EVIDENCE IS ALL YOU NEED—BUT NOT FOREVER: OPEN-WORLD CONTINUAL IMAGE FORENSICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image forensics underpins multimedia security. Research is moving from detection under fixed conditions toward open-world forensics, where generators and image processing conditions continually change. Their joint effects can alter how visual features distinguish real from fake images, so past decision rules may become unreliable. Broader training coverage and continual updates improve detection. However, a theoretical gap remains: what should representations retain to reduce label requirements for future adaptation as decision rules change? To address this gap, we propose a new research framework and establish key theoretical guarantees. In a Gaussian model, we prove that coordinates learned from finite history reduce adaptation labels compared with retaining optimal past decision rules alone. These coordinates retain information that past decisions omit and future learning can use. Our bounds separately account for learning the coordinates from history and reaching a specified detection quality with current labels. We establish *shared-space sufficiency*, identifying when a shared space preserves full-space optimal AUROC as generation and processing change together. These results motivate *Continual Updating of Evidence* (CUE). With its visual encoder and initial detector fixed, CUE estimates within-class variation from historical and current statistics; current labels determine the class difference. One statistical objective yields the coordinates and classifier. We bound how statistical estimation error affects the classifier. Adding or subtracting window statistics supports continual learning and unlearning. Verified labels separately determine when fixed score contributions are suspended or restored. With 16 labels per class, CUE achieves state-of-the-art AUROC against 22 pretrained detectors across six standard benchmarks. CUE further turns accumulated experience into more data-efficient adaptation to new conditions. Together, these results establish a learning-theoretic foundation for open-world forensics.

1 INTRODUCTION

Image forensics distinguishes real images from AI-generated or manipulated ones. In open-world forensics, detectors face unseen generators and changing processing conditions, including compression, transcoding, and re-digitization (Baracchi et al., 2024; Park & Owens, 2025; Li et al., 2025a). Robustness to either change alone does not ensure reliable detection when both occur together. The challenge is to learn from finite history while both continue to change (Figure 1).

Methods detect unseen generators using diffusion reconstruction errors or pretrained visual features (Wang et al., 2023; Ojha et al., 2023; Guillaro et al., 2025). Datasets spanning thousands of generators support training on diverse sources (Yang et al., 2025; Park & Owens, 2025; Qin et al., 2026). Few-shot methods adapt to unseen generators with limited labeled examples (Wu et al., 2025; Wang et al., 2026). A feature is forensic evidence if it distinguishes real from fake images under current conditions. *Its value as evidence can change even when the feature remains detectable.* However, better detection alone does not explain what a representation should retain to reduce the labels needed for future adaptation.

We ask when finite history can yield a representation that reduces future label requirements. Can those requirements depend on a shared space's dimension without growing with the number of generation–processing combinations? Optimal past decisions may omit information useful for future learning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

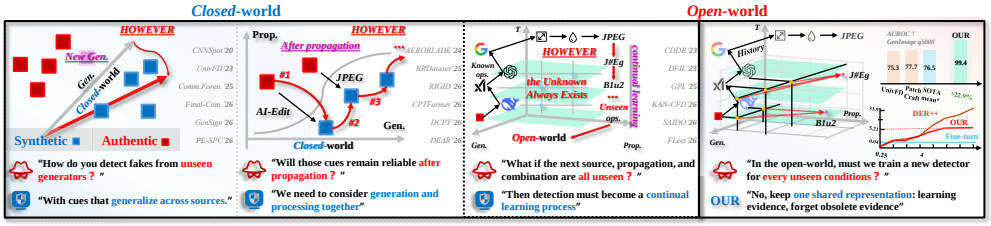


Figure 1: Open-world forensic learning: reusable coordinates for changing decisions.

In a Gaussian model, we prove that coordinates learned from finite history reduce adaptation labels compared with retaining optimal past decision rules alone. These coordinates encode a historical estimate of a direction omitted by past decisions but useful under new conditions. Separate bounds quantify historical data for coordinate learning and current labels for a specified detection quality. We establish **shared-space sufficiency**, identifying when a shared space preserves full-space optimal AUROC under joint changes. The label bound depends on its dimension rather than the number of generator–processing combinations. Under suitable identifiability conditions, more accurate historical statistics tighten our bound on generalized-eigenvector coordinate estimation error.

Our model, **Continual Updating of Evidence (CUE)**, uses historical statistics to calibrate coordinates for each new decision. It combines frozen visual features with an initial detector's score, pooling their joint statistics across historical and current observations to estimate within-class variation. Current labels determine the class difference. One squared-error objective yields both the coordinates and classifier, and we bound the classifier's sensitivity to statistical estimation error. Adding or subtracting window statistics updates both, supporting continual learning and unlearning. Audits separately determine whether fixed score contributions remain active, are suspended, or are restored.

With Community Forensics features and 16 labels per class, CUE **achieves state-of-the-art AUROC in the evaluated comparison**, exceeding the 22 compared pretrained detectors on all six standard benchmarks. With DINOv2 features, fixed current support, and the same update rule, historical statistics improve mean normalized area under the label-budget curve by 1.66 percentage points across four conditions. Both fixed front ends use the same statistical learning rule. Appendix A distinguishes these configurations from neural-adaptation and stored-classifier reference results.

We make three contributions.

- **Research framework.** We formulate finite-history learning to reduce future label requirements under joint generation and processing changes (Section 3.1).
- **Learning theory.** We prove a finite-history label advantage over optimal past decisions in a Gaussian model and characterize shared-space sufficiency under joint changes (Section 3.2).
- **Model and evaluation.** CUE supports continual learning and unlearning through one objective. It exceeds the compared pretrained detectors with few labels; matched controls show a benefit from history (Sections 4–5).

2 Related Work

Forensics under changing generation and processing conditions. Generalizable detectors use pretrained visual features, reconstruction errors, structural cues, and training that reduces dataset bias (Ojha et al., 2023; Wang et al., 2023; Tan et al., 2024; Guillaro et al., 2025). Datasets spanning many generators broaden source coverage (Yang et al., 2025; Park & Owens, 2025; Qin et al., 2026), while benchmarks evaluate detection under realistic processing changes (Zhu et al., 2023; Li et al., 2025b; Du et al., 2025). Few-shot and continual detectors adapt to unseen generators (Li et al., 2023; Pan et al., 2023; Wu et al., 2025; Wang et al., 2026).

Shared representations and selective forgetting. Representation-learning theory explains how sharing low-dimensional structure across related tasks can improve transfer (Tripuraneni et al., 2021; Cao et al., 2022). Concept drift concerns changing input–label relations (Gama et al., 2014), while

continual learning seeks to preserve performance on earlier tasks (Kirkpatrick et al., 2017; Chaudhry et al., 2019). Recoverable forgetting allows reuse of retained capabilities (Ye et al., 2022; Zhao et al., 2024). CUE adds or removes retained statistics to update its coordinates and classifier, and separately suspends or restores fixed score contributions.

3 Learning from History in Open-World Forensics

3.1 Open-World Setting and Learning Objective

Changing generation and processing. Open-world forensics retains binary real/fake labels as sources and processing change; unlike open-world recognition, it adds no semantic classes (Bendale & Boult, 2015; Fontanel et al., 2021). An image $X \in \mathcal{X}$ has label $Y \in \{0, 1\}$, where $Y = 1$ denotes generated or manipulated content. A forensic world $\omega = (g, e)$ combines a real/fake source pair $g = (g_0, g_1)$ (e.g., cameras and generators) with a label-preserving processing and propagation chain e. Robustness to generation or processing changes separately need not extend to their combination (Appendix E).

What should history retain? Historical learning should reduce current label requirements for a given detection quality. With a frozen visual encoder ϕ, historical labeled observations $\mathcal{H}$ yield a reusable representation $\mathcal{R}(\mathcal{H})$. Adaptation fits a score on this representation with support $\mathcal{A}_\omega$: K real and K fake images ($2K$ current labels). We count historical observations separately.

Novelty across sources and processing. Let $\mathcal{G}_0$ and $\mathcal{E}_0$ contain development source pairs and propagation chains. The balanced world model has $P_\omega(Y = y) = \frac{1}{2}$ and $P_\omega(X \mid Y = y) = P(X \mid Y = y, g_y, e)$. Define

$$q(\omega) = \big(\mathbf{1}[g \notin \mathcal{G}_0], \mathbf{1}[e \notin \mathcal{E}_0]\big), \qquad Q_{ab} = \{\omega : q(\omega) = (a, b)\}. \tag{1}$$

Only sources are new in Q_{10}, only processing in Q_{01}, both in Q_{11}, and neither in Q_{00}.

Learning as conditions change. After scoring queries, the detector updates from a separate *audit batch* of at most K verified images per class. An *audit window* retains one batch's information, indexed by arrival order. Windows may combine worlds; worlds may recur across windows. The detector observes audit arrivals and authenticity labels, without source or processing identities or change indicators. Query labels are available only for evaluation. Few-shot evaluation measures detection after adaptation to a fixed support set (Section 5).

3.2 A Provable Reduction in Adaptation Labels

In a Gaussian family, we prove that finite history reduces current label requirements by preserving information omitted by optimal past decisions. In this construction, $X \in \mathbb{R}^d$ is a feature vector and $y = 2Y - 1 \in \{-1, +1\}$ its signed label.

Information beyond past decisions. For $d > 4$, let $j \in \{-1, +1\}$ index two historical Gaussian distributions with unknown orthogonal unit directions u, w. Their class means $\pm(u + jw)$ contain both directions. We choose class-shared positive-definite covariances with bounded condition numbers to give exact posterior log-odds $2jw^\top x$ (Appendix K.1). Optimal historical probabilities reveal w but omit u, which separates the current classes $\mathcal{N}(yu, I_d)$. Historical observations retain this direction: their balanced signed average cancels the opposing jw terms to estimate u.

A *rule-memory learner* retains exact historical posterior functions but no historical observations or marginal statistics. It may fit any measurable current score, including nonlinear ones, using all feature coordinates. The representation learner estimates a direction and normalization scales from finite history. Neither receives additional current unlabeled data or the current marginal covariance.

Learning reusable coordinates. Using T independent historical observations balanced across distributions and classes, we estimate

$$\widehat{u}_T = \frac{1}{T} \sum_{i=1}^{T} y_i X_i, \qquad \mathbb{E}\|\widehat{u}_T - u\|_2^2 = \frac{d+1}{T}. \tag{2}$$

Normalize $\widehat{v}_T = \widehat{u}_T/\|\widehat{u}_T\|_2$. An independent, balanced historical sample estimates the mean squared feature norm $\widehat{D}$ and mean squared projection $\widehat{\kappa}$ along $\widehat{v}_T$. We retain raw features and add the historical projection:

$$\mathcal{R}_T(x) = \left[x^\top/\sqrt{\widehat{D}}, \quad x^\top \widehat{v}_T/\sqrt{\widehat{\kappa}}\right]^\top. \tag{3}$$

Ridge regression fits current labels in these coordinates. Raw features preserve the linear score class; the added projection lowers the effective ridge penalty along the historical direction (Equation 115). The direction and scales require $O(d)$ storage. With two current labels, the AUROC lower bound is nondecreasing with history (Corollary 3).

Proposition 1 (Finite history reduces adaptation labels)**.** *In the Gaussian family above with $d = 1024$, the coordinates in Equation 3 allow the ridge readout specified in Appendix K to reach expected population AUROC 0.75 with **two current labels**, one per class. The construction uses $20{,}480$ independent historical observations: $16{,}384$ to estimate the direction and $4{,}096$ to estimate its normalization scales. Any rule-memory learner needs **at least** 202 **current labels in total**, balanced between the classes to guarantee the same target, even when granted w. Both guarantees are uniform over unknown orthogonal unit directions u, w. The constructive expectation includes historical and current sampling; evaluation uses independent future examples after fitting the score.*

Shared-space sufficiency. When can a shared space remain sufficient as generation and processing change together? Assume Gaussian classes with a common covariance within each world and mean differences in a shared r-dimensional space. This alone is insufficient: directions outside the space can help cancel correlated noise. If source covariances, processing operators, and added-noise covariances preserve the space and its orthogonal complement, we prove that it attains the full features' optimal AUROC. Corollary 4 separates historical coordinate error from errors in current class means and covariance estimated from the same labeled support. With sufficiently accurate historical coordinates and a fixed signal range bounded away from zero, the label bound depends on r and the target AUROC gap. It has no term for generator–processing combinations (Appendix L).

To bound this historical coordinate error, we analyze a generalized-eigenvector estimate of the shared space. Under identifiability, small-error, and retained-rank conditions, we bound its subspace error by historical moment errors (Appendix M). For fixed windows with balanced Gaussian samples, the bound is nonincreasing as samples per class in each window grow.

CUE (Section 4) likewise learns coordinates from history and fits decisions to current labels, using a different construction: whitening a regularized within-class covariance and solving one joint-moment objective. Proposition 2 bounds this classifier's sensitivity to joint covariance estimation error with current class means and regularization fixed.

4 METHODOLOGY

CUE uses history to estimate within-class variation and current labels to determine the class difference. One squared-error objective yields calibrated coordinates and a closed-form classifier; adding or removing retained statistics updates both (Figure 2).

4.1 LEARNING FORENSIC COORDINATES

A frozen extractor with fixed preprocessing and input projection maps images to $z \in \mathbb{R}^d$; the frozen initial detector supplies $a(x)$. For fixed $\kappa > 0$, write $g(x) = a(x)/\kappa$ and $y = 2Y - 1$, with $Y = 1$ for fake images. CUE adds $z^\top w + b$ to g, fitting the residual $y - g$ from joint statistics of z and g. M_B fixes the regularization geometry at initialization. The fixed orthonormal columns of R specify directions available to the correction; $R = I$ allows all directions.

For each class and audit window, CUE retains the count, vector sum and outer-product sum of $u = [z; g]$. Current support provides K verified images per class ($n = 2K$); its class means define midpoint $\mu = (\mu_z, \mu_g)$ and half-difference $\delta = (c, c_g)$. The within-class covariance W pools history and current support, centering each class within its own window (Appendix B.1).

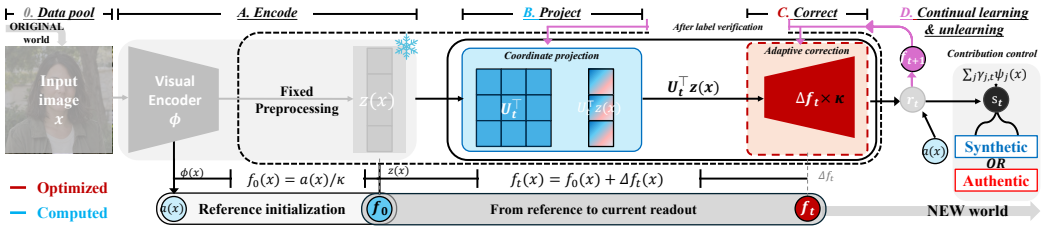


Figure 2: CUE (A–D): updated coordinates and correction, with fixed encoding and preprocessing.

These moments define a balanced two-class model with conditional means $\mu + y\delta$ and covariance W; $\widehat{\mathbb{E}}$ denotes expectation under this model. For fixed $\lambda > 0$ and positive-definite M_B, CUE minimizes

$$\min_{w \in \mathrm{range}(R),\, b} \widehat{\mathbb{E}}\big[(y - g - z^\top w - b)^2\big] + \alpha w^\top M_B w, \qquad \alpha = \lambda/n. \tag{4}$$

The risk depends only on these moments and requires no Gaussian assumption.

Partition W into feature and initial-score blocks. A Cholesky factorization gives a basis U satisfying

$$G = W_{zz} + \alpha M_B, \qquad U^\top G U = I, \qquad \mathrm{range}(U) = \mathrm{range}(R). \tag{5}$$

Thus $U^\top z$ measures features on a scale set by within-class variation and regularization (Figure 3B). Writing $\tilde{c} = U^\top c$ and $q = U^\top W_{zg}$, the fitted coefficients are

$$\beta = -q + \tilde{c}\,\frac{1 - c_g + \tilde{c}^\top q}{1 + \|\tilde{c}\|_2^2}, \qquad w = U\beta, \qquad b = -\mu_g - \mu_z^\top w. \tag{6}$$

Here W_{zg} is the within-class feature–score cross-covariance. The term $-q$ uses this covariance to minimize within-class variation of the corrected score plus regularization; the $\tilde{c}$ term trades this cost against fitting current labels. The intercept b centers the normalized score at zero in the balanced model. Appendix B.2 derives their joint optimum from Equation 4.

The Community Forensics configuration uses its pretrained features and official initial score; DINOv2 uses the same statistical update with a different fixed input representation. Appendix A separates these evaluations from the neural-adaptation, generalized-eigenvector and stored-classifier references.

4.2 Detection and Update

CUE stores classifiers with descriptors of their support feature distributions. It matches each unlabeled query batch to the closest stored descriptor and uses that classifier for all images, without query labels or source/processing identities (Appendix C). Without current support, the current classifier uses $w = b = 0$; the initial detector remains available.

Contribution control tests whether a fixed score term still helps distinguish real from fake images. Each contribution $\psi_j(x)$ combines a fixed image-pattern sensor and fixed calibrated weight; only its use indicator $\gamma_j \in \{0, 1\}$ changes. The correction basis R excludes directions reserved for these contributions. For the selected classifier (w_k, b_k), the score is

$$s(x) = a(x) + \kappa\,[z(x)^\top w_k + b_k] + \sum_j \gamma_j \psi_j(x). \tag{7}$$

With the selected classifier fixed, switching off γ_j removes exactly $\psi_j(x)$ while retaining the sensor and weight for restoration.

After detection, a separate audit supplies verified authenticity labels. Before fitting a new classifier, CUE selects an existing classifier for the audit and holds its pre-update score fixed while comparing mean binary cross-entropy with and without each contribution. Accumulated audit results and a false-positive check govern activation, suspension and restoration.

CUE then moves the previous current window into history, installs the new labeled window, and solves the same objective (Figure 3D). It saves the fitted classifier and its support descriptor for later predictions. Each observation enters current or historical statistics, never both; omitting history gives the matched current-only learner.

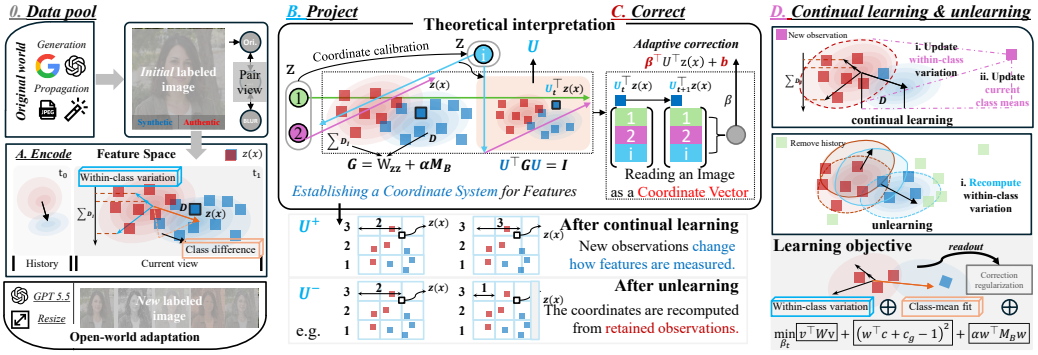


Figure 3: Coordinates (B), readout (C), and statistical updates (D). The displayed objective is Equation 4 minimized over b, with $w = U\beta$ and $v = [w; 1]$.

4.3 Removing Retained Observations

CUE removes a window by subtracting its counts and moment sums, recomputing W and solving the same objective. With current support and initialization fixed, we prove that this reproduces the classifier rebuilt from retained statistics in exact arithmetic (Proposition 5). Every directly or indirectly dependent classifier and descriptor is invalidated. Removing current support resets the current classifier to the initial detector; other valid classifiers remain available. Initialization and previous contribution audits remain unchanged. Appendix C.5 gives the dependency and reconstruction rules.

4.4 Statistical Estimation Guarantee

Proposition 2 (Classifier stability)**.** *Fix the current class means, α, M_B and an orthonormal R with at least one column. Let $\widehat{w}, w_*$ solve Equation 4 using positive-semidefinite joint covariances $\widehat{W}, W_*$, and set $E = \widehat{W} - W_*$. Then*

$$m = \lambda_{\min}\left[R^\top(\widehat{W}_{zz} + cc^\top + \alpha M_B)R\right], \qquad \|\widehat{w} - w_*\|_2 \le \frac{\|E\|_{\mathrm{op}}\sqrt{\|w_*\|_2^2 + 1}}{m}. \tag{8}$$

The fitted normalized core scores differ by $(z - \mu_z)^\top(\widehat{w} - w_*)$.

The bound links pooled feature and feature–score covariance errors to classifier error, with amplification quantified by $1/m$. Appendix B.3 gives the proof and score and squared-risk bounds.

5 Experiments

We evaluate detection under joint generation–processing changes, adaptation, retention, reversible contribution use, and deployment cost. Matched controls test the benefit of historical statistics.

5.1 Experimental Setup

Datasets and configurations. UFD and GenImage (Ojha et al., 2023; Zhu et al., 2023), AIGC-DB and AIGIBench (Zhong et al., 2023; Li et al., 2025b), and WildFake and Chameleon (Hong et al., 2025; Yan et al., 2025) supply 32 worlds covering all four seen/unseen source–processing combinations (Section 3). RRDataset (Li et al., 2025a) adds multi-platform transmission and re-digitization to continual evaluation. The 32-world study uses Community Forensics ViT-S/16; the twelve-world stream uses DINOv2-L features and image statistics. Table 10 and Appendix A specify each configuration and adaptation rule.

Baselines. We compare 22 benchmark detectors and use UnivFD (Ojha et al., 2023), NPR (Tan et al., 2024), and IAPL (Li et al., 2026) for few-shot comparisons. Continual baselines are frozen and fine-tuned models, SimpleShot (Wang et al., 2019), SLDA (Hayes & Kanan, 2020), RanPAC (McDonnell

et al., 2023), LayUP (Ahrens et al., 2024), budget-matched A-GEM (Chaudhry et al., 2019), and frozen SigLIP 2 So400m (Tschannen et al., 2025) with an updated linear head. Control baselines use always-on contributions or irreversible withdrawal.

Protocol. Benchmark comparisons follow their original protocols; adaptation comparisons share queries and audit budgets, with matched generator-first and propagation-first orders for continual streams. Budgets are K images per class ($2K$ total), with $K = 64$ in continual streams. Continual methods predict before each audit; few-shot queries follow support-set adaptation. Audits provide images and authenticity labels, without source or processing identities, change indicators, or query labels.

Metrics. Detection uses AUROC, AP, and FPR95. Post-audit and end-of-stream queries measure adaptation and retention; pooled AUROC summarizes all worlds. Control and deployment report evidence-state errors, transition delays, storage, and latency. Detection confidence intervals resample base-image identities. Appendices G and J give full budgets, audit/query assignments, and statistical procedures.

5.2 Detection and Few-Shot Adaptation

Table 1: **Detection across six standard benchmarks.** AUROC (%) for 22 baselines. CUE zero-shot: frozen Community Forensics (CF); few-shot: CF prototype/kernel adaptation with 16 labels per class per benchmark, including model selection. Gray: CUE; bold/underline: best/second-best per row.

	Trace-centric								Representation-centric							Diversity-centric							CUE	
Benchmark AUROC (%) ↑	CNN Det	DIRE	Patch Craft	Freq Net	NPR	DRCT	SAFE	Ferret Net	Univ FD	Omni AID	C2P-CLIP	RINE	Effort	IAPL	DDA	AIDE	CO-SPY	D^3	GAPL	Comm. Foren.	LOTA	DEAR	Zero-shot	Few-shot
UFD	91.2	56.6	76.8	91.5	80.2	68.7	80.9	90.6	93.8	95.7	70.3	98.1	81.0	98.3	90.7	92.9	83.2	97.6	**98.5**	97.6	68.6	77.1	97.6	98.4
GenImage	56.6	49.6	78.8	54.7	46.8	97.5	52.0	44.3	74.3	92.8	91.2	93.2	90.5	79.4	98.2	71.5	52.3	**99.6**	98.5	98.9	55.8	97.7	98.9	98.9
AIGC-DB	71.6	49.5	84.2	61.1	42.0	88.6	59.5	38.3	85.2	91.0	87.7	94.2	87.1	79.4	95.7	76.9	47.9	97.7	97.9	**99.1**	52.7	85.5	**99.1**	**99.1**
AIGIBench	67.2	50.9	80.3	70.3	72.0	85.1	82.5	86.2	67.8	93.4	77.2	75.6	82.2	75.4	91.2	71.1	77.5	89.5	95.7	96.0	78.5	81.2	96.0	**96.7**
WildFake	57.5	55.8	83.7	78.8	88.3	82.2	89.1	90.7	68.6	92.3	84.4	87.1	90.2	87.2	91.1	81.5	88.6	94.4	98.4	**99.4**	90.2	88.6	**99.4**	**99.4**
Chameleon	37.8	63.1	43.2	53.2	52.5	73.1	54.3	42.8	43.0	91.3	56.0	51.9	52.5	31.4	91.6	44.5	52.8	69.0	80.9	92.6	50.3	86.1	92.6	**93.7**

CUE reaches state-of-the-art benchmark performance. With 16 labels per class, joint-moment CUE exceeds all 22 pretrained detectors on shared query subsets of six benchmarks (Appendix A). The CF prototype/kernel reference ranks first (or tied first) on four benchmarks and second on two (Table 1; full metrics: Table 9).

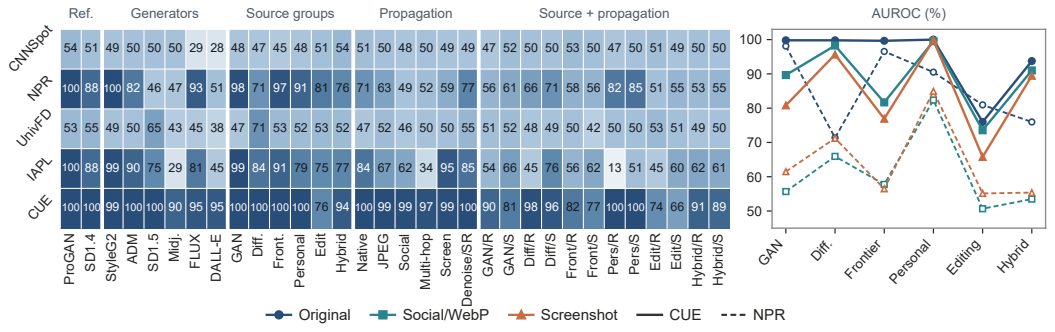


Figure 4: **Zero-shot detection across 32 forensic worlds.** Left: AUROC (%, rounded); darker cells indicate higher values. Right: frozen Community Forensics (CUE, solid) and NPR (dashed). R: Social/WebP; S: screenshot recapture.

Before audits, CUE's frozen CF ViT-S/16 scores 93.27 mean AUROC across 32 worlds, versus 70.12 for the strongest alternative; worst-world AUROC rises from 46.09 to 65.78 (Figure 4). Joint changes remain challenging: personalized generation stays near 100, but screenshot recapture lowers frontier-generation AUROC from 99.66 to 76.91. Full metrics: Appendix H.1 and Figure 7.

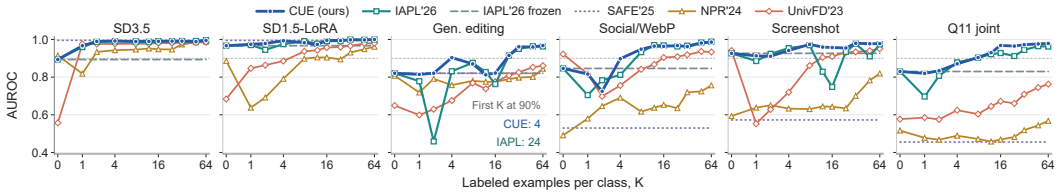


Figure 5: **Few-shot AUROC on six initially unseen conditions.** CUE: IAPL/CLIP neural-adaptation reference. Fixed queries are disjoint from nested audits (K labels per class; $2K$ total). Frozen IAPL and SAFE are references; editing-panel annotations mark the first tested K reaching 0.90 AUROC.

Few labels improve detection in initially unseen worlds. The neural-adaptation reference (Figure 5) improves macro AUROC from 88.05 without audits to 94.94 and 98.42 with 16 and 64 labels per class; worst-target AUROC rises from 82.07 to 96.75 at 64 labels per class. All targets first reach 0.90 AUROC by eight labels per class; editing requires four, versus 24 for IAPL (Appendix H.2).

History improves adaptation at matched current-label budgets. For DINOv2 CUE (Equation 4), current supports, initialization, calibration, and contribution control are fixed. Across four conditions and eight budgets ($K = 1$–128 per class), adding 256 historical labels in total raises the mean normalized AUROC–$\log_2 K$ area by **1.66 percentage points** (paired 95% CI: 1.36–1.99; Appendix A).

5.3 Continual Adaptation and Retention

Table 2: **Continual adaptation and retention.** Each audit supplies 64 labels per class. Gen./Prop.: generator-/propagation-first. Final AUROC (%); pooled: all 12 conditions; Max. drop: largest per-condition post-audit-to-final decline (percentage points). Full results: Table 14. Gray: CUE stored-classifier reference; bold/underline: best/second-best per order.

