Stability-Weighted Direction Regularization Disentangles Generator Shortcuts from Detection Signal

Jang-Ho Choi[1,2] **Seongho Kim**[1] **Jaehyun Choi**[1]

Dahye Kim[1] **Sungwon Yi**[1] **Eunho Yang**[2*]

[1]**AI Safety Institute, Electronics and Telecommunications Research Institute (ETRI)**
[2]Korea Advanced Institute of Science and Technology (KAIST)
{janghochoi,sbde500,pre6ent,dhaisi,sungyi}@etri.re.kr, eunhoy@kaist.ac.kr

Abstract

AI-generated image detectors often fail on unseen generators because the classifier relies on generator-specific feature directions that are predictive only on the training set. These generator-discriminative linear discriminant analysis (LDA) directions also carry real-vs-fake signal, so removing all of them after training discards detection evidence together with the shortcuts. We introduce **StaR**, a training-time regularizer that scores each LDA direction by its stability under leaving one training generator out and penalizes the classifier's alignment with low-stability directions, in proportion to their instability (continuous form) or below a stability threshold (thresholded form). On a CLIP detector trained on a generator-diverse corpus, the continuous form significantly improves the average out-of-distribution AUC over four benchmarks relative to the same detector trained without the penalty, and also improves over EFFORT. On a ResNet detector, the thresholded form makes out-of-distribution performance far less sensitive to the training seed, with no significant change in the mean. Post-hoc projection of the same directions onto frozen features, on the other hand, does not reproduce these training-time effects. StaR moves the classifier away from unstable generator directions while concentrating, rather than erasing, generator-discriminative structure in the features.

1 Introduction

Binary classifiers for AI-generated image detection can be highly accurate on generators seen during training, yet fail on generators that appear after deployment [18, 8]. A common failure mode is a *generator shortcut*, a feature direction that separates the training generators from real images but does not describe synthetic content in a generator-invariant way. When the classifier relies on such a direction, it may exploit generator-specific patterns rather than detection cues that generalize to unseen generators.

A direct remedy is to remove generator-discriminative directions from the feature space. This post-hoc strategy is unreliable because generator identity and detection signal are *entangled*. The linear discriminant analysis (LDA) directions that separate known generators also carry real-vs-fake evidence. We denote the same detector trained without the penalty as ERM (empirical risk minimization) in the tables and figures. On frozen ERM features, LDA projection, stability-filtered projection (SLDA,

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

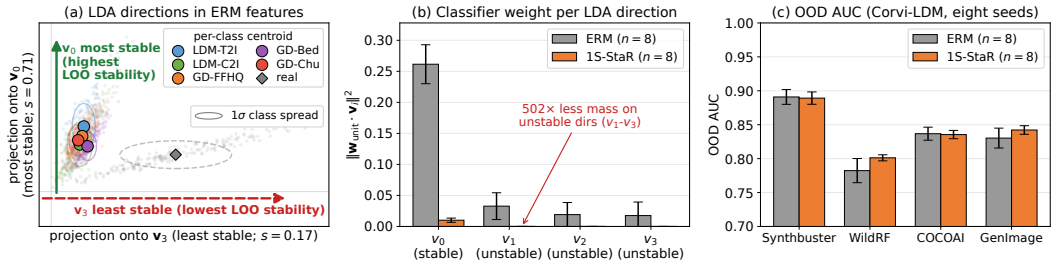


Figure 1: **Overview** (ResNet, Corvi-LDM). (a) Training features of one ERM model (seed 42) projected on its most and least stable LDA directions, with class centroids and 1σ spreads. (b) Squared projection of the unit classifier weight on each LDA direction, ERM vs. thresholded StaR, eight seeds per model: StaR strongly suppresses classifier mass on the unstable directions v_1–v_3 and concentrates the residual on v_0 (Table 12). (c) OOD AUC on Synthbuster, WildRF, COCOAI and GenImage, the ResNet Avg. (4) benchmarks, eight seeds per model: StaR has a substantially smaller across-seed SD while maintaining a similar mean (Table 8).

TSLDA) and LEACE-style erasure do not reproduce the training-time effect. On CLIP, none of these post-hoc methods improves over the baseline, and on ResNet the thresholded projection recovers the mean of training-time StaR but not its across-seed spread (Table 5, Appendix B.1). The remedy should therefore be both selective and training-time, keeping generator-shared detection evidence while suppressing directions that depend strongly on individual generators, before the classifier learns the shortcut.

Not every generator-discriminative direction is equally spurious (Figure 1). Recomputing the LDA basis after leaving out one training generator (leave-one-generator-out, LOO) assigns a stability score to every direction. Directions that persist across LOO subsets reflect structure shared across generators, while directions that change substantially when a generator is removed are more generator-dependent. This criterion identifies, within the same LDA subspace, the directions whose estimate depends on one training generator (§3). It is the basis of **StaR** (Stability-Weighted Direction Regularization), a training-time penalty that estimates LDA directions and their LOO stability online from exponential moving average (EMA) generator centroids and suppresses the classifier's projection along low-stability directions, weighted by instability or masked by a threshold.

We make three contributions.

- **Diagnosis: the entanglement problem and a stability criterion.** Generator-discriminative LDA directions entangle shortcut information with real-vs-fake detection signal, and post-hoc projection of these directions does not reproduce the training-time effect. We introduce LOO stability, which scores each direction by how consistently it is identified across held-out generators rather than by how much detection signal it carries. On the baseline model the most stable direction carries the least single-direction signal on two of the four benchmarks. After StaR training, the classifier weight concentrates on this direction, whose probe AUC increases on three of the four (§3).

- **StaR: a stability-weighted training-time regularizer.** The penalty acts on the linear classifier and is computed online from EMA-updated LDA statistics (Figure 2), so the basis follows the features as they change during training. It requires generator labels only at training time (§4).

- **Results on two backbones and six OOD benchmarks.** On CLIP, the continuous form improves over the baseline and over EFFORT on Avg. (4), the mean out-of-distribution (OOD) AUC over Synthbuster, WildRF, COCOAI and LDM-FakeDetect, with the largest gain on WildRF (Table 1). On ResNet, the thresholded form reduces the across-seed spread several-fold without a significant change in the mean (Table 8, Appendix B.1). StaR rotates the classifier weight largely out of the LDA subspace and concentrates the small residual on the most stable direction (Table 12).

2 Related work

Background. An AI-generated image detector is a binary classifier trained on real photographs and images from a set of known generators, and it is deployed on images from previously unseen generators [28, 18]. Detectors often reach near-perfect accuracy on seen generators but degrade substantially on unseen ones [8, 33]. One major cause is a generator shortcut, a feature direction that separates images from the training generators and real images without capturing synthetic-content cues that generalize across generators, the detection analogue of shortcut learning in general vision [12]. Existing mitigations act on the data, features or objective: broader training sets [33, 19], pretrained weights kept largely frozen so that the features are harder to overfit [18, 31], reconstruction or upsampling cues that transfer across generators [29, 24, 27], constraints on which features the classifier may use [21], and, outside detection, group-robust training and post-hoc erasure of known nuisance directions [26, 3]. StaR is closest to the classifier constraints and group-robust methods. It identifies the low-stability directions from the generator labels during training and penalizes the classifier for using them.

AI-generated image detection. Early detectors trained binary classifiers on GAN images [28, 11, 6]. Recent work targets diffusion models [8], using reconstruction error [29, 24] or upsampling artifacts [27]. Several recent detectors build on CLIP features [18, 17, 15]. Training-side approaches include diffusion reconstruction contrast [7], hybrid fusion of CLIP and DCT/SRM features [30], and causal feature masking in CausalCLIP [16].

EFFORT/OSD. EFFORT [31] decomposes the pretrained CLIP self-attention weights by SVD, fine-tunes only the residual component, and keeps the principal component that carries pretrained semantics frozen. In our matched-corpus comparison, continuous two-stage StaR improves over EFFORT on Avg. (4) (Table 1).

Stay-Positive and post-hoc weight constraints. Stay-Positive [21] shares our goal most directly. It argues that detectors should ignore real-image features and enforces this by clamping last-layer weights to be non-negative. An earlier paper from the same group [22] introduces the LDM benchmark that Stay-Positive extends to LDM-FakeDetect, which we use as one of our six OOD benchmarks. Both methods restrict which features the classifier may use. Stay-Positive constrains the sign of every last-layer weight, whereas StaR identifies low-stability generator-discriminative directions through LOO LDA and penalizes the classifier's projection onto those directions during training.

Spurious correlations and subspace methods. ERM-trained models exploit spurious correlations [12, 26]. Mitigations include invariant risk minimization (IRM) [1], group distributionally robust optimization (Group DRO) [26], deep feature reweighting (DFR) [14], and post-hoc concept erasure such as LEACE [3]. LDA underlies our diagnostic basis [10], and orthogonal projection has been used to debias word embeddings [4] and classifiers [23]. Our setting is distinct because generator-discriminative directions overlap with useful detection signal. Consequently, post-hoc removal does not reproduce the effect of training-time regularization (Tables 5 and 8).

3 Diagnosis

Setup. A detector maps an image to a feature vector $\phi(x) \in \mathbb{R}^D$ and applies a linear classifier $\mathbf{w}$. Training images come from K known generators plus real images. Evaluation uses unseen generators and shifted real-image sources.

Generator shortcuts are linearly available. A multinomial probe on intermediate features predicts the generator label with 63–94% accuracy across three backbones on our GenImage-based corpus (K=7 generator classes), far above chance (Appendix C.2), showing that generator identity is linearly accessible in the learned features.

Generator and detection directions are entangled. LDA on generator labels gives a generator-discriminative basis. Removing this subspace from the classifier after training does not reproduce the training-time effect. On ResNet, projecting out the full LDA subspace lowers Avg. (4), while the stability-filtered projections (SLDA, TSLDA) and LEACE-style erasure leave the mean near the baseline and retain substantially larger across-seed variation than training-time StaR. On CLIP, no post-hoc method improves on the baseline, and the full projection lowers Avg. (4) (Table 5, Appendix B.1).

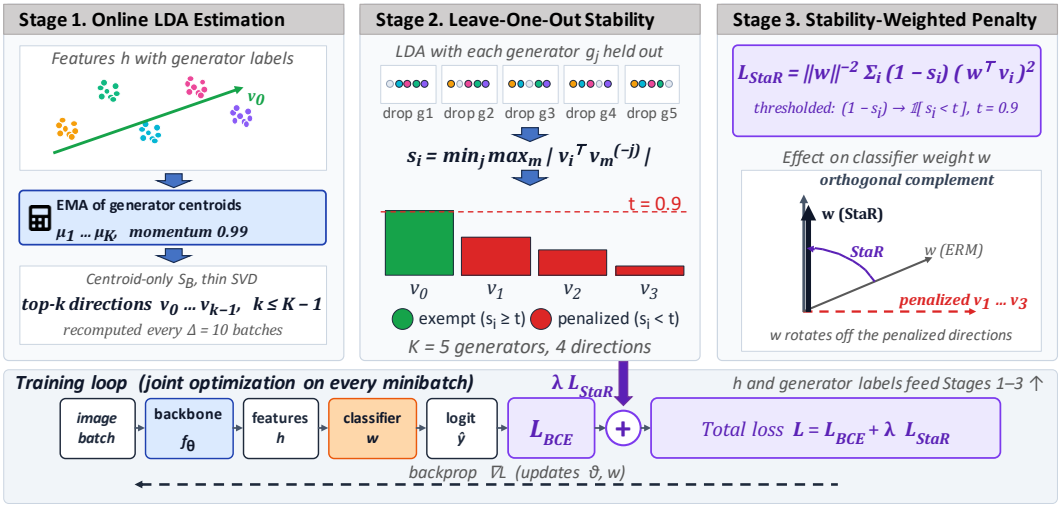


Figure 2: **StaR pipeline.** EMA-updated generator centroids give the centroid-only between-class scatter, whose top directions are scored by leave-one-generator-out stability (Eq. 1). In Stage 2 the stable direction (green) keeps its orientation across the held-out recomputations and an unstable one (red) rotates. The penalty (Eq. 2) weights the classifier's projection onto each direction by its instability, and Table 12 measures the rotation in Stage 3.

The loss under full projection on both backbones indicates that the generator-discriminative subspace contains useful detection evidence as well as shortcuts.

Stability measures consistency, not single-direction signal. The stability score measures how consistently a direction is identified when each training generator is held out. It does not order the directions by the detection signal each carries on its own. A one-dimensional real-vs-fake probe on each LDA direction of the baseline model, averaged over eight seeds, shows that the most stable direction carries the least single-direction signal on Synthbuster and COCOAI, significantly so, while the four directions are similar on WildRF and LDM-FakeDetect (Table 13). StaR therefore does not select the direction with the most signal. It penalizes the classifier's use of the directions whose estimate depends on a single generator. After training, the classifier weight concentrates on the most stable direction (Table 12), and that direction's probe AUC rises on Synthbuster, WildRF and LDM-FakeDetect but not on COCOAI (Table 14, Appendix B.6).

4 StaR: stability-weighted regularization

4.1 Leave-one-out stability

For each LDA direction $\mathbf{v}_i$ computed from all K training generators, we recompute the basis with generator g_j held out, yielding directions $\{\mathbf{v}_m^{(-j)}\}_m$. We define

$$s_i = \min_{j \in [K]} \max_m |\mathbf{v}_i^\top \mathbf{v}_m^{(-j)}|. \qquad (1)$$

The maximum over the $k' = \min(k, K-2)$ directions of the recomputed basis matches $\mathbf{v}_i$ to its closest one, so the score does not depend on the ordering of that basis. A high score indicates a direction estimated from structure shared across generators, whereas a low score indicates sensitivity to leaving out one generator. The minimum over j lets one held-out generator mark a direction as unstable, and Appendix C.8 compares it with

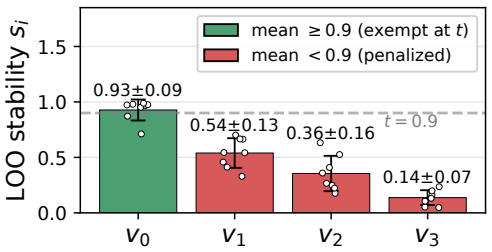


Figure 3: **Leave-one-generator-out stability on Corvi-LDM.** Scores of the four LDA directions of the ERM features, mean±SD over eight seeds with each seed as a circle. The dashed line is the threshold t=0.9 of the thresholded form, which exempts only v_0 on the eight-seed mean.

