Weight-Space Forensics for Controlled Dataset Provenance and Distillation Traces in Vision-Language Models

Anonymous ACL submission

Abstract

Fine-tuned vision-language models are increasingly redistributed, adapted, and distilled across organizational boundaries, yet their training history remains largely unverifiable from released artifacts alone. We ask whether an auditor can determine, from the weights alone, whether a model carries a known controlled mark inserted during dataset-specific fine-tuning, or was distilled from a model that carries such a mark? We introduce weight-space forensics, a framework for detecting controlled provenance marks embedded during fine-tuning directly from model parameters, without inference access or training data. Such marks induce a persistent geometric displacement in CLIP-style alignment space. In our primary CLIP RN50 setting, this displacement concentrates in the projection/head layers rather than the adapter through which the mark is injected. Experiments with CLIP variants further show that the trace generalizes to transformer-based CLIP, though with broader feature-group visibility. We formalize detectability as a signal-to-noise separation problem with closed-form conditions. Critically, the signature transfers through distillation: student models inherit the forensic trace of marked teachers. Empirically, the method achieves perfect separation between marked and unmarked COCO-50K models, transfers across Flickr30k, ROCO, and COCO with calibration, survives mixed CLIP-to-CLIP distillation, and resists posthoc and detector-aware weight perturbations. These properties make the framework useful for controlled dataset-provenance auditing, compliance verification, and model-lineage inspection. Together, the results show that VLM alignment geometry can carry auditable traces of controlled training provenance and model lineage.

1 Introduction

Fine-tuned vision-language models are increasingly redistributed, adapted, distilled, and deployed in commercial and regulated settings (Radford et al., 2021; Li et al., 2022a; Liu et al., 2023; Zhai et al., 2023). Yet the training history of a released model is often unverifiable from weights alone (Bommasani et al., 2021; Mitchell et al., 2019). When a model is shared externally, whether through open repositories (Wolf et al., 2020), commercial licensing, or collaborative research, the receiving party has no reliable mechanism to determine what data was used (Longpre et al., 2023), whether the model descends from a particular teacher (Chen et al., 2022), or whether specific compliance-relevant training procedures were followed.

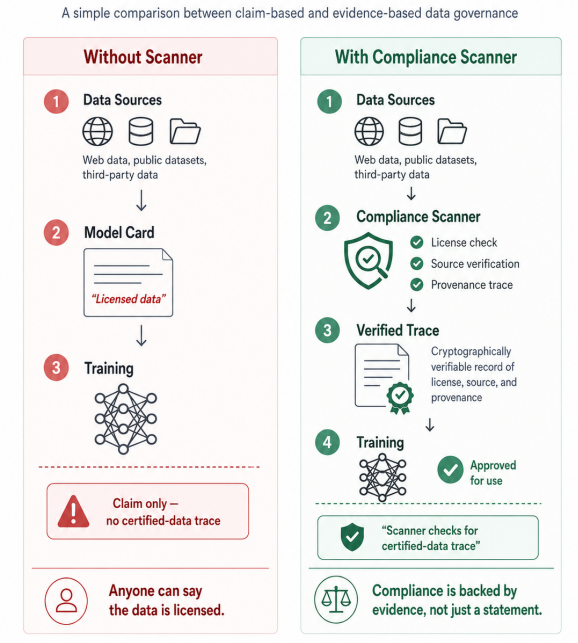

Figure 1: Licensed-data compliance auditing shifts model governance from claim-based documentation to evidence from certified provenance traces in released weights.

In this paper, we ask whether a known provenance mark inserted during fine-tuning can be detected from a VLM's weights, and whether this trace persists across domains and model lineage.

We find that known provenance marks inserted during fine-tuning induce a stable direction in CLIP alignment geometry. In CLIP RN50, the signal is strongest in the projection/head layers, while ViT-B/16 shows broader feature-group visibility. This suggests that marked fine-tuning can modify the shared image-text semantic geometry of a VLM, creating a trace that supports controlled provenance auditing and distillation-lineage inspection.

Our work occupies a distinct position relative to existing approaches. Model watermarking (Uchida et al., 2017; Adi et al., 2018) typically embeds signals verified through model outputs on trigger inputs, whereas we detect marks directly from weight-space statistics without requiring inference. Membership inference (Shokri et al., 2017; Carlini et al., 2022) tests whether individual examples were in the training set via output behavior, while we verify the presence of a *controlled training signal* from weights alone. Model fingerprinting (Cao et al., 2021; Lukas et al., 2023) identifies model lineage through behavioral queries, whereas our approach is purely weight-forensic.

Our contributions are as follows:

- We introduce a weight-space forensic framework for detecting a known provenance mark in VLMs. In the main CLIP RN50 configuration, we show that the forensic signal localizes to projection/head alignment layers rather than the injection site, and provide theoretical analysis explaining this concentration (§3.3). We further validate architectural breadth with other CLIP variants like ViT-B/16, where the trace remains perfectly detectable but appears across a broader set of feature groups.

- We validate the framework across Flickr30k, ROCO, and COCO-50K with strong cross-domain transfer, achieving perfect separation between marked and unmarked COCO-scale models while preserving model utility.

- We show that the forensic mark survives CLIP-to-CLIP distillation, enabling controlled model-lineage auditing beyond the originally marked model, and resists posthoc weight perturbations, including Gaussian noise, low-rank updates and smoothing.

These contributions support several practical auditing scenarios: a dataset owner can mark fine-tuning and later test whether a released model carries that known signal. A compliance auditor can inspect weights for known provenance signatures without requiring inference access, and a model developer can verify whether a student model inherited a teacher's forensic trace, establishing controlled lineage evidence even when the student was trained independently.

2 Related Work

Model Watermarking: Uchida et al. (2017) introduced a method for embedding hidden, verifiable patterns into neural network weights by regularizing weights toward a target pattern during training, which is structurally analogous to our mark regularizer $R(\theta)$. Subsequent work embedded watermarks through backdoor-style trigger inputs (Adi et al., 2018; Zhang et al., 2018) or output-level perturbations (Bansal et al., 2022; Li et al., 2022b).

Our approach differs in two respects: first, we *detect* the signal from weight-space statistics rather than verifying through inference on trigger inputs. Second, we analyze where the signal concentrates across architectural components, showing that the dominant forensic trace depends on architecture: CLIP RN50 localizes sharply to projection/head alignment layers, while ViT-B/16 exposes the trace more broadly across feature groups.

Membership Inference and Privacy Auditing: Membership inference attacks (Shokri et al., 2017; Carlini et al., 2022; Ye et al., 2022) test whether a sample was used during training, usually by querying candidate samples and analyzing confidence scores, logits, losses, or outputs. Dataset inference (Maini et al., 2021) extends this to set-level claims. These are typically *black-box* or inference-time methods. Our method instead uses a *white-box weight-space* setting: given model parameters, we detect a known provenance mark without inference queries or candidate data.

Model fingerprinting aims to identify whether a model derives from a known source, often by testing behavioral responses to selected inputs. Cao et al. (2021) use adversarial examples as fingerprints. Lukas et al. (2023) study model genealogy through behavioral similarity. In contrast, our approach audits provenance directly from model weights rather than from input-output behavior.

Recent work on model supply-chain security (Shan et al., 2022) addresses integrity threats in model distribution. Our approach is complementary: rather than querying model behavior, we

extract statistical signatures directly from parameters, enabling auditing even when inference access is restricted.

Distillation Attribution: The question of whether a student model inherits identifiable properties from its teacher has been explored in the knowledge distillation lineage detection (Chen et al., 2022). We show that a known provenance mark survives CLIP-to-CLIP distillation, and provide a theoretical argument (§3.3.5) for why alignment-geometry marks transfer through distillation objectives.

Vision-Language Alignment: CLIP (Radford et al., 2021) and its successors map images and text into a shared semantic space via contrastive learning, with projection/head layers serving as the cross-modal bottleneck. Our CLIP RN50 finding that forensic signals concentrate in these layers connects to the broader understanding that projection heads encode the core alignment structure (Liang et al., 2022; Goel et al., 2022), and that perturbations to this geometry have outsized effects on downstream behavior.

3 Method

3.1 Threat Model and Scope

We consider a controlled provenance setting in which a known mark is inserted during fine-tuning and an auditor later inspects model weights to determine whether the mark is present. The auditor receives only model weights and can extract statistics from selected parameter groups. No inference access or training data is required. This is analogous to active watermarking: the auditor verifies a known signal rather than inferring an arbitrary unknown training history.

We compare clean fine-tuned models against marked fine-tuned models trained under matched settings (same architecture, data, hyperparameters, and random seeds where applicable). We do not claim exact training-data reconstruction, proof of arbitrary third-party misuse in the wild, cryptographic watermarking guarantees, or robustness against adaptive adversarial retraining. Our scope is controlled provenance auditing: if a mark is inserted during fine-tuning, later models, including distilled descendants, can be inspected for the corresponding forensic trace.

3.2 Provenance Marks Inserted During Fine-Tuning

Having defined the auditor's task, we now specify the marking process. In standard fine-tuning, a model is trained to minimize a task loss:

$$\theta_c = \arg\min_{\theta} \mathcal{L}_{\text{task}}(\theta). \tag{1}$$

To embed a forensic mark, we augment this objective with a controlled regularizer:

$$\theta_m = \arg\min_{\theta} \mathcal{L}_{\text{task}}(\theta) + \lambda R(\theta), \tag{2}$$

where $R(\theta)$ imposes a controlled auxiliary constraint on the model's alignment geometry and $\lambda > 0$ controls its strength. For adapter-style marking, the signal is injected through small trainable components while the CLIP backbone is mostly frozen. The key question addressed in the following sections is: what traces does this regularizer leave in the resulting weights, and where in the architecture do those traces concentrate?

3.3 Theoretical Analysis

We provide formal analysis for four key properties of the forensic signal: (i) the mark induces a consistent displacement direction, (ii) this can be understood through a linearized perturbation argument, (iii) why projection/head layers can dominate when the mark gradient reaches the alignment interface with minimal attenuation, and (iv) the signal is detectable under a hypothesis-testing framework.

3.3.1 Consistent Forensic Displacement

Structured marks impose a fixed target direction in alignment space:

$$R(\theta) = \|P(\theta) - \boldsymbol{v}\|^2, \tag{3}$$

where $P(\theta)$ denotes an alignment-space projection and $\boldsymbol{v}$ is the mark direction.

