NOT ALL CHANNELS ARE EQUAL: PERTURBATION-INVARIANT CHANNEL SELECTION FOR ROBUST AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image detectors often lose accuracy after compression, blurring, noise, or resizing. We ask whether this sensitivity is concentrated in particular feature channels of a pretrained detector, and on a forensic-trained CNN we find that it is: the channels whose spatial activation patterns survive perturbation are not the ones the classifier relies on most. We propose PICS (Perturbation-Invariant Channel Selection), which scores each channel by the Pearson correlation between its clean and perturbed activation maps, keeps the union of the highest-ranked channels across perturbations, and retrains the linear head on the selected channels. Under a common training protocol, PICS improves robustness not only beyond retraining the head with augmentation alone, but also beyond magnitude-based, random, attention-based, and weight-clamping alternatives, as well as end-to-end fine-tuning of the backbone. The recipe applies to CNN and ViT detectors, with the largest gains over the released detectors on forensic-trained backbones whose robust and discriminative channels are most separated, and we evaluate it on cross-dataset benchmarks and social-media platforms. These results support perturbation stability as a practical criterion for selecting the features of forensic detectors.

1 INTRODUCTION

Recent advances in generative modeling (Rombach et al., 2022; Esser et al., 2024; Dhariwal & Nichol, 2021) have raised significant concerns about misinformation and fraud, motivating forensic detectors (Corvi et al., 2023; Rajan et al., 2025; Wang et al., 2020b) that distinguish real from synthetic images based on subtle generative artifacts. While these detectors achieve high accuracy on clean test images, their performance degrades substantially when images undergo common post-processing (JPEG and WebP compression, Gaussian blur, additive noise, resizing, color jitter) that is ubiquitous in real-world deployment such as social-media sharing.

Existing approaches address this fragility from two directions: *augmentation-based* training (Wang et al., 2020b; Gragnaniello et al., 2021), which exposes detectors to perturbed inputs but does not explicitly identify which learned features are reliable, and *channel-selection* methods such as Stay-Positive's non-negative weights (Rajan & Lee, 2025), magnitude pruning (He et al., 2017; Li et al., 2017), and channel attention (Hu et al., 2018; Woo et al., 2018; Wang et al., 2020a), which change which features the classifier uses. None of them directly targets a channel's *robustness*: how well its spatial activation pattern survives perturbation.

We hypothesize a complementary explanation for forensic-detector fragility: **the channels currently used by the classifier are not the most perturbation-robust**, and selecting for robustness (rather than for sparsity, magnitude, or accuracy) is a generic recipe for improving robustness across architectures. Figure 1 illustrates the underlying phenomenon: given the same pretrained Corvi detector, a robust channel preserves its spatial activation pattern after Telegram transmission, while a fragile one is severely disrupted. Among the channels of the final convolutional layer, perturbation sensitivity varies dramatically. When the classifier relies on fragile channels, perturbations distort its evidence.

We test this hypothesis with PICS (Perturbation-Invariant Channel Selection), a three-step recipe. First, we quantify per-channel robustness by measuring Pearson correlation of spatial activation maps between clean and perturbed images across diverse perturbation categories. Second, we select the

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

union of the top-$K\%$ most robust channels. Third, we train only the fully-connected (FC) classifier on the masked features with the backbone frozen (for the general-purpose CLIP backbone, the final LayerNorm is co-trained, Section D.5). PICS adds no parameters and, for CNN detectors, reduces at inference to a fixed mask on the pooled features (ViT instantiations in Sections D and E).

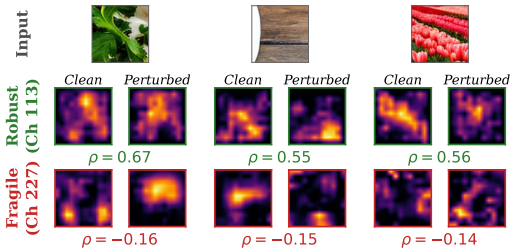

Figure 1: **Not all feature channels are equally robust.** Layer-4 activation maps of a pretrained detector for three AI-generated images before and after Telegram transmission. A robust channel (top, green) preserves its spatial pattern ($\rho \geq 0.55$). A fragile channel (bottom, red) is disrupted ($\rho < 0$).

Under a common training protocol on a forensic-trained CNN backbone, PICS outperforms FC-magnitude pruning and random selection at a matched channel budget, as well as SE channel attention (Hu et al., 2018), Stay-Positive's non-negative weight clamping, and the unmasked control (Table 2), and it also outperforms the evaluated end-to-end fine-tuning of the backbone. A separately configured bottom-ranked-channel control is reported in the appendix. The same procedure applies to two CNN detector checkpoints (Corvi (Corvi et al., 2023), Rajan (Rajan et al., 2025)) and two ViT backbones (UniversalFakeDetect (Ojha et al., 2023) on CLIP ViT-L/14, where the gain turns out to come from adapting the final LayerNorm rather than from selection, and B-Free (Guillaro et al., 2025) on DINOv2 ViT-B/14), and we evaluate it on social-media platforms, cross-dataset benchmarks, and additional unseen generators. The largest gain is on multi-step social-media-style compression (+6.3 pp over Stay-Positive, Table 6). On fifteen perturbation families outside the scoring set, PICS preserves baseline performance on most conditions (Table S26), with regressions concentrated on rotation and occlusion, which we characterize as a scope boundary of the spatial-correspondence criterion (Section 6).

Our contributions are as follows:

1. A channel-level diagnosis: perturbation stability and clean discriminativeness are negatively associated on the Corvi backbone and only weakly related on Rajan and on fine-tuned checkpoints (Section 3).
2. PICS, a selection rule based on clean-to-perturbed spatial correlation with a retrained head, which outperforms augmentation-only retraining, magnitude and random selection, attention-based reweighting, non-negative clamping, and end-to-end fine-tuning under a common protocol (Section 5.3).
3. An evaluation across four detector backbones, fifteen additional perturbation families, cross-dataset benchmarks, and social-media platforms, with a documented scope boundary on rotation and occlusion and a classifier constraint that improves that subset (Section 6).

2 Related Work

AI-generated image detection. Training-based detectors learn to distinguish real from fake images using neural network classifiers. Wang et al. (2020b) showed that data augmentation during training improves cross-generator generalization. Gragnaniello et al. (2021) introduced ResNet50-NoDown, a ResNet-50 without down-sampling in the first block trained with strong augmentation, which Corvi et al. (2023) combined with patch-wise training (Chai et al., 2020) and extended to latent diffusion model (LDM) detection. Rajan et al. (2025) proposed training on autoencoder reconstructions of real images to synthesize complementary fake data. Another direction leverages strong, general-purpose visual encoders for cross-generator transfer. UniversalFakeDetect (Ojha et al., 2023) adopts CLIP (Radford et al., 2021) features with linear probing, and CLIPDet (Cozzolino et al., 2024a) further refines this paradigm. Frequency- and artifact-based cues (Odena et al., 2016; Zhang et al., 2019; Chai et al., 2020; Karageorgiou et al., 2025) and zero-shot detection (Cozzolino et al., 2024b) are complementary directions. PICS is a post-hoc robustness enhancement for detectors that expose spatially indexed features and a retrainable head.

Robustness of forensic detectors. Grommelt et al. (2024) revealed that common detection benchmarks contain JPEG-compression and image-size biases that detectors learn, distorting reported accuracy. Rajan & Lee (2025) showed that detectors trained on LSUN images (Yu et al., 2015) spuriously associate WebP compression artifacts with real images, and proposed Stay-Positive, constraining FC weights to be non-negative, to eliminate reliance on real-image features. Stay-Positive mitigates spurious correlations but does not explicitly identify which features are perturbation-robust, the gap our channel-level criterion addresses.

Channel pruning and feature selection. Structured channel pruning removes entire channels for efficiency, guided by weight magnitude or sensitivity surrogates (He et al., 2017; Li et al., 2017), and channel-wise *selection* is also used as a representational mechanism: channel-attention modules (*e.g.*, SE (Hu et al., 2018), CBAM (Woo et al., 2018), and ECA (Wang et al., 2020a)) learn to reweight channels to emphasize task-relevant features, and related approaches learn explicit gates or sparse masks (*e.g.*, L_0-style feature masking) to select a subset of informative representations (Louizos et al., 2018). Both are optimized for efficiency or predictive performance under the training distribution, not for the robustness of activations under post-processing. Our goal is robustness rather than compression: we rank channels by the Pearson correlation of their activation maps before and after perturbation and mask out fragile ones, a criterion distinct from magnitude-based pruning or accuracy-driven attention and gating. FVRCP (He et al., 2020) uses Pearson correlation as a compression statistic, whereas ours correlates the same channel's clean and perturbed maps to measure robustness. In AI-generated image detection, Gaintseva et al. (2024) remove components of frozen CLIP embeddings to improve robustness. Closest to our motivation, the recent DEAR (Kim et al., 2026) measures how strongly each channel's activation aligns with inpainting masks (mean activation inside versus outside the mask) and prunes channels at both extremes, which it shows to be the least robust to post-processing. PICS instead measures robustness directly through clean–perturbed correlation on paired inputs and selects a per-perturbation top-$K\%$ union, differing in both the measured quantity and the aggregation rule. The recent DRIFT (Ameta et al., 2026) learns projections onto robust and fragile feature subspaces and detects via perturbation drift at inference, unlike PICS's fixed channel selection with no test-time perturbation.

3 Diagnosis: Channel Robustness and Discriminativeness

We diagnose at the channel level, on a frozen backbone ϕ producing a layer-4 feature map $\phi(x) \in \mathbb{R}^{C\times H\times W}$ (C=2,048 for ResNet-50 (He et al., 2016)) followed by global average pooling and a linear classifier. We measure two properties per channel.

Robustness: the Pearson correlation $\bar{\rho}_c^{(\tau_a)}$ between spatial activation maps of clean and perturbed images, averaged over training images and a set of scoring transforms spanning compression, blur, noise, resize, and color changes (Eq. (1)), where the diagnosis uses the 18-level analysis set of Section B and selection in Section 4 its 11-setting subset.

Discriminativeness: $\max(\mathrm{AUC}, 1-\mathrm{AUC})$, where AUC treats the channel's global-average-pooled activation as a standalone binary classifier on clean inputs. The two scores are computed independently and capture distinct channel properties.

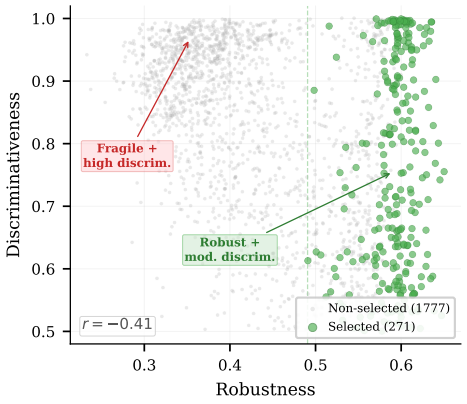

Figure 2: **Robustness vs. discriminativeness** on the Corvi backbone, one point per channel. The two properties are negatively correlated ($r = -0.41$): the most discriminative channels tend to be the most fragile under perturbation.

Stability and discriminativeness differ across backbones. Figure 2 shows that on the Corvi backbone the two properties are negatively correlated: the most discriminative channels tend to be the most fragile. Only a minority of channels fall in the high-robustness, high-discriminativeness quadrant, and channel robustness is not well predicted by the clean-trained classifier's weight magnitudes (Spearman $\rho = +0.13$, Section F.1). The Rajan backbone lacks this negative association ($r = +0.06$), and so do the end-to-end fine-tuned

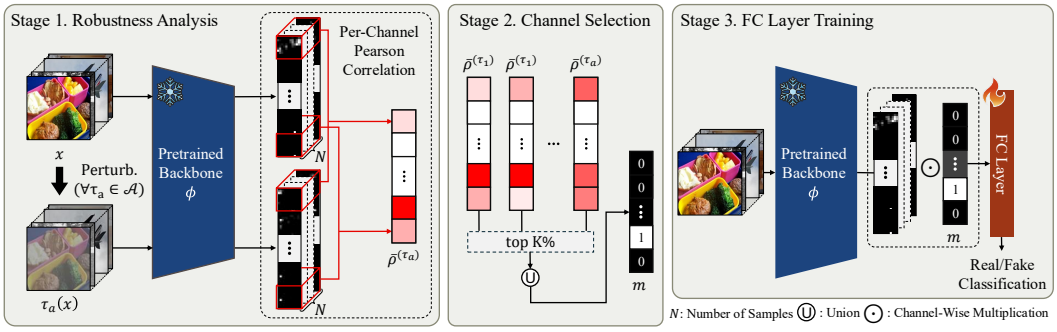

Figure 3: **Method overview.** (1) For N training images and a set of perturbations, compute per-channel Pearson correlation $\bar{\rho}_c^{(\tau_a)}$ between clean and perturbed spatial activation maps. (2) Rank channels per perturbation, take the union of the top-$K\%$. (3) Train only the FC classifier on the masked features with the backbone frozen (the CLIP variant also adapts its final LayerNorm).

Corvi checkpoints of Section F.2 (r between -0.14 and $+0.33$). This pattern is consistent with the larger selection gains we observe on Corvi than on Rajan and on the fine-tuned checkpoints (Tables 1 and 2).

Robustness sets are partially disjoint across perturbations. The distribution of $\bar{\rho}_c$ varies dramatically across perturbation categories (Figure S2, appendix): mild perturbations (grayscale, jitter) leave most channels intact, while destructive ones (strong blur, heavy compression) push the majority to low correlation. Per-perturbation top-$K\%$ sets are therefore partially disjoint, a property that motivates the union aggregation we adopt in Section 4.

4 METHOD

The diagnosis motivates a generic recipe: select channels by robustness rather than by accuracy or sparsity. PICS executes this in three stages (Figure 3) with no added parameters.

Stage 1: per-channel Pearson criterion. For N training images $\{x_n\}$ and perturbations $\mathcal{A} = \{\tau_1, \ldots, \tau_{|\mathcal{A}|}\}$ spanning compression, blur, noise, resize, and color transforms, let $f_c{=}\phi_c(x_n) \in \mathbb{R}^{HW}$ and $g_c{=}\phi_c(\tau_a(x_n))$ denote the flattened clean and perturbed spatial activations of channel c for image x_n (index omitted inside the sum). The robustness score averages their Pearson correlation over images, per transform:

$$\bar{\rho}_c^{(\tau_a)} = \frac{1}{N}\sum_{n=1}^{N} \frac{\sum_{i=1}^{HW}\left(f_{c,i}-\bar{f}_c\right)\left(g_{c,i}-\bar{g}_c\right)}{\left\|f_c-\bar{f}_c\right\|_2 \cdot \left\|g_c-\bar{g}_c\right\|_2}, \tag{1}$$

with $\bar{f}_c, \bar{g}_c$ the spatial means. A value $\bar{\rho}_c^{(\tau_a)} \approx 1$ indicates spatial-pattern preservation, while values near 0 or negative indicate strong disruption.

Why Pearson? The criterion is invariant to per-channel affine rescaling, $\rho(af{+}b,\, g) = \rho(f,\, g)$ for $a{>}0$, a useful property because compression, blur, and additive noise largely rescale or shift channel activations while preserving their spatial layout. Distance-based criteria (relative-L_2, GAP-magnitude difference) penalize benign rescaling and reject spatially robust channels. Per-channel AUC on $\mathrm{GAP}(f)$ ignores spatial structure and does not measure perturbation stability, so it can reward channels whose pooled activation discriminates clean inputs but changes under perturbation. Pearson sits between these failure modes, and in the tested setting it outperforms relative-L_2, the AUC threshold, and their three-metric union in Overall AP (Table S17). The criterion measures preservation of a channel's spatial pattern, not of its pooled value: an offset that leaves ρ unchanged still shifts GAP and the logit, so Pearson is an empirical proxy for decision-level stability, which we test in Section F.1.

Stage 2: channel selection by union-of-top-$K\%$. For each perturbation τ_a we rank channels by $\bar{\rho}_c^{(\tau_a)}$ and take the top-$K\%$, then union across scoring transforms: $\mathcal{S} = \bigcup_{\tau_a \in \mathcal{A}} \mathrm{top}_{K\%}\{c \mid \bar{\rho}_c^{(\tau_a)}\}$. The union retains every channel that is among the highest-scoring ones for at least one scoring transform, so the retained fraction exceeds $K\%$ (Table 5), whereas pure intersection becomes impractically small (Figure S5). At a matched 271-channel budget the union also outperforms average-Pearson, minimum-Pearson, mean-rank, and k-of-n aggregations in Overall AP (Table S21), so the choice is not merely a channel-count effect.

Choosing K. K is a hyperparameter that we sweep once per backbone and training protocol and then fix (Table 5, with $K{=}1\%$ for the LDM-trained Corvi protocol, Table S22). The validation split of the head-training data saturates (balanced accuracy above 0.99 for every K on the CNN backbones), so it does not discriminate between values of K. The sweeps in Tables S7, S12, S19 and S20 report the primary benchmark, on which the operating K was chosen, and every sweep sits on a plateau within 1 pp over most of its range. The primary-benchmark results are therefore development results, and all transfer evaluations reuse the selected configuration without further tuning. The selected fraction is narrower for the forensic-trained CNNs and the DINOv2 backbone than for the CLIP representation, which is consistent with, but does not by itself establish, a more distributed discriminative signal in CLIP.

