REAL IMAGES, FAKE VERDICTS: SYNTHETIC-IMAGE DETECTORS FAIL INDEPENDENTLY ON FAKE-LOOKING REAL PHOTOGRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic-image detectors are evaluated almost exclusively on benchmarks whose real images are typical photographs. We study the opposite tail: genuine photographs that look fake to people — extreme lighting, improbable geometry, coincidences, unusual post-processing. We introduce FLORIDA, 795 such curated real photographs, and evaluate ten detectors across six families after a shared control-benchmark sanity check. Seven pass; six nevertheless flag FLORIDA as fake at $1.8\times$–$33\times$ their control false-positive rate (all $p < 10^{-4}$), up to 80% for the strongest detector. A provenance-matched control shows web provenance alone elevates FPR for five of seven detectors, while a genuine content effect survives matching for four of seven (up to $5.5\times$). The failures are also nearly independent across detectors (median $\phi = 0.04$): 95.5% of FLORIDA fools at least one detector, yet none fools all seven. One deployed detector shows no significant elevation, and retraining a linear probe with 396 fake-looking reals as hard negatives cuts FPR from 55.9% to 1.0%: robustness is attainable and cheap. A nine-way taxonomy of *why* images look fake finds no safe category (union FPR 91–98%), with each detector sensitive to a different one; fake-looking reals are more atypical in frequency space than the fakes themselves. "Fake-looking" is not one shared cue detectors accidentally learn, but a broad region of real-image space where each detector fails differently — with direct consequences for deploying detectors as evidence of authenticity.

1 INTRODUCTION

The reliability of AI-generated-image detection is usually summarized by accuracy on benchmarks that pair generated images with *typical* real photographs. Deployment, however, exposes detectors to the full distribution of real photography — including its strange tail: single-exposure photographs of rare coincidences, forced perspective, extreme lighting, heavy color grading, and improbable-but-real geometry (Figure 1). Such images are precisely the ones users submit to detection services ("is this real?"), and precisely the ones where a false "AI-generated" verdict is most damaging: extraordinary real photographs are the currency of photojournalism, and public accusations of fakery attach to exactly this material.

We make this tail measurable. We introduce FLORIDA (**F**ake-**LO**oking **R**eal **I**mage **DA**taset), 795 manually curated genuine photographs that viewers routinely doubt, and use it to answer three questions that determine whether the tail matters for deployment. **(1) Scope:** is misclassification of such images a quirk of particular architectures, or does it span modern detector families — CNN-artifact classifiers, CLIP-feature probes, frequency and upsampling artifact detectors, reconstruction-based diffusion detectors, and deployed commercial-style classifiers? **(2) Validity:** a detector can score poorly on FLORIDA simply by being a poor detector; without a simultaneous sanity check on a standard real/fake benchmark, "fails on FLORIDA" is uninterpretable. **(3) Structure:** do detectors fail on the *same* images — suggesting one shared spurious cue — or independently, suggesting the tail is broadly hard?

We answer all three with a controlled protocol. Every detector is evaluated on a balanced control benchmark (500 ordinary real photographs from COCO, 500 diffusion-generated images from Dif-

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



Figure 1: **Real photographs that look fake** — FLORIDA examples labeled with their category in our "why it looks fake" taxonomy (§4.6). Every image is a genuine, single-exposure photograph; $k/7$ = number of sanity-passing synthetic-image detectors that flag it as AI-generated at control-calibrated thresholds.

fusionDB and DALL-E 3) and on FLORIDA, using a single decision threshold calibrated on the control set and then held fixed. Detectors that fail the control sanity check (AUROC < 0.75; three pre-diffusion-era detectors fall here) are reported but excluded from inference, since a detector that calls *everything* real trivially "passes" FLORIDA.