Metric / condition	Frozen		Fine-tune		SimpleShot		SLDA		RanPAC		LayUP		A-GEM		SigLIP 2		CUE	
AUROC (%) ↑	Gen.	Prop.	Gen.	Prop.	Gen.	Prop.	Gen.	Prop.	Gen.	Prop.	Gen.	Prop.	Gen.	Prop.	Gen.	Prop.	Gen.	Prop.
Initial condition	96.9	96.9	94.0	92.6	78.4	78.4	96.6	96.6	87.7	87.7	57.8	72.8	96.8	96.8	**99.1**	**99.2**	96.9	96.9
All conditions (pooled)	76.7	76.7	82.6	82.2	66.3	66.3	82.9	82.9	75.8	75.8	65.1	81.0	78.5	78.5	93.2	93.0	**94.3**	**94.3**
Max. drop (pp) ↓	**0.0**	**0.0**	5.9	4.6	0.7	0.7	0.2	2.4	0.3	1.0	25.2	11.7	1.2	0.7	2.0	2.4	**0.0**	**0.0**
Generator changes																		
SD3.5	73.4	73.4	87.6	88.8	95.0	95.0	85.7	85.7	90.4	90.4	72.3	93.6	80.7	79.7	93.1	92.4	**96.6**	**96.6**
FLUX.1	41.7	41.7	49.4	48.8	44.4	44.4	47.8	47.8	49.1	49.1	66.2	67.5	43.4	43.3	**86.8**	**87.0**	83.0	83.0
SD1.5-LoRA	87.3	87.3	91.8	92.5	65.4	65.4	91.8	91.8	79.2	79.2	69.5	86.1	88.2	88.4	**99.3**	**99.5**	**99.3**	99.3
SDXL-DB	90.7	90.7	92.1	98.1	92.9	92.9	97.6	97.6	95.7	95.7	74.1	96.8	90.2	90.5	99.6	99.8	**100.0**	**100.0**
Propagation changes																		
JPEG	93.9	93.9	93.9	91.3	83.3	83.3	93.6	93.6	85.3	85.3	63.0	83.3	93.8	93.9	**97.5**	**97.8**	96.6	96.6
Resize	92.7	92.7	94.1	92.5	85.6	85.6	94.6	94.6	87.3	87.3	67.4	83.4	92.7	92.8	**98.8**	**98.8**	95.3	95.3
Screenshot	93.5	93.5	91.7	89.8	86.9	86.9	94.1	94.1	85.7	85.7	62.0	83.9	93.9	94.0	**96.8**	**96.9**	95.8	95.8
Multi-platform	73.7	73.7	85.9	83.8	47.8	47.8	78.2	78.2	64.2	64.2	68.5	81.6	76.0	75.9	91.8	91.2	**92.6**	**92.6**
Re-digitization	67.0	67.0	81.4	78.9	45.1	45.1	76.9	76.9	67.9	67.9	61.2	74.9	69.9	69.9	86.1	85.1	**90.3**	**90.3**

The stored-classifier reference **combines adaptation with retention**: 94.3% pooled AUROC in both twelve-condition streams, 11.4 points above SLDA, while retaining 96.9% on the initial condition (Table 2). Frozen has zero drop but only 76.7% pooled AUROC.

With an updated linear head, frozen SigLIP 2 So400m trails CUE by 1.11 pooled AUROC points (95% CI: 0.33–1.87) and 1.23 (0.47–2.00) in the two orders. It retains the initial condition better; the encoders differ. SigLIP 2 scores 96.8–98.8 on isolated transformations and 85.1–91.8 on multi-platform transmission and re-digitization (CUE: 90.3–92.6). Robustness to single transformations need not compose along a propagation chain (Appendix I).

In the generalized-eigenvector ablation (Table 3, left), the current audit alone yields 93.12 macro Q_{11} AUROC (complete core: 92.94), with a 1.35-point Q_{00} AUROC drop and a 2.49-point historical-threshold FPR increase. Complete-core changes are -0.51 and $+0.13$ points. Accumulating all audits yields 87.77 macro Q_{11} AUROC (Table 26).

Table 3: **Adaptation and contribution-control ablations.** Left (generalized-eigenvector reference): mean/worst Q_{11} AUROC (%) and changes in Q_{00} AUROC and historical-threshold FPR (pp). Right: action accuracy, FPR, AUROC (%); R/A: withdrawal/restoration delays (audits). Gray: CUE; bold/underline: best/second-best.

Adaptation	Mean↑	Worst↑	ΔQ_{00}↑	ΔFPR↓
No adaptation	89.05	77.17	0.00	0.00
Accumulate all audits	87.77	78.75	−0.04	+0.15
No window balancing	91.18	80.79	**+0.47**	**−1.48**
No head regularization	91.46	80.05	−0.97	+1.23
Use current audit only	**93.12**	**83.08**	−1.35	+2.49
CUE (ours)	92.94	83.02	−0.51	+0.13

Control	Accuracy↑	R/A↓	FPR↓	AUROC↑
Core detector only	–	–	49.96	82.14
Always-on evidence	65.22	–	51.44	81.52
No confidence reset	95.65	0/–	**48.37**	82.26
No reactivation	82.61	0/–	49.96	82.14
No conservative bound	**100.0**	**0/0**	49.96	**82.30**
CUE (ours)	**100.0**	**0/0**	**48.37**	**82.30**

5.4 Reversible Evidence Use and Deployment

A retained contribution can become harmful and later useful again. Across three independent cycles, CUE suspends harmful fixed contributions and restores them when conditions return, recovering its initial error (Figure 6). At a fixed threshold on the same 256 real and 256 fake images, screenshot withdrawal reduces balanced error from 97.1% to 22.9%. Permanent deletion matches CUE during invalidation but forfeits reuse: editing-return error is 49.4%, versus 2.7% for CUE (Appendix H.3).

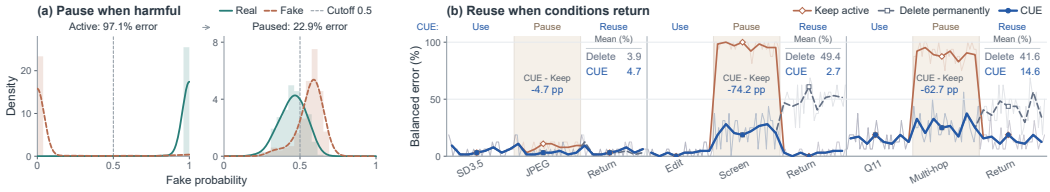


Figure 6: **Reversible contribution use.** Left: screenshot fake probabilities at cutoff 0.5. Right: balanced error over three independent cycles; shading marks changed conditions. Faint/bold: batch values/non-overlapping four-batch means. Annotations: all 32 batches per phase; state labels: CUE.

On four controller histories and five query trajectories (Table 3, right), CUE takes all 23 actions correctly at the corresponding audit, with zero phase-weighted exposure (active score influence during invalid or dormant phases). Full recovery requires a post-withdrawal confidence reset. Against always-on use at the shared threshold, false positives fall from 51.44% to 48.37%; sensitivity changes from 95.33% to 93.98% (Appendix H.3). Appendix H.4 tests dormancy and reuse with fewer labels.

Small mutable state. The generalized-eigenvector reference uses three windows and six label–window cells, changes **0 of 304M backbone parameters**, and stores **no replay images**. Mutable state totals at most **5.385 MiB**: a 5.217-MiB core plus three contributions (Table 4; Appendix I.2). After audit-image encoding, two consecutive core updates take **38.4 ms** at p95. Saved controller state restores exactly (Appendix J).

Table 4: **Contribution-control overhead.** J: contributions; Extra: state beyond the core; Log: audit log. Predict: per image; Update: per control audit, excluding encoding and core computation.

Setting	J	Extra/MiB	Log/KiB	Predict/μs	Update/ms
No control	0	< 0.001	0.00	0.000	–
Active	3	0.144	5.97	0.353	0.386
Dormant	3	0.152	9.01	0.397	0.473
Reactivated	3	0.168	15.01	0.356	0.616
Active	16	0.888	72.94	0.855	1.444
Stress test	64	3.545	287.29	2.717	5.829

6 Conclusion

We prove that finite history reduces adaptation labels relative to optimal past decisions alone in a Gaussian construction, and establish conditions for shared-space reuse under joint generation–processing changes. CUE derives coordinates and a classifier from one objective, with an estimation-error bound; adding or removing statistics supports continual learning and unlearning. CUE reaches state-of-the-art detection; history improves adaptation at matched current-label budgets. Suspended contributions can be reused when conditions return.

AI USE STATEMENT

Generative AI tools assisted with manuscript polishing, literature search and synthesis, drafting parts of the appendix, result analysis, reproduction and comparison of existing methods, and code refinement. The authors developed the mathematical proofs. Multiple AI agents separately wrote and ran numerical or symbolic verification programs to check the derivations. We did not use generative AI to create datasets for this study. The authors reviewed the AI-assisted text, code, analyses, and verification results and take responsibility for the final content of this submission.

REPRODUCIBILITY STATEMENT

Section 5 describes the matched evaluation protocols and metrics. Appendices K, L, and M prove the Gaussian label separation, shared-space sufficiency, and estimation guarantee for the generalized-eigenvector construction. Appendices B and C give CUE's joint-moment objective, statistical-error bound, and executable update and removal rules. Appendix A assigns the reported results to their evaluated implementations. The remaining appendices specify the estimators and structural guarantees, prediction and audit algorithms, world construction, and observation permissions. They also document identity-isolated splits, audit budgets, statistical units, implementation dimensions, experiment-specific hyperparameters, and resource measurements. We retain per-image scores, metric definitions, and independent recomputation records, as described in Appendix J. Permitted derived artifacts can be distributed subject to their source-dataset licenses.

REFERENCES

Kyra Ahrens, Hans Hergen Lehmann, Jae Hee Lee, and Stefan Wermter. Read between the layers: Leveraging multi-layer representations for rehearsal-free continual learning with pre-trained models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ZTcxp9xYr2.

Daniele Baracchi, Giulia Boato, Francesco De Natale, Massimo Iuliani, Andrea Montibeller, Cecilia Pasquini, Alessandro Piva, and Dasara Shullani. Toward open-world multimedia forensics through media signature encoding. *IEEE Access*, 12:59930–59952, 2024. doi: 10.1109/ACCESS.2024.3391809. URL https://doi.org/10.1109/ACCESS.2024.3391809.

Abhijit Bendale and Terrance E. Boult. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1893–1902, 2015. doi: 10.1109/CVPR.2015.7298799. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Bendale_Towards_Open_World_2015_CVPR_paper.html.

Xinyuan Cao, Weiyang Liu, and Santosh Vempala. Provable lifelong learning of representations. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 6334–6356. PMLR, 2022. URL https://proceedings.mlr.press/v151/cao22a.html.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hkf2_sC5FX.

Bo Du, Xuekang Zhu, Xiaochen Ma, Chenfan Qu, Kaiwen Feng, Zhe Yang, Chi-Man Pun, Jian Liu, and Ji-Zhe Zhou. ForensicHub: A unified benchmark & codebase for all-domain fake image detection and localization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2167. URL https://papers.neurips.cc/paper_files/paper/2025/hash/5d8a47fc1cb0a143f149ee07321621c9-Abstract-Datasets_and_Benchmarks_Track.html.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x. URL https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Dario Fontanel, Fabio Cermelli, Massimiliano Mancini, and Barbara Caputo. On the challenges of open world recognition under shifting visual domains. *IEEE Robotics and Automation Letters*, 6 (2):604–611, 2021. doi: 10.1109/LRA.2020.3047777. URL https://doi.org/10.1109/LRA.2020.3047777.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–37, 2014. doi: 10.1145/2523813. URL https://doi.org/10.1145/2523813.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general AI-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18685–18694, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Guillaro_A_Bias-Free_Training_Paradigm_for_More_General_AI-generated_Image_Detection_CVPR_2025_paper.html.

Tyler L. Hayes and Christopher Kanan. Lifelong machine learning with deep streaming linear discriminant analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 887–896, June 2020. doi: 10.1109/CVPRW50498.2020.00118. URL https://openaccess.thecvf.com/content_CVPRW_2020/html/w15/Hayes_Lifelong_Machine_Learning_With_Deep_Streaming_Linear_Discriminant_Analysis_CVPRW_2020_paper.html.

Yan Hong, Jianming Feng, Haoxing Chen, Jun Lan, Huijia Zhu, Weiqiang Wang, and Jianfu Zhang. WildFake: A large-scale and hierarchical dataset for AI-generated images detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):3500–3508, 2025. doi: 10.1609/aaai.v39i4.32363. URL https://ojs.aaai.org/index.php/AAAI/article/view/32363.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://doi.org/10.1073/pnas.1611835114.

Chunxiao Li, Xiaoxiao Wang, Meiling Li, Boming Miao, Peng Sun, Yunjian Zhang, Xiangyang Ji, and Yao Zhu. Bridging the gap between ideal and real-world evaluation: Benchmarking AI-generated image detection in challenging scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20379–20389, October 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Li_Bridging_the_Gap_Between_Ideal_and_Real-world_Evaluation_Benchmarking_AI-Generated_ICCV_2025_paper.html.

Chuqiao Li, Zhiwu Huang, Danda Pani Paudel, Yabin Wang, Mohamad Shahbazi, Xiaopeng Hong, and Luc Van Gool. A continual deepfake detection benchmark: Dataset, methods, and essentials. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1339–1349, January 2023. URL https://openaccess.thecvf.com/content/WACV2023/html/Li_A_Continual_Deepfake_Detection_Benchmark_Dataset_Methods_and_Essentials_WACV_2023_paper.html.

Yiheng Li, Zichang Tan, Guoqing Xu, Zhen Lei, Xu Zhou, and Yang Yang. Towards generalizable AI-generated image detection via image-adaptive prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21262–21271, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Li_Towards_Generalizable_AI-Generated_Image_Detection_via_Image-Adaptive_Prompt_Learning_CVPR_2026_paper.pdf.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-5739. URL https://papers.nips.cc/paper_files/paper/2025/hash/fb693c67f61e5321746ffce8b6fdd2d0-Abstract-Datasets_and_Benchmarks_Track.html.

Mark D. McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. RanPAC: Random projections and pre-trained models for continual learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0526. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2793dc35e14003dd367684d93d236847-Abstract-Conference.html.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ojha_Towards_Universal_Fake_Image_Detectors_That_Generalize_Across_Generative_Models_CVPR_2023_paper.html.

Kun Pan, Yifang Yin, Yao Wei, Feng Lin, Zhongjie Ba, Zhenguang Liu, Zhibo Wang, Lorenzo Cavallaro, and Kui Ren. DFIL: Deepfake incremental learning by exploiting domain-invariant forgery clues. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8035–8046. ACM, 2023. doi: 10.1145/3581783.3612377. URL https://doi.org/10.1145/3581783.3612377.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Park_Community_Forensics_Using_Thousands_of_Generators_to_Train_Fake_Image_CVPR_2025_paper.pdf.

Ziheng Qin, Yuheng Ji, Renshuai Tao, Yuxuan Tian, Yuyang Liu, Yipu Wang, and Xiaolong Zheng. Scaling up AI-generated image detection with generator-aware prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Qin_Scaling_Up_AI-Generated_Image_Detection_with_Generator-Aware_Prototypes_CVPR_2026_paper.pdf.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tan_Rethinking_the_Up-Sampling_Operations_in_CNN-based_Generative_Network_for_Generalizable_CVPR_2024_paper.html.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10434–10443. PMLR, 2021. URL https://proceedings.mlr.press/v139/tripuraneni21a.html.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv:2502.14786, 2025. URL https://arxiv.org/abs/2502.14786. Preprint.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010. doi: 10.48550/arXiv.1011.3027. URL https://arxiv.org/abs/1011.3027. Version 7, revised 2011; Chapter 5 in Compressed Sensing, Theory and Applications, Cambridge University Press, 2012.

Jiaan Wang, Sirui Liu, Yu Li, Kaiyuan Yang, Juan Cao, and Sheng Tang. Fleet: Few shots lead effective AI-generated image detection. *arXiv preprint arXiv:2606.31082*, 2026. URL https://arxiv.org/abs/2606.31082.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Yan Wang, Wei-Lun Chao, Kilian Q. Weinberger, and Laurens van der Maaten. SimpleShot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019. doi: 10.48550/arXiv.1911.04623. URL https://arxiv.org/abs/1911.04623.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_DIRE_for_Diffusion-Generated_Image_Detection_ICCV_2023_paper.html.

Shiyu Wu, Jing Liu, Jing Li, and Yequan Wang. Few-shot learner generalizes across AI-generated image detection. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67449–67460. PMLR, 2025. URL https://proceedings.mlr.press/v267/wu25r.html.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b0303773962ea1b5394c3a83cc7dd066-Abstract-Conference.html.

Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. D^3: Scaling up deepfake detection by learning from discrepancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Yang_D3_Scaling_Up_Deepfake_Detection_by_Learning_from_Discrepancy_CVPR_2025_paper.pdf.

Jingwen Ye, Yifang Fu, Jie Song, Xingyi Yang, Songhua Liu, Xin Jin, Mingli Song, and Xinchao Wang. Learning with recoverable forgetting. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, volume 13671 of *Lecture Notes in Computer Science*, pp. 87–103, Cham, 2022. Springer. doi: 10.1007/978-3-031-20083-0_6. URL https://doi.org/10.1007/978-3-031-20083-0_6.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *arXiv preprint arXiv:1405.0680*, 2014. URL https://arxiv.org/abs/1405.0680.

Hongbo Zhao, Bolin Ni, Junsong Fan, Yuxi Wang, Yuntao Chen, Gaofeng Meng, and Zhaoxiang Zhang. Continual forgetting for pre-trained vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28631–28642, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhao_Continual_Forgetting_for_Pre-trained_Vision_Models_CVPR_2024_paper.html.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. PatchCraft: Exploring texture patch for efficient AI-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023. doi: 10.48550/arXiv.2311.12397. URL https://arxiv.org/abs/2311.12397.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting AI-generated image. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77771–77782, 2023. doi: 10.52202/075280-3398. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f4d4a021f9051a6c18183b059117e8b5-Abstract-Datasets_and_Benchmarks.html.

A Implementation Configurations and Result Assignment

The statistical CUE model is the joint-moment learner in Equation 4. Its frozen front end, initialization, retained historical windows, and contribution bank are specified for each evaluation. The generalized-eigenvector, neural-adaptation, and stored-classifier references use different learning procedures. Their results are assigned below rather than treated as interchangeable instances of the statistical update.

Statistical CUE on standard benchmarks. The Community Forensics front end supplies its released 384-dimensional features and original linear logit. The fixed standardizer and 256-dimensional projection use 8,000 initialization records, with 2,000 separate records for scale calibration. The six benchmark episodes each use 16 real and 16 fake support images, no added historical windows, and no controlled contributions ($R = I$, $J = 0$). All use the same joint-moment solver and the initialization in Appendix C.1. The compared released detectors receive no current support labels. Each benchmark uses a common query subset for all compared scores, with 2,997 query records in total. Appendix N.2 gives the protocol and all 22 comparisons.

Benchmark	Statistical CUE	Prior CF adaptation	Best released detector
UFD	99.38	98.37	98.45
GenImage	99.67	98.89	99.56
AIGC-DB	99.40	99.06	99.06
AIGIBench	98.84	96.65	95.99
WildFake	99.48	99.30	99.31
Chameleon	98.49	93.68	92.60

These scores belong to the joint-moment update; the CF support-selected prototype/kernel configuration in Table 1 uses a different adaptation rule. Figure 4 evaluates the frozen CF starting detector without current adaptation.

Historical reuse with a matched statistical update. A DINOv2 front end augmented with image statistics evaluates the same joint-moment solver under two fixed histories: no additional historical windows and 256 additional historical labels. The current support, initial detector, JPEG contribution, and controller are shared. Four conditions cross two generators with original and processed images, using six fixed support orders and eight budgets, for 384 fitted episodes. The mean normalized area under the AUROC curve against $\log_2 K$ increases by 1.663926 percentage points, with a paired 95% interval of [1.355028, 1.985257]. The interval resamples linked image-identity groups; it does not represent uncertainty over arbitrary future generators. Appendix N.1 defines the metric, splits, and separate historical and current label costs. This result establishes historical reuse for the DINOv2 configuration, not an already measured CF history benefit.

Neural adaptation and continual reference results. The six-scene curves in Figure 5 use IAPL/CLIP neural adaptation with the specified gradient-clipping and fixed combination rules. The prompt, visual adapter, conditional modules, and classifier are updated for each support budget. They are therefore distinct from a fixed-feature joint-moment learner. On a common 2,712-query subset, the fixed DINOv2 statistical configuration reaches 86.34% and 88.23% mean AUROC at 16 and 64 labels per class; the corresponding neural-adaptation reference reaches 94.98% and 98.43%. The original full-query reference values remain 94.94% and 98.42%. The six-scene CF statistical configuration has not been evaluated.

The continual results in Table 2 use the stored-classifier reference with its fixed combination and unlabeled batch-selection rules. They do not measure the joint-moment solver in Equation 4. The original generalized-eigenvector ablation and three-window resource measurements similarly refer to the construction in Appendix D. Contribution-control evaluations test their stated fixed score terms and audit trajectories. Their numerical results and resource measurements retain these implementation-specific scopes.

B THE JOINT-MOMENT OBJECTIVE AND ITS PROPERTIES

We use authenticity labels $Y \in \{0, 1\}$ and signed labels $y = 2Y - 1$. The frozen pipeline supplies $z \in \mathbb{R}^d$ and the initial score $a(x)$. Write $g = a/\kappa$, $u = [z; g]$, and $v = [w; 1]$, with fixed $\kappa > 0$. The normalized core score is $f(x) = g(x) + z(x)^\top w + b$. The current support contains $K \geq 1$ observations per class, so $n = 2K$. Its joint class means have midpoint $\mu = [\mu_z; \mu_g]$ and half-difference $\delta = [c; c_g]$. We fix $M_B \succ 0$, $\lambda > 0$, and an allowed-space basis $R \in \mathbb{R}^{d \times k}$ with $R^\top R = I_k$, where $1 \leq k \leq d$. The correction satisfies $w \in \mathrm{range}(R)$ and $\alpha = \lambda/n$.

B.1 A Single Risk from Joint Moments

A nonempty class-window cell stores (n_j, s_j, Q_j), where $s_j = \sum_i u_i$ and $Q_j = \sum_i u_i u_i^\top$. Its centered scatter is $Q_j - s_j s_j^\top / n_j$. Let SS_H sum these scatters over the retained historical cells, let $\nu_H = \sum_{j \in H}(n_j - 1)$, and let SS_S sum the two current class scatters. The estimator uses

$$W = \frac{SS_H + SS_S}{\nu_H + n}. \tag{9}$$

Historical cells and current support are disjoint. Every scatter is positive semidefinite, hence $W \succeq 0$.

Consider a balanced binary distribution $\widetilde{P}$ with

$$\mathbb{E}_{\widetilde{P}}[u \mid y] = \mu + y\delta, \qquad \mathrm{Cov}_{\widetilde{P}}(u \mid y) = W, \qquad \widetilde{P}(y = +1) = \widetilde{P}(y = -1) = \tfrac{1}{2}. \tag{10}$$

These moments are realizable: take $u = \mu + y\delta + \xi$, where ξ is independent of y, has mean zero, and has covariance W. No Gaussian assumption is needed. Equation 10 uses current class means and the pooled variation estimate; it defines the moment model used for fitting.

Proposition 3 (Joint-moment risk identity)**.** *The regularized squared risk under $\widetilde{P}$ equals*

$$\begin{aligned} \mathcal{L}(w, b) &:= \mathbb{E}_{\widetilde{P}}\big[(y - v^\top u - b)^2\big] + \alpha w^\top M_B w \\ &= v^\top W v + (w^\top c + c_g - 1)^2 + (w^\top \mu_z + \mu_g + b)^2 + \alpha w^\top M_B w. \end{aligned} \tag{11}$$

Proof. Write $\xi = u - \mu - y\delta$. Its conditional mean is zero, and

$$y - v^\top u - b = y(1 - v^\top \delta) - (v^\top \mu + b) - v^\top \xi.$$

All cross terms vanish after taking expectations: $\mathbb{E}[y] = 0$ and $\mathbb{E}[\xi \mid y] = 0$. The three squared terms give $(1 - v^\top \delta)^2$, $(v^\top \mu + b)^2$, and $v^\top W v$, respectively. Adding the regularizer proves the identity. □

The same identity holds when the two class covariances differ but their average is W. With no history, $W = SS_S/n$, and Equation 11 is exactly the current empirical squared loss plus its regularizer. With history, the pooled estimate replaces the current within-class second moment while retaining the current class means.

B.2 The Closed-Form Solution and Its Coordinates

Partition W according to $u = [z; g]$, and define

$$G = W_{zz} + \alpha M_B, \qquad B = G + cc^\top, \qquad h = (1 - c_g)c - W_{zg}. \tag{12}$$

Both G and B are positive definite.

Proposition 4 (Equivalent constrained and coordinate solutions)**.** *The unique minimizer of Equation 11 over $w \in \mathrm{range}(R)$ and $b \in \mathbb{R}$ is*

$$w = R(R^\top B R)^{-1} R^\top h, \qquad b = -\mu_g - \mu_z^\top w. \tag{13}$$

Let $LL^\top = R^\top G R$ be the Cholesky factorization, with positive diagonal entries in L, and set

$$U = RL^{-\top}, \qquad \widetilde{c} = U^\top c, \qquad q = U^\top W_{zg}. \tag{14}$$

Then $U^\top G U = I_k$, and the same solution is

$$\begin{aligned} \beta &= (I_k + \widetilde{c}\widetilde{c}^\top)^{-1}\big[(1 - c_g)\widetilde{c} - q\big] \\ &= -q + \widetilde{c}\,\frac{1 - c_g + \widetilde{c}^\top q}{1 + \|\widetilde{c}\|_2^2}, \qquad w = U\beta. \end{aligned} \tag{15}$$

Proof. Minimizing over b sets the score-center term to zero. Write $w = R\theta$. Differentiating the remaining objective gives

$$R^\top B R\,\theta = R^\top h.$$

Since $R^\top BR \succ 0$, this has the unique solution in Equation 13. The Cholesky definition gives $U^\top GU = I_k$ and $\mathrm{range}(U) = \mathrm{range}(R)$. After substituting $w = U\beta$ and the minimizing intercept, the objective is

$$\|\beta\|_2^2 + 2q^\top\beta + (\widetilde{c}^\top\beta + c_g - 1)^2 + W_{gg}.$$

Its normal equations can be written as

$$(I_k + \widetilde{c}\widetilde{c}^\top)(\beta + q) = \widetilde{c}(1 - c_g + \widetilde{c}^\top q).$$

Since $(I_k + \widetilde{c}\widetilde{c}^\top)\widetilde{c} = (1 + \|\widetilde{c}\|_2^2)\widetilde{c}$, this gives both expressions in Equation 15. □

Thus the same normalized core score can be evaluated as

$$f(x) = g(x) - \mu_g + \beta^\top U^\top\big(z(x) - \mu_z\big). \tag{16}$$

In these coordinates, within-class score variance plus the regularizer is $\|\beta + q\|_2^2 + W_{gg} - \|q\|_2^2$, whose unique minimizer is $-q$. The remaining adjustment in Equation 15 lies along the current contrast $\widetilde{c}$ and trades off this cost against matching the signed class targets. Both terms belong to the same optimum; U and q depend on historical and current within-class statistics. The coordinate expression changes neither the allowed space nor the fitted score.

B.3 Joint-Moment Error and the Fitted Score

Fix the feature pipeline, the initial detector, κ, the current μ, c, c_g, and α, M_B, R. Compare two positive-semidefinite joint variation matrices $W_\star$ and $\widehat{W}$, with $E = \widehat{W} - W_\star$. Use Equation 13 to define their respective coefficients $w_\star = R\theta_\star$, $\widehat{w} = R\widehat{\theta}$ and intercepts $b_\star, \widehat{b}$. The reference replaces only the joint variation matrix; it keeps the realized current-support means and contrasts.

To prove Proposition 2, write

$$\widehat{B}_R = R^\top(\widehat{W}_{zz} + cc^\top + \alpha M_B)R, \qquad m = \lambda_{\min}(\widehat{B}_R).$$

Then $m \geq \alpha\lambda_{\min}(R^\top M_B R) > 0$.

Proof of Proposition 2. The two reduced normal equations give

$$\widehat{B}_R(\widehat{\theta} - \theta_\star) = -R^\top(E_{zz}w_\star + E_{zg}). \tag{17}$$

Consequently,

$$\begin{aligned}\|\widehat{w} - w_\star\|_2 &\leq \frac{\|R^\top(E_{zz}w_\star + E_{zg})\|_2}{m} \\ &\leq \frac{\|E\|_{\mathrm{op}}\sqrt{\|w_\star\|_2^2 + 1}}{m}.\end{aligned} \tag{18}$$

Here $w_\star = R\theta_\star$. Because R has orthonormal columns, $\|\widehat{w} - w_\star\|_2 = \|\widehat{\theta} - \theta_\star\|_2$. The first bound follows by applying $\widehat{B}_R^{-1}$. For the second, $E_{zz}w_\star + E_{zg}$ is the feature block of $E[w_\star; 1]$, and multiplication by $R^\top$ cannot increase its norm. The lower bound on m follows from $\widehat{W}_{zz} \succeq 0$ and $cc^\top \succeq 0$. For the normalized core scores $\widehat{f}$ and $f_\star$, let $\Delta w = \widehat{w} - w_\star$ and $\Delta f = \widehat{f} - f_\star$. The intercept formula gives

$$\Delta f(x) = (z(x) - \mu_z)^\top \Delta w. \tag{19}$$

□

The argument is deterministic; it requires no independence between the current contrasts and the variation estimate.

For an evaluation distribution $\mathcal{D}$ with finite second moments, set $T_\mathcal{D} = \mathbb{E}_\mathcal{D}[(z - \mu_z)(z - \mu_z)^\top]$. The mean squared score change satisfies

$$\begin{aligned}\varepsilon_f &:= \mathbb{E}_\mathcal{D}[(\Delta f)^2] = \Delta w^\top T_\mathcal{D}\Delta w \\ &\leq \frac{\|T_\mathcal{D}\|_{\mathrm{op}}\|E\|_{\mathrm{op}}^2(\|w_\star\|_2^2 + 1)}{m^2}.\end{aligned} \tag{20}$$

For squared risk $\mathcal{R}_{\mathcal{D}}(f) = \mathbb{E}_{\mathcal{D}}[(y-f)^2]$, direct expansion yields

$$\mathcal{R}_{\mathcal{D}}(\widehat{f}) - \mathcal{R}_{\mathcal{D}}(f_\star) = \varepsilon_f - 2\mathbb{E}_{\mathcal{D}}[(y-f_\star)\Delta f], \tag{21}$$

$$\left|\mathcal{R}_{\mathcal{D}}(\widehat{f}) - \mathcal{R}_{\mathcal{D}}(f_\star)\right| \le \varepsilon_f + 2\sqrt{\mathcal{R}_{\mathcal{D}}(f_\star)\varepsilon_f}. \tag{22}$$

The inequality is Cauchy–Schwarz. The residual cross term remains: $f_\star$ uses the same empirical current means and contrasts and need not minimize risk under $\mathcal{D}$.

B.4 Moment Addition, Removal, and Core Reconstruction

For a finite observation block A, write $\mathcal{M}(A) = (|A|, \sum_{i\in A} u_i, \sum_{i\in A} u_i u_i^\top)$. For disjoint blocks A, D,

$$\mathcal{M}(A \cup D) = \mathcal{M}(A) + \mathcal{M}(D), \qquad \mathcal{M}(A \cup D) - \mathcal{M}(D) = \mathcal{M}(A). \tag{23}$$

Subtraction uses the moments of the block being removed. After changing a cell, its scatter is recomputed from its updated (n, s, Q); an empty cell is omitted. For a whole historical window, removing its cells removes exactly their contributions to SS_H and ν_H.