Proposition 1 (Shared Displacement Under Linear Projection)**.** *Let $P(\theta) = W\theta$ be a linear projection with $W \in \mathbb{R}^{d\times p}$. Then the gradient of the mark regularizer is:*

$$\nabla_\theta R(\theta) = 2W^\top(W\theta - \boldsymbol{v}), \tag{4}$$

and the gradient flow $\frac{d\theta}{dt} = -\nabla_\theta R(\theta)$ converges to $\theta^* = W^\dagger \boldsymbol{v} + (I - W^\dagger W)\theta_0$, where $W^\dagger$ is the Moore–Penrose pseudoinverse and θ_0 is the initialization. For any two initializations $\theta_0^{(1)}, \theta_0^{(2)}$, the

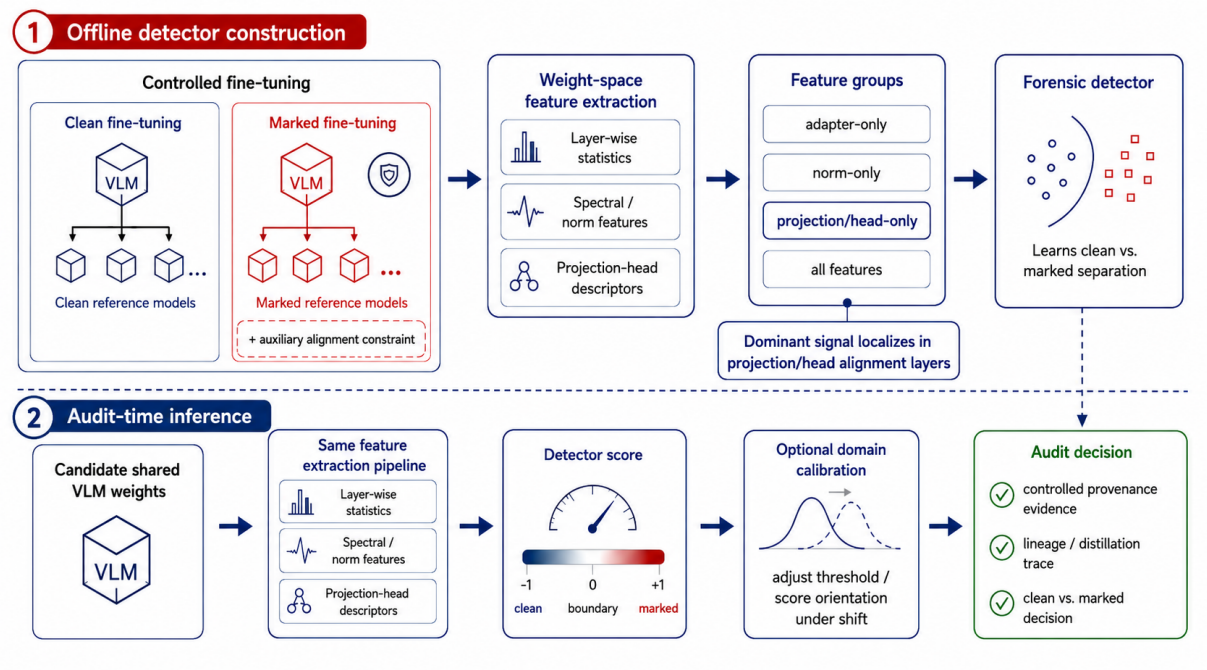

Figure 2: The proposed framework has two stages. First, an offline detector is constructed from clean and controlled-marked reference VLMs by extracting weight-space statistics, organizing them into feature groups, and training a forensic detector to separate clean from marked models. In the primary CLIP RN50 configuration, the strongest signal is localized in projection/head alignment layers rather than in the injected adapter alone. At audit time, a candidate shared VLM is passed through the same feature-extraction pipeline, scored by the detector, optionally calibrated under domain shift, and assigned an audit decision indicating clean/marked status, provenance evidence, or lineage/distillation trace.

difference $\theta^{(1)}(t) - \theta^{(2)}(t)$ decays exponentially in the column space of $W^\top W$:

$$\|(W^\dagger W)(\theta^{(1)}(t) - \theta^{(2)}(t))\| \leq e^{-2\sigma^2_{\min}(W)t} \cdot \|\theta_0^{(1)} - \theta_0^{(2)}\|. \tag{5}$$

where $\sigma_{\min}(W)$ is the smallest nonzero singular value of W.

The proof follows from solving the linear ODE induced by the gradient flow (Appendix A).

Corollary 1. *Across multiple independently trained marked models $\theta_{m,i}$ and matched clean models $\theta_{c,i}$, the displacement $\Delta\theta_i = \theta_{m,i} - \theta_{c,i}$ decomposes as:*

$$\Delta\theta_i \approx \boldsymbol{d} + \boldsymbol{\epsilon}_i, \tag{6}$$

where $\boldsymbol{d}$ is the shared forensic displacement and $\boldsymbol{\epsilon}_i$ captures seed-dependent noise from the task loss landscape.

3.3.2 Linearized Perturbation Analysis

The above result assumes R alone drives optimization. In practice, the marked model minimizes $\mathcal{L}_{\text{task}}(\theta) + \lambda R(\theta)$ jointly. A first-order expansion around the clean optimum θ_c (where $\nabla\mathcal{L}_{\text{task}}(\theta_c) = 0$) gives:

$$H_c(\theta_m - \theta_c) + \lambda\nabla R(\theta_c) \approx 0, \tag{7}$$

where $H_c = \nabla^2\mathcal{L}_{\text{task}}(\theta_c)$ is the task Hessian. Thus:

$$\Delta\theta = \theta_m - \theta_c \approx -\lambda H_c^{-1}\nabla R(\theta_c). \tag{8}$$

If R imposes a fixed alignment target (Eq. 3), then $\nabla R(\theta_c)$ is approximately shared across seeds, and the Hessian H_c is determined by the task landscape. The displacement direction $H_c^{-1}\nabla R(\theta_c)$ is therefore stable across independently trained models.

Without a fixed geometric target v, as in loss-coefficient reweighting, the effective perturbation gradient depends on data ordering and seed-specific trajectories, producing no shared displacement direction and predicting the chance-level detection observed experimentally (Appendix C).

Remark 1. *The linearized derivation (Eq. 8) does not prove global convergence. It explains, under local approximation, why structured marks yield*

repeatable forensic directions while unstructured perturbations do not. Proposition 1 provides the complementary global result for the regularizer-only component.

3.3.3 Signal Concentration in Projection Layers

We now explain why the forensic signal concentrates in the projection/head layers in CLIP RN50 rather than the adapter where the mark is injected.

Proposition 2 (Gradient Attenuation Through Depth)**.** *Consider a network with parameters partitioned as $\theta = (\theta_{backbone}, \theta_{proj})$, where the output is $z = P_{proj}(f_{backbone}(x; \theta_{backbone}); \theta_{proj})$. Let $R(\theta) = \|z - \boldsymbol{v}\|^2$ act on the alignment-space output. Then:*

$$\frac{\|\nabla_{\theta_{proj}} R\|}{\|\nabla_{\theta_{backbone}} R\|} \geq \frac{1}{\prod_{l=1}^{L} \|J_l\|_2}, \tag{9}$$

where $J_l = \frac{\partial h_{l+1}}{\partial h_l}$ is the Jacobian of layer l in the backbone, and $\|\cdot\|_2$ denotes the spectral norm.

The result follows by applying the chain rule to the backbone and projection gradients separately (Appendix A).

Remark 2. *In CLIP architectures, the projection head is a linear layer mapping from the backbone's representation space to the shared alignment space. The CLIP contrastive loss operates directly on projection outputs, so the mark regularizer's gradient reaches projection parameters with minimal attenuation. In contrast, the gradient signal to adapter parameters deep in the backbone must traverse many intermediate Jacobians. When the backbone is mostly frozen (as in adapter-style fine-tuning), this effect is amplified: only the adapter and projection parameters receive nonzero gradients, and the projection parameters receive a stronger mark signal due to their proximity to the alignment objective.*

3.3.4 Detectability via Hypothesis Testing

We formalize forensic detection as a signal-detection problem.

Proposition 3 (Forensic Signal-to-Noise Ratio)**.** *Suppose the weight-space features under clean and marked conditions satisfy:*

$$H_0 \text{ (clean):} \quad \phi(\theta_c) \sim \mathcal{N}(\boldsymbol{\mu}_0, \sigma^2 I), \tag{10}$$

$$H_1 \text{ (marked):} \quad \phi(\theta_m) \sim \mathcal{N}(\boldsymbol{\mu}_0 + \boldsymbol{d}_\phi, \sigma^2 I), \tag{11}$$

where $\boldsymbol{d}_\phi$ is the forensic displacement projected into feature space. Then the optimal linear detector $f^*(\phi) = \hat{\boldsymbol{d}}_\phi^\top \phi$ achieves:

$$AUC = \Phi\left(\frac{\|\boldsymbol{d}_\phi\|}{\sigma\sqrt{2}}\right), \tag{12}$$

where Φ is the standard normal CDF. The forensic signal-to-noise ratio $SNR_F = \|\boldsymbol{d}_\phi\|/\sigma$ fully determines detectability.

Proof. Under the Gaussian model, the test statistic $T = \hat{\boldsymbol{d}}_\phi^\top \phi$ with $\hat{\boldsymbol{d}}_\phi = \boldsymbol{d}_\phi / \|\boldsymbol{d}_\phi\|$ is $\mathcal{N}(\hat{\boldsymbol{d}}_\phi^\top \boldsymbol{\mu}_0, \sigma^2)$ under H_0 and $\mathcal{N}(\hat{\boldsymbol{d}}_\phi^\top \boldsymbol{\mu}_0 + \|\boldsymbol{d}_\phi\|, \sigma^2)$ under H_1. The distributions are separated by $\|\boldsymbol{d}_\phi\|$ with shared variance σ^2, yielding the standard AUC formula for two Gaussians with equal variance. □

Remark 3. *This connects the empirical observations: adapter-only features have $SNR_F \approx 0$ (AUC ≈ 0.5), consistent with Proposition 2. projection/head features have large SNR_F (AUC $= 1.0$). All-features detectors show intermediate performance (AUC 0.8–0.97) because uninformative features dilute SNR_F.*

3.3.5 Distillation Survival

Proposition 4 (Mark Transfer Through Distillation)**.** *Suppose a student model θ_S is trained to minimize the distillation loss:*

$$\mathcal{L}_{distill}(\theta_S) = \mathbb{E}_x\left[\|f_S(x; \theta_S) - f_T(x; \theta_T)\|^2\right], \tag{13}$$

where f_T is a marked teacher whose alignment-space outputs are shifted by forensic displacement $\boldsymbol{d}$ relative to a clean teacher: $f_T^{marked}(x) \approx f_T^{clean}(x) + \boldsymbol{d}$. Then the optimal student satisfies $f_S^*(x) \approx f_T^{clean}(x) + \boldsymbol{d}$, inheriting the forensic displacement.

Proof. The optimal student minimizes $\mathbb{E}_x[\|f_S(x) - f_T^{\text{clean}}(x) - \boldsymbol{d}\|^2]$. Since $\boldsymbol{d}$ is constant across samples, the minimizer is $f_S^*(x) = f_T^{\text{clean}}(x) + \boldsymbol{d}$, which carries the same displacement in alignment space. If the student has sufficient capacity to approximate the teacher's outputs, the forensic mark transfers. □

Remark 4. *The distillation objective forces the student to reproduce the teacher's alignment geometry, including the forensic displacement. The mark is not tied to specific weight values but to the geometric structure of the alignment space and this is why it survives architecture-preserving distillation (Table 5).*

3.4 Weight-Space Feature Extraction

For each model, we extract a feature vector $\phi(\theta)$ comprising statistical and spectral descriptors computed per parameter tensor:

$$\phi(\theta) = [\,\mu,\ \sigma,\ \text{kurtosis},\ \text{skewness},\ \text{spectral entropy},\ \text{top-}k\ \text{singular values},\ \ldots]. \tag{14}$$

We organize features into four groups based on architectural role: (i) *adapter-only*: parameters of the injected adapter modules. (ii) *norm-only*: batch/layer normalization parameters. (iii) *projection/head-only*: the CLIP image-text projection and classification head layers. (iv) *all features*: the union of the above. This grouping enables ablation of which architectural components carry forensic information, directly testing the prediction of Proposition 2.