Stage 3: FC-only training with channel mask. We construct a binary mask $\mathbf{m} \in \{0,1\}^C$ ($m_c{=}1$ iff $c \in \mathcal{S}$) and train a linear head $z = \mathbf{w}^\top[\mathbf{m} \odot \mathrm{GAP}(\phi(x))] + b$ with the backbone frozen (for CLIP, the final LayerNorm is adapted as well, Section D). The recipe extends to vision transformers by treating patch-token channels as the analogue of CNN feature maps (instantiations on CLIP ViT-L/14 and DINOv2 ViT-B/14 in Table 1). Training hyperparameters and augmentation schedule are in Section B.

5 Experiments

5.1 Experimental setup

Backbones. We evaluate four detector backbones spanning forensic-trained CNNs and ViTs initialized from general-purpose pretraining (B-Free is additionally forensic fine-tuned): (1) Corvi (Corvi et al., 2023), ResNet50-NoDown (Gragnaniello et al., 2021) trained on LDM-generated fakes, (2) Rajan (Rajan et al., 2025), the same architecture trained on autoencoder reconstructions of real images, (3) UniversalFakeDetect (UFD) (Ojha et al., 2023), a CLIP ViT-L/14 backbone with a linear FC head trained on ProGAN, and (4) B-Free (Guillaro et al., 2025), a DINOv2 ViT-B/14 backbone with register tokens and a linear head. All four use Pearson scoring and union selection, with backbone-specific details in Sections B, D and E.

Training protocols and data. For Corvi/Rajan, we train the FC classifier using two types of fake data: (1) paired autoencoder (AE) reconstructions with synchronized augmentations, and (2) unpaired LDM-generated fakes with independent augmentations. The paired protocol is the primary configuration (both are reported in Table S22). Both protocols share 180k real images from COCO (Lin et al., 2014) and LSUN (Yu et al., 2015) paired with 180k fakes. For UFD, we retrain the linear head on a GenImage/BigGAN subset (Zhu et al., 2023), following the frozen-backbone linear-probing protocol of Ojha et al. (2023) with a different training set. For B-Free, we retrain the linear head on the released B-Free training set of Guillaro et al. (2025). Dataset sizes and splits are in Sections B, D and E.

Test benchmarks. The Corvi, Rajan, and B-Free families are evaluated on the LDMFakeDetect benchmark (Rajan et al., 2025) (nine unseen generators with balanced real/fake splits).[1] The UFD family is evaluated on the eight GenImage generators (Zhu et al., 2023), following GenImage's cross-generator protocol. Cross-dataset evaluation swaps the two benchmarks (Section C.3, appendix).

[1] Generators: Midjourney (Midjourney, 2024), aMUSEd (Patil et al., 2024), FLUX (Black Forest Labs, 2024), Kandinsky (Razzhigaev et al., 2023), LCM (Luo et al., 2023), PixArt-α (Chen et al., 2024), Playground (Li et al., 2024), Stable Diffusion (Rombach et al., 2022), Würstchen (Pernías et al., 2024).

Table 1: **AP (%) across four detector backbones** on Clean and per-perturbation categories of each backbone's primary benchmark. **Pert.** is the mean over 35 single-perturbation conditions (Combo conditions are reported separately). **Overall** is the mean over all 51 conditions. ± is the std over training seeds, and rows without ± are single runs. **Bold** marks the best within each backbone group.

Method	Ch.	Clean	JPEG	WebP	Blur	Noise	Jitter	Resize	Combo	Pert.	Overall
ResNet-50 (Corvi). Test: LDMFakeDetect 9-gen × 51 cond.											
Corvi	2,048	92.7	83.3	81.6	78.9	76.7	90.0	70.4	73.9	81.3	79.3
Corvi⊕	27	**98.5**	87.8	85.1	83.2	78.3	95.1	77.3	80.3	85.3	84.1
Ours: top-5%	271	97.9±0.2	**91.2**	**88.3**	**85.6**	**82.9**	**95.4**	**79.2**	**82.6**	**88.0±0.3**	**86.6±0.3**
ResNet-50 (Rajan). Test: LDMFakeDetect 9-gen × 51 cond.											
Rajan	2,048	91.3	85.2	74.7	**86.9**	73.9	89.7	73.8	76.7	81.4	80.2
Rajan⊕	128	93.9	87.0	**81.7**	83.5	80.4	92.0	**75.7**	79.5	84.2	83.0
Ours: top-2%	147	**94.1±0.2**	**88.7**	81.1	84.8	**81.1**	**92.5**	75.3	**80.0**	**84.8±0.1**	**83.6±0.0**
CLIP ViT-L/14 (UFD). Test: GenImage 8-gen × 51 cond.											
UFD (FC retrained)	1,024	85.1	70.3	70.5	67.4	75.0	78.5	69.5	65.8	72.1	70.5
UFD (LayerNorm+FC)	1,024	**95.1±0.3**	**78.4**	74.1	67.9	**78.2**	**83.7**	**75.5**	**71.0**	**76.5±0.5**	**75.3±0.5**
Ours: top-20%	423	94.8±0.2	77.8	**74.6**	**68.8**	77.4	83.2	75.3	70.9	**76.5±0.7**	75.2±0.7
DINOv2 ViT-B/14 (B-Free). Test: LDMFakeDetect 9-gen × 51 cond.											
B-Free	768	96.9	95.5	95.5	95.0	94.2	96.0	93.4	94.7	95.1	95.1
Ours: top-7%	76	**97.6**	**96.5**	**96.2**	**96.0**	**95.0**	**96.7**	**95.2**	**95.6**	**96.0**	**96.0**

Perturbation conditions. We evaluate under 51 total conditions: clean, 35 single-perturbation conditions spanning 7 categories (JPEG, WebP, blur, noise, color jitter, resize, grayscale) at multiple severity levels, and 15 multi-step combinations (*e.g.*, resize+JPEG, blur+noise+JPEG), simulating real-world scenarios. Exact parameters are listed in Section K, and Table 1 omits the Grayscale column for space.

Metrics. We report Average Precision (AP) following prior work (Wang et al., 2020b; Ojha et al., 2023; Rajan & Lee, 2025). AP is computed per generator on its balanced split and macro-averaged over generators, and "Overall AP" is the mean over all 51 conditions. For real-world deployment evaluation (Table 3), we additionally report real and fake accuracy (R.Acc, F.Acc) at logit threshold 0.

Baselines. For each backbone we compare against the pretrained detector and, for the CNN backbones, Stay-Positive's non-negative weight clamping (Rajan & Lee, 2025) (⊕). Baseline detectors are released checkpoints or our reproductions (Sections B and E). The two Stay-Positive reproductions match the released checkpoints within 0.2 pp, and Rajan⊕ in Table 1 is the released checkpoint. All share identical frozen backbones.

5.2 Main results

Table 1 reports perturbation robustness across four backbones spanning both forensic-trained CNNs (Corvi, Rajan) and general-purpose ViTs (CLIP and DINOv2, instantiated through UFD (Ojha et al., 2023) and B-Free (Guillaro et al., 2025)). PICS gives the best Pert. AP in the three forensic-trained groups while staying close to the best Clean AP of each group, whereas on CLIP it only ties an unmasked control, as discussed below. The Corvi-family gain is large and paired-significant over the nine generators (Section B). The Rajan-family gains are smaller and more uneven, consistent with the weaker robustness–discriminativeness association on that backbone (Section 3). PICS improves on the strongest non-masking baseline on every Corvi-family generator and on all but five of the 51 conditions (Tables S1 and S31, with per-severity curves in Figure S1).

The two ViT backbones extend the same recipe to general-purpose representations. On CLIP ViT-L/14, all three rows of Table 1 are heads retrained on the same GenImage/BigGAN data, and the decisive factor is the final LayerNorm: a LayerNorm+FC head trained without any mask matches PICS, so on this backbone the margin over the FC-only head comes from the adaptation and channel selection adds nothing measurable. We report this as a null result for selection on CLIP. On DINOv2 ViT-B/14, PICS retains only a tenth of the channels and improves every category over B-Free (Table 1),

Table 2: **Channel-suppression methods and end-to-end fine-tuning compared under a common protocol on the Corvi backbone (aggressive augmentation, no clamp unless noted).** Hard-selection controls are budget-matched to PICS, and SE reweights all channels (Ch. column). The Corvi⊕ row is the Stay-Positive reproduction under its own schedule, with its aggressive-schedule counterpart in Table S18. End-to-end fine-tuning updates all backbone parameters with the PICS head's data, augmentation distribution, and stopping criterion (Section F.2). ± is the std over training seeds, and rows without ± are single runs.

Method	Ch.	Clean	Pert.	Combo	Overall
Corvi⊕ (Stay-Positive)	27	**98.5**	85.3	80.3	84.1
FC-magnitude pruning	271	95.4	84.0	78.1	82.5
SE channel attention	—	97.5	85.4	78.6	83.6
No-mask (aggressive aug)	2,048	97.6±0.3	85.4±0.1	78.6±0.2	83.6±0.1
Random-271	271	95.0±1.2	83.2±1.1	77.2±0.9	81.6±1.1
End-to-end fine-tuning, no selection	2,048	95.9±0.5	84.2±0.5	77.6±1.0	82.5±0.6
PICS top-5% on fine-tuned backbone	250–270	94.8±1.2	84.4±0.3	78.5±0.7	82.9±0.4
PICS top-5% (ours)	271	97.9±0.2	**88.0±0.3**	**82.6±0.4**	**86.6±0.3**

and most of the gain comes from FLUX (Black Forest Labs, 2024), the one generator on which the baseline is far from saturation (Table S13). Cross-dataset clean AP improves for Corvi and stays within a point for B-Free and Rajan, while PICS-CLIP does not transfer (Table S2). Implementation details for the ViT instantiations are in Sections D and E.

5.3 Isolating the criterion: matched-budget channel-selection and reweighting baselines

To isolate the contribution of the robustness criterion from generic feature suppression, we compare PICS against five alternative channel-selection and reweighting methods under a common protocol, with hard-selection alternatives budget-matched to PICS and Stay-Positive on its own schedule (Table 2). The alternatives span four selection rationales: Stay-Positive's non-negative weight clamping (an implicit selection by predictive utility), FC-magnitude pruning, SE channel attention (learned reweighting for accuracy), a random-selection control, and an unmasked control trained with the same augmentation. PICS outperforms every alternative on Pert., Combo, and Overall AP. SE attention, despite its added parameters, matches the unmasked baseline, so the tested accuracy-driven reweighting did not improve robustness. Random selection underperforms even the unmasked baseline, so the gain is not reproduced by a matched random subset. Stay-Positive's clamping reaches a Pert. AP comparable to the unmasked control and is clearly below PICS. Holding the channel count fixed under the WebP protocol, uncentered cosine similarity, logit-contribution stability, a joint robustness-discriminativeness product, and random selection all trail Pearson by a wide margin, so the gain comes from which channels are kept rather than from how many (Table 4).

Because Corvi, Rajan, and B-Free were already trained end-to-end with augmentation (Corvi and Rajan inherit the recipe of Gragnaniello et al. (2021)), the backbones we select over are not perturbation-naive. We nevertheless fine-tune the full Corvi backbone a second time with the same data, augmentation distribution, and stopping criterion used for the PICS head (Section F.2). Averaged over training seeds, it lands below PICS on the frozen backbone and even below the frozen-backbone unmasked control (Table 2), across the learning rates and stopping settings we tested. Applying PICS to the fine-tuned backbones adds a small further gain (Table 2), and B-Free, an end-to-end augmentation-trained ViT, also improves (Table 1).

5.4 Real-world deployment evaluation

We evaluate on Telegram (LDMFakeDetect images that we upload to and download from Telegram ourselves, Section B) and WildRF (Cavia et al., 2024) (real and fake images scraped from Reddit, X, and Facebook). At threshold zero, PICS raises fake-image accuracy on all four platforms for both families, at the cost of lower real-image accuracy (Table 3). Balanced accuracy rises only on some platforms (X and Facebook for Corvi, Telegram for Rajan) and falls on the others. AP is also

Table 3: **Real-world deployment evaluation** on Telegram and WildRF (Reddit, X, Facebook). AP (%) and real and fake accuracy (R.Acc, F.Acc, %) at decision threshold 0. **Bold** marks the best per metric.

	Telegram			Reddit			X (Twitter)			Facebook		
Method	AP	R.Acc	F.Acc	AP	R.Acc	F.Acc	AP	R.Acc	F.Acc	AP	R.Acc	F.Acc
Corvi family												
Corvi	**97.1**	**93.7**	84.7	77.3	**91.2**	48.8	**86.9**	**95.6**	53.8	85.8	**91.2**	60.6
Corvi⊕	90.9	83.9	78.5	**80.5**	86.8	52.3	82.6	85.6	59.1	90.1	86.9	71.9
Ours: top-5%	96.6	81.2	**92.5**	78.2	77.3	**57.5**	86.7	81.2	**71.3**	**91.6**	78.1	**85.6**
Rajan family												
Rajan	**83.5**	**96.2**	32.8	**78.3**	88.0	51.3	87.8	**95.0**	58.8	92.0	**93.1**	69.4
Rajan⊕	74.4	92.8	32.1	78.0	**88.3**	50.8	**88.0**	92.1	64.9	92.3	90.6	77.5
Ours: top-2%	81.2	91.1	**48.0**	77.0	82.8	**53.2**	**88.0**	90.3	**66.1**	**92.3**	84.4	**82.5**

Table 4: **Count-matched selection criteria under the WebP protocol** (Overall AP, %, 51 conditions). All masks retain 271 channels (273 for the logit criterion, the closest achievable union) and all heads train under the identical schedule.

Criterion (same schedule)	Ch.	Overall
Pearson union top-5% (ours)	271	**86.4**
Cosine similarity, uncentered	271	81.3
Logit-contribution stability $\lvert w \cdot \Delta\mathrm{GAP}\rvert$	273	81.7
Random	271	80.4
Joint product $\bar{\rho} \times A$	271	78.6

mixed. These results show an operating-point shift rather than uniform deployment superiority. On the perturbed conditions of the primary benchmark, threshold-level metrics do improve (PICS lowers the false-positive rate at 95% recall and raises balanced accuracy at threshold 0, Table S15), while probability calibration trails Corvi⊕ and is restored by a single temperature (Table S16).

5.5 Ablation studies

All ablations use the Corvi protocol unless noted. **Channel selection metric:** Pearson outperforms relative L_2, per-channel AUC, and a combined-union baseline under the same protocol (channel counts differ, Table S17, and count-matched comparisons are in Section F.1). Per-channel AUC (selecting by clean discriminativeness alone) underperforms Pearson despite directly optimizing for utility, indicating that robustness and discriminativeness are distinct properties. **Masking vs. clamping** (Table S18, factorial under aggressive augmentation): masking gives the largest gain, clamping helps only without masking, and combining them falls back to the unmasked level in this experiment, and the interaction depends on the evaluation conditions (Section J.2). **Channel identity:** PICS also outperforms random channel selection at a matched budget (Table 2), and severity curves with a separately configured bottom-ranked-channel control are in Figure S4. K **plateau** (Table 5): every per-backbone sweep sits within 1.0 pp of its optimum over most of the swept range.

6 Discussion and Limitations

Perturbations outside the scoring set. The eleven scoring transforms leave open whether the selected channels generalize beyond them. We evaluate fifteen additional families (Table 6, and all fifteen with group means in Tables S25 and S26). They are excluded from channel scoring but not from head training, whose augmentation shares operators with several of them (Section B). The largest gain is on JPEG-to-WebP recompression, and photometric and filtering families are on par with Corvi⊕.

Table 5: K **sits on a plateau for every backbone.** Swept range, selected K (retained channels), and the range within 1.0 pp Overall AP of the best setting. Full sweeps in Tables S7, S12, S19 and S20.

Backbone	Swept K	Selected K (channels, fraction)	Within 1.0 pp of best
Corvi (ResNet-50)	3–10%	5% (271, 13.2%)	entire sweep
Rajan (ResNet-50)	1–5%	2% (147, 7.2%)	entire sweep
UFD (CLIP ViT-L/14)	5–25%	20% (423, 41.3%)	10–25%
B-Free (DINOv2 ViT-B/14)	1–20%	7% (76, 9.9%)	5–20%
End-to-end fine-tuned Corvi	2–10%	5% used (optimum 7%)	entire sweep

Table 6: **Six real-world perturbation families** that capture common pre-publication processing pipelines, a deployment-relevant subset of the fifteen families outside the scoring set. Mean AP (%) over the 9 generators. **Bold** marks the best per family, and Δ is PICS minus Corvi$\oplus$. The families with regressions (geometric and occlusion) are in Table S26.

