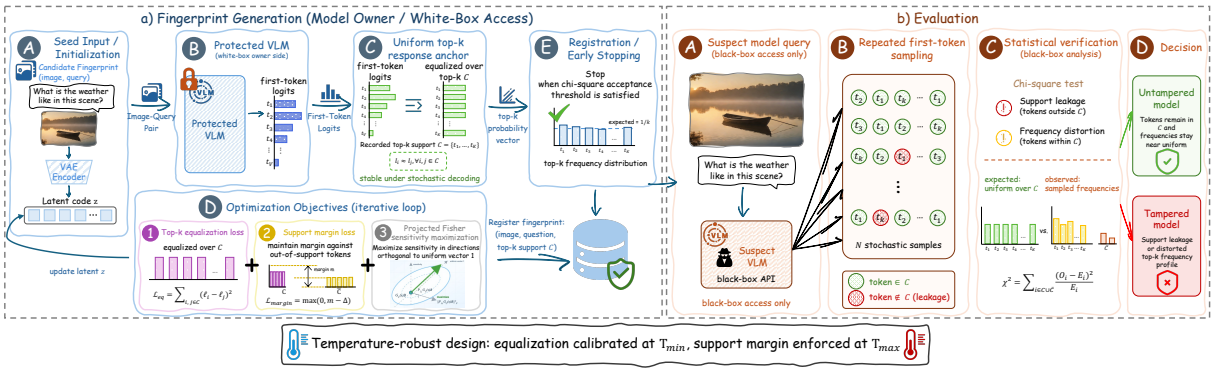

Figure 1: LVSF overview: one weighted objective combines top-k distribution calibration (TDC), support-leakage margin (SLM), and full-parameter Fisher sensitivity (FS).

ken positions, uses Randomness-Set Consistency Checking (RSCC) for output stochasticity, and selects diverse fingerprints with a Max Coverage Strategy (MCS) (Bai et al., 2025b). RESF further frames LLM tamper detection as a hypothesis test with support and Pearson distributional checks under temperature variation (Hu et al., 2025). These methods move fingerprinting from deterministic output matching to distribution-aware verification, but they remain text-based. In particular, RESF fingerprints are high-perplexity prompts distinct from normal text queries, and text-only verification leaves the visual pathway untested: a perturbation confined to the vision encoder may not affect a text-only fingerprint at all.

VLM fingerprinting through continuous visual inputs. VLMs introduce a new fingerprint carrier: the image (Liu et al., 2023). Unlike text tokens, an image or its latent representation is continuous, so it can be optimized by gradient-based methods while the accompanying text remains fluent and task-like. A useful VLM fingerprint should be sensitive to model tampering, statistically stable under stochastic decoding, and visually inconspicuous. LVSF addresses these requirements with normal image-query API calls, a stable top-k support with a uniform conditional null, and full-parameter Fisher sensitivity in a single weighted objective.

3 Threat Model

We consider integrity verification for a VLM deployed behind a cloud API. Following prior fingerprinting works (He et al., 2019; Wang et al., 2023; Bai et al., 2025b; Hu et al., 2025), LVSF adopts a white-box generation and black-box detection setting. Before deployment, the model owner has white-box access to the protected model and uses it to generate and register private fingerprints. A fingerprint consists of an optimized but natural-looking image $I^\star$, a natural-language query q, a recorded top-k token support $C^\star$, and reference testing parameters. During auditing, the verifier has only black-box API access to the suspect model: it repeatedly submits the ordinary VLM query $(I^\star, q)$, records first-token samples, and performs the statistical test. Because $(I^\star, q)$ has the same interface and natural surface form as a normal image-text request, tamper-detection queries are not directly singled out from ordinary API traffic. No logits, hidden states, gradients, model parameters, sampling seeds, or VAE access are assumed at verification time.

The adversary may be an untrusted service provider or an external party that modifies the hosted model while preserving normal API functionality. We consider practical tampering types including fine-tuning, safety or preference alignment, poisoning, post-training quantization or compression, and model replacement. The adversary may know the LVSF algorithm and verification protocol, but does not know the active fingerprints or their registered reference statistics. As in prior fingerprinting work, we do not claim protection against a fully adaptive adversary who has obtained the private fingerprints and explicitly optimizes the suspect model to reproduce their first-token distributions. Under non-leaked fingerprints, LVSF provides a practical MLaaS-oriented audit mechanism without modifying the protected model.

4 Sensitivity Theory for VLM Fingerprinting

This section derives the design principle behind LVSF. The key message is that black-box first-token tamper sensitivity is controlled by two factors: a top-k distribution term and a Fisher sensitiv-

ity term. LVSF optimizes both factors jointly in a single weighted latent-space objective.