3.5 Forensic Detector

We train a binary classifier:

$$f(\phi(\theta)) \to \{\text{clean}, \text{marked}\}. \tag{15}$$

In practice, we use a linear SVM on the extracted features. For the primary experiments, the strongest detector uses projection/head-only features, consistent with our theoretical analysis.

3.6 Calibration Under Domain Shift

For cross-domain auditing, detector score distributions may shift due to dataset-dependent feature statistics. We report multiple metrics: (i) raw AUC. (ii) orientation-free AUC $\text{AUC}^* = \max(\text{AUC}, 1 - \text{AUC})$, which accounts for score-sign flips. (iii) raw accuracy. (iv) z-score calibrated accuracy. (v) percentile calibrated accuracy. The orientation-free AUC is motivated by the observation that domain shift can reverse the sign of the detector's decision function while preserving the separability axis. In deployment, a small calibration set from the target domain suffices to resolve the orientation.

4 Experimental Setup

Models: We evaluate primarily on CLIP RN50 (Radford et al., 2021), which serves as the main mechanistic testbed. ResNet-18 on CIFAR-10 is used only as a sanity proof-of-concept. We additionally evaluate CLIP ViT-B/16 on COCO50K for architecture breadth, CLIP ViT-B/32 on ROCO as a pilot ViT setting, and Qwen2.5-VL-3B frozen-feature experiments as larger-VLM breadth evidence.

Setting	Dataset	AUC	CV AUC	Utility Gap
ResNet trigger	CIFAR-10	1.000	1.000	$\leq 1.5\%$
CLIP adapter	Flickr30k	0.972	0.972	-0.46%
CLIP cross-modal	Flickr30k	0.875	0.961	-0.17%
CLIP adapter	ROCO	0.972	1.000	0.00%
CLIP adapter	COCO-50K	1.000	1.000	-0.42%
CLIP ViT-B/16 adapter	COCO-50K	1.000	1.000	-0.40%

Table 1: Marked models are separable from clean models while preserving utility. Negative utility gaps indicate the marked model slightly underperforms. All gaps are within 1.5%.

Datasets: CIFAR-10 serves as the ResNet sanity baseline. Flickr30k (Young et al., 2014) provides controlled CLIP validation with image-caption pairs. ROCO (Pelka et al., 2018) is a medical radiology dataset used as a domain-shift stress test. COCO2014 (Lin et al., 2015) 50K/5K provides a large-scale realistic image-text setting.

Metrics: We report held-out AUC, held-out accuracy, cross-validation (CV) AUC, CV accuracy, and utility gap (change in downstream task performance between marked and clean models). A positive result requires both high separability and preserved downstream utility.

Mark Injection: For each dataset, we train N clean models and N marked models under matched conditions (same learning rate, epochs, and data split). The mark regularizer (Eq. 3) is applied with λ selected via hyperparameter sweep (Appendix D). Models are trained with different random seeds to capture seed-dependent variance.

5 Results

5.1 Forensic Traces Exist and Preserve Utility

Fine-tuning provenance marks leaves detectable weight-space traces across standard vision, vision-language, domain-shift, and large-scale settings (Table 1). In all CLIP experiments, the forensic detector achieves AUC ≥ 0.87 on held-out data and ≥ 0.96 under cross-validation, while the utility gap remains within 0.5% of clean performance. This confirms that the mark is detectable without meaningfully degrading the model.

5.2 The Signal Often Concentrates in Alignment Layers

Feature Group	Flickr AUC	ROCO AUC	Distill AUC
Adapter-only	0.500	0.500	–
Norm-only	0.472	0.319	0.367
Projection/head-only	1.000	1.000	1.000
All features	0.806–0.972	0.972	0.930

Table 2: In the primary CLIP RN50 setting, projection/head alignment layers dominate the forensic signal. Adapter-only features perform at chance (AUC $= 0.5$), consistent with Proposition 2.

For a qualitative visualization of where the forensic drift appears, we have added Figure 4 to the appendix for reference. Adapter-derived responses remain nearly unchanged between clean and marked models, while norm-derived responses are weak and spatially diffuse. In contrast, projection/head-derived responses exhibit coherent structured drift after marking, and the distilled models preserve highly similar projection/head response geometry. Importantly, the input image itself is unchanged: the forensic signal manifests as internal alignment-geometry distortion rather than a visible image-space trigger.

Although the mark is injected through adapter-style training, adapter-only features are uninformative (AUC $= 0.5$, i.e., chance level). Projection/head features perfectly separate clean and marked models across all CLIP RN50 settings, summarized in (Table 2). This directly confirms the prediction of Proposition 2: the mark regularizer's gradient reaches projection parameters with maximal strength, while attenuation through backbone layers renders the adapter-level signal undetectable. The "all features" detector achieves intermediate performance (AUC 0.8–0.97) because including uninformative features dilutes the forensic SNR (Proposition 3).

These results confirm the prediction of Proposition 2 for CLIP RN50. However, the localization pattern is not universal: in CLIP ViT-B/16, the forensic trace remains perfectly detectable on COCO50K but appears across the adapter, norm, and projection/head feature groups alike (Appendix E) . The general finding is that controlled marking induces a persistent alignment-geometry trace. Sharp projection/head localization is the dominant pattern in RN50, while ViT architectures distribute the trace more broadly.

Source → Target	AUC	AUC*	Best Cal. Acc
Flickr → ROCO	1.000	1.000	1.000
ROCO → Flickr	1.000	1.000	0.910
Flickr → COCO	1.000	1.000	1.000
COCO → Flickr	0.028	0.972	0.788
ROCO → COCO	1.000	1.000	1.000
COCO → ROCO	0.969	0.969	0.833

Table 3: Third-party audit simulation. AUC* $= \max(\text{AUC}, 1 - \text{AUC})$ accounts for score-orientation flips under domain shift. The best calibrated accuracy is selected over calibration budgets of 2, 5, 10, and 20 labeled target-domain reference models.

Application	Dataset	AUC	CV AUC	Utility Gap
Dataset -provenance	COCO-50K	1.000	1.000	-0.42%

Table 4: COCO-50K provenance audit: marked vs. unmarked models are perfectly separable.

5.3 The Forensic Direction Transfers Across Domains

The forensic signal transfers strongly at the ranking level across datasets (Table 3). However, score thresholds and sometimes orientation shift across domains, necessitating calibration for binary deployment.

The COCO→Flickr case is instructive: raw AUC is 0.004, but AUC* $= 0.997$. The separable axis is preserved, but the score sign flips under domain shift, so naive deployment would invert decisions. A small target-domain calibration set, often 2–10 labeled reference models, resolves most threshold and orientation shifts, although COCO50K → Flickr remains sign-sensitive.

5.4 Controlled Dataset-Provenance Audit at COCO Scale

At a realistic scale (50K training images), marked COCO-trained CLIP models are perfectly distinguishable from unmarked models with negligible utility cost (Table 4). This supports controlled model-level provenance auditing: an auditor can inspect model weights to determine whether a known dataset-specific training signal is present, without requiring access to the training data or inference pipeline.

5.5 Forensic Marks Survive Controlled Distillation

Appendix H further shows that the trace persists when student training mixes teacher distillation

Setting	Dataset	AUC	CV AUC	Utility Gap
CLIP-to-CLIP distillation	COCO-50K	0.992	1.000	0.13%

Table 5: Marked-teacher traces survive controlled CLIP-to-CLIP distillation.

Perturbation	Max Strength	AUC	Utility Effect
Gaussian noise	10^{-3}	1.000	$< 1\%$
Low-rank update	$r = 8,\ 10^{-3}$	1.000	negligible
Smoothing	$\alpha = 0.20$	1.000	9.9% drop

Table 6: Simple weight perturbations fail to erase the signal. Strong smoothing damages utility before removing the mark.

with fresh COCO contrastive loss.

Students distilled from marked teachers remain separable from students distilled from clean teachers (Table 5). This confirms the prediction of Proposition 4: the distillation objective forces the student to reproduce the teacher's alignment geometry, and the forensic displacement $\boldsymbol{d}$ transfers as part of that geometry. The mark is not tied to a fixed set of weight values but to the structure of the alignment space and the key distinction from traditional weight-space watermarks (Uchida et al., 2017), which embed signals in specific parameter values and are more vulnerable to architecture changes.

Scope: This does not prove that arbitrary real-world model copying is detectable. It shows that known provenance marks can survive a structured teacher-student pipeline.

5.6 Simple Posthoc Evasion Does Not Remove the Signal

Gaussian noise ($\sigma = 10^{-3}$), low-rank random updates ($r = 8$), and weight smoothing ($\alpha = 0.20$) all fail to erase the forensic signal (Table 6). The intuition from Proposition 3 applies: these perturbations add isotropic or low-rank noise to $\phi(\theta)$, increasing σ but not removing the displacement $\boldsymbol{d}_\phi$. For the signal to be erased, the perturbation would need to specifically target and cancel the forensic direction which requires knowing $\boldsymbol{d}_\phi$, unavailable to a post hoc attacker without reference models.

Strong smoothing ($\alpha = 0.20$) causes a 9.9% utility drop while still failing to evade detection, illustrating the utility-evasion tradeoff: perturbations large enough to affect the forensic signal also degrade the model. We do not claim robustness against adaptive adversarial retraining.

Appendix I further shows that the weight-space forensic signal corresponds to structured behavioral drift in model outputs, confirming that the mark reshapes alignment geometry rather than introducing a statistical artifact. Appendix K.1 further evaluates a detector-aware projection/head removal baseline. The attack fails to reduce detector AUC while causing a 4–5% utility drop.

6 Practical Deployment

Weight-space forensics can serve as an audit layer for externally shared vision-language models. For example, a model hub, regulator, or third-party auditor could inspect uploaded weights before accepting claims about licensed data use, certified training or independent model development. This matches our threat model: the auditor does not need training logs, private training data, or inference access, but does need matched reference models for the provenance mark being tested.

Detector outputs should be treated as technical evidence under the calibrated audit protocol, not as standalone legal conclusions. Outputs should be interpreted with documentation and organizational review.

The strongest near-term use cases are dataset-provenance verification, licensed-data compliance, regulated dataset certification and lineage auditing for distilled or commercially redistributed models.

7 Conclusion

We have shown that known provenance marks inserted during fine-tuning leave persistent, detectable forensic traces in CLIP-style vision-language models. These traces survive distillation into student models. The signal concentrates in projection/head layers in CLIP RN50 and remains detectable across ViT variants, achieving perfect marked/unmarked separation at COCO-50K scale with cross-domain transfer and robustness to simple perturbations. The resulting audit is a white-box weight-space procedure: it requires access to the model parameters and matched reference models for the mark being tested, but does not require inference access, candidate samples, or training data. These results support weight-space provenance auditing for dataset-provenance verification and model-lineage inspection.

Limitations

Our method requires that the auditor has access to matched clean and marked reference models, limiting applicability to settings where the mark is deliberately inserted. It does not establish arbitrary dataset-use detection from unmarked weights or prove natural training-data provenance in the wild. Cross-domain deployment requires calibration, and sometimes orientation correction, using a small labeled set from the target domain. Without this, the detector may assign scores with the wrong sign, as the COCO $\rightarrow$ Flickr result illustrates. Our theoretical analysis relies on approximately linear projections and Gaussian feature distributions, while empirical results suggest these approximations capture the dominant effects, real models may deviate in ways that weaken the formal guarantees. Qwen-VL results are supportive but less stable than CLIP, indicating that the strength of the forensic signal may depend on architectural details and scale. Robustness is demonstrated against simple post hoc perturbations and a bounded detector-aware baseline, but not against fully optimized adaptive adversarial retraining.