Family	#cnd	Corvi	Corvi$\oplus$	PICS (ours)	Δ
2-step compression (JPEG70→WebP70)	1	84.6	83.5	**89.8**	$+6.3$
Bilateral filter (d=5, 9)	2	89.0	89.0	**90.4**	$+1.4$
Sharpen (9 severities r=0.25–4)	9	93.7	97.2	**97.6**	$+0.3$
Median blur (k=3, 5)	2	83.6	87.8	**88.0**	$+0.2$
Brightness ($\Delta \in \{\pm 25, \pm 50\}$)	4	92.3	97.4	**97.5**	$+0.1$
Contrast ($f \in \{0.5, 0.7, 1.5\}$)	3	91.9	**97.3**	97.2	-0.1

Scope of the criterion and a classifier constraint. The largest regressions are on rotation and occlusion (cutout, sticker) for the three hardest unseen generators (aMUSEd, FLUX, Playground). This subset scores lower than the remaining cells of the novel-perturbation set (exploratory permutation $p < 10^{-4}$ over cells, Section J.2). Rotation lowers the pattern correlation of the selected channels, consistent with geometric displacement of content across feature-map positions, whereas cutout preserves it (Table S28), so the two failures differ, and the criterion-side variants we tested do not close the gap (Section J.2). Applying the non-negative weight constraint of Stay-Positive to the head, with the same PICS mask, raises AP on this subset to the clamp-only level (Table S29) at the cost of Overall AP on the primary benchmark (Table S18). The excluded channels remain masked, so the improvement comes from the constraint on the classifier weights, whose interaction with these transforms we leave unresolved. PICS without clamp remains the general-purpose configuration, and mask+clamp is an opt-in where such transforms are expected.

Limitations. K is set per backbone and protocol on the primary benchmark and then fixed (Table 5). Learning it jointly with the classifier is future work. A fixed soft weighting by robustness score underperforms the hard mask (Section F.1), but because a positive rescaling can be absorbed by the linear head this does not rule out learned gating. On CLIP, channel selection adds nothing beyond LayerNorm adaptation (Section 5.2). At extreme perturbations all models collapse to near chance, and absolute performance remains limited for generators far from the backbone's pretraining distribution. Audio and video detectors, which expose indexed activations and a retrainable head, are untested next targets.

7 Conclusion

We asked whether the fragility of forensic detectors under post-processing is concentrated in particular feature channels, and found that on a forensic-trained CNN it is. PICS turns this diagnosis into a recipe that scores channels by perturbation stability, keeps the union of the most stable ones, and retrains the linear head, with no added parameters. Under a common protocol it improves robustness beyond augmentation-only retraining and beyond magnitude-based, random, attention-based, weight-clamping, and fine-tuning alternatives, holds on most perturbation families outside the scoring set, and transfers to cross-dataset benchmarks, with an operating-point shift on social media. The recipe extends to a forensic-trained ViT detector, and its gains over the released detectors are largest where robust and discriminative channels are most separated. Perturbation stability is thus a practical criterion for selecting the features of forensic detectors.

AI USE STATEMENT

In this work, we used generative AI tools to help design, implement, and analyze the additional control experiments reported in the paper, and to help interpret their results. Because AI-generated images are the object of study, one evaluation set was generated with a public SDXL-Turbo checkpoint (Section C.5). We did not use generative AI tools to develop theoretical models, mathematical claims, or proofs, or to propose or refine the research hypotheses. Translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to help identify related work, for writing assistance and language editing of the manuscript, and for formatting bibliography entries, which the authors verified against their sources. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The supplementary material includes code for the PICS pipeline, covering channel scoring, channel selection, head training, and evaluation. Channel masks and configuration files for every reported experiment will be released. Sections B, D and E specify the datasets, splits, scoring sets, augmentation schedules, seeds, and the statistical protocol, Section K lists the 51 evaluation conditions, and Sections F.1, F.2 and G give the settings of the control experiments.

REFERENCES

Abhishek Ameta, Sayan Banerjee, Shreyas Pandith, Harshit, Ankita Chatterjee, Akshay Janardan Bankar, and Amit Satish Unde. DRIFT: From robustness gaps to invariance manifolds for AI-generated image detection. *arXiv preprint arXiv:2606.06918*, 2026.

Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024. doi: 10.1109/OJSP.2023.3337714.

Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024. doi: 10.48550/arXiv.2406.09398.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *ECCV*, 2020.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP*, 2023.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of AI-generated image detection with CLIP. In *CVPRW*, 2024a.

Davide Cozzolino, Giovanni Poggi, Matthias Nießner, and Luisa Verdoliva. Zero-shot detection of AI-generated images. In *ECCV*, 2024b.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL `https://proceedings.mlr.press/v235/esser24a.html`.

Tatiana Gaintseva, Laida Kushnareva, German Magai, Irina Piontkovskaya, Sergey Nikolenko, Martin Benning, Serguei Barannikov, and Gregory Slabaugh. Improving interpretability and robustness for the detection of AI-generated images. *arXiv preprint arXiv:2406.15035*, 2024.

Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In *ICME*, 2021.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or JPEG? revealing common biases in generated image detection datasets. In *ECCV Workshops*, 2024.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general AI-generated image detection. In *CVPR*, pp. 18685–18694, June 2025.

Junjie He, Bohua Chen, Yinzhang Ding, and Dongxiao Li. Feature variance ratio-guided channel pruning for deep convolutional network acceleration. In *Asian Conference on Computer Vision (ACCV)*, pp. 170–186, 2020. doi: 10.1007/978-3-030-69538-5_11.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *ICCV*, 2017.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution AI-generated image detection by spectral learning. In *CVPR*, pp. 18706–18717, June 2025.

Dahye Kim, Jaehyun Choi, Hyun Seok Seong, Seongho Kim, Donghun Lee, Sungwon Yi, and Jang-Ho Choi. Dissect and prune: Enhancing robustness in AI-generated image detection. In *International Conference on Machine Learning (ICML)*, 2026.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient ConvNets. In *ICLR*, 2017.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through L_0 regularization. In *International Conference on Learning Representations (ICLR)*, 2018. URL `https://openreview.net/forum?id=H1Y8hhg0b`.

Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. Seeing is not always believing: Benchmarking human and model perception of AI-generated images. *arXiv preprint arXiv:2304.13023*, 2023.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Midjourney. Midjourney. `https://www.midjourney.com/`, 2024.

Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8245–8257, 2025.

Suraj Patil, William Berman, Robin Rombach, and Patrick von Platen. aMUSEd: An open MUSE reproduction. *arXiv preprint arXiv:2401.01808*, 2024.

Pablo Pernías, Dominic Rampas, Mats L. Richter, Christopher Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *ICLR*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Anirudh Sundara Rajan and Yong Jae Lee. Stay-Positive: A case for ignoring real image features in fake image detection. In *ICML*, 2025.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *ICLR*, 2025.

Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023. doi: 10.48550/arXiv.2310.03502.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Rajarshi Roy, Ashhar Aziz, Shashwat Bajpai, Nasrin Imanpour, Gurpreet Singh, Shwetangshu Biswas, Kapil Wanaskar, Parth Patwa, Subhankar Ghosh, Shreyas Dixit, Nilesh Ranjan Pal, Vipula Rawte, Ritvik Garimella, Amitava Das, Amit Sheth, Gaytri Jena, Vasu Sharma, Aishwarya Naresh Reganti, Vinija Jain, and Aman Chadha. A comprehensive dataset for human vs. AI generated image detection. *arXiv preprint arXiv:2601.00553*, 2026.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11531–11539, 2020a. doi: 10.1109/CVPR42600.2020.01155.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *CVPR*, 2020b.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Computer Vision – ECCV 2018*, volume 11211 of *Lecture Notes in Computer Science*, pp. 3–19. Springer, 2018. doi: 10.1007/978-3-030-01234-2_1.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *ICLR*, 2025.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. In *IEEE WIFS*, 2019.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting AI-generated image. In *NeurIPS*, 2023.

A FIGURES AND TABLES REFERENCED FROM THE MAIN TEXT

This section collects the figures and tables that the main text references.

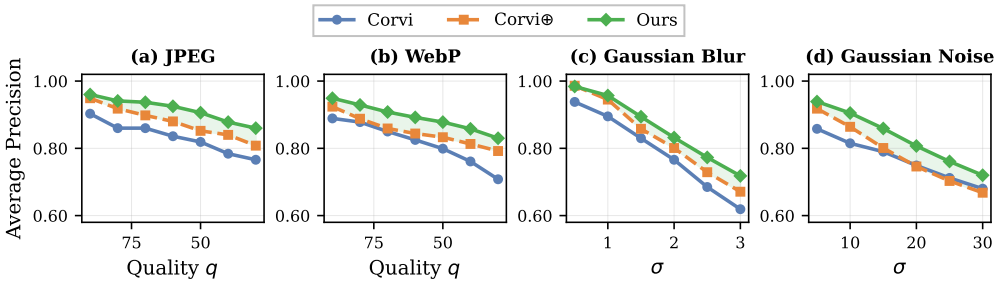

Figure S1: **AP vs. perturbation severity** for four major perturbation categories. PICS (green) consistently outperforms both Corvi (blue) and Corvi⊕ (orange), with the advantage widening at more severe levels. Shaded region denotes gain over Corvi⊕.

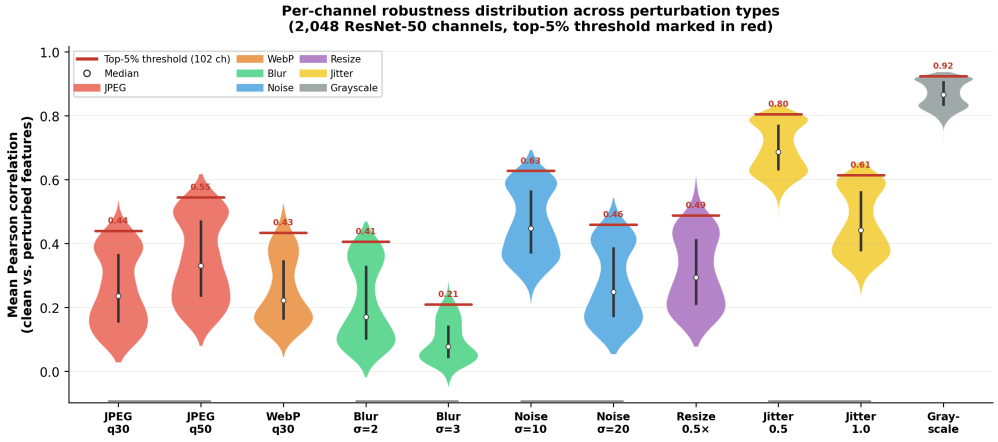

Figure S2: **Per-channel robustness distribution across perturbation categories.** Each violin shows the distribution of mean Pearson correlation between clean and perturbed spatial activations. The red line marks the top-5% threshold (102 channels) for each perturbation category. Destructive perturbations (blur σ=3, JPEG q=30) push most channels to low correlation, while mild perturbations (grayscale, jitter) leave most channels intact.

B IMPLEMENTATION DETAILS

Training hyperparameters. The FC layer is initialized to zero (weights and bias). We train with binary cross-entropy loss using Adam (lr=10^{-4}, β_1=0.9, β_2=0.999) with batch size 512. Early stopping monitors validation balanced accuracy with patience 5 and δ=10^{-3}. Convergence typically occurs within 20–25 epochs.

Aggressive augmentation schedule. Three head-training schedules appear in the paper. All use the same operators in the same order, applied at full resolution before the 96×96 crop: random resized crop (area fraction 0.08–1.0, aspect ratio 0.75–1.33), color jitter (brightness, contrast, saturation 0.4, hue 0.1), random grayscale, cutout (one 48×48 hole), Gaussian noise (variance 10–50), Gaussian blur ($\sigma \in [0, 3]$), compression, and a random rotation by a multiple of 90° (p=1). The *standard* schedule (the schedule of the released detectors) uses JPEG only ($q \in [30, 100]$, p=0.5) with probabilities 0.2/0.8/0.2/0.2/0.2/0.5 for resize/jitter/grayscale/cutout/noise/blur. The *WebP* schedule keeps these probabilities and adds WebP to the compression codecs. The *aggressive* schedule, used for all

Table 1 and Table 2 CNN results unless noted, uses JPEG and WebP with $q \in [10, 100]$ at p=0.8, blur at p=0.8, noise at p=0.5, resize at p=0.5, jitter at p=0.8, grayscale at p=0.3, and cutout at p=0.3. The head-training data are 180k real/fake pairs split into 162k training and 18k validation pairs; one default seed and two replicate seeds are used (values in the code release). Seeds change data order and augmentation sampling only: the FC head is zero-initialized and the channel mask is fixed across seeds.

Perturbation analysis set. For channel selection we use $|\mathcal{A}|$=11 transform settings from seven categories: JPEG compression (q=30, 50), WebP compression (q=30), Gaussian noise (σ=10, 20), Gaussian blur (σ=2.0, 3.0), bicubic resize (0.5×), grayscale conversion, and color jitter (α=0.5, 1.0). The diagnosis of Figure 2 averages robustness over a superset of 18 severity levels of the same seven categories (JPEG q=30, 50, 70; WebP q=30, 50, 70; blur σ=1, 2, 3; noise σ=5, 10, 20; jitter α=0.2, 0.5, 1.0; resize 0.5×, 0.75×; grayscale), which is also the analysis set used for the DINOv2 backbone (Section E). With robustness averaged over the 11-setting selection subset instead, the correlation of Figure 2 is −0.39.

Pearson correlation details. We add ϵ=10^{-12} to the denominator of the Pearson correlation for numerical stability. Channels with constant activation maps (zero variance) are assigned ρ=0 and treated as non-robust. Since the layer-4 output follows ReLU activation, activations are non-negative and negative correlations are rare, so we do not take absolute values. Clean and perturbed crops are paired by a fixed crop seed.

Patch evaluation protocol. For the CNN protocols, test images are center-cropped to 96×96 pixels, yielding a single patch per image (CLIP uses a 224 center crop and B-Free its native 504 resolution). Perturbations are applied to the full image before cropping, and image-space resizing is undone by resampling back to the original size, so clean and perturbed feature maps share the same shape. Average Precision (AP) is computed at the image level using the single-patch logit as the detection score.

Cross-dataset protocol. We evaluate on two cross-dataset benchmarks. **Synthbuster** (Bammey, 2024) uses 1,000 real images from the RAISE-1k dataset and fake images from 9 generators (DALL-E 2/3, Firefly, GLIDE, Midjourney v5, Stable Diffusion 1.3/1.4/2/XL). AP is averaged across generators. **Chameleon** (Yan et al., 2025) is a test-only benchmark of about 26k images (14,863 real, 11,170 AI-generated) whose AI-generated images passed a human perception test and whose real images were retrieved with the same query words, so content and semantics are weak cues.

Baseline reproduction. Both Stay-Positive (Rajan & Lee, 2025) reproductions (Corvi⊕ and Rajan⊕) match the authors' released checkpoints within 0.2 pp Overall AP (0.13 and 0.10 pp), and per-category deviations are at most 0.2 pp for Corvi⊕ and 0.6 pp for Rajan⊕. Appendix tables mark these reproductions with †.

Statistical protocol. The ± values are standard deviations over training seeds (ddof 0). Confidence intervals for gains are paired bootstraps over the benchmark generators: the per-generator differences of seed-mean AP between PICS and the strongest non-PICS row are resampled with replacement (10,000 draws) and the 2.5/97.5 percentiles are reported. The intervals are conditional on the benchmark generators and shared evaluation images. The hotspot test in Section J.2 pools the 126 cells of the novel-perturbation set (14 conditions × 9 generators), permutes the labels of the 15 rectangle cells against the remaining 111 (10,000 permutations, one-sided), and reports a bootstrap interval for the rectangle mean. Cells share generators and images, so exchangeability is approximate, and the rectangle was identified from the full result matrix, so the test is exploratory: it quantifies an observed pattern rather than testing a pre-registered hypothesis, and its p-value should not be read as confirmatory. Ablation tables without ± report the default-seed run (Overall 87.0 for PICS top-5% under the aggressive schedule), whereas rows with ± are three-seed means (86.6).

Telegram round trip. The Telegram condition of Table 3 uses 1,000 real images and 1,000 Würstchen-generated fakes from LDMFakeDetect. Each image is uploaded to and downloaded from Telegram, which re-encodes it as JPEG, and the detector is evaluated on the downloaded copies with

the same 96×96 center-crop protocol as the main benchmark, and the images before upload serve as the clean reference.

Computational cost. Channel masking adds no parameters. For the CNNs the mask multiplies the pooled features and can be absorbed into the head weights, so backbone computation at inference is unchanged, while scoring, the K sweep, and head retraining are offline costs. Backbone parameter counts are 23.5M for ResNet50-NoDown, 427.6M for the full CLIP ViT-L/14 model (about 304M for the visual tower used at inference), and 86.5M for DINOv2 ViT-B/14.

C Per-Generator and Cross-Dataset Results

C.1 Per-Generator Analysis

Table S1 reports per-generator Overall AP (mean over the 51 conditions) for both CNN families, with PICS averaged over training seeds as in Table 1. Channel masking improves all nine test generators in the Corvi family over Corvi$\oplus$, least on PixArt-α and most on Würstchen. Gains are largest for compression and noise perturbations. For the Rajan family, gains are more modest and three generators (aMUSEd, Playground, and LCM) regress slightly against the released Rajan$\oplus$ checkpoint, consistent with the robustness–discriminativeness analysis (Section 3) showing that Rajan's backbone lacks the strong negative association observed for Corvi.