Proposition 5 (Exact reconstruction from retained moments)**.** *Fix the feature pipeline, initial detector, κ, λ, M_B, R, the retained class-window partition, and the designation of the current window. Recomputing the core after removing a historical window gives the same W, μ, δ, coordinates U, coefficients w, b, and scores as rebuilding them from the remaining observations. The equality is in exact arithmetic.*

Proof. Equation 23 gives the exact moments of every retained cell. They determine its count, mean, and scatter, hence SS_H, ν_H, SS_S, n and the current μ, δ. Therefore they determine the same W and α. The constrained minimizer is unique, and Cholesky factorization with positive diagonal entries is unique, proving equality of the derived coordinates, coefficients, and scores. □

The result concerns the core reconstructed for the specified retained windows and current support. Removing all history gives the current-only learner. If the current support is absent, the implemented fallback sets $w = b = 0$ and returns $a(x)$. For an empty allowed space, the nonempty-support solution is $w = 0,\ b = -\mu_g$.

C Classifier Selection, Contribution Control, and Window Removal

The online state contains retained joint feature–score statistics, saved classifiers, fixed contributions ψ_j, and their use states γ_j. Each saved classifier has immutable coefficients, a batch descriptor, and references to the windows used to construct it. Prediction reads this state; verified labels change it only at an update.

C.1 Fixed Initialization

Initialization uses a fixed split of 8,000 support records and 2,000 calibration records. The split is obtained by hashing image identifiers separately within each class, with seed 20260824, and reserving the first fifth for calibration. The stored allocation fixes the exact records. A feature standardizer and unwhitened 256-dimensional PCA are fitted on the 8,000 support records; PCA uses seed 20260824. Both remain fixed throughout subsequent updates.

For the DINOv2-L representation augmented with 78 image statistics, the initial score $a(x)$ is a logistic-regression logit. Its fit starts with equal total weight for each class. Two rounds identify the highest-scoring 15% of real images and multiply their balanced weights by five; a final fit uses the resulting weights. Logistic regression uses $C = 1$, at most 3,000 iterations, and seed 20260824. For the Community Forensics representation, $a(x)$ is the released linear logit on the original 384-dimensional features. This logit does not pass through the newly fitted standardizer or PCA.

The regularization matrix M_B is also fixed at initialization. For each initial class Y, multiply its count, feature sum and outer-product sum by $\min\{1, 256/n_Y\}$; this rescales its statistical weight

without subsampling its images. Let $\mathbb{E}_B$ denote the resulting weighted empirical average in the PCA space, and define

$$\begin{aligned} \mu_B &= \mathbb{E}_B[z], \qquad T_B = \mathbb{E}_B[zz^\top], \qquad \Sigma_B = T_B - \mu_B\mu_B^\top, \\ W_B &= \sum_{Y=0}^{1} \pi_Y \operatorname{Cov}_B(z \mid Y), \qquad \Delta_B = \mathbb{E}_B[z \mid Y=1] - \mathbb{E}_B[z \mid Y=0]. \end{aligned} \tag{24}$$

Here π_Y is the weighted class proportion. The single initial window gives a rank-one generalized discriminant direction: normalize $(2W_B + 10^{-3}I)^{-1}\Delta_B$ to obtain the unit vector q_B. This is the direction returned by the initial generalized eigensystem; the construction is performed once, rather than during online updates. Set

$$\begin{aligned} D_B = \operatorname{tr}(T_B), \qquad k_B = q_B^\top T_B q_B, \qquad \sigma_B^2 = q_B^\top \Sigma_B q_B, \\ M_B = I - \frac{D_B}{D_B + k_B} q_B q_B^\top, \qquad m_B = \frac{k_B}{D_B + k_B}, \\ \lambda = \frac{(10 + \epsilon_B)\sigma_B^2}{16 m_B}. \end{aligned} \tag{25}$$

where the recorded successful Cholesky jitter is $\epsilon_B = 10^{-6}$. Thus M_B has eigenvalue $m_B > 0$ along q_B and eigenvalue one in orthogonal directions.

The multiplier $1/16$ is inherited from a historical leave-one-source-out calibration completed before the present evaluation. Its four sources supply 768 distinct labeled identities, separate from the initial 8,000 support and 2,000 calibration records. The first block contains 256 identities for historical statistics; the other two blocks contain 512 identities for calibration support and queries. These 512 labels are a shared calibration cost for every arm, including the control without added historical geometry. Each held-out source is excluded from its fold's historical statistics. The six grid values $1/16, 1/4, 1, 4, 16, 64$ multiply the base reference strength $(10 + \epsilon_B)\sigma_B^2/m_B$. Selection uses the mean area under the AUROC curve against $\log_2 K$ over the four held-out sources. The control, historical, and random-orientation arms all select $1/16$ and share this frozen value in the reported evaluation. Community Forensics transfers this multiplier and recomputes its own base reference strength; it does not repeat the historical selection.

The score scale κ is determined on the separate calibration split. Write $\bar{y}_B = \mathbb{E}_B[y]$ and $c_B = \mathbb{E}_B[(z - \mu_B)(y - \bar{y}_B)]$. The initial one-direction ridge score and its calibration scale are

$$\begin{aligned} h_B(z) &= \bar{y}_B + \frac{q_B^\top c_B}{(11 + \epsilon_B)\sigma_B^2} q_B^\top (z - \mu_B), \\ \kappa &= \frac{\operatorname{Cov}_{\mathrm{cal}}(a, h_B)}{\operatorname{Var}_{\mathrm{cal}}(h_B)}. \end{aligned} \tag{26}$$

The frozen DINOv2-L configuration gives $\kappa = 195.28429704757548$ and $\lambda = 393.06698064271103$; the Community Forensics configuration gives $\kappa = 132.8073255390887$ and $\lambda = 240.74777421095598$. Subsequent labels update the joint moments and current classifier, while M_B, κ, λ, the projection, and the initial score remain fixed.

C.2 Selecting a Classifier from an Unlabeled Batch

Selection uses the distribution of the incoming batch, without its labels or source and processing identities. In the continual implementation, let $\varphi(x) \in \mathbb{R}^{1102}$ contain the frozen visual features and image statistics. A feature standardizer is fitted once on the initial support pool and then frozen. Its zero-variance coordinates have scale one. The standardized features are projected to 64 dimensions:

$$\xi(x) = \Omega^\top \frac{\varphi(x) - \mu_{\mathrm{route}}}{\sigma_{\mathrm{route}}}, \qquad \Omega_{ab} \sim \mathcal{N}(0, 1/1102). \tag{27}$$

Division is coordinatewise. The implementation fixes NumPy's default random generator seed to 20260824 and stores Ω in float32. For a batch $\mathcal{B}$, concatenate four coordinatewise statistics:

$$h(\mathcal{B}) = \big[\operatorname{mean}(\xi),\ \log(1 + \operatorname{std}(\xi)),\ Q_{0.25}(\xi),\ Q_{0.75}(\xi)\big] \in \mathbb{R}^{256}. \tag{28}$$

Standard deviations use divisor $|\mathcal{B}|$; quantiles use linear interpolation. Each new classifier stores the descriptor of its complete support batch. The initial classifier stores a descriptor from 64 real and 64 fake initial support images, selected by the fixed hash ordering used for initialization.

Let $\mathcal{C}_t$ contain the valid classifiers available at time t, including the frozen initial classifier. With stored descriptors h_i, set $a_{t,k} = \max\{\mathrm{std}_{i\in\mathcal{C}_t}(h_{i,k}), 0.10\}$ and select

$$i_t(\mathcal{B}) = \arg\min_{i\in\mathcal{C}_t} \frac{1}{256}\sum_{k=1}^{256}\left(\frac{h_k(\mathcal{B}) - h_{i,k}}{a_{t,k}}\right)^2. \tag{29}$$

Ties select the first classifier in storage order. All images in the batch use this classifier. Batch boundaries are therefore part of the input protocol. Selection neither averages classifiers nor updates their coefficients. The initial classifier has zero adaptive correction and remains available when no learned classifier is valid.

C.3 Prediction, Verification, and Learning

For a selected classifier i, its score is $r_i(x) = a(x) + \kappa\{w_i^\top z(x) + b_i\}$, where a is the frozen initial score and z is the fixed projected representation. The complete score is

$$s_t(x;\mathcal{B}) = r_{i_t(\mathcal{B})}(x) + \sum_j \gamma_{j,t}\psi_j(x). \tag{30}$$

The fixed contributions are added once, after classifier selection. Only an ACTIVE contribution has $\gamma_{j,t} = 1$; PROVISIONAL and DORMANT contributions have $\gamma_{j,t} = 0$.

At update t, execution follows this order:

1. Score the incoming query batch using the state available before update t. Query labels are not supplied to selection or learning.
2. Receive a separate batch $\mathcal{A}_t$ with verified labels $Y \in \{0,1\}$, and use $y = 2Y - 1$ when fitting the classifier. For contribution verification, exclude identities already used in earlier updates or reserved for contribution construction and calibration. Select a classifier for the eligible audit images from $\mathcal{C}_{t^-}$, before admitting the new classifier.
3. Compute all contribution audits from the same selected classifier and the same pre-update use states, as specified below. If the eligible images do not contain both classes, retain the controller state unchanged.
4. Move the previous current window into historical statistics, install the balanced current support batch, and solve for the new classifier. Save its coefficients, support descriptor, arrival index, and dependencies. It becomes available only after this update.

The same current labels may verify contributions and fit the new classifier; they are counted once. Verification precedes fitting, so it does not assess a classifier on the labels just used to train it. Core and controller changes are staged and published only after both operations succeed. Immediate post-update evaluation uses the newly fitted classifier; later batch prediction uses Equation 29.

The current window remains balanced. The stream implementation avoids counting repeated observations twice: an exactly repeated, same-label class cell moves from its previous historical owner to the current window. Other class cells remain stored. Partial overlaps and inconsistent labels are rejected. This transfer changes current statistical ownership, while saved classifiers retain their original training dependencies.

C.4 Auditing a Fixed Contribution

The fixed JPEG sensor uses 48 historical image identities, each with an original and a JPEG view. Its direction is the normalized mean of the paired projected-feature differences; its center is the mean midpoint of the two views along that direction. Authenticity labels do not enter this estimation. A disjoint set of 48 labeled identities calibrates its scalar weight by logistic loss with the frozen initial score as an offset. Calibration uses real originals and JPEG views of fake images, creating a controlled historical correlation. The scalar response is scaled to unit standard deviation for fitting,

with an L_2 coefficient of 0.001 and coefficient range $[-100, 100]$; its scale is then absorbed into the fixed weight. The resulting $\psi_j(x)$ is the weighted, centered sensor response. These construction and calibration identities remain ineligible for later contribution audits. Let B_{sensor} contain the fixed sensor directions in the standardized feature space preceding PCA, and let C be the fixed PCA component matrix. The correction uses

$$R = \operatorname{orth}\big(\ker\left((CB_{\text{sensor}})^\top\right)\big). \tag{31}$$

Thus sensor directions remain excluded from the adaptive correction in every use state. The fixed initial score and batch selector are unchanged.

Freeze the selected classifier and all γ_{j,t^-} throughout the audit. For each contribution, form its off and on scores from the common pre-update score s^{pre}:

$$s^{-j} = s^{\text{pre}} - \gamma_{j,t^-}\psi_j, \qquad s^{+j} = s^{-j} + \psi_j. \tag{32}$$

With signed labels, binary cross-entropy is $\ell(y, s) = \log(1 + \exp(-ys))$. For n eligible audit images, the observed gains are $d_i = \ell(y_i, s_i^{-j}) - \ell(y_i, s_i^{+j})$. Let $\bar{d}_j$ be their mean and $\widehat{v}_j = \operatorname{var}_{n-1}(d)/n + 10^{-12}$. The controller maintains a Gaussian estimate of the mean gain. Given prior mean $\mu_{j,t}^-$ and variance $v_{j,t}^-$, it computes

$$\begin{aligned} v_{j,t} &= \left(\frac{\rho}{v_{j,t}^-} + \frac{1}{\widehat{v}_j}\right)^{-1}, \\ \mu_{j,t} &= v_{j,t}\left(\frac{\rho\mu_{j,t}^-}{v_{j,t}^-} + \frac{\bar{d}_j}{\widehat{v}_j}\right), \\ p_{j,t} &= \Phi(\mu_{j,t}/\sqrt{v_{j,t}}). \end{aligned} \tag{33}$$

where $\rho = 0.25$ and Φ is the standard normal CDF. Initialization uses $(\mu_{j,t}^-, v_{j,t}^-) = (0, 1)$.

An additional check limits false positives. In each off/on branch separately, the threshold is the $\lceil 0.95n_1 \rceil$-th largest fake-image score, where n_1 is the number of fake audit images. Real images at or above that threshold count as false positives. The check passes when the on branch adds at most one false positive relative to the off branch.

Current state	Condition	Next state
Provisional	$p_{j,t} > 0.8$ and check passes	Active
Active	$p_{j,t} < 0.2$ or check fails	Dormant
Dormant	$p_{j,t} > 0.8$ and check passes	Active
Any	otherwise	unchanged

On the first eligible audit after withdrawal, the prior resets to $(0, 1)$; subsequent audits accumulate evidence within the new episode. All contribution updates use independent copies of their previous controller states and are committed together. No update reads another contribution's proposed new state.

The controller also returns a conservative score $\underline{s}_t = s_t + \sum_{j:\gamma_{j,t}=0} \min\{0, \psi_j\}$ for thresholded positive decisions. The integrated adaptation and continual evaluations compute AUROC from the ordinary score s_t, not from $\underline{s}_t$. Switching a contribution changes the ordinary score by exactly ψ_j when the selected classifier and other state are fixed. The classifier selector and frozen initial score remain separate from the direction constraint imposed on the adaptive correction.

C.5 Removing a Window and Its Dependent Classifiers

Window removal operates on both statistics and saved classifiers. Each classifier records its training-window dependencies, the window supplying its descriptor, and any parent classifiers from which it was derived. Deletion first removes the requested window's retained class cells and re-solves the live core. If the requested window is the current window, the live core has no current support and returns to the frozen initial score. Earlier valid classifiers remain candidates for batch selection.

Every classifier directly dependent on the removed window is invalidated, together with all of its dependent descendants. Their coefficients and descriptors are removed from the callable state. Invalidated descriptors also leave the reference set used to compute $a_{t,k}$ in Equation 29; they cannot influence subsequent selection through its scale estimate.

A classifier can be reconstructed when its exact creation-time sufficient statistics are available. These include the class-cell counts, sums, and cross-products, their window and identity ownership, and the fixed initialization, scales, metric, and direction constraint. Its current support and descriptor source must still be retained. The implementation removes deleted historical cells from that saved statistical state and solves again. The result receives a new identifier; the old classifier remains invalid. Classifiers with unavailable statistics or unsupported parent dependencies remain invalid, and prediction selects among the remaining classifiers, including the initial classifier.

Deletion is staged on a copy and committed after checking dependency closure and successful reconstruction. Unknown or repeated deletions are rejected without changing the state. A separate record of previously used identities is retained, so deletion cannot make old labels eligible for a fresh audit. Observations legitimately owned by another retained window remain available. The operation leaves the shared initialization, fixed contributions, and controller's earlier audit state unchanged. Its scope is removal of the specified window's retained statistics and dependent callable classifiers.

D Generalized-Eigenvector Reference and Contribution Control

This appendix defines the generalized-eigenvector reference used in the coordinate-identification analysis and the corresponding reference evaluations. It differs from the joint-moment update in Section 4: historical class contrasts determine its separation operator, whereas the joint-moment update fits the current contrast using pooled within-class statistics. Appendix A identifies the evaluated configurations.

The reference learner retains historical feature statistics from which it can learn new forensic coordinates as conditions change. A frozen visual encoder supplies the features; new labels and retained statistics determine the updated coordinates and classifier. Continual learning adds a batch's statistics, while unlearning subtracts them and recomputes the learned state. The reference learner also suspends or restores selected fixed contributions according to their current detection value.

D.1 A Unified Forensic Score

The reference learner combines an adaptive core with selected feature contributions. The core updates its coordinates and classifier from retained statistics. Each contribution is a fixed score term whose use the reference learner can suspend or restore.

Let $\phi : \mathcal{X} \to \mathbb{R}^d$ be the frozen encoder, and let $\bar{\boldsymbol{z}}(x)$ denote features standardized using fixed statistics estimated before deployment. The adaptive classifier uses a fixed projection $\boldsymbol{z}$ of these features. The frozen initial score is $s_0(x)$. Classifiers f_0 and f_t give the initial and current outputs on $\boldsymbol{z}$; their difference corrects $s_0(x)$ with a fixed scale κ. For contribution j, $\psi_j(x)$ is its signed score term and $\gamma_{j,t} \in \{0, 1\}$ indicates whether it is used:

$$s_t(x) = \underbrace{s_0(x) + \kappa\big[f_t(\boldsymbol{z}) - f_0(\boldsymbol{z})\big]}_{r_t(x)\text{: adaptive core}} + \underbrace{\sum_{j=1}^{J} \gamma_{j,t}\psi_j(x)}_{c_t(x)\text{: contributions}} . \tag{34}$$

Thus updating f_t changes the core while preserving the initial score. Changing $\gamma_{j,t}$ suspends or restores a fixed contribution.

A fixed linear sensor measures an image pattern, such as a compression trace. Its weighted output defines a contribution. Let $\boldsymbol{B} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_J]$ collect these sensor directions. We use $\boldsymbol{z} = \boldsymbol{P}_\perp \bar{\boldsymbol{z}}(x)$, where $\boldsymbol{P}_\perp$ projects onto the complement of $\mathrm{span}(\boldsymbol{B})$. This projection prevents the adaptive linear classifier from reconstructing a suspended contribution along its controlled direction.

D.2 Learning and Updating Forensic Coordinates

After making predictions, the reference learner receives an audit batch $\mathcal{A}_t$ of verified images. The reference learner retains class-wise feature statistics for the corresponding audit window. Using the same pre-audit score, the reference learner first evaluates all fixed contributions and applies their use states simultaneously. It then updates the core's coordinates and classifier from the retained statistics. Each image passes through the frozen encoder once. Here t counts audit arrivals, while w indexes a retained window. For label y and window w, the learner stores the count $n_{y,w}$ and the first two unnormalized moments:

$$\mathcal{M}_{y,w} = (n_{y,w}, \boldsymbol{q}_{y,w}, \boldsymbol{S}_{y,w}), \qquad \boldsymbol{q}_{y,w} = \sum_i \boldsymbol{z}_i, \qquad \boldsymbol{S}_{y,w} = \sum_i \boldsymbol{z}_i \boldsymbol{z}_i^\top. \tag{35}$$

The sums run over images with label y in window w. The retained collection $\mathcal{M}_t$ keeps these statistics separately for each label and window. They describe feature means and variation without storing the images, and provide the inputs for learning the coordinates and classifier.

We seek directions that separate real from fake images while limiting within-class variation. For a unit direction $\boldsymbol{u}$ in the adaptive feature space and regularizer $\varepsilon > 0$, the criterion is

$$\rho_t(\boldsymbol{u}) = \frac{\mathrm{Sep}_t(\boldsymbol{u})}{\mathrm{Var}_t(\boldsymbol{u}) + \varepsilon}. \tag{36}$$

Here Sep_t measures real–fake separation, whereas Var_t measures within-class variation. Both quantities are estimated from the retained statistics. The separation matrix $\boldsymbol{A}_t$ combines global separation with separation within individual audit windows. The matrix $\boldsymbol{W}_t$ measures within-class variation across the retained windows. With weight η, setting $\mathrm{Sep}_t(\boldsymbol{u}) = \boldsymbol{u}^\top \boldsymbol{A}_t \boldsymbol{u}$ and $\mathrm{Var}_t(\boldsymbol{u}) = \eta \boldsymbol{u}^\top \boldsymbol{W}_t \boldsymbol{u}$ gives

$$\rho_t(\boldsymbol{u}) = \frac{\boldsymbol{u}^\top \boldsymbol{A}_t \boldsymbol{u}}{\boldsymbol{u}^\top (\eta \boldsymbol{W}_t + \varepsilon \boldsymbol{I}) \boldsymbol{u}}, \qquad \boldsymbol{A}_t \boldsymbol{u} = \lambda(\eta \boldsymbol{W}_t + \varepsilon \boldsymbol{I}) \boldsymbol{u}. \tag{37}$$

The generalized eigenvalue λ equals $\rho_t(\boldsymbol{u})$. We form $\boldsymbol{U}_t$ from the leading eigenvectors under this Fisher-like criterion (Fisher, 1936). We normalize them so that $\boldsymbol{U}_t^\top (\eta \boldsymbol{W}_t + \varepsilon \boldsymbol{I}) \boldsymbol{U}_t = \boldsymbol{I}$, setting the scale of the learned coordinates. A regularized linear classifier f_t is fitted on $\boldsymbol{U}_t^\top \boldsymbol{z}$ using the feature covariance and feature–label covariance recovered from the same statistics. Appendix F.1 gives the estimators, weighting rules, and classifier solution.

The stored statistics cover the full adaptive feature space, including directions absent from the current coordinates. Later updates can therefore select different directions without retrieving earlier images. Adding a window incorporates its observations; removing a window subtracts its counts and moment sums. After either operation, the reference learner recomputes the weights, $\boldsymbol{A}_t$, $\boldsymbol{W}_t$, $\boldsymbol{U}_t$, and f_t from the retained windows. Continual learning and unlearning act on the same statistical representation and use the same recomputation step. With the encoder, projection, initialization, hyperparameters, and solution procedure fixed, removing a window gives the same result as recomputing the core from the remaining windows. Appendix F.1 proves the exact, order-independent updates.

Historical contrasts can reveal the same useful direction even when their signs differ across windows: their outer products accumulate the direction without canceling it. The following result shows when this coordinate-learning step recovers a shared space from finite historical statistics.

Proposition 6 (Identification from historical moments)**.** *Let A_*, W_* be the population counterparts of the reference learner's historical operators. Suppose A_* has rank r and range $\mathcal{S}_*$, and W_* preserves this space and its orthogonal complement. Let $\gamma > 0$ be their smallest positive generalized eigenvalue with metric $G_* = \eta W_* + \varepsilon I$. For the reference learner's estimates, set $\widehat{G} = \eta \widehat{W} + \varepsilon I$ and $\mathcal{E}_H = \|\widehat{A} - A_*\|_{\mathrm{op}} + \|\widehat{G} - G_*\|_{\mathrm{op}}$. For sufficiently small $\mathcal{E}_H$, if the reference learner retains at least the leading r generalized eigenvectors, its retained-space projector $\widehat{P}$ satisfies*

$$\|(I - \widehat{P}) P_*\|_{\mathrm{op}} \le \min\left\{1, C_G \frac{\mathcal{E}_H}{\gamma}\right\}, \tag{38}$$

where P_* projects onto $\mathcal{S}_*$ and C_G depends only on the population operators.

The bound separates historical estimation error from how clearly history identifies the shared directions, measured by γ. Appendix M gives the explicit constant, identification conditions, proof, and finite-history sample bound. It then substitutes this coordinate error into the future Gaussian adaptation bound, keeping the historical and current data costs separate.

D.3 SUSPENDING AND RESTORING CONTRIBUTIONS

The preceding updates change what the core learns from retained observations. The reference learner separately tests whether each controlled contribution should affect the current decision. We fit each sensor h_j from paired views of the same image that differ only in a specified intervention, such as compression. This learns to measure the image pattern without using authenticity labels. An initial audit then jointly fits the scoring weights $\boldsymbol{\beta}$, defining each contribution as $\psi_j(x) = \beta_j h_j(x)$. The sensor and weight remain fixed thereafter, so later audits evaluate the same contribution under changing conditions. Each contribution remains provisional until a later, disjoint audit supports its activation.

At audit t, let $\mathcal{A}_t$ be the labeled audit batch and hold the pre-audit score s_{t^-} fixed. For each contribution j, we form scores without and with it: $s_t^{-j} = s_{t^-} - \gamma_{j,t^-}\psi_j$ and $s_t^{+j} = s_t^{-j} + \psi_j$. Using binary cross-entropy loss ℓ, its current value is the population risk reduction

$$V_{j,t} = \mathbb{E}_{(X,Y)\sim P_{\omega_t}}\left[\ell\Big(Y, s_t^{-j}(X)\Big) - \ell\Big(Y, s_t^{+j}(X)\Big)\right]. \tag{39}$$

A positive $V_{j,t}$ means that the contribution reduces loss; a negative value means that using it increases loss. The audit estimates $V_{j,t}$ by averaging these loss differences over $\mathcal{A}_t$ (Equation 51). For a mixed-world audit, this average estimates the contribution's value under the audit mixture. The controller combines estimates across successive audits into a discounted Gaussian confidence score $b_{j,t}$ for positive current value. It also applies a separate check $R_{j,t}$ to limit the increase in false positives on the audited real images. Appendix F specifies the confidence model and this check.

Each contribution has a use state $a_{j,t} \in \{\text{PROVISIONAL}, \text{ACTIVE}, \text{DORMANT}\}$, initialized as PROVISIONAL. Only active contributions affect the score; dormant contributions remain available for restoration. The controller uses a higher confidence threshold for activation (τ_+) than for suspension (τ_-). This gap reduces repeated switching:

$$a_{j,t} = \begin{cases} \text{ACTIVE}, & b_{j,t} > \tau_+ \text{ and } R_{j,t} = 1, \\ \text{DORMANT}, & a_{j,t-1} = \text{ACTIVE and } (b_{j,t} < \tau_- \text{ or } R_{j,t} = 0), \\ a_{j,t-1}, & \text{otherwise.} \end{cases} \tag{40}$$

The use indicator is $\gamma_{j,t} = \mathbf{1}[a_{j,t} = \text{ACTIVE}]$. Suspension removes the contribution from the score while preserving its sensor and scoring weight for restoration. This changes the use of a retained contribution, whereas removing a window changes the statistics used to learn the core.

Proposition 7 (Exact withdrawal without linear re-entry)**.** *For the linear core in Equation 34, changing $\gamma_{j,t}$ from one to zero removes exactly $\psi_j(x)$ from the score, while $\boldsymbol{P}_\perp \boldsymbol{v}_j = \mathbf{0}$ prevents the adaptive core from linearly reconstructing that contribution along the controlled trace direction.*

Appendix F.3 proves the statement. The guarantee concerns the explicit score contribution and the adaptive linear core; the frozen initial score remains unchanged.

Algorithm 2 specifies the corresponding contribution-control update.

E FROM TRACE TO EVIDENCE: CONCEPTUAL AND THEORETICAL FOUNDATIONS

A detector observes an image pattern, assigns it a score contribution, and must determine whether that contribution is useful under the current conditions. This appendix distinguishes these objects, then explains why robustness to source changes and processing changes need not compose:

$$\text{visual trace} \longrightarrow \text{fixed score contribution} \longrightarrow \text{current value} \longrightarrow \text{use decision.}$$

Trace, evidence, and contribution. A trace is an observable image pattern. The sensor response $h_j(x)$ measures how strongly trace j is present in image x. The initial audit uses authenticity labels to learn how that response should affect the forensic score. The signed contribution $\psi_j(x) = \beta_j h_j(x)$ gives its scoring orientation. Whether that orientation is useful depends on the current world: including ψ_j is justified only when it reduces decision risk there. The controller estimates this

Table 5: Objects that a static forensic score usually collapses. The image reveals a trace response; only a verified audit can support a decision about its current use.

Level	Object	Question	Available when?
Image	$h_j(x)$	How strongly is trace j present?	Prediction
Score	$\psi_j(x)$	How would that trace change this score?	Prediction
World	$V_{j,t}$	Would including it reduce expected risk now?	Latent
Audit	$b_{j,t}$	How strongly do verified samples support $V_{j,t} > 0$?	After audit
Action	$\gamma_{j,t}$	May the contribution affect the current verdict?	After audit

value from post-prediction audits and records whether the contribution may enter the score. Table 5 separates the observation, score contribution, current value, audit confidence, and use decision.

Detecting a trace and deciding to use it are therefore different operations. A trace can remain visible after its contribution becomes harmful, and a withdrawn contribution can become useful again. The current world determines evidential value, the audit estimates it, and the use state controls the score.

Why a forensic world has two coordinates. The source pair g specifies where real and fake content comes from. The processing chain e describes label-preserving operations such as compression, transcoding, recapture, and restoration. These processes are distinct, but their forensic effects interact. A propagation chain can erase a generator trace or induce a similar response in real images. The world is therefore indexed by $\omega = (g, e)$ even when evaluation summarizes novelty with the coarser cells $Q_{00}, Q_{10}, Q_{01}, Q_{11}$.

Consider a detector contribution that responds to a loss of fine-scale detail. After platform transcoding, its sensor output may remain large, so the trace is still present. If the platform induces the same response in real images, however, the contribution can increase false accusations. The trace remains detectable, but its contribution no longer supports the same decision. A later world may make that fixed contribution useful again.

Why marginal robustness does not compose. Let $\boldsymbol{Z} = \phi(X)$ denote encoded features. For a unit direction $\boldsymbol{u}$, define the signed class gap

$$\Delta_\omega(\boldsymbol{u}) = \mathbb{E}[\boldsymbol{u}^\top \boldsymbol{Z} \mid Y = 1, \omega] - \mathbb{E}[\boldsymbol{u}^\top \boldsymbol{Z} \mid Y = 0, \omega].$$

Orient the direction so that larger projections support $Y = 1$ in the reference world. Because the gap depends jointly on (g, e), it can weaken or reverse as the world changes.

Proposition 8 (Marginal robustness need not compose)**.** *Without a factorization or bounded-interaction assumption on $\Delta_\omega(\boldsymbol{u})$, positive gaps when only the source changes and when only processing changes place no sign constraint on the gap when both change.*

The signed gap $\Delta_{g,e}(\boldsymbol{u})$ admits the exact decomposition

$$\Delta_{g,e}(\boldsymbol{u}) = \Delta_{00}(\boldsymbol{u}) + \alpha_g(\boldsymbol{u}) + \beta_e(\boldsymbol{u}) + \chi_{g,e}(\boldsymbol{u}), \tag{41}$$

where α_g and β_e are the changes observed along the two axes and $\chi_{g,e}$ is their interaction. This is an accounting identity, not a model assumption: choose the reference source pair g_0 and processing chain e_0, define $\alpha_g = \Delta_{g,e_0} - \Delta_{00}$ and $\beta_e = \Delta_{g_0,e} - \Delta_{00}$, and assign the remainder to χ.