Findings. Across the seven detectors that pass the sanity check:

- **The failure is general, not architectural.** Six of seven detectors show significantly elevated false-positive rates on FLORIDA relative to ordinary real photographs — CLIP-feature probes ($2.3\times$ and $33\times$), a state-of-the-art CNN artifact detector ($10\times$), a training-free diffusion-reconstruction detector ($1.8\times$), and two deployed transformer classifiers ($3.5\times$ and $4.3\times$); all $p < 10^{-4}$. The best detector (control AUROC 0.994, 2.4% control FPR) flags **80%** of FLORIDA at the same threshold.
- **It has two layers: provenance and content.** Re-encoding ordinary COCO images to FLORIDA's resolution and JPEG-quality profile (content unchanged) reproduces *all* of some detectors' elevation (Corvi-2023) but leaves a large significant content effect for four of seven — $5.5\times$ for CLIPDet (14.6% matched vs. 80% FLORIDA).
- **Failures are nearly independent across detectors.** Pairwise error correlations on FLORIDA are small (median $\phi = 0.04$, maximum 0.26) and co-error rates match the independence prediction (lift 0.96–2.0). Consequently 95.5% of FLORIDA images are flagged by at least one detector while no image fools all seven. "Fake-looking" is not one cue; it is a region of image space where every detector has its own failure surface.
- **Robustness is attainable — and cheap.** One deployed SwinV2 classifier (control AUROC 0.983) shows a statistically insignificant elevation (3.2% $\rightarrow$ 4.9%, $p = 0.09$); and adding 396 FLORIDA-train images as hard negatives to a linear probe on the CLIPDet backbone cuts held-out FLORIDA FPR from 55.9% to 1.0% with control AUROC unchanged (§4.9). The tail is not intrinsically hard — detectors have simply never seen it labeled real.
- **The failure has structure, and the structure differs per detector.** Labeling every image with the dominant reason it looks fake shows no category is safe (union FPR $\geq$91% in every category with $n \geq 20$), yet five of seven detectors are significantly category-sensitive — each to different categories — and in radial frequency space FLORIDA is more separable from ordinary real photographs (AUROC 0.74) than the actual fakes are (0.66).

Our contributions are: (i) FLORIDA, the first dataset making the fake-looking tail of real photography an explicit axis of detector evaluation, with per-image "why it looks fake" taxonomy labels; (ii) a measurement protocol — control-calibrated thresholds, sanity-check gating, independence analysis

— under which the tail can be measured meaningfully; and (iii) a characterization of how the current detector ecosystem behaves on it, released together with per-image scores for all ten detectors.

2 Related Work

Synthetic-image detection. Detection approaches divide roughly into (i) CNN classifiers trained on generator output, which generalize surprisingly well across GAN architectures (Wang et al., 2020) and, retrained on latent diffusion data with architectural care, across diffusion models (Corvi et al., 2023); (ii) detectors of low-level generation artifacts: spectral signatures of upsampling (Zhang et al., 2019; Durall et al., 2020), learned frequency-domain features (Tan et al., 2024a), and neighboring pixel relationships (Tan et al., 2024b); (iii) linear probes on frozen foundation-model features (Ojha et al., 2023; Cozzolino et al., 2024), which currently dominate cross-generator generalization; and (iv) reconstruction-based methods that exploit the autoencoder or inversion error of diffusion models themselves (Wang et al., 2023a; Ricker et al., 2024). Public detection services deploy fine-tuned transformer classifiers of type (i)/(iii). Our study samples all of these families.

Evaluation beyond typical benchmarks. Detector evaluations report strong in-distribution results, but robustness studies consistently find sharp degradation under distribution shift: post-processing and re-compression (Wang et al., 2020), cross-generator transfer (Ojha et al., 2023), and real-world web data. A complementary line asks when *real* images are misclassified: Grommelt et al. (2024) show that JPEG compression and resolution act as shortcut features separating "real" from "fake" in common training sets; Yan et al. (2025) study the converse tail to ours — generated images that look real to humans (Chameleon) — and find detectors collapse there too. The tail we study — real images that look fake — has to our knowledge not been made an explicit evaluation axis before: no existing benchmark curates it, and no prior work measures modern detector families on it under a calibrated, sanity-checked protocol or analyzes the cross-detector structure of the resulting errors.