Table S1: **Per-generator Overall AP (%) on LDMFakeDetect, CNN families.** Mean over the 51 conditions per generator. Ours is the mean over three training seeds. Corvi† and Corvi$\oplus^\dagger$ are the reproduced models of Table 1 and Rajan$\oplus$ the released checkpoint. Δ is Ours minus the Stay-Positive row of the same family. Best per row within each family in **bold**.

	Corvi family				Rajan family			
Generator	Corvi†	Corvi$\oplus^\dagger$	Ours (top-5%)	Δ	Rajan†	Rajan$\oplus$	Ours (top-2%)	Δ
Midjourney	81.6	83.1	**86.1**	+3.0	86.5	87.4	**88.4**	+1.0
aMUSEd	72.8	77.1	**80.3**	+3.2	68.2	**76.0**	74.7	−1.3
FLUX	61.7	75.2	**77.0**	+1.8	62.5	65.1	**66.3**	+1.2
Kandinsky	83.3	85.1	**88.1**	+3.1	83.8	87.2	**87.5**	+0.3
LCM	78.0	84.6	**87.3**	+2.8	88.1	**91.4**	91.2	−0.2
PixArt-α	90.9	94.6	**95.7**	+1.1	94.8	96.0	**96.1**	+0.1
Playground	69.7	84.4	**85.7**	+1.3	79.4	**87.8**	87.2	−0.6
Stable Diffusion	89.4	88.7	**90.3**	+1.6	90.0	90.4	**90.7**	+0.3
Würstchen	86.5	84.0	**88.9**	+4.9	68.7	65.8	**69.9**	+4.1
Mean	79.3	84.1	**86.6**	+2.5	80.2	83.0	**83.6**	+0.6

C.2 Cross-Generator Robustness of Channel Robustness

A natural concern is whether per-channel robustness measured on the *training* distribution (autoencoder-reconstructed fakes) transfers to images produced by entirely different generators. To test this, we compute the mean Pearson robustness of every channel on both the training set and the test set (9 unseen latent generators from LDMFakeDetect, eight diffusion or flow models and the masked-image model aMUSEd) and compare the two rankings.

Figure S3 plots per-channel robustness measured on the training set against robustness measured on the test set for the Corvi backbone with paired AE fakes (2,048 channels). The two measurements are strongly rank-correlated (Figure S3), indicating that the relative ordering of channels by robustness is largely preserved across generator distributions. Moreover, two thirds of the top-5% channels selected from the training set also appear in the top-5% when ranked by test-set robustness.

This result suggests that channel robustness is largely a property of the backbone architecture rather than an artifact of the specific generator used during analysis. Because the training set contains only autoencoder fakes while the test set spans nine unseen generators, the high rank correlation indicates that channel selection performed on a single, convenient source of fake data transfers well within the

latent-diffusion generator family without re-analysis. Architecturally distant generators (e.g., GANs) may exhibit different robustness profiles. Evaluating this remains future work.

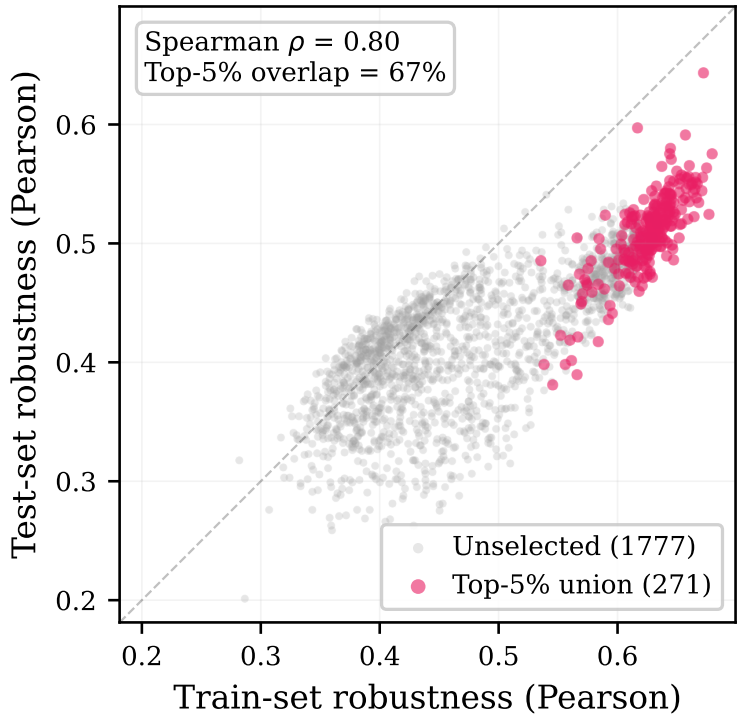

Figure S3: **Cross-generator robustness of per-channel robustness** (Corvi backbone, 2,048 channels). Each point is one channel. The x-axis shows mean Pearson robustness measured on training-distribution fakes (autoencoder), and the y-axis shows robustness measured on 9 unseen test-set generators. Pink: channels in the top-5% union mask. Spearman ρ_s=0.80, and 67% of the training-selected top-5% channels remain in the test-set top-5%.

C.3 Cross-Dataset Clean AP

Table S2 reports clean AP on the held-out cross-dataset benchmark for each backbone group: GenImage's 7 distinct generators (adm, biggan, glide, midjourney, sd, vqdm, wukong) for Corvi/Rajan/B-Free (whose in-distribution evaluation uses LDMFakeDetect), and LDMFakeDetect's 9 generators for UFD (whose in-distribution evaluation uses GenImage).

PICS improves cross-dataset clean AP on Corvi and matches B-Free, and on Rajan it lands just below Rajan⊕ (Table S2). The CLIP/UFD comparison in this table uses the released UFD head (trained on ProGAN), whereas the in-distribution rows of Table 1 use a head retrained on GenImage/BigGAN, so this comparison is corpus-confounded and is reported for completeness rather than as a head-to-head. A sweep over $K \in \{3, 5, 7, 10, 15, 20, 25\}\%$ does not change the conclusion (best variant union-7% reaches 67.9 AP, still below UFD's 70.4). On the CLIP backbone, the masked LayerNorm+FC head improves over the FC-only head in-distribution, but the unmasked control in Table 1 shows that this gain comes from LayerNorm adaptation rather than channel selection, and the masked head does not transfer cross-dataset.

C.4 Perturbed Cross-Dataset Evaluation

To verify that robustness improvements transfer to external datasets, we run the full 51-condition perturbation sweep on both cross-dataset benchmarks for the three Corvi-family models. Tables S3 and S4 report per-category results (clean AP is included as the first row of each table).

Table S2: **Cross-dataset clean AP (%)** per backbone group. Each row's mask budget matches the in-distribution Table 1 configuration. **Bold** marks the best within each backbone group. † marks our reproductions. Rajan⊕ here is our Stay-Positive reproduction, whereas Table 1 uses the released checkpoint.

Method	Held-out benchmark	Clean AP
ResNet50-NoDown (Corvi). Held-out: GenImage 7-gen		
Corvi†	GenImage 7-gen	89.9
Corvi⊕†	GenImage 7-gen	90.8
Ours: top-5%	GenImage 7-gen	**91.9**
ResNet50-NoDown (Rajan). Held-out: GenImage 7-gen		
Rajan†	GenImage 7-gen	85.9
Rajan⊕†	GenImage 7-gen	**88.0**
Ours: top-2%	GenImage 7-gen	87.5
CLIP ViT-L/14 (UFD). Held-out: LDMFakeDetect 9-gen		
UFD (released head)	LDMFakeDetect 9-gen	**70.4**
Ours: top-20%	LDMFakeDetect 9-gen	65.2
DINOv2 ViT-B/14 (B-Free). Held-out: GenImage 7-gen		
B-Free†	GenImage 7-gen	**95.2**
Ours: top-7%	GenImage 7-gen	**95.2**

Table S3: Perturbed cross-dataset evaluation on **Synthbuster**. AP (%) averaged over severity levels within each category. Best per row in **bold**.

Category	# Cond.	Corvi	Corvi⊕	Ours (top-5%)
Clean	1	86.2	87.1	**88.9**
JPEG	7	83.8	82.8	**85.7**
WebP	7	76.7	79.4	**81.6**
Blur	6	80.6	**82.0**	81.0
Noise	6	78.3	77.4	**81.0**
Jitter	5	83.9	86.2	**87.6**
Resize	3	72.0	73.9	**75.0**
Grayscale	1	85.3	86.7	**88.0**
Combo	15	79.2	78.7	**79.7**
All perturbed	50	79.7	80.2	**81.8**
Overall	51	79.8	80.4	**81.9**

PICS leads in 8 of 9 categories on Synthbuster and wins overall on both benchmarks. On Chameleon, the pretrained Corvi baseline remains competitive on noise and combo perturbations, possibly because its broader channel set retains more signal under severe corruption, a hypothesis the table does not test. Across both benchmarks, the perturbation-category ranking mirrors the in-distribution findings.

C.5 Held-Out Generators (20 in Total, Seven Source Groups)

To further stress-test cross-dataset generalization, we construct an extended benchmark spanning **20 generators** drawn from seven source groups (six external sources plus locally generated SDXL-Turbo images). All fakes are paired with the same RAISE-1k real images, so source-specific content and preprocessing cues are not separated from generator architecture. The benchmark is a stress test, not an isolation of architecture effects. The 9 original Synthbuster generators (Bammey, 2024) (DALL-E 2/3, Firefly, GLIDE, Midjourney v5, SD 1.3/1.4/2/XL) are augmented with 11 additional generators covering a broad range of architectures and release periods: Wukong from GenImage (Zhu et al., 2023), StyleGAN3 and CogView2 from Fake2M (Lu et al., 2023), SD 3 and Midjourney v6 from the MS COCOAI dataset of Defactify 4.0 (Roy et al., 2026), PixArt-α (Chen et al., 2024) from its official evaluation set, SDXL-Turbo (Sauer et al., 2023) (1,000 images generated locally with the public checkpoint from 30 generic photographic prompts), and four generators from CommunityForensics-Eval (Park & Owens, 2025): DeciDiffusion v2, Firefly Image 2, Ideogram v2,

Table S4: Perturbed cross-dataset evaluation on **Chameleon**. AP (%) averaged over severity levels within each category. Best per row in **bold**.

Category	# Cond.	Corvi	Corvi⊕	Ours (top-5%)
Clean	1	67.6	67.8	**69.4**
JPEG	7	65.0	62.3	**65.1**
WebP	7	47.5	54.8	**57.6**
Blur	6	55.3	**60.0**	56.4
Noise	6	**56.9**	49.7	53.8
Jitter	5	63.6	63.6	**65.3**
Resize	3	49.5	**51.1**	**51.1**
Grayscale	1	69.4	67.9	**70.1**
Combo	15	**55.5**	53.9	54.9
All perturbed	50	56.6	56.5	**57.9**
Overall	51	56.8	56.7	**58.1**

and Imagen 3. Each generator contributes 1,000 fake images (999 for Ideogram v2), paired with the same 1,000 RAISE-1k real images used in the original Synthbuster protocol. The collection spans GANs (StyleGAN3), autoregressive models (CogView2), pixel-space diffusion (GLIDE), latent diffusion models (SD 1.3/1.4/2/XL, SDXL-Turbo, DeciDiffusion v2, Wukong), DiT/flow-matching models (SD 3, PixArt-α, Imagen 3), and proprietary commercial generators (DALL-E 2/3, Midjourney v5/v6, Firefly/Firefly Image 2, Ideogram v2). None of these 11 additional generators appear in the head-training data or in the Corvi pretraining data.

Table S5 reports per-generator clean AP. All three models struggle most on DALL-E 2/3, while achieving near-perfect detection on SD variants and most modern generators. Our model achieves the highest macro-average, indicating transfer to this heterogeneous collection of generators and sources.

Table S6 extends this evaluation to the full 51-condition perturbation sweep, showing that the robustness improvements observed in-distribution carry over to this broader cross-dataset setting with 20 generators.

D CLIP ViT-L/14 Extension Details

This section provides implementation details and additional results for the UniversalFakeDetect (Ojha et al., 2023) extension reported in the main paper (Section 5.2).

D.1 Architecture and Masking Points

UniversalFakeDetect uses a frozen CLIP ViT-L/14 (Radford et al., 2021) backbone that produces 257 tokens (1 CLS + 256 patch tokens) of dimension 1,024 from the final transformer block. The CLS token passes through a final LayerNorm (`ln_post` in the CLIP implementation) and a linear projection ($1{,}024 \to 768$), then through a single FC layer ($768 \to 1$) for binary classification.

We investigated three masking points: **Point A** (primary): a binary mask on the 1,024 channels of the last transformer block's output, applied by a forward hook to all 257 tokens (the CLS token and the 256 patch tokens) before the final LayerNorm, the projection, and the FC head. The CLS readout is therefore masked before `ln_post`, and the channels are scored on the 256 patch tokens of the same tensor. **Point B**: mask on the 1,024-dim CLS token after the final LayerNorm. **Point C**: mask on the 768-dim CLS token after projection. Point A performed best in preliminary runs (not shown) and is used for all main results.

D.2 Channel Robustness Analysis

We performed the same Pearson-correlation-based analysis as for ResNet-50, treating the 256 patch-token positions of the last block's output as the analogue of the 12×12 CNN feature maps, using 500 images per class from the GenImage/BigGAN validation set.

Table S5: **Held-out generators: per-generator clean AP (%).** 20 generators from seven source groups, all paired with the same 1,000 RAISE-1k real images (the original Synthbuster protocol). The original Synthbuster benchmark contributes 9 generators. The additional 11 are sourced from CommunityForensics-Eval, GenImage, Defactify, Fake2M, the official PixArt-α evaluation set, and locally generated SDXL-Turbo images. Best per row in **bold**.

Generator	Corvi	Corvi$\oplus$	Ours (top-5%)
Synthbuster (Bammey, 2024) (9 generators)			
DALL-E 2	**61.8**	50.6	61.4
DALL-E 3	51.8	**61.8**	61.4
Firefly	89.5	89.5	**95.0**
GLIDE	75.3	**83.0**	**83.0**
Midjourney v5	99.4	99.7	**99.9**
SD 1.3	**100.0**	**100.0**	**100.0**
SD 1.4	**100.0**	**100.0**	**100.0**
SD 2	99.5	**99.8**	**99.8**
SD XL	98.1	99.1	**99.7**
CommunityForensics-Eval (Park & Owens, 2025) (4 generators)			
DeciDiffusion v2	99.7	99.6	**99.9**
Firefly Image 2	**95.4**	87.1	95.3
Ideogram v2	95.4	**98.6**	98.1
Imagen 3	98.2	99.0	**99.5**
GenImage (Zhu et al., 2023) (1 generator)			
Wukong	**100.0**	99.9	**100.0**
MS COCOAI (Defactify 4.0) (Roy et al., 2026) (2 generators)			
Midjourney v6	87.3	81.4	**91.2**
SD 3	93.3	86.4	**94.1**
Other sources (2 generators)			
PixArt-α (Chen et al., 2024)	96.9	98.5	**98.6**
SDXL-Turbo (Sauer et al., 2023)	94.5	97.6	**98.2**
Fake2M (Lu et al., 2023) (2 generators)			
StyleGAN3	**75.7**	69.1	75.6
CogView2	**100.0**	99.6	99.9
Macro average (20 gen.)	90.6	90.0	**92.5**

Table S6: **Held-out 20-generator set: perturbed cross-dataset evaluation.** AP (%) averaged over severity levels within each category, across all 20 held-out generators. Best per row in **bold**.

Category	# Cond.	Corvi	Corvi$\oplus$	Ours (top-5%)
Clean	1	90.6	90.0	**92.5**
JPEG	7	85.4	83.3	**86.4**
WebP	7	76.6	80.2	**84.6**
Blur	6	81.6	**84.1**	81.2
Noise	6	78.4	74.9	**80.4**
Jitter	5	87.2	88.0	**90.4**
Resize	3	72.3	73.3	**74.5**
Grayscale	1	89.8	89.2	**91.3**
Combo	15	78.7	78.2	**78.8**
All perturbed	50	80.4	80.4	**82.3**
Overall	51	80.6	80.6	**82.5**

The headline union-20% mask is computed from five single-level transforms (JPEG compression, Gaussian blur, Gaussian noise, color jitter, and horizontal flip; flipped maps are compared without realignment, so that setting scores the left-right symmetry of a channel's pattern rather than its stability) on GenImage/BigGAN validation images. Richer analysis sets (seven categories at one level, or the 18-level set of Section B) change the Clean and Pert. AP of a union-10% mask by at most

0.9 pp under an earlier evaluation protocol that applied 19 perturbation settings on the fly (superseded by the 51-condition protocol).

D.3 Masking Threshold ($K\%$)

Unlike ResNet-50, where $K{=}5\%$ is optimal, the CLIP sweep is flat above 10% and no K exceeds the unmasked LayerNorm+FC head (Tables 1 and S7). This is consistent with general-purpose features distributing forensic information more evenly across channels.

Table S7: Masking threshold $K\%$ for CLIP ViT-L/14 (LayerNorm+FC training, eight GenImage generators). "Pert." is the 35-cell single-perturbation mean, "Combo" the 15-cell multi-step mean, and "Overall" the mean across all 51 conditions (AP in %). Best per column in **bold**.