4.1 Setup

Let $x = (I, q)$ be an image-query fingerprint. For model parameter θ, let $\ell_\theta(x) \in \mathbb{R}^{|\mathcal{V}|}$ be the first-token logit vector over 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

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

Let u_C be the uniform distribution on C, and define

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

The matrix $J(\pi)$ is the covariance matrix of a one-hot categorical draw with probabilities π, and it is also the derivative geometry 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

Tamper detection can be written as the hypothesis test

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

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 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*

$$\begin{aligned} &D_{\chi^2}\big(\pi^C_{\theta_0+\Delta\theta,T}(\cdot \mid x), \pi^C_{\theta_0,T}(\cdot \mid x)\big) \\ &= \Delta\theta^\top F_{C,T}(x)\Delta\theta + o(\|\Delta\theta\|^2), \end{aligned} \tag{4}$$

where

$$F_{C,T}(x) = \frac{1}{T^2} G_C(x)^\top J\big(\pi^C_{\theta_0,T}(\cdot \mid x)\big) G_C(x). \tag{5}$$

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,

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

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 The top-k distribution term: why uniform anchors

Only relative top-k logit changes are visible, because $J(\pi)\mathbf{1} = 0$: 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 finite second moment $\Sigma = \mathbb{E}[AA^\top]$. For $\pi = \mathrm{softmax}(z/T)$, define*

$$\bar{s}_T(\pi; Q) = \frac{1}{T^2}\,\mathrm{tr}\,\big(J(\pi)\Sigma\big). \tag{7}$$

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

$$\Sigma = \sigma^2 P_{1^\perp}, \qquad \sigma^2 > 0. \tag{8}$$

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

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

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

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

$$\bar{s}_T(\pi; Q) = \frac{\sigma^2}{T^2}\,\mathrm{tr}\,J(\pi) = \frac{\sigma^2}{T^2}\big(1 - \|\pi\|_2^2\big), \tag{10}$$

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*

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

It is γ-stable if, in addition,

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

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

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

and the local sensitivity becomes

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

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,

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

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

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

The KL deviation $D_{\mathrm{KL}}(\pi_T \| u_C)$ 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

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

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 the generation rule used by LVSF: 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.

5 Latent Visual Sensitive Fingerprinting

5.1 Overview

Given an initial natural image I_0 and a natural-language query q, LVSF optimizes 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 the token identities are selected once from the protected model's top-k first-token logits on the initial pair (I_0, q). This support is then kept fixed while LVSF optimizes the image so that these selected tokens have nearly equal logits. After optimization, the final support $C^\star$, image $I^\star$, and query q are registered and kept private.

Figure 1 summarizes the pipeline. The important distinction from text-only LLM fingerprinting is that the optimized variable is a continuous visual latent. The text query can remain natural, and the verification query activates both visual and language components of the VLM.

5.2 Single-stage fingerprint optimization

LVSF optimizes one weighted objective with three goals: calibrating the fixed top-k support toward a uniform null, keeping that support valid against tail-token leakage, and increasing Fisher sensitivity with respect to all model parameters. We name the first loss term according to its theoretical role: it is not merely a numerical logit-equalization penalty, but a *top-k distribution calibration* loss that drives the distribution term $J(\pi)$ toward its average-sensitivity maximizer under isotropic unknown relative perturbations. Let C be the fixed

support selected from (I_0, q) and let

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

The temperature-normalized centered logit $\tilde{d}_i$ matches the calibration rule in Section 4: uniformity is hardest to certify at $T_{\min}$. We therefore use the **top-k distribution calibration (TDC) loss**

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

where

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

This term is a Huber relaxation of $\pi^C_{\theta,T_{\min}}(\cdot \mid x) \to u_C$: it precisely aligns logits near the anchor while avoiding unstable gradients when the initial gaps are large. Its name is chosen to align with the top-k distribution term in the sensitivity decomposition.

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 $\ell_{\mathrm{tail}}(z) = \max_{r\notin C}\ell_r(z)$ and let $s_{\min}$ be a differentiable soft-min approximation of the weakest in-support logit:

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

Given a desired gap $\gamma_{\sup}$ calibrated for $T_{\max}$, define

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

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

$$\gamma_{\sup} \ge T_{\max}\log\frac{M(1-\eta_{\sup})}{k\eta_{\sup}}, \tag{23}$$

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**

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

This naming separates the two validity duties: $\mathcal{L}_{\mathrm{TDC}}$ calibrates the distribution term, while $\mathcal{L}_{\mathrm{SLM}}$ enforces support validity so that the registered top-k conditional test remains meaningful. Their weighted sum $\mathcal{L}_{\mathrm{VAC}}$ is therefore a validity loss, not the full LVSF objective.