Shortcut learning. Our motivating hypothesis — that detectors absorb "visual plausibility" as a proxy for authenticity — is an instance of shortcut learning (Geirhos et al., 2020): when the training distribution's real images are overwhelmingly ordinary, "looks unusual" becomes correlated with "generated." The near-independence of failures we observe (§4.4) refines this: different training distributions induce *different* plausibility shortcuts rather than one shared one.

3 Benchmark and Protocol

3.1 Datasets

FLORIDA. We collected 795 real photographs curated for appearing artificial: unusual lighting and color, improbable geometry and perspective, rare coincidences, and striking natural phenomena. Candidates were gathered from public web communities where viewers explicitly discuss photographs as unbelievable — photography forums, Q&A platforms, and curated collections of extraordinary photography — and manually screened, retaining only genuine photographs (excluding composites, staged digital manipulations, and generated content) whose appearance plausibly triggers doubt. Each image additionally carries a primary "why it looks fake" category from a nine-way taxonomy (§4.6, Appendix A.4). Integrity checks: 795/795 decodable, no exact duplicates by MD5; smallest side 184 px, largest image 2448×3264, median $\sim700\times619$. All images are real; every "fake" verdict on FLORIDA is a false positive.

Control benchmark. A detector could "fail" on FLORIDA by being broken. We therefore measure every detector on a balanced control set: 500 real photographs sampled uniformly from COCO `val2017` (Lin et al., 2014) and 500 generated images (250 Stable Diffusion 1.x prompts from DiffusionDB (Wang et al., 2023b), 250 DALL-E 3 images (Christoforu, 2023)). The control set plays two roles: it gates which detectors are interpretable at all, and it supplies the decision threshold.

3.2 Detectors

We evaluate ten publicly available detectors spanning six families, using each author's released weights and test-time pipeline (verified against the original inference code): **Corvi et al. 2023** (Corvi et al., 2023), a ResNet-50 without downsampling in the first block, trained on latent-diffusion data (CNN-artifact family); **NPR** (Tan et al., 2024b) (upsampling-artifact) and **FreqNet**

Table 1: **Main result.** Synthetic-image detectors evaluated on the control benchmark (500 COCO real / 500 generated images) and on FLORIDA (795 fake-looking *real* photographs). FPR = fraction of real images flagged as fake at the threshold that maximizes balanced accuracy on the control set (held fixed for FLORIDA). Brackets: bootstrap 95% CIs (2,000 resamples). Ratio = FLORIDA FPR / control-real FPR. Significance of the elevation by one-sided Fisher exact test: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$. Gray: detectors failing the sanity check (control AUROC < 0.75); their FLORIDA scores are vacuous.

Detector	Family	Control AUROC	Control real FPR (%)	FLORIDA FPR (%)	Ratio
CLIPDet (latent10k+)	CLIP feature	0.994	2.4	80.0^{***} [77.1, 82.8]	33.3×
Organika SDXL	Deployed (Swin)	0.903	12.2	42.9^{***} [39.4, 46.4]	3.5×
UnivFD	CLIP feature	0.776	17.6	40.0^{***} [36.6, 43.5]	2.3×
AEROBLADE	Diffusion reconstruction	0.859	19.6	35.7^{***} [32.4, 39.1]	1.8×
Ateeqq	Deployed (SigLIP)	0.957	7.4	31.6^{***} [28.4, 34.8]	4.3×
Corvi et al. 2023	CNN artifact	0.994	2.0	20.1^{***} [17.3, 22.8]	10.0×
Haywoodsloan	Deployed (SwinV2)	0.983	3.2	4.9 [3.4, 6.5]	1.5×
umm-maybe	Deployed (Swin)	0.551	8.4	16.4^{***} [13.8, 18.9]	1.9×
FreqNet	Frequency	0.540	4.0	0.3 [0.0, 0.6]	0.1×
NPR	Upsampling artifact	0.570	0.0	0.0 [0.0, 0.0]	0.0×