Algorithm 1 One StaR training step. The single-stage schedule initializes the backbone with its pretrained weights and the two-stage schedule starts from the ERM checkpoint. The LDA directions and penalty weights are held fixed between recomputations and receive no gradient. The pseudocode shows the steady state: a centroid's first observation sets it to the batch mean, no basis is computed until three centroids are initialized, and before all K centroids are initialized, the number of initialized centroids is used in place of K when computing the basis and k'.

Require: backbone ϕ, linear classifier $\mathbf{w}$, generator centroids $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$ with EMA momentum α, number of directions $k \le K-1$, regularization strength λ, penalty form (continuous, or thresholded at t), recompute interval Δ

run the batch, compute $\mathcal{L}_{\text{BCE}}$ on the logits, and detach a copy of the features $\phi(x)$ for the centroid update
for each training generator g present in the batch **do** ▷ real images are not used
    $\boldsymbol{\mu}_g \leftarrow \alpha \boldsymbol{\mu}_g + (1-\alpha)\,\text{mean}\{\phi(x) : x \text{ from } g\}$
end for
if the batch index is a multiple of Δ **then**
    $\mathbf{d}_g \leftarrow \boldsymbol{\mu}_g - \frac{1}{K}\sum_h \boldsymbol{\mu}_h$ for all g
    $\mathbf{v}_1, \dots, \mathbf{v}_k \leftarrow$ top-k right singular vectors of the $K \times D$ matrix with rows $\mathbf{d}_g$
    for $j = 1, \dots, K$ **do**
        $\mathbf{v}_1^{(-j)}, \dots, \mathbf{v}_{k'}^{(-j)} \leftarrow$ top-k' right singular vectors of the re-centered centroids without $\boldsymbol{\mu}_j$, where $k' = \min(k, K-2)$
    end for
    $s_i \leftarrow \min_j \max_m |\mathbf{v}_i^\top \mathbf{v}_m^{(-j)}|$ for $i = 1, \dots, k$ ▷ Eq. 1
    $\omega_i \leftarrow 1 - s_i$ (continuous) or $\mathbf{1}[s_i < t]$ (thresholded)
end if
$\mathcal{L}_{\text{StaR}} \leftarrow \|\mathbf{w}\|^{-2} \sum_{i=1}^{k} \omega_i (\mathbf{w}^\top \mathbf{v}_i)^2$ ▷ Eq. 2
one optimizer step on $\mathcal{L}_{\text{BCE}} + \lambda \mathcal{L}_{\text{StaR}}$ for the trainable parameters of ϕ and for $\mathbf{w}$

the mean and the median. Appendix D proves two one-sided bounds, that a direction whose scatter is dominated by one generator cannot align strongly with a held-out direction of comparable eigenvalue, and that a direction whose eigengap is large relative to the centroid norms cannot score low. The appendix also validates the criterion on synthetic centroids with known ground truth.

On Corvi-LDM, the K=5 generator centroids yield four directions. Their scores, averaged over eight baseline seeds, spread widely (Figure 3), and at the threshold $t = 0.9$ used below only v_0 is left unpenalized on the eight-seed mean.

4.2 StaR loss and online updates

StaR penalizes the linear classifier where it aligns with unstable LDA directions, weighted by their instability:

$$\mathcal{L}_{\text{StaR}} = \frac{1}{\|\mathbf{w}\|^2} \sum_{i=1}^{k} (1 - s_i)(\mathbf{w}^\top \mathbf{v}_i)^2, \qquad \mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda \mathcal{L}_{\text{StaR}}. \tag{2}$$

The factor $1/\|\mathbf{w}\|^2$ makes the penalty depend on the direction of $\mathbf{w}$ rather than on its norm. We call the $(1 - s_i)$ weighting of Eq. 2 the *continuous* form. Replacing it by the indicator $\mathbf{1}[s_i < t]$ with $t = 0.9$ gives the *thresholded* form, which fully penalizes directions below the threshold and leaves the others unpenalized. Each table states which form it reports. The requested k is capped at $K - 1$, the maximal rank of the LDA subspace. For a new corpus we suggest $k \le K - 2$ (§6).

Both the post-hoc baselines and online StaR use the centroid-only between-class scatter $\mathbf{S}_B$, so no streaming estimate of the within-class covariance is needed. Generator centroids are updated by EMA every batch (Figure 2). LDA directions and stability scores are recomputed every $\Delta = 10$ batches and held fixed between updates, which smooths the per-batch estimates. Each update requires $K + 1$ thin SVDs of a centroid matrix with at most K rows and D columns (the full basis and the K leave-one-out bases), so the overhead is small relative to the backbone forward/backward pass. Recomputing online keeps the basis and the scores aligned with the features as they rotate under the penalty. Algorithm 1 lists one training step.

5 Experiments

5.1 Setup

Data and models. ResNet experiments use the five-generator Corvi-LDM corpus [8] with the ImageNet-initialized ResNet-50-NoDown backbone ($D = 2048$), a ResNet-50 variant without early downsampling, introduced by Gragnaniello et al. [13] and used by Corvi et al., hereafter ResNet. CLIP experiments use CLIP ViT-L/14 [20] ($D = 1024$), hereafter CLIP, trained on a 50% GenImage [33] subset. OOD benchmarks are Synthbuster [2], WildRF [5], COCOAI [25], LDM-FakeDetect [22, 21], Chameleon [30], and CommunityForensics [19]. For the ResNet models, which are not trained on GenImage, GenImage is an additional OOD benchmark. The baselines are the same backbone trained with the same recipe and no penalty (ERM), EFFORT [31], the released detectors of Table 1, and the frozen-feature post-hoc methods of Table 5. CLIP training uses six GenImage subsets plus 4,000 ProGAN images from CNNDetection [28] (K=7 generator classes) and excludes real images from the LDA centroids. Code is available at `https://github.com/anti-fake/DeepfakeDetector/tree/main/StaR`.

Training and statistics. ResNet is trained with Adam (weight decay 0) for 20 epochs. CLIP uses AdamW for 20 epochs with batch size 32. StaR uses EMA momentum 0.99, k=5 capped at $K-1$ (four directions on the five-generator Corvi-LDM corpus, five on the K=7 CLIP corpus), and t=0.9 in the thresholded form. λ was selected on the reported OOD benchmarks (no separate validation split). For CLIP, λ=19 is the value used in the main tables, and the sweep in Appendix C.4 (thresholded form) shows no detectable differences across $\lambda \in [5, 50]$. For ResNet, λ=15 comes from a single-seed sweep with a standard ResNet-50 trained on GenImage, and a four-seed check with $\lambda \in \{10, 15, 25\}$ gives the highest Avg. (4) at 15 (Appendix A). Every multi-run entry is a mean±SD over runs, with eight runs per method unless a table states otherwise. The exceptions are the ResNet ablation companion (Table 10) and the ResNet λ check, both using four seeds. Differences between independently trained runs are tested with two-sided Welch t-tests and reported as plain p-values. Post-hoc methods, which act on the same checkpoints as their baseline, are tested on the within-checkpoint differences. Variance ratios use one-sided F-tests. SD uses $n - 1$. Avg. (6) averages the six OOD benchmarks of Table 1. Avg. (4) averages Synthbuster, WildRF, COCOAI and a fourth benchmark, LDM-FakeDetect for CLIP and GenImage for ResNet. ResNet rows use the final checkpoint, whereas CLIP rows use the checkpoint with the best in-domain validation AUC, measured on the held-out part of the training corpus (10% for ResNet, 50% for CLIP). Full hyperparameters are in Appendix A.

5.2 Main results: OOD generalization

Table 1 compares StaR with released detectors and with CLIP baselines trained on the same GenImage corpus, on six OOD benchmarks. Among the GenImage-trained CLIP models, the continuous form of 2S-StaR achieves the highest Avg. (6), significantly above the baseline. It is followed by the thresholded form, the baseline, single-stage StaR (λ=2, chosen in a single-seed sweep) and EFFORT. The largest gain over the baseline is on WildRF.

On Avg. (4), the continuous form improves over the baseline by 1.8 points (Welch p=0.0013) and improves over EFFORT by a larger margin, with the higher mean on every benchmark. The thresholded form is slightly above the baseline, though the difference is not significant. The continuous form outperforms the thresholded form on both averages and is our recommended CLIP configuration. A classifier-only L2 penalty at one tuned strength on the same warm-start continuation (StaR penalty off) matches the mean of the continuous form, while the continuous form has a smaller across-run spread (Table 15, Appendix C.1). The dependence of gain on the warm start is examined in §6 (Table 6).

On ResNet, the main effect is a smaller across-seed spread with a small mean difference. Thresholded StaR has a 4.1× smaller across-seed SD of Avg. (4) than the baseline (one-sided F, p=0.0007) while its mean is 0.7 points higher, which is not significant (Welch p=0.14, Table 8, Appendix B.1). The continuous form reduces the spread much less on this backbone and has a lower mean than the thresholded form, so the thresholded form is our recommended ResNet configuration. Part of the spread reduction is not specific to StaR. A classifier-only L2 penalty on the baseline reduces the

Table 1: **OOD AUC on six benchmarks.** Released detectors are evaluated using their public checkpoints. The CLIP rows report eight runs each: two-stage (2S) StaR at λ=19 for both weighting forms and single-stage (1S) StaR at λ=2. The two-stage runs share one warm start, the CLIP-ERM run (seed 42) and vary the stage-2 seed. Avg. (4) averages Synthbuster, WildRF, COCOAI and LDM-FakeDetect and Avg. (6) averages all six benchmarks. **Bold**: best in column.

Method	Synth.	WildRF	COCOAI	LDM-FD	Chameleon	CommFor	Avg. (4)	Avg. (6)
Released checkpoints (one model each)								
CNNSpot [28]	0.525	0.431	0.548	0.476	0.344	0.529	0.495	0.476
UnivFD [18]	0.622	0.659	0.618	0.595	0.487	0.584	0.624	0.594
NPR [27]	0.526	0.695	0.717	0.834	0.558	0.798	0.693	0.688
AIDE [30]	0.648	0.595	0.641	0.883	0.594	0.857	0.692	0.703
DRCT-ConvNeXt-B [7]	0.840	0.783	0.808	**0.945**	0.724	0.848	0.844	0.825
GenImage-trained CLIP, eight runs per row, mean±SD								
EFFORT [31]	$0.890_{\pm.018}$	$0.829_{\pm.025}$	$0.966_{\pm.014}$	$0.871_{\pm.023}$	$0.638_{\pm.073}$	$0.900_{\pm.015}$	$0.889_{\pm.015}$	$0.849_{\pm.019}$
CLIP-ERM (ours)	$0.913_{\pm.015}$	$0.858_{\pm.025}$	$0.984_{\pm.004}$	$0.922_{\pm.014}$	$0.820_{\pm.026}$	$0.934_{\pm.012}$	$0.919_{\pm.009}$	$0.905_{\pm.011}$
1S-StaR, λ=2 (ours)	$0.907_{\pm.019}$	$0.848_{\pm.031}$	$0.981_{\pm.010}$	$0.909_{\pm.019}$	$0.792_{\pm.040}$	$0.932_{\pm.019}$	$0.911_{\pm.016}$	$0.895_{\pm.018}$
2S-StaR, thresholded (ours)	$0.922_{\pm.016}$	$0.883_{\pm.034}$	$0.982_{\pm.005}$	$0.919_{\pm.017}$	$0.795_{\pm.028}$	$0.937_{\pm.008}$	$0.926_{\pm.014}$	$0.906_{\pm.013}$
2S-StaR, continuous (ours)	$\mathbf{0.928_{\pm.017}}$	$\mathbf{0.905_{\pm.016}}$	$\mathbf{0.987_{\pm.001}}$	$0.929_{\pm.010}$	$\mathbf{0.838_{\pm.021}}$	$\mathbf{0.944_{\pm.009}}$	$\mathbf{0.937_{\pm.009}}$	$\mathbf{0.922_{\pm.008}}$

across-seed SD by two- to three-fold at approximately the baseline mean, and StaR is slightly above these controls (Table 2, details in Appendix B.5).

Table 2: **Classifier-only L2 controls on ResNet** (Corvi-LDM, Avg. (4), final checkpoint, n in parentheses). SD ratio: ERM SD divided by the row SD, with the one-sided F p. Δ: mean difference from ERM, in points, with the two-sided Welch p.

Method (n)	Avg. (4)	SD ratio (p)	Δ vs ERM (p)
ERM (8)	$0.835_{\pm.012}$	ref.	ref.
1S-StaR, thresholded (λ=15) (8)	$0.842_{\pm.003}$	4.1× (<.001)	+0.7 (.143)
ERM + classifier-only L2 10^{-3} (8)	$0.836_{\pm.004}$	3.1× (.004)	+0.1 (.784)
ERM + classifier-only L2 10^{-2} (8)	$0.838_{\pm.006}$	2.0× (.039)	+0.3 (.515)

5.3 Ablation studies

StaR differs from generic regularization in two choices, the LDA directions it targets and the leave-one-out stability weighting. We ablate each while holding the rest fixed to the configuration of Table 1 (two-stage CLIP on the GenImage corpus, λ=19, k=5, and t=0.9 in the thresholded form). Table 3 reports Avg. (4) over Synthbuster, WildRF, COCOAI and LDM-FakeDetect for both weighting forms. Each ablation, with eight runs, is compared with the eight-run full method of the same form using a two-sided Welch test (four comparisons per block). The ResNet companion with four seeds is Table 10 (Appendix B.2).

The two forms respond differently to the ablations. In the continuous form, the recommended CLIP configuration, every ablation lowers Avg. (4), with the largest per-benchmark losses on WildRF. Replacing the LDA directions (perpendicular, random) and removing the stability weighting (uniform, anti-StaR) both lower the average, and uniform weighting is the only ablation whose decrease is significant. The thresholded form is not significantly above CLIP-ERM on this backbone (Table 1), and accordingly no ablation lowers its Avg. (4). At t=0.9 it exempts only the top direction (Table 19) and penalizes the other four at full weight, so it differs from uniform weighting only in that exemption. On ResNet, where the thresholded form is the recommended configuration, every ablation returns Avg. (4) to the ERM level (Table 10, Appendix B.2). Sweeps of k and t (Appendix C.3) show only small differences across settings.