Proof of Proposition 8. For a new source pair g_1 and processing chain e_1, the two single-axis tests reveal $\Delta_{g_1,e_0} = \Delta_{00} + \alpha_{g_1}$ and $\Delta_{g_0,e_1} = \Delta_{00} + \beta_{e_1}$. Neither expression contains χ_{g_1,e_1}. For any fixed positive values of the two observed gaps, choosing a sufficiently negative interaction makes $\Delta_{g_1,e_1} = \Delta_{00} + \alpha_{g_1} + \beta_{e_1} + \chi_{g_1,e_1} < 0$. Thus the joint-world sign is unconstrained unless the interaction is removed by factorization or bounded by an additional assumption. □

The role of Q_{11} is therefore to evaluate the interaction that the two single-axis tests leave unidentified.

Relation to neighboring settings. Open-world recognition usually opens the semantic label space (Bendale & Boult, 2015; Fontanel et al., 2021); our label space remains binary while the mechanisms that generate observations remain open. Domain generalization tests transfer without post-deployment labels, whereas our protocol reveals a small verified audit only after prediction. Continual learning preserves competence across tasks, often through regularization, constraints, or replay (Chaudhry et al., 2019); continual test-time adaptation tracks a changing target distribution (Wang et al., 2022). Neither objective alone determines whether a retained contribution should still support the same verdict. Here, withdrawal suspends an identified score contribution while retaining its sensor and weight; it makes no claim of training-data deletion or machine unlearning. This control operation is distinct from removing an audit window's moments when unlearning its influence on the core. Recoverable forgetting has a related control intent, but acts on a different object and provides a different guarantee (Ye et al., 2022; Zhao et al., 2024).

Table 6: Notation used in the method and appendices.

Symbol	Meaning	Symbol	Meaning
$\omega_t = (g_t, e_t)$	Forensic world at audit t	$\mathcal{A}_t$	Verified audit batch received after prediction
$r_t(x)$	Adaptive core score	$c_t(x)$	Sum of contributions
$h_j(x)$	Response of sensor j	$\psi_j(x)$	Fixed signed contribution of trace j
$V_{j,t}$	Population risk reduction from including j	$b_{j,t}$	Audit-based confidence that $V_{j,t} > 0$
$a_{j,t}$	Provisional, active, or dormant state	$\gamma_{j,t}$	Binary permission to include ψ_j
$\mathcal{M}_{y,w}$	Aggregate moments for label y and window w	$[\underline{s}_t, \overline{s}_t]$	Envelope for unresolved contributions

F Reference Estimators and Contribution-Control Details

This appendix specifies the generalized-eigenvector reference in Appendix D, followed by the contribution-control definitions used in the reference evaluations. The joint-moment objective and its exact solution are specified separately in Appendix B.

F.1 Learning Forensic Coordinates

Let $z = P_\perp \tilde{z}(x)$. For label y and anonymous audit window w, store only $\mathcal{M}_{y,w} = (n_{y,w}, q_{y,w}, S_{y,w})$, where $q_{y,w} = \sum_i z_i$ and $S_{y,w} = \sum_i z_i z_i^\top$. To cap each cell's effective sample count, set $c_{y,w} = \min\{1, n_{\max}/n_{y,w}\}$. The resulting counts and means are

$$N_y = \sum_w c_{y,w} n_{y,w}, \qquad \mu_y = N_y^{-1} \sum_w c_{y,w} q_{y,w}, \qquad N = \sum_y N_y, \qquad \mu = N^{-1} \sum_y N_y \mu_y,$$
$$\mu_{y,w} = q_{y,w}/n_{y,w}, \qquad N_w = \sum_y c_{y,w} n_{y,w}. \tag{42}$$

Let $\mathcal{W}_\pm$ contain the audit windows in which both labels occur. The separation and within-class variation operators are

$$A_t = \sum_{y=0}^{1} \frac{N_y}{N} (\mu_y - \mu)(\mu_y - \mu)^\top + \sum_{w \in \mathcal{W}_\pm} \frac{N_w}{N} (\mu_{1,w} - \mu_{0,w})(\mu_{1,w} - \mu_{0,w})^\top, \tag{43}$$

$$W_t = \frac{1}{N} \sum_{y=0}^{1} \left(\sum_w c_{y,w} S_{y,w} - N_y \mu_y \mu_y^\top \right). \tag{44}$$

The first term of A_t measures global real–fake separation; the second measures separation within audit windows. The operator W_t penalizes within-class variation. We form the columns of U_t from the leading generalized eigenvectors of $A_t u = \lambda G_t u$, where $G_t = \eta W_t + \varepsilon I$, and normalize them so that $U_t^\top G_t U_t = I$. All terms are reconstructed from the additive moments of retained audit windows. This makes it possible to add or remove windows without storing individual feature vectors.

Exact window algebra. For a set of retained windows $\mathcal{R}$, write $\mathcal{M}_y(\mathcal{R}) = \sum_{w\in\mathcal{R}} \mathcal{M}_{y,w}$, where addition acts componentwise on (n, q, S). The estimators above read only these per-window cells and their deterministic weights.

Lemma 1 (Exact and order-independent moment updates)**.** *For any retained window set $\mathcal{R}$, permuting the order of additions does not change its aggregate moments, A_t, or W_t. Removing window w by subtracting its cell is identical to recomputing these quantities from $\mathcal{R} \setminus \{w\}$.*

Proof. Counts, first moments, and second moments form a commutative additive tuple. Each cap $c_{y,w}$ depends only on its own cell, so the sums in Equations 42–44 depend on the set of retained cells, not their insertion order. Subtracting one cell leaves exactly the sums indexed by $\mathcal{R} \setminus \{w\}$. □

The same cells also determine the classifier fitted on the learned coordinates. Let $\widetilde{Y} = 2Y - 1$, and let μ_z, $\mu_{\tilde{y}}$, C_t, and c_t denote the capped-weight means, feature covariance, and feature–label cross-covariance:

$$C_t = \mathbb{E}_c[(z - \mu_z)(z - \mu_z)^\top], \qquad c_t = \mathbb{E}_c[(z - \mu_z)(\widetilde{Y} - \mu_{\tilde{y}})]. \tag{45}$$

Here $\mathbb{E}_c$ uses the cell weights in Equation 42. The vector c_t is the feature–label cross-covariance, distinct from the controlled score $c_t(x)$. Writing $C_{U,t} = U_t^\top C_t U_t$, $r = \dim(U_t)$, and $\tau_t = \mathrm{tr}(C_{U,t})/r$, the regularized moment readout is

$$\boldsymbol{\theta}_t = [C_{U,t} + (\lambda_h + \xi_t)\tau_t I]^{-1} U_t^\top c_t, \quad \beta_{0,t} = \mu_{\tilde{y}} - \boldsymbol{\theta}_t^\top U_t^\top \mu_z, \quad f_t(z) = \boldsymbol{\theta}_t^\top U_t^\top z + \beta_{0,t}. \tag{46}$$

The implementation selects the first $\xi_t \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$ whose Cholesky solution passes a relative-residual tolerance of 10^{-8}; when $r = 0$, it sets $f_t \equiv 0$. Thus the stored moments determine both the forensic coordinates and the classifier fitted on them.

F.2 Controlling Evidence Use

Identifiable trace construction. For trace j, let (x_i^-, x_i^+) be same-identity views that differ only by the intervention of interest, and assign paired-view indicators $v_i^- = 0$ and $v_i^+ = 1$. The sensor is fitted without authenticity labels:

$$(u_j, c_j) = \arg\min_{u,c} \sum_{i,q\in\{-,+\}} \ell\big(v_i^q, u^\top \bar{z}(x_i^q) + c\big) + \frac{\lambda_u}{2}\|u\|_2^2, \qquad h_j(x) = u_j^\top \bar{z}(x) + c_j. \tag{47}$$

To determine how the sensor responses affect the forensic score, the initial audit jointly fits their signed scoring weights:

$$\boldsymbol{\beta} = \arg\min_{b} \sum_{(x_i,y_i)\in\mathcal{A}_0} \ell\Big(y_i, r_0(x_i) + \textstyle\sum_j b_j h_j(x_i)\Big) + \frac{\lambda_\beta}{2}\|b\|_2^2, \qquad \psi_j(x) = \beta_j h_j(x). \tag{48}$$

The implementation stores the equivalent centered sensor

$$h_j(x) = v_j^\top (\bar{z}(x) - \bar{m}_j), \qquad \psi_j(x) = \alpha_j h_j(x). \tag{49}$$

This parameterization has $u_j = v_j$, $c_j = -v_j^\top \bar{m}_j$, and $\beta_j = \alpha_j$. The serialized contribution contains $(v_j, \bar{m}_j, \alpha_j)$ and bindings to the encoder, scaler, paired intervention data, and initial audit. Its identifier is the SHA-256 digest of a canonical byte representation. Deserialization recomputes both the bytes and the digest; a mismatch raises an error.

Sensor fitting and contribution activation use different observations. Same-identity intervention pairs teach the sensor to detect a trace without authenticity labels. The initial audit determines its signed scoring weight, but the contribution remains provisional until a later, disjoint audit supports activation. The sample defining the contribution is therefore separate from the sample authorizing its first use.

Risk reduction and discounted confidence. The implementation uses binary cross-entropy on logits,

$$\ell_{\mathrm{BCE}}(y, s) = \log(1 + \exp s) - ys. \tag{50}$$

For a frozen audit batch, let $d_{i,j,t} = \ell_{\mathrm{BCE}}(y_i, s_t^{-j}(x_i)) - \ell_{\mathrm{BCE}}(y_i, s_t^{+j}(x_i))$ and

$$\bar{d}_{j,t} = \frac{1}{n_t}\sum_i d_{i,j,t}, \qquad \nu_{j,t} = \frac{\mathrm{Var}_i(d_{i,j,t})}{n_t} + \varepsilon_b. \tag{51}$$

Let the preceding confidence state be $\mathcal{N}(\mu^V_{j,t-1}, (\sigma^V_{j,t-1})^2)$. Under the Gaussian working model, discounting its precision by $\delta \in (0,1]$ and combining it with the audit observation gives

$$(\sigma^V_{j,t})^{-2} = \delta(\sigma^V_{j,t-1})^{-2} + \nu_{j,t}^{-1},$$

$$\mu^V_{j,t} = (\sigma^V_{j,t})^2\left[\delta\mu^V_{j,t-1}(\sigma^V_{j,t-1})^{-2} + \bar{d}_{j,t}\nu_{j,t}^{-1}\right], \qquad b_{j,t} = \Phi\left(\frac{\mu^V_{j,t}}{\sigma^V_{j,t}}\right). \tag{52}$$

The score $b_{j,t}$ expresses confidence in positive current value under the Gaussian working model. The empirical risk gate below checks a different quantity: the change in audited false positives. Withdrawal closes the current reliability episode. At the next audit, the trace starts from $\mathcal{N}(0,1)$ with a new epoch identifier. The ledger retains the preceding posterior, but the new episode does not use it. This confidence model organizes sequential decisions; it does not provide a distribution-free guarantee of positive query utility. The budget study evaluates its behavior as the available audit changes.

Empirical FPR constraint. Average loss and false accusations are tested separately. Let $\mathrm{FP}_{95}(y, s)$ be the number of audited real images above the largest threshold that attains at least 95% true-positive rate. The implemented gate is

$$R_{j,t} = \mathbf{1}\left[\mathrm{FP}_{95}(y, s_t^{+j}) - \mathrm{FP}_{95}(y, s_t^{-j}) \le 1\right]. \tag{53}$$

Activation therefore requires both confidence in positive average gain and an acceptable increase in audited false positives. This requirement applies to provisional and dormant contributions alike.

Score envelope. Let $\mathcal{U}_t = \{j : a_{j,t} \neq \textsc{Active}\}$ collect the contributions currently excluded from the score. Allowing each of them to be included or excluded gives the exact algebraic envelope

$$\underline{s}_t(x) = s_t(x;\boldsymbol{\gamma}_t) + \sum_{j\in\mathcal{U}_t}\min\{0,\psi_j(x)\}, \qquad \overline{s}_t(x) = s_t(x;\boldsymbol{\gamma}_t) + \sum_{j\in\mathcal{U}_t}\max\{0,\psi_j(x)\}. \tag{54}$$

It is a deterministic score range, not a calibrated predictive interval or a reported detection metric.

F.3 Online Execution and Structural Results

CUE separates prediction from learning through two operations. PREDICT reads images and the current state without changing it. AUDIT uses one verified batch to update all contribution states simultaneously, then updates the adaptive core. The audit index records update arrivals; source identities, processing identities, and indicators of condition changes are not inputs.

Algorithm 1 Prediction from the current evidence state

Require: Image batch X; core r_t; contributions $\{\psi_j, \mathrm{id}_j\}_{j=1}^J$; states $\{a_{j,t}\}_{j=1}^J$

```
s ← r_t(X); s_lo ← r_t(X); s_hi ← r_t(X)
for j = 1, ..., J do
    e_j ← ψ_j(X)
    if a_{j,t} = ACTIVE then
        s ← s + e_j; s_lo ← s_lo + e_j; s_hi ← s_hi + e_j
    else
        s_lo ← s_lo + min(0, e_j); s_hi ← s_hi + max(0, e_j)
    end if
end for
return forensic score s, conservative accusation score s_lo, interval [s_lo, s_hi], and contribution states
```

An active contribution is added to the forensic score and to both endpoints of the score envelope. A provisional or dormant trace is excluded from the ordinary forensic score, but the envelope still includes its two possible effects. The lower endpoint is used for a positive accusation. Prediction reads no labels and leaves the controller state unchanged.

Algorithm 2 Post-prediction audit with simultaneous state updates

Require: Audit index t; verified batch $\mathcal{A}_t$; core scores r_t; contributions $\{\psi_j\}$; current confidence states and use states

Fix $r_t(\mathcal{A}_t)$, all $\psi_j(\mathcal{A}_t)$, and all current states for this audit
$s^{\text{pre}} \leftarrow r_t(\mathcal{A}_t) + \sum_{j:a_{j,t^-}=\text{ACTIVE}} \psi_j(\mathcal{A}_t)$
for $j = 1, \ldots, J$ **do**
    Make an exact copy of controller state j
    $s^{-j} \leftarrow s^{\text{pre}} - \gamma_{j,t^-}\psi_j$; $s^{+j} \leftarrow s^{-j} + \psi_j$
    Compute BCE gains, update $(\mu^V_{j,t}, \sigma^V_{j,t}, b_{j,t})$, and evaluate $R_{j,t}$
    Propose BIRTH, RETIRE, REACTIVATE, or HOLD on the copy
end for
Apply all proposed controller states simultaneously
Record all contribution updates, linked by a hash to the preceding audit record
Update the adaptive core only after all controller states have been updated

State transitions. The dead band $\tau_- \le b \le \tau_+$ prevents an audit near the decision boundary from repeatedly toggling a contribution. The transition rules are

Current state	Audit condition	Next state and event
Provisional	$b > \tau_+$ and $R = 1$	Active, BIRTH
Active	$b < \tau_-$ or $R = 0$	Dormant, RETIRE
Dormant	$b > \tau_+$ and $R = 1$	Active, REACTIVATE
Any state	otherwise	unchanged, HOLD

The lifecycle experiments use $\delta = 0.25$, $\tau_- = 0.2$, and $\tau_+ = 0.8$. These values are fixed before evaluating the stream.

Persistent state and audit trail. Each stored contribution has a fixed content-derived identifier, its own confidence state, and a reliability episode. A bank event records the state vector before and after the update, every contribution update, and the digest of the preceding event. Each contribution update records the observed gain and variance, posterior, risk counts, action, state, episode, contribution identity, and preceding hash. Restoration succeeds only when the contribution bytes, identities, controller state, and both hash chains reproduce exactly.

Withdrawal changes the use state and the resulting score while preserving the contribution bytes. Restoration reactivates the same contribution, with its sensor and contribution identity unchanged.

Structural consequences. The score decomposition gives two exact properties. Withdrawal removes the identified contribution, and simultaneous state updates produce the same result regardless of the order in which contributions are evaluated.

Proof of Proposition 7. Consider two state vectors that differ only at coordinate j. Subtracting the corresponding instances of Equation 34 cancels the core and every $k \neq j$ term, leaving $(\gamma'_{j,t} - \gamma_{j,t})\psi_j(x)$. Switching from one to zero therefore removes exactly $\psi_j(x)$. For the second part, let Q_B span the controlled trace directions and $P_\perp = I - Q_B Q_B^\top$. For every $b \in \text{span}(B)$, $P_\perp b = 0$. A linear core coefficient of the form $P_\perp^\top w$ consequently has inner product $w^\top P_\perp b = 0$ with every controlled direction. The adaptive core cannot linearly reintroduce the withdrawn contribution along that span. □

Lemma 2 (Order-independent audit update)**.** *If every proposed contribution update is computed from the same fixed pre-action score, the order of evaluating contributions does not change the resulting state vector.*

Proof. The counterfactual pair for j depends only on the fixed pre-action score, the pre-action state γ_{j,t^-}, and ψ_j. It never reads another proposed update. Each update is therefore a coordinate-wise function of fixed inputs, and applying all updates simultaneously is permutation invariant. □

Proposition 9 (Exact score envelope)**.** *For unresolved contributions $\mathcal{U}_t$, every score obtained by assigning binary use states to those contributions lies in $[\underline{s}_t(x), \overline{s}_t(x)]$ from Equation 54.*

Proof. Each unresolved contribution is either 0 or $\psi_j(x)$ and therefore lies in $[\min\{0, \psi_j(x)\}, \max\{0, \psi_j(x)\}]$. Summing these coordinate-wise bounds with the active score proves the result. No distributional approximation is involved. □

G Complete Evaluation Protocol

Query labels are used only by a separate evaluator to compute metrics. In continual operation, each incoming query batch is scored before the next labeled audit batch updates the detector. The query scores, configuration, model state, and data-access record are fixed at prediction time. Post-audit and end-of-stream evaluations separately measure adaptation and retention. Few-shot evaluations adapt to each support budget before scoring held-out query images.

G.1 Worlds and Information Interface

Table 7: Information available at prediction and audit time.

Phase	Images	Audit labels	Query labels	World identity	Change point
Predict	yes	no	no	no	no
Audit	designated support only	yes	no	no	no
Independent evaluation	sealed scores	no	yes	evaluator only	evaluator only

What the model is allowed to observe. All splits are formed at the base-image identity level. Transformed views of the same image form one statistical unit and stay within the same support, query, or stream role.

Thirty-two open worlds. The coverage experiment defines 32 worlds by their source and processing conditions. Each contains 256 real and 256 generated or manipulated images. All reproduced methods use the same image identities within a world. Table 8 groups the worlds by what changes.

Table 8: Composition of the complete open-world evaluation.

World group	Count	What changes
Reference worlds	2	Familiar generator and clean export conditions
Specific generators	6	Generator identity and generation family
Industrial content sources	6	GAN, diffusion, frontier generation, personalization, generative editing, and hybrid content
Processing conditions	6	Clean export, web processing, platform transcoding, repeated forwarding, recapture, and trace cleaning
Joint worlds	12	A content source and a processing chain change together
Total	32	g, e, and their interaction

CNNSpot, NPR, UnivFD, IAPL, and CUE are evaluated on the same query images. The six-benchmark comparison follows the query protocol of each benchmark and uses different query sets from the additional AIGC-DB and GenImage comparison in Appendix H.1. Each baseline retains its native configuration. Results are grouped by query protocol and are not averaged across incompatible protocols. Per-image scores are retained for the metrics defined in Section G.3.

G.2 Adaptation, Continual Streams, and Evidence Lifecycle

Few-shot adaptation. A K-shot support set contains K real and K fake images, giving $2K$ labeled images in total. We use

$$K \in \{0, 1, 2, 4, 8, 12, 16, 24, 32, 48, 64\}. \tag{55}$$

Support sets are nested prefixes of one fixed hash ordering. Support and query identities remain disjoint. The six target worlds cover a frontier generator, a fine-tuned generator, generative editing, social-platform transcoding, screen recapture, and a jointly new Q_{11} world. UnivFD adapts a linear head, NPR follows its standard fine-tuning route, IAPL adapts its prompt, and CUE updates from the same audit budget. After adaptation, query AUROC is plotted against K. Area under the adaptation curve and the first K reaching AUROC 0.90 summarize label efficiency; FPR95 measures false positives at high detection sensitivity.

Continual worlds without condition identities. The continual experiment changes the arrival order of the same worlds. Generators arrive before propagation conditions in the generator-first path; the propagation-first path reverses this order. The displayed stream contains four new generators, two fine-tuned generators, four propagation operations, and two compound propagation pipelines. The latter use RRDataset's released internet-transmission and re-digitization classes (Li et al., 2025a). The public archive provides transmission and re-digitization classes without verifiable platform-route identities. We therefore label these worlds *Multi-platform* and *Re-digitized*. Every method receives 64 real and 64 fake images in each audit batch and no world identifier. Frozen, fine-tuning, SimpleShot, SLDA, RanPAC, LayUP, byte-matched A-GEM, the frozen SigLIP 2 linear probe, and CUE share the same audit batches, update opportunities, world order, and query identities.

For each world, the score archive gives AUROC immediately after its audit and at the end of the stream. The first measures adaptation on arrival; the change between the two measures retention through later updates. Table 2 summarizes final AUROC on the initial condition, AUROC over pooled query images from all conditions, and the maximum post-audit-to-final decline in percentage points, alongside selected conditions. Table 14 reports final AUROC for every generator and propagation condition. We also compute final average performance, worst-world AUROC, forgetting, backward transfer, and FPR95 from the same score archive.

Fine-tuned lineage transfer. We evaluate two parent–child lineages using 256 real and 256 fake query identities. Fake images are paired by prompt and seed. We estimate parent–child differences with 10,000 label-stratified paired bootstrap resamples. In the controlled SDXL pair, base weights, prompts, seeds, sampler, inference settings, and PNG output are fixed, and only the DreamBooth-LoRA adapter is toggled. The SD1.5/LCM-LoRA pair provides a second lineage shift with a looser visual pairing. Appendix I.1 reports the paired results and their interpretation.

Mixed-world audit stress test. We relax the one-world-per-window convention using three world pairs: SD3.5/FLUX.1, JPEG/WebP, and Multi-platform/Re-digitized. Each anonymous window contains 16 audited images per class. We test seven mixture ratios over four disjoint identity draws, comparing matched audit/query mixtures with asynchronous compositions. All comparisons use 10,000 stratified paired bootstrap resamples. Appendix I.1 reports the resulting interference boundary.

Fixed-contribution lifecycle trajectories. The additional controller evaluation follows two fixed contributions across five query trajectories. Firefly Image 3, Ideogram V1, and LaVi-Bridge share one blur contribution, with separate controller histories. The LaVi-Bridge stream uses generated images and Flickr8k photographs from the DFBench test split.

EPAIQA-15K supplies paired original and edited images. Its shared contribution measures JPEG chroma subsampling: 4:4:4 versus 4:2:0 at quality 90. The direction is learned from projected features of the two JPEG views of each image using view labels. Its scoring weight is fitted on the first labeled audit and then frozen. The query groups Flickr-farm and numeric contain images whose original filename stems begin with `farm` or consist of digits, respectively. These groups share the same contribution and controller; filenames serve only to group evaluation results.

Action metrics use the four distinct controller histories, containing 23 audits. Query metrics retain all five trajectories. The three blur histories use the epochal Gaussian controller; the EPAIQA-15K history uses a Schmitt switching rule. Both preserve a fixed contribution while changing its use.

The control ablations replay these recorded histories and paired score branches. They test individual use, withdrawal, restoration, episode reset, and the conservative accusation bound. The recorded trajectories describe separate control decisions, so this comparison does not include a synchronized global switch. The three-contribution bank in the resource study tests implementation cost and

transactions. Appendix H.4 specifies the audit inputs and held-out query identities used in the additional budget study.

G.3 Metrics and Statistical Analysis

Lifecycle measurements. The evaluator, not the model, knows the trace-validity schedule. For contribution j at time t, the effect of including it is measured by the paired score change $|s_t^{+j}(x) - s_t^{-j}(x)| = |\psi_j(x)|$. Unsupported evidence exposure is

$$\mathrm{UEE} = \sum_{t,j} \gamma_{j,t}\, \mathbb{E}_x |\psi_j(x)|\, [-V_j(t)]_+, \tag{56}$$

This quantity weights score influence by the magnitude of negative utility. The control ablation measures a phase-weighted score-exposure proxy instead, using $B = 640$ fixed query batches:

$$E_{\mathrm{phase}} = \frac{1}{B} \sum_{b=1}^{B} \gamma_b \left(\frac{1}{|Q_b|} \sum_{x \in Q_b} |s_{\mathrm{on}}(x) - s_{\mathrm{off}}(x)| \right) \mathbf{1}\{b \text{ belongs to an invalid or dormant phase}\}. \tag{57}$$

This is the "Exp." column in Table 26; it uses a phase indicator and a batch mean rather than the magnitude of negative population utility. False withdrawal uses the 13 valid audits. Action accuracy uses all 23 audits after deduplicating the shared EPAIQA-15K controller history. Withdrawal and restoration delays count completed audits from the first corresponding validity transition. A state change at that audit has delay zero; exposure before the audit arrives is outside this delay measure. Query AUROC and shared-threshold FPR are reported separately from state decisions. Restoration reuses the original contribution identity and its unchanged serialized parameters. In Table 3, shading identifies the full CUE configuration; bold and underlined values mark the best and second-best results for each metric. R/A lists withdrawal and restoration delays in that order.

Statistical unit. Detection comparisons with intervals use 10,000 paired bootstrap resamples of base-image identities, stratified by world and label. Transformed views and repeated scores remain grouped by identity. The control table gives point estimates from four controller histories and five query trajectories, without trajectory-level confidence intervals. The budget study repeatedly samples audits within fixed streams, so its variation describes audit selection rather than variation across independent streams. The evaluation records identify the images used for audits, post-audit scoring, and held-out query evaluation.

H Extended Results

This appendix gives the complete conventional and open-world results, few-shot and continual evaluations, and contribution lifecycle measurements. Open-world metrics are computed from fixed query scores under the protocols in Appendix G.

H.1 Conventional and Open-World Performance

Table 9: **Complete results on six forensic benchmarks.** All values are percentages. AUROC and average precision (AP) are higher-is-better; F95 is the false-positive rate at 95% true-positive rate (lower is better). The AUROC values also appear in Table 1. Shaded rows identify CUE; bold and underlined values mark the best and second-best results, respectively.

Method	Generator						Forgery						Domain					
	UFD			GenImage			AIGC-DB			AIGIBench			WildFake			Chameleon		
	AUC↑	AP↑	F95↓	AUC↑	AP↑	F95↓	AUC↑	AP↑	F95↓	AUC↑	AP↑	F95↓	AUC↑	AP↑	F95↓	AUC↑	AP↑	F95↓
Trace-Centric Detectors																		
CNNDet *(CVPR'20)*	91.2	92.4	52.0	56.6	56.8	100.0	71.6	73.9	100.0	67.2	68.6	100.0	57.5	60.2	100.0	37.8	44.5	100.0
DIRE *(ICCV'23)*	56.6	53.7	85.9	49.6	49.9	96.1	49.5	49.8	94.5	50.9	50.5	97.3	55.8	53.2	85.5	63.1	57.8	76.6
PatchCraft *(arXiv'23)*	76.8	69.0	48.0	78.8	76.0	69.9	84.2	84.7	63.3	80.3	85.3	91.8	83.7	87.3	81.2	43.2	47.7	91.8
FreqNet *(AAAI'24)*	91.5	89.0	31.2	54.7	53.3	100.0	61.1	59.9	100.0	70.3	69.6	100.0	78.8	81.5	100.0	53.2	55.2	100.0
NPR *(CVPR'24)*	80.2	71.6	35.5	46.8	49.3	100.0	42.0	49.4	100.0	72.0	68.5	100.0	88.3	90.3	100.0	52.5	56.9	100.0
DRCT *(ICML'24)*	68.7	70.9	89.8	97.5	97.5	13.7	88.6	89.9	55.9	85.1	87.7	66.4	82.2	85.5	81.2	73.1	74.5	86.3
SAFE *(KDD'25)*	80.9	68.9	32.0	52.0	53.4	88.7	59.5	55.4	93.8	82.5	84.7	86.3	89.1	92.6	80.9	54.3	54.3	91.0
FerretNet	90.6	89.8	32.8	44.3	48.8	97.3	38.3	43.9	99.2	86.2	90.1	79.7	90.7	93.9	69.1	42.8	49.3	94.5
Representation-Centric Detectors																		
UnivFD *(CVPR'23)*	93.8	94.5	33.6	74.3	75.8	79.3	85.2	87.6	71.5	67.8	73.4	88.3	68.6	73.0	86.3	43.0	45.2	97.7
OmniAID	95.7	96.4	24.6	92.8	93.9	43.8	91.0	91.1	37.5	93.4	94.5	41.0	92.3	94.1	48.4	91.3	93.3	59.0
C2P-CLIP *(AAAI'25)*	70.3	68.9	66.0	91.2	90.9	39.1	87.7	85.8	36.3	77.2	72.8	63.3	84.4	85.4	65.6	56.0	58.1	93.0
RINE	98.1	98.4	10.2	93.2	91.9	21.9	94.2	94.6	26.6	75.6	79.5	100.0	87.1	90.4	75.0	51.9	53.5	100.0
Effort *(ICML'25)*	81.0	76.5	51.6	90.5	90.4	39.8	87.1	86.6	45.7	82.2	82.8	66.0	90.2	92.5	59.4	52.5	53.9	94.9
IAPL *(CVPR'26)*	98.3	**98.8**	**3.5**	79.4	81.1	77.7	79.4	85.8	90.2	75.4	80.2	84.4	87.2	90.0	66.8	31.4	42.7	100.0
DDA	90.7	92.0	47.3	98.2	98.4	11.7	95.7	96.3	32.8	91.2	93.5	58.6	91.1	93.5	55.5	91.6	92.8	48.8
Diversity-Centric Detectors																		
AIDE *(ICLR'25)*	92.9	93.7	49.6	71.5	71.5	68.0	76.9	77.4	64.1	71.1	75.9	85.9	81.5	74.9	68.0	44.5	48.1	95.3
CO-SPY *(CVPR'25)*	83.2	75.5	31.6	52.3	53.7	100.0	47.9	55.6	100.0	77.5	79.6	74.2	88.6	91.5	100.0	52.8	57.3	100.0
D^3 *(CVPR'25)*	97.6	97.7	11.7	**99.6**	**99.6**	**2.3**	97.7	97.9	10.9	89.5	90.0	49.6	94.4	95.4	35.2	69.0	67.1	83.2
GAPL *(CVPR'26)*	**98.5**	98.7	9.4	98.5	98.9	4.3	97.9	98.2	9.8	95.7	96.5	29.7	98.4	98.8	5.5	80.9	82.9	74.2
CommunityForensics	97.6	98.1	17.6	98.9	99.2	3.1	**99.1**	**99.1**	**5.9**	96.0	96.1	21.1	**99.4**	**99.4**	**4.3**	92.6	93.3	40.2
LOTA	68.6	62.4	66.0	55.8	56.8	100.0	52.7	51.4	100.0	78.5	70.8	54.3	90.2	93.6	100.0	50.3	53.9	100.0
DEAR	77.1	75.7	73.0	97.7	98.2	16.8	85.5	88.9	80.5	81.2	86.4	86.3	88.6	90.7	71.9	86.1	87.8	50.4
CUE *(ours, zero-shot)*	97.6	98.1	17.6	98.9	99.2	3.1	**99.1**	**99.1**	**5.9**	96.0	96.1	21.1	**99.4**	**99.4**	**4.3**	92.6	93.3	40.2
CUE *(ours, few-shot)*	98.4	98.7	5.5	98.9	99.2	3.1	**99.1**	**99.1**	**5.9**	**96.7**	**96.7**	**16.0**	**99.4**	**99.4**	**4.3**	**93.7**	**94.4**	**31.2**

Experiment-to-configuration map. Table 10 specifies the representation and update permissions for each experiment. In continual evaluation, prediction uses the state available before the current audit. The last column identifies which state can change after that audit.