$K\%$	Ch.	Clean	Pert.	Combo	Overall
5	127	88.4	72.0	68.8	71.4
7	177	90.1	73.7	70.0	72.9
10	249	92.3	74.9	**70.5**	74.0
15	331	94.3	75.3	69.8	74.1
20	**423**	**94.7**	**75.7**	70.0	**74.4**
25	511	94.6	75.4	69.9	74.1

D.4 Training Configuration

Table S8 lists the training hyperparameters used for the CLIP ViT-L/14 experiments.

Table S8: Training hyperparameters for CLIP ViT-L/14 experiments.

Parameter	Value
Backbone	CLIP ViT-L/14 (frozen)
Trainable layers	Final LayerNorm (`ln_post`) + FC (2,817 params)
FC initialization	Zero-initialized
Optimizer	AdamW (lr=10^{-3}, no weight decay)
Scheduler	Cosine ($10^{-3} \rightarrow 10^{-6}$)
Epochs	50
Batch size	128
Training data	GenImage/BigGAN: 10k real + 10k fake
Validation data	GenImage/BigGAN: 1k real + 1k fake
Input augmentation	RandomResizedCrop(224), RandomHorizontalFlip

D.5 FC-Only vs. LayerNorm+FC Training

Unlike ResNet-50, FC-only training with channel masking leads to severe overfitting on CLIP: validation AP peaks early and collapses to 61.7 Clean / 51.5 Pert. AP by epoch 50 under the earlier 19-setting on-the-fly protocol, far below the unmasked UFD head (85.1 / 72.1 on the 51-condition protocol, so this comparison across two protocols is qualitative). Co-training the final LayerNorm with the FC head removed the collapse in the tested configuration, as its gain and bias can adapt to the masked representation: the union-10% mask then reaches 92.3 Clean / 74.9 Pert. AP on the 51-condition protocol. The same LayerNorm+FC training without any mask reaches 95.1 Clean / 76.5 Pert. AP over three seeds (Table S9), matching the masked heads, so on CLIP the adaptation rather than the mask carries the gain over the FC-only head.

This highlights an important practical consideration: when masking removes a large fraction of channels, downstream normalization layers may need to recalibrate. For the forensic-trained ResNet-50, the FC layer alone suffices because the preceding batch normalization operates per-channel and is unaffected by zeroed channels. For the general-purpose CLIP backbone, the final LayerNorm operates across all channels, so masking alters its statistics, and adapting it removed the collapse in the tested configuration.

D.6 AUGMENTATION TRAINING ON CLIP

Augmented training applies random perturbations (JPEG, WebP, blur, noise, jitter, resize) with probability 0.5 during training. On CLIP, augmented training improves perturbed and combo AP. For the masked head it also costs clean AP, whereas for the unmasked head it raises clean AP as well (Table S9). The rows differ in trainable layers, as the caption states. The unmasked LayerNorm+FC head matches the masked ones, so channel selection adds nothing measurable on CLIP at matched adaptation and data; the released ProGAN-trained head scores above the FC-only head retrained on BigGAN. Combining augmentation with masking yields the best perturbed and combo AP in this table, but without an unmasked LayerNorm+FC head trained with augmentation the mask's contribution under augmentation is untested, and the clean AP trade-off remains.

Table S9: Augmentation ablation on CLIP ViT-L/14 (eight GenImage generators, 51-condition protocol). The first block trains the FC only (LayerNorm frozen) on GenImage/BigGAN, without and with augmentation. The second block co-trains the final LayerNorm with the FC head on the same data: without a mask, with the union-20% mask, and with the mask plus augmentation. The last row is the released UFD head evaluated under the same protocol. Best per column in **bold**. The unmasked LayerNorm+FC row is the mean over three seeds. The other rows are single runs.

Aug	Mask	Clean AP	Pert. AP	Combo AP
		85.1	72.1	65.8
✓		88.5	76.5	71.8
LayerNorm+FC head, trained on GenImage/BigGAN				
		95.1	76.5	71.0
	✓	94.7	75.7	70.0
✓	✓	89.8	**76.8**	**72.4**
Released UFD head (FC trained on ProGAN), reference				
		87.8	74.3	70.9

D.7 PER-GENERATOR CLEAN AP BREAKDOWN

Table S10 shows per-generator clean AP. The Δ column is measured against the FC-only head and therefore includes the LayerNorm-adaptation effect, which the matched control in Table 1 shows to account for the margin, so it is not a selection effect.

Table S10: Per-generator clean AP on GenImage (CLIP ViT-L/14, 8 generators). Union-20% is a single training run (seed means in Table 1). Δ = Union-20% − Baseline.

Generator	Baseline	Union-20%	Aug., no mask	Δ
BigGAN	98.6	**99.9**	99.1	+1.4
ADM	75.8	**91.4**	78.1	+15.6
GLIDE	96.2	**99.6**	96.3	+3.4
Midjourney	85.8	**89.0**	85.1	+3.1
SD v1.4	77.6	**93.4**	84.3	+15.8
SD v1.5	78.6	**93.1**	84.2	+14.5
VQDM	85.2	**96.8**	94.7	+11.6
Wukong	83.1	**94.0**	85.8	+11.0
Mean (8 gen.)	85.1	**94.7**	88.5	+9.5

D.8 COMPARISON: FORENSIC-TRAINED VS. GENERAL-PURPOSE BACKBONE

The forensic-trained ResNet-50 backbones and the general-purpose CLIP ViT-L/14 backbone differ in training objective (real/fake classification vs. image-text contrastive), feature type (layer-4 maps vs. patch tokens), and channel count (2,048 vs. 1,024). The core Pearson-correlation-based channel analysis and union selection transfer directly. The main adaptations are the wider mask fraction and the co-training of the final LayerNorm, which on CLIP carries the gain by itself (Section 5.2). The

forensic-trained B-Free ViT behaves like the CNNs (Section E), which suggests that the training objective rather than the CNN-vs-transformer architecture drives the difference, but objective, architecture, preprocessing, and normalization all differ between these backbones, so the source of the difference is not isolated.

E B-FREE (DINOV2 VIT-B/14) EXTENSION DETAILS

This section provides implementation details and additional results for the B-Free (Guillaro et al., 2025) extension reported in the main paper (Section 5.2).

E.1 ARCHITECTURE AND MASKING IMPLEMENTATION

B-Free fine-tunes a DINOv2 ViT-B/14 (Oquab et al., 2024) backbone end-to-end for binary real/fake classification. The backbone produces 768-dimensional patch tokens from the final transformer block. We apply the binary channel mask to the output of the final layer normalization for all tokens, zeroing out perturbation-sensitive dimensions before classification. The classifier reads the CLS token of this masked output, so the classified CLS vector and the patch tokens used for scoring share the same 768 channel indices. As in the CNN protocol, we freeze the trained B-Free backbone and train only the single-layer linear classifier (769 parameters: 768 weights + 1 bias). Unlike CLIP ViT-L/14, no normalization layer co-training was needed in the tested configuration. Two possible explanations, which we did not ablate, are: (1) the B-Free backbone has already been fine-tuned on forensic data, so the final layer normalization already produces well-adapted features, and (2) there is no learned projection layer between the backbone output and the classifier, unlike CLIP's LayerNorm-then-projection head, so masking does not create a mismatch with a downstream projection that expects all channels.

E.2 CHANNEL ROBUSTNESS ANALYSIS

We apply the same Pearson-correlation-based per-channel analysis described in Section 3 to DINOv2 ViT-B/14, on $1{,}000$ images from the B-Free training set. Features are extracted from the final layer normalization output (768 dimensions). We use the per-patch spatial tokens (discarding the CLS and register tokens) for the per-channel correlation computation. We evaluate robustness over the same seven perturbation categories used for ResNet (JPEG, WebP, blur, noise, jitter, resize, grayscale), at the 18 severity levels listed in Section B. The analysis shows that a subset of dimensions maintains high spatial correlation under perturbation, so the channel-selection pipeline applies without modification.

E.3 TRAINING CONFIGURATION

Table S11 summarizes the training hyperparameters for both the B-Free baseline reproduction and our channel-masked model.

E.4 MASKING THRESHOLD ($K\%$)

Table S12 shows the full threshold sweep. Performance shows a broad peak in K over the tested range: selecting too few channels (1%) discards discriminative signal and drops below the baseline, while selecting too many (20%) retains perturbation-sensitive channels and yields no Overall improvement at the reported precision. The optimal threshold is $K{=}7\%$ (76 of 768 channels), which achieves the highest overall AP across all 51 evaluation conditions. Compared to CLIP ViT-L/14 ($K{=}20\%$, Table S7), B-Free requires a much narrower mask, consistent with the observation that forensic-trained backbones concentrate discriminative signal into fewer channels.

E.5 PER-GENERATOR AP BREAKDOWN

Table S13 reports per-generator overall AP (averaged across all 51 conditions) for the B-Free baseline and our top-7% masked model. The largest gain appears on FLUX (Black Forest Labs, 2024), the generator on which the baseline struggles most, suggesting that channel masking helps expose more transferable cues for unseen generators. Most other generators already operate near saturation (>99%

Table S11: Training hyperparameters for DINOv2 ViT-B/14 experiments. B-Free† is our reproduction of the baseline. Ours applies channel masking on the frozen B-Free† backbone.

Parameter	B-Free†	Ours
Backbone	DINOv2 ViT-B/14	DINOv2 ViT-B/14 (frozen)
Trainable layers	All (86.5M params)	Linear head only (769 params)
Channel mask	—	Union-of-top-$K\%$
Head initialization	Pre-trained	Zero-initialized
Optimizer	Adam	Adam
Learning rate	10^{-6}	10^{-4}
Weight decay	10^{-6}	10^{-6}
Epochs	100	50
Early stopping	patience 5	patience 5
Batch size	24	24
Training data	B-Free official (51k COCO real + 309k SD 2.1 fake)	
Augmentation	*inpainted*++ (blur, JPEG, cutout, noise, jitter, crop, flip)	
Input resolution	504×504	

Table S12: Masking threshold $K\%$ for DINOv2 ViT-B/14 (B-Free). Pearson union, frozen backbone. "Pert." is the 35-cell single-perturbation mean, "Combo" the 15-cell multi-step mean, and "Overall" the mean across all 51 conditions (AP in %). Best per column in **bold**.

$K\%$	Ch.	Clean	Pert.	Combo	Overall
1	8	96.0	94.0	93.5	93.9
3	31	96.2	94.4	94.1	94.4
5	52	**97.6**	**96.0**	**95.6**	95.9
7	**76**	**97.6**	**96.0**	**95.6**	**96.0**
10	109	97.5	95.7	95.3	95.6
15	167	97.1	95.3	94.9	95.2
20	221	97.0	95.2	94.8	95.1
100	768	96.9	95.1	94.7	95.1

AP), leaving little room for improvement. By perturbation category, the largest gains are on resize ($+1.8$ pp) and blur ($+1.0$ pp), the two most destructive spatial transforms, whereas on Corvi the largest gains are on compression and blur (Table S18).

Table S13: Per-generator overall AP on LDMFakeDetect (DINOv2 ViT-B/14, B-Free). Overall AP (%) is the mean across all 51 conditions (clean and 50 perturbed). Δ denotes the difference between Ours (top-7%) and B-Free baseline. Best per row in **bold**.

Generator	B-Free	Ours (top-7%)	Δ
Midjourney	98.4	**98.5**	+0.1
aMUSEd	**100.0**	**100.0**	0.0
FLUX	60.0	**68.0**	+8.0
Kandinsky	**99.4**	**99.4**	0.0
LCM	**99.7**	**99.7**	0.0
PixArt-α	**99.9**	**99.9**	0.0
Playground	**100.0**	99.9	-0.1
Stable Diffusion	98.5	**98.6**	+0.1
Würstchen	**99.7**	**99.7**	0.0
Mean	95.1	**96.0**	+0.9

F CONTROLS FOR THE SELECTION CRITERION

F.1 COUNT-MATCHED CRITERION CONTROLS

Table 2 compares selection rationales under the aggressive-augmentation protocol. Here every alternative is held to the same channel count (Table 4) as PICS (within two channels for the logit criterion, whose union-of-top-k rule cannot hit 271 exactly) and trained under the identical WebP-augmentation schedule, so the only difference is which channels are kept.

The joint product, which multiplies robustness by clean discriminativeness $A = \max(\mathrm{AUC}, 1 - \mathrm{AUC})$, shares only 26 of 271 channels with the Pearson mask, is skewed toward clean-discriminative channels (median A of 0.95 against 0.73), and scores lowest. The logit criterion preferentially selects near-zero-weight channels (median $|w|$ of $1.9{\times}10^{-2}$ against $3.9{\times}10^{-2}$). Five further joint robustness-and-AUC criteria at budgets from 49 to 133 channels score between 83.1 and 84.6 under the same schedule, and a sixth 356-channel joint criterion under the standard schedule scores 83.9 against 85.6 for Pearson. The gain therefore comes from which channels are kept rather than from how many. A fixed soft mask that scales every channel by a normalized robustness score (z-scored average Pearson correlation rescaled to $[0, 1]$, with the lowest 10% of channels zeroed) reaches 84.0 Overall under the same WebP schedule, 2.4 pp below the hard top-5% mask. Because a fixed positive rescaling can be absorbed by the linear head, this difference reflects the zeroed decile and optimization rather than a general limitation of soft selection, and it does not rule out learned gating.

Decision-level stability. Pearson robustness is a spatial-pattern criterion, whereas the classifier consumes the pooled value. On the $2{,}048$ Corvi channels we therefore also computed, over the 11 scoring transforms and $5{,}000$ images, the relative change of the pooled activation, $|\mathrm{GAP}(g_c) - \mathrm{GAP}(f_c)|/\max(\mathrm{GAP}(f_c), \epsilon)$, and its absolute change. Defining relative GAP stability as the negative relative change, Pearson robustness correlates positively with it (Spearman $\rho = +0.70$) and also positively with the absolute GAP change ($\rho = +0.61$), because absolute changes scale with activation magnitude and the pattern-stable channels tend to be the strongly activated ones. Selection criteria built directly on decision-level stability underperform Pearson at matched channel counts (Table 4), and the retrained head down-weights selected channels with low clean discriminativeness: of the 271 selected channels, the 32 with $A < 0.55$ carry a median $|w|$ of 0.032 against 0.070 for the rest ($2.2\times$, Mann–Whitney $p < 10^{-3}$, aggressive-schedule head). Selection provides stability and the retrained head restores class separability, but the criterion carries no guarantee of decision-level stability. Clean-trained classifier weight magnitudes are only weakly correlated with channel robustness (Spearman $\rho = +0.13$ with the robustness scores of Section 3), so robustness is not well predicted by the classifier's weight magnitudes.

F.2 END-TO-END FINE-TUNING CONTROLS

This section reports the end-to-end fine-tuning baseline summarized in Section 5.3 and the controls that address under-tuning as its explanation. All numbers follow the main protocol (LDMFakeDetect, 9 generators, 51 conditions, mean AP, $\pm$ over three training seeds, paired bootstrap over generators for confidence intervals).

Setup. Starting from the Corvi detector of Table 1, we unfreeze all 23.5M backbone parameters and train on the same 162k training pairs, the same aggressive augmentation schedule, and the same early-stopping protocol used to train the PICS head, with backbone learning rate 10^{-5} and batch size 128. Each seed converges on its own validation set (best validation balanced accuracy 99.4% after 13 epochs for the first seed). We then run the unchanged PICS pipeline on each fine-tuned checkpoint, with $K{=}5\%$ carried over from the original backbone as a transfer test, and compare every PICS-on-fine-tuned run against its own backbone's fine-tuned baseline.

Results. End-to-end fine-tuning lands below PICS on the frozen backbone (Table S14, 95% CI on the gap $[+1.9, +6.7]$) with a larger seed spread (std 0.59 vs 0.33), and it also trails the frozen-backbone no-selection control. Per fine-tuned backbone, PICS adds $+0.31/+0.04/+0.91$ pp (mean $+0.4$, 95% CI $[-0.1, +1.0]$), positive in each case but not statistically conclusive. The K sweep on the fine-tuned backbone is flat across 2–10% (gains of $+0.10/+0.31/+0.94/+0.68$ pp at $K{=}2/5/7/10\%$ on the first backbone, nominal optimum 7%), so the carried-over $K{=}5\%$ did not

Table S14: **Decomposition at matched perturbation training on the Corvi backbone.** Rows 2 and 5 isolate selection at the same head-training schedule. ± is the std over training seeds.

Configuration	Clean	Pert.	Combo	Overall
Corvi detector (Table 1), starting point	92.7	81.3	73.9	79.3
Perturbation-trained FC on the original backbone, no selection	97.6±0.3	85.4±0.1	78.6±0.2	83.6±0.1
End-to-end fine-tuned backbone, no selection	95.9±0.5	84.2±0.5	77.6±1.0	82.5±0.6
PICS on each fine-tuned backbone (K=5% carried over)	94.8±1.2	84.4±0.3	78.5±0.7	82.9±0.4
PICS on the original backbone (ours)	97.9±0.2	88.0±0.3	82.6±0.4	**86.6±0.3**

need retuning. Measured on the fine-tuned checkpoints, the strong anti-correlation between channel robustness and discriminativeness that Section 3 reports for the original backbone is absent, which is the regime in which the diagnosis predicts small selection gains.