Weaker grouping signals. StaR needs generator labels only at training time. Table 4 replaces the seven generator labels of the CLIP corpus with four coarse family labels or k-means pseudo-labels computed from CLIP-ERM features, using the same warm start. Family labels retain most of the continuous-form gain and remain significantly above CLIP-ERM, and in the thresholded form they are similar to generator labels. k-means pseudo-labels, which split generators across clusters, fall below CLIP-ERM in the thresholded form at four runs, although the difference is not significant (Appendix C.5).

Table 3: **Ablation of StaR components** (CLIP, two-stage, λ=19, k=5). *anti-StaR* reverses the stability weighting so that stable directions are penalized, *perpendicular-to-LDA* targets the orthogonal complement of the LDA subspace, *random directions* replaces the LDA basis by random orthonormal directions (all penalized, since their stability is near zero), and *uniform weights* penalizes all k directions equally. Δ is the Avg. (4) difference from the full-method row of the same form (the Table 1 row, repeated here), in points, with the two-sided Welch p.

Variant	Synth.	WildRF	COCOAI	LDM-FD	Avg. (4)	Δ Avg. (4) (p)
Thresholded form						
2S-StaR, thresholded, full method (n=8)	$0.922_{\pm.016}$	$0.883_{\pm.034}$	$0.982_{\pm.005}$	$0.919_{\pm.017}$	$0.926_{\pm.014}$	ref.
anti-StaR (n=8)	$0.929_{\pm.011}$	$0.889_{\pm.019}$	$0.986_{\pm.003}$	$0.920_{\pm.026}$	$0.931_{\pm.011}$	+0.5 (.456)
perpendicular-to-LDA (n=8)	$0.928_{\pm.014}$	$0.876_{\pm.011}$	$0.982_{\pm.005}$	$0.923_{\pm.021}$	$0.927_{\pm.009}$	+0.1 (.884)
random directions (n=8)	$0.923_{\pm.017}$	$0.886_{\pm.025}$	$0.983_{\pm.008}$	$0.924_{\pm.026}$	$0.929_{\pm.015}$	+0.2 (.738)
uniform weights (n=8)	$0.918_{\pm.019}$	$0.889_{\pm.017}$	$0.981_{\pm.004}$	$0.919_{\pm.023}$	$0.927_{\pm.011}$	0.0 (.958)
Continuous form						
2S-StaR, continuous, full method (n=8)	$0.928_{\pm.017}$	$0.905_{\pm.016}$	$0.987_{\pm.001}$	$0.929_{\pm.010}$	$0.937_{\pm.009}$	ref.
anti-StaR (n=8)	$0.923_{\pm.018}$	$0.891_{\pm.036}$	$0.986_{\pm.005}$	$0.917_{\pm.023}$	$0.929_{\pm.018}$	−0.8 (.287)
perpendicular-to-LDA (n=8)	$0.917_{\pm.023}$	$0.883_{\pm.021}$	$0.986_{\pm.004}$	$0.913_{\pm.023}$	$0.925_{\pm.014}$	−1.3 (.053)
random directions (n=8)	$0.927_{\pm.015}$	$0.885_{\pm.028}$	$0.984_{\pm.005}$	$0.913_{\pm.017}$	$0.927_{\pm.014}$	−1.0 (.098)
uniform weights (n=8)	$0.911_{\pm.027}$	$0.875_{\pm.039}$	$0.983_{\pm.004}$	$0.913_{\pm.020}$	$0.921_{\pm.018}$	−1.7 (.038)

Table 4: **Weaker grouping signals on CLIP** (2S-StaR, λ=19). Family labels merge the seven training generators into four coarse families. k-means pseudo-labels cluster CLIP-ERM features without any generator label. Δ is the Avg. (4) difference from the eight-run CLIP-ERM mean, in points, with the two-sided Welch p.

Grouping signal (n)	Avg. (4)	Δ vs CLIP-ERM (p)
Generator labels, thresholded (8)	$0.926_{\pm.014}$	+0.7 (.265)
Family labels, thresholded (8)	$0.928_{\pm.009}$	+0.9 (.078)
Generator labels, continuous (8)	$0.937_{\pm.009}$	+1.8 (.001)
Family labels, continuous (8)	$0.934_{\pm.012}$	+1.4 (.019)
k-means pseudo-labels (K=7), thresholded (4)	$0.909_{\pm.009}$	−1.1 (.112)
CLIP-ERM (8)	$0.919_{\pm.009}$	ref.

Post-hoc projection on the same checkpoints. Table 5 on CLIP and Table 8 on ResNet (Appendix B.1) compare training-time StaR with projection and erasure methods applied to the frozen features of the same eight baseline checkpoints (LDA projection, SLDA, TSLDA and LEACE-style erasure). The ResNet training-time rows repeat the result above. On ResNet, among the post-hoc rows, projecting the classifier off the full LDA subspace lowers Avg. (4), the stability-filtered projections raise it slightly, and LEACE-style erasure leaves it at the ERM level (paired tests over the eight shared checkpoints in Table 8). Their across-seed SDs of Avg. (4) remain at least twice StaR's. On CLIP, no post-hoc method improves on the baseline, and continuous 2S-StaR is significantly above the best post-hoc row (Table 5). Therefore, stability-filtered post-hoc projection can recover the ResNet mean, but it recovers neither the ResNet across-seed spread nor the CLIP gain.

5.4 Rotation, not erasure: how StaR reshapes the classifier

StaR changes which feature directions the classifier uses without removing generator information from the representation. Table 12 (Appendix B.4) shows that thresholded StaR reduces the classifier mass on the unstable LDA directions (v_1–v_3) by more than two orders of magnitude and the total LDA-subspace projection by more than an order of magnitude, leaving the residual concentrated on the most stable direction.

This rotation is not feature erasure. At the penultimate layer, generator-probe accuracy is essentially unchanged and the between-class scatter concentrates on the top LDA direction, whose squared singular value grows while the second falls by more than an order of magnitude (one seed pair, Appendix B.4). Two weight-regularization controls help distinguish the mean and variance effects in the ResNet result. Weight decay on all parameters lowers OOD AUC at every value tried (single seed,

Table 5: **Post-hoc vs. training-time regularization on CLIP** (GenImage). Post-hoc rows apply LDA projection, SLDA, TSLDA or LEACE-style erasure (Appendix B.1) to the frozen features and classifier of the eight CLIP-ERM runs. Δ: Avg. (4) difference from CLIP-ERM in points, with paired tests over the shared runs for the post-hoc rows and Welch tests for the independently trained rows. The ResNet comparison is Table 8 in Appendix B.1.

Method	Synth.	WildRF	COCOAI	LDM-FD	Avg. (4)	Δ vs CLIP-ERM (p)
Post-hoc, on the frozen features of the eight CLIP-ERM runs						
CLIP-ERM (frozen features)	$0.913_{\pm.015}$	$0.858_{\pm.025}$	$0.984_{\pm.004}$	$0.922_{\pm.014}$	$0.919_{\pm.009}$	ref.
LDA projection of $\mathbf{w}$	$0.896_{\pm.023}$	$0.854_{\pm.025}$	$0.984_{\pm.004}$	$0.915_{\pm.016}$	$0.912_{\pm.011}$	-0.7 (.011)
SLDA (stability-weighted projection)	$0.912_{\pm.016}$	$0.857_{\pm.024}$	$0.984_{\pm.003}$	$0.920_{\pm.014}$	$0.918_{\pm.009}$	-0.1 (.106)
TSLDA (projection off directions with $s_i < 0.5$)	$0.911_{\pm.016}$	$0.855_{\pm.024}$	$0.983_{\pm.003}$	$0.917_{\pm.014}$	$0.917_{\pm.009}$	-0.3 (.100)
LEACE-style (erase subspace, refit probe)	$0.855_{\pm.044}$	$0.810_{\pm.047}$	$0.979_{\pm.006}$	$0.893_{\pm.044}$	$0.884_{\pm.028}$	-3.5 (.004)
Training-time, two-stage from the CLIP-ERM seed-42 warm start (eight runs)						
2S-StaR, continuous	$0.928_{\pm.017}$	$0.905_{\pm.016}$	$0.987_{\pm.001}$	$0.929_{\pm.010}$	$0.937_{\pm.009}$	$+1.8$ (.001)
2S-StaR, thresholded	$0.922_{\pm.016}$	$0.883_{\pm.034}$	$0.982_{\pm.005}$	$0.919_{\pm.017}$	$0.926_{\pm.014}$	$+0.7$ (.265)

Appendix B.5), whereas a classifier-only L2 penalty reduces the across-seed spread at the baseline mean (Table 2, Appendix B.5).

5.5 Training schedule on the two backbones

ResNet uses the single-stage schedule, with StaR trained from the ImageNet initialization. CLIP uses the two-stage schedule, a baseline warm start followed by StaR fine-tuning. Two-stage ResNet runs from the ERM seed-42 checkpoint are similar to the single-stage runs on Avg. (4) (Table 11, Appendix B.3), and single-stage CLIP runs (λ=2) average slightly below the baseline on Avg. (6) (Table 1).

6 Discussion

Practical guidance for k, λ, and t. The reported runs use k=5 capped at $K-1$, t=0.9 in the thresholded form, and λ=15 for ResNet and λ=19 for CLIP (§5). In the CLIP k and t sweeps, all settings lie within 1.2 points of the default (Table 16), and the λ sweep of the thresholded form shows no detectable differences across $\lambda \in [5, 50]$ (Table 17). For a new backbone we suggest $k = \min(5, K-2)$, which excludes the rank-boundary direction of the $(K-1)$-dimensional LDA subspace. The ResNet runs reported here used $k = K-1 = 4$. We recommend the continuous form on CLIP and the thresholded form on ResNet, matching the configurations of our main results. We advise tuning λ on held-out data, since its value differed between our backbones while k, t, and the EMA momentum did not.

Checkpoint selection. CLIP checkpoints are selected by the best in-domain validation AUC, for comparability with prior work. On the single-stage CLIP runs traced at every epoch, a deployable early-stopping rule, the first epoch with in-domain AUC of at least 0.99, performs no worse on Synthbuster, and part of the apparent λ effect in single-stage sweeps is epoch-selection variance (Appendix C.6, Table 18).

Limitations. (i) StaR requires generator labels at training time only, as in Group DRO [26], and the online LDA recomputation adds $K+1$ thin SVDs every ten batches. Coarse family labels retain most of the gain, while k-means pseudo-labels that split generators across clusters do not recover it (Table 4). (ii) The gain differs in size and form across backbones. On ResNet the result is a reduction in the across-seed spread with a mean difference from the baseline that is not significant, and the continuous form reduces the spread much less. The CLIP gain is significant from one baseline warm start. From three further warm starts the continuous form is above its own warm start on two of the three and shows smaller, non-significant differences from the eight-run baseline, and the thresholded form is at or below the baseline on them (Table 6, Appendix C.1). (iii) On both backbones a classifier-only L2 penalty at one tested strength reproduces much of the observed effect. On ResNet it reduces the across-seed SD by two- to three-fold at the baseline mean, compared with about four-fold for StaR, and StaR is slightly above these controls (Table 2). On CLIP it matches the continuous form's mean on the same warm-start continuation, and the continuation without the penalty already accounts for part of the gain (Table 6). (iv) The per-epoch study (Appendix C.6)

Table 6: **Warm starts and continuation controls on CLIP** (2S, λ=19, Avg. (4), runs in parentheses). Upper rows: the continuous form from four CLIP-ERM warm starts. Middle rows: the thresholded form from the same warm starts. Lower rows: continuations of the seed-42 warm start with the penalty off, with and without a classifier-only L2 penalty. Δ_{ws} and Δ are the differences from the row's warm start and from the eight-run CLIP-ERM mean, respectively, with the two-sided Welch p reported for Δ (Table 15).

Method (n)	Warm start	Avg. (4)	Δ_{ws}	Δ vs CLIP-ERM (p)
CLIP-ERM, eight independent runs (8)	–	$0.919_{\pm.009}$	–	–
Continuous, from CLIP-ERM seed 42 (8)	0.927	$0.937_{\pm.009}$	+1.0	+1.8 (.001)
Continuous, from CLIP-ERM seed 1 (6)	0.928	$0.927_{\pm.007}$	−0.2	+0.8 (.108)
Continuous, from CLIP-ERM seed 2 (6)	0.902	$0.924_{\pm.010}$	+2.2	+0.5 (.391)
Continuous, from CLIP-ERM seed 3 (6)	0.914	$0.922_{\pm.014}$	+0.8	+0.3 (.655)
Thresholded, from CLIP-ERM seed 42 (8)	0.927	$0.926_{\pm.014}$	−0.1	+0.7 (.265)
Thresholded, from CLIP-ERM seed 1 (6)	0.928	$0.912_{\pm.012}$	−1.6	−0.7 (.261)
Thresholded, from CLIP-ERM seed 2 (6)	0.902	$0.913_{\pm.016}$	+1.0	−0.7 (.371)
Thresholded, from CLIP-ERM seed 3 (6)	0.914	$0.919_{\pm.017}$	+0.5	−0.1 (.943)
Continuation of seed 42, λ=0 (8)	0.927	$0.928_{\pm.013}$	+0.1	+0.9 (.138)
Continuation of seed 42 + classifier-only L2 10^{-2} (8)	0.927	$0.936_{\pm.014}$	+0.8	+1.6 (.018)

shows an active LDA subspace of effective rank 1–3, but this does not determine the number of directions worth penalizing. Our ResNet runs used k=$K-1$ whereas we suggest $K-2$, a setting we did not test on ResNet. (v) EFFORT's leave-one-out protocol is not part of our evaluation, and we do not claim to beat EFFORT in JPEG robustness. On CLIP, EFFORT loses less AUC than StaR between the released images and the heaviest compression tested on Synthbuster and COCOAI, where it scores above both StaR forms at that compression, while the continuous form stays above EFFORT at every quality factor on WildRF and LDM-FakeDetect (Appendix C.7). (vi) λ was selected on the reported OOD benchmarks, without a separate validation split. Selecting on Chameleon instead, which was not used for selection, would pick λ=50, within the range of the sweep, with a similar Avg. (4) (Table 17).

7 Conclusion

Generator-discriminative feature directions in AI-generated image detectors carry both cross-generator detection signal and generator-specific shortcuts. StaR scores each LDA direction by its leave-one-generator-out stability and penalizes the classifier along the unstable directions during training. This shifts the classifier weight out of the unstable component while the generator-discriminative structure of the features concentrates on one direction. StaR reduces the across-seed spread of Avg. (4) on ResNet by about four-fold and, in the continuous form, improves Avg. (4) over the baseline and over EFFORT on GenImage-trained CLIP. A classifier-only L2 penalty reproduces part of both effects (§6). Post-hoc projection of the same directions recovers the ResNet mean but neither the ResNet across-seed spread nor the CLIP gain.