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

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

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. The projection removes common-logit shifts that cannot be observed through the conditional top-k softmax. Combining validity and sensitivity gives the single LVSF objective

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

The weights in $\mathcal{L}_{\mathrm{VAC}}$ and λ_{FS} control the trade-off between preserving the statistical null and increasing parameter sensitivity. The final registered fingerprint is the best optimized iterate that satisfies the uniformity criterion and the support-leakage margin condition for $C^\star$.

Chi-square-aligned stopping. In practice, optimization cannot drive the top-k logits to be exactly equal, so the protected model's conditional distribution $\pi^C(z)$ never reaches the ideal uniform null u_C exactly. We accept an optimized fingerprint once the residual deviation from u_C would be statistically negligible for the verifier's chi-square test. Since chi-square deviations scale with the number of samples, using a generation-side calibration budget n_{gen} substantially larger than the detection query budget n_{det} makes the remaining protected-model bias much smaller than the verifier's sampling noise. With significance level α, we use the coordinate-wise tolerance

$$\Delta_{n_{\mathrm{gen}}} = \sqrt{\frac{\chi^2_{k-1,1-\alpha}}{(k-1)\,n_{\mathrm{gen}}\,k^2}}. \tag{27}$$

The latent is accepted as a valid anchor when

$$\max_{i\in C}\left|\pi_i^C(z) - \frac{1}{k}\right| \le \Delta_{n_{\mathrm{gen}}}. \tag{28}$$

Because $n_{\mathrm{gen}} \gg n_{\mathrm{det}}$, 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. The extra $(k-1)$ factor makes this a conservative subset of the coordinate-wise sufficient bound for Pearson acceptance, serving as a safety margin rather than a tight approximation to the boundary.

5.3 Black-box verification

For each registered fingerprint $(I^\star, q, C^\star)$, the verifier queries the suspect model n_{det} times and

records the first generated token. Verification remains a hypothesis test against the protected-model null: tokens outside $C^\star$ should not appear under the registered support, while tokens inside $C^\star$ should follow the uniform conditional distribution. Thus, if any sampled token falls outside $C^\star$, the verifier treats it as support leakage and immediately rejects the null hypothesis without requiring more queries. Otherwise, it applies a Pearson chi-square test to empirical counts over $C^\star$ against the uniform distribution. Let N_i be the number of times token $i \in C^\star$ appears. The test statistic is

$$X^2 = \sum_{i \in C^\star} \frac{(N_i - n_{\text{det}}/k)^2}{n_{\text{det}}/k}. \tag{29}$$

The suspect model is declared tampered if $X^2 > \chi^2_{k-1,1-\alpha}$; otherwise, the verifier reports no evidence of tampering. Verification uses only ordinary API calls and sampled first tokens.

6 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(Team, 2025a), and Kimi-VL-A3B-Instruct(Team, 2025b). These models cover different parameter scales and architectural families which allows us to evaluate whether LVSF generalizes across heterogeneous VLM designs rather than being tailored to a single architecture.

We consider four common tampering scenarios: model fine-tuning, model poisoning, model compression, and model replacement. For fine-tuning, we use the Alpaca dataset (Taori et al., 2023) for supervised finetune scenario and the SPA-VL dataset(Zhang et al., 2025) for multimodal safety alignment scenario. For poisoning, we follow the setting of (Xu et al., 2024a), where the training data contains 10% poisoned samples. We apply freeze finetuning (freezing approximately 70% of parameters closest to the input, noted as Freeze) and Low-Rank Adaptation (LoRA) in both settings, and additionally use Direct Preference Optimization (DPO) (Rafailov et al., 2023) for the benign task adaptation, all conducted with the ms-swift framework(Zhao et al., 2025). For model compression, we apply post-training quantization with 4-bit and 8-bit precision by bitsandbytes(Dettmers et al., 2022). All LoRA experiments are fixed with rank of 8 and alpha of 32.

Fingerprint Generation and Detection Settings. We construct 100 image–query pairs and generate 100 corresponding initial images with SDXL (Podell et al., 2024), yielding 100 distinct fingerprints per setting. Unless otherwise specified, LVSF is optimized in the VAE latent space using Adam (Kingma and Ba, 2015) with learning rate 7.5×10^{-4}, support size $k = 5$, and significance level $\alpha = 0.05$. The uniformity acceptance rule uses a generation-side calibration budget $n_{\text{gen}} = 1000$ in Eq. 28, an order of magnitude larger than the detection budget below, so that the residual deviation from the uniform anchor is negligible for the chi-square detector. The single LVSF objective combines anchor calibration, support-margin preservation, and full-parameter Fisher sensitivity with weighted loss terms.

For each detection, the verifier uses one fingerprint, queries the suspect model $n_{\text{det}} = 100$ times, and applies the chi-square test at $\alpha = 0.05$ for a tampered/intact decision. Rather than an ensemble verdict, our 100 fingerprints yield 100 independent single-fingerprint decisions; we report the average *rejection rate*—the fraction rejecting the intact-model null—which is the true detection rate for tampered models and the false-positive rate for intact controls.