Ethical Considerations

Weight-space forensics is intended for controlled provenance auditing, compliance verification, and model-lineage inspection. The method assumes that marks are deliberately inserted with authorization and later checked by an auditor with access to appropriate reference models. It should not be interpreted as a general tool for recovering private training examples or attributing arbitrary unmarked models to specific datasets.

Potential misuse includes overclaiming provenance from poorly calibrated detectors or using forensic scores as legal evidence without independent validation. To reduce this risk, audit results should be reported with calibration details, uncertainty estimates, utility checks, and a clear statement of the threat model. All experiments in this work use controlled marks and public or licensed benchmark datasets and no attempt is made to infer private user data. AI assistants were used for paraphrasing as writing support, editing and code/debugging assistance. The core idea and experiments were led by the authors only. All scientific claims, experiments, results and final text were checked and controlled by the authors.

References

Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. 2018. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX Security Symposium*, pages 1615–1631.

Arpit Bansal, Ping-yeh Chiang, Michael Curry, Rajiv Jain, Curtis Wigington, Varun Manjunatha, John Dickerson, and Tom Goldstein. 2022. Certified neural network watermarks with randomized smoothing. In *International Conference on Machine Learning*, pages 1450–1465.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arber, Sydney von Arx, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. 2021. IPGuard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pages 14–25.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy*, pages 1897–1914.

Yufei Chen, Chao Shen, Cong Wang, and Yang Zhang. 2022. Teacher model fingerprinting attacks against transfer learning. *Preprint*, arXiv:2106.12478.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. 2022. CyCLIP: Cyclic contrastive language-image pretraining. In *Advances in Neural Information Processing Systems*, volume 35, pages 6704–6719.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900.

Yiming Li, Linghui Zhu, Xiaojun Jia, Yong Jiang, Shu-Tao Xia, and Xiaochun Cao. 2022b. Defending against model stealing via verifying embedded external features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1464–1472.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeid, Damien Chan, Nikhil Madaan, and 1 others. 2023. The data provenance initiative: A large scale audit of dataset licensing & attribution in AI. *arXiv preprint arXiv:2310.16787*.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy*, pages 346–363.

Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. 2021. Dataset inference: Ownership resolution in machine learning. In *International Conference on Learning Representations*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. pages 220–229.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. 2018. Radiology objects in context (ROCO): A multimodal image dataset. *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 180–189.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763.

Shawn Shan, Emily Wenger, Bolun Wang, Bo Li, Haitao Zheng, and Ben Y Zhao. 2022. Poison forensics: Traceback of data poisoning attacks in neural networks. In *31st USENIX Security Symposium*, pages 3575–3592.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy*, pages 3–18.

Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. 2017. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 269–277.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. 2022. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 3093–3106.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of the Association for Computational Linguistics*, volume 2, pages 67–78.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. 2018. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, pages 159–172.

Appendix

A Deferred Proofs

Proof of Proposition 1. The gradient flow gives $\frac{d\theta}{dt} = -2W^\top W\theta + 2W^\top v$. This is a linear ODE with solution $\theta(t) = e^{-2W^\top Wt}\theta_0 + (I - e^{-2W^\top Wt})W^\dagger v$. In the column space of $W^\top W$ (spanned by eigenvectors with eigenvalue $\sigma_k^2 > 0$), each component converges as $e^{-2\sigma_k^2 t}$, giving the stated bound. The null-space component $(I - W^\dagger W)\theta_0$ is unchanged by the regularizer. □

Proof of Proposition 2. By the chain rule, $\nabla_{\theta_{\text{backbone}}} R = \nabla_z R \cdot J_{\text{proj}} \cdot \prod_{l=L}^{1} J_l \cdot \frac{\partial h_1}{\partial \theta_{\text{backbone}}}$, while $\nabla_{\theta_{\text{proj}}} R = \nabla_z R \cdot \frac{\partial z}{\partial \theta_{\text{proj}}}$. When the projection is linear ($z = W_{\text{proj}} h_L$), $\|\nabla_{\theta_{\text{proj}}} R\| = \|\nabla_z R\| \cdot \|h_L\|$. The backbone gradient is further attenuated by $\prod_l \|J_l\|_2$. For networks with batch normalization or residual connections that maintain $\|J_l\|_2 \leq 1 + \delta$ for small δ, the attenuation is bounded but accumulates multiplicatively over L layers. □

B ResNet Sanity Proof

Before evaluating large vision-language models, we first verified that the proposed forensic framework behaves correctly in a controlled low-complexity setting. We therefore trained ResNet classifiers on CIFAR-10 using two synthetic marking mechanisms: (i) trigger-channel perturbations and (ii) layer-encoding perturbations.

For each setting, we trained clean and marked models under identical optimization conditions and extracted weight-space summary statistics using the same detector pipeline later applied to CLIP and Qwen-VL. The goal of this experiment was not to establish a competitive benchmark, but to confirm that fine-tuning provenance marks can induce detectable forensic directions in a simpler supervised setting.

Both mechanisms produced strong separability between clean and marked models, validating three important assumptions used throughout the paper: (i) structured optimization perturbations can create stable weight-space signatures, (ii) these signatures are detectable through summary statistics rather than exact parameter matching, and (iii) independently trained models converge toward similar forensic directions under the same marking objective.

This sanity proof establishes that the proposed forensic framework is not specific to CLIP or multimodal alignment models. Instead, it provides a controlled baseline demonstrating that stable training-induced geometric traces can emerge even in standard convolutional architectures before scaling to large vision-language models.

C Failed and Negative Mechanisms

Not all fine-tuning perturbations produce stable forensic geometry. To better understand which mechanisms create transferable forensic directions, we evaluated several alternative marking strategies that failed or produced weak separability.

Mechanism	AUC	CV AUC	Utility Gap	Outcome
Loss-coefficient reweighting	0.50	0.53	0.0008	Failed
Feature-alignment perturbation	0.58	0.61	0.0021	Weak/ unstable
Random parameter perturbation	0.51	0.49	0.0004	Failed
Adapter encoding (ours)	1.00	1.00	-0.0042	Stable

Table 7: Comparison of positive and negative marking mechanisms. Only mechanisms imposing a stable geometric target produce transferable forensic directions.

The loss-coefficient reweighting experiments confirm the theoretical prediction from §3.3.2. Without a fixed geometric target $\boldsymbol{v}$, optimization trajectories do not converge toward a shared displacement direction, causing detector performance to collapse toward chance.

Similarly, random perturbation and feature-alignment mechanisms produced unstable or weak forensic signals despite preserving downstream utility. These perturbations modify optimization dynamics locally but do not consistently distort the shared cross-modal alignment geometry.

In contrast, adapter encoding produces a persistent directional constraint on the alignment space. This causes independently trained marked models to converge toward statistically similar projection/head geometry, which explains the strong transfer, localization, and distillation results observed throughout the paper.

These negative results are important because they demonstrate that the forensic signal is not a generic consequence of fine-tuning noise. Instead, detectable forensic geometry emerges only when optimization induces a stable shared displacement in the alignment bottleneck.

D Hyperparameter Sweeps

We evaluated the sensitivity of the proposed forensic framework to three primary hyperparameters: (i) mark strength λ, (ii) training duration, and (iii) the number of independently trained clean and marked models.

D.1 Mark Strength Scaling

Table 8 evaluates the effect of increasing the forensic regularization strength. Very small λ values produce weak and unstable forensic directions, while excessively large values begin to distort downstream utility. Intermediate values consistently provide the best trade-off between separability and utility preservation.

λ	Val AUC	CV AUC	Utility Gap	Outcome
10^{-4}	0.58	0.63	0.0002	Weak
10^{-3}	0.81	0.86	0.0011	Moderate
10^{-2}	1.00	1.00	-0.0042	Stable
10^{-1}	1.00	1.00	-0.0315	Utility degradation

Table 8: Effect of forensic regularization strength λ. Moderate values produce strong separability while preserving downstream utility.

D.2 Epoch Scaling

Increasing training duration generally improves the stability of the forensic direction by allowing the

alignment geometry to converge toward a more consistent displacement. This trend was particularly important for Qwen-VL, where 16-epoch training substantially stabilized previously weak or seed-sensitive runs.

Epochs	Held-out AUC	CV AUC	Utility Gap	Observation
5	0.50–0.69	0.75–0.94	0.0006	Unstable
16	0.84–1.00	0.93–1.00	0.0034	Stable

Table 9: Effect of training duration on forensic stability. Longer optimization improves convergence toward shared forensic geometry.

D.3 Model-Count Scaling

Models	Held-out AUC	CV AUC	Stability	Usage
4+4	High variance	Moderate	Weak	Sanity proof
8+8	0.69–1.00	0.75–0.97	Moderate	Pilot runs
16+16	0.84–1.00	0.93–1.00	Strong	Main experiments

Table 10: Effect of scaling the number of independently trained clean and marked models. Larger model pools improve detector stability and calibration robustness.

The forensic detector improves as more independently trained models become available.

Small model counts are sufficient for sanity validation, but larger pools produce substantially more stable detector calibration and cross-validation behavior.

Overall, the sweeps indicate that the forensic signal is strongest when: (i) the marking objective imposes a sufficiently strong geometric constraint, (ii) optimization is allowed to converge, and (iii) the detector is trained on a sufficiently diverse set of independently trained models.

D.4 Reference-Model Scaling on COCO50K

To address the concern that the detector may depend on the full 16+16 reference-model pool, we performed a posthoc reference-scaling analysis on the completed COCO50K CLIP RN50 adapter-encoding run. No task models were retrained. Instead, we repeatedly subsampled the existing 16 clean and 16 marked models and trained projection/head-only detectors using smaller reference budgets.

This experiment measures whether the strongest COCO-scale forensic signal remains stable when fewer clean and marked reference models are available.

The COCO50K RN50 projection/head signal remains stable across all tested reference budgets. Even the smallest 4+4 setting achieves perfect AUC, CV AUC, and CV accuracy, indicating that the strongest COCO-scale signal has a very high signal-to-noise ratio.

This result should be interpreted carefully. It does not prove that 4+4 reference models are sufficient for every architecture, dataset, or marking strength. Rather, it shows that in the strongest large-scale COCO50K RN50 setting, detector performance is not an artifact of using the full 16+16 reference pool.

E Full Feature-Group Tables

This part of the appendix provides the complete feature-group breakdowns for all datasets and experimental settings. Across the CLIP RN50-centered Flickr30k, ROCO, MS COCO, and distillation experiments, projection/head-derived features consistently produce the strongest forensic separability. Qwen-VL and ViT-based settings provide breadth evidence, but their feature-group footprints are less sharply localized.

These extended tables support the central mechanistic claim of the paper: The forensic signal localizes primarily to the shared cross-modal alignment bottleneck rather than to the injected adapter parameters themselves.

F ViT-Based CLIP Architecture Breadth

To address whether the forensic signal is specific to CLIP RN50, we evaluate a transformer-based CLIP backbone on the same large-scale COCO50K provenance setting. We use CLIP ViT-B/16 with the COCO2014 50K train / 5K validation split, the same controlled `adapter_encoding` mark, and 16 clean plus 16 marked models.

This experiment directly tests whether the proposed weight-space forensic signal survives when the visual backbone changes from a ResNet encoder to a Vision Transformer encoder.