Acknowledgments and Disclosure of Funding

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. RS-2025-02263841, Development of a Realtime Multimodal Framework for Comprehensive Deepfake Detection Incorporating Common Sense Error Analysis, and No. RS-2025-02215344, Development of AI Technology with Robust and Flexible Resilience Against Risk Factors) and by the Artificial Intelligence Industrial Convergence Cluster Development Project funded by the Ministry of Science and ICT (MSIT, Korea) and Gwangju Metropolitan City.

References

[1] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Q. Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024.

[3] N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff, and S. Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 66044–66063, 2023.

[4] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

[5] B. Cavia, E. Horwitz, T. Reiss, and Y. Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024.

[6] L. Chai, D. Bau, S.-N. Lim, and P. Isola. What makes fake images detectable? Understanding properties that generalize. In *European Conference on Computer Vision (ECCV)*, pages 103–120, 2020.

[7] B. Chen, J. Zeng, J. Yang, and R. Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 7621–7639, 2024.

[8] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva. On the detection of synthetic images generated by diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[9] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[10] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936.

[11] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 3247–3258, 2020.

[12] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[13] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021.

[14] P. Kirichenko, P. Izmailov, and A. G. Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations (ICLR)*, 2023.

[15] C. Koutlis and S. Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision (ECCV)*, pages 394–411, 2024.

[16] B. Liu, Q. Qin, and Q. He. CausalCLIP: Causally-informed feature disentanglement and filtering for generalizable detection of generated images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(9):7069–7077, 2026.

[17] H. Liu, Z. Tan, C. Tan, Y. Wei, J. Wang, and Y. Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10770–10780, 2024.

[18] U. Ojha, Y. Li, and Y. J. Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24480–24489, 2023.

[19] J. Park and A. 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 (CVPR)*, pages 8245–8257, 2025.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021.

[21] A. S. Rajan and Y. J. Lee. Stay-Positive: A case for ignoring real image features in fake image detection. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 58008–58028, 2025.

[22] A. S. Rajan, U. Ojha, J. Schloesser, and Y. J. Lee. Aligned datasets improve detection of latent diffusion-generated images. In *International Conference on Learning Representations (ICLR)*, 2025.

[23] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7237–7256, 2020.

[24] J. Ricker, D. Lukovnikov, and A. Fischer. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9130–9140, 2024.

[25] R. Roy, A. Aziz, S. Bajpai, N. Imanpour, G. Singh, S. Biswas, K. Wanaskar, P. Patwa, S. Ghosh, S. Dixit, N. R. Pal, V. Rawte, R. Garimella, A. Das, A. Sheth, G. Jena, V. Sharma, A. N. Reganti, V. Jain, and A. Chadha. A comprehensive dataset for human vs. AI generated image detection. *arXiv preprint arXiv:2601.00553*, 2026.

[26] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

[27] C. Tan, H. Liu, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. 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 (CVPR)*, pages 28130–28139, 2024.

[28] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8695–8704, 2020.

[29] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22445–22455, 2023.

[30] S. Yan, O. Li, J. Cai, Y. Hao, X. Jiang, Y. Hu, and W. Xie. A sanity check for AI-generated image detection. In *International Conference on Learning Representations (ICLR)*, 2025.

[31] Z. Yan, J. Wang, P. Jin, K.-Y. Zhang, C. Liu, S. Chen, T. Yao, S. Ding, B. Wu, and L. Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 70268–70288, 2025.

[32] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

[33] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang. GenImage: A million-scale benchmark for detecting AI-generated image. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, pages 77771–77782, 2023.

A Dataset and training details

This appendix expands on the brief setup paragraphs of §5.

Training corpora. The Corvi-LDM corpus [8] contains five generators (`ldm_text2img`, `ldm_class2img`, `gd_ffhq`, `gd_bedrooms`, `gd_churches`). We train at native resolution with one tenth of the corpus held out for in-domain validation. The CLIP training corpus uses six GenImage [33] subsets (ADM, GLIDE, SD 1.4, Midjourney, VQDM, Wukong) plus ProGAN images from CNNDetection [28], $K{=}7$ generator classes. We train on half of this corpus and hold out the other half for in-domain validation. A class-balanced sampler makes real and fake images equally likely in every batch. The image counts per class are listed in the code release.

OOD benchmarks. Synthbuster [2] has nine generators (DALL-E 2/3, GLIDE, Firefly, SD 1.3/1.4/2/XL, Midjourney v5) over RAISE-1k real photos. None is in the Corvi-LDM corpus, while GLIDE, SD 1.4 and Midjourney share a generator name with the GenImage training corpus of the CLIP models (different image sets), so six of the nine Synthbuster generators are unseen by the CLIP models. WildRF [5] has real and fake images from Reddit, Facebook, and Twitter, with realistic JPEG and platform-specific post-processing and mixed generators. COCOAI [25] has five generators (SD 2.1, SDXL, SD 3, DALL-E 3, Midjourney, the last sharing its name with a CLIP training generator) over COCO real photos. We use the Hugging Face validation split. Its real images are COCO photographs, many of which also appear in the Corvi-LDM training corpus of our ResNet models. Excluding these overlapping real images lowers the ResNet SD ratio from $4.1\times$ to $3.9\times$ (one-sided F, $p = 0.001$) and does not change the conclusion. LDM-FakeDetect [22, 21] has nine latent-space generators and is treated as in-distribution-like for the Corvi-LDM training corpus and as OOD for the GenImage training corpus. Chameleon [30] is a mixed real and fake benchmark, and CommunityForensics (CompEval split) [19] spans 21 generators paired with real photos from RAISE, COCO, FFHQ, and LAION.

Hyperparameters. Table 7 lists the training and StaR settings for both backbones. All ResNet runs use the Corvi stage-2 augmentation pipeline (random resized crop, color jitter, grayscale, cutout, Gaussian noise and blur, JPEG recompression, rotation and horizontal flip), applied at the native resolution before the crop with each operation drawn independently. The two-stage ResNet runs of Table 11 use the same recipe but start from the ERM seed-42 checkpoint. For CLIP, the fc1 and fc2 weights of every ViT MLP block remain frozen (about two thirds of the vision parameters) in every model, CLIP-ERM and StaR alike. The EFFORT runs use the same transform for training and evaluation and are evaluated with EFFORT's SVD-residual parameters loaded. The requested k is capped at $K{-}1$, the maximal rank of the LDA subspace, and real images are excluded from the LDA centroids. The LDA eigenvectors and stability scores are recomputed every ten batches and held fixed between updates, while the centroids are updated every batch by EMA. For CLIP, λ was chosen on a grid with eight runs per value (Appendix C.4) and the continuous form uses the same value. For ResNet, λ comes from a single-seed sweep with a standard ResNet-50 trained on GenImage and is used unchanged, and a four-seed check on Corvi-LDM (Table 7) confirms it. The augmentation parameters, the sweep grids and the per-epoch recompute counts are listed in the code release.

Compute. Each run uses one NVIDIA H100 80GB GPU. A ResNet run of any method takes about 12 GPU-hours, and a CLIP 2S-StaR run about 1.7 GPU-hours plus a CLIP-ERM warm start of similar cost. The reported runs total about $2{,}100$ GPU-hours and all training runs of the project about $2{,}700$.

B ResNet results on Corvi-LDM

This appendix gathers the ResNet results cited in §5. All runs use the Corvi-LDM corpus and the recipe of Appendix A.

B.1 Seven benchmarks and post-hoc comparison

This section compares training-time StaR with four post-hoc methods that operate on frozen features using the same LDA basis and leave-one-generator-out stability scores (TSLDA thresholds at $s_i < 0.5$). We evaluate them on ResNet (eight ERM checkpoints, final epoch, Table 8) and on CLIP

Table 7: **Training and StaR settings** of the two backbones. The upper block is the training recipe, the lower block the StaR defaults shared by both forms. The λ check reports each backbone's Avg. (4) at the reported λ and two neighbors, mean$\pm$SD over four seeds for ResNet and over the eight runs of Table 17 for CLIP.

	ResNet (Corvi-LDM)	CLIP (GenImage)
Backbone	ResNet-50-NoDown, ImageNet initialization	CLIP ViT-L/14, MLP fc1 and fc2 frozen
Optimizer	Adam (β_1=0.9, β_2=0.999)	AdamW
Learning rate	10^{-4}	2×10^{-4} (classifier), $0.1\times$ for the backbone
Weight decay	0	5×10^{-4}
Epochs	20	20
Batch size	64	32
Input	random 200×200 crop at native resolution	Resize-224, CenterCrop-224
Augmentation	Corvi stage-2 pipeline	horizontal flip
Schedule	single-stage from ImageNet weights	two-stage from the CLIP-ERM seed-42 run
k (requested 5, capped at $K-1$)	4	5
EMA momentum α	0.99	0.99
Recompute interval Δ	10 batches	10 batches
Threshold t (thresholded form)	0.9	0.9
λ	15	19
λ selection	single-seed sweep, standard ResNet-50 on GenImage	grid at eight runs per value (Table 17)
λ check, neighboring λ	10 / 15 / 25, four seeds	10 / 19 / 30, eight runs
λ check, Avg. (4)	$0.834_{\pm.009}$ / $0.843_{\pm.003}$ / $0.839_{\pm.002}$	$0.929_{\pm.013}$ / $0.926_{\pm.014}$ / $0.928_{\pm.018}$

(eight CLIP-ERM runs, Table 5 in §5.3). The comparison holds the backbone, the LDA directions and the stability scores fixed and varies only the mechanism. LDA, SLDA and TSLDA ($s_i < 0.5$) project the classifier weight out of generator-discriminative directions, and the LEACE-style row removes the LDA-spanned subspace from the features by a hard orthogonal projection (rather than the covariance-aware LEACE map of Belrose et al. [3]) and refits a linear probe on training data. Training-time StaR (1S-StaR on ResNet, 2S-StaR on CLIP) changes the features during training. Because the post-hoc rows act on the same checkpoints as their ERM row, their differences to ERM are tested on the within-checkpoint differences across the eight checkpoints. The training-time rows are independent runs and use Welch tests.

Table 8: **ResNet on seven benchmarks** (Corvi-LDM, final checkpoint, eight runs per row, mean$\pm$SD). Post-hoc rows act on the frozen features and classifier of the eight ERM checkpoints with the four methods of Appendix B.1. Lower panel: Avg. (4) averages Synthbuster, WildRF, COCOAI and GenImage, Avg. (6) every benchmark except GenImage. Δ: difference from ERM in points, with paired tests over the shared checkpoints for the post-hoc rows and Welch tests for the independently trained rows. SD ratio: ERM SD divided by the row SD, with the one-sided F p. Both are given for Avg. (4) and for Avg. (6). **Bold**: best in column.

Method	Synth.	WildRF	COCOAI	GenImage	LDM-FD	Chameleon	CommFor
Post-hoc, on the frozen features of the eight ERM checkpoints							
ERM (frozen features)	$0.891_{\pm.011}$	$0.782_{\pm.018}$	$0.837_{\pm.010}$	$0.830_{\pm.015}$	$0.953_{\pm.018}$	$0.725_{\pm.023}$	$0.867_{\pm.024}$
LDA projection	$0.865_{\pm.015}$	$0.737_{\pm.021}$	$0.825_{\pm.009}$	$0.807_{\pm.012}$	$0.881_{\pm.037}$	$0.700_{\pm.019}$	$0.799_{\pm.029}$
SLDA	$0.895_{\pm.009}$	$0.794_{\pm.011}$	$0.839_{\pm.008}$	$0.835_{\pm.012}$	$0.962_{\pm.015}$	$0.731_{\pm.021}$	$0.881_{\pm.019}$
TSLDA ($s_i < 0.5$)	$\mathbf{0.898_{\pm.008}}$	$\mathbf{0.803_{\pm.012}}$	$\mathbf{0.840_{\pm.008}}$	$0.839_{\pm.011}$	$0.968_{\pm.015}$	$0.733_{\pm.020}$	$0.894_{\pm.019}$
LEACE-style	$0.889_{\pm.009}$	$0.786_{\pm.012}$	$\mathbf{0.840_{\pm.007}}$	$0.831_{\pm.008}$	$0.965_{\pm.010}$	$0.746_{\pm.012}$	$0.868_{\pm.008}$
Training-time, single-stage (eight seeds)							
1S-StaR, thresholded	$0.889_{\pm.009}$	$0.801_{\pm.004}$	$0.835_{\pm.006}$	$\mathbf{0.842_{\pm.006}}$	$\mathbf{0.978_{\pm.006}}$	$\mathbf{0.748_{\pm.017}}$	$\mathbf{0.900_{\pm.012}}$
1S-StaR, continuous	$0.890_{\pm.007}$	$0.787_{\pm.020}$	$0.836_{\pm.008}$	$0.829_{\pm.013}$	$0.953_{\pm.022}$	$0.722_{\pm.030}$	$0.878_{\pm.022}$

Method	Avg. (4)	Δ vs ERM (p)	SD ratio (p)	Avg. (6)	Δ vs ERM (p)	SD ratio (p)
Post-hoc, on the frozen features of the eight ERM checkpoints						
ERM (frozen features)	$0.835_{\pm.012}$	ref.	ref.	$0.842_{\pm.015}$	ref.	ref.
LDA projection	$0.809_{\pm.013}$	-2.6 (<.001)	$0.9\times$ (.594)	$0.801_{\pm.020}$	-4.1 (<.001)	$0.8\times$ (.766)
SLDA	$0.841_{\pm.008}$	$+0.6$ (.011)	$1.4\times$ (.175)	$0.850_{\pm.011}$	$+0.8$ (.014)	$1.4\times$ (.216)
TSLDA ($s_i < 0.5$)	$\mathbf{0.845_{\pm.007}}$	$+1.0$ (.019)	$1.8\times$ (.078)	$0.856_{\pm.010}$	$+1.4$ (.024)	$1.5\times$ (.164)
LEACE-style	$0.836_{\pm.007}$	$+0.1$ (.636)	$1.7\times$ (.099)	$0.849_{\pm.006}$	$+0.6$ (.213)	$2.7\times$ (.009)
Training-time, single-stage (eight seeds)						
1S-StaR, thresholded	$0.842_{\pm.003}$	$+0.7$ (.143)	$4.1\times$ (<.001)	$\mathbf{0.859_{\pm.006}}$	$+1.6$ (.020)	$2.5\times$ (.013)
1S-StaR, continuous	$0.836_{\pm.007}$	$+0.1$ (.911)	$1.6\times$ (.119)	$0.844_{\pm.015}$	$+0.2$ (.791)	$1.0\times$ (.498)

Projecting the classifier weight off all LDA directions (LDA) lowers Avg. (4) relative to ERM, consistent with the entanglement, since removing generator-discriminative directions from a frozen feature space discards detection signal alongside the shortcut. The stability-filtered projections raise the mean, with both SLDA and TSLDA significantly above ERM, while LEACE-style erasure is similar to ERM. TSLDA reaches the mean of training-time 1S-StaR, but no post-hoc row reproduces its across-seed spread of Avg. (4).