(Tan et al., 2024a) (learned frequency features), both GAN-era; **UnivFD** (Ojha et al., 2023), a linear probe on frozen OpenAI CLIP ViT-L/14 features, and **CLIPDet** (Cozzolino et al., 2024) (`latent10k_plus`), a linear head on CommonPool-pretrained ViT-L/14 (CLIP-feature family); **AEROBLADE** (Ricker et al., 2024), training-free reconstruction error through Stable Diffusion autoencoders (diffusion-reconstruction family); and four **deployed** open transformer classifiers from HuggingFace with 17k–150k downloads (Organika SDXL-detector and umm-maybe, Swin; Ateeqq, SigLIP; Haywoodsloan, SwinV2), standing in for detection-as-a-service systems. As a smaller-scale preliminary check, we also manually evaluated two consumer-facing platforms through their ordinary upload interfaces rather than downloaded weights; both performed poorly (Appendix A.11). A fully calibrated version of this idea — a vision–language model's zero-shot real-vs-AI-generated verdict, evaluated as an eleventh detector under the identical protocol of §3.3 — is reported in Appendix A.12. A separate blind study with ordinary human raters, measuring the same elevation directly rather than through a model proxy, is reported in Appendix A.13.

3.3 Protocol and statistics

For each detector we compute a score $s(x) \in [0, 1]$ (probability-of-fake) per image. On the control set we report AUROC; detectors with AUROC < 0.75 fail the sanity check and are excluded from inference (they are shown grayed in Table 1). For each remaining detector we fix the threshold t^* that maximizes balanced accuracy on the control set, then apply it unchanged to FLORIDA. We report the false-positive rate $\mathrm{FPR} = \Pr[s(x) > t^*]$ on control-real and on FLORIDA, with bootstrap 95% confidence intervals (2,000 resamples) and a one-sided Fisher exact test for elevation. As a secondary operating point we also report each detector's native threshold (0.5 unless the author specifies otherwise; Appendix A.1). For cross-detector structure we form the binary error matrix on FLORIDA and compute pairwise co-error rates against the independence baseline $\hat{p}_a\hat{p}_b$ (lift) and ϕ correlations.

4 Results

4.1 Detectors that fail the sanity check

NPR, FreqNet, and umm-maybe reach control AUROC of only 0.54–0.57: trained against GAN generators (or, for umm-maybe, early art-focused generators), they do not detect modern diffusion output. The near-zero FLORIDA FPR of NPR and FreqNet is therefore vacuous — a detector that never says "fake" cannot be fooled — and umm-maybe's verdicts are uninterpretable at chance-level control discrimination; we exclude all three from subsequent analysis. This gating is essential: without it, the two never-fire detectors would appear to be the two most robust (Appendix Figures 6 and 7 show the FPR shifts and the most-flagged images).

Table 2: **Content vs. provenance.** FPR (%) at the original calibrated thresholds on ordinary COCO images (control), on the same images re-encoded to match FLORIDA's resolution and JPEG-quality distribution (matched), and on FLORIDA. Residual = FLORIDA FPR / matched FPR — the elevation attributable to *content* beyond web provenance; significance by one-sided Fisher exact test: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

Detector	Control FPR	Matched FPR	FLORIDA FPR	Content residual
CLIPDet	2.4	14.6	80.0	$5.5\times^{***}$
Ateeqq	7.4	10.2	31.6	$3.1\times^{***}$
UnivFD	17.6	29.2	40.0	$1.4\times^{***}$
Organika	12.2	31.4	42.9	$1.4\times^{***}$
AEROBLADE	19.6	32.8	35.7	$1.1\times$ (n.s.)
Haywoodsloan	3.2	4.6	4.9	$1.1\times$ (n.s.)
Corvi et al. 2023	2.0	19.8	20.1	$1.0\times$ (n.s.)