7 Results

7.1 Main Results

Table 1 reports the LVSF rejection rate across four VLMs, four tampering categories, and three decoding temperatures. Overall, LVSF achieves consistently high rejection rates under fine-tuning, poisoning, model replacement, and compression, showing that individual optimized fingerprints provide reliable evidence of integrity violations. The diagonal entries in the replacement block correspond to intact-model controls and show that LVSF introduces no false alarms in our evaluation.

A notable advantage of LVSF is its robustness to decoding temperature. Across all tampered settings, the average rejection rate remains at 100% when the temperature is 0.1, 99.8% at 0.5, and 97.7% even at 0.9. Although performance drops slightly at higher temperatures, especially under the milder 8-bit compression setting, the overall rejection rate remains high. These results indicate that the low-temperature fingerprint generation strategy produces fingerprints that retain strong statistical sensitivity under different inference temperatures.

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**
Finetune	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: Rejection rates. Each cell averages 100 decisions; higher is better for tampered rows, and the replacement-block diagonals are intact controls (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.

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
Finetune	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: Rejection rate; 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.

We also find that larger top-k supports consistently yield higher rejection rates, since they provide more degrees of freedom for the chi-square test and a richer uniform response anchor; full numbers and the ablation protocol are reported in Appendix B.1. Tighter stopping targets on the uniform top-k anchor likewise improve the rejection rate at the cost of longer generation time; the full trade-off is reported in Appendix B.2.

7.2 Baseline Comparison

Table 2 presents a comparison between LVSF and RESF-based baselines, with the evaluation protocol detailed in Appendix B.3. RESF is the original text-based fingerprinting method, whose high-perplexity prompts are distinct from normal text queries; RESF-image is our image-input variant that applies the same RESF approach to images, like LVSF. The LVSF variant without the Fisher-sensitivity term already achieves a 99.3% average rejection rate, substantially outperforming RESF (53.1%) and RESF-image (42.1%). Full LVSF increases the average rejection rate to 99.9%, showing that the gain comes from both the uniform response anchor and the full-parameter Fisher sensitivity term in the single weighted objective. LVSF also preserves the visual appearance of the source images (SSIM 0.946, PSNR 33.67 dB, VLM-judge consistency $4.97/5$); the full evaluation, examples, and a contrast with text-optimized fingerprints are reported in Appendix B.4.

8 Conclusion

We presented LVSF, a latent-space visual fingerprinting framework for black-box VLM tamper detection. LVSF constructs uniform top-k response anchors from image–query pairs and verifies model integrity through stochastic first-token statistics, using a single weighted objective that combines top-k distribution calibration, support-leakage margin preservation, and full-parameter Fisher sensitivity. Across four VLMs and tampering scenarios spanning fine-tuning, poisoning, compression, and replacement, LVSF achieves a high rejection rate, is robust to decoding temperature, preserves visual semantics, and substantially outperforms RESF-based baselines, showing the promise of visual response anchors for black-box VLM auditing.

9 Limitations

LVSF is designed for a practical MLaaS integrity-auditing setting in which the model owner can generate fingerprints before deployment and later verify a hosted VLM through ordinary black-box API queries. This assumption is realistic for model owners, licensors, and auditors who control the protected model before outsourcing deployment, although it does mean that fingerprint generation requires white-box access to the original VLM, while post-deployment verification itself only relies on repeated first-token samples from the suspect API. Importantly, this design keeps the deployed model unchanged and turns LVSF into a lightweight audit layer rather than an intrusive defense mechanism.

Like prior fingerprinting-based verification methods, LVSF assumes that the active fingerprints and their registered reference statistics remain private. The adversary may know the algorithm and detection protocol, but is not assumed to have access to the exact active fingerprints and to explicitly optimize the hosted model to mimic their first-token distributions. This threat model matches many practical cloud-hosting scenarios, where the main concern is unauthorized fine-tuning, compression, replacement, poisoning, or local module modification rather than a fully adaptive attack with leaked audit secrets.

The current empirical evaluation covers four open-source instruction-tuned VLMs and a broad set of tampering types, including fine-tuning, poisoning, compression, and model replacement. Nevertheless, future work can further expand the evaluation to larger proprietary VLMs, longer deployment lifecycles, and more diverse API policies. These extensions would help characterize how LVSF behaves under real commercial constraints such as restricted sampling control, hidden decoding parameters, or changing service-side inference configurations.

Finally, LVSF maximizes Fisher sensitivity with respect to all model parameters rather than requiring the defender to choose a target layer or module. This avoids assuming prior knowledge about where tampering occurs, but it can increase optimization cost because the fingerprint must balance full-parameter sensitivity with the validity terms that preserve the uniform null. Future work may explore more efficient sensitivity estimators and multi-fingerprint ensembles.