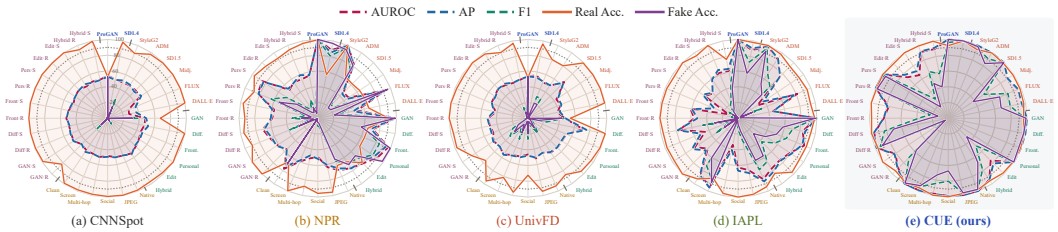


Figure 7: **Complete metric profiles for the 32-world comparison.** The original profiles preserve AUROC, AP, F1, real-image accuracy, and fake-image accuracy for all five methods and all 32 worlds. Figure 4 aligns the AUROC comparison by world; no methods or conditions are removed.

Table 10: Configuration map for the main experimental claims. "Native" means that a reproduced baseline retains its released representation and update interface; it does not mean that methods receive different queries or audit budgets.

Experiment	Frozen at prediction	Common evaluation unit	State allowed to change after audit
32-world coverage	Released detector; Community Forensics ViT-S/16 for the CUE core	512 fixed query identities per world	None; zero-shot evaluation
Six-world few-shot	Method-native representation and fixed query set	Nested K real and K fake audit identities	Method-native lightweight adaptation state
Twelve-world stream	DINOv2-L plus 78-D forensic representation for CUE; released representation for each baseline	Identical audits, queries, and two arrival orders	Continual head/state; no backbone update for CUE
SigLIP 2 probe	SigLIP 2 So400m/14@384	Same twelve-world stream	Linear head only
Evidence lifecycle	Frozen forensic scores and immutable trace contribution	Five fixed score trajectories and four controller histories	Confidence and contribution-use state
Mixed-window stress	Same deployed continual detector	16 labeled images per class per window, mixed at the tested ratios	Anonymous aggregate moments

Additional evaluation on fixed query sets. We additionally evaluate fixed query subsets of AIGC-DB and GenImage, containing 4,000 and 5,000 images, respectively, with equal numbers of real and fake images. PatchCraft and AIDE are the strongest reproduced comparators on these respective subsets. The query images differ from those in the six-benchmark comparison in Table 1. On these subsets, CUE reaches 98.57% and 99.36% AUROC, exceeding the strongest reproduced comparators by 14.11 and 20.39 percentage points, respectively.

Exact performance across 32 forensic worlds. Tables 11 and 12 give the exact AUROC values behind Figure 4. Within each row, all five methods use the same 256 real and 256 fake identities. CUE has the highest AUROC in 29 of the 32 worlds, including all twelve joint source–processing shifts.

Table 11: Exact AUROC (%) for the 20 reference and single-axis worlds in Figure 4. Bold marks the best result in each identical-query row.

World	CNNSpot	NPR	UnivFD	IAPL	CUE
Reference: ProGAN, clean	53.78	**100.00**	53.31	**100.00**	99.97
Reference: SD1.4, clean	50.60	88.42	55.46	88.21	**100.00**
StyleGAN2	49.22	**99.85**	49.43	98.69	99.45
ADM	50.17	82.38	49.62	90.26	**99.64**
SD1.5	50.01	46.09	65.03	75.39	**99.99**
Midjourney	49.57	46.96	43.24	29.18	**90.30**
FLUX.1-schnell	28.78	93.25	44.84	81.36	**95.43**
DALL-E 3	27.71	50.68	38.24	45.24	**94.87**
Traditional GAN content	47.97	98.12	46.79	99.22	**99.80**
Conventional diffusion content	47.40	71.38	71.21	83.69	**99.80**
Frontier generation content	45.16	96.55	52.74	90.69	**99.66**
Personalized fine-tuning	47.69	90.51	52.02	79.08	**100.00**
Generative editing	50.54	**80.96**	52.80	75.02	75.98
Hybrid synthesis	53.67	75.98	52.39	76.74	**93.74**
Original export	51.23	71.38	47.23	83.69	**99.80**
JPEG and resize	50.27	63.32	52.05	67.08	**99.35**
Social/WebP	47.77	49.09	45.84	62.39	**98.58**
Multi-hop compression	49.99	51.87	49.53	34.30	**97.39**
Screenshot recapture	48.73	59.32	50.37	95.04	**98.71**
Denoise and super-resolution	48.75	77.25	55.03	84.99	**99.85**

Table 12: Exact AUROC (%) for the twelve joint source–processing worlds. These are the joint source–processing columns in Figure 4.

Source × processing	CNNSpot	NPR	UnivFD	IAPL	CUE
Traditional GAN × Social/WebP	46.58	55.68	50.93	54.40	**89.66**
Traditional GAN × screenshot	51.88	61.49	52.12	66.01	**80.82**
Conventional diffusion × Social/WebP	50.25	65.93	47.59	45.21	**98.26**
Conventional diffusion × screenshot	49.92	71.21	49.14	76.05	**95.64**
Frontier generation × Social/WebP	52.52	57.76	50.36	55.65	**81.69**
Frontier generation × screenshot	50.49	56.48	41.51	61.69	**76.91**
Personalized fine-tuning × Social/WebP	47.06	82.37	50.37	13.34	**99.54**
Personalized fine-tuning × screenshot	49.78	84.93	50.47	50.80	**99.85**
Generative editing × Social/WebP	51.35	50.67	53.49	44.62	**73.56**
Generative editing × screenshot	48.54	55.13	50.71	60.47	**65.78**
Hybrid synthesis × Social/WebP	50.43	53.48	48.52	62.32	**91.02**
Hybrid synthesis × screenshot	49.93	55.37	49.90	61.17	**89.45**

H.2 Few-Shot and Continual Adaptation

All six targets cross 0.90 AUROC by $K = 8$, and every target exceeds 0.96 by $K = 64$ (Table 13). These are selected points from Figure 5. Each budget K uses a nested support prefix, but the resulting curve need not be monotone: a small audit can temporarily move the decision rule in an unhelpful direction.

Table 13: CUE few-shot AUROC (%) on the fixed query identities. $K_{.90}$ is the smallest tested audit budget per class reaching 0.90 AUROC.

Target world	$K = 0$	$K = 8$	$K = 16$	$K = 64$	$K_{.90}$
SD3.5	89.34	98.89	99.02	99.50	1
SD1.5-LoRA	96.68	98.24	99.56	99.99	0
Generative editing	82.07	87.01	82.18	96.75	4
Social/WebP	84.62	94.84	96.32	98.60	8
Screenshot recapture	92.63	97.39	95.70	97.80	0
Jointly new Q_{11}	82.99	90.38	96.84	97.87	8
Macro mean	88.05	94.46	94.94	98.42	—
Worst world	82.07	87.01	82.18	96.75	—

Complete continual-stream results. Table 14 reports every condition and method from the twelve-condition streams. Table 2 retains selected conditions for compact presentation; its summaries use all conditions and are unchanged.

Table 14: **Complete twelve-condition results under two arrival orders.** AUROC (%) for generator-first (Gen.) and propagation-first (Prop.) streams. Pooled AUROC uses all query images jointly; Max. drop is the largest per-world post-audit-to-final decline (percentage points). Gray marks CUE; bold and underlining mark the best and second-best results within each order.

Metric / condition	Frozen		Fine-tune		SimpleShot		SLDA		RanPAC		LayUP		A-GEM		SigLIP 2		CUE	
AUROC (%) ↑	Gen.	Prop.	Gen.	Prop.	Gen.	Prop.	Gen.	Prop.	Gen.	Prop.	Gen.	Prop.	Gen.	Prop.	Gen.	Prop.	Gen.	Prop.
Initial condition	96.9	96.9	94.0	92.6	78.4	78.4	96.6	96.6	87.7	87.7	57.8	72.8	96.8	96.8	**99.1**	**99.2**	96.9	96.9
All conditions (pooled)	76.7	76.7	82.6	82.2	66.3	66.3	82.9	82.9	75.8	75.8	65.1	81.0	78.5	78.5	93.2	93.0	**94.3**	**94.3**
Max. drop (pp) ↓	**0.0**	**0.0**	5.9	4.6	0.7	0.7	0.2	2.4	0.3	1.0	25.2	11.7	1.2	0.7	2.0	2.4	**0.0**	**0.0**
Generator changes																		
SD3.5	73.4	73.4	87.6	88.8	95.0	95.0	85.7	85.7	90.4	90.4	72.3	93.6	80.7	79.7	93.1	92.4	**96.6**	**96.6**
FLUX.1	41.7	41.7	49.4	48.8	44.4	44.4	47.8	47.8	49.1	49.1	66.2	67.5	43.4	43.3	**86.8**	**87.0**	83.0	83.0
Ideogram	77.7	77.7	84.6	84.9	43.1	43.1	76.3	76.3	58.6	58.6	55.8	70.6	80.4	80.3	91.0	90.3	**95.9**	**95.9**
Seedream	74.1	74.1	79.2	81.9	39.7	39.7	74.3	74.3	61.3	61.3	58.3	71.1	76.2	76.3	86.1	86.4	**92.1**	**92.1**
SD1.5-LoRA	87.3	87.3	91.8	92.5	65.4	65.4	91.8	91.8	79.2	79.2	69.5	86.1	88.2	88.4	**99.3**	**99.5**	**99.3**	99.3
SDXL-DB	90.7	90.7	92.1	98.1	92.9	92.9	97.6	97.6	95.7	95.7	74.1	96.8	90.2	90.5	99.6	99.8	**100.0**	**100.0**
Propagation changes																		
JPEG	93.9	93.9	93.9	91.3	83.3	83.3	93.6	93.6	85.3	85.3	63.0	83.3	93.8	93.9	**97.5**	**97.8**	96.6	96.6
WebP	91.8	91.8	94.1	92.2	84.4	84.4	92.4	92.4	85.7	85.7	63.3	81.5	92.3	92.4	**97.1**	**97.1**	95.8	95.8
Resize	92.7	92.7	94.1	92.5	85.6	85.6	94.6	94.6	87.3	87.3	67.4	83.4	92.7	92.8	**98.8**	**98.8**	95.3	95.3
Screenshot	93.5	93.5	91.7	89.8	86.9	86.9	94.1	94.1	85.7	85.7	62.0	83.9	93.9	94.0	**96.8**	**96.9**	95.8	95.8
Multi-platform	73.7	73.7	85.9	83.8	47.8	47.8	78.2	78.2	64.2	64.2	68.5	81.6	76.0	75.9	91.8	91.2	**92.6**	**92.6**
Re-digitization	67.0	67.0	81.4	78.9	45.1	45.1	76.9	76.9	67.9	67.9	61.2	74.9	69.9	69.9	86.1	85.1	**90.3**	**90.3**

Table 15 compares CUE with a modern frozen encoder and an updated linear head. SigLIP 2 closes most of the performance gap and retains the original world slightly better. CUE achieves higher pooled all-world AUROC in both arrival orders, with paired confidence intervals above zero. Because the encoders differ, these results compare the complete systems rather than isolate their update rules.

Table 15: Paired comparison with a frozen SigLIP 2 So400m representation and updated linear head. Confidence intervals use 10,000 identity-paired bootstrap resamples.

Arrival order	SigLIP 2	CUE	Δ (95% CI)
Generator first	93.18	94.28	$+1.11$ $(+0.33, +1.87)$
Propagation first	93.05	94.28	$+1.23$ $(+0.47, +2.00)$

H.3 Evidence Lifecycle and Ablations

Figure 6 displays three independent fixed-contribution cycles: SD3.5–JPEG–SD3.5, editing–screenshot–editing, and Q_{11}–multi-hop–Q_{11}. Each phase uses 256 real and 256 fake query images,

with paired identities and batch order across policies. On return, each policy reuses the original evidence identity and scoring weight where its retained state allows restoration. Balanced error uses a fixed probability threshold of 0.5. The left panels use all 512 screenshot queries. Their 26-bin density histograms and reflection kernel estimates (bandwidth 0.045) compare scores before and after withdrawal on the same images. The density axes are labeled separately. In the right panel, faint lines show 32 query batches of eight real and eight fake images per phase; prominent lines show non-overlapping means of four consecutive batches. Shaded intervals mark changed conditions, and the three independent cycles are drawn separately. Use-state labels refer to CUE. Phase annotations use all 32 batches: pause annotations give the mean error difference between CUE and keeping the contribution active; return annotations compare the mean errors of permanent deletion and CUE.

Table 16 reports an additional evaluation on five identity-disjoint query worlds built from two immutable evidence contributions. In every world, the retained contribution follows the intended ACTIVE→DORMANT →DORMANT→ACTIVE state sequence. The controller uses audited decision risk and a false-accusation constraint. The AUROC differences separately describe changes in ranking, including a decline after restoration in one world.

Table 16: Per-world evidence lifecycle. Δ_{off} is CUE minus always-on AUROC in the invalidated phase; Δ_{return} is CUE minus permanent deletion after the trace returns. All values are percentage points.

Query world	Controlled contribution	Δ_{off}	Δ_{return}	State sequence
Firefly Image 3	Blur contribution	$+0.20$	$+0.39$	A–D–D–A
Ideogram V1	Blur contribution	$+0.59$	$+0.78$	A–D–D–A
LaVi-Bridge	Blur contribution	$+0.81$	$+0.56$	A–D–D–A
EPAIQA-15K Flickr-farm	JPEG subsampling contribution	$+5.92$	-1.52	A–D–D–A
EPAIQA-15K numeric	JPEG subsampling contribution	$+2.52$	$+2.91$	A–D–D–A

On the three blur-based trajectories, the controller issues RETIRE and REACTIVATE actions. The two EPAIQA-15K partitions evaluate the same use-state history of a second frozen contribution. The five query-world cycles therefore involve two learned contribution identities: one shared by the first three trajectories and another shared by the last two.

What the ablations isolate. The left panel of Table 3 compares complete update strategies and two matched component ablations. Window balancing and head regularization are each removed while the other selected settings are held fixed. Current-audit-only and accumulate-all strategies change several components together and compare the resulting balance between adaptation and retention. The right panel replays control operations on the archived trajectories. The no-reset intervention affects the epochal blur controller; the EPAIQA-15K Schmitt history has no corresponding episode reset and is unchanged.

Operational effect of the conservative score. Table 17 reports false accusations together with sensitivity. For each group/phase, we select the Always-on threshold attaining at least 95% sensitivity and apply it to every policy. Rates are macro means over ten invalid/dormant group–phase cells. The bound-violation rate is the fraction of scores exceeding $\min(s_{\mathrm{on}}, s_{\mathrm{off}})$ in those phases. A zero violation rate checks the score construction. The false-accusation and sensitivity columns describe the corresponding detection behavior.

Table 17: Shared-threshold accusation tradeoff (%). These are descriptive rates for the selected invalid/dormant phases.

Policy	False accusations	Sensitivity	Bound violations
Always-on	51.44	95.33	52.57
No episode reset	48.37	93.98	0.00
Irreversible withdrawal	49.96	94.73	47.43
No conservative bound	49.96	94.73	47.43
CUE	48.37	93.98	0.00

H.4 Audit Budget, Dormancy, and Recovery

A retained contribution can become useful again after a period of dormancy. We examine how the audit budget affects the decision to reuse it through an exploratory replay of the recorded Ideogram and LaVi-Bridge streams. The two streams share one fixed blur contribution. Before subsampling, we reproduce the full-budget gains and controller states for all ten audits from the recorded inputs.

Paired audit-budget protocol. For each audit, we draw $K \in \{8, 16, 32\}$ identities per class without replacement. Two hundred fixed random orderings define nested subsets across K. All policies receive the same subsets. The available return audit limits the common budget to 32 per class. Each controller processes the original five audits in order. The stage names are evaluator metadata, never policy inputs. The full controller retains its original thresholds and discount. The sign gate uses the contribution when the current audit gain is positive and the same FPR guard passes. Signed reweighting minimizes current-audit binary cross-entropy for $s_{\text{off}} + w(s_{\text{on}} - s_{\text{off}})$ over $w \in [-1, 1]$; a coefficient is replaced by zero if it fails the same FPR guard. Thresholds and the coefficient range are kept fixed.

Control decisions are fixed before query labels are used for evaluation. Query loss is measured on three subsequent LaVi-Bridge query sets that share no base-image identities with any audit set. The 200 draws measure sensitivity to audit selection within these fixed streams, not variation across 200 independently collected streams. This exploratory analysis was specified after the original full-budget results were known.

Separating uncertain evidence from returning evidence. Table 18 reports use in the weak-evidence stage N and the return stage. In LaVi-Bridge, increasing K from 8 to 32 reduces positive use at N from 40.0% to 19.5% for CUE, while increasing return-stage use from 83.0% to 98.5%. A larger audit therefore reduces use in the weak-evidence stage while allowing more frequent reuse after return. The sign gate and signed coefficient yield different operating points; the table reports both streams at every budget. Stage N follows the original protocol and does not specify the sign of population utility.

Table 18: Contribution use across 200 paired audit subsets (%). $N+$ denotes a positive coefficient at the weak-evidence stage; $N{\neq}0$ includes negative coefficients. Return+ denotes a positive coefficient after return.

Stream	K	Policy	$N+$	$N{\neq}0$	Return+
Ideogram	8	CUE	46.0	46.0	67.5
Ideogram	8	Sign gate	40.5	40.5	77.5
Ideogram	8	Signed weight	42.0	100.0	78.0
Ideogram	8	Qualified signed	13.5	41.0	53.0
Ideogram	16	CUE	27.0	27.0	58.0
Ideogram	16	Sign gate	34.5	34.5	86.0
Ideogram	16	Signed weight	37.5	100.0	86.5
Ideogram	16	Qualified signed	7.0	30.0	46.0
Ideogram	32	CUE	12.5	12.5	84.5
Ideogram	32	Sign gate	34.5	34.5	100.0
Ideogram	32	Signed weight	43.5	100.0	100.0
Ideogram	32	Qualified signed	9.5	26.5	100.0
LaVi-Bridge	8	CUE	40.0	40.0	83.0
LaVi-Bridge	8	Sign gate	47.5	47.5	87.5
LaVi-Bridge	8	Signed weight	50.0	99.5	89.5
LaVi-Bridge	8	Qualified signed	20.5	37.5	73.0
LaVi-Bridge	16	CUE	36.5	36.5	83.5
LaVi-Bridge	16	Sign gate	50.0	50.0	97.5
LaVi-Bridge	16	Signed weight	55.0	99.0	99.5
LaVi-Bridge	16	Qualified signed	25.0	36.0	82.5
LaVi-Bridge	32	CUE	19.5	19.5	98.5
LaVi-Bridge	32	Sign gate	58.0	58.0	99.5
LaVi-Bridge	32	Signed weight	63.5	96.5	99.5
LaVi-Bridge	32	Qualified signed	18.5	23.0	97.5

State control and predictive risk. Table 19 reports held-out query loss, including the reversal phase where a negative coefficient can exploit the opposite orientation. Lower contribution use alone does not imply lower query loss. The policies trade off predictive risk across phases. This distinction between use frequency and prediction loss is relevant when evaluating them in a joint stream.

Table 19: Mean held-out LaVi-Bridge binary cross-entropy across the same 200 audit subsets; lower is better. No query identities enter the audits.

K	Policy	Reversal	Weak evidence	Return
8	CUE	0.79586	0.47923	0.28089
8	Sign gate	0.78278	0.47889	0.27990
8	Signed weight	0.74801	0.48223	0.28222
8	Qualified signed	0.75298	0.48148	0.28332
16	CUE	0.78972	0.47940	0.28078
16	Sign gate	0.78199	0.47877	0.27771
16	Signed weight	0.74642	0.48150	0.27788
16	Qualified signed	0.74704	0.48080	0.28100
32	CUE	0.78258	0.48018	0.27749
32	Sign gate	0.78179	0.47840	0.27727
32	Signed weight	0.74830	0.48023	0.27729
32	Qualified signed	0.74830	0.48059	0.27771

An exploratory extension: orientation-aware use. The phase-specific losses suggest distinguishing a weak contribution from a reversed one. We evaluate one additional candidate, *qualified signed use*, on the same paired audit subsets, guided by these phase-specific observations. It retains the fixed trace score and chooses $a \in \{-1, 0, 1\}$. For each nonzero orientation, it estimates current-audit BCE reduction relative to $a = 0$ and its Gaussian confidence under a unit-normal prior. An orientation can be used only above 0.8 confidence and under the same empirical FPR guard. If neither orientation qualifies, the contribution is zero; if both qualify, the larger estimated gain determines the choice. This prototype uses only the current audit, with its threshold fixed.

At $K = 16$, its reversal-stage query loss is 0.74704 versus 0.78972 for CUE. The corresponding weak-evidence losses are 0.48080 and 0.47940, and return losses are 0.28100 and 0.28078. Signed reweighting remains a strong comparator in Table 19. These operating points distinguish orientation from whether a contribution should be used: a reversed contribution may remain informative, whereas an uncertain one may warrant continued dormancy. This candidate was introduced after the first comparison and remains an exploratory extension requiring independent confirmation. The main results use the original controller.

Full-budget sign-gate diagnostic. A separate retrospective replay uses all recorded audit statistics on the three blur trajectories. The sign gate matches CUE in 14 of 15 use states and 13 of 15 control actions. Scores are identical in 11 of 12 query evaluations. Its sole state difference is earlier reuse at LaVi-Bridge's N stage, where query AUROC is 90.53% versus 90.41% for CUE. The EPAIQA-15K partitions are excluded because the required risk-guard inputs are absent. This diagnostic locates the difference between the policies at a weak-evidence stage and motivates a closer study of audit uncertainty and phase-specific risk.

I Boundary Studies and Scaling

The main experiments examine how the learner acquires, retains, and controls evidence as forensic worlds change. This section asks when these capabilities fail to transfer. We study three cases: changes within a model lineage, compositions of individually tested transformations, and audit windows that mix observations from different worlds.

I.1 When Forensic Transfer Breaks

Generator genealogy is not forensic familiarity. Table 20 pairs parent and adapter-tuned child images by prompt and seed. In the controlled SDXL pair, changing only the adapter lowers IAPL by

31.63 points, raises CUE by 9.18, and leaves NPR nearly unchanged. These opposing changes show that the child is not uniformly harder; the adapter changes which detector evidence remains useful. The SD1.5/LCM pair also produces method-dependent changes, although its looser visual pairing supports only a lineage-shift interpretation.

Table 20: Zero-shot parent-to-child AUROC (%). Differences and 95% intervals use 10,000 prompt/seed-paired bootstrap resamples.

Lineage	Method	Parent	Child	Child − parent (95% CI)
SD1.5 → LCM-LoRA	NPR	99.98	90.51	-9.48 $(-12.24, -6.98)$
	IAPL	96.93	79.08	-17.85 $(-21.55, -14.42)$
	CUE	99.25	86.25	-13.00 $(-16.31, -9.88)$
SDXL → DreamBooth-LoRA	NPR	98.86	98.44	-0.42 $(-1.18, +0.23)$
	IAPL	91.88	60.26	-31.63 $(-36.70, -26.64)$
	CUE	82.93	92.11	$+9.18$ $(+6.92, +11.73)$

Robust transformations do not automatically form a robust chain. The continual stream separates four individual transformations from two compound propagation pipelines. The SigLIP 2 linear baseline reaches 96.8–98.8 AUROC on the individual transformations but 85.1–91.8 after multi-platform transmission and re-digitization. CUE reaches 95.3–96.6 and 90.3–92.6, respectively. Successive operations in the compound worlds can erase one trace while creating another. Testing JPEG or resizing separately does not capture these interactions, so robustness to individual operations provides no composition guarantee.

An identity-free window must still represent its query world. An audit window must represent the query distribution even when its world identity is unknown. Table 21 mixes two worlds inside one 16-per-class window and compares the update with a component-matched oracle. JPEG/WebP shows a small transient dilution. The two compound environments do not cross the simultaneous-confidence boundary. The largest failure occurs when a FLUX.1-only audit is followed by SD3.5 queries. A shared update interface does not make observations from one generator representative of another.

Table 21: Mixed-window interference penalty in AUROC points relative to the component-matched oracle. "First boundary" is the first tested audit composition whose simultaneous 95% upper band is below zero; none means the point loss does not satisfy that criterion.

World pair	Matched, post	Async., post	Async., end	First boundary	Direction relation
JPEG / WebP	-3.76	-3.55	0.00	4/12	Aligned; transient dilution
Multi-platform / re-digitized	-2.48	-3.15	-0.75	none	Evidence dilution
SD3.5 / FLUX.1	-3.88	-30.88	-25.71	0/16	Near-orthogonal evidence

I.2 Robustness, Scaling, and Scope

What grows with the stream. Table 22 distinguishes the fixed representation, moments stored by window, adaptive classifier, and controlled contributions. The complete continual core stores its projection, moment ledger, forensic coordinates, and classifier once. We report evidence-control state as an increment beyond this core, so shared state is not counted twice.

Table 22: Implementation dimensions and scaling. State excludes the shared frozen backbone and raw images. Evidence-control values are incremental beyond the complete 5.217-MiB continual core.

Component	Implementation and reported setting	Measured state or time	Scaling
Frozen representation	1024-D DINOv2-L embedding plus a 78-D forensic vector; $D = 1102$ and the 304M-parameter encoder remains fixed	Shared across all worlds	Independent of stream length
Aggregate projection	Fixed $1102 \to 256$ projection applied before moment updates	Included in the continual core	Fixed after initialization
Moment ledger	$(n, \sum z, \sum zz^\top)$ for each label–window cell; $W = 3$ windows and $2W = 6$ cells, with effective cap $n_{\max} = 256$ per cell	Included in the 5.217-MiB core	$\mathcal{O}(2Wd^2)$ for binary labels
Coordinates and classifier	Adaptive generalized eigenspace and residual head; $r \leq 64$, with rank 3 in the reported Q_{11} streams	Included in the continual core	$\mathcal{O}(rd)$ beyond the ledger
Controlled contributions	Fixed trace contributions with confidence states and an audit ledger; $J = 3$ in the resource reference and $J \in \{0, 1, 3, 8, 16, 32, 64\}$ in scaling tests	$+0.144$ MiB when active, $+0.168$ MiB after restoration; $+3.545$ MiB at $J = 64$	Approximately linear in J
Update and serving	Evidence operations after image encoding; two core updates and a controller stress test at $J = 64$	38.4 ms update p95; $+2.717\,\mu$s/image and 5.829 ms/control audit at $J = 64$	Measured operating points

Table 23 gives all measured operating points, including the intermediate contribution counts omitted from Table 4.

Table 23: **Complete contribution-control measurements.** J: retained contributions. Extra: serialized state beyond the 5.217-MiB core; Log: separate audit log. Prediction/update times are per image/control audit, excluding encoding and shared-core computation.

Setting	J	**Extra/MiB**	**Log/KiB**	**Predict/μs**	**Update/ms**
No control	0	< 0.001	0.00	0.000	–
Active	1	0.058	6.12	0.270	0.110
Active	3	0.144	5.97	0.353	0.386
Dormant	3	0.152	9.01	0.397	0.473
Reactivated	3	0.168	15.01	0.356	0.616
Active	8	0.445	37.28	0.544	0.736
Active	16	0.888	72.94	0.855	1.444
Active	32	1.772	144.04	1.484	2.878
Stress test	64	3.545	287.29	2.717	5.829

Toward joint evidence maintenance. Several extensions follow from these results. An audit window provides a local snapshot without a world identity label; arbitrary mixtures of unrelated worlds may require detecting latent mixture boundaries before updating. Proposition 7 applies to the projected linear core. Extending exact control to nonlinear adaptive heads requires a stronger separation mechanism. Aggregate moments also avoid historical-image replay, but do not by themselves provide a formal privacy guarantee.

The controlled trace identities are defined before deployment. Discovering new identities and deciding how to partition and control them remain open problems. A joint stream would test how updates to the adaptive core change the value of other controlled contributions. Persistent reversal may instead call for signed reweighting, which the budget study compares with temporary withdrawal. Because the Gaussian confidence is a working model, evaluating sparse or delayed audits must consider query risk and recovery together.

J REPRODUCIBILITY AND ARTIFACT INTEGRITY

Each reported metric can be traced to its query images, permitted inputs, and model state through retained records. Predictions are fixed before reference labels are used for evaluation.

Standardized result records. The retained records support independent recomputation and checks of data use:

Table 24: Records retained for reproducing the reported results.

Artifact	Required contents
`manifest.json`	Identities, roles, data hashes, checkpoint hash, configuration hash, and permitted model inputs
`scores.npz`	One score record per base identity, before metric aggregation
`summary.csv`	Verified metric summaries for the figures and tables
`provenance.json`	Reproduced or paper-reported source, protocol, and metric definition
`ledger.jsonl`	Time-ordered confidence, risk, action, state, evidence ID, and event hashes
`verification.json`	Independent recomputation, overlap checks, replay status, and file digests

Each displayed value is associated with a run identifier, query manifest, checkpoint hash, metric definition, and source type. A value is marked *reproduced* only when it comes from the common local protocol, *reported* only when the original paper provides a compatible value, and *N/A* when neither is available. Values from incompatible protocols are not averaged.