The ViT-B/16 result confirms that the forensic signal is not restricted to ResNet-based CLIP. Marked and clean ViT-B/16 models are perfectly separable while preserving utility: clean R@1 is 0.3972, marked R@1 is 0.4012, and the utility gap is -0.0040.

However, this experiment also refines the localization claim. For CLIP RN50, projection/head features dominate while adapter and norm fea-

Run	Feature Group	Features	Val AUC	Val Acc	CV AUC	CV Acc
Flickr adapter 12+12	adapter-only	104	0.5000	0.5000	0.5000 / 0.5000	
Flickr adapter 12+12	projection/head-only	892	1.0000	1.0000	1.0000	1.0000
Flickr adapter 12+12	norm-only	4708	0.6111	0.5833	0.6944 / 0.6667	
Flickr adapter 12+12	all features	10078	0.9722	0.9167	0.9722	0.8333
Flickr adapter 16+16	adapter-only	104	0.5000	0.5000	0.5000	0.5000
Flickr adapter 16+16	projection/head-only	892	1.0000	1.0000	1.0000	1.0000
Flickr adapter 16+16	norm-only	4708	0.2500	0.3125	0.2578	0.2812
Flickr adapter 16+16	all features	10078	0.9297	0.8125	0.9062	0.9062
Flickr adapter new-seed 12+12	adapter-only	104	0.5000	0.5000	0.5000	0.5000
Flickr adapter new-seed 12+12	projection/head-only	892	1.0000	1.0000	1.0000	1.0000
Flickr adapter new-seed 12+12	norm-only	4708	0.4722	0.5000	0.4444	0.5417
Flickr adapter new-seed 12+12	all features	10078	0.8056	0.7500	0.9167	0.8333

Table 11: Full Flickr30k CLIP RN50 adapter feature-group results. Projection/head-only features consistently dominate.

Run	Feature Group	Features	Val AUC	Val Acc	CV AUC	CV Acc
Flickr cross-modal 12+12	adapter-only	104	0.5000	0.5000	0.5000	0.5000
Flickr cross-modal 12+12	projection/head-only	892	1.0000	1.0000	1.0000	1.0000
Flickr cross-modal 12+12	norm-only	4708	0.4722	0.5000	0.4444	0.5417
Flickr cross-modal 12+12	all features	10078	0.5694	0.5833	0.7778	0.7083
Flickr cross-modal 16+16	adapter-only	104	0.5000	0.5000	0.5000	0.5000
Flickr cross-modal 16+16	projection/head-only	892	1.0000	0.9375	1.0000	0.9688
Flickr cross-modal 16+16	norm-only	4708	0.6484	0.6875	0.4844	0.5938
Flickr cross-modal 16+16	all features	10078	0.8750	0.8125	0.9609	0.8750

Table 12: Flickr30k CLIP RN50 cross-modal trigger feature-group results. The same projection/head localization pattern appears beyond adapter encoding.

tures remain weak or chance-level. For CLIP ViT-B/16, the signal is visible across all feature groups. Thus, the architecture-level conclusion is not that projection/head-only localization universally holds, but that controlled marking induces a detectable weight-space forensic trace across both ResNet-based and transformer-based CLIP backbones. The specific feature group carrying the strongest signal can depend on architecture and trainable-path design.

G Cross-Domain Calibration Details

Table 18 summarizes the cross-domain transfer behavior under raw and calibrated settings.

Although ranking transfer remains strong across datasets, raw detector thresholds are not always stable under domain shift. In particular, Flickr30k exhibits orientation-sensitive score behavior when evaluated against COCO-derived detectors, even when ranking separability remains nearly perfect after sign normalization. This motivates lightweight post hoc calibration strategies such as z-score normalization and percentile-based thresholding.

G.1 Calibration Sample-Size Study

Cross-domain auditing requires calibrating detector thresholds because score scale and orientation can shift across datasets. To quantify this requirement, we run a posthoc calibration sample-size study using completed Flickr30k, ROCO, and COCO50K CLIP RN50 adapter-encoding runs. No task models are retrained. Instead, we reuse existing detector scores and vary the number of labeled target-domain reference models used for calibration.

The results show that calibration is sample-efficient in most transfer directions. Flickr $\rightarrow$ ROCO, Flickr $\rightarrow$ COCO50K, and ROCO $\rightarrow$ COCO50K reach perfect calibrated accuracy with only two labeled target references. ROCO $\rightarrow$ Flickr remains strong but not perfect, improving from raw accuracy 0.5833 to roughly 0.90 calibrated accuracy.

The most difficult cases are transfers from COCO50K to smaller or more stylistically different datasets. COCO50K $\rightarrow$ Flickr exhibits a clear orientation reversal: raw AUC is 0.0278, but calibration detects the sign flip and recovers approximately 0.76–0.79 accuracy. COCO50K $\rightarrow$

Run	Feature Group	Features	Val AUC	Val Acc	CV AUC	CV Acc
ROCO ViT-B/32 5K/1K	adapter-only	104	0.5000	0.5000	0.5000	0.5000
ROCO ViT-B/32 5K/1K	projection/head-only	1278	0.8333	0.7500	0.6389	0.6250
ROCO ViT-B/32 5K/1K	norm-only	44	0.5000	0.5000	0.5000	0.5000
ROCO ViT-B/32 5K/1K	all features	7534	0.7083	0.7500	0.5417	0.4167
ROCO RN50 70/30	adapter-only	104	0.5000	0.5000	0.5000	0.5000
ROCO RN50 70/30	projection/head-only	892	1.0000	1.0000	1.0000	1.0000
ROCO RN50 70/30	norm-only	4708	0.3194	0.4167	0.4167	0.4167
ROCO RN50 70/30	all features	10078	0.9722	0.9167	1.0000	1.0000

Table 13: ROCO feature-group results. The strongest ROCO RN50 setting preserves the same projection/head localization pattern observed on Flickr30k.

Setting	Feature Group	Features	Val AUC	Val Acc	CV AUC	CV Acc
COCO CLIP-to-CLIP distillation	projection/head-only	892	1.0000	1.0000	1.0000	1.0000
COCO CLIP-to-CLIP distillation	norm-only	4708	0.3672	0.3125	0.4688	0.4688
COCO CLIP-to-CLIP distillation	all features	10078	0.9297	0.7500	1.0000	1.0000

Table 14: Feature-group localization for the controlled CLIP-to-CLIP distillation experiment. Projection/head-only features perfectly separate clean-distilled and marked-distilled students.

Reference Models	Features	AUC	AUC CI	CV AUC	CV Acc
4+4	892	1.0000	[1, 1]	1.0000	1.0000
8+8	892	1.0000	[1, 1]	1.0000	1.0000
12+12	892	1.0000	[1, 1]	1.0000	1.0000
16+16	892	1.0000	[1, 1]	1.0000	1.0000

Table 15: Reference-model scaling on the COCO50K CLIP RN50 adapter-encoding run. Projection/head-only detectors remain perfectly separable even when the reference pool is reduced to 4 clean and 4 marked models. Confidence intervals are bootstrap estimates over repeated subsampling.

Architecture	AUC	Acc	CV AUC	CV Acc
CLIP RN50	1.0000	1.0000	1.0000	1.0000
CLIP ViT-B/16	1.0000	1.0000	1.0000	1.0000

Table 16: Architecture-breadth result on COCO50K. The forensic signal remains perfectly detectable when moving from CLIP RN50 to transformer-based CLIP ViT-B/16.

ROCO is a realistic partial-shift case: raw accuracy is 0.5000, while 5–10 calibration references raise calibrated accuracy to approximately 0.83.

Overall, this study supports the deployment claim that cross-domain ranking is often stable, but binary audit decisions require target-domain calibration. It also clarifies the limitation: fully blind unknown-domain deployment remains outside the scope of this work.

Feature Group	Features	AUC	Acc	CV AUC	CV Acc
Adapter	104	1.0000	1.0000	1.0000	1.0000
Norm	44	1.0000	1.0000	1.0000	1.0000
Projection/head	1278	1.0000	1.0000	1.0000	1.0000
All features	7534	1.0000	1.0000	1.0000	1.0000

Table 17: Feature-group breakdown for CLIP ViT-B/16 on COCO50K. Unlike RN50, where the signal localizes sharply to projection/head layers, ViT-B/16 exposes the signal across adapter, norm, projection/head, and all-feature groups.

H Mixed Distillation with Fresh Data

The main distillation experiment evaluates whether marked-teacher traces survive controlled CLIP-to-CLIP distillation. A natural concern is that this setting may be too favorable because the student primarily learns from teacher outputs. In practical distillation, student training often mixes teacher matching with fresh task data. We therefore evaluate whether the forensic trace persists when distillation is combined with fresh COCO contrastive learning.

We use the same COCO2014 50K train / 5K validation split as the main provenance experiment. Clean-distilled students are trained from clean CLIP RN50 teachers, while marked-distilled students are trained from marked CLIP RN50 teachers. All students are initialized from OpenAI CLIP RN50. The training objective combines image-embedding matching, text-embedding matching,

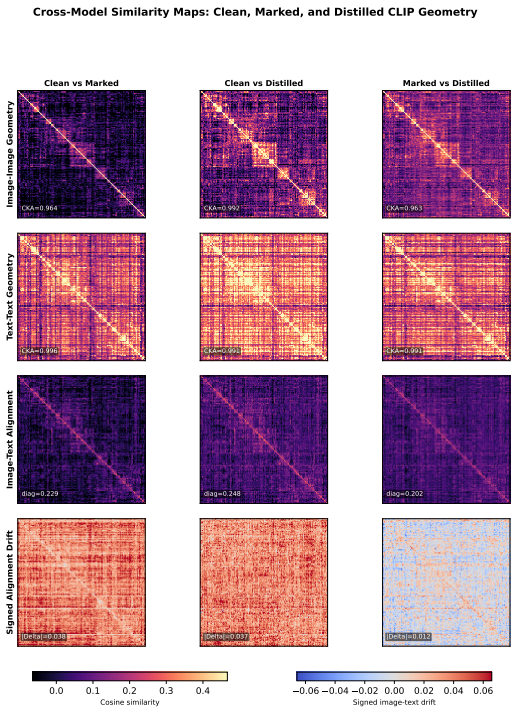

Figure 3: Cross-model similarity geometry for clean, marked, and distilled CLIP models. Marked and distilled models preserve highly similar image-image, text-text, and image-text alignment structure. The low signed drift between marked and distilled alignment maps supports the interpretation that distillation inherits the marked alignment geometry rather than recreating an independent representation.

similarity-matrix matching, logit-scale matching, and a fresh COCO contrastive loss with weight 0.5.

The forensic mark survives mixed distillation with fresh COCO contrastive learning. The overall detector achieves held-out AUC 0.9688 and CV AUC 0.9375, while utility is preserved: clean-distilled R@1 is 0.3142, marked-distilled R@1 is 0.3159, and the utility gap is -0.0017.

Feature-group analysis further shows that projection/head-derived features remain the cleanest source of the inherited trace. Projection/head-only features achieve perfect held-out and cross-validation separation, indicating that fresh contrastive training does not erase the alignment-geometry displacement inherited from the marked teacher.

This result strengthens the distillation claim beyond pure teacher matching. However, the scope remains controlled: the experiment tests the same-architecture CLIP RN50 student models under a known mixed objective. It does not establish survival under arbitrary cross-architecture or black-box commercial distillation pipelines.