Per-benchmark variance tests. Table 9 reports the one-sided F-test of the eight ERM seeds against the eight thresholded 1S-StaR seeds on every benchmark and on Avg. (4). The reduction in across-seed SD is significant on WildRF, GenImage, LDM-FakeDetect and CommunityForensics and is not significant on Synthbuster, COCOAI or Chameleon. The classifier-only L2 controls, which reduce the across-seed SD at the ERM mean, are reported in Appendix B.5 (Table 2).

Table 9: **Across-seed variance tests on ResNet** (Corvi-LDM, eight ERM seeds against eight thresholded 1S-StaR seeds, final checkpoint). One-sided F-statistic (ERM variance over StaR variance, df_1=df_2=7) per benchmark and on Avg. (4).

Benchmark	F	p
Synthbuster	1.4	.323
WildRF	15.8	<.001
COCOAI	2.5	.129
GenImage	5.3	.021
LDM-FakeDetect	9.9	.004
Chameleon	1.7	.249
CommunityForensics	4.0	.043
Avg. (4)	17.0	<.001

Chameleon, CommunityForensics and Avg. (6). Table 8 also reports Chameleon and CommunityForensics for every row, and Avg. (6) averages Synthbuster, WildRF, COCOAI, LDM-FakeDetect, Chameleon and CommunityForensics. On Avg. (6) the thresholded form has a smaller spread and a higher mean than ERM, both significantly, while the continuous form stays near the ERM level. On Chameleon and CommunityForensics LDA projection lowers both, SLDA and TSLDA raise both slightly, and LEACE-style erasure raises Chameleon and leaves CommunityForensics at the ERM level. LEACE-style erasure has a smaller across-checkpoint SD than 1S-StaR on these two benchmarks, so on Avg. (6) the two have similar spreads, with 1S-StaR higher (Welch p=0.005). The spread reduction that separates training-time StaR from post-hoc erasure is therefore on the Avg. (4) benchmarks and LDM-FakeDetect, not on Chameleon or CommunityForensics.

B.2 Direction ablations

Table 10 repeats the direction ablations on the single-stage ResNet (λ=15, k=5, t=0.9, four seeds per row, with the same recipe except for the ablated component). StaR has the highest Avg. (4). Removing stability weighting, replacing the LDA directions with perpendicular or random ones, or penalizing the stable directions instead returns the average to the ERM level. The continuous form also stays near the ERM level on this backbone, which is why we use the thresholded form for ResNet. The losses are largest on WildRF and GenImage.

B.3 Single-stage vs. two-stage schedule

Table 11 lists the single-stage ResNet runs next to two-stage ResNet runs that start from the ERM seed-42 checkpoint under the matched recipe (crop 200, batch 64, 20 epochs). The matched two-stage runs reduce the across-seed SD less than the single-stage runs and are similar to them on Avg. (4), not significantly different (Welch p=0.12).

B.4 Rotation, not erasure

Figure 4 shows two feature-geometry indicators for the seed-42 ERM and 1S-StaR models trained on Corvi-LDM, the per-class generator-probe accuracy (Figure 4a) and the LDA singular values

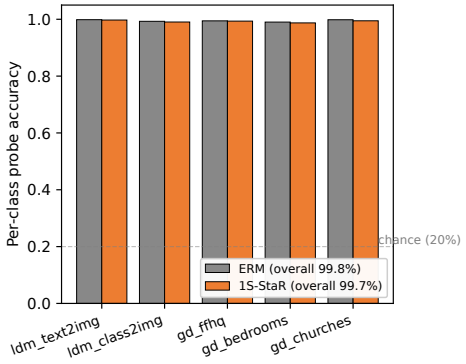
Table 10: Direction ablations on ResNet (Corvi-LDM, single-stage, λ=15, k=5 capped at four, t=0.9, four seeds per row). Δ is against the baseline, in points. **Bold**: best in column.

Variant	Synth.	WildRF	COCOAI	GenImage	Avg. (4)	Δ
Baseline (no penalty)	0.891	0.780	0.837	0.829	$0.834_{\pm.008}$	ref.
StaR (full method)	0.890	**0.804**	0.836	**0.840**	$\mathbf{0.843_{\pm.003}}$	+0.8
StaR, continuous (no threshold)	0.889	0.776	0.834	0.825	$0.831_{\pm.006}$	−0.3
Uniform weights (no stability weighting)	0.893	0.783	**0.838**	0.821	$0.834_{\pm.004}$	0.0
Random directions + uniform	**0.894**	0.780	0.833	0.823	$0.833_{\pm.007}$	−0.2
Perpendicular-to-LDA + uniform	0.889	0.774	0.836	0.834	$0.833_{\pm.005}$	−0.1
Anti-StaR (penalize stable directions)	0.885	0.781	0.832	0.820	$0.830_{\pm.008}$	−0.4

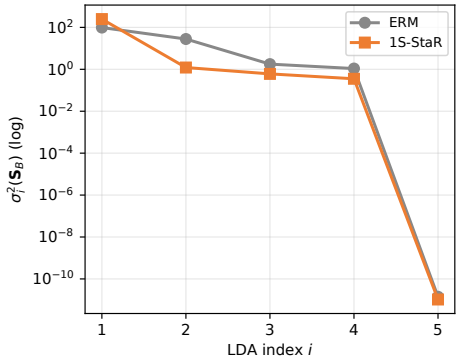
Table 11: **Single-stage vs. two-stage StaR on ResNet** (Corvi-LDM, thresholded, λ=15, t=0.9, final checkpoint). The two-stage runs start from the ERM seed-42 checkpoint under the single-stage recipe. Avg. (6) as in Table 8. SD ratio: ERM SD of Avg. (4) over the row's, with the one-sided F p.

Method (n)	Synth.	WildRF	COCOAI	GenImage	Avg. (4)	Avg. (6)	SD ratio (p)
ERM (8)	$0.891_{\pm.011}$	$0.782_{\pm.018}$	$0.837_{\pm.010}$	$0.830_{\pm.015}$	$0.835_{\pm.012}$	$0.842_{\pm.015}$	ref.
1S-StaR, thresholded (8)	$0.889_{\pm.009}$	$0.801_{\pm.004}$	$0.835_{\pm.006}$	$0.842_{\pm.006}$	$0.842_{\pm.003}$	$0.859_{\pm.006}$	4.1× (<.001)
2S-StaR, matched recipe, ERM warm start (8)	$0.882_{\pm.010}$	$0.795_{\pm.011}$	$0.835_{\pm.008}$	$0.843_{\pm.013}$	$0.839_{\pm.005}$	$0.856_{\pm.006}$	2.5× (.014)

(Figure 4b). Throughout the paper σ_i denotes the singular values of $\mathbf{S}_B^{1/2}$ ($\eta_i = \sigma_i^2$ in Appendix D), and λ is reserved for the StaR penalty coefficient.



(a) Per-class probe accuracy on the five training generators (chance 20%).



(b) Eigenvalues σ_i^2 of the LDA between-class scatter $\mathbf{S}_B$ (log scale, index 5 is the null direction of the five-centroid LDA).

Figure 4: **Mechanism: rotation, not erasure** (Corvi-LDM, seed 42, single seed pair). Generator information remains linearly decodable and the between-class scatter concentrates on one direction, while the classifier weight leaves the LDA subspace (Table 12).

B.5 Classifier-only L2 controls

To test whether StaR's effect is generic weight regularization, we trained ERM with the 1S-StaR recipe (Corvi-LDM corpus, NoDown architecture, crop 200, batch size 64, 20 epochs) with an L2 penalty on the classifier weight vector alone at 10^{-3} and 10^{-2}, eight seeds each (seeds 42 and 1 to 7), and compared them with the eight-seed ERM and 1S-StaR runs (Table 2 in §5.2). Weight decay on all network parameters (Adam with coupled L2) at six values from 10^{-4} to 10^{-1}, using a single seed, lowered Avg. (4) at every value, possibly because it also shrinks the ImageNet-pretrained backbone, and reduced the model to chance at 10^{-1}. The CLIP counterpart, a classifier-only L2 penalty during the λ=0 continuation from the CLIP-ERM warm start, is reported in Table 15 in Appendix C.1.

Table 12: Classifier-weight geometry relative to the Corvi-LDM LDA subspace (ResNet, thresholded 1S-StaR vs. ERM, final checkpoint, mean over seeds). Unstable directions are the fixed set v_1–v_3 in both models, the three directions below the threshold on the eight-seed ERM mean. Reduction is the ratio of the ERM mean to the StaR mean.

Quantity	ERM (n=8)	1S-StaR, thresholded (n=8)	Reduction
Mass on unstable directions v_1–v_3	6.9%	0.014%	502×
Total LDA-subspace projection	33.1%	1.02%	32×
Unstable / stable mass ratio	0.28	0.022	13×

Both controls reduce the across-seed SD significantly at the ERM mean, by less than 1S-StaR (Table 2). 1S-StaR is above both controls, significantly compared with the 10^{-3} control but not relative to the 10^{-2} control.

B.6 Per-direction k-sweeps behind the diagnosis

This section provides the per-direction evidence behind the diagnosis of §3, computed on the models reported in the paper. PRE denotes the eight ERM checkpoints (seeds 42 and 1 to 7) and POST the eight 1S-StaR checkpoints (thresholded form, λ=15, seeds 42 and 1 to 7, final checkpoint), all trained on Corvi-LDM. Every value is reported as a mean±SD over the eight checkpoints, with the LDA basis and the stability scores recomputed per checkpoint from the training corpus (k_{eff}=4).

PRE (ERM, eight seeds). For each OOD benchmark and each $k \in \{1, \ldots, 4\}$ (the LDA basis on K=5 generators has four signal-bearing directions), we project features onto k LDA directions and train a logistic-regression real-vs-fake probe on the projection (Tables 13 and 14). "Top-k stable" uses the k highest-stability directions, "Bot-k unstable" uses the k lowest-stability directions, and "Random k" the mean±SD over five random subsets of size k. Over the eight PRE checkpoints the stability scores are those of Figure 3. The single-direction probe AUC increases from the most stable to the third most stable direction on Synthbuster and COCOAI, so the most stable direction carries the least signal on these two benchmarks (most stable against least stable direction, t-test on the eight within-checkpoint differences, p=0.002 and p<0.001). On WildRF and LDM-FakeDetect the four directions are similar (p=0.76 and p=0.48). With three directions (Table 13, k=3) the top-stable set is at or above the bottom-unstable set on every benchmark.

Table 13: Probe AUC on k LDA directions of the ERM checkpoints (PRE, ResNet, Corvi-LDM, mean±SD over eight seeds). For each checkpoint the LDA basis and the stability scores are recomputed from the training corpus (k_{eff}=4). For each benchmark a logistic real-vs-fake probe is fit on features projected onto k directions: the k most stable, the k least stable, or random k-subsets (mean over five). Full: all four directions.

		$k=1$			$k=3$		
Benchmark	full	top-stable	bot-unstable	random	top-stable	bot-unstable	random
Synthbuster	$0.993_{\pm.001}$	$0.594_{\pm.051}$	$0.806_{\pm.102}$	$0.713_{\pm.080}$	$0.893_{\pm.014}$	$0.882_{\pm.026}$	$0.867_{\pm.025}$
WildRF	$0.983_{\pm.001}$	$0.647_{\pm.017}$	$0.641_{\pm.054}$	$0.619_{\pm.043}$	$0.786_{\pm.011}$	$0.766_{\pm.032}$	$0.763_{\pm.042}$
COCOAI	$0.993_{\pm.001}$	$0.598_{\pm.037}$	$0.780_{\pm.065}$	$0.746_{\pm.029}$	$0.830_{\pm.015}$	$0.808_{\pm.030}$	$0.808_{\pm.016}$
LDM-FakeDetect	$1.000_{\pm.000}$	$0.812_{\pm.048}$	$0.825_{\pm.054}$	$0.742_{\pm.081}$	$0.983_{\pm.009}$	$0.935_{\pm.045}$	$0.943_{\pm.029}$

POST (1S-StaR, eight seeds). Table 14 repeats the sweep on the eight 1S-StaR checkpoints. After StaR training the most stable direction scores close to one and the other three keep their ERM-level scores. Relative to PRE, the probe AUC of the most stable direction rises on Synthbuster, WildRF and LDM-FakeDetect and is unchanged on COCOAI, where it remains the direction with the least signal (p<0.001 compared with the least stable direction). The other three directions and the full four-direction probe move little on every benchmark. With three directions the top-stable set is again at or above the bottom-unstable set on every benchmark.

Table 14: As Table 13, on the eight thresholded 1S-StaR checkpoints (POST, λ=15).

		$k=1$			$k=3$		
Benchmark	full	top-stable	bot-unstable	random	top-stable	bot-unstable	random
Synthbuster	$0.991_{\pm.001}$	$0.745_{\pm.046}$	$0.819_{\pm.037}$	$0.766_{\pm.044}$	$0.893_{\pm.009}$	$0.874_{\pm.011}$	$0.880_{\pm.014}$
WildRF	$0.978_{\pm.004}$	$0.696_{\pm.040}$	$0.683_{\pm.036}$	$0.681_{\pm.024}$	$0.815_{\pm.010}$	$0.782_{\pm.040}$	$0.806_{\pm.015}$
COCOAI	$0.992_{\pm.002}$	$0.592_{\pm.091}$	$0.791_{\pm.022}$	$0.757_{\pm.028}$	$0.843_{\pm.008}$	$0.842_{\pm.015}$	$0.835_{\pm.008}$
LDM-FakeDetect	$1.000_{\pm.000}$	$0.890_{\pm.077}$	$0.901_{\pm.051}$	$0.841_{\pm.049}$	$0.986_{\pm.004}$	$0.956_{\pm.029}$	$0.969_{\pm.013}$

C CLIP results on GenImage

This appendix gathers the CLIP results beyond Tables 1, 3 and 5.