Identity and leakage checks. Canonical pixel hashes and dataset-native identity keys are both retained. Support and query identities are disjoint, as are historical data and held-out evaluation identities. Distinct audit batches within a continual run use disjoint identities. Few-shot budgets use nested subsets of one support pool; these budget comparisons are separate evaluations, not successive audit times. All files derived from one base image share a group identity. Image decoding errors and duplicate assignments are resolved before feature encoding or score inspection.

Fixed predictions and independent evaluation. Query scores are computed without query labels. We save the scores, configuration, model state, resource measurements, and data-access record together with their hashes. An independent evaluator verifies the hashes before using query labels to recompute metrics from per-image scores. The recomputed summaries provide the reference values for checking the figures and tables.

State preservation and update checks. The implementation checks exact restoration of saved states, unique contribution identities, increasing audit indices, and detection of changes to linked audit records. For multiple contributions, the tests verify that toggling one leaves the others unchanged and that all proposed state changes take effect together. They also verify that a failed update leaves the prior controller state byte-for-byte unchanged and that query prediction does not change confidence. The controller receives only an audit index, fixed on/off or component scores, and audit labels. Source, generator, environment, data-split, and stage labels are excluded.

Deployment measurements. Resource tests vary the evidence-bank size over $J \in \{0, 1, 3, 8, 16, 32, 64\}$. They record serialized state size, ledger growth, batch-1 and batch-8 prediction latency, audit-update latency, and exact equality after state restoration and replay. Image encoding and evidence-bank operations are timed separately to measure the additional cost of contribution control. The audit-budget study uses $K \in \{8, 16, 32\}$ labeled images per class. For each of 200 fixed random orderings, the subsets are nested across budgets, while the query streams stay fixed. Appendix H.4 gives the sampling protocol.

Implementation dimensions. Appendix I.2 reports the feature dimensions, retained state, and measured growth with audit windows and controlled contributions. The 5.217-MiB core and 38.4-ms p95 are measured with three retained audit windows and six label–window cells. The timing covers two consecutive core updates after feature extraction. This resource instance is separate from the

twelve-world accuracy evaluation; storing more windows increases moment storage as specified in Appendix I.2. The table below gives reference settings for the continual core and epochal controller. Static coverage and the selected open-world ablation use the experiment-specific configurations described below.

Fixed hyperparameters. Table 25 lists the reference settings. The twelve-world stream uses $n_{\max} = 256$ and $\lambda_h = 10$. The full method in the left ablation panel uses $n_{\max} = 224$ and $\lambda_h = 1000$, selected on development data and fixed before evaluation. The matched component ablations remove the cap or set $\lambda_h = 0$, respectively. Static coverage leaves both geometry and the controller unchanged. The EPAIQA-15K switching history uses a separate configuration from the epochal Gaussian controller.

Table 25: Reference hyperparameters for the continual core and epochal controller; the selected open-world ablation overrides are stated above. The residual scale is estimated once from the initial-world validation split; it is not fitted to query labels.

Block	Quantity	Value	Selection or role
Adaptive core	Projected dimension d	256	Fixed before the stream
Adaptive core	Effective samples per label–window cell $n_{\max}$	256	Caps a window's influence, not its stored count
Adaptive core	Within-class variation weight η	2.0	Development-selected, then frozen
Adaptive core	Diagonal regularizer ε	10^{-3}	Stabilizes the generalized eigensystem
Adaptive core	Relative eigenvalue cutoff; maximum rank $r_{\max}$	10^{-6}; 64	Adaptive retained rank
Adaptive core	Residual-head regularization λ_h	10.0	Development-selected, then frozen
Adaptive core	Residual scale κ	$\mathrm{Cov}(s_0, f_0)/\mathrm{Var}(f_0)$	Computed once on Q_{00} validation scores
Evidence control	Initial confidence prior	$\mathcal{N}(0, 1)$	Shared across contributions
Evidence control	Previous-precision discount δ	0.25	Discounts accumulated confidence precision
Evidence control	Dormancy and activation thresholds τ_-, τ_+	0.2, 0.8	Hysteretic state transition
Evidence control	Audit FPR allowance	$1/n_{\mathrm{real}}$	At most one additional false positive on audited real images
Evidence control	Audit size in the budget study	$K \in \{8, 16, 32\}$ per class	Nested subsets of archived audits
Evidence control	Resource reference and stress-test bank sizes J	3; $\{0, 1, 3, 8, 16, 32, 64\}$	Resource study; separate from the four recorded control trajectories

Ablation switches. Table 26 gives the component settings and all results for the ablations summarized in Table 3. The adaptation switches are updates (Upd), historical memory (Mem), learned geometry (Geo), window balancing (Bal), and head regularization (Reg). The control switches are individual contribution control (Sep), withdrawal (Wdr), episode reset (Rst), reactivation (Rea), and the conservative bound (Bnd).

Table 26: **Complete ablation results and component settings.** Left: mean/worst Q_{11} AUROC (%) and changes in Q_{00} AUROC and historical-threshold FPR (percentage points). Right: action accuracy, withdrawal/restoration delay (R/A, audits), phase-weighted exposure (Exp.), bound violations, FPR, and AUROC. Except R/A and Exp., right-panel values are percentages. Gray marks full CUE; bold/underlined values mark best/second-best results. Appendix J defines the switches.

Adaptation	Upd	Mem	Geo	Bal	Reg	Macro↑	Worst↑	ΔQ_{00} ↑	ΔFPR↓
No adaptation	×	×	×	×	×	89.05	77.17	0.00	0.00
Accumulate all audits	✓	✓	×	×	×	87.77	78.75	−0.04	+0.15
No window balancing	✓	✓	✓	×	✓	91.18	80.79	**+0.47**	**−1.48**
No head regularization	✓	✓	✓	✓	×	91.46	80.05	−0.97	+1.23
Use current audit only	✓	×	×	×	×	**93.12**	**83.08**	−1.35	+2.49
CUE (ours)	✓	✓	✓	✓	✓	92.94	83.02	−0.51	+0.13

Evidence control	Sep	Wdr	Rst	Rea	Bnd	Act.↑	R/A↓	Exp.↓	Viol.↓	FPR↓	AUC↑
Core detector only	×	×	×	×	×	–	–	**0.000**	–	49.96	82.14
Always-on evidence	×	×	×	×	×	65.22	–	0.336	52.57	51.44	81.52
No reset after withdrawal	✓	✓	×	✓	✓	95.65	0/–	**0.000**	**0.00**	**48.37**	82.26
No reactivation	✓	✓	✓	×	✓	82.61	0/–	**0.000**	47.43	49.96	82.14
No conservative bound	✓	✓	✓	✓	×	**100.0**	**0/0**	**0.000**	47.43	49.96	**82.30**
CUE (ours)	✓	✓	✓	✓	✓	**100.0**	**0/0**	**0.000**	**0.00**	**48.37**	**82.30**

K Proof of the historical learning separation

The proof compares two learners through the historical information they retain. First, we identify the information omitted by exact historical posterior functions and bound the performance of every learner restricted to that memory. We then learn coordinates from finite historical observations and bound the population AUROC of a ridge classifier fitted with two current labels. The current support and evaluation examples are independent of the historical observations.

Readout and numerical guarantee. At $d = 1024$, the construction uses $16{,}384$ historical observations for the direction and $4{,}096$ independent historical observations for normalization. The ridge readout uses the augmented coordinates and an intercept to fit signed targets. The penalty 0.2 applies to all coefficients, including the intercept. Its expected population AUROC is greater than 0.76789 with two current labels, one per class. The direction denoted e below corresponds to $\widehat{v}_T$ in the main text. The expectations average over the two historical samples and the current support, all of which are independent of the evaluation examples. Corollary 2 gives the complete numerical bound and its comparison with 202 balanced current labels.

K.1 What exact historical decisions retain

We first construct historical distributions whose exact posterior functions do not identify the direction needed to distinguish the current classes. Class labels in this proof use the signed convention introduced in the main analysis. Fix a unit vector $w \in \mathbb{R}^d$, write $\mathcal{H} = w^\perp$ and $q = d - 1$, and let the unknown unit vector u range over $\mathbb{S}(\mathcal{H})$. For historical distribution j, the class-conditional law is $X \mid (j, Y = y) \sim \mathcal{N}(y\theta_j, \Sigma_j)$, where

$$\theta_j = jw + u, \qquad \Sigma_j = I_d + uu^\top + j(wu^\top + uw^\top), \qquad j \in \{-1, +1\}. \tag{58}$$

The class priors are balanced in both historical distributions. Since $\Sigma_j(jw) = \theta_j$, the exact posterior functions are

$$m_j(x) = \Pr(Y = +1 \mid x, j) = \frac{1}{1 + \exp(-2jw^\top x)}. \tag{59}$$

Thus the rule memory $M_R = (m_{-1}, m_{+1})$ is identical for every u. The covariance is positive definite: on $\mathrm{span}(w, u)$ its matrix is $\left(\begin{smallmatrix} 1 & j \\ j & 2 \end{smallmatrix}\right)$, with eigenvalues $(3 \pm \sqrt{5})/2$, and all other eigenvalues equal one. The historical covariance matrices therefore have bounded condition numbers throughout the construction.

Under the current distribution,

$$X_i = Y_i s u + \varepsilon_i, \qquad \varepsilon_i \text{ independent } \mathcal{N}(0, I_d), \tag{60}$$

where $s > 0$ is known and the n fixed labels contain $n/2$ of each sign. Here n counts current labels in total, so $n = 2K$ in the per-class budget notation. Denote this current labeled sample by D_n. The learner observes all raw coordinates of these labeled examples. We also grant it w, the distribution family, s and n. Its historical information consists of M_R. It receives no historical observations or historical marginal statistics, and no additional current unlabeled pool. The training procedure may be any measurable function and may use independent algorithmic randomization.

For a fitted score f, define its population AUROC by

$$\mathrm{AUC}_u(f) = \Pr\{f(X^+) > f(X^-)\} + \tfrac{1}{2}\Pr\{f(X^+) = f(X^-)\}, \tag{61}$$

where the independent positive and negative test observations are drawn after f is fitted. We bound the best performance achievable under this protocol, including by nonlinear scores. The learner may fit a new detector rather than simply apply an old one.

Lemma 3 (Exact rule-memory minimax AUROC)**.** *Let $q \geq 2$, $b = s\sqrt{n}$, and let $Z \sim \mathcal{N}(0,1)$ and $Q \sim \chi^2_{q-1}$ be independent. Define*

$$C_{q,n} = \frac{b+Z}{\sqrt{(b+Z)^2+Q}}, \qquad F_q(n,s) = \mathbb{E}\Phi(\sqrt{2}sC_{q,n}), \tag{62}$$

where Φ is the standard normal CDF. Then

$$\sup_{\mathcal{A}} \inf_{u\in\mathbb{S}(\mathcal{H})} \mathbb{E}\,\mathrm{AUC}_u\big(\mathcal{A}(D_n, M_R)\big) = F_q(n,s). \tag{63}$$

The same value is the optimal average AUROC under the uniform prior on $\mathbb{S}(\mathcal{H})$. It is attained by the score

$$f_T(x) = T^\top P_{\mathcal{H}} x, \qquad T = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i P_{\mathcal{H}} X_i = bu + G_0, \quad G_0 \sim \mathcal{N}(0, I_{\mathcal{H}}). \tag{64}$$

Proof. We identify a sufficient statistic for the training data, find the best way to rank a positive–negative test pair given that statistic, and then show that its average performance under a uniform prior equals the minimax value.

Put $Z_i = Y_i P_{\mathcal{H}} X_i$. These observations are independent $\mathcal{N}(su, I_{\mathcal{H}})$. Expanding their joint density shows that its dependence on u is precisely the factor $\exp(su^\top \sum_i Z_i) = \exp(bu^\top T)$. The residuals around their sample mean and the w-coordinates are independent of T and have distributions independent of u. The memory in equation 59 likewise does not depend on u. Consequently, these additional observations can provide randomization, but they add no information about the unknown direction beyond T.

Let U have the uniform distribution on $\mathbb{S}(\mathcal{H})$ and set

$$J_q(v) = \mathbb{E}\exp(v^\top U) = j_q(\|v\|_2). \tag{65}$$

Rotational invariance gives the radial form. For $r > 0$, symmetry yields

$$j_q'(r) = \mathbb{E}[U_1 \exp(rU_1)] = \mathbb{E}[U_1 \sinh(rU_1)] > 0. \tag{66}$$

The posterior density of U given the full training sample is proportional to $\exp(bT^\top U)$.

Consider an independent test pair with the positive observation assigned to either position with equal probability. For realized points (x_1, x_2), write $d_x = P_{\mathcal{H}}(x_1 - x_2)$. Under the two possible orders, the test density factors depending on u are $\exp(su^\top d_x)$ and $\exp(-su^\top d_x)$, respectively. All remaining factors coincide. After integrating over the posterior, the likelihood ratio for these two order hypotheses is therefore

$$\frac{J_q(bT + sd_x)}{J_q(bT - sd_x)}. \tag{67}$$

By equation 66 and

$$\|bT + sd_x\|_2^2 - \|bT - sd_x\|_2^2 = 4bs\,T^\top d_x, \tag{68}$$

the Bayes-optimal order decision has the sign of $T^\top d_x$.

Every fitted scoring function orders the pair by comparing its two scores, with fair randomization for ties. Its probability of choosing the correct order is exactly its expected AUROC. Allowing arbitrary decisions about the pair order can only enlarge the class of procedures, so the Bayes optimum upper-bounds the performance of every scoring procedure. The score f_T realizes that optimum. This proves prior-average optimality without any linearity assumption on competing scores. Independent algorithmic randomization cannot exceed this Bayes optimum.

Conditional on nonzero T, the difference of the two scores $f_T(X^+) - f_T(X^-)$ is Gaussian with mean $2sT^\top u$ and variance $2\|T\|_2^2$. Thus its conditional AUROC is $\Phi(\sqrt{2}sT^\top u/\|T\|_2)$. Rotating coordinates so that u is the first axis gives equation 62. This expected value is identical for every u. For any algorithm, its infimum over u is no larger than its uniform-prior average, which is at most $F_q(n,s)$. Conversely, f_T attains $F_q(n,s)$ at every u. These two inequalities establish equation 63. □

Corollary 1 (Necessary current label budget)**.** *If $q > 3$, then*

$$F_q(n,s) \le \min\left\{\Phi(\sqrt{2}s),\ \frac{1}{2} + s^2\sqrt{\frac{n}{\pi(q-3)}}\right\}. \tag{69}$$

A rule-memory procedure guaranteeing expected AUROC at least $1/2 + \Delta$ for every u must therefore use

$$n \ge \frac{\pi(q-3)\Delta^2}{s^4}. \tag{70}$$

In particular, when $d = 1024$, $s = 1$, and the target is 0.75, at least 202 balanced current labels are necessary.

Proof. For $k > 0$, $a \in \mathbb{R}$ and $v > 0$, define

$$h_k(a,v) = \Phi\left(\frac{ka}{\sqrt{a^2+v}}\right) - \frac{1}{2}. \tag{71}$$

This is odd in a, so $\mathbb{E}h_k(Z,Q) = 0$. Writing ϕ for the standard normal density, differentiation gives

$$0 \le \partial_a h_k(a,v) = \phi\left(\frac{ka}{\sqrt{a^2+v}}\right)\frac{kv}{(a^2+v)^{3/2}} \le \frac{k}{\sqrt{2\pi v}}. \tag{72}$$

Integration over the shift $b = s\sqrt{n}$ and then expectation imply

$$0 \le \mathbb{E}h_k(Z+b,Q) \le \frac{kb}{\sqrt{2\pi}}\mathbb{E}Q^{-1/2}. \tag{73}$$

For $Q \sim \chi^2_{q-1}$, direct integration of its density and the identity $\Gamma(a+1) = a\Gamma(a)$ give

$$\mathbb{E}Q^{-1} = \frac{1}{q-3}, \qquad \mathbb{E}Q^{-1/2} \le (\mathbb{E}Q^{-1})^{1/2} = \frac{1}{\sqrt{q-3}}, \tag{74}$$

where the inequality is Cauchy–Schwarz. Taking $k = \sqrt{2}s$ in equation 73 proves the second term of equation 69. The first is the oracle optimum when u is known: the likelihood-ratio score $u^\top x$ has AUROC $\Phi(\sqrt{2}s)$. Rearranging the second bound proves equation 70.

Finally $q - 3 = d - 4 = 1020$ and $\Delta = 1/4$, hence

$$n \ge \frac{\pi(1020)}{16} = 200.2765316663493\ldots. \tag{75}$$

The next permitted even sample size is 202. This budget is necessary for a guarantee uniform over the unknown direction. An algorithm hard-coded to a particular direction may still perform well when that direction is the correct one. □

The comparison concerns the information retained in exact historical posterior functions. Weights fitted from finite samples may also encode historical marginal information through their fluctuations, so they need not satisfy the same restriction. The constructive learner in Lemma 8 estimates a direction and its normalization statistics from a finite historical sample whose cost is counted explicitly. Lemma 9 then bounds the performance of the ridge readout fitted with two current labels.

K.2 A STATE LEARNED FROM FINITE HISTORICAL OBSERVATIONS

For the historical family in equation 58, we construct the coordinates from observations and signed labels. Neither unknown direction u nor w is supplied to the learner. The current distribution is equation 60 with $s = 1$, namely $X \mid y \sim \mathcal{N}(yu, I_d)$. One historical sample estimates the direction; an independent sample estimates the mean squared norms used for normalization. Both samples count toward the historical data budget and are independent of the later labeled support and evaluation observations.

Estimating the direction. Collect N independent observations $X_{j,y,i}$ for each of the four combinations of historical distribution and class, for a total of $T = 4N$ observations. Write

$$\overline{X}_{j,y} = \frac{1}{N}\sum_{i=1}^{N} X_{j,y,i}, \qquad \widehat{\theta}_j = \frac{\overline{X}_{j,+1} - \overline{X}_{j,-1}}{2}, \qquad h = \frac{\widehat{\theta}_{-1} + \widehat{\theta}_{+1}}{2}, \qquad e = \frac{h}{\|h\|}.$$

The event $h = 0$ has probability zero; a fixed unit vector defines the estimator on this event. Equal sample counts imply

$$h = \frac{1}{T}\sum_{j,y,i} y X_{j,y,i}.$$

The estimator therefore requires an $O(d)$ running sum and a count. Computing it uses signed labels, but neither the historical distribution index nor the coordinates of u or w. Equal counts across historical distributions and classes remain an assumption of the sampling design. The retained direction is the unsigned line spanned by e: changing its sign leaves the kernel of the augmented coordinates unchanged.

Lemma 4 (Exact estimation law)**.** *Under the balanced acquisition scheme,*

$$h - u \sim \mathcal{N}\left(0, \frac{I_d + uu^\top}{T}\right), \qquad \mathbb{E}\|h - u\|^2 = \frac{d+1}{T}. \tag{76}$$

Equivalently,

$$h = \left(1 + \sqrt{\frac{2}{T}}\, Z\right) u + \frac{G_\perp}{\sqrt{T}},$$

where $Z \sim \mathcal{N}(0,1)$ and $G_\perp$ is a standard Gaussian on the $(d-1)$-dimensional space $u^\perp$, independently.

Proof. The independent cell means give $\mathbb{E}\widehat{\theta}_j = \theta_j$ and $\mathrm{Cov}(\widehat{\theta}_j) = \Sigma_j/(2N)$. The opposing jw terms cancel in the mean, while

$$\mathrm{Cov}(h) = \frac{\Sigma_{-1} + \Sigma_{+1}}{8N} = \frac{I_d + uu^\top}{T}.$$

As a linear combination of independent Gaussian observations, h is Gaussian. Its covariance has eigenvalue $2/T$ along u and $1/T$ on $u^\perp$. This proves the decomposition and, by taking the trace, the mean-squared error. In particular, no projection onto $w^\perp$ is performed: the estimation noise has $d-1$ perpendicular coordinates. □

K.3 Concentration of the estimated line and its normalization

To bound the normalization error, we use a quadratic concentration inequality for independent Gaussian observations. It allows fixed, balanced counts for each combination of world and class without requiring the observations to be identically distributed.

Lemma 5 (Noncentral Gaussian quadratic concentration)**.** *Let $Z_i \sim \mathcal{N}(\mu_i, \Gamma_i)$ be independent Gaussian vectors, with positive definite covariance matrices. Set*

$$Q = \sum_i \|Z_i\|^2, \qquad V = \sum_i \left\{\mathrm{tr}(\Gamma_i^2) + 2\mu_i^\top \Gamma_i \mu_i\right\}, \qquad b = \max_i \|\Gamma_i\|_{\mathrm{op}}.$$

For every $x > 0$,

$$\mathbb{P}\Big(Q - \mathbb{E}Q > 2\sqrt{Vx} + 2bx\Big) \le e^{-x}, \tag{77}$$

$$\mathbb{P}\Big(Q - \mathbb{E}Q < -2\sqrt{Vx}\Big) \le e^{-x}. \tag{78}$$

Proof. Diagonalizing each covariance matrix and evaluating the Gaussian integral gives, for $0 < t < 1/(2b)$,

$$\log \mathbb{E}\exp\{t(Q - \mathbb{E}Q)\} = \sum_i \Big\{ -\frac{1}{2}\log\det(I - 2t\Gamma_i) - t\,\mathrm{tr}(\Gamma_i) + 2t^2 \mu_i^\top \Gamma_i (I - 2t\Gamma_i)^{-1}\mu_i \Big\}.$$

For $0 \le z < 1$, $-\log(1-z) - z \le z^2/[2(1-z)]$. Applying this inequality to each eigenvalue, and using $(I - 2t\Gamma_i)^{-1} \preceq (1-2bt)^{-1} I$, yields

$$\log \mathbb{E} \exp\{t(Q - \mathbb{E}Q)\} \le \frac{t^2 V}{1 - 2bt}.$$

Chernoff's inequality with $t = \sqrt{x}/(\sqrt{V} + 2b\sqrt{x})$ proves equation 77. For the opposite tail, the same integral has $I + 2t\Gamma_i$ in place of $I - 2t\Gamma_i$. The inequalities $z - \log(1+z) \le z^2/2$ for $z \ge 0$ and $(I + 2t\Gamma_i)^{-1} \preceq I$ give

$$\log \mathbb{E} \exp\{-t(Q - \mathbb{E}Q)\} \le t^2 V \quad (t > 0).$$

Taking $t = \sqrt{x/V}$ proves equation 78. □

Lemma 6 (Finite-sample alignment)**.** *For $x > 0$ and $T > 4x$, define*

$$A_T = 1 - 2\sqrt{\frac{x}{T}}, \qquad B_T = \frac{\sqrt{d-1} + \sqrt{2x}}{\sqrt{T}}, \qquad \rho_T = \frac{A_T}{\sqrt{A_T^2 + B_T^2}}.$$

Then $\mathbb{P}(|u^\top e| \ge \rho_T) \ge 1 - 2e^{-x}$. In particular, for $d = 1024$, $T = 16384$, and $x = \log(1000)$, $\rho_T > 0.96$, with failure probability at most 0.002.

Proof. In Lemma 4, the scalar Gaussian tail gives $\mathbb{P}(Z < -\sqrt{2x}) \le e^{-x}$. Applying equation 77 to a standard Gaussian vector of dimension $k = d - 1$ gives

$$\mathbb{P}\Big(\|G_\perp\|^2 > k + 2\sqrt{kx} + 2x \Big) \le e^{-x}.$$

Since $\sqrt{k + 2\sqrt{kx} + 2x} \le \sqrt{k} + \sqrt{2x}$, the axial coefficient of h is at least $A_T > 0$, and its perpendicular norm is at most B_T, outside an event of probability at most $2e^{-x}$. Normalizing h proves the claimed angle bound.

The numerical containment follows from rational bounds: $x < 7$, $\sqrt{1023} < 32$, $\sqrt{7} < 8/3$, and $\sqrt{14} < 15/4$. As $\sqrt{T} = 128$, these give

$$A_T > \frac{23}{24}, \qquad B_T < \frac{143}{512}, \qquad \frac{B_T}{A_T} < \frac{429}{1472} < \frac{7}{24}.$$

Therefore $\rho_T > 1/\sqrt{1 + (7/24)^2} = 24/25 = 0.96$. □

Estimating normalization scales. Independently of the direction sample, collect N_b observations $B_{j,y,i}$ for each combination of world and class in equation 58, for a total $M = 4N_b$. Define

$$\widehat{D} = \frac{1}{M} \sum_{j,y,i} \|B_{j,y,i}\|^2, \qquad \widehat{\kappa} = \frac{1}{M} \sum_{j,y,i} (e^\top B_{j,y,i})^2, \qquad L = \sqrt{\frac{\widehat{D}}{\widehat{\kappa}}}\, e. \tag{79}$$

These empirical energies are mean squared norms and are positive almost surely. Computing them uses the normalization observations and the estimated line; it requires neither labels in the calculation nor population moments. The normalization sample nevertheless counts toward the historical data budget and has equal counts for each combination of world and class. The normalized raw feature $z(x) = x/\sqrt{\widehat{D}}$ and its projection satisfy

$$\frac{1}{M} \sum_{j,y,i} \|z(B_{j,y,i})\|^2 = 1, \qquad \frac{1}{M} \sum_{j,y,i} \big(z(B_{j,y,i})^\top L\big)^2 = 1.$$

Lemma 7 (Concentration of empirical energies)**.** *Let $b = (3 + \sqrt{5})/2$. For any $x > 0$, with probability at least $1 - 2e^{-x}$,*

$$d + 3 - 2\sqrt{\frac{(d+15)x}{M}} \;\le\; \widehat{D} \;\le\; d + 3 + 2\sqrt{\frac{(d+15)x}{M}} + \frac{2bx}{M}. \tag{80}$$

Conditional on every unit line representative e,

$$\mathbb{P}\left(\widehat{\kappa} > 3 + 2\sqrt{\frac{9x}{M}} + \frac{6x}{M} \,\middle|\, e \right) \le e^{-x}. \tag{81}$$

Proof. The two-dimensional covariance block gives, for every world and class,

$$\mathbb{E}\|B_{j,y,i}\|^2 = d+3, \qquad \operatorname{tr}(\Sigma_j^2) = d+5, \qquad \theta_j^\top \Sigma_j \theta_j = 5, \qquad \|\Sigma_j\|_{\mathrm{op}} = b.$$

Thus the quadratic concentration proxy per observation is $(d+5)+2\cdot 5 = d+15$. Apply Lemma 5 to the independent normalization observations and divide its two bounds by M. This proves equation 80.

For the second bound, condition on e, which is independent of the normalization sample, and write $\rho = e^\top u$, $\xi = e^\top w$. Orthogonality implies $\rho^2+\xi^2 \le 1$, but ξ need not vanish. Each scalar projection has law

$$e^\top B_{j,y,i} \mid e \sim \mathcal{N}(y(\rho + j\xi),\, v_j), \qquad v_j = 1+\rho^2+2j\rho\xi.$$

The equal world counts cancel the terms that are odd in j, so

$$\mathbb{E}[\widehat{\kappa} \mid e] = 1+2\rho^2+\xi^2 \le 3.$$

To bound the quadratic proxy, put $r=\rho^2$, $t=\xi^2$ and $\mu_j^2 = (\rho+j\xi)^2$. Its average over the two worlds is

$$\begin{aligned}
\overline{V} &= \frac{1}{2}\sum_{j\in\{-1,+1\}} \left(v_j^2 + 2\mu_j^2 v_j\right) \\
&= 1+4r+3r^2+2t+14rt \\
&\le 3+16r-11r^2 \\
&= \frac{97}{11} - 11\left(r-\frac{8}{11}\right)^2 \le \frac{97}{11} < 9.
\end{aligned}$$

The inequality uses $t \le 1-r$ and $2+14r \ge 0$. Every projected Gaussian variance is bounded by $v_j \le \|\Sigma_j\|_{\mathrm{op}} = b < 3$. Lemma 5, applied conditionally to these independent scalar Gaussians, therefore has $V < 9M$ and maximum variance below 3. Using their mean bound 3 proves equation 81, uniformly in e. The calculation applies separately to the fixed counts from each Gaussian component, without treating their mixture as a single Gaussian. □

Lemma 8 (A finite historical good event)**.** *In dimension $d = 1024$, take $N = 4096$ observations per world–class cell for the direction and $N_b = 1024$ independent observations per cell for normalization. Thus $T = 16384$ and $M = 4096$. The state in equation 79 satisfies*

$$\mathbb{P}\Big(|u^\top e| \ge 0.96, \quad 1016.73 \le \widehat{D} \le 1037.27, \quad 0 < \widehat{\kappa} \le 3.3\Big) \ge 0.995. \tag{82}$$

The bound is uniform over all orthogonal unit u, w.

Proof. Set $x = \log(1000)$, so $e^{-x} = 0.001$. Lemma 6 gives the required alignment outside an event of probability at most 0.002. For $\widehat{D}$, Lemma 7 has failure probability at most 0.002. Its interval is contained in the stated one: using $x < 7$ and $b < 3$,

$$2\sqrt{\frac{1039x}{4096}} < 2\sqrt{\frac{7273}{4096}} < 3, \qquad \frac{2bx}{4096} < \frac{42}{4096} < 1.$$

Consequently that interval lies inside $(1024, 1031)$, which is itself contained in $[1016.73, 1037.27]$. For the projected energy,

$$3 + 2\sqrt{\frac{9x}{4096}} + \frac{6x}{4096} < 3 + \frac{1}{4} + \frac{42}{4096} < 3.3.$$

Its conditional failure probability is at most 0.001 for every e, and hence its unconditional failure probability is at most 0.001. Positivity holds almost surely by the positive projected Gaussian variances. The union bound gives total failure at most $0.002 + 0.002 + 0.001 = 0.005$. The two energies use the same normalization sample and need not be independent; the union bound does not require a product of their probabilities. □

Fitting the classifier with two current labels. The retained line and its two empirical energies require $O(d)$ storage. For a new observation, the ridge readout uses the augmented feature vector

$$\left[\frac{x}{\sqrt{\widehat{D}}},\ \frac{x^\top e}{\sqrt{\widehat{\kappa}}},\ 1\right].$$

With one current observation from each class, all its coefficients, including the intercept, are fitted to signed targets in $\{-1,+1\}$ by ridge regression with penalty 0.2. The current observations are independent of the retained direction and scales. On the event in Lemma 8, these quantities meet the assumptions of Lemma 9. That conditional lemma only requires the isotropic future law equation 60 with $s=1$; it does not require $e \perp w$. The historical data budget is $T+M=20480$ observations, counted separately from the two current labels. These estimates provide the state required by the finite Gaussian construction. The construction does not assume that an arbitrary historical optimizer or a combination of learned residual heads would estimate the same state.