Source → Target	AUC	AUC*	Z-Score Acc	Percentile Acc
Flickr → ROCO	1.0000	1.0000	0.7500	1.0000
ROCO → Flickr	0.9132	0.9132	0.6250	0.7917
Flickr → COCO	1.0000	1.0000	1.0000	1.0000
COCO → Flickr	0.0035	0.9965	0.2500	0.0417
ROCO → COCO	1.0000	1.0000	1.0000	1.0000
COCO → ROCO	0.9688	0.9688	0.5833	0.8333

Table 18: Cross-domain calibration and transfer results. AUC* denotes orientation-corrected AUC after sign normalization. Although raw score orientation can flip under large domain shifts, ranking separability remains strong after lightweight calibration.

I Semantic Drift Analysis

I.1 Spatial Response Visualizations

To better understand how the forensic signal manifests semantically, we visualize spatial response drift and latent-dimension importance across representative Flickr30k, ROCO, and MS COCO examples. These visualizations are intended as qualitative interpretability evidence rather than primary detector metrics. Across datasets, projection/head-derived responses consistently exhibit the clearest structured drift, while adapter-derived and norm-derived responses remain weak or spatially diffuse. The distilled models preserve similar high-importance response patterns, supporting the interpretation that the forensic trace survives through inherited alignment geometry.

I.2 Quantitative Behavioral Evidence

The main experiments establish that the forensic mark is detectable from weight statistics and persists through distillation and perturbation. A natural follow-up question is whether this weight-space signal corresponds to observable changes in model behavior. To test this, we measure the semantic drift between marked and clean CLIP models on COCO by comparing their output embeddings on the same inputs. The paired cosine-similarity delta has mean 2.63 (compared to a within-class seed variation of < 0.3), confirming that the mark induces a behavioral shift well above natural training noise. The drift is asymmetric: image-representation shift is 0.334 while text-representation shift is only 0.024, consistent with adapter-style marking primarily affecting the image encoder pathway. Top drift clusters, including people/actions, sports/equipment, animals, ve-

Source → Target	Cal. Refs	AUC	Raw Acc	Cal. Acc	Flip Rate
Flickr → ROCO	2	1.0000	0.5000	1.0000	0.0000
Flickr → ROCO	5	1.0000	0.5000	1.0000	0.0000
Flickr → ROCO	10	1.0000	0.5000	1.0000	0.0000
Flickr → ROCO	20	1.0000	0.5000	1.0000	0.0000
Flickr → COCO50K	2	1.0000	0.5000	1.0000	0.0000
Flickr → COCO50K	5	1.0000	0.5000	1.0000	0.0000
Flickr → COCO50K	10	1.0000	0.5000	1.0000	0.0000
Flickr → COCO50K	20	1.0000	0.5000	1.0000	0.0000
ROCO → Flickr	2	1.0000	0.5833	0.8970	0.0000
ROCO → Flickr	5	1.0000	0.5833	0.9095	0.0000
ROCO → Flickr	10	1.0000	0.5833	0.9061	0.0000
ROCO → Flickr	20	1.0000	0.5833	0.9012	0.0000
ROCO → COCO50K	2	1.0000	0.5938	1.0000	0.0000
ROCO → COCO50K	5	1.0000	0.5938	1.0000	0.0000
ROCO → COCO50K	10	1.0000	0.5938	1.0000	0.0000
ROCO → COCO50K	20	1.0000	0.5938	1.0000	0.0000
COCO50K → Flickr	2	0.0278	0.3750	0.7255	0.9400
COCO50K → Flickr	5	0.0278	0.3750	0.7626	1.0000
COCO50K → Flickr	10	0.0278	0.3750	0.7879	1.0000
COCO50K → Flickr	20	0.0278	0.3750	0.7550	1.0000
COCO50K → ROCO	2	0.9688	0.5000	0.8082	0.0200
COCO50K → ROCO	5	0.9688	0.5000	0.8274	0.0100
COCO50K → ROCO	10	0.9688	0.5000	0.8332	0.0000
COCO50K → ROCO	20	0.9688	0.5000	0.8325	0.0000

Table 19: Calibration sample-size study under cross-domain transfer. A small number of labeled target-domain reference models is often sufficient to recover useful calibrated decisions. COCO50K → Flickr exposes a strong score-orientation reversal, reflected by high flip rate, while COCO50K → ROCO shows partial threshold recovery with 5–10 calibration references.

Item	Value
Dataset	COCO2014 50K / 5K
Teacher	Clean vs. marked CLIP RN50
Student	CLIP RN50
Models	8 clean-distilled + 8 marked-distilled
Epochs	15
Fresh contrastive weight	0.5
Detector	Weight-space detector

Table 20: Mixed-distillation setup. Students are trained with both teacher matching and fresh COCO contrastive learning.

Setting	AUC	Acc	CV AUC	CV Acc	Utility Gap
Mixed distillation	0.9688	0.8750	0.9375	0.9375	−0.0017

Table 21: Mixed-distillation detection result. The forensic trace remains detectable even when student training includes fresh COCO contrastive loss.

Feature Group	Features	AUC	Acc	CV AUC
Projection/head	892	1.0000	1.0000	1.0000
Norm	4708	0.8750	0.7500	0.8750
All features	10078	1.0000	1.0000	0.9375

Table 22: Feature-group results for mixed distillation. Projection/head features remain the strongest forensic source under fresh-data mixing.

hicles/outdoor, and food/kitchen, suggest the mark creates structured, domain-correlated perturbations rather than uniform noise. This serves two purposes. First, it provides independent behavioral confirmation that the weight-space forensic signal reflects a real change in model computation, not a statistical artifact of the feature extraction. Second, the structured nature of the drift (concentrated in specific semantic categories rather than spread uniformly) is consistent with the theoretical prediction that the mark reshapes alignment geometry in a coherent direction rather than adding random noise. We emphasize that this is evidence of model distortion, not data reconstruction: we observe that the mark changes what the model represents, not what data it memorized.

J Qwen-VL Breadth Experiments

To test whether the forensic behavior observed in CLIP generalizes beyond a single architecture family, we evaluated Qwen2.5-VL-3B under frozen-feature fine-tuning settings. Unlike the CLIP experiments, these runs use a large multimodal transformer with separate image/text alignment components and substantially different optimization dynamics.

We evaluated two settings: (i) tiny frozen-feature adapter tuning and (ii) frozen-feature projection/head fine-tuning. The goal of these experiments is not to establish a new state-of-the-art forensic benchmark, but to evaluate whether the same alignment-localized forensic trend appears in a structurally different VLM.

J.1 Frozen-Feature Adapter Experiments

Table 24 summarizes the adapter-only Qwen experiments. Scaling from 8+8 to 16+16 independently trained models substantially improved stability, particularly on ROCO.

Although Qwen-VL exhibits greater seed sensitivity than CLIP RN50, the overall trend remains consistent: Larger model pools and longer optimization improve convergence toward stable forensic geometry while preserving downstream retrieval utility.

J.2 Projection/Head Fine-Tuning Experiments

We next evaluated whether the forensic signal survives beyond tiny adapter-only tuning. Instead of training small adapters, these experiments fine-tune the full-rank image/text projection heads while keeping the underlying Qwen-VL backbone frozen.

Feature-group analysis shows that projection/head features remain consistently informative, although localization is less clean than CLIP because norm-derived features also carry partial signal.

Overall, these experiments support a conservative breadth claim: The forensic phenomenon is not unique to CLIP RN50 and can appear in larger multimodal transformer architectures. However, CLIP remains the cleaner and more stable mechanistic setting, particularly for projection/head localization and cross-domain transfer.

K Posthoc Weight-Evasion Robustness

A practical forensic system should remain stable under mild posthoc weight modifications. We therefore evaluated whether simple evasion strategies could suppress the forensic signal without substantially degrading downstream utility.

We considered three perturbation families: (i) additive Gaussian noise, (ii) weight smoothing, (iii) low-rank perturbations.

The perturbations were applied directly to trained marked CLIP RN50 weights after optimization. Detector performance was then reevaluated without retraining the forensic classifier.

Across all tested perturbation families, the forensic detector remained fully separable under mild perturbations. Importantly, the strongest smoothing attacks began to damage downstream utility before reducing detector performance, indicating that simple post hoc weight manipulation does not easily erase the forensic geometry while maintaining model quality.

These results support the interpretation that the forensic signal is distributed across shared alignment structure rather than isolated to a small set of fragile parameters. Consequently, naive perturbation-based evasion appears substantially less effective than directly retraining the alignment geometry itself.

K.1 Adaptive Detector-Removal Baseline

The previous robustness experiments evaluate non-adaptive perturbations such as Gaussian noise, smoothing, and low-rank updates. We additionally test a stronger detector-aware baseline in which the attacker knows that the detector relies on projection/head features.

Starting from marked COCO50K CLIP RN50 models, we fine-tune only the projection/head weights toward the clean reference centroid while preserving COCO contrastive utility. This is not full adversarial retraining, but it tests whether a simple detector-aware update can erase the forensic direction without damaging the model.

The adaptive centroid-removal attack fails to evade the detector: AUC and accuracy remain 1.0000 for all tested strengths. Moreover, the attack is not stealthy, since utility drops by roughly 4–5%, exceeding the 2% preservation target even at the weakest setting.

This strengthens the robustness claim beyond random perturbations. In this controlled COCO50K RN50 setting, simple detector-aware projection/head removal does not erase the forensic signal without measurable utility degradation. The result remains a bounded adaptive baseline rather

Run	Feature Group	Features	Held-out AUC	Held-out Acc	CV AUC	CV Acc
Qwen projection/head seed99	projection/head-only	60	0.9531	0.8125	0.8047	0.7500
Qwen projection/head seed99	norm-only	88	0.7812	0.6250	0.8750	0.7500
Qwen projection/head seed99	all trainable head features	148	0.9844	0.6875	0.8750	0.8125
Qwen projection/head seed2027	projection/head-only	60	0.6562	0.6250	0.8906	0.7188
Qwen projection/head seed2027	norm-only	88	0.6484	0.6250	0.8594	0.7500
Qwen projection/head seed2027	all trainable head features	148	0.6953	0.6250	0.8750	0.7500

Table 23: Qwen-VL frozen-feature projection/head feature-group results. Qwen provides optional breadth evidence, but localization is less clean than CLIP because norm-only features also carry signal.

Dataset	Seed	Models	Held-out AUC	Held-out Acc	CV AUC	CV Acc	Clean R@1	Poison R@1	Utility Gap
Flickr30k	seed99	8+8	0.9375	0.6250	1.0000	0.9375	0.1419	0.1263	0.0156
Flickr30k	seed2027	8+8	0.2500	0.2500	0.7500	0.6250	0.1394	0.1406	−0.0013
ROCO	seed99	8+8	0.6875	0.6250	0.9375	0.8125	0.0294	0.0288	0.0006
ROCO	seed2027	8+8	1.0000	0.8750	0.9375	0.9375	0.0394	0.0369	0.0025
Flickr30k	seed99	16+16	0.8438	0.7500	0.9297	0.8125	0.1425	0.1338	0.0087
Flickr30k	seed2027	16+16	0.6719	0.7500	0.8281	0.6562	0.1425	0.1375	0.0050
ROCO	seed99	16+16	0.8359	0.8125	0.9688	0.8750	0.0275	0.0316	−0.0041
ROCO	seed2027	16+16	1.0000	1.0000	1.0000	0.9688	0.0403	0.0369	0.0034

Table 24: Frozen-feature Qwen-VL adapter experiments. Scaling from 8+8 to 16+16 models substantially improves forensic stability, particularly on ROCO.

than a proof against fully optimized adversarial retraining.