Learning rate and training length. Holding the seed fixed, the fine-tuned Overall AP decreases monotonically with the backbone learning rate: 83.5 at 10^{-6} (indistinguishable from the frozen-backbone control, as expected when the backbone barely moves), 83.2 at 10^{-5}, and 82.5 at 10^{-4}, with Clean AP falling in step (97.3/96.6/95.3). Tripling the early-stopping patience makes it worse (81.9, with the run continuing to epoch 36 without improving on its epoch-6 checkpoint). Within the tested learning rates and stopping settings the deficit persists, and we do not claim to have exhausted the fine-tuning design space (optimizer, learning-rate ratio, batch size, and regularization all differ from the head-only recipe, including a batch size of 128 against 512). Partial adaptation shows the same direction: unfreezing only the last residual block on top of the PICS mask under the WebP protocol scores 85.4 Overall against 86.4 for the frozen-backbone head.

Second family, held-out perturbations, and deployment. On the Rajan backbone, end-to-end fine-tuning under the same protocol reaches 81.0 Overall against 83.6 ± 0.0 for PICS (95% CI [+1.5, +3.9]). On the 14 held-out perturbation conditions used for the hotspot analysis (six families including cutout and rotate, 9-generator protocol), the fine-tuned model scores 94.9/93.6/93.6 mean AP across seeds and 92.5/90.8/90.6 on cutout and rotate, below PICS (96.7 and 94.5) in every seed. On the hotspot protocol of Section J.2 (three hardest generators) it is the weakest of the three models (84.2 against PICS 86.7 and Corvi⊕ 92.3), so backbone adaptation does not offer a route out of the scope boundary. On WildRF the fine-tuned model reaches a mean AP of 85.7, comparable to PICS, but a fake accuracy of 60.2 at threshold 0 against 71.5 for PICS.

G Threshold-Level Metrics and Calibration

Average precision is threshold-free. To characterize the operating point, we compute FPR at 95% TPR, TPR at 1% FPR, balanced accuracy at logit threshold 0 (probability 0.5), and expected calibration error (ECE, 15 equal-width bins) from per-image scores. Every metric is computed per generator on the balanced 3,000 real / 3,000 fake split and averaged over the 9 generators, the protocol of Table 1, over the 50 perturbed conditions. Thresholds for FPR@95TPR and TPR@1%FPR are swept within each generator and condition as the definitions require, and no deployment threshold is fitted on test labels.

Table S15: **Threshold-level metrics on perturbed LDMFakeDetect (Corvi family).** Mean over 9 generators and 50 perturbed conditions. PICS improves every operating-characteristic metric. The Rajan family replicates with smaller margins (PICS-Rajan 53.4/40.1/74.9, best on all three).

Model	FPR@95TPR ↓	TPR@1%FPR ↑	BAcc@thr0 ↑
Corvi	61.7	31.5	67.5
Corvi⊕	61.9	36.1	74.3
No-mask aggressive aug	58.4	39.1	71.3
End-to-end fine-tuning	57.0	38.3	73.1
PICS	**51.4**	**41.8**	**77.5**

Table S16: **Expected calibration error on perturbed LDMFakeDetect.** Lower is better. The right column applies one temperature fitted by NLL on the perturbed scores.

Model	ECE	ECE after one fitted temperature
PICS	0.157	**0.113**
Corvi⊕	0.132	0.122
No-mask aggressive aug	0.130	—
End-to-end fine-tuning	0.106	—
PICS on the fine-tuned backbone	**0.103**	—

The ranking gains carry over to operating-characteristic metrics (Table S15), and augmentation alone does not deliver them (identical schedule, no mask). Probability calibration is the one metric PICS does not improve: its ECE trails Corvi⊕ (Table S16). The deficit is consistent with under-confidence, since the NLL-optimal temperature for PICS is 0.47, and a single global temperature fitted by NLL on the perturbed scores removes the gap: 0.113 when fitted and evaluated on the same scores (in-sample, the value in Table S16), 0.114 when fitted on half of the conditions and evaluated on the other half, and 0.114 under leave-one-generator-out fitting (Corvi⊕: 0.122 and 0.123). The same images appear under every condition, so these are condition- and generator-transfer results, not independent-image calibration. Temperature scaling leaves rankings and the threshold-0 decision unchanged, so it explains none of the AP or accuracy gains. The best-calibrated model in the set is PICS applied to the fine-tuned backbone, so masking does not inherently degrade calibration. We treat probability calibration as an orthogonal post-hoc scaling step.

H Ablation Studies

H.1 Channel Selection Metric under Aggressive Augmentation

Section 5.5 in the main paper summarizes the channel selection metric comparison. Here we report it in full: all four masked variants and the no-mask baseline are retrained under aggressive augmentation, the setting used in all other ablation tables. Table S17 reports the results per category together with the Pert. and Overall aggregates.

The "Combined union" baseline takes the union of channels selected by all three individual metrics (Pearson correlation, Relative L_2, and Per-channel AUC at their respective operating points), testing whether combining heterogeneous selection criteria improves coverage.

Pearson correlation is the best overall metric, achieving the best or tied-best AP in 8 of 9 categories. Its advantage over Relative L_2 is most pronounced on spatially destructive perturbations, where L_2 penalizes magnitude rescaling that Pearson ignores. Per-channel AUC performs worst, indicating that clean-image discriminativeness is a poor proxy for perturbation robustness. Combined union outperforms each non-Pearson metric individually but still falls short of Pearson alone, indicating that adding noisy selections dilutes the mask quality.

Table S17: **Channel selection metric comparison under aggressive augmentation: per-category and aggregate AP (%).** All use Corvi backbone, aggressive augmentation, no clamp. Union top-5% for Pearson, Relative L_2, and Combined. AUC > 0.9 threshold for Per-channel AUC. Best per column in **bold**.

Selection Metric	Clean	JPEG	WebP	Blur	Noise	Jitter	Resize	Gray	Combo	Pert.	Overall
No mask (all 2,048 ch)	97.3	88.1	84.7	82.2	80.4	94.7	75.8	98.3	78.3	85.2	83.5
Per-channel AUC (> 0.9)	97.7	90.9	86.2	84.2	82.1	95.2	77.4	98.7	81.8	87.0	85.7
Combined union	97.9	**91.5**	88.5	85.8	83.0	95.5	**79.2**	98.8	82.9	88.2	86.8
Relative L_2	**98.1**	90.6	88.7	83.6	82.4	**95.7**	77.3	**98.9**	81.5	87.4	85.9
Pearson correlation (ours)	**98.1**	**91.5**	**89.2**	**86.0**	**83.2**	**95.7**	79.1	**98.9**	**83.1**	**88.4**	**87.0**

H.2 Mask × Clamp Factorial

This factorial separates the effect of the channel mask from that of the non-negative weight constraint under the aggressive schedule. The mask carries the gain, and the two do not combine on this benchmark (their interaction on the novel-perturbation set is analyzed in Section J.2).

Table S18: **Ablation: mask × clamp factorial, per category.** Corvi backbone, aggressive augmentation. Mask without clamp dominates Pert. and Overall. Clamping helps only without masking. Combining mask and clamp hurts performance, particularly on the compression (JPEG, WebP) and Noise categories.

Mask	Clamp	Clean	JPEG	WebP	Blur	Noise	Jitter	Resize	Gray	Combo	Overall
		97.3	88.1	84.7	82.2	80.4	94.7	75.8	98.3	78.3	83.5
	✓	**98.6**	88.9	85.8	83.8	79.0	95.4	78.1	**99.1**	81.3	84.9
✓		98.1	**91.5**	**89.2**	**86.0**	**83.2**	**95.7**	**79.1**	98.9	**83.1**	**87.0**
✓	✓	98.1	86.5	83.0	83.3	77.3	94.4	78.4	98.7	80.0	83.4

H.3 Seed Variability

To quantify run-to-run variability, we train each configuration with three random seeds, varying only data order and augmentation sampling (the FC head is zero-initialized and the mask is fixed). Standard deviations across three seeds are 0.3 pp or below for the three CNN configurations (Corvi and Rajan with paired AE fakes, and Corvi with LDM fakes), showing low variability from data order and augmentation sampling. For the CLIP configuration (union-20%, three seeds) the spread is larger (Overall 74.4/75.0/76.1, std 0.7 pp), and the unmasked LayerNorm+FC control shows an overlapping spread over its three seeds (std 0.5 pp). Across the nine CNN runs (3 configurations × three seeds), the worst Overall AP still exceeds the respective baseline.

The random-mask control (271 randomly selected channels, same aggressive augmentation) was also trained with three seeds. Its standard deviation (±1.1 pp) is 3× larger than the Pearson-selected mask, and even the best random seed falls 3.4 pp below the worst Pearson seed. Pearson-based selection thus scores higher and varies less across seeds than random channel subsampling. The no-selection control of Table 2 was likewise trained with three seeds, so its comparison with PICS is seed-matched, and the end-to-end fine-tuned backbones show a larger seed spread than PICS with non-overlapping ranges (Section F.2). Figure S4 shows the same comparison across perturbation severity. Top-Pearson selection outperforms random selection at the matched 271-channel budget under the aggressive schedule (86.6 vs. 81.6 Overall over three seeds, Table 2) and bottom-5% selection, which is a separately configured control (340 channels, WebP schedule, 75.4 Overall against 86.4 for top-5% under the same WebP schedule), and the gap widens as severity increases.

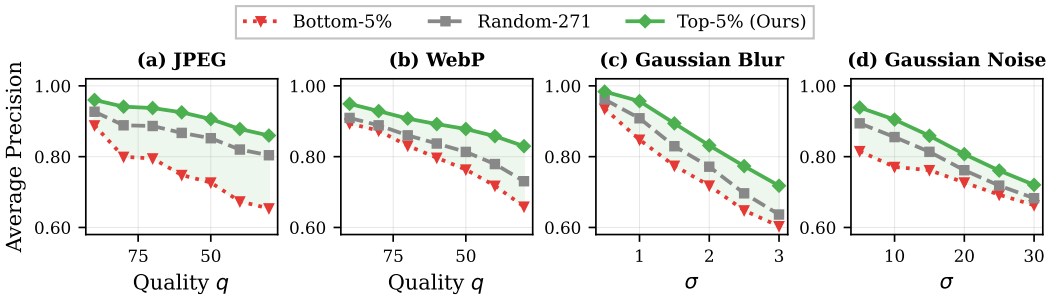

Figure S4: **Channel-identity validation.** AP vs. perturbation severity on the Corvi backbone for top-5% Pearson selection (271 channels, aggressive schedule, single run), random-271 selection (aggressive schedule, single run), and bottom-5% Pearson selection (340 channels, WebP schedule, single run). The bottom-selection curve therefore differs in budget and schedule from the other two. The shading marks the gap between the top and bottom curves and is not a confidence interval.

H.4 MASKING THRESHOLD ($K\%$)

Table S19 shows the full threshold sweep. Performance shows a broad peak in K over the tested range: selecting too few channels discards discriminative signal, while selecting too many retains perturbation-sensitive ones.

Table S19: Masking threshold $K\%$. Pearson union, WebP augmentation, no clamp. "Overall" denotes the mean of all 51 conditions (AP in %).

$K\%$	Ch.	Clean	Combo	Pert.	Overall
3	182	97.4	81.5	86.9	85.5
5	**271**	**97.8**	**82.4**	**87.7**	**86.4**
7	321	97.6	**82.4**	87.5	86.2
10	414	97.5	81.6	87.1	85.7
100	2,048	97.7	79.3	85.7	84.1

Table S20 reports the analogous sweep for the Rajan backbone (aggressive augmentation, no clamp). Performance peaks at $K{=}2\%$, well above the no-mask baseline ($K{=}100\%$ row).

Table S20: Masking threshold $K\%$ for the Rajan backbone. Pearson union, aggressive augmentation, no clamp. "Pert." is the 35-cell single-perturbation mean, "Combo" the 15-cell multi-step mean, and "Overall" the mean across all 51 conditions. Best per column in **bold**.

$K\%$	Ch.	Clean	Pert.	Combo	Overall
1	79	**94.4**	84.6	79.6	83.3
2	**147**	93.8	**84.7**	**80.1**	**83.5**
3	214	93.1	83.9	79.6	82.8
5	333	93.5	84.3	79.6	83.1
100	2,048	92.0	80.9	76.5	79.8

H.5 AGGREGATION STRATEGY

Table S21 compares aggregation strategies in two settings. *(top)* Union consistently outperforms stricter strategies under standard augmentation when channel budgets differ. *(bottom)* At a *matched* 271-channel budget under aggressive augmentation, the union still leads, ahead of average-Pearson, minimum-Pearson, mean-rank, and k-of-n aggregations in Overall AP, which does not support the possibility that the union's advantage is merely an artifact of having more channels.

Figure S5 visualizes why intersection fails: each perturbation category contributes a substantial number of channels that are robust only under that specific category, indicating that robustness to different perturbation categories is distributed across partially disjoint channel subsets. The union strategy captures this complementarity by retaining channels that are robust under any perturbation category.

H.6 BEST CONFIGURATION PER TRAINING PROTOCOL

We evaluate PICS under four training protocols formed by combining two CNN backbones with two fake-data sources: *paired AE* uses autoencoder-reconstructed fakes with paired real/fake loading and synchronized augmentation, while *LDM* uses prompt-generated latent-diffusion fakes with unpaired loading and independent augmentation. Table S22 shows that the optimal masking threshold varies by backbone, but channel masking improves performance across all protocols regardless of the specific threshold.

H.7 LAYER-WISE ANALYSIS

To justify the use of layer-4 (final convolutional block), we apply the same Pearson-based channel analysis and masking at each of the four residual blocks of ResNet50-NoDown, testing both $K{=}5\%$

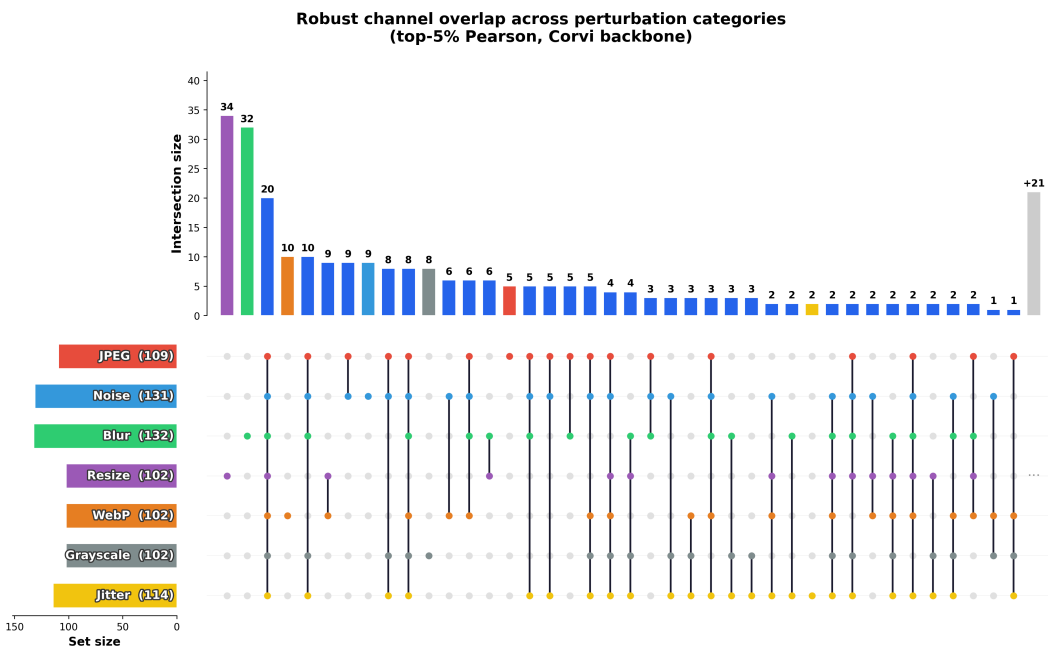

Figure S5: **UpSet plot of robust channel overlap across perturbation categories** (Corvi backbone, top-5% Pearson). Each row represents a perturbation category. The set size (left) shows the number of channels in that category's top-5%. Bar heights show the number of channels in each intersection pattern. The largest bars correspond to category-exclusive channels (e.g., 34 resize-only, 32 blur-only), indicating that different perturbation categories require partially disjoint sets of robust channels. The full 7-category intersection (all dots connected) yields only 20 channels, motivating the union strategy.

Table S21: Aggregation strategy comparison. "Overall" denotes the mean of all 51 conditions (AP in %). *Top:* standard augmentation, varying channel budget. *Bottom:* aggressive augmentation, matched 271-channel budget, isolating the effect of the aggregation criterion from the channel-count effect.