K.4 A FINITE CERTIFICATE FOR THE TWO-LABEL RIDGE READOUT

We now bound the performance of the classifier fitted with two current labels. Throughout the following lemma, we condition on a historical state consisting of a unit representative e of a line and positive scales D,κ. The state is independent of the current support. The proof requires no relation between e and w, so it applies to the unprojected historical estimator in Lemma 8. The current classes are $\mathcal{N}(yu, I_d)$ with $\|u\|_2=1$.

Lemma 9 (Finite two-label ridge certificate)**.** *Suppose $d=1024$ and the historical state satisfies*

$$|u^\top e| \ge 0.96, \qquad D \in [1016.73, 1037.27], \qquad 0<\kappa\le 3.3.$$

Fit ridge regression to one independent current observation from each class, with signed targets $t=(+1,-1)^\top$ and features

$$\psi(x)=\left[\frac{x^\top}{\sqrt{D}}, \frac{x^\top e}{\sqrt{\kappa}}, 1\right] \tag{83}$$

and objective $\|\Psi\gamma - t\|_2^2 + 0.2\|\gamma\|_2^2$. The intercept is included in the penalty. Uniformly over all such states,

$$\mathbb{E}[\mathrm{AUC}_u(\widehat{f}) \mid e, D, \kappa] \ge C_{\mathrm{cert}} > 0.7713584756601982. \tag{84}$$

The expectation is over the two current support observations; AUROC is the population ranking probability on independent test examples.

Proof. With two support observations, the ridge solution reduces to a two-by-two matrix calculation. We bound its AUROC over rectangles in the remaining Gaussian coordinates, then certify the resulting finite sum.

The fitted slope and a lower bound valid for every support sample. For clarity, initially allow general d, signal $s>0$, and penalty $\lambda>0$, with $X_i = t_i s u + \varepsilon_i$. Let the rows of X be the two raw support vectors and put $v=Xe$. The fitted score has intercept $\mathbf{1}^\top\alpha$ and slope in the original feature space

$$\alpha = \left(\frac{XX^\top}{D} + \frac{vv^\top}{\kappa} + \mathbf{1}\mathbf{1}^\top + \lambda I_2\right)^{-1} t, \tag{85}$$

$$\beta = \frac{X^\top\alpha}{D} + \frac{e\,v^\top\alpha}{\kappa}. \tag{86}$$

These equations give the ridge solution on equation 83. Under the stated sum-of-squares objective, the dual penalty is λ, not 2λ. Conditional on a nonzero slope, the difference between independent positive and negative test scores is normal with mean $2su^\top\beta$ and variance $2\|\beta\|_2^2$. Consequently,

$$A(\beta) = \Phi\left(\frac{\sqrt{2}su^\top\beta}{\|\beta\|_2}\right) \ge f_0 := \Phi(-\sqrt{2}s). \tag{87}$$

A zero slope gives AUROC $1/2$ and satisfies the same floor. This bound accounts for every event outside the sharper analysis below.

Exact reduction with two support observations. Replacing e by $-e$ leaves the kernel and fitted score unchanged, so for the proof orient it such that $\rho = u^\top e \geq 0$. Write $u = \rho e + \eta f$, where $\eta = \sqrt{1-\rho^2}$ and f is a unit vector orthogonal to e; when $\eta = 0$, choose any such f. Rotate the two support rows by the orthogonal matrix with rows $t^\top/\sqrt{2}$ and $\mathbf{1}^\top/\sqrt{2}$. The target becomes $(\sqrt{2}, 0)^\top$ and the intercept kernel becomes $\mathrm{diag}(0, 2)$. In these rotated coordinates, retain X for the support matrix and write

$$v = (a, b)^\top, \qquad a \sim \mathcal{N}(\sqrt{2}s\rho, 1), \qquad b \sim \mathcal{N}(0, 1), \tag{88}$$

$$Y = XP_{e^\perp} = E + \sqrt{2}s\eta\,(1, 0)^\top f^\top, \qquad W = YY^\top. \tag{89}$$

Here E has two rows and $d-1$ independent standard normal coordinates per row. Conditional on the historical state, a, b, E are independent. Set

$$h = \frac{1}{D} + \frac{1}{\kappa}, \qquad B = \frac{W}{D} + \mathrm{diag}(\lambda, 2+\lambda), \qquad G = B + hvv^\top. \tag{90}$$

The matrix G is positive definite. Its explicit inverse gives

$$\alpha = \frac{\sqrt{2}}{\det G}\begin{pmatrix} B_{22} + hb^2 \\ -B_{12} - hab \end{pmatrix}, \tag{91}$$

$$c := e^\top \beta = \frac{h\sqrt{2}}{\det G}(aB_{22} - bB_{12}), \qquad \beta_\perp = P_{e^\perp}\beta = Y^\top \alpha / D. \tag{92}$$

Put $R = B_{12}/B_{22}$. Whenever $a - bR > 0$, we have $c > 0$ and

$$\frac{\alpha_1}{c} = A_{\mathrm{exact}} := \frac{1/h + b^2/B_{22}}{a - bR}, \qquad \frac{\alpha_2}{c} = -\frac{b}{B_{22}} - RA_{\mathrm{exact}}. \tag{93}$$

The random determinant therefore cancels from the ratios that determine the direction of the fitted slope.

Bounding the fitted component orthogonal to the historical direction. Suppose $\ell I_2 \preceq W \preceq U I_2$, $D \in [D_-, D_+]$, and $0 < \kappa \leq \kappa_+$. Then $|W_{12}| \leq (U-\ell)/2$ and the constants

$$B_0 = 2 + \lambda + \frac{\ell}{D_+}, \qquad R_0 = \frac{U-\ell}{2D_-B_0}, \qquad h_0 = \frac{1}{D_+} + \frac{1}{\kappa_+} \tag{94}$$

give $B_{22} \geq B_0$, $|R| \leq R_0$, and $h \geq h_0$. For $a > BR_0$ and $B \geq 0$, define the scalar functions

$$A_0(a, B) = \frac{1/h_0 + B^2/B_0}{a - BR_0}, \tag{95}$$

$$r_0(a, B) = \frac{\sqrt{U}}{D_-}\sqrt{A_0(a, B)^2 + \left(B/B_0 + R_0 A_0(a, B)\right)^2}. \tag{96}$$

The scalar argument B in these two functions denotes $|b|$, not the matrix in equation 90. Equation 93 implies $0 < \alpha_1/c \leq A_0(a, |b|)$ and $|\alpha_2/c| \leq |b|/B_0 + R_0 A_0(a, |b|)$. Together with $\|Y^\top\|_{\mathrm{op}} \leq \sqrt{U}$, this proves

$$c > 0, \qquad \frac{\|\beta_\perp\|_2}{c} \leq r_0(a, |b|). \tag{97}$$

The function r_0 decreases with a and increases with its nonnegative second argument on the positive-denominator region.

Let $\rho \geq \rho_0 \geq 0$ and $\eta_0 = \sqrt{1-\rho_0^2}$. Since $u^\top\beta \geq \rho c - \eta\|\beta_\perp\|_2$,

$$A(\beta) \geq \Phi\left(\sqrt{2}s\,\frac{\rho_0 - \eta_0 r_0(a, |b|)}{\sqrt{1 + r_0(a, |b|)^2}}\right). \tag{98}$$

Indeed $(\rho - \eta r)/\sqrt{1+r^2}$ decreases in $r \geq 0$, with derivative $-(\eta + \rho r)/(1+r^2)^{3/2}$, and increases with ρ. This argument also applies when the lower bound on alignment is negative.

Bounding the Gram matrix independently of the axial coordinates. Set $p = d - 1$ and, for $x_0 > 0$, define

$$\zeta = \sqrt{2} + \sqrt{2x_0} + \sqrt{2}s\eta_0, \qquad \ell = (\sqrt{p} - \zeta)_+^2, \qquad U = (\sqrt{p} + \zeta)^2. \tag{99}$$

Apply the bounds on the largest and smallest singular values to the $p \times 2$ standard Gaussian matrix $E^\top$. With probability at least $1 - 2e^{-x_0}$,

$$\sqrt{p} - \sqrt{2} - \sqrt{2x_0} \le \sigma_{\min}(E^\top) \le \sigma_{\max}(E^\top) \le \sqrt{p} + \sqrt{2} + \sqrt{2x_0}. \tag{100}$$

We use the standard Gaussian bound stated by Vershynin (Vershynin, 2010, Corollary 5.35). The deterministic mean added in equation 89 has operator norm $\sqrt{2}s\eta \le \sqrt{2}s\eta_0$. The singular-value perturbation inequality therefore yields

$$\ell I_2 \preceq W \preceq U I_2 \quad \text{on an event } \mathcal{E}_G \text{ with } \quad \Pr(\mathcal{E}_G \mid e, D, \kappa) \ge 1 - 2e^{-x_0}. \tag{101}$$

We take $\mathcal{E}_G$ to be the event in equation 100, which depends only on E. It is therefore independent of (a, b) conditional on history.

Combining local bounds into a finite analytic sum. Choose disjoint rectangles with $a \in [a_i, a_i + \Delta]$ and $|b| \in [b_j, b_j + \Delta]$, where $a_i, b_j \ge 0$ and $\Delta > 0$. For $\mu_0 = \sqrt{2}s\rho_0$ and $\mu_1 = \sqrt{2}s$, put

$$p_i^- = \Delta\, \phi\left(\max\{|a_i - \mu_1|, |a_i + \Delta - \mu_0|\}\right), \qquad q_j^- = 2\Delta\, \phi(b_j + \Delta). \tag{102}$$

Each bin probability is bounded below by its width times the minimum density over the bin. The first minimum also ranges over every $\mathbb{E}a \in [\mu_0, \mu_1]$, so it is uniform in the unknown alignment. If $a_i > (b_j + \Delta)R_0$, define

$$g_{ij} = \max\left\{ f_0, \Phi\left(\sqrt{2}s\, \frac{\rho_0 - \eta_0 r_0(a_i, b_j + \Delta)}{\sqrt{1 + r_0(a_i, b_j + \Delta)^2}} \right) \right\}; \tag{103}$$

otherwise set $g_{ij} = f_0$. Monotonicity in equation 96 makes g_{ij} a lower bound throughout each valid rectangle on $\mathcal{E}_G$. Every excess $g_{ij} - f_0$ is nonnegative. Independence and equation 101 now give the finite inequality

$$\begin{aligned} \mathbb{E}[A(\beta) \mid e, D, \kappa] &\ge \mathcal{L}(\rho_0, D_-, D_+, \kappa_+) \\ &:= f_0 + (1 - 2e^{-x_0}) \sum_{i,j} p_i^- q_j^- (g_{ij} - f_0). \end{aligned} \tag{104}$$

Outside the selected rectangles, on rectangles where the denominator is not positive, and when $\mathcal{E}_G$ fails, we retain f_0 from equation 87. We account for this probability mass without renormalizing the selected bins. Equation 104 therefore bounds the expected AUROC analytically; it is not a quadrature approximation to that expectation.

Explicit constants and certified arithmetic. To obtain the numerical guarantee in the lemma, set

$$\begin{gathered} d = 1024, \quad s = 1, \quad \lambda = 0.2, \quad \rho_0 = 0.96, \quad D_- = 1016.73, \quad D_+ = 1037.27, \quad \kappa_+ = 3.3, \\ x_0 = \log(100), \quad \Delta = 1/50, \qquad a_i = i/50 \ (0 \le i < 300), \quad b_j = j/50 \ (0 \le j < 200). \end{gathered} \tag{105}$$

There are $60{,}000$ rectangles in $a \in [0, 6]$, $|b| \in [0, 4]$. The released calculation certifies a positive gain on $57{,}795$ rectangles; the others contribute only the global floor.

Interval arithmetic at 32 decimal digits bounds the constants, densities, denominators, radii, and sum. To control numerical error in each CDF value, for $|z| \le 1.5$ the calculation uses

$$\Phi(z) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{13} \frac{(-1)^k z^{2k+1}}{2^k k!(2k+1)} + R_{14}(z), \tag{106}$$

with the explicit bound

$$|R_{14}(z)| \le \frac{1.5^{29}}{\sqrt{2\pi}\, 2^{14} 14!\, 29} < 1.232 \times 10^{-12} < 3 \times 10^{-12}. \tag{107}$$

The term magnitudes decrease in this range, so the first omitted term bounds the alternating remainder; negative arguments follow by symmetry. All needed CDF arguments lie in $[-\sqrt{2}, \sqrt{2}] \subset [-1.5, 1.5]$. The implementation includes the full error interval $[-3 \times 10^{-12}, 3 \times 10^{-12}]$.

Each step uses the endpoint that preserves a lower bound: lower probability masses and CDF values, lower positive denominators, and upper bounds on the perpendicular radii. Bins whose denominator or gain is not certified positive contribute no excess above the global floor. Replacing f_0 by its lower certified value is also conservative: the sum of the retained probability lower bounds, multiplied by $1 - 2e^{-x_0}$, is at most one, so the coefficient of the floor remains nonnegative. Thus the computed quantity $\widetilde{\mathcal{L}}$ is itself a lower certificate for equation 104.

The resulting certificate lies within the following outward-rounded decimal endpoints:

$$0.77135847566019821731621099463574130 \le \widetilde{\mathcal{L}} \le 0.77135847566019821731621099463603790. \tag{108}$$

Set C_{cert} to the lower endpoint in equation 108. This interval bounds the arithmetic certificate itself. It does not provide a two-sided interval for the true AUROC or the exact-CDF expression in equation 104. The guarantee follows from the one-sided inequality $\mathbb{E}A(\beta) \ge \mathcal{L} \ge C_{\mathrm{cert}}$.

For reproduction, run `certify_n2_bound.py` with `--case interval`. The certificate and per-a-bin contributions are saved in:

```
interval_normalization_certificate.json
interval_normalization_rows.csv
```

The calculation uses `mpmath` version 1.3.0; ordinary endpoint extraction has 75 decimal digits of precision. The saved endpoints are serialized exactly from their dyadic representations using integer arithmetic: $m2^{-k} = (m5^k)10^{-k}$. The final JSON serialization therefore cannot round a lower endpoint upward or an upper endpoint downward. This calculation supplies the conditional certificate; Lemma 8 supplies the historical event probability. This completes the finite conditional proof. □

Corollary 2 (Finite learned history and current label separation)**.** *For the historical estimator and independent normalization sample in Lemma 8,* 16,384 *historical direction-learning observations and* 4,096 *normalization observations suffice for the readout in Lemma 9, using two current labels, to satisfy*

$$\inf_{\substack{\|u\|_2=\|w\|_2=1 \\ u^\top w=0}} \mathbb{E}\,\mathrm{AUC}_u(\widehat{f}) > 0.767894931 > 0.75. \tag{109}$$

For the same current family, guaranteeing expected AUROC 0.75 *uniformly over* u *from exact rule memory requires at least* 202 *balanced current labels, even when* w *is given.*

Proof. Lemma 8 supplies the joint historical event $\mathcal{E}_H = \{|u^\top e| \ge .96,\ D \in [1016.73, 1037.27],\ 0 < \kappa \le 3.3\}$ with probability at least .995. It is independent of the current support. On that event apply Lemma 9; on its complement use $F = \Phi(-\sqrt{2})$ from equation 87. No independence between D, κ and e is required. Since $C_{\mathrm{cert}} > F$,

$$\mathbb{E}\,\mathrm{AUC}_u(\widehat{f}) \ge F + .995(C_{\mathrm{cert}} - F). \tag{110}$$

For an explicitly conservative decimal evaluation, the CDF remainder bound above gives $F > 0.07864960352$ and the conditional certificate gives $C_{\mathrm{cert}} > 0.7713584756601982$. Hence

$$\begin{aligned} \mathbb{E}\,\mathrm{AUC}_u(\widehat{f}) &> .005(0.07864960352) + .995(0.7713584756601982) \\ &= 0.767894931299497209 > 0.767894931. \end{aligned} \tag{111}$$

The historical event and the conditional certificate are uniform in the unknown orthogonal unit vectors, so the same is true of this average over direction-estimation, normalization, and current support samples. Corollary 1 supplies the comparison budget. The historical data cost is 20,480 observations, counted separately from current adaptation. The saving concerns current labels; it does not imply a lower total information cost for learning the unknown direction from scratch. □

Corollary 3 (Historical budget and two-label quality)**.** *For $d = 1024$, use the historical estimator in Lemma 4 with a predetermined direction budget $T \in \{4, 8, 12, \ldots\}$, an independent balanced normalization sample of 4096 observations, and the readout in Lemma 9. Write $F = \Phi(-\sqrt{2})$, $C = C_{\mathrm{cert}}$, and $[z]_+ = \max\{z, 0\}$. Define*

$$a = \frac{24}{7}\sqrt{1023}, \qquad b = 2 + \frac{24}{7}\sqrt{2},$$
$$x_T = \left(\frac{(\sqrt{T} - a)_+}{b}\right)^2, \qquad p_T = [.997 - 2e^{-x_T}]_+. \tag{112}$$

Uniformly over unknown orthogonal unit directions u, w,

$$\mathbb{E}\,\mathrm{AUC}_u(\widehat{f}_T) \geq Q_2(T) := F + p_T(C - F). \tag{113}$$

This lower bound is nondecreasing in the historical budget, with two current labels held fixed. For $F < q < F + .997(C - F)$, let

$$\epsilon_q = .997 - \frac{q - F}{C - F}, \qquad T_q = 4\left\lceil \frac{\left[a + b\sqrt{\log(2/\epsilon_q)}\right]^2}{4} \right\rceil. \tag{114}$$

Then T_q direction observations and 4096 normalization observations suffice for expected AUROC at least q.

Proof. Lemma 6 gives alignment at least .96 when $\sqrt{T} \geq a + b\sqrt{x}$, with failure probability at most $2e^{-x}$. For $\sqrt{T} > a$, choosing $x = x_T$ satisfies $T > 4x_T$ since $b > 2$. With $M = 4096$ and $x = \log(1000)$, Lemma 7 bounds normalization failure by .003, uniformly over the learned direction. A union bound therefore gives probability at least p_T for the same historical event used in Corollary 2. Apply the conditional certificate C on that event and the global AUROC floor F elsewhere. When $p_T = 0$, the global floor alone proves the claim. The expectation includes historical and current sampling. Monotonicity follows from that of x_T and p_T; solving $Q_2(T) \geq q$ and rounding upward to a multiple of four gives T_q. □

How historical coordinates change the penalty. The coordinates preserve all affine linear scores while changing their regularization. Fix a unit direction e and positive scales D, κ. For any support matrix X, target vector t, and $\lambda > 0$, ridge regression on $[x/\sqrt{D},\ e^\top x/\sqrt{\kappa},\ 1]$ is equivalent to

$$\min_{\beta, c}\left\{\|X\beta + c\mathbf{1} - t\|_2^2 + \lambda\left[D\|\beta_\perp\|_2^2 + \frac{D\kappa}{D + \kappa}(e^\top\beta)^2 + c^2\right]\right\}, \qquad \beta_\perp = \beta - (e^\top\beta)e. \tag{115}$$

To see this, write the augmented score as $a^\top x/\sqrt{D} + b\, e^\top x/\sqrt{\kappa} + c = \beta^\top x + c$. Every (β, c) is representable through the raw-feature block. For a fixed score, the coefficient constraint gives $a_\perp = \sqrt{D}\,\beta_\perp$ and $a_\parallel/\sqrt{D} + b/\sqrt{\kappa} = e^\top\beta$. Cauchy–Schwarz, with equality attained, gives $\min(a_\parallel^2 + b^2) = (e^\top\beta)^2/(D^{-1} + \kappa^{-1})$. Adding the perpendicular and intercept penalties proves the identity. Thus the penalty coefficient along e decreases from D to $D\kappa/(D + \kappa)$, while the perpendicular coefficient remains D. The identity holds for any support sample; the preceding Gaussian analysis establishes when this choice yields the stated label saving.

L HISTORICAL COORDINATES UNDER JOINT CHANGES

The historical separation establishes that past observations can reduce the cost of learning a new decision. We now allow both the class means and the covariance to change. The analysis separates information missing from historical coordinates from the error of fitting a current score, and gives a family of generator and processing changes for which the required coordinate dimension stays fixed.

L.1 What must the coordinates retain?

Fix a current world with

$$X \mid y \sim \mathcal{N}(m + y\theta, \Sigma), \qquad y \in \{-1, +1\}, \quad \Sigma \succ 0. \tag{116}$$

Let $U \in \mathbb{R}^{d\times r}$ have orthonormal columns, with $1 \le r < d$, and let V be an orthonormal complement. The coordinates $U^\top X$ may be learned from historical observations independent of the current support. For this subsection, condition on their realized value. Write the mean contrast and covariance in these coordinates as

$$a = U^\top\theta, \quad b = V^\top\theta, \qquad A = U^\top\Sigma U, \quad B = U^\top\Sigma V, \quad C = V^\top\Sigma V. \tag{117}$$

The full and retained squared signals are

$$s^2 = \theta^\top\Sigma^{-1}\theta, \qquad s_U^2 = a^\top A^{-1} a. \tag{118}$$

These quantities account for both class separation and feature noise.

Lemma 10 (Information retained by historical coordinates)**.** *Let $S = C - B^\top A^{-1}B \succ 0$ and $h = b - B^\top A^{-1}a$. Then*

$$\underbrace{s^2}_{\text{full signal}} = \underbrace{s_U^2}_{\text{retained signal}} + \underbrace{h^\top S^{-1} h}_{\Delta_U^2\text{: omitted signal}}. \tag{119}$$

The best population AUROCs using X and $U^\top X$ are, respectively, $\mathcal{A}^* = \Phi(\sqrt{2}s)$ and $\mathcal{A}_U^* = \Phi(\sqrt{2}s_U)$. When $s > 0$, their gap satisfies

$$G_U := \mathcal{A}^* - \mathcal{A}_U^* \le \frac{\Delta_U^2}{\sqrt{\pi}(s + s_U)}. \tag{120}$$

The gap vanishes exactly when $\Sigma^{-1}\theta \in \mathrm{range}(U)$.

Proof. Block elimination gives the standard Schur-complement identity

$$\theta^\top\Sigma^{-1}\theta = a^\top A^{-1}a + (b - B^\top A^{-1}a)^\top S^{-1}(b - B^\top A^{-1}a).$$

For the two Gaussian classes, the log likelihood ratio is $2\theta^\top\Sigma^{-1}(x - m)$. Its ranking is optimal at every false-positive rate; the difference between its independent positive and negative scores yields AUROC $\Phi(\sqrt{2}s)$. The same argument in the projected coordinates gives $\Phi(\sqrt{2}s_U)$, including optimality among nonlinear scores of $U^\top X$. Since $\sup_t \Phi'(t) = 1/\sqrt{2\pi}$,

$$G_U \le \frac{s - s_U}{\sqrt{\pi}} = \frac{\Delta_U^2}{\sqrt{\pi}(s + s_U)}.$$

Finally, the V block of $\Sigma^{-1}\theta$ equals $S^{-1}h$. It is zero exactly when $h = 0$, which is equivalent to retaining the full signal. If $s = 0$, both optimal AUROCs equal $1/2$. □

The mean contrast alone does not determine which coordinates are sufficient. Even if $b = 0$, correlations B can make $h \neq 0$: a feature outside U can help remove noise from a feature inside it. This is why changing processing conditions can require new directions without creating a new mean contrast.

L.2 Learning the current score from the same labeled support

Both the mean contrast and the covariance can be estimated from the current labels. Take $K \ge 2$ independent observations per class, project them as $z_{y,i} = U^\top X_{y,i}$, and form

$$\widehat{a} = \frac{\bar{z}_+ - \bar{z}_-}{2}, \qquad \nu = 2K - 2,$$

$$\widehat{A} = \frac{1}{\nu}\sum_{y\in\{-1,+1\}}\sum_{i=1}^{K}(z_{y,i} - \bar{z}_y)(z_{y,i} - \bar{z}_y)^\top, \qquad \widehat{f}(x) = \widehat{a}^\top\widehat{A}^{-1}U^\top x. \tag{121}$$

Under the sample-size condition below, $\widehat{A}$ is invertible almost surely. The same $2K$ labeled observations supply both estimates. The learner receives neither a population covariance nor an additional current unlabeled sample. Estimating the common midpoint only changes the score's intercept and hence does not affect AUROC.

Proposition 10 (Finite labels under changing covariance)**.** *Fix $\delta \in (0,1)$ and set*

$$q = \left(\sqrt{r} + \sqrt{2\log(3/\delta)}\right)^2, \qquad \epsilon_K = \sqrt{\frac{q}{2K}}, \quad \eta_K = \sqrt{\frac{q}{2K-2}},$$
$$D_K = \frac{\epsilon_K + s_U(2\eta_K + \eta_K^2)}{(1-\eta_K)^2}. \tag{122}$$

Suppose $2K - 2 > q$ and $s_U > 0$. Conditional on the historical coordinates, with probability at least $1-\delta$ over the current support,

$$0 \le \mathcal{A}^* - \mathrm{AUROC}(\widehat{f}) \le G_U + \min\left\{1, \frac{2D_K^2}{\sqrt{\pi} s_U}\right\}. \tag{123}$$

If $s_U = 0$, every score of the projected features has population AUROC $1/2$, so the gap equals G_U.

Proof. Gaussian sample means are independent of the within-class residuals. Consequently,

$$\widehat{a} \sim \mathcal{N}\left(a, \frac{A}{2K}\right), \qquad \nu\widehat{A} \sim \mathrm{Wishart}_r(A, \nu).$$

For analysis, write

$$\alpha = A^{-1/2}a, \quad e = A^{-1/2}(\widehat{a} - a), \quad W = A^{-1/2}\widehat{A}A^{-1/2}, \quad v = A^{1/2}\widehat{A}^{-1}\widehat{a} = W^{-1}(\alpha + e).$$

This whitening is used only in the proof. Here $\|\alpha\| = s_U$, $e \sim \mathcal{N}(0, I_r/(2K))$, and W has the law $G^\top G/\nu$ for a $\nu \times r$ standard Gaussian matrix G. Gaussian concentration and singular-value bounds give, jointly with probability at least $1-\delta$,

$$\|e\| \le \epsilon_K, \qquad (1-\eta_K)^2 I_r \preceq W \preceq (1+\eta_K)^2 I_r. \tag{124}$$

The norm event has failure probability at most $\delta/3$, and the two singular-value tails sum to at most $2\delta/3$ (Vershynin, 2010, Proposition 5.34 and Corollary 5.35). Since $\eta_K < 1$, this event implies

$$\|v - \alpha\| \le \|W^{-1}\|\,\|e\| + \|W^{-1}(I_r - W)\|\,\|\alpha\| \le D_K.$$

For $v \neq 0$, the fitted score has AUROC $\Phi(\sqrt{2}\alpha^\top v/\|v\|)$. The elementary normalization inequality gives

$$s_U - \frac{\alpha^\top v}{\|v\|} = \frac{s_U}{2}\left\|\frac{\alpha}{s_U} - \frac{v}{\|v\|}\right\|^2 \le \frac{2\|v-\alpha\|^2}{s_U}.$$

Applying the same bound on Φ' yields $\mathcal{A}_U^* - \mathrm{AUROC}(\widehat{f}) \le 2D_K^2/(\sqrt{\pi}s_U)$. The fitted slope is nonzero almost surely; a zero slope is assigned AUROC $1/2$. Adding $\mathcal{A}^* - \mathcal{A}_U^* = G_U$ proves the result. If $s_U = 0$, the two projected class distributions coincide. □

For s_U bounded away from zero and above, and $\eta_K \le 1/2$, the current estimation term is $O((r + \log(1/\delta))/K)$. This follows from $\epsilon_K^2 = q/(2K)$ and $\eta_K^2 = q/(2K-2)$, with the denominator in D_K bounded away from zero. The result concerns a projected Gaussian readout. The augmented ridge construction in Equation 3 retains the raw features and has its own two-label guarantee; Equation 123 instead quantifies the sufficiency of historical directions when the current covariance must be estimated.

L.3 A FIXED COORDINATE DIMENSION ACROSS CHANGING WORLDS

To make generation and processing explicit, consider source features and a common processing operation for both classes:

$$Z_g \mid y \sim \mathcal{N}(m_g + y\theta_g, \Sigma_g), \qquad X = T_e Z_g + c_e + \xi_e, \quad \xi_e \sim \mathcal{N}(0, \Gamma_e). \tag{125}$$

The noise ξ_e is independent of Z_g and the label, and $\Gamma_e \succeq 0$. The current mean contrast and covariance are

$$\theta_{g,e} = T_e\theta_g, \qquad \Sigma_{g,e} = T_e\Sigma_g T_e^\top + \Gamma_e \succ 0. \tag{126}$$

The following sufficient family permits changing decisions within a shared space. It places a structural condition on the changes, rather than a bound on how many generators or processing conditions may occur.

Corollary 4 (Shared coordinates under joint changes)**.** *Let $P_* = U_* U_*^\top$ be an orthogonal projection of rank r. Suppose, for every source and processing condition,*

$$P_*\theta_g = \theta_g, \qquad \Sigma_g P_* = P_* \Sigma_g, \qquad T_e P_* = P_* T_e, \qquad \Gamma_e P_* = P_* \Gamma_e. \tag{127}$$

Then U_* retains the full Gaussian detection information for every pair (g, e). For a historical estimate $\widehat{U}$ of the same rank, put $\varepsilon_H = \|\widehat{U}\widehat{U}^\top - P_*\|_{\mathrm{op}}$. In any fixed current world, let $s^2 = \theta^\top \Sigma^{-1}\theta$ and $\kappa_\Sigma = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$. Its omitted signal obeys

$$\Delta_{\widehat{U}}^2 \le \kappa_\Sigma \varepsilon_H^2 s^2, \qquad G_{\widehat{U}} \le \frac{\kappa_\Sigma \varepsilon_H^2 s}{\sqrt{\pi}}. \tag{128}$$

If $s_0 \le s \le s_1$ for fixed $0 < s_0 \le s_1 < \infty$, $\kappa_\Sigma \varepsilon_H^2 \le 1/2$, and $K \ge 2q+1$, the score in Equation 121 fitted in $\widehat{U}$ satisfies

$$\mathcal{A}^* - \mathrm{AUROC}(\widehat{f}) \le \frac{s_1}{\sqrt{\pi}} \kappa_\Sigma \varepsilon_H^2 + C(s_0, s_1) \frac{r + \log(3/\delta)}{K} \tag{129}$$

with conditional probability at least $1 - \delta$ over the current support. The constant depends only on the stated signal bounds, not on d or the number of possible pairs (g, e).

Proof. The commutation conditions say that the source covariance and processing preserve the two blocks $\mathrm{range}(U_*)$ and its orthogonal complement. Since P_* is symmetric, $T_e^\top$ also commutes with it. Thus $\theta_{g,e}$ lies in U_* and $\Sigma_{g,e}$ commutes with P_*. Its inverse preserves the same blocks, so $\beta_* = \Sigma^{-1}\theta$ lies in U_*. Lemma 10 gives exact retention.