Overall, these limitations define the intended deployment scope of LVSF rather than diminishing its practical value: it provides a model-owner-friendly, black-box, and visually inconspicuous mechanism for auditing the integrity of cloud-hosted VLMs.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan 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.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3.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*.

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. 2018. Verideep: Verifying integrity of deep neural networks through sensitive-sample fingerprinting. *CoRR*, abs/1808.03277.

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 *International Conference on Learning Representations*.

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.

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*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

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*.

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`.

Gemma Team. 2025a. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Kimi Team. 2025b. Kimi-vl technical report. *CoRR*, abs/2504.07491.

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 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Shuo Wang, Sharif Abuadbba, Sidharth Agarwal, Kristen Moore, Ruoxi Sun, Minhui Xue, Surya Nepal, Seyit Camtepe, and Salil S. Kanhere. 2023. Publicheck: 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.

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 38: 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19867–19878. Computer Vision Foundation / IEEE.

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.

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

$$q_r = \mathrm{softmax}\left(\frac{z + ra}{T}\right). \tag{30}$$

The softmax derivative gives

$$q_r - \pi = \frac{r}{T} J(\pi)a + O(r^2). \tag{31}$$

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

$$D_{\chi^2}(q, \pi) = \sum_i \frac{\delta_i^2}{\pi_i} = \delta^\top \mathrm{diag}(1/\pi)\delta. \tag{32}$$

Substituting Eq. (31),

$$D_{\chi^2}(q_r, \pi) = \frac{r^2}{T^2} a^\top J(\pi)\,\mathrm{diag}(1/\pi) J(\pi) a + o(r^2). \tag{33}$$

It remains to note that

$$J(\pi)\,\mathrm{diag}(1/\pi) J(\pi) = J(\pi). \tag{34}$$

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

$$JD^{-1}J = (D - \pi\pi^\top)D^{-1}(D - \pi\pi^\top) \tag{35}$$

Since $\pi^\top D^{-1}\pi = \sum_i \pi_i = 1$, Eq. (34) follows. Therefore

$$\lim_{r \to 0} \frac{1}{r^2} D_{\chi^2}(q_r, \pi) = \frac{1}{T^2} a^\top J(\pi) a. \tag{36}$$

Finally,

$$a^\top J(\pi) a = \sum_i \pi_i a_i^2 - \left(\sum_i \pi_i a_i\right)^2 \tag{37}$$

This proves Eq. (5).

For parameter perturbations, set

$$G_{\mathcal{C}}(x) = \frac{\partial \ell_{\theta,\mathcal{C}}(x)}{\partial \theta}, \qquad a = G_{\mathcal{C}}(x)\Delta\theta. \tag{38}$$

Assumption 1 gives

$$\ell_{\theta+\Delta\theta,\mathcal{C}}(x) = \ell_{\theta,\mathcal{C}}(x) + G_{\mathcal{C}}(x)\Delta\theta + O(\|\Delta\theta\|^2), \tag{39}$$

and hence

$$\mathcal{S}_{\mathcal{C},T}(x; \Delta\theta) = \Delta\theta^\top \left[\frac{1}{T^2} G_{\mathcal{C}}(x)^\top J(\pi) G_{\mathcal{C}}(x)\right] \Delta\theta + o(\|\Delta\theta\|^2), \tag{40}$$

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

$$n\mathcal{S}_{\mathcal{C},T}(x; \Delta\theta_n) \to v^\top F_{\mathcal{C},T}(x) v, \tag{41}$$

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

A.2 Average-sensitivity optimality of the uniform distribution

By Definition 7,

$$\bar{s}_T(\pi; \mathcal{Q}) = \frac{1}{T^2}\mathbb{E}[A^\top J(\pi) A] = \frac{1}{T^2}\,\mathrm{tr}(J(\pi)\Sigma), \tag{42}$$

where $\Sigma = \mathbb{E}[AA^\top]$. Under Assumption 8, $\Sigma = \sigma^2 P_{\mathbf{1}^\perp}$. Since $J(\pi)\mathbf{1} = 0$, we have $J(\pi)P_{\mathbf{1}^\perp} = J(\pi)$, and therefore

$$\bar{s}_T(\pi; \mathcal{Q}) = \frac{\sigma^2}{T^2}\,\mathrm{tr}\, J(\pi). \tag{43}$$

Now

$$\begin{aligned} \mathrm{tr}\, J(\pi) &= \mathrm{tr}(\mathrm{diag}(\pi)) - \mathrm{tr}(\pi\pi^\top) \\ &= 1 - \|\pi\|_2^2. \end{aligned} \tag{44}$$

For any probability vector π on k points,

$$\|\pi\|_2^2 = \sum_{i=1}^k \pi_i^2 \geq \frac{1}{k}\left(\sum_{i=1}^k \pi_i\right)^2 = \frac{1}{k}, \tag{45}$$