L Implementation Details

L.1 Model Architectures

The primary experiments use CLIP RN50 as the main forensic testbed. Additional breadth experiments use: (i) CLIP ViT-B/16 on COCO50K, (ii) CLIP ViT-B/32 on ROCO and (iii) Qwen2.5-VL-3B under frozen-feature fine-tuning settings.

For Qwen-VL, the backbone model remains frozen while trainable components are restricted to: (i) lightweight adapters or (ii) full-rank image/text projection heads.

L.2 Datasets and Splits

The main experiments use Flickr30k, ROCO, and MS COCO2014.

Flickr30k and ROCO experiments primarily use:

- 500 training samples
- 200 validation samples

The large-scale COCO experiment uses:

- 50K training samples
- 5K validation samples

All clean and marked models are trained using identical dataset splits and optimization settings.

L.3 Model Counts

We evaluate multiple independently trained clean and marked model pools:

- 4+4 models for ResNet sanity validation
- 8+8 models for pilot CLIP/Qwen studies
- 12+12 models for early CLIP experiments
- 16+16 models for the main CLIP, COCO, robustness, and Qwen evaluations

Increasing model count consistently improves detector stability and cross-validation robustness.

L.4 Feature Extraction

Forensic detection operates entirely in weight space. For each trained model, we extract summary statistics from:

- projection/head layers
- normalization layers
- adapter layers (when applicable)

The extracted statistics include:

- parameter norms
- spectral statistics
- singular value summaries

Run	AUC	Acc	CV AUC	CV Acc	Utility Gap	Status
seed99	0.9844	0.8750	0.9531	0.8438	-0.0009	Strong
seed2027	0.7578	0.7500	0.9375	0.7812	0.0103	Moderate

Table 25: Qwen-VL frozen-feature projection/head fine-tuning results on ROCO. The forensic signal survives beyond tiny adapters and remains detectable under full-rank alignment-head optimization.

Attack	Strength	AUC	Acc	Utility Drop	Verdict
Gaussian	1×10^{-5}	1.0000	1.0000	0.0004	Robust
Gaussian	5×10^{-5}	1.0000	1.0000	0.0000	Robust
Gaussian	1×10^{-4}	1.0000	1.0000	-0.0006	Robust
Gaussian	5×10^{-4}	1.0000	1.0000	0.0026	Robust
Gaussian	1×10^{-3}	1.0000	1.0000	0.0083	Robust
Smoothing	0.01	1.0000	1.0000	0.0031	Robust
Smoothing	0.05	1.0000	1.0000	0.0134	Robust
Smoothing	0.10	1.0000	1.0000	0.0403	Weakened utility
Smoothing	0.20	1.0000	1.0000	0.0990	Weakened utility
Low-rank ($r = 1$)	1×10^{-4}	1.0000	1.0000	0.0001	Robust
Low-rank ($r = 1$)	5×10^{-4}	1.0000	1.0000	-0.0003	Robust
Low-rank ($r = 1$)	1×10^{-3}	1.0000	1.0000	-0.0004	Robust
Low-rank ($r = 4$)	1×10^{-4}	1.0000	1.0000	0.0004	Robust
Low-rank ($r = 4$)	5×10^{-4}	1.0000	1.0000	0.0001	Robust
Low-rank ($r = 4$)	1×10^{-3}	1.0000	1.0000	0.0004	Robust
Low-rank ($r = 8$)	1×10^{-4}	1.0000	1.0000	0.0003	Robust
Low-rank ($r = 8$)	5×10^{-4}	1.0000	1.0000	0.0001	Robust
Low-rank ($r = 8$)	1×10^{-3}	1.0000	1.0000	0.0006	Robust

Table 26: Posthoc weight-evasion robustness experiments. The forensic detector remains fully stable under Gaussian and low-rank perturbations. Strong smoothing begins to reduce downstream utility before detector performance degrades, suggesting that mild posthoc perturbations are insufficient to remove the forensic signal while preserving model quality.

Item	Value
Dataset	COCO2014 50K / 5K
Base model	CLIP RN50
Base condition	`adapter_encoding`, $\lambda = 2.0$
Attacked models	8 marked models
Edited weights	Projection/head only
Adaptive objective	Move toward clean reference centroid
Utility preservation	COCO contrastive loss
Detector	Projection/head-only detector

Table 27: Adaptive detector-removal setup. The attacker is given the detector feature family and attempts to remove the marked projection/head signature while preserving utility.

- entropy-based descriptors
- distribution moments

Feature-group experiments separately evaluate: (i) adapter-only, (ii) norm-only, (iii) projection/head-only, and (iv) all-feature settings.

Strength	AUC	Acc	Utility Drop	Verdict
$\lambda = 0.01$	1.0000	1.0000	0.0496	Not evaded
$\lambda = 0.05$	1.0000	1.0000	0.0408	Not evaded
$\lambda = 0.10$	1.0000	1.0000	0.0439	Not evaded
$\lambda = 0.50$	1.0000	1.0000	0.0493	Not evaded
$\lambda = 1.00$	1.0000	1.0000	0.0523	Not evaded

Table 28: Adaptive detector-removal results. The detector remains perfectly separable under all tested removal strengths, while utility drops by roughly 4–5%.

L.5 Detector Configuration

Unless otherwise specified, the forensic detector uses a linear SVM with default hyperparameters from scikit-learn. Cross-validation uses stratified splits over independently trained clean and marked models.

We additionally evaluate:

- logistic regression
- random forests
- linear SVMs

on the main CLIP experiments to verify that the forensic signal is not classifier-specific.

L.6 Optimization Settings

The primary CLIP experiments use:

- AdamW optimization
- 15–16 training epochs
- mark strength $\lambda = 2.0$

Qwen-VL experiments use:

- 16 epochs
- adapter hidden size 128
- frozen-feature optimization

All clean and marked runs use identical optimization settings except for the marking objective.

L.7 Compute Budget

The experiments were conducted using multi-GPU training across independently trained model pools. The largest run trains 32 total CLIP RN50 models (16 clean + 16 marked) on the 50K-scale MS COCO subset.

Feature extraction, detector training, calibration, robustness analysis, and visualization pipelines are lightweight relative to model training and run on standard CPU/GPU infrastructure.

L.8 Reproducibility

All experiments use fixed random seeds when reported. For Qwen-VL, we explicitly evaluate multiple seeds due to increased optimization sensitivity.

All reported detector evaluations use held-out testing or stratified cross-validation unless otherwise specified.

M Extended Theoretical Discussion

M.1 Relaxing the Linear Projection Assumption

Proposition 1 assumes $P(\theta) = W\theta$ is linear. For nonlinear projections $P(\theta) = W\theta + g(\theta)$ with $\|g\| \ll \|W\theta\|$ near the optimum, the shared displacement is $\boldsymbol{d} \approx W^{\dagger}\boldsymbol{v}$ plus a higher-order correction. The empirical success of our method suggests this approximation is adequate for CLIP projection heads, which are linear or nearly linear in practice.

M.2 Connection to Uchida et al. Watermarking

The regularizer in Eq. 3 is structurally similar to the embedding loss in Uchida et al. (2017): $L_w = \|X\theta - \boldsymbol{b}\|^2$, where X is a secret projection matrix and $\boldsymbol{b}$ is the watermark bit string. The key difference is that Uchida et al. verify by projecting weights through X and checking bits, while we detect via statistical features $\phi(\theta)$ without knowing the secret projection. Our theoretical framework explains why statistical detection succeeds: the mark induces a displacement large enough to be detectable through summary statistics even without access to the exact mark direction.

M.3 Why the Signal Often Concentrates in Alignment Layers

The empirical concentration of the forensic signal in projection/head layers follows naturally from the geometric interpretation of the marking objective. In CLIP-like architectures, projection layers form the cross-modal alignment bottleneck: image and text embeddings are forced into a shared semantic space where contrastive matching is optimized.

Under the linearized approximation from Proposition 1, the marking objective induces a shared displacement direction $\boldsymbol{d}$ in the alignment geometry. Although the perturbation may originate through adapters or auxiliary modules, optimization propagates the resulting geometric constraint toward the shared embedding interface where image-text similarity is computed.

This explains the CLIP RN50 results, where projection/head-derived features consistently dominate adapter-only and norm-only features. However, the ViT-B/16 COCO50K experiment shows that the exact localization pattern is architecture-dependent: the forensic signal remains perfectly detectable, but it is visible across adapter, norm, projection/head, and all-feature groups. Thus, projection/head localization should be interpreted as the dominant RN50 mechanism, not as a universal law across all CLIP backbones.

This supports four empirical observations:

1. CLIP RN50 projection/head-derived features consistently outperform other feature groups,
2. RN50 adapter-only features remain near chance even when the mark is injected through adapters,

3. the forensic direction transfers across datasets despite substantial domain shift,

4. and the signal survives controlled distillation into new models.

Together, these results suggest that the forensic trace is not merely a localized parameter artifact. It is best understood as a persistent distortion of shared alignment geometry, whose observed feature-group footprint depends on architecture and trainable-path design.

M.4 Connection to Distillation Inheritance

The distillation experiments provide additional evidence that the forensic signal behaves as inherited geometry rather than memorized parameter structure.

Let $f_T(x)$ denote the teacher embedding function and $f_S(x)$ the student embedding function learned through distillation. Under standard contrastive or representation-matching objectives, the student minimizes:

$$L_{\text{distill}} = \|f_S(x) - f_T(x)\|^2.$$

If the teacher embedding space contains a forensic displacement $\boldsymbol{d}$ induced by the marking objective, then minimizing the distillation loss implicitly transfers this displaced geometry into the student representation space.

Consequently, the student inherits:

- the marked embedding structure,
- the shifted similarity geometry,
- and the associated forensic direction.

This interpretation is strongly supported by the empirical results: projection/head-derived features remain highly separable after distillation, while cross-model similarity maps show that distilled models preserve the marked alignment structure rather than converging toward an independent geometry.

Thus, the forensic signal appears to persist at the level of shared semantic organization rather than exact parameter reuse.