C.1 Warm-start dependence of the two-stage comparison

The eight two-stage StaR runs of each form in Table 1 branch from one warm start, the CLIP-ERM run with seed 42, and vary the stage-2 seed. To test whether the comparison depends on that warm start, we branched stage-2 runs from three further warm starts, A to C, the CLIP-ERM runs with seeds 1 to 3 (six runs of each form from each). Table 6 reports the per-warm-start means. Under the continuous form every warm start lies above the eight-run CLIP-ERM mean, by less than the seed-42 warm start does, and warm start A ends below its own CLIP-ERM run. Pooled over A to C, the continuous form is not significantly above CLIP-ERM (Welch p=0.23 over 18 runs, which share three warm starts, so the test overstates the independent sample). Under the thresholded form the three further warm starts lie at or below the CLIP-ERM mean, so the thresholded difference of Table 1 does not replicate from other warm starts. The CLIP gain is therefore a continuous-form result whose size depends on the warm start.

Continuation without the penalty. Continuing the seed-42 warm start for the same number of epochs without the penalty (λ=0, eight runs) gives the continuation row of Table 15. The continuous-form gain over CLIP-ERM decomposes into the advantage of the seed-42 warm start over the eight-run mean, the continuation itself, which adds little, and the penalty, which accounts for about half (Table 6). The continuous form is above the continuation on Avg. (4), although not significantly (p=0.13), and significantly above it on WildRF (p=0.039). The thresholded form is at the continuation's level (p=0.78).

Classifier-only L2 during the continuation. Table 15 adds an L2 penalty on the classifier weight vector to the same continuation (seed-42 warm start, λ=0, same number of epochs). At 10^{-2} the control matches the mean of the continuous form (Welch p=0.76), while 10^{-3} and 10^{-1} stay close to CLIP-ERM. The continuous form's gain over CLIP-ERM is therefore matched by an L2 penalty on the classifier alone at one tuned strength. The continuous form has a smaller across-run spread than the control (one-sided F, p=0.11). The thresholded form lies below the 10^{-2} control.

Table 15: Classifier-only L2 control on CLIP. Continuation rows start from the CLIP-ERM seed-42 warm start and train for the stage-2 schedule with the StaR penalty off (λ=0), optionally with an L2 penalty on the classifier weight vector. n in parentheses. Δ is the Avg. (4) difference from the eight-run CLIP-ERM mean, in points, with the two-sided Welch p.

Method (n)	Synth.	WildRF	COCOAI	LDM-FD	Avg. (4)	Δ vs CLIP-ERM (p)
CLIP-ERM (independent runs) (8)	$0.913_{\pm.015}$	$0.858_{\pm.025}$	$0.984_{\pm.004}$	$0.922_{\pm.014}$	$0.919_{\pm.009}$	ref.
Continuation, λ=0 (no penalty) (8)	$0.926_{\pm.015}$	$0.884_{\pm.020}$	$0.985_{\pm.005}$	$0.917_{\pm.021}$	$0.928_{\pm.013}$	+0.9 (.138)
Continuation + classifier-only L2 10^{-3} (8)	$0.920_{\pm.019}$	$0.875_{\pm.025}$	$0.985_{\pm.004}$	$0.914_{\pm.016}$	$0.923_{\pm.010}$	+0.4 (.423)
Continuation + classifier-only L2 10^{-2} (8)	$0.933_{\pm.014}$	$0.893_{\pm.031}$	$0.987_{\pm.004}$	$0.929_{\pm.013}$	$0.936_{\pm.014}$	+1.6 (.018)
Continuation + classifier-only L2 10^{-1} (8)	$0.922_{\pm.016}$	$0.885_{\pm.028}$	$0.982_{\pm.008}$	$0.911_{\pm.044}$	$0.925_{\pm.019}$	+0.6 (.467)
2S-StaR, continuous (8)	$0.928_{\pm.017}$	$0.905_{\pm.016}$	$0.987_{\pm.001}$	$0.929_{\pm.010}$	$0.937_{\pm.009}$	+1.8 (.001)
2S-StaR, thresholded (8)	$0.922_{\pm.016}$	$0.883_{\pm.034}$	$0.982_{\pm.005}$	$0.919_{\pm.017}$	$0.926_{\pm.014}$	+0.7 (.265)

C.2 Linear probe per layer depth

Figure 5 reports linear-probe accuracy for the generator label as a function of depth on the CLIP training corpus (K=7 generator classes), the measurement behind §3.

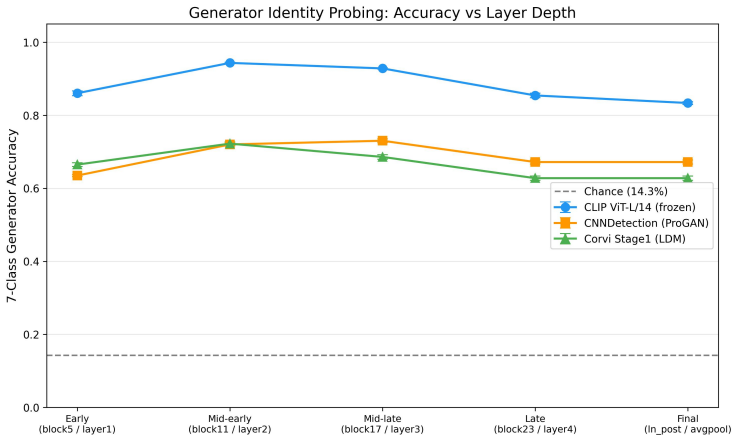


Figure 5: Linear-probe accuracy for the generator label at five depths of three frozen detectors (CLIP, the CNNDetection ResNet-50 and the Corvi stage-1 ResNet-50), on the CLIP training corpus (K=7 classes, chance about 14%).

C.3 k and t sweeps

Table 3 reports the four CLIP ablations. This section supplies the k and t sweeps.

Table 16 reports the k and t sweeps for the thresholded form with the Table 1 configuration as the default row. The k=6 row is the cap $K-1$ on the K=7 CLIP corpus. All settings are close to the default, and k=3 is the row that differs the most, but not significantly. At k=1 the single penalized direction is classified as stable at every epoch, so the penalty is inactive.

Table 16: CLIP k and t sweeps (thresholded 2S-StaR, λ=19, from the CLIP-ERM seed-42 warm start). k is the number of LDA directions the penalty considers, capped at $K-1=6$. t is the stability threshold of $\mathbf{1}[s_i < t]$. Δ and p compare each row with the default row (k=5, t=0.9) on Avg. (4), using a two-sided Welch test.

Variant	Synth.	WildRF	COCOAI	LDM-FD	Avg. (4)	Δ	p
k-sweep (t=0.9 fixed)							
k=1	$0.925_{\pm.018}$	$0.881_{\pm.028}$	$0.984_{\pm.005}$	$0.921_{\pm.021}$	$0.928_{\pm.013}$	+0.1	.832
k=3	$0.902_{\pm.037}$	$0.865_{\pm.047}$	$0.981_{\pm.003}$	$0.911_{\pm.023}$	$0.915_{\pm.025}$	−1.2	.270
k=5 (default)	$0.922_{\pm.016}$	$0.883_{\pm.034}$	$0.982_{\pm.005}$	$0.919_{\pm.017}$	$0.926_{\pm.014}$	ref.	
k=6 ($K-1$)	$0.907_{\pm.032}$	$0.896_{\pm.023}$	$0.980_{\pm.006}$	$0.915_{\pm.023}$	$0.925_{\pm.018}$	−0.2	.824
t-sweep (k=5 fixed)							
t=0.5	$0.923_{\pm.016}$	$0.890_{\pm.020}$	$0.985_{\pm.004}$	$0.925_{\pm.020}$	$0.931_{\pm.009}$	+0.4	.472
t=0.7	$0.922_{\pm.028}$	$0.884_{\pm.034}$	$0.984_{\pm.004}$	$0.927_{\pm.011}$	$0.929_{\pm.015}$	+0.3	.721
t=0.9 (default)	$0.922_{\pm.016}$	$0.883_{\pm.034}$	$0.982_{\pm.005}$	$0.919_{\pm.017}$	$0.926_{\pm.014}$	ref.	

C.4 λ sweep

Table 17 reports the sweep over $\lambda \in \{5, 10, 15, 16, 17, 18, 19, 20, 30, 50\}$ for the thresholded form with t=0.9, k=5, stage 2 initialized from the CLIP-ERM seed-42 warm start, eight runs per value.

Summary. Avg. (4) shows no differences detectable at eight runs per value over λ from 5 to 50 (one-way ANOVA p=0.69). λ=19, the value used in the main tables, is not significantly above

Table 17: CLIP λ sweep (thresholded 2S-StaR, t=0.9, k=5, from the CLIP-ERM seed-42 warm start, n per row in parentheses). Δ is the Avg. (4) difference from the eight CLIP-ERM runs, in points, with the two-sided Welch p. λ was selected on these benchmarks. Chameleon was not used for selection.

λ (n)	Synth.	WildRF	COCOAI	LDM-FD	Avg. (4)	Δ vs ERM (p)	Chameleon
5 (8)	$0.917_{\pm.013}$	$0.878_{\pm.027}$	$0.980_{\pm.007}$	$0.897_{\pm.022}$	$0.918_{\pm.011}$	-0.1 (.820)	$0.806_{\pm.043}$
10 (8)	$0.921_{\pm.016}$	$0.892_{\pm.023}$	$0.983_{\pm.005}$	$0.920_{\pm.017}$	$0.929_{\pm.013}$	$+1.0$ (.106)	$0.821_{\pm.048}$
15 (8)	$0.912_{\pm.028}$	$0.892_{\pm.027}$	$0.980_{\pm.006}$	$0.918_{\pm.029}$	$0.926_{\pm.021}$	$+0.6$ (.469)	$0.783_{\pm.049}$
16 (8)	$0.906_{\pm.020}$	$0.878_{\pm.032}$	$0.981_{\pm.006}$	$0.924_{\pm.011}$	$0.922_{\pm.013}$	$+0.3$ (.644)	$0.807_{\pm.054}$
17 (8)	$0.924_{\pm.015}$	$0.887_{\pm.031}$	$0.981_{\pm.008}$	$0.919_{\pm.018}$	$0.928_{\pm.014}$	$+0.8$ (.181)	$0.787_{\pm.070}$
18 (8)	$0.915_{\pm.017}$	$0.885_{\pm.022}$	$0.979_{\pm.011}$	$0.914_{\pm.030}$	$0.923_{\pm.017}$	$+0.4$ (.562)	$0.791_{\pm.026}$
19 (8)	$0.922_{\pm.016}$	$0.883_{\pm.034}$	$0.982_{\pm.005}$	$0.919_{\pm.017}$	$0.926_{\pm.014}$	$+0.7$ (.265)	$0.795_{\pm.028}$
20 (8)	$0.890_{\pm.073}$	$0.861_{\pm.075}$	$0.979_{\pm.014}$	$0.909_{\pm.031}$	$0.910_{\pm.046}$	-1.0 (.584)	$0.804_{\pm.028}$
30 (8)	$0.914_{\pm.024}$	$0.880_{\pm.040}$	$0.986_{\pm.005}$	$0.932_{\pm.009}$	$0.928_{\pm.018}$	$+0.9$ (.255)	$0.797_{\pm.041}$
50 (8)	$0.924_{\pm.014}$	$0.884_{\pm.024}$	$0.983_{\pm.003}$	$0.928_{\pm.015}$	$0.930_{\pm.010}$	$+1.1$ (.051)	$0.824_{\pm.041}$
CLIP-ERM (8)	$0.913_{\pm.015}$	$0.858_{\pm.025}$	$0.984_{\pm.004}$	$0.922_{\pm.014}$	$0.919_{\pm.009}$	ref.	$0.820_{\pm.026}$
EFFORT (8)	$0.890_{\pm.018}$	$0.829_{\pm.025}$	$0.966_{\pm.014}$	$0.871_{\pm.023}$	$0.889_{\pm.015}$	-3.0 (<.001)	$0.638_{\pm.073}$

Table 18: **Epoch-selection rules** on 11 single-stage StaR CLIP runs of 20 epochs (Synthbuster OOD AUC, mean±SD across runs). Deployable rules are OOD-blind. The oracle epoch is an upper bound. Ties: runs where the rule's pick is within 10^{-3} of the oracle.

Selection rule	Mean OOD	Std	Gap to oracle	#ties oracle
Oracle (max over 20 epochs)	0.9255	0.009	0.0000	11/11
First epoch with in-dom. AUC ≥ 0.99	**0.9186**	0.014	-0.0070	**6/11**
Fixed ep_0 (after one epoch)	**0.9182**	0.015	-0.0074	**6/11**
$\arg\max$ in-domain validation (default)	0.9107	0.015	-0.0149	3/11
$\arg\max\ \sigma_1$	0.8999	0.019	-0.0256	1/11
$\arg\max\ \bar{s}_i$ (mean stability)	0.8927	0.026	-0.0329	1/11
Fixed ep_2	0.8938	0.022	-0.0317	1/11

CLIP-ERM, and the largest mean over the grid is at λ=50. Chameleon, which was not used for selection, is at or below the CLIP-ERM level at every λ except 10 and 50, so selecting by the highest Chameleon mean would also choose λ=50. WildRF contributes most clearly to the improvement over CLIP-ERM and lies above it at every λ. The threshold t and the number of penalized directions k matter less (Table 16), and the per-epoch study (Appendix C.6) shows that the active LDA subspace has effective rank 1–3 during training.

C.5 Weaker grouping signals

StaR needs generator labels at training time. Table 4 (§5.3) reports runs in which the seven generator labels of the CLIP corpus are replaced by weaker grouping signals (2S-StaR, λ=19, same warm start). Family labels group the generators into four coarse families (GAN, pixel diffusion, latent diffusion, commercial), and k-means pseudo-labels are computed on CLIP-ERM features without any generator label. Under the continuous form, family labels keep most of the generator-label gain and remain significantly above CLIP-ERM. Under the thresholded form, family labels are similar to generator labels. k-means pseudo-labels, which split generators across clusters, do not recover the gain and fall below CLIP-ERM.