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

$$\bar{s}_T(\pi; \mathcal{Q}) \leq \frac{\sigma^2}{T^2}\left(1 - \frac{1}{k}\right), \tag{46}$$

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 $P_{\mathbf{1}^\perp}$, the objective is $\mathrm{tr}(J(\pi)\Sigma)$ and the uniform distribution need not be globally optimal. For a fixed known vector a, maximizing $\mathrm{Var}_{i\sim\pi}(a_i)$ 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

$$J(u_{\mathcal{C}}) = \mathrm{diag}(u_{\mathcal{C}}) - u_{\mathcal{C}} u_{\mathcal{C}}^\top = \frac{1}{k} P_{\mathbf{1}^\perp}. \tag{47}$$

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

$$\mathcal{S}_{\mathcal{C},T}(x; \Delta\theta) = \frac{\|\delta_{\mathcal{C}}(x, \Delta\theta)\|_2^2}{kT^2} + o(\|\Delta\theta\|^2). \tag{48}$$

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

$$d_i = \ell_i - \frac{1}{k}\sum_{j\in\mathcal{C}} \ell_j, \qquad \sum_{i\in\mathcal{C}} d_i = 0, \tag{49}$$

and write

$$\pi_T(i) = \frac{\exp(d_i/T)}{\sum_{j\in\mathcal{C}} \exp(d_j/T)}. \tag{50}$$

Theorem 3 (Temperature-robust calibration rule)**.** *The KL deviation $D_{\mathrm{KL}}(\pi_T \| u_{\mathcal{C}})$ is non-increasing in T. Hence, for any $\varepsilon > 0$ and for all $T_d \in [T_{\min}, T_{\max}]$*

$$D_{\mathrm{KL}}(\pi_{T_{\min}} \| u_{\mathcal{C}}) \le \varepsilon \implies D_{\mathrm{KL}}(\pi_{T_d} \| u_{\mathcal{C}}) \le \varepsilon. \tag{51}$$

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

$$D_{\mathrm{KL}}(\pi_{T_g} \| u_{\mathcal{C}}) < \varepsilon \text{ but } D_{\mathrm{KL}}(\pi_{T_{\min}} \| u_{\mathcal{C}}) > \varepsilon. \tag{52}$$

Consequently, among single calibration temperatures inside the deployment interval,

$$T_g^\star = T_{\min} \tag{53}$$

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 $\max_{i\in\mathcal{C}} |\ell_i - \bar{\ell}_{\mathcal{C}}| \le \epsilon$, then for any temperature $T > 0$,

$$\|\pi_T^{\mathcal{C}} - u_{\mathcal{C}}\|_\infty \le \frac{e^{2\epsilon/T} - 1}{k}. \tag{54}$$

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

$$\mathbb{P}_{\theta,T}(Y \notin \mathcal{C} \mid x) \le \frac{m e^{-\gamma/T}}{k + m e^{-\gamma/T}}, \tag{55}$$

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

$$\inf_{T_d\in\mathcal{T}} \mathcal{S}_{\mathcal{C},T_d}(x;\Delta\theta) = \frac{\|\delta_{\mathcal{C}}(x,\Delta\theta)\|_2^2}{kT_{\max}^2} + o(\|\Delta\theta\|^2). \tag{56}$$

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

$$\pi_\beta(i) = \frac{\exp(\beta d_i)}{Z(\beta)}, \quad Z(\beta) = \sum_{i\in\mathcal{C}} \exp(\beta d_i). \tag{57}$$

Since $u_i = 1/k$,

$$\begin{aligned} D_{\mathrm{KL}}(\pi_\beta \| u_{\mathcal{C}}) &= \sum_i \pi_\beta(i) \log \frac{\pi_\beta(i)}{1/k} \\ &= \beta \mathbb{E}_{\pi_\beta}[d] - \log Z(\beta) + \log k. \end{aligned} \tag{58}$$

We use the identities $\partial_\beta \log Z(\beta) = \mathbb{E}_{\pi_\beta}[d]$ and $\partial_\beta \mathbb{E}_{\pi_\beta}[d] = \mathrm{Var}_{\pi_\beta}(d)$. Differentiating Eq. (58) gives

$$\partial_\beta D_{\mathrm{KL}}(\pi_\beta \| u_{\mathcal{C}}) = \beta\, \mathrm{Var}_{\pi_\beta}(d) \ge 0. \tag{59}$$

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. (51) 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

$$\phi(s) = D_{\mathrm{KL}}(\mathrm{softmax}(s d_0) \| u_{\mathcal{C}}), \; s \ge 0. \tag{60}$$