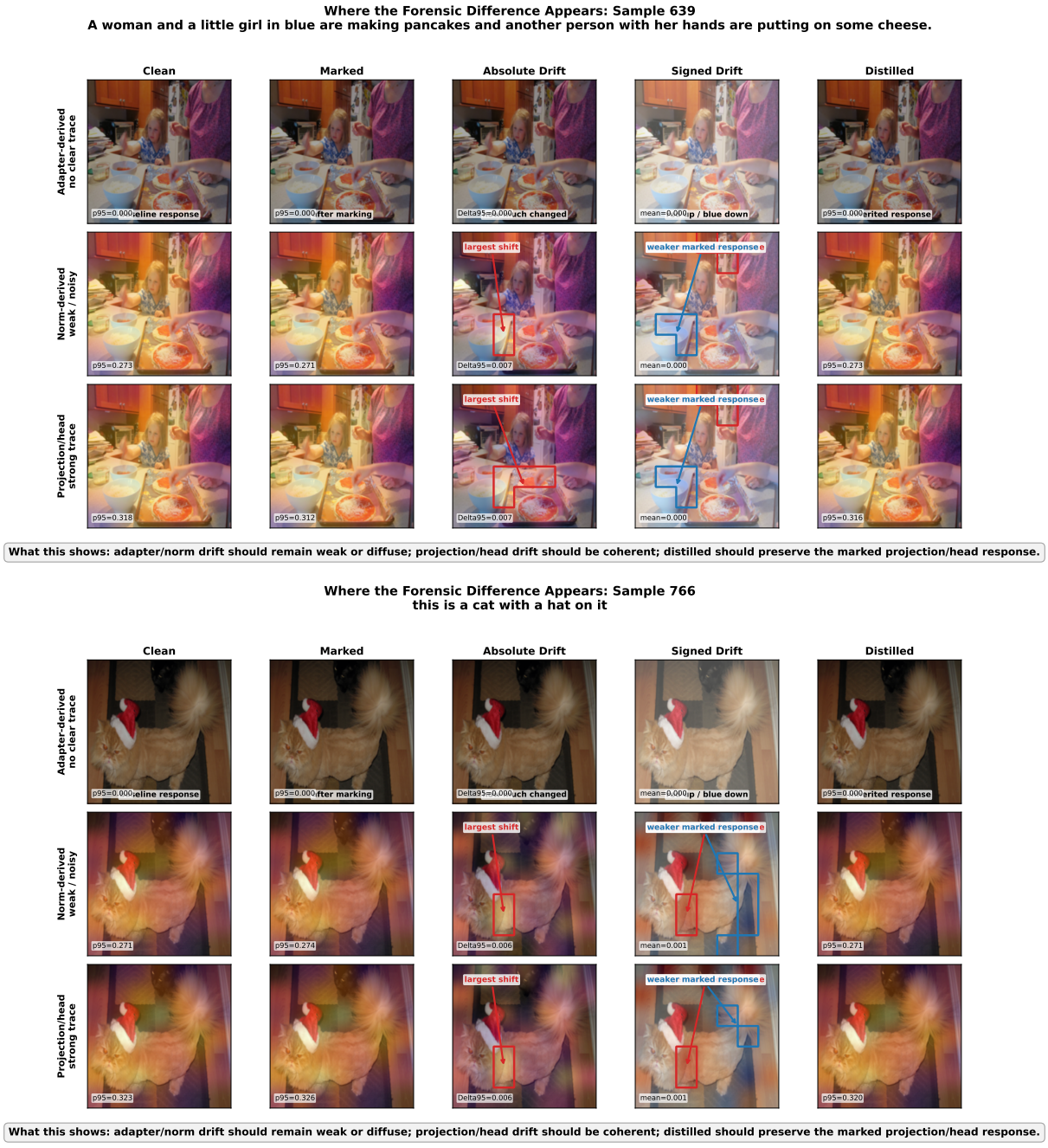

Figure 4: Visualization of forensic representation drift across feature groups. Rows correspond to adapter-derived, norm-derived, and projection/head-derived features. Columns show clean responses, marked responses, absolute drift magnitude, signed drift direction, and distilled-model responses. Adapter and norm features exhibit weak or diffuse drift, while projection/head features show coherent structured changes after marking. The distilled model preserves similar projection/head response geometry, supporting the claim that the forensic signal survives through inherited alignment structure rather than visible image perturbations.

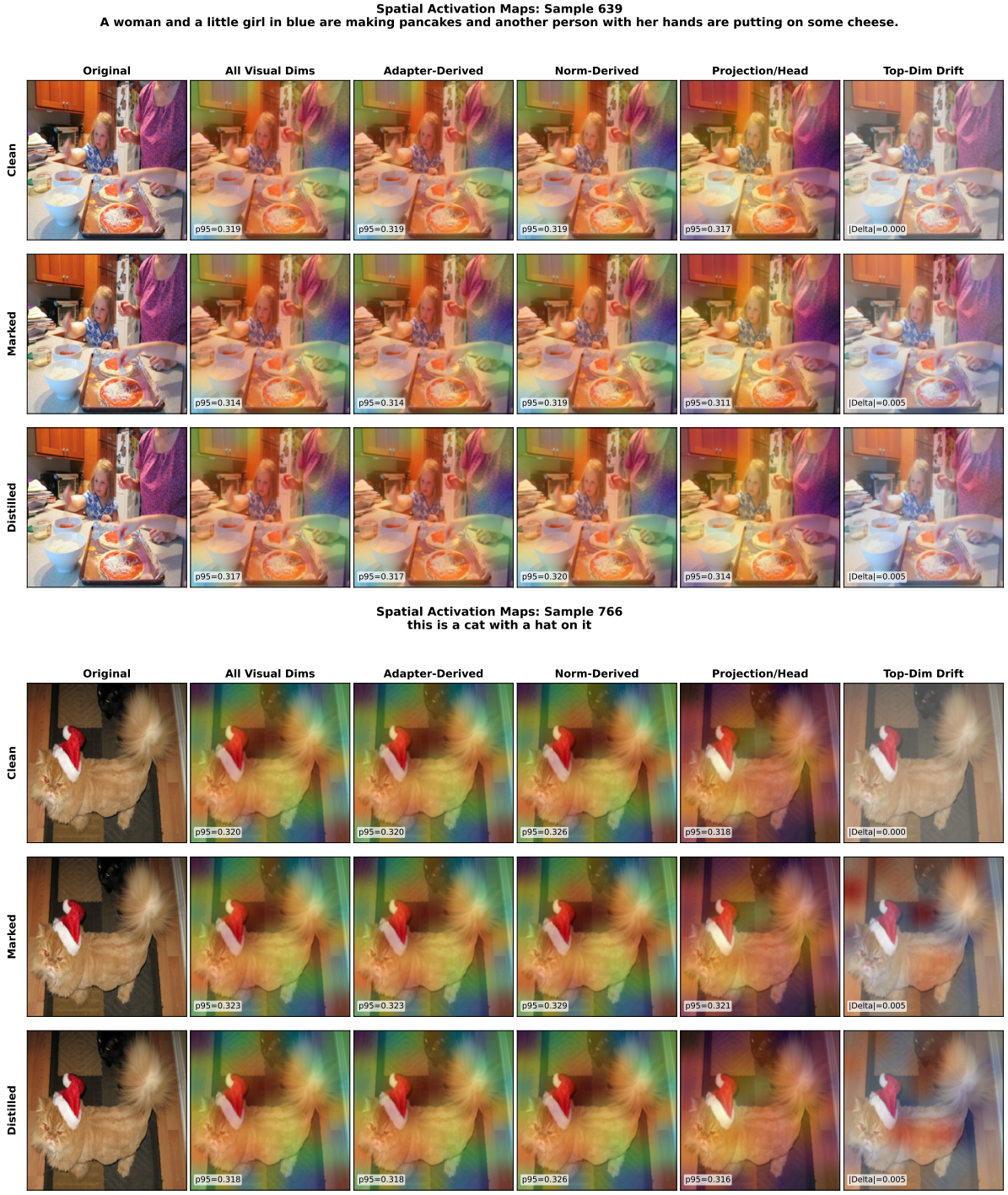

Figure 5: Spatial activation maps showing where forensic representation drift appears across the image. Rows correspond to adapter-derived, norm-derived, and projection/head-derived feature groups. Projection/head responses exhibit coherent structured drift after marking, while adapter and norm responses remain weak or spatially diffuse. These results support the claim that the forensic signal manifests primarily in cross-modal alignment geometry rather than in the injected adapter parameters themselves.

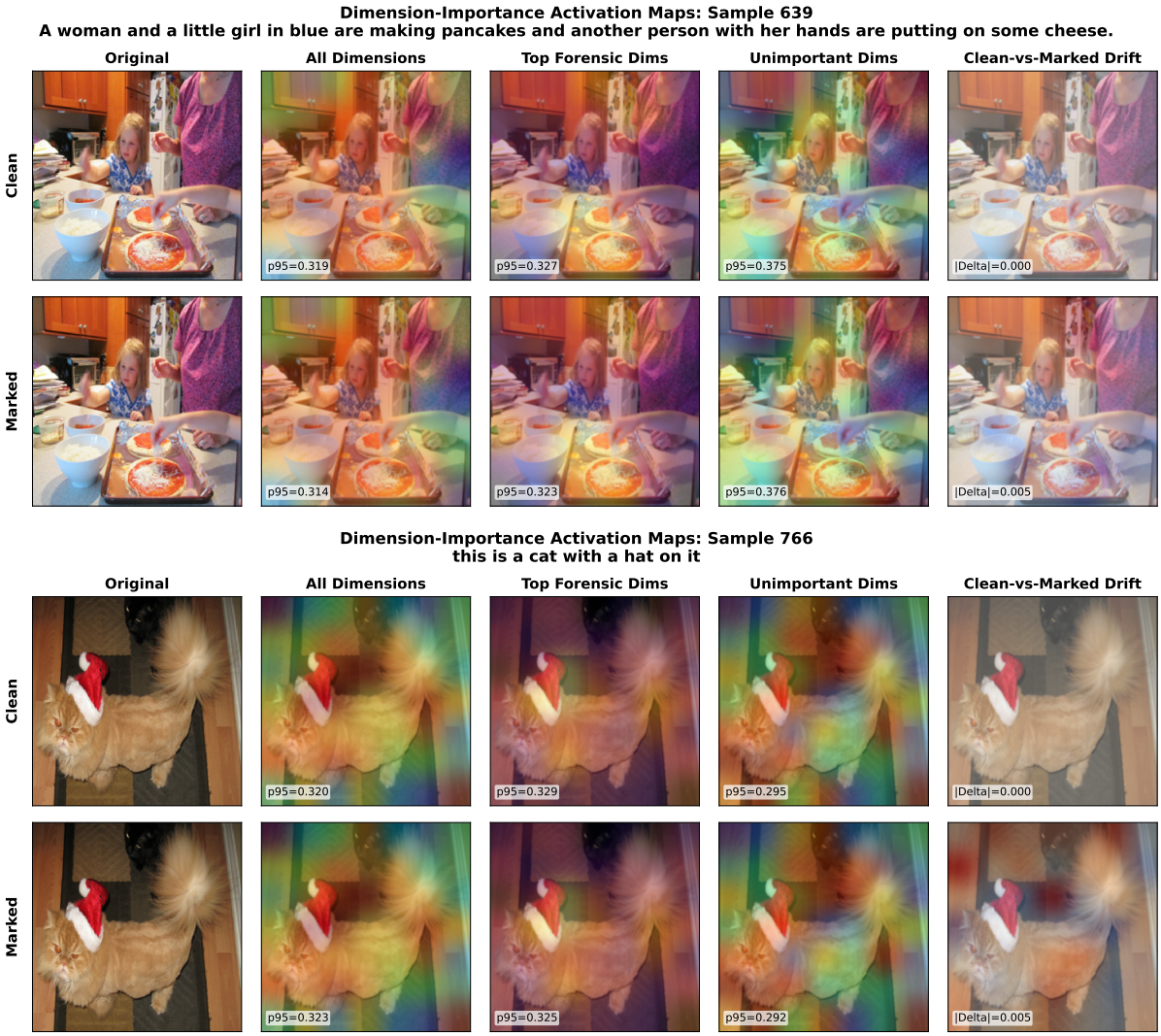

Figure 6: Dimension-importance activation maps visualizing which latent dimensions contribute most strongly to the forensic signal. Top forensic dimensions produce coherent structured drift between clean and marked models, whereas low-importance dimensions remain diffuse and weakly informative. The distilled model preserves similar high-importance response patterns, supporting the interpretation that the forensic trace survives through inherited alignment geometry.