C.6 Per-epoch trajectory analysis: k, λ, and epoch selection

To complement the two-stage λ sweep of Appendix C.4 with per-epoch evidence, we ran 11 long-train (20-epoch) 1S-StaR runs (thresholded form) on CLIP over the GenImage corpus at $\lambda \in \{2, 3, 4, 5, 6, 7, 8, 10, 12, 15\}$ (seed 42) plus a λ=2 run at seed 1, with all checkpoints saved. For each saved checkpoint we evaluated Synthbuster OOD AUC, giving a per-epoch trajectory for each configuration, and the online LDA log provides the singular values σ_i and the stability scores at every recomputation.

Effective LDA rank vs. the number of penalized directions. We track two quantities, (i) the *effective rank* of the active LDA subspace at training time (the number of σ_i with $\sigma_i > 0.5\,\sigma_1$ at each LDA recomputation), a structural measurement of the basis itself, and (ii) the number of penalized directions k, a hyperparameter whose effect is measured by the sweep of Table 16. Across the snapshots of the 11 1S-StaR CLIP runs, the effective rank is one or two in nearly every snapshot and never above five. In three two-stage runs at λ=19 it is one in almost every snapshot. The k sweep, in turn, shows no significant differences among k=3, 5 and 6 on Avg. (4) (Table 16). The active LDA basis is low-rank, and the choice of k within $\{3, \ldots, 6\}$ has no significant effect.

The λ-curve flattens at the oracle epoch. Synthbuster OOD AUC at the $\arg\max$-in-domain checkpoint varies by several points across $\lambda \in \{2, \ldots, 15\}$. With OOD-oracle epoch selection the span contracts by about a third, and all but one run then lie within a narrow band, so part of the apparent tuning effect is epoch-selection variance. The seed gap at fixed λ=2 is comparable to the entire λ-effect range at the oracle epoch.

For the single-stage configuration of Table 1, the epoch-0 rule and $\arg\max$-in-domain give similar Avg. (6) and Avg. (4) over the eight runs with per-epoch checkpoints. Epoch-0 checkpoints exist only for these eight runs, so all tables report the $\arg\max$-in-domain checkpoint for CLIP. Selecting per-OOD oracle epochs would inflate numbers above what a deployable, OOD-blind training procedure can deliver.

C.7 JPEG robustness against EFFORT

We evaluate CLIP-ERM, both forms of 2S-StaR (λ=19) and EFFORT on Synthbuster, WildRF, COCOAI and LDM-FakeDetect, each re-compressed at JPEG quality factor (QF) $\in \{50, 70, 90\}$, eight runs per method (Figure 6). All methods are evaluated on the same re-compressed images. WildRF is already platform-compressed, so its re-compression is a second compression.

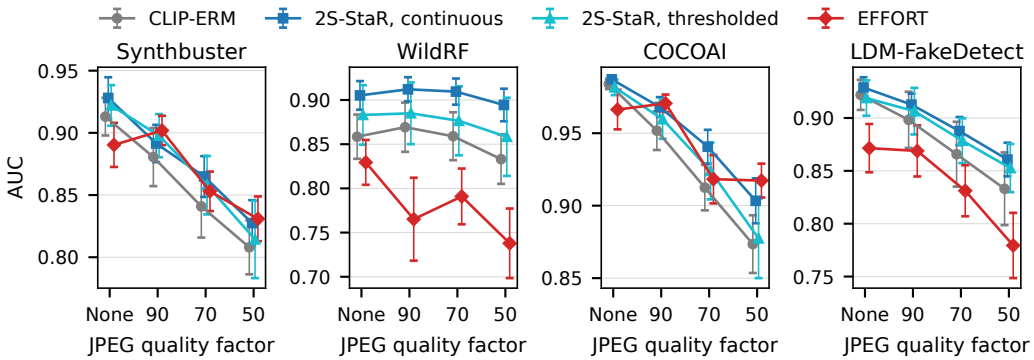


Figure 6: **JPEG robustness on four benchmarks** (CLIP, eight runs per method, mean±SD). "None" denotes the benchmark as released. The other points re-encode every image at the given JPEG quality factor (QF) after the resize and crop, so WildRF, which is already platform-compressed, is compressed twice.

Summary. At each tested JPEG quality factor (QF = 90, 70, and 50), both StaR forms achieve higher mean AUC than CLIP-ERM on all four benchmarks. On WildRF and LDM-FakeDetect the continuous form scores significantly above EFFORT at every quality factor, and EFFORT scores below CLIP-ERM throughout. On Synthbuster and COCOAI, EFFORT loses less AUC from the released images to QF=50 than the other three methods. On these two benchmarks, it scores below both StaR forms without compression, above both at QF=90 and QF=50, and below the continuous form at QF=70. The difference at QF=50 is significant against the thresholded form on COCOAI. Our claims concern generalization to unseen generators on the benchmarks as released, not compression robustness (§6).

C.8 Aggregation over held-out generators

The minimum in Eq. 1 lets a single held-out generator mark a direction as unstable. Table 19 compares it with the mean and the median of the same best-match alignments on the ERM checkpoints of both backbones. The scores there are recomputed offline from the features of 1,500 training images per generator, whereas the training-time penalty and Figure 3 use the online EMA estimate, so the seed-42 ResNet values differ from that seed's circles in Figure 3. For several directions one or two held-out generators set the minimum while the median stays near one. On CLIP v_1 and on ResNet v_2 have a minimum far below the threshold while their median stays near one (Table 19). The mean and the median average this dependence away, and it is the dependence on individual generators that the criterion targets, so the minimum is the conservative choice. On CLIP, mean aggregation leaves three directions above t=0.9, compared with one under the minimum, so it also weakens the thresholded penalty. Retraining the thresholded form with mean aggregation inside the regularizer (Table 20) does not improve performance. Avg. (4) is slightly lower, not significantly, and no benchmark differs significantly.

Table 19: **Aggregation over held-out generators.** Per-direction stability of the ERM checkpoints (seed 42), recomputed offline from 1,500 training images per generator, when the best-match alignment over the K leave-one-out bases is aggregated by the minimum (Eq. 1), the mean or the median. All $K-1$ directions are listed, one more than the k=5 the CLIP penalty uses. The last column marks the directions the thresholded form leaves unpenalized at t=0.9 under the minimum.

Direction	min	mean	median	$s_i \geq t$
CLIP-ERM, GenImage (K=7)				
v_0	0.99	1.00	1.00	yes
v_1	0.54	0.91	0.99	no
v_2	0.75	0.90	0.90	no
v_3	0.66	0.86	0.96	no
v_4	0.34	0.80	0.96	no
v_5	0.14	0.75	0.99	no
ResNet ERM, Corvi-LDM (K=5)				
v_0	0.86	0.95	0.97	no
v_1	0.37	0.83	0.99	no
v_2	0.11	0.72	0.99	no
v_3	0.13	0.64	0.86	no

Table 20: **Mean aggregation inside the regularizer.** Two-stage thresholded StaR (λ=19, t=0.9, k=5) with the minimum or the mean over held-out generators in the stability score, four matched seeds per row, mean±SD. The last row reports the two-sided Welch p for the comparison between aggregations.

Aggregation	Synth.	WildRF	COCOAI	LDM-FD	Avg. (4)
min (Eq. 1, the paper)	$0.915_{\pm.012}$	$0.890_{\pm.046}$	$0.980_{\pm.006}$	$0.913_{\pm.018}$	$0.925_{\pm.019}$
mean	$0.914_{\pm.047}$	$0.866_{\pm.021}$	$0.981_{\pm.003}$	$0.914_{\pm.015}$	$0.919_{\pm.018}$
p, min vs. mean	.955	.397	.809	.955	.667

D Why leave-one-generator-out stability separates shared and generator-specific directions

This appendix states the assumptions behind the stability score of Eq. 1, proves two one-sided bounds on it, and reports a synthetic study with known ground truth. The bounds say that a direction whose scatter is dominated by one generator cannot align strongly with a held-out direction of comparable eigenvalue, and that a direction whose eigengap is large relative to the centroid norms cannot score low. They do not cover tied eigenvalues or a generator with an outsized loading on a shared direction, and the synthetic study shows what happens there.

D.1 Setting

Let $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^D$ be the centroids of the K training generators (real images are not used) and $\mathbf{d}_g = \boldsymbol{\mu}_g - \frac{1}{K}\sum_h \boldsymbol{\mu}_h$ the centered centroids, so $\sum_g \mathbf{d}_g = \mathbf{0}$. The implementation takes the top right singular vectors of the $K \times D$ matrix with rows $\mathbf{d}_g$, which are the top eigenvectors of the centroid-only between-class scatter $\mathbf{S}_B = \sum_g \mathbf{d}_g \mathbf{d}_g^\top$. Write $(\eta_i, \mathbf{v}_i)$ for its eigenpairs with $\eta_1 \geq \cdots \geq \eta_{K-1} > 0 = \eta_K$. We assume the K centroids are affinely independent, so that $\mathbf{S}_B$ has rank $K-1$, and that every set of $K-1$ centroids is affinely independent as well, so that the held-out scatter matrices have rank $K-2$. Holding out generator j re-centers the remaining centroids on their own mean, $\mathbf{d}_g^{(-j)} = \boldsymbol{\mu}_g - \frac{1}{K-1}\sum_{h\neq j} \boldsymbol{\mu}_h = \mathbf{d}_g + \frac{1}{K-1}\mathbf{d}_j$, and gives $\mathbf{S}_B^{(-j)} = \sum_{g\neq j} \mathbf{d}_g^{(-j)} \mathbf{d}_g^{(-j)\top}$ with eigenpairs $(\mu_m^{(j)}, \mathbf{u}_m^{(j)})$, $\mu_1^{(j)} \geq \cdots \geq \mu_{K-2}^{(j)} > 0$. The implementation keeps the top $k' = \min(k, K-2)$ of them and scores $s_i = \min_j \max_{m \leq k'} |\mathbf{v}_i^\top \mathbf{u}_m^{(j)}|$.

D.2 Holding out a generator is a rank-one downdate

Lemma 1. *For every j, $\mathbf{S}_B^{(-j)} = \mathbf{S}_B - \frac{K}{K-1}\mathbf{d}_j \mathbf{d}_j^\top$.*

Proof. Expanding the definition and using $\sum_{g\neq j} \mathbf{d}_g = -\mathbf{d}_j$,

$$\begin{aligned}\mathbf{S}_B^{(-j)} &= \sum_{g\neq j} \mathbf{d}_g \mathbf{d}_g^\top + \frac{1}{K-1}\Big(\sum_{g\neq j} \mathbf{d}_g\Big)\mathbf{d}_j^\top + \frac{1}{K-1}\mathbf{d}_j\Big(\sum_{g\neq j} \mathbf{d}_g\Big)^\top + \frac{K-1}{(K-1)^2}\mathbf{d}_j \mathbf{d}_j^\top \\ &= \mathbf{S}_B - \Big(1 + \frac{2}{K-1} - \frac{1}{K-1}\Big)\mathbf{d}_j \mathbf{d}_j^\top = \mathbf{S}_B - \frac{K}{K-1}\mathbf{d}_j \mathbf{d}_j^\top.\end{aligned}$$

□

The perturbation $\Delta\mathbf{S}_j = \frac{K}{K-1}\mathbf{d}_j \mathbf{d}_j^\top$ is positive semidefinite of rank one with $\|\Delta\mathbf{S}_j\|_{\mathrm{op}} = \frac{K}{K-1}\|\mathbf{d}_j\|^2$, so the eigenvalues interlace, $\eta_{m+1} \leq \mu_m^{(j)} \leq \eta_m$. The held-out bases are therefore rank-one perturbations of one matrix, which is what makes the two bounds below possible.

D.3 A generator-dominated direction cannot score high

Let $\rho_{ij} = \frac{K}{K-1}(\mathbf{d}_j^\top \mathbf{v}_i)^2/\eta_i \in [0, 1]$ be the fraction of the scatter along $\mathbf{v}_i$ that holding out generator j removes. The unscaled shares $(\mathbf{d}_j^\top \mathbf{v}_i)^2/\eta_i$ sum to one over j, because $\sum_j (\mathbf{d}_j^\top \mathbf{v}_i)^2 = \eta_i$.

Proposition 1. *For every $i \leq K-1$, every j and every m with $\mu_m^{(j)} > 0$,*

$$\mathbf{v}_i^\top \mathbf{S}_B^{(-j)} \mathbf{v}_i = (1 - \rho_{ij})\eta_i \qquad \textit{and} \qquad |\mathbf{v}_i^\top \mathbf{u}_m^{(j)}| \leq \sqrt{\frac{(1-\rho_{ij})\eta_i}{\mu_m^{(j)}}}.$$

Proof. The identity is Lemma 1 applied to $\mathbf{v}_i^\top \mathbf{S}_B \mathbf{v}_i = \eta_i$. Expanding $\mathbf{v}_i$ in the eigenbasis of $\mathbf{S}_B^{(-j)}$ gives $\mathbf{v}_i^\top \mathbf{S}_B^{(-j)} \mathbf{v}_i = \sum_m \mu_m^{(j)} (\mathbf{v}_i^\top \mathbf{u}_m^{(j)})^2$, and every term is non-negative, so each term is at most the sum. □

If holding out generator j removes most of the scatter along $\mathbf{v}_i$ (ρ_{ij} near one), then $\mathbf{v}_i$ cannot align with any held-out direction whose eigenvalue is comparable to η_i, and since s_i is a minimum over j it is at most the bound for that generator. The bound is informative for the retained directions of comparable eigenvalue and loose for a held-out direction with a much smaller one. Conversely, a high score bounds the fraction every held-out generator removes, through $\rho_{ij} \leq 1 - s_i^2 \, \mu_m^{(j)}/\eta_i$ for the matched m.

D.4 A direction with a large eigengap cannot score low

Let $\gamma_i = \min(\eta_{i-1} - \eta_i, \eta_i - \eta_{i+1})$ with $\eta_0 = \infty$ be the eigengap of $\mathbf{S}_B$ at i.