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, $\mathrm{softmax}(s d_0)$ converges to a point mass as $s \to \infty$, hence

$$\lim_{s\to\infty} \phi(s) = \log k. \tag{61}$$

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

$$\alpha \in (T_{\min} s_\varepsilon,\, T_g s_\varepsilon). \tag{62}$$

Set $d = \alpha d_0$. Then

$$\frac{\alpha}{T_g} < s_\varepsilon < \frac{\alpha}{T_{\min}}. \tag{63}$$

Strict monotonicity of ϕ yields

$$D_{\mathrm{KL}}(\pi_{T_g}\|u_{\mathcal{C}}) = \phi(\alpha/T_g) < \varepsilon, \quad D_{\mathrm{KL}}(\pi_{T_{\min}}\|u_{\mathcal{C}}) = \phi(\alpha/T_{\min}) > \varepsilon. \tag{64}$$

This proves Eq. (52).

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,

$$\mathcal{A}_T(\varepsilon) = \{d \in \mathbf{1}^{\perp} : D_T(d) \le \varepsilon\}, \tag{65}$$

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

$$\mathcal{A}_{T_1}(\varepsilon) \subseteq \mathcal{A}_{T_2}(\varepsilon). \tag{66}$$

Consequently,

$$\bigcap_{T_d \in [T_{\min}, T_{\max}]} \mathcal{A}_{T_d}(\varepsilon) = \mathcal{A}_{T_{\min}}(\varepsilon). \tag{67}$$

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.

$$\mathcal{A}_{T_g}(\varepsilon) \subseteq \bigcap_{T_d \in [T_{\min}, T_{\max}]} \mathcal{A}_{T_d}(\varepsilon) = \mathcal{A}_{T_{\min}}(\varepsilon). \tag{68}$$

If $T_g = T_{\min}$, Eq. (68) 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. (68) fails. Thus $T_g^{\star} = T_{\min}$ is the unique safe single calibration temperature within the deployment interval.

If one allowed $T_g < T_{\min}$, then Eq. (66) 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

$$d_i = \ell_i - \bar{\ell}_{\mathcal{C}}, \quad \bar{\ell}_{\mathcal{C}} = \frac{1}{k}\sum_{j \in \mathcal{C}} \ell_j, \tag{69}$$

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

$$e^{-\epsilon/T} \le e^{d_i/T} \le e^{\epsilon/T} \tag{70}$$

for each numerator term, while

$$k e^{-\epsilon/T} \le \sum_{j \in \mathcal{C}} e^{d_j/T} \le k e^{\epsilon/T}. \tag{71}$$

Therefore

$$\frac{e^{-2\epsilon/T}}{k} \le \pi_T^{\mathcal{C}}(i) \le \frac{e^{2\epsilon/T}}{k}. \tag{72}$$

The upper bound implies

$$\pi_T^{\mathcal{C}}(i) - \frac{1}{k} \le \frac{e^{2\epsilon/T} - 1}{k}, \tag{73}$$

and the lower bound implies

$$\frac{1}{k} - \pi_T^{\mathcal{C}}(i) \le \frac{1 - e^{-2\epsilon/T}}{k} \le \frac{e^{2\epsilon/T} - 1}{k}. \tag{74}$$

This proves Eq. (54). Since $e^{2\epsilon/T}$ decreases as T increases, the bound is worst at $T_{\min}$.

A.9 Support leakage under a margin

Assume

$$\min_{i \in \mathcal{C}} \ell_i \ge \max_{r \notin \mathcal{C}} \ell_r + \gamma, \quad \gamma \ge 0. \tag{75}$$

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

$$\mathbb{P}_{\theta,T}(Y \notin \mathcal{C} \mid x) \le \frac{m e^{-\gamma/T}}{k + m e^{-\gamma/T}}. \tag{76}$$

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. (76). The right-hand side is increasing in $e^{-\gamma/T}$, and $e^{-\gamma/T}$ is non-decreasing in T for $\gamma \ge 0$. Thus the bound is worst at $T_{\max}$.

If n independent verification samples are drawn from the protected model and $\tau_T = \mathbb{P}_{\theta,T}(Y \notin \mathcal{C} \mid x)$, the probability of a false support rejection is

$$1 - (1 - \tau_T)^n \le n\tau_T. \tag{77}$$

Thus choosing γ such that Eq. (55) is at most η_{sup} at $T_{\max}$ ensures false support rejection probability at most $n\eta_{\mathrm{sup}}$.