4.2 Elevated false-positive rates on fake-looking real images

Table 1 shows the main result. Every sanity-passing detector except Haywoodsloan flags FLORIDA images at a significantly higher rate than ordinary real photographs (one-sided Fisher exact, all $p < 10^{-4}$): CLIPDet rises from 2.4% to 80.0% (33×), Corvi-2023 from 2.0% to 20.1% (10×), Ateeqq from 7.4% to 31.6% (4.3×), Organika from 12.2% to 42.9% (3.5×), UnivFD from 17.6% to 40.0% (2.3×), and AEROBLADE from 19.6% to 35.7% (1.8×). The two most severe failures are the two detectors with the *best* control AUROC (0.994 both), so the failure is not an artifact of weak detectors: the better a detector separates ordinary real from fake, the more of that separation appears to rest on features that fake-looking real images violate. The AEROBLADE result is mechanistically notable: it uses no classifier at all, scoring images by their reconstruction error through Stable Diffusion autoencoders — fake-looking real photographs literally sit closer to the diffusion autoencoder manifold than pristine COCO photographs do (though §4.3 shows this proximity is largely conferred by web re-encoding rather than content).

The elevation is robust to the choice of operating point. At native thresholds ($t = 0.5$), CLIPDet still flags 36.9% of FLORIDA versus 0.4% of control real images (92×), Ateeqq 36.6% versus 10.4%, and Organika 31.7% versus 6.8%. Corvi-2023's native threshold is very conservative (FPR 0.0% control, 0.6% FLORIDA) at the cost of control accuracy (85.8% versus 96.6% calibrated), illustrating that the visible failure depends on where a deployment sets its operating point — but the score distributions themselves have shifted wholesale (Figure 2a): FLORIDA scores spread across the full range and largely track the *generated* distribution rather than the ordinary-real one, so no threshold separates them.

4.3 Content or provenance? A matched control

FLORIDA is web imagery with heterogeneous resolution and compression, whereas our control reals come from COCO; and compression and resolution are known shortcut features for real/fake classifiers (Grommelt et al., 2024). To separate the two, we re-encode each control COCO image to match FLORIDA's provenance statistics — rescaled (seeded, aspect-preserving) to a FLORIDA-sampled pixel area and JPEG-compressed at that image's quality, estimated from its quantization tables — leaving content untouched, and score the result at the *unchanged* calibrated thresholds (Table 2).

Both factors are real, and they matter differently per detector. Provenance alone significantly elevates FPR for five of seven detectors; for Corvi-2023 it explains essentially the *entire* FLORIDA elevation (19.8% matched vs. 20.1% FLORIDA, n.s.), and for AEROBLADE most of it (residual $1.1\times$, n.s.). But a significant content effect survives provenance matching for four of seven detectors — most strikingly the strongest one: CLIPDet rises from 2.4% to 14.6% under matching alone, yet FLORIDA content pushes it to 80.0%, a $5.5\times$ residual ($p < 10^{-3}$); Ateeqq's elevation is mostly content ($3.1\times$ residual). Haywoodsloan is immune to both. The failure therefore has two layers — a web-provenance shortcut and a content-typicality shortcut — and which layer dominates depends on the detector. For deployment the distinction is immaterial (extraordinary photographs arrive *with* their provenance); for understanding detectors it is central, and subsequent sections use it.