Strategy	Ch.	Clean	Pert.	Overall
Standard augmentation, varying budget				
Union top-5% (ours)	271	96.9	**86.9**	**85.6**
Union top-3%	182	96.8	86.6	85.3
k-of-n (k=3)	143	97.0	86.0	84.7
k-of-n (k=5)	105	**97.5**	86.1	84.7
Avg z-score top-5%	102	97.4	86.3	84.9
Aggressive augmentation, matched 271-channel budget				
Union top-5% (ours)	271	98.1	**88.4**	**87.0**
Avg Pearson top-271	271	97.8	88.1	86.7
Min Pearson top-271	271	97.8	87.8	86.5
Mean rank top-271	271	97.7	87.8	86.4
k-of-n (k=2, ∼271 ch)	269	—	—	86.1

and K=10%. Table S23 reports the results: layer-4 masking substantially outperforms all earlier layers, while layers 1–2 perform near chance. (The layer-4 row shows 275 channels vs. 271 in the main results due to a rerun for cross-layer consistency. The two masks overlap by 97% with <0.2 pp AP difference.)

Masks at earlier blocks alter the inputs of the subsequent frozen blocks, so this result concerns the masking location rather than the absence of stable channels in early layers. With that caveat, the frozen downstream blocks do not recover from early-layer masking. Layer-4 features exhibit

Table S22: Best configuration per training protocol. Optimal $K\%$ varies but the method improves performance across protocols. "All pert." is the mean AP (%) over the 50 perturbed conditions and "Overall" the mean over all 51 conditions (1 clean, 35 single-perturbation, 15 combo).

Protocol	Backbone	$K\%$	Channels	Aug.	Clean	All pert.	Overall
Paired AE	Corvi	5	271	Aggr.	98.1	86.8	**87.0**
LDM	Corvi	1	74	Aggr.	98.2	85.9	86.2
Paired AE	Rajan	2	147	Aggr.	93.8	83.3	83.5
LDM	Rajan	2	147	Aggr.	94.5	83.3	83.5

Table S23: **Layer-wise masking analysis** on Corvi backbone with aggressive augmentation. Each row applies the union-of-top-$K\%$ Pearson channel mask at a single residual block. "Ch." denotes the number of selected channels out of total channels at that layer. Layer-4 dominates. Best per column in **bold**.

Layer	$K\%$	Ch.	Total	Clean	JPEG	WebP	Blur	Noise	Jitter	Resize	Gray	Combo	Overall
Layer-1	5	40	256	58.4	57.6	56.9	58.7	58.4	57.6	58.0	58.7	57.5	57.7
Layer-1	10	68	256	59.1	61.1	58.6	60.2	63.9	58.3	59.2	58.1	59.5	60.0
Layer-2	5	72	512	42.0	45.0	44.5	43.2	41.7	44.7	44.0	49.6	46.6	44.8
Layer-2	10	135	512	50.0	50.8	50.2	44.7	49.0	50.7	47.3	53.3	49.8	49.3
Layer-3	5	126	1,024	68.0	65.5	58.8	63.1	56.8	64.6	64.4	68.4	63.0	62.5
Layer-3	10	216	1,024	81.6	72.3	68.4	64.9	63.8	77.5	68.3	85.3	67.9	69.3
Layer-4	**5**	**275**	**2,048**	**98.1**	**91.7**	**89.2**	**86.3**	**83.3**	**95.7**	**79.6**	**98.9**	**83.3**	**87.2**
Layer-4	10	422	2,048	98.0	91.2	89.0	85.9	82.9	95.6	78.9	98.9	83.0	86.9

channel-specific perturbation sensitivity, enabling meaningful selection. The performance gap across layers is far larger than any threshold effect, reinforcing layer-4 as the clear default.

I Sensitivity of the Selection to the Analysis Set

I.1 Channel Selection Robustness

To verify that the channel selection is not sensitive to the sample used for analysis, we rerun the robustness analysis with N=500 and N=2,000 images (the default analysis uses N=5,000 for the Corvi backbones and N=1,000 for Rajan). At each family's operating $K\%$, the Jaccard overlap between the default analysis and the N=2,000 rerun exceeds 0.91 in all cases: J=0.978 for Corvi with paired AE fakes at 5%, J=0.961 for Corvi with LDM fakes at 1%, and J=0.915 for Rajan at 2%. Changing the sample size by a factor of two or more changes fewer than 10% of selected channels, indicating near-converged selection.

I.2 Perturbation Set Sensitivity

We test the sensitivity of the selected channels to the composition of the perturbation analysis set. Removing WebP or JPEG from the analysis set changes the selection modestly (Jaccard overlap above 0.93), whereas removing blur or resize changes it more (0.69–0.92). Adding extra WebP quality levels barely alters the set ($J \geq 0.99$). However, restricting to JPEG-only perturbations causes substantial divergence (J=0.30–0.40), indicating that both spatial and frequency-domain perturbations contribute complementary information. Diversity across perturbation categories is essential (JPEG-only analysis produces a substantially different mask), but the method is insensitive to the exact number or severity levels within each category.

I.3 Robustness to Omitting a Category from Channel Scoring

To test whether the selected channels generalize beyond the perturbation categories used for analysis, we perform a leave-one-category-out experiment. For each of the 7 perturbation categories, we construct the channel mask using the remaining categories, retrain the FC classifier with the resulting mask (the head-training augmentation is unchanged, so the category is excluded from scoring only), and evaluate on all 51 conditions including the excluded category. Table S24 reports the results.

Overall AP is robust across all held-out configurations, and performance on the held-out category itself remains strong. This indicates substantial redundancy: robust channels tend to be robust to multiple perturbation categories, not just the ones used to identify them. WebP and grayscale contribute the most unique information, followed by blur, whereas JPEG and jitter can be removed without degradation, as their signals are captured by other categories (Table S24). Overall, the mask is robust to omitting any one category from scoring. Because the omitted category remains in the head-training augmentation, this does not test categories absent from every training stage.

Table S24: **Leave-one-category-out evaluation.** Each row excludes one perturbation category from the channel analysis and reports AP (%) on the held-out category (mean over all its severity levels) and Overall (mean over the 51 conditions). Δ columns show change vs. the full configuration (first row).

Held-out Category	Ch.	Held-out AP	Δ_{held}	Overall AP	Δ_{all}
None (full)	271	—	—	87.0	—
JPEG	266	91.7	+0.1	87.3	+0.3
WebP	261	88.0	−1.2	86.6	−0.5
Noise	262	83.2	+0.1	87.1	+0.1
Blur	239	85.5	−0.5	86.7	−0.4
Resize	237	79.5	+0.5	86.9	−0.1
Jitter	269	95.8	+0.1	87.3	+0.2
Grayscale	263	98.9	−0.0	86.6	−0.5

J Held-Out Perturbation Families and the Hotspot

Table S25: **Perturbation families outside the scoring set, grouped.** Mean AP (%) over each group's cells (conditions × 9 generators, and groups differ in size and are an evaluation suite, not deployment prevalence). Δ is PICS minus Corvi$\oplus$, computed before rounding. Per-family values are in Table S26.

Group	#cnd	Corvi	Corvi$\oplus$	PICS (ours)	Δ
Photometric (brightness, contrast, gamma, hue)	26	92.1	**97.7**	97.5	−0.2
Filtering (bilateral, median, motion, sharpen)	15	89.9	94.0	**94.3**	+0.4
Near chance (salt-pepper, pixelize)	5	47.5	**52.4**	50.4	−2.0
Recompression (JPEG70→WebP70)	1	84.6	83.5	**89.8**	+6.3
Geometric/occlusion (crop, sticker, cutout, rotate)	16	88.8	**97.2**	95.9	−1.3

J.1 Held-Out Perturbation Families: Full Breakdown

Table S26 reports mean AP on LDMFakeDetect across the 15 held-out perturbation families summarized in Section 6 (63 unique conditions × 9 generators = 567 cells). The deployment-relevant subset (six families) appears in Table 6. Here we show all families including the failure cases analyzed in the hotspot dissection (Section J.2). Grouping the families by whether they preserve spatial alignment, the photometric and filtering groups are on par with Corvi$\oplus$ on average, the near-chance and alignment-breaking groups lose a little, and multi-step compression gains the most (Table S25, with group-mean rows in Table S26).

PICS preserves (within 0.5 pp) or improves over Corvi$\oplus$ on nine of the fifteen families, with the largest deployment-relevant gain on multi-step compression. The negative tail concentrates on the geometric and occlusion transforms (the failure subset dissected in Section J.2), plus salt-pepper noise and pixelization, where all detectors approach chance.

J.2 Hotspot Mechanism: A Refuted Hypothesis

The cutout/rotate hotspot characterized in Section 6 (mean $\Delta = -5.6$ pp on the 15-cell rectangle, with a permutation $p < 10^{-4}$ that is exploratory because the rectangle was chosen after inspecting

Table S26: **Held-out perturbation families: full breakdown on LDMFakeDetect.** Mean AP (%) across 9 generators per family. **#cnd** is the number of severity levels per family. Δ is PICS top-5% minus Corvi$\oplus$. Italic rows give the mean over each group's cells. Below the final midrule are the geometric and occlusion transforms on which the largest regressions occur (analyzed in Section J.2). All families are excluded from channel scoring. Cutout, brightness, contrast, and hue overlap with head-training augmentation types (cutout, color jitter), and the training schedule applies $90°$ rotations whereas the test rotations are 2–$20°$.

Family	#cnd	Corvi	Corvi$\oplus$	PICS (top-5%)	Δ
Photometric					
Brightness ($\Delta \in \{\pm 25, \pm 50\}$)	4	92.32	97.41	**97.48**	+0.07
Contrast ($f \in \{0.5, 0.7, 1.5\}$)	3	91.93	**97.29**	97.22	−0.07
Gamma ($\gamma \in [0.3, 2.5]$, 16 levels)	16	92.24	**97.85**	97.63	−0.22
Hue ($\Delta \in \{60°, 120°, 180°\}$)	3	90.91	**97.62**	97.09	−0.53
Group mean	26	92.07	97.69	97.50	−0.19
Filtering / enhancement					
Bilateral filter ($d \in \{5, 9\}$)	2	88.98	88.98	**90.43**	+1.45
Median blur ($k \in \{3, 5\}$)	2	83.58	87.76	**88.00**	+0.24
Motion blur ($k \in \{3, 5\}$)	2	80.10	**90.31**	89.90	−0.41
Sharpen ($r \in [0.25, 4]$, 9 levels)	9	93.73	97.25	**97.57**	+0.32
Group mean	15	89.93	93.96	94.32	+0.36
Near-chance degradations (all detectors close to chance)					
Salt-pepper ($p \in \{0.01, 0.05\}$)	2	49.17	**56.83**	52.93	−3.90
Pixelize ($b \in \{4, 8, 12\}$ px)	3	46.42	**49.42**	48.66	−0.77
Group mean	5	47.52	52.39	50.36	−2.02
Multi-step compression					
JPEG70→WebP70	1	84.56	83.51	**89.84**	+6.33
Geometric and occlusion transforms					
Crop offset ($\delta \in \{4, 8, 16, 24, 32\}$)	5	92.58	**98.52**	98.07	−0.45
Sticker (16 px occlusion)	1	91.41	**98.07**	97.52	−0.55
Cutout ($p \in \{5, 10, 20, 25\}\%$)	4	86.90	**96.97**	95.27	−1.70
Rotate ($\theta \in \{2°, 5°, 7°, 10°, 15°, 20°\}$)	6	86.38	**96.07**	94.16	−1.91
Group mean	16	88.76	97.19	95.87	−1.32
All cells (per-cell median)	63	—	—	—	+0.10
All cells ($\lvert\Delta\rvert \leq 2$ pp)	63	—	—	—	77%
All cells ($\Delta \geq 0$)	63	—	—	—	62%

the result matrix) initially appeared to admit a simple channel-identity explanation. We tested that explanation directly and found it does not hold. We report the failed test here both for transparency and because it shows that one specific intervention does not repair the gap.

Initial hypothesis. On the Corvi backbone, PICS top-5% selects $|\mathcal{S}| = 271$ channels. In our reproduction Corvi$\oplus$, Stay-Positive's non-negative weight clamping leaves only 27 classifier weights non-zero (a 98.7% implicit sparsity rate). The two channel sets are only partially aligned: the intersection is $14/27$ (52%), so 13 channels are active in Corvi$\oplus$ but absent from PICS top-5%. We initially conjectured that these 13 channels carried perturbation-robust signal under cutout/rotate that PICS misses, and that retaining them via a union mask would close the hotspot.

Test 1: union mask training. We trained a 284-channel union mask, $\mathcal{S}_\cup = \mathcal{S}_{\text{Pearson top-5\%}} \cup \mathcal{S}_{\text{Corvi}\oplus\text{-active}}$, with the same FC-only recipe used by PICS. The result reproduces PICS top-5% within 0.1 pp Overall on the 51 LDMFakeDetect conditions (86.9 vs 87.0) and on every one of the 14 novel-perturbation conditions, including the cutout/rotate hotspot cells (Table S29). The 13 added channels contribute essentially zero predictive value when trained jointly with the 271 PICS channels: the optimizer prefers the dominant 271 and the 13 additions get tiny FC weights.

Test 2: per-channel robustness re-check on cutout/rotate. Independently, we re-measured per-channel Pearson robustness on the cutout ($10\%, 20\%$) and rotate ($5°, 10°, 15°$) perturbations

themselves, the very transforms that define the hotspot (Table S27). The 13 Corvi⊕ extras are *not* specially robust under these transforms: their mean Pearson is consistently slightly *lower* than that of the typical PICS-selected channel.

Table S27: **Pearson robustness of the three channel groups under hotspot transforms.** Mean $\bar{\rho}_c$ over 1,000 training images. Higher is more robust. The 13 Corvi⊕ extras are *slightly less* robust than typical PICS-selected channels, which does not support the conjecture that they carry special cutout/rotate signal.

Channel group	cutout-10%	cutout-20%	rotate-5°	rotate-10°	rotate-15°
PICS top-5% (271 ch)	0.960	0.884	0.533	0.361	0.283
Corvi⊕ extras (13 ch)	0.958	0.879	0.526	0.336	0.253
All 2,048 channels	0.947	0.865	0.416	0.245	0.172

What the negative result shows. Neither test supports the channel-identity explanation: the tested channel-addition intervention did not explain or repair the gap, and the 13 extras are not more pattern-stable under these transforms. This does not prove that the added channels carry no useful information under every readout, only that adding them to the unconstrained masked head does not help.

Test 3: the hotspot is two distinct failure modes, not one. A direct measurement of where the spatial-correspondence assumption holds vs. breaks (Table S28) reveals that "cutout" and "rotate" (both included in the hotspot rectangle) behave very differently with respect to the criterion. **Rotate** progressively destroys spatial correspondence: at 15° rotation the median PICS-channel $\bar{\rho}_c$ falls below every analysis-set value except the strongest blur (Table S28), consistent with geometric displacement of content across feature-map positions (low correlation alone does not prove broken correspondence, since blur also lowers it without moving content). **Cutout**, by contrast, preserves the criterion's assumption: median $\bar{\rho}_c$ stays as high as under the cleanest analysis-set perturbations (mild jitter, grayscale). Cutout zeros a small spatial region but leaves the surrounding pattern intact, so the per-channel pattern correlation barely moves, yet PICS still loses to Stay-Positive on cutout-on-hard-generator cells.

The cutout failure mode is therefore not explained by either the channel-identity hypothesis or the assumption-violation hypothesis.

Test 4 (positive result): hotspot evaluation of mask+clamp. We took the existing mask+clamp checkpoint (Table S18 bottom row, the same PICS top-5% mask trained with the additional non-negative weight clamp) and evaluated it on the 14-condition novel-perturbation set. The results disambiguate whether clamping per se closes the hotspot independently of channel selection (Table S29).

The combination + clamp closes the hotspot: mask+clamp matches Corvi⊕ (clamp only) on four of the five hotspot conditions and exceeds it on cutout-20%, a large gain over either mask-only configuration (PICS top-5% or the union mask of PICS top-5% with Corvi⊕-active channels, 284 channels in total). Because the same binary mask is retained, the excluded channels contribute nothing in either configuration, so this improvement cannot be attributed to restoring their contributions. The result instead indicates that the sign constraint on the readout matters on these conditions, and Tests 1–3 show that adding the 13 channels active only in the clamp baseline does not help the unconstrained masked head. We leave the precise interaction between signed evidence and these transformations unresolved.

Trade-off: hotspot vs. overall. The mask+clamp combination closes the hotspot but *loses* Overall AP on the standard 51-condition LDMFakeDetect benchmark (Table S18). The two effects point in opposite directions: the sign constraint helps on the alignment-breaking subset but costs AP on the geometry-preserving perturbations that make up most of the benchmark. A practical deployment would either keep PICS top-5% for general robustness, or opt into mask+clamp if cutout/rotate-style transforms are expected. Adaptive switching (e.g., a hotspot detector at test time) is a natural future direction.

Table S28: **Where the spatial-correspondence assumption holds and breaks.** Median per-channel Pearson robustness $\bar{\rho}_c$ across the 271 PICS-selected channels, for each analysis-set and held-out perturbation. Lower values indicate weaker preservation of the channel's spatial pattern. Rotation has a geometric reason to challenge pointwise comparison, whereas blur lowers correlation without moving content. The upper block lists the 11 selection settings plus JPEG q=70 and blur σ=1 from the 18-level diagnosis set (Section B).