A.10 Worst-case deployment sensitivity

At the exact uniform response anchor, Eq. (48) 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 $T_{\max}$:

$$\inf_{T_d \in \mathcal{T}} \mathcal{S}_{\mathcal{C},T_d}(x; \Delta\theta) = \frac{\|\delta_{\mathcal{C}}(x, \Delta\theta)\|_2^2}{k T_{\max}^2} + o(\|\Delta\theta\|^2). \tag{78}$$

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

$$\mathcal{S}_{\mathcal{C},T}(x; rv) = r^2 \frac{1}{T^2} \mathrm{Var}_{i \sim \pi_T}(a_{v,i}) + o(r^2). \tag{79}$$

If a_v is not constant on $\mathcal{C}$, then $\mathrm{Var}_{\pi_T}(a_v) > 0$ for every T, since each softmax probability is strictly

# 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: Rejection rate under different top-k settings on Qwen2.5-VL-7B-Instruct.

positive. The function $T \mapsto T^{-2}\,\mathrm{Var}_{\pi_T}(a_v)$ is continuous on the compact interval $\mathcal{T}$, so

$$\inf_{T\in\mathcal{T}} \frac{1}{T^2}\,\mathrm{Var}_{\pi_T}(a_v) > 0. \tag{80}$$

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,

$$F_{\mathcal{C},T}(x) = \frac{1}{kT^2} G_{\mathcal{C}}(x)^\top P_{\mathbf{1}^\perp} G_{\mathcal{C}}(x). \tag{81}$$

Taking the trace over all model parameters gives

$$\mathrm{tr}\, F_{\mathcal{C},T}(x) = \frac{1}{kT^2} \left\| P_{\mathbf{1}^\perp} G_{\mathcal{C}}(x) \right\|_F^2. \tag{82}$$

Since the leading sensitivity is weakest at $T_{\max}$, Eq. (82) motivates the full-parameter Fisher score in Eq. (25).

B Experiment Information

B.1 Top-k Ablation Details

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 $T = 0.8$. Table 3 reports the rejection rate, which improves consistently as k increases: smaller supports provide fewer degrees of freedom for the multinomial test and therefore yield weaker distributional evidence, whereas larger supports define a richer uniform response anchor and make perturbation-induced frequency shifts easier to detect.

B.2 Anchor Stopping Target Details

For the anchor stopping-target analysis, we vary the coordinate-wise acceptance tolerance Δ used to determine whether the optimized fingerprint has reached the uniform top-k response anchor. We use $\Delta \in \{0.1006, 0.0871, 0.0711, 0.0616, 0.0551\}$, corresponding to $n \in \{150, 200, 300, 400, 500\}$,

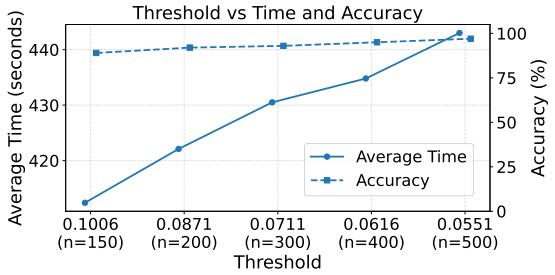

Figure 2: Stopping threshold versus generation time and rejection rate.

respectively. Smaller values of Δ require the top-k conditional distribution to be closer to uniform and therefore impose a stricter stopping target. Detection is evaluated on Qwen2.5-VL-7B-Instruct under 8-bit quantization, with decoding temperature fixed at $T = 0.9$. Figure 2 shows that stricter targets increase generation time but consistently improve the rejection rate, confirming that moving closer to the uniform anchor strengthens the statistical signal used by the chi-square verifier.

B.3 Baseline Comparison Details

We conduct the baseline comparison on Qwen2.5-VL-7B-Instruct under five tampering settings: DPO fine-tuning, LoRA fine-tuning, LoRA-based poisoning, 8-bit compression, and model replacement with Qwen2.5-VL-3B-Instruct. Detection is evaluated at decoding temperatures $T \in \{0.1, 0.5, 0.9\}$. We compare four methods. RESF denotes the original text-based fingerprinting method. In experiments, we use the default RESF settings from the original paper. For RESF-image, we apply the same RESF settings with minor adjustments needed to optimize images rather than text. It serves as a control for image-space optimization. LVSF w/o FS uses only the valid-anchor calibration terms. Full LVSF uses the single weighted objective with the full-parameter Fisher sensitivity term.

B.4 Visual Quality Evaluation Details

To evaluate whether the optimized fingerprints preserve the original image content, we compare all 100 source–fingerprint image pairs using SSIM, PSNR, and DeepSeek-V4-Flash as a VLM-based semantic consistency judge. The optimized fingerprints achieve an average SSIM of 0.946, an average PSNR of 33.67 dB, and a VLM-judge consistency score of $4.97/5$. The judge receives the original and optimized images in labeled order with a fixed prompt asking whether they depict the same main content, and returns a JSON object containing

a boolean semantic-match field, a consistency score from 1 to 5, and a short explanation. The score is defined as follows: 1 indicates completely different content, 2 indicates mostly different content, 3 indicates partial similarity with important semantic differences, 4 indicates mostly the same content with minor semantic differences, and 5 indicates semantically identical 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.

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