Proposition 2. *For every $i \le k'$ with $\gamma_i > 0$ and every j, $\sin\theta(\mathbf{v}_i, \mathbf{u}_i^{(j)}) \le \frac{2\|\Delta\mathbf{S}_j\|_{\mathrm{op}}}{\gamma_i} = \frac{2K\|\mathbf{d}_j\|^2}{(K-1)\gamma_i}$, and hence*

$$s_i \;\ge\; \min_j \sqrt{1 - \min\Big\{1, \Big(\frac{2K\|\mathbf{d}_j\|^2}{(K-1)\gamma_i}\Big)^2\Big\}}.$$

Proof. The first inequality is the Davis–Kahan $\sin\theta$ theorem [9] in the variant of Yu et al. [32], whose eigengap involves only the unperturbed matrix, applied to the symmetric matrices $\mathbf{S}_B$ and $\mathbf{S}_B^{(-j)}$ and the i-th eigenvector. The best-match alignment over the retained directions is at least the index-matched one, $\max_{m \le k'} |\mathbf{v}_i^\top \mathbf{u}_m^{(j)}| \ge |\mathbf{v}_i^\top \mathbf{u}_i^{(j)}| = \cos\theta(\mathbf{v}_i, \mathbf{u}_i^{(j)})$, and taking the minimum over j gives the second. □

When the centroid norms are comparable, $\|\mathbf{d}_j\|^2$ is about $\mathrm{tr}(\mathbf{S}_B)/K$ for every j, so the bound is informative for a direction whose eigengap exceeds about $2\,\mathrm{tr}(\mathbf{S}_B)/(K-1)$, a direction that aggregates comparable contributions from all generators and is separated from its neighbors. Both propositions were checked numerically on random centroid configurations (K from 3 to 11, D up to 300, Gaussian, heavy-tailed, near-rank-two and one-dominant-generator draws): the identity and the quadratic form hold to machine precision, and no instance violates either inequality.

D.5 Synthetic validation with known ground truth

The bounds are one-sided, so we measure the criterion's separation in a model where every direction has a known type. Centroids are $\boldsymbol{\mu}_g = \sum_{r=1}^{2} a_{gr}\sigma_r \mathbf{e}_r + \beta\mathbf{f}_g + \boldsymbol{\epsilon}_g$ in D=256 dimensions, with a rank-two shared subspace $\{\mathbf{e}_1, \mathbf{e}_2\}$ loaded by every generator ($a_{gr} \sim \mathcal{N}(0,1)$, $\sigma = (3,2)$), one specific direction $\mathbf{f}_g$ per generator (β=1), all $\mathbf{e}_r$ and $\mathbf{f}_g$ orthonormal, and isotropic noise of norm about ϵ. An LDA direction $\mathbf{v}$ is shared-dominated when $\|[\mathbf{e}_1\,\mathbf{e}_2]^\top \mathbf{v}\|^2$ exceeds $\max_g(\mathbf{f}_g^\top \mathbf{v})^2$ and specific-dominated otherwise. Figure 7 pools 200 draws at K=5, ϵ=0.05 (800 directions, 1,000 held-out events) and sweeps K and ϵ with 100 draws per cell.

The bound of Proposition 2 is never violated and is informative, below one, in about half of the held-out events. Stability separates the two types, with the median score of the shared-dominated directions well above that of the specific-dominated ones. At t=0.9 no specific-dominated direction passes the gate, while only about a quarter of the shared-dominated ones do, so the threshold errs toward penalizing. In these simulations it protects no specific-dominated direction, at the cost of penalizing most shared ones, which is consistent with the threshold exempting at most one direction on both backbones (Figure 3, Table 19) and is why we suggest $k \le K - 2$ and a per-configuration threshold. Figure 7(d) shows that the separation improves with the number of generators and degrades with noise.

E Broader impacts

Detectors that generalize to unseen generators help platforms and newsrooms screen synthetic images. The same detectors can mislabel legitimate AI-assisted or heavily edited photographs, and a false positive can cost its author. Scores should feed human review rather than automatic action, with the false-positive rate on the deployment distribution measured and disclosed. We use public benchmarks and release training and evaluation code but no detector weights.

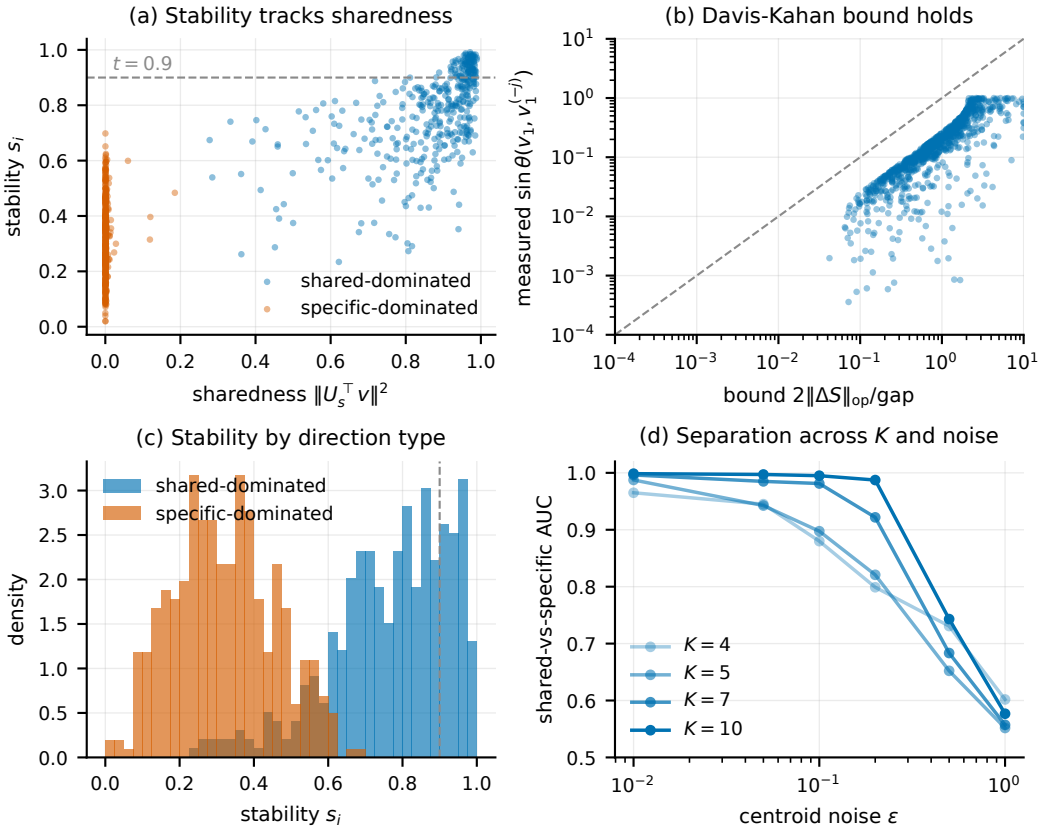


Figure 7: **Synthetic validation of the stability criterion.** (a) Stability against sharedness for shared-dominated and specific-dominated directions. (b) Measured rotation of the top direction under each held-out generator against the Davis–Kahan bound of Proposition 2. (c) Stability distributions by direction type with the threshold t=0.9. (d) Shared-versus-specific separation AUC of s_i over the number of generators K and the centroid noise ϵ. At K=5 and ϵ=0.05 the bound is below one in 555 of 1,000 events, the median scores are 0.80 (shared) and 0.32 (specific), and 24% of the shared-dominated directions pass t=0.9.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 state three claims: the entanglement diagnosis with the leave-one-out stability criterion (§3), the training-time StaR loss (§4), and the empirical results, namely a $+1.8$-point Avg. (4) gain of the continuous form over CLIP-ERM on CLIP ViT-L/14 (Welch $p = 0.0013$, eight runs per method) and a $4.1\times$ reduction of the across-seed SD of Avg. (4) by the thresholded form on ResNet-50-NoDown (one-sided F, $p = 0.0007$, eight seeds per model). A classifier-only L2 penalty at a tuned strength reaches the same CLIP mean (Table 15) and reduces the ResNet across-seed SD 2–3$\times$ without StaR's mean difference of $+0.7$ points (Welch $p = 0.14$, Appendix B.5). The mechanism analysis of §5.4 describes the feature geometry of the trained models. The L2 controls on both backbones bound what is specific to StaR (§6). The post-hoc comparison is scoped to what Tables 5 and 8 show. Post-hoc projection does not recover the training-time gain on CLIP, and on ResNet it recovers the mean but not the across-seed spread. What is not claimed (the ResNet mean difference, the thresholded CLIP gain, and JPEG robustness against EFFORT, whose leave-one-out protocol is not part of our evaluation) is stated in the Limitations paragraph of §6.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §6 discusses the limitations. Generator labels are required at training time (coarse family labels keep most of the gain, Appendix C.5), and the online LDA recomputation adds overhead. A classifier-only L2 penalty at one strength matches the CLIP mean gain and most of the ResNet SD reduction. On ResNet the controls reduce the across-seed SD 2–3$\times$ at the ERM mean (eight seeds, StaR $4.1\times$, and $+0.4$ to $+0.6$ points above the controls, Welch $p = 0.12$ and $p = 0.005$, Appendix B.5), and on CLIP a classifier-only L2 penalty of 10^{-2} during the same continuation reaches the continuous form's mean (0.936 vs. 0.937, $p = 0.76$, Table 15). The form of the gain differs by backbone, an across-seed SD reduction on ResNet whose mean difference against ERM is not claimed ($p = 0.14$) and a mean gain on CLIP that is significant from one ERM warm start and smaller and not significant from three further warm starts (Appendix C.1). We make no claim of superiority to EFFORT in JPEG robustness, where EFFORT scores above both StaR forms at the heaviest compression we test on two of the four benchmarks (Appendix C.7), and EFFORT's leave-one-out protocol is not part of our evaluation. λ was selected on the reported OOD benchmarks without a separate validation split, which is why we report the full λ sweep (eight runs per value, no detectable differences across λ) and a Chameleon cross-check.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Appendix D states the assumptions (centroid-only between-class scatter centered on the centroid mean, real images excluded, affinely independent centroids, an eigengap at the direction in question) and proves the rank-one downdate identity, the exact alignment bound for a generator-dominated direction and the Davis-Kahan bound for a direction with a large eigengap, citing the perturbation theorem used. Both bounds are one-sided and the appendix says so. A synthetic study with known ground truth reports where the bounds are informative and shows that the threshold t=0.9 errs toward penalizing.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: §5 (Setup) gives corpora, backbones, optimizers, epochs and the StaR defaults. Appendix A gives batch sizes, learning rates, weight decay, crops, augmentation, the frozen CLIP MLP weights, the online-LDA cadence, the λ selection procedure and the EFFORT training and evaluation transform.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code release (`https://github.com/anti-fake/DeepfakeDetector/tree/main/StaR`) contains the training/evaluation pipeline, configs, and result-aggregation scripts. All training datasets are publicly released and cited in §5. Preprocessing pointers are in Appendix A and the code release. Detector weights are not released (see Appendix E). Runs are not bitwise reproducible, and the two-stage CLIP results depend on the ERM warm start (Appendix C.1), which is why we report means over eight runs. A single run trains on one H100-class GPU.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: §5 (Setup) and Appendix A report optimizers (Adam for ResNet, AdamW for CLIP), learning rates, weight decay (0 for ResNet), epochs, batch sizes, both 1S and 2S schedules, the frozen CLIP MLP weights, and the StaR defaults (k=5, t=0.9, momentum 0.99). λ was selected on the reported OOD benchmarks (15 for ResNet, confirmed by a four-seed check over $\{10, 15, 25\}$ in Appendix A, and 19 for CLIP), and there is no separate validation split. The λ sweep (eight runs per value, no detectable differences, one-way ANOVA $p = 0.69$) is Table 17 (Appendix C.4) and the k and t sweeps (eight runs per setting) are Table 16 (Appendix C.3).

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Every multi-run entry uses eight runs unless its table states otherwise. Eight runs per method are used in Tables 1 and 3, in the λ sweep (Table 17), in the k and t sweeps (Table 16), in the post-hoc panels (Tables 5 and 8), in the classifier-only L2 controls on ResNet (Table 2) and on CLIP (Table 15) and in the matched two-stage ResNet runs (Table 11). Six runs per warm start are used in Table 6, and four runs per row in the ResNet ablation companion (Table 10) and the ResNet λ check. Cells are mean$\pm$SD with $n - 1$. Differences between independently trained runs use two-sided unpaired Welch t-tests with plain p-values. The post-hoc rows, which act on the same checkpoints as their ERM row, are tested on the within-checkpoint differences. Variance ratios use one-sided F-tests. The conventions are stated in §5. The continuous-form CLIP gain over CLIP-ERM is $+1.8$ points (Welch $p = 0.0013$), and the thresholded form's $+0.7$ ($p = 0.26$) is reported, not claimed. The ResNet across-seed SD ratio is $4.1\times$ ($F = 17.0$, $p = 0.0007$), significant per benchmark on WildRF, GenImage, LDM-FakeDetect and CommunityForensics and not on Synthbuster, COCOAI or Chameleon. The ResNet mean difference against ERM ($+0.7$ points, $p = 0.14$) is not claimed.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: Appendix A reports hardware (NVIDIA H100 80GB), per-run wall-clock (ResNet 1S-StaR about 12 GPU-hours, CLIP 2S-StaR about 1.7 GPU-hours), and total training compute (about 2,100 GPU-hours for the reported runs, about 2,700 GPU-hours for all training runs of the project, the sum of the recorded per-epoch wall times over all training runs, including sweeps, the per-epoch trajectory study, the seed extensions and ablations not surfaced in the main paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public benchmarks without identifiable individuals as analysis targets, does not involve human subjects, and addresses dual-use considerations in Appendix E.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix E discusses positive (screening synthetic images on platforms and in journalism) and negative (mislabeling legitimate AI-assisted work, hence the need for human review in consequential decisions) impacts.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We release training/evaluation code and per-recipe configs but not deployable detector checkpoints, to limit adversarial-evasion fine-tuning. We do not redistribute the underlying image benchmarks. Users obtain them from the original sources under the original licenses.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets are cited in §5: Corvi-LDM [8], GenImage [33], Synthbuster [2], WildRF [5], COCOAI [25], LDM-FakeDetect [22, 21], Chameleon [30], CommunityForensics [19], and the ProGAN images of CNNDetection [28]. The CLIP ViT-L/14 backbone [20] is cited there as well. Baseline implementations and checkpoints (CNNSpot, UnivFD, NPR, AIDE, DRCT, EFFORT, and the ResNet-50-NoDown architecture [13, 8]) are credited at first use. The code release lists the source URL of each asset and the license its source states, or that none is stated.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new asset is a code release (StaR training pipeline and result-aggregation scripts), provided as a public code release with a README documenting environment setup, recipe-by-recipe reproduction commands, and dataset-acquisition pointers.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human-subjects research.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, so IRB review is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [N/A]

    Justification: LLMs are not part of the core methodology. The detector backbones are ResNet-50 and CLIP ViT-L/14 (used as image feature extractors), and no LLM is used for training, regularization, or evaluation.