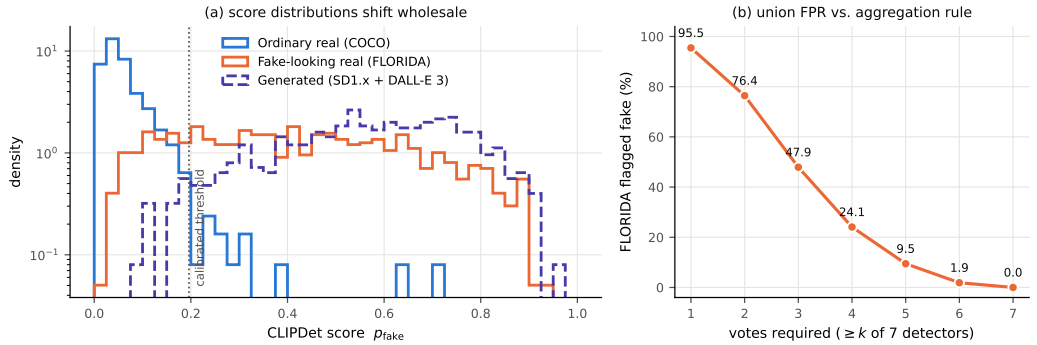


Figure 2: **(a)** CLIPDet score distributions (log density): FLORIDA spreads across the full score range and tracks the generated distribution, so the failure is not an artifact of threshold placement. **(b)** Fraction of FLORIDA flagged as a function of the ensemble aggregation rule: any-fires policies misfire on 95.5% of the tail, unanimity on none.

Table 3: textbfWhy the images look fake, and how often detectors agree. Each FLORIDA image received one primary category (VLM-assisted labeling, rubric in Appendix A.4). Union = flagged fake by $\geq$1 of the 7 sanity-passing detectors at calibrated thresholds; mean = average number of detectors (of 7) flagging an image in the category.

Primary category	n	Share	Union FPR	Mean #detectors
Atypical subject	193	24%	95%	2.83
Geometry / scale	189	24%	97%	2.42
Coincidence / timing	178	22%	95%	2.47
Lighting / sky	70	9%	97%	2.76
Reflection / transparency	63	8%	97%	2.35
Texture too perfect	45	6%	91%	2.33
Color grading	41	5%	98%	2.59
Motion / long exposure	10	1%	100%	2.30
Other	4	1%	50%	1.25

4.4 Failures are nearly independent across detectors

Raw error overlap between detectors is substantial (Jaccard up to 0.38), but this is exactly what high marginal error rates predict: observed co-error rates match the independence baseline closely (lift 0.96–2.0 across all 21 pairs), and pairwise ϕ correlations are small (median 0.04, maximum 0.26). In other words, **no shared "fake-looking" cue drives the failures**: each detector is fooled by a largely different subset of FLORIDA. The consequences are stark in both directions: 95.5% of FLORIDA images are flagged as AI-generated by at least one of the seven well-functioning detectors, 76.4% by at least two, 47.9% by at least three — yet not a single image fools all seven, and only 1.9% fool six (Figure 2b). An ensemble's behavior on this tail therefore depends critically on its aggregation rule: any-detector-fires policies (common in moderation pipelines) misfire on essentially the whole tail, while unanimity rules would be nearly immune. The full per-image binary error matrix and pairwise ϕ-correlations underlying these figures are released and visualized directly in Appendix A.10 (Figure 11).

4.5 Robustness is attainable

Haywoodsloan (SwinV2) combines strong control performance (AUROC 0.983, 95.3% accuracy) with a statistically insignificant elevation on FLORIDA (3.2% $\rightarrow$ 4.9%, $p = 0.09$). Whatever combination of its training distribution and augmentation produced this, it demonstrates that the fake-looking tail is not intrinsically undecidable — unusual-but-real is learnable as *real* without sacrificing fake-detection, a proof of existence that motivates hard-negative training with unusual real photographs.

4.6 No category of "fake-looking" is safe, but each detector has its own weakness

To ask *which kinds* of fake-looking images fool detectors, we assign every FLORIDA image one primary category from a nine-way taxonomy of the dominant reason it reads as fake (VLM-assisted labeling under a fixed rubric; taxonomy, protocol, and released labels in Appendix A.4). Three causes dominate the dataset: atypical subjects (24%), improbable geometry or scale (24%), and rare coincidences (22%), with extreme lighting, reflections, too-perfect texture, and heavy color grading covering the rest (Table 3).