Category	Perturbation	median $\bar{\rho}_c$ (PICS-271)	IQR	median (all 2048)
analysis set	JPEG q=70	0.637	[0.614, 0.655]	0.443
	JPEG q=50	0.535	[0.509, 0.555]	0.331
	JPEG q=30	0.430	[0.403, 0.450]	0.236
	WebP q=30	0.421	[0.390, 0.449]	0.223
	Blur σ=1	0.691	[0.675, 0.705]	0.476
	Blur σ=2	0.395	[0.366, 0.412]	0.170
	Blur σ=3	0.201	[0.168, 0.218]	0.078
	Noise σ=10	0.619	[0.599, 0.636]	0.448
	Noise σ=20	0.451	[0.428, 0.469]	0.248
	Resize 0.5×	0.480	[0.453, 0.505]	0.293
	Grayscale	0.922	[0.916, 0.926]	0.866
	Jitter 0.5	0.802	[0.789, 0.808]	0.686
	Jitter 1.0	0.608	[0.591, 0.620]	0.442
held-out hotspot	Cutout 10%	0.960	[0.958, 0.962]	0.953
	Cutout 20%	0.882	[0.877, 0.888]	0.872
	Rotate 5°	0.551	[0.503, 0.584]	0.407
	Rotate 10°	0.369	[0.329, 0.413]	0.230
	Rotate 15°	0.289	[0.245, 0.331]	0.153

Table S29: **Test 4: hotspot AP under the hard mask, the non-negative readout constraint (clamp), and their combination.** Mean AP (%) on the cutout/rotate cells of the three hardest unseen generators (aMUSEd, FLUX, Playground), 15 cells total. Adding non-negative clamp on top of the same PICS top-5% mask closes the hotspot to within 0.2 pp of the clamp-only baseline.

Configuration	Cutout 10%	Cutout 20%	Rotate 5°	Rotate 10°	Rotate 15°	Hotspot avg.
PICS top-5% (mask only)	91.2	85.7	88.5	83.9	84.1	86.7
PICS ∪ Corvi⊕-active union, 284 ch (mask only)	91.1	85.7	88.5	84.0	84.1	86.7
Corvi⊕ (clamp only)	94.8	92.0	**93.5**	**91.2**	**90.0**	92.3
Mask + clamp	**94.9**	**93.1**	93.3	91.1	**90.0**	**92.5**

Test 5: extended analysis set. A complementary fix holds the criterion form fixed (flat-spatial Pearson) and instead extends the analysis set with the would-be hotspot transforms themselves: cutout (10%, 20%) and rotate (5°, 10°, 15°). The union of top-5% per perturbation now spans 16 perturbations and selects 379 channels (vs. PICS's 271). Table S30 reports the result: extending coverage helps only modestly and does not close the gap to clamp-based methods, so extending coverage alone, even at a larger channel count, does not close it.

Table S30: **Test 5: hotspot AP for analysis-set extension.** Extending the analysis set with cutout and rotate themselves modestly helps (+0.9 pp on the hotspot rectangle) but does not close the gap to clamp-based methods (4.7 pp gap to Stay-Positive remains). Extending coverage does not close the gap, whereas the non-negative readout constraint does (Test 4).

Configuration	Cutout 10%	Cutout 20%	Rotate 5°	Rotate 10°	Rotate 15°	Hotspot avg.
PICS top-5% (11 pert., 271 ch)	91.2	85.7	88.5	83.9	84.1	86.7
Extended top-5% (16 pert., 379 ch)	91.7	86.3	89.7	85.2	85.1	87.6
Corvi⊕ (clamp, for reference)	94.8	92.0	93.5	91.2	90.0	92.3

Test 6: alignment-tolerant criterion. A criterion-side alternative replaces the Pearson correlation by its maximum over all integer shifts $(\Delta x, \Delta y) \in \{-2, \ldots, 2\}^2$ of the perturbed 12×12 map, so that a channel whose pattern survives up to a small displacement is still scored as robust. The

resulting union-of-top-5% mask has 284 channels, 231 of them shared with the 271-channel PICS mask. Under the WebP head-training protocol (so the values differ from the aggressive-schedule ones in Section F.2) it scores 86.0 Overall against 86.4 for the flat criterion, and 93.9 against 94.1 mean AP over the cutout and rotate conditions of the 14-condition novel-perturbation set (9-generator protocol). This variant addresses small translations of the feature map only and is not a rotation-aligned comparison. Tolerating small shifts in the score does not recover the alignment-breaking cells, consistent with Test 3: the rotation deficit is not a matter of small displacements, and for cutout the readout constraint of Test 4 helps where the criterion-side variants tested here do not.

K Full Per-Condition Results

Tables S31 and S32 report the mean AP (%) across 9 generators for every perturbation condition, covering the three primary models in each backbone family. Table S33 and Table S34 report per-condition AP for CLIP ViT-L/14 and DINOv2 ViT-B/14 (B-Free), respectively.

Table S31: **Full per-condition AP (%): Corvi family.** Mean AP across 9 generators on LDM-FakeDetect. Ours is a single training run (seed means in Table 1). Best per row in **bold**.

Condition	Corvi	Corvi⊕	Ours	Condition	Corvi	Corvi⊕	Ours
Clean	92.7	**98.5**	98.1	Jitter 0.2	92.2	**97.7**	97.5
JPEG 90	90.3	94.9	**96.0**	Jitter 0.4	91.2	**96.9**	**96.9**
JPEG 80	86.0	91.8	**94.1**	Jitter 0.6	90.5	96.2	**96.4**
JPEG 70	86.0	89.8	**93.7**	Jitter 0.8	89.3	94.6	**95.2**
JPEG 60	83.6	88.0	**92.5**	Jitter 1.0	86.7	90.4	**92.5**
JPEG 50	81.9	85.2	**90.6**				
JPEG 40	78.4	84.0	**87.8**	Resize 0.25	46.7	49.1	**51.6**
JPEG 30	76.6	80.9	**86.0**	Resize 0.5	74.9	86.5	**89.2**
WebP 90	88.9	92.4	**94.9**	Resize 0.75	89.6	96.3	**96.4**
WebP 80	87.8	88.8	**92.9**				
WebP 70	85.0	85.9	**90.8**	Grayscale	95.7	**99.0**	98.9
WebP 60	82.5	84.4	**89.2**				
WebP 50	79.9	83.3	**87.8**	Resize 0.75 + JPEG 70	81.2	86.4	**90.5**
WebP 40	76.1	81.3	**85.8**	Resize 0.50 + JPEG 50	68.3	77.2	**79.6**
WebP 30	70.8	79.2	**83.0**	Resize 0.75 + WebP 70	72.0	79.1	**82.0**
Blur 0.5	93.8	**98.5**	98.4	Resize 0.50 + JPEG 30	64.9	73.2	**75.0**
Blur 1.0	89.5	94.5	**95.7**	Resize 0.75 + WebP 50	65.3	75.8	**77.7**
Blur 1.5	83.0	85.8	**89.4**	Blur 0.5 + JPEG 70	85.7	88.2	**92.7**
Blur 2.0	76.6	80.1	**83.2**	Blur 1.0 + JPEG 50	69.8	77.6	**80.2**
Blur 2.5	68.5	72.9	**77.3**	Jitter 0.2 + JPEG 70	84.8	88.9	**92.9**
Blur 3.0	61.9	67.1	**71.8**	Jitter 0.6 + JPEG 50	78.6	81.7	**86.4**
Noise 5	85.8	91.9	**93.9**	Gray + JPEG 70	87.9	90.0	**94.1**
Noise 10	81.5	86.4	**90.5**	Blur 1.0 + Noise 5	76.3	82.3	**84.2**
Noise 15	79.0	80.1	**85.9**	Blur 1.5 + Noise 10	59.6	**68.4**	67.5
Noise 20	74.9	74.6	**80.7**	Resize 0.75 + JPEG 50 + Noise 5	75.5	82.2	**85.2**
Noise 25	71.2	70.2	**76.1**	Blur 1.0 + Noise 5 + JPEG 50	68.9	76.4	**78.9**
Noise 30	68.0	66.8	**72.0**	Resize 0.50 + Blur 0.5 + JPEG 50	69.3	77.1	**79.5**
				Overall (51 cond.)	79.3	84.1	**87.0**

Table S32: **Full per-condition AP (%): Rajan family.** Mean AP across 9 generators on LDM-FakeDetect. Rajan⊕ is the released checkpoint and Ours a single training run (seed means in Table 1). Best per row in **bold**.

Condition	Rajan	Rajan⊕	Ours	Condition	Rajan	Rajan⊕	Ours
Clean	91.3	**93.9**	93.8	Jitter 0.2	91.9	93.9	**94.0**
JPEG 90	90.6	91.8	**92.7**	Jitter 0.4	91.3	93.2	**93.4**
JPEG 80	87.2	89.2	**90.4**	Jitter 0.6	90.5	92.6	**92.8**
JPEG 70	87.5	89.0	**90.7**	Jitter 0.8	89.3	91.6	**91.9**
JPEG 60	86.2	88.3	**90.1**	Jitter 1.0	85.7	88.9	**89.8**
JPEG 50	84.9	86.6	**88.6**				
JPEG 40	80.9	82.9	**84.8**	Resize 0.25	**47.8**	46.5	44.7
JPEG 30	79.0	81.0	**83.1**	Resize 0.5	81.5	**86.7**	86.3
WebP 90	87.6	89.8	**90.7**	Resize 0.75	92.1	93.9	**94.0**
WebP 80	81.9	86.6	**87.2**				
WebP 70	77.4	83.4	**83.6**	Grayscale	94.3	**95.4**	**95.4**
WebP 60	74.2	**81.4**	81.1				
WebP 50	71.2	**79.8**	78.9	Resize 0.75 + JPEG 70	87.9	87.8	**89.4**
WebP 40	67.8	**76.9**	75.7	Resize 0.50 + JPEG 50	77.6	77.0	**78.6**
WebP 30	63.1	**74.0**	71.8	Resize 0.75 + WebP 70	63.8	**77.0**	74.3
Blur 0.5	93.8	94.6	**94.8**	Resize 0.50 + JPEG 30	71.6	70.4	**71.9**
Blur 1.0	93.2	92.6	**93.5**	Resize 0.75 + WebP 50	56.3	**71.1**	67.1
Blur 1.5	**90.5**	86.3	87.9	Blur 0.5 + JPEG 70	89.8	88.5	**90.5**
Blur 2.0	**86.5**	80.8	82.2	Blur 1.0 + JPEG 50	78.5	78.5	**80.2**
Blur 2.5	**81.1**	75.5	76.0	Jitter 0.2 + JPEG 70	86.4	88.3	**89.8**
Blur 3.0	**76.2**	71.3	71.7	Jitter 0.6 + JPEG 50	80.9	83.2	**85.1**
Noise 5	88.2	91.0	**91.9**	Gray + JPEG 70	87.8	89.1	**90.9**
Noise 10	83.0	87.4	**88.3**	Blur 1.0 + Noise 5	79.9	82.0	**82.3**
Noise 15	76.2	82.6	**83.4**	Blur 1.5 + Noise 10	64.3	**65.8**	65.7
Noise 20	69.7	77.7	**78.2**	Resize 0.75 + JPEG 50 + Noise 5	69.6	**79.4**	78.3
Noise 25	64.9	73.7	**73.8**	Blur 1.0 + Noise 5 + JPEG 50	77.7	76.9	**78.5**
Noise 30	61.4	**70.2**	70.0	Resize 0.50 + Blur 0.5 + JPEG 50	78.5	77.0	**78.7**
				Overall (51 cond.)	80.2	83.0	**83.5**

Table S33: **Full per-condition AP (%): CLIP ViT-L/14 (UniversalFakeDetect).** Mean AP across the eight GenImage generators of Section 5.1. Ours is a single training run at $K=20\%$ (seed means in Table 1). Best per row in **bold**. The matched LayerNorm+FC control (Table 1) is not shown, so the margin over UFD is not a selection effect.

Condition	UFD	Ours	Condition	UFD	Ours
Clean	85.1	**94.7**	Jitter 0.2	84.1	**90.8**
JPEG 90	74.0	**82.9**	Jitter 0.4	81.4	**87.3**
JPEG 80	72.3	**80.6**	Jitter 0.6	79.5	**84.1**
JPEG 70	69.3	**77.1**	Jitter 0.8	77.2	**80.1**
JPEG 60	69.9	**77.6**	Jitter 1.0	**70.1**	69.9
JPEG 50	69.8	**75.6**			
JPEG 40	67.9	**74.8**	Resize 0.25	62.0	**62.4**
JPEG 30	68.8	**73.4**	Resize 0.5	68.4	**74.6**
WebP 90	79.0	**80.2**	Resize 0.75	78.1	**87.2**
WebP 80	73.3	**73.8**			
WebP 70	70.6	**72.9**	Grayscale	84.1	**88.9**
WebP 60	69.9	**73.1**			
WebP 50	68.6	**73.0**	Resize 0.75 + JPEG 70	69.2	**74.8**
WebP 40	66.9	**72.2**	Resize 0.50 + JPEG 50	63.8	**65.9**
WebP 30	65.0	**71.3**	Resize 0.75 + WebP 70	65.4	**69.9**
Blur 0.5	85.9	**91.9**	Resize 0.50 + JPEG 30	64.0	**65.1**
Blur 1.0	74.8	**76.8**	Resize 0.75 + WebP 50	62.0	**68.5**
Blur 1.5	**63.4**	61.8	Blur 0.5 + JPEG 70	68.4	**75.8**
Blur 2.0	**62.0**	59.3	Blur 1.0 + JPEG 50	63.6	**66.3**
Blur 2.5	**60.0**	59.3	Jitter 0.2 + JPEG 70	68.3	**77.4**
Blur 3.0	58.4	**59.1**	Jitter 0.6 + JPEG 50	63.3	**69.3**
Noise 5	80.9	**84.3**	Gray + JPEG 70	71.2	**78.0**
Noise 10	78.7	**80.3**	Blur 1.0 + Noise 5	69.3	**73.2**
Noise 15	76.3	**77.6**	Blur 1.5 + Noise 10	63.1	**64.0**
Noise 20	73.8	**74.7**	Resize 0.75 + JPEG 50 + Noise 5	70.5	**73.2**
Noise 25	71.2	**71.8**	Blur 1.0 + Noise 5 + JPEG 50	62.6	**64.5**
Noise 30	69.0	**69.1**	Resize 0.50 + Blur 0.5 + JPEG 50	62.8	**64.4**
			Overall (51 cond.)	70.5	**74.4**

Table S34: **Full per-condition AP (%): DINOv2 ViT-B/14 (B-Free).** Mean AP across 9 generators on LDMFakeDetect. Ours is PICS at $K{=}7\%$ ($76/768$ channels), the run of Table 1. Best per row in **bold**.

Condition	B-Free	Ours	Condition	B-Free	Ours
Clean	96.9	**97.6**	Jitter 0.2	96.8	**97.5**
JPEG 90	96.3	**97.1**	Jitter 0.4	96.4	**97.2**
JPEG 80	95.9	**96.9**	Jitter 0.6	96.1	**96.8**
JPEG 70	95.8	**96.8**	Jitter 0.8	95.7	**96.4**
JPEG 60	95.7	**96.6**	Jitter 1.0	95.0	**95.6**
JPEG 50	95.4	**96.4**			
JPEG 40	94.8	**95.8**	Resize 0.25	89.8	**92.4**
JPEG 30	94.6	**95.7**	Resize 0.5	94.7	**96.3**
WebP 90	96.6	**97.3**	Resize 0.75	95.8	**96.9**
WebP 80	96.5	**97.2**			
WebP 70	96.1	**96.8**	Grayscale	96.5	**96.9**
WebP 60	95.8	**96.4**			
WebP 50	95.3	**96.1**	Resize 0.75 + JPEG 70	95.4	**96.5**
WebP 40	94.7	**95.4**	Resize 0.50 + JPEG 50	94.6	**95.7**
WebP 30	93.7	**94.3**	Resize 0.75 + WebP 70	95.4	**96.2**
Blur 0.5	96.8	**97.5**	Resize 0.50 + JPEG 30	93.7	**94.8**
Blur 1.0	95.9	**96.7**	Resize 0.75 + WebP 50	94.4	**95.2**
Blur 1.5	95.5	**96.5**	Blur 0.5 + JPEG 70	95.6	**96.6**
Blur 2.0	94.8	**95.8**	Blur 1.0 + JPEG 50	94.5	**95.6**
Blur 2.5	94.1	**95.3**	Jitter 0.2 + JPEG 70	95.8	**96.7**
Blur 3.0	93.2	**94.4**	Jitter 0.6 + JPEG 50	94.9	**95.7**
Noise 5	96.1	**96.9**	Gray + JPEG 70	95.4	**96.1**
Noise 10	95.3	**96.2**	Blur 1.0 + Noise 5	94.9	**95.8**
Noise 15	94.7	**95.5**	Blur 1.5 + Noise 10	92.8	**93.5**
Noise 20	94.0	**94.8**	Resize 0.75 + JPEG 50 + Noise 5	94.9	**95.8**
Noise 25	93.2	**94.0**	Blur 1.0 + Noise 5 + JPEG 50	94.1	**95.1**
Noise 30	92.2	**92.9**	Resize 0.50 + Blur 0.5 + JPEG 50	94.6	**95.6**
			Overall (51 cond.)	95.1	**96.0**