For an estimated space, let $\widehat{P} = \widehat{U}\widehat{U}^\top$, $q_* = \Sigma^{-1/2}\theta$, and $\mathcal{W} = \Sigma^{1/2}\,\mathrm{range}(\widehat{U})$. The retained squared signal equals the squared norm of the Euclidean projection of q_* onto $\mathcal{W}$. Since $\Sigma^{1/2}\widehat{P}\beta_* \in \mathcal{W}$ and $P_*\beta_* = \beta_*$,

$$\begin{aligned}
\Delta_{\widehat{U}}^2 &= \mathrm{dist}^2(q_*, \mathcal{W}) \\
&\le \|\Sigma^{1/2}(I - \widehat{P})\beta_*\|^2 \\
&\le \lambda_{\max}(\Sigma)\varepsilon_H^2 \|\beta_*\|^2 \le \kappa_\Sigma \varepsilon_H^2 s^2.
\end{aligned}$$

Equation 120, with $s + s_{\widehat{U}} \ge s$, gives the second bound in Equation 128. The condition $\kappa_\Sigma \varepsilon_H^2 \le 1/2$ implies $s_{\widehat{U}} \ge s/\sqrt{2} \ge s_0/\sqrt{2}$. Also $K \ge 2q+1$ ensures $\eta_K \le 1/2$. On the event in Equation 124,

$$\frac{2D_K^2}{\sqrt{\pi} s_{\widehat{U}}} \le \frac{1}{\sqrt{\pi}} \left(\frac{32}{s_{\widehat{U}}} + 400 s_{\widehat{U}} \right) \frac{q}{K}.$$

Using $s_{\widehat{U}} \le s_1$ and $q \le 4(r + \log(3/\delta))$ proves Equation 129, for example with $C(s_0, s_1) = 4(32\sqrt{2}/s_0 + 400 s_1)/\sqrt{\pi}$. □

The guarantee is for each fixed current world, conditional on the historical estimate. If a bound on ε_H holds with historical probability $1 - \delta_H$, the combined success probability is at least $1 - \delta_H - \delta$. Historical coverage is needed to identify U_*; Equation 129 keeps that estimation cost separate from the current labels. The finite-history construction in Proposition 1 gives one explicit case in which useful coordinates can be learned. The shared-space conditions here allow different rotations, scales, and correlations within each block, so the optimal current score need not remain fixed.

L.4 Why coverage of mean directions is insufficient

An explicit two-dimensional example isolates the interaction between generation and processing. Let $U = e_1$, $\theta_g = e_1$, $\Sigma_g = \mathrm{diag}(1, \sigma_g^2)$, and

$$T_e = \begin{pmatrix} 1 & t_e \\ 0 & 1 \end{pmatrix}, \qquad \Gamma_e = 0, \qquad \Sigma_{g,e} = \begin{pmatrix} 1 + t_e^2\sigma_g^2 & t_e\sigma_g^2 \\ t_e\sigma_g^2 & \sigma_g^2 \end{pmatrix}. \tag{130}$$

Every mean contrast remains e_1, but

$$s^2 = 1, \qquad s_U^2 = \frac{1}{1 + t_e^2\sigma_g^2}, \qquad \Delta_U^2 = \frac{t_e^2\sigma_g^2}{1 + t_e^2\sigma_g^2}.$$

Take old and new source noise scales 0.1 and 10, and old and new mixing coefficients 0 and 1. The optimal AUROC in U is 0.92135 initially, 0.92135 after changing only the source, and 0.92032 after changing only the processing. Changing both reduces it to 0.55595. These values are direct evaluations of $\Phi(\sqrt{2/(1+t_e^2\sigma_g^2)})$, not image experiments. The full score $X_1 - t_e X_2$ removes the mixed noise and keeps AUROC 0.92135 in every case. Thus even relearning the best score inside the old space cannot replace the missing noise-removal direction.

M Learning a Shared Subspace from Finite Historical Moments

The shared-space result in Appendix L takes historical coordinate error as an input. Here we bound that error for the generalized eigenvectors in Appendix D. The argument uses the separation and variation operators in Equations 43–44, including their window weights. It identifies which historical directions can be recovered, accounts for errors in both operators, and then gives an explicit finite-sample guarantee.

M.1 Which historical distributions identify the space?

Work in fixed coordinates on the adaptive feature space after the projection $P_\perp$; write d for its dimension. Fix H retained windows with both labels present and equal effective counts for the two labels within each window. The capped counts define weights $\rho_w = N_w/N > 0$ with $\sum_w \rho_w = 1$. All counts and windows in this analysis are fixed independently of their observed feature values. Write the population cell means and covariances as

$$\mu_{y,w} = m_w + (2y-1)\theta_w, \qquad \mathrm{Cov}(z \mid y, w) = \Sigma_{y,w}, \qquad y \in \{0,1\}. \tag{131}$$

Let A_*, W_* be the operators obtained by substituting these population moments into the reference formulas, keeping the same count weights. They are population counterparts, not expectations of the nonlinear sample estimators. For $\bar\theta = \sum_w \rho_w \theta_w$ and weighted covariance Cov_ρ, direct expansion gives

$$\begin{aligned} A_* &= \bar\theta\bar\theta^\top + 4\sum_w \rho_w \theta_w \theta_w^\top, \\ W_* &= \frac{1}{2}\sum_w \rho_w(\Sigma_{0,w} + \Sigma_{1,w}) + \mathrm{Cov}_\rho(m_w) + \mathrm{Cov}_\rho(\theta_w). \end{aligned} \tag{132}$$

The midpoint term matters: pooling windows can introduce correlations even when each individual window has block-diagonal covariance.

Lemma 11 (Historical coverage identifies the population subspace)**.** *Let P_* be the orthogonal projector onto an r-dimensional space $\mathcal{S}_*$, $1 \le r < d$. Suppose all θ_w lie in $\mathcal{S}_*$ and span it, each $\Sigma_{y,w}$ commutes with P_*, and $\mathrm{Cov}_\rho(m_w)$ commutes with P_*. Then A_* has rank r and range $\mathcal{S}_*$, and W_* preserves $\mathcal{S}_*$ and its orthogonal complement. For fixed $\eta > 0$ and $\varepsilon > 0$, the leading r generalized eigenvectors of $A_* u = \lambda G_* u$, $G_* = \eta W_* + \varepsilon I$, span $\mathcal{S}_*$. Moreover, if*

$$\nu_H = \lambda_r\left(\sum_w \rho_w \theta_w \theta_w^\top\right) > 0, \qquad C_* = G_*^{-1/2} A_* G_*^{-1/2},$$

then the positive-to-zero eigengap obeys

$$\gamma := \lambda_r(C_*) \ge \frac{4\nu_H}{\lambda_{\max}(G_*)} > 0. \tag{133}$$

Proof. Balanced effective counts give population class means $\bar m \pm \bar\theta$ and global mean $\bar m$. Their between-class scatter is $\bar\theta\bar\theta^\top$, while the within-window contrast is $2\theta_w$. For W_*, averaging the covariance of $m_w + \theta_w$ and of $m_w - \theta_w$ cancels the cross terms and gives Equation 132. The separation operator is supported on $\mathcal{S}_*$ and is positive definite on that space because the contrasts span it. All terms of W_* preserve the two blocks, as does G_* and its inverse square root. Thus C_* has exactly r positive eigenvalues and range $\mathcal{S}_*$. Mapping its eigenvectors back by $G_*^{-1/2}$ preserves this range. For $v \in \mathcal{S}_*$, $v^\top A_* v \ge 4\nu_H \|v\|^2$ and $v^\top G_* v \le \lambda_{\max}(G_*)\|v\|^2$, giving the stated gap. □

The conditions allow different mean contrasts and different covariance matrices in every window. For example, midpoints can vary within $\mathcal{S}_*$ while their component in $\mathcal{S}_*^\perp$ stays fixed; covariances can change independently within both blocks. No individual historical decision needs to use every direction in $\mathcal{S}_*$. Opposite contrasts, θ_w and $-\theta_w$, contribute the same outer product and therefore reinforce the same direction. The quantity ν_H measures coverage of the weakest shared direction, rather than the number of recorded sources.

M.2 From moment error to coordinate error

Let $\widehat{A}, \widehat{W}$ be the reference sample operators and $\widehat{G} = \eta\widehat{W} + \varepsilon I$. For this subsection, assume the population conclusions of Lemma 11; the deterministic argument also applies whenever those conclusions hold directly. Set

$$m_G = \lambda_{\min}(G_*), \qquad \kappa_G = \frac{\lambda_{\max}(G_*)}{m_G}, \qquad L = \lambda_1(C_*),$$
$$\Delta_H = \|\widehat{A} - A_*\|_{\mathrm{op}} + L\|\widehat{G} - G_*\|_{\mathrm{op}}. \tag{134}$$

Here L and γ describe the population separation after accounting for within-class variation. The scale m_G is positive because of the regularizer. For the unweighted error $\mathcal{E}_H$ in Proposition 6, an explicit choice is

$$C_G = \frac{4\sqrt{\kappa_G}\max\{1, L\}}{m_G}, \qquad \mathcal{E}_H \le \frac{m_G\gamma}{4\max\{1, L\}}. \tag{135}$$

We prove the sharper weighted-error bound under $\Delta_H \le m_G\gamma/4$.

Proof of Proposition 6. Use the fixed population metric to define

$$E_A = G_*^{-1/2}(\widehat{A} - A_*)G_*^{-1/2}, \qquad E_G = G_*^{-1/2}(\widehat{G} - G_*)G_*^{-1/2}.$$

Write $a = \|E_A\|_{\mathrm{op}}$, $b = \|E_G\|_{\mathrm{op}}$, and $x = \Delta_H/m_G$. Then $a + Lb \le x \le \gamma/4$ and, since $L \ge \gamma$, $b \le x/\gamma \le 1/4$. In particular, $I + E_G$ is positive definite. Let

$$R = (I + E_G)^{-1/2}, \qquad J = R(C_* + E_A)R.$$

The eigenvalues of J are the generalized eigenvalues of $(\widehat{A}, \widehat{G})$. The bounds $\|R\|_{\mathrm{op}} \le (1-b)^{-1/2}$ and $\|R - I\|_{\mathrm{op}} \le (1-b)^{-1/2} - 1$ give

$$\|J - C_*\|_{\mathrm{op}} \le \frac{a + Lb}{1 - b} =: e \le \frac{x}{1-b} \le \frac{\gamma}{3}. \tag{136}$$

Indeed, write $RC_*R - C_* = (R - I)C_*R + C_*(R - I)$; its norm is at most $Lb/(1-b)$.

Let $\widehat{V}_r$ contain orthonormal eigenvectors for the largest r eigenvalues of J, with diagonal eigenvalue matrix $\widehat{\Lambda}_r$. The variational characterization of eigenvalues yields $\lambda_r(J) \ge \gamma - e > 0$ and $\lambda_{r+1}(J) \le e$. Since $(I - P_*)C_* = 0$, projecting the eigenvector equation gives

$$(I - P_*)\widehat{V}_r\widehat{\Lambda}_r = (I - P_*)(J - C_*)\widehat{V}_r, \qquad \|(I - P_*)\widehat{V}_r\|_{\mathrm{op}} \le \frac{e}{\gamma - e} \le \frac{2x}{\gamma}.$$

This is the rank-r spectral perturbation argument underlying subspace-angle bounds; see also Yu et al. (2014).

Generalized eigenvectors in the original feature coordinates span the columns of $Z = G_*^{-1/2}R\widehat{V}_r$. This change of metric must be included when measuring Euclidean subspace error. Because $G_*^{-1/2}$ commutes with P_*,

$$\|(I - P_*)Z\|_{\mathrm{op}} \le \frac{1}{\sqrt{m_G}}\left[\frac{e}{\gamma - e} + (1-b)^{-1/2} - 1\right], \qquad \sigma_{\min}(Z) \ge \frac{1}{\sqrt{\lambda_{\max}(G_*)(1+b)}}.$$

For $Q = Z(Z^\top Z)^{-1/2}$, therefore,

$$\|QQ^\top - P_*\|_{\mathrm{op}} \le \sqrt{\kappa_G(1+b)}\left[\frac{e}{\gamma - e} + (1-b)^{-1/2} - 1\right] \le \frac{4\sqrt{\kappa_G}\Delta_H}{m_G\gamma}. \tag{137}$$

The last step uses $(1-b)^{-1/2}-1 \le b \le x/\gamma$ for $b \le 1/4$ and $3\sqrt{5/4} < 4$. For two rank-r spaces, the projector difference equals the largest sine of their principal angles. If the reference learner retains $k \ge r$ leading directions, their space contains $\mathrm{range}(Q)$, so projection onto that larger space cannot increase $\|(I-\widehat{P})P_*\|_{\mathrm{op}}$. All these distances are at most one. Finally, $\Delta_H \le \max\{1, L\}\mathcal{E}_H$ gives the main-text statement with Equation 135. □

Retained rank and coordinate normalization. The result concerns the span of generalized eigenvectors, not the particular basis or its normalization. For the generalized-eigenvector learner's basis $\widehat{U}$ with $\widehat{U}^\top \widehat{G}\widehat{U} = I$, its Euclidean projector is $\widehat{P} = \widehat{U}(\widehat{U}^\top\widehat{U})^{-1}\widehat{U}^\top$, not $\widehat{U}\widehat{U}^\top$. The adaptive rank rule retains leading eigenvalues above a threshold and then applies the rank cap. With computed threshold $\widehat{\tau}$, the sufficient conditions $r_{\max} \ge r$ and $\widehat{\tau} < \gamma - e$ ensure at least r retained directions. If additionally $\widehat{\tau} \ge e$, exactly r survive the threshold. The implemented threshold is

$$\widehat{\tau} = \max\left\{\tau_{\mathrm{rel}}\widehat{\lambda}_1, d\epsilon_{\mathrm{mach}}\max(\widehat{\lambda}_1, 1)\right\},$$

with strict retention $\widehat{\lambda}_i > \widehat{\tau}$. These conditions account for the existing rank rule; they do not assume that a fixed threshold always recovers the population rank.

M.3 An explicit finite-history bound

We first relate cell-level moment errors to the two operators. Suppose all population cell means have norm at most M, and, simultaneously over cells,

$$\|\widehat{\mu}_{y,w} - \mu_{y,w}\| \le a_n, \qquad \|\widehat{Q}_{y,w} - Q_{y,w}\|_{\mathrm{op}} \le b_n, \quad Q_{y,w} = \mathbb{E}[zz^\top \mid y, w].$$

Fixed capped weights and balanced cells imply

$$\|\widehat{A} - A_*\|_{\mathrm{op}} \le 10Ma_n + 5a_n^2, \qquad \|\widehat{W} - W_*\|_{\mathrm{op}} \le b_n + 2Ma_n + a_n^2. \tag{138}$$

To see this, a class contrast has norm at most $2M$ and error at most $2a_n$. The difference of its outer products is at most $8Ma_n + 4a_n^2$. The global between-class term has weight $1/4$ and the within-window terms have weights summing to one, giving the first bound. For W, average the raw second-moment errors and use $2Ma_n + a_n^2$ for the error in each pooled class-mean outer product.

Corollary 5 (Finite Gaussian history learns the coordinates)**.** *Assume the conditions of Lemma 11. In each of the $2H$ Gaussian cells, take $n \ge 2$ independent observations with $\|\mu_{y,w}\| \le M$ and $\|\Sigma_{y,w}\|_{\mathrm{op}} \le \sigma^2$. Here n counts observations per class per historical window, giving $2Hn$ historical observations in total. Use the same n in every cell, so capped weighting gives $\rho_w = 1/H$ at every sample size. For $\delta_H \in (0,1)$, set*

$$\begin{aligned} h_H &= \sqrt{d} + \sqrt{2\log(6H/\delta_H)}, \qquad a_n = \frac{\sigma h_H}{\sqrt{n}}, \qquad t_n = \frac{h_H}{\sqrt{n-1}}, \\ b_n &= \sigma^2\left(2t_n + t_n^2 + \frac{1}{n}\right) + 2Ma_n + a_n^2, \\ D_n &= 10Ma_n + 5a_n^2 + L\eta(b_n + 2Ma_n + a_n^2). \end{aligned} \tag{139}$$

With probability at least $1-\delta_H$, $\Delta_H \le D_n$. If $D_n \le m_G\gamma/4$ and the reference learner retains at least r leading directions, its historical coordinate error satisfies

$$\varepsilon_H := \|(I-\widehat{P})P_*\|_{\mathrm{op}} \le B_n := \min\left\{1, \frac{4\sqrt{\kappa_G}D_n}{m_G\gamma}\right\}. \tag{140}$$

Define $e_n = D_n/[m_G(1 - D_n/(m_G L))]$ and $T(\ell) = \max\{\tau_{\mathrm{rel}}\ell, d\epsilon_{\mathrm{mach}}\max(\ell, 1)\}$. The rank condition holds throughout this event if $r_{\max} \ge r$ and $T(L+e_n) < \gamma - e_n$.

Proof. Each cell mean error is Gaussian with covariance $\Sigma_{y,w}/n$. Gaussian norm concentration bounds it by a_n, with failure probability at most $\delta_H/(6H)$. For the unbiased centered sample covariance $\widehat{\Sigma}$, Gaussian singular-value bounds give

$$\|\widehat{\Sigma} - \Sigma_{y,w}\|_{\mathrm{op}} \le \sigma^2(2t_n + t_n^2)$$

except on an event of probability at most $2\delta_H/(6H)$; these are the same bounds used in Proposition 10 (Vershynin, 2010). For $t_n \geq 1$, positive semidefiniteness supplies the lower bound and the upper singular-value bound still suffices. The identity $\widehat{Q} = (n-1)\widehat{\Sigma}/n + \widehat{\mu}\widehat{\mu}^\top$ then gives the stated b_n. A union bound over $2H$ cells establishes the simultaneous moment event. Apply Equation 138 and $\widehat{G} - G_* = \eta(\widehat{W} - W_*)$ to get $\Delta_H \leq D_n$, followed by Proposition 6. For the final rank condition, Equation 136 gives $e \leq e_n$. Hence $\widehat{\lambda}_1 \leq L + e_n$, $\widehat{\lambda}_r \geq \gamma - e_n$, and the monotone threshold function gives $\widehat{\tau} \leq T(L + e_n) < \widehat{\lambda}_r$. □

For fixed historical windows and fixed population operators, D_n is nonincreasing in n. Once the stated small-error conditions hold, B_n is therefore a nonincreasing upper bound on coordinate error. If n is at least a fixed multiple of $d + \log(H/\delta_H)$, then $B_n^2 = O((d + \log(H/\delta_H))/n)$, with constants depending on the signal, metric, coverage gap, and moment bounds. The historical label count is $N_H = 2Hn$; the result keeps this cost separate from future labels. Adding new windows can also change the population space and its coverage gap, so increasing the number of sources alone does not imply a monotone improvement.

M.4 Connecting the learned space to future adaptation

Suppose a fixed future world obeys the shared-space conditions of Corollary 4 with this same $\mathcal{S}_*$. The covariance is positive definite in the adaptive feature coordinates defined above; the optimal-AUROC comparator refers to these features. Freeze the coordinates learned from historical observations, and let $k \geq r$ be their retained rank. Use an orthonormal basis of their span to fit the Gaussian score in Equation 121 from an independent current support. Although the basis differs from the generalized-eigenvector learner's metric-normalized basis, an invertible change of projected coordinates leaves this unregularized Gaussian score unchanged.

Take K independent observations per class, or $2K$ current labels in total. With fixed $0 < s_0 \leq s \leq s_1 < \infty$, define $q_k = (\sqrt{k} + \sqrt{2\log(3/\delta_C)})^2$. If $\kappa_\Sigma B_n^2 \leq 1/2$ and $K \geq 2q_k + 1$, then, on the historical event and the retained-rank condition above, the current score satisfies

$$\mathcal{A}^* - \mathrm{AUROC}(\widehat{f}) \leq \underbrace{\frac{s_1\kappa_\Sigma}{\sqrt{\pi}} B_n^2}_{\text{learning coordinates from history}} + \underbrace{C(s_0, s_1)\frac{k + \log(3/\delta_C)}{K}}_{\text{learning the current decision}} \tag{141}$$

The conditional current success probability is at least $1 - \delta_C$. If the rank and current-budget conditions hold throughout the historical event, joint success is at least $1 - \delta_H - \delta_C$. The current budget may be chosen after observing the historical rank k; a fixed choice $K \geq 2q_{r_{\max}} + 1$ also suffices for the dimension condition.

For completeness, the information argument also covers $k > r$. The optimal current coefficient $\beta_* = \Sigma^{-1}\theta$ lies in $\mathcal{S}_*$, hence $\|(I - \widehat{P})\beta_*\| \leq B_n\|\beta_*\|$. The proof of Equation 128 therefore gives $\Delta_{\widehat{U}}^2 \leq \kappa_\Sigma B_n^2 s^2$ without requiring equal ranks. Apply Proposition 10 in dimension k to finish the bound.

Equation 141 connects the generalized-eigenvector learner's coordinate estimator to the shared-space adaptation analysis. Its current readout is the Gaussian estimator above; the generalized-eigenvector learner's regularized, pooled-moment classifier in Equation 46 is a different readout. For a target AUROC gap ζ larger than the historical term $H_n = s_1\kappa_\Sigma B_n^2/\sqrt{\pi}$, a sufficient current budget is

$$K \geq \max\left\{2q_k + 1, \frac{C(s_0, s_1)[k + \log(3/\delta_C)]}{\zeta - H_n}\right\}.$$

At a fixed retained rank, reducing the historical error decreases this sufficient label budget until the current-estimation requirement dominates.

Learning and removing history. The same identification result applies to any fixed retained-window set satisfying its conditions. Lemma 1 ensures that subtracting windows reconstructs exactly the operators for the remaining set. Their population coverage and moment errors determine the new guarantee. For a window set chosen from observed feature values, a uniform or selection-aware probability bound is additionally needed; the fixed-set concentration result above does not supply one.

N STATISTICAL CUE: EVALUATION PROTOCOLS AND COMPLETE COMPARISONS

These protocols accompany the two statistical CUE results assigned in Appendix A. The DINOv2 configuration tests historical reuse; the Community Forensics configuration tests standard-benchmark detection. Both use the joint-moment solver, with separate frozen front ends and initialization-derived scales.

N.1 HISTORICAL REUSE

The DINOv2-L extractor produces 1024 features, concatenated with 78 fixed image statistics. An initial 8000-row allocation fits the starting detector, standardization and 256-dimensional PCA; a separate 2000-row allocation sets the residual scale. These 10000 records contain 9999 unique images. The frozen scale is $\kappa = 195.28429704757548$. The feature transform and initial detector are shared by every paired history comparison.

Historical data contain four sources: SD3.5, FLUX.1, Ideogram and Seedream. Each source and class has three disjoint blocks of 32 images. The first blocks supply 256 unique observations for retained class-wise statistics, with 248 total within-cell degrees of freedom. The other two blocks supply 512 distinct historical calibration labels. A leave-one-source-out calibration used the second blocks as support and third blocks as evaluation data to choose one regularization strength from six fixed multiples of an initialization-derived reference: $\{1/16, 1/4, 1, 4, 16, 64\}$. Current-only, historical and random-orientation controls all selected $1/16$, giving $\lambda = 393.06698064271103$. This setting was then fixed for the reported target evaluations. The 512 calibration labels are additional to the initial 8000/2000 allocations; they are not the effective count cap used when initializing moments. The current-only control therefore removes retained statistical history, while sharing prior initialization and calibration.

The targets are Z-Image-Turbo and FLUX.2-klein-9B, with DOCCI real images. Each of the four source–processing conditions contains 256 real and 256 generated query images. Original images and processed views share identities; the real queries are also shared across generator conditions. The two-upload chain converts to RGB, resizes the longest side to at most 768 pixels with Lanczos interpolation, applies JPEG quality 70 with 4:2:0 chroma subsampling, resizes to at most 512 pixels, then applies JPEG quality 85 with the same subsampling. Saving the result as PNG preserves the resulting decoded image.

For each of six fixed support orders, budgets are nested within each class at $K \in \{1, 2, 4, 8, 16, 32, 64, 128\}$. The orders use seeds 20260918–20260923 and remain fixed across methods and processing variants. The support pools contain 128 real identities shared by both generators and 128 generated identities per generator, giving 384 unique support identities at the largest budget. Views and repeated orders do not add annotation costs. The two integrated arms give 384 prescribed fits: four conditions, six orders, eight budgets and two history settings. This is an evaluation on previously used development queries with its formula and settings fixed before the integrated comparison.

Both integrated arms use the same fixed JPEG contribution, sensor-excluding correction space and controller. The sensor uses 48 unlabelled paired identities; fitting its fixed scoring weight uses 48 separate labels. These costs are shared by both arms and are separate from current annotation savings. Each condition, order and budget starts with no current support and a provisional contribution. The current labels first audit the common initial score and then fit the classifier; they count once. Audit evidence is not accumulated across the nested-budget episodes. The controller uses discount 0.25, activation threshold 0.8 and suspension threshold 0.2.

Let $A(K)$ be AUROC averaged over the six fixed support orders. AULC is its trapezoidal integral against $\log_2 K$, divided by seven:

$$\mathrm{AULC} = \frac{A(1)/2 + A(2) + A(4) + A(8) + A(16) + A(32) + A(64) + A(128)/2}{7}.$$

Thus the two endpoint budgets have half weight, and AULC remains on the AUROC scale. The macro result gives equal weight to the four conditions. The paired bootstrap samples association groups within the real, Z-Image and FLUX.2 source strata. It retains all images in each sampled

group, with shared multiplicities across arms, budgets, support orders and processed views. AUROC remains image-weighted when groups differ in size. The 1000 fixed draws use seed 20260918 for these development queries. Support orders and the four linked conditions are averaged, not treated as independent experimental replicates. Table 27 gives every tested budget.

Table 27: **Complete learning curves with contribution control.** AUROC (%) averages six fixed support orders; K counts current labels per class. Each budget is a separate adaptation episode.

Model / K	1	2	4	8	16	32	64	128
Z-Image-Turbo / Original								
Current only	62.15	70.39	79.97	85.66	91.48	94.15	95.22	95.91
CUE	64.40	74.31	82.57	87.79	92.02	94.72	95.65	96.41
Z-Image-Turbo / Two uploads								
Current only	63.64	71.46	81.12	85.14	91.18	94.01	95.36	95.88
CUE	66.10	75.27	83.67	87.99	92.16	94.78	95.88	96.46
FLUX.2-klein-9B / Original								
Current only	61.49	72.25	84.51	88.11	94.37	96.70	97.21	97.75
CUE	63.74	75.76	87.10	91.06	94.22	96.28	97.07	97.52
FLUX.2-klein-9B / Two uploads								
Current only	63.82	70.21	84.11	85.81	93.60	96.31	96.86	97.50
CUE	65.73	74.85	86.64	90.09	93.89	96.08	97.03	97.55

N.2 Standard-Benchmark Comparison

The official Community Forensics ViT-S/16 model supplies 384-dimensional pre-logit features and the frozen initial score. The official linear detector is preserved; it is not retrained on the benchmark supports. Standardization and PCA to 256 dimensions use the fixed 8000-row initial allocation, and 2000 separate rows determine the scale. The transferred regularization multiplier is $1/16$; applying the same initialization recipe in this feature space gives $\lambda = 240.74777421095598$ and $\kappa = 132.8073255390887$. Each benchmark starts with empty retained history, $R = I$ and no controlled contributions, and fits one classifier using 16 real and 16 generated support images. No benchmark-specific parameter or classifier selection is performed.

All 22 released detectors use their saved checkpoint scores with zero current labels. Their query identities are matched to CUE after excluding known support, initialization, byte and pixel aliases. The retained counts are 491, 512, 512, 511, 459 and 512 for UFD, GenImage, AIGC-DB, AIGIBench, WildFake and Chameleon, respectively: 2997 images in 2993 known identity groups. No benchmark support image collides with the initial allocation under these records. The same masks are applied to every detector.

The comparison matches evaluation images, while current labels and pretraining differ between methods. Table 28 reports AUROC for all 22 released detectors and the statistical update. The 1.99-point mean gain over CUE's frozen initial detector has a paired 95% interval of 1.61–2.39 points. Benchmark-wise intervals are [1.08, 2.87], [0.25, 1.53], [0.09, 0.67], [1.92, 4.01], $[-0.13, 0.50]$ and [4.19, 7.65] points in the same order. These intervals use 1000 identity-group draws with seed 20260919 and condition on each benchmark's fixed support.

Table 28: **Complete AUROC comparison (%).** Released detectors use zero current labels; CUE uses 16 per class. Higher is better. Bold and underline mark the best and second-best values.

Method	UFD	GenImage	AIGC-DB	AIGIBench	WildFake	Chameleon
CNNDet	91.09	56.62	71.57	67.32	60.07	37.77
DIRE	56.27	49.63	49.49	50.84	54.58	63.09
PatchCraft	76.96	78.79	84.22	80.46	83.56	43.17
FreqNet	91.37	54.68	61.09	70.34	78.39	53.24
NPR	79.81	46.82	42.03	71.98	88.04	52.49
DRCT	70.05	97.55	88.58	85.39	85.76	73.12
SAFE	80.51	52.03	59.51	82.74	89.63	54.30
FerretNet	90.19	44.29	38.32	86.18	91.24	42.81
UnivFD	93.50	74.29	85.18	67.98	70.56	42.96
OmniAID	95.58	92.76	90.96	93.42	92.84	91.30
C2P-CLIP	70.81	91.16	87.73	77.15	88.70	55.96
RINE	98.04	93.16	94.22	75.46	87.72	51.86
Effort	81.17	90.50	87.12	82.28	91.27	52.53
IAPL	98.15	79.35	79.37	75.26	89.26	31.39
DDA	91.39	98.16	95.74	91.22	91.57	91.62
AIDE	92.66	71.50	76.94	70.96	86.20	44.46
CO-SPY	82.97	52.34	47.89	77.45	89.73	52.83
D^3	97.53	<u>99.56</u>	97.69	89.57	95.28	68.96
GAPL	<u>98.45</u>	98.53	97.93	95.67	98.66	80.90
CommunityForensics	97.50	98.89	<u>99.06</u>	<u>95.99</u>	<u>99.31</u>	<u>92.60</u>
LOTA	69.32	55.81	52.68	78.75	90.97	50.28
DEAR	77.28	97.70	85.52	81.40	87.98	86.13
CUE (16 labels/class)	**99.38**	**99.67**	**99.40**	**98.84**	**99.48**	**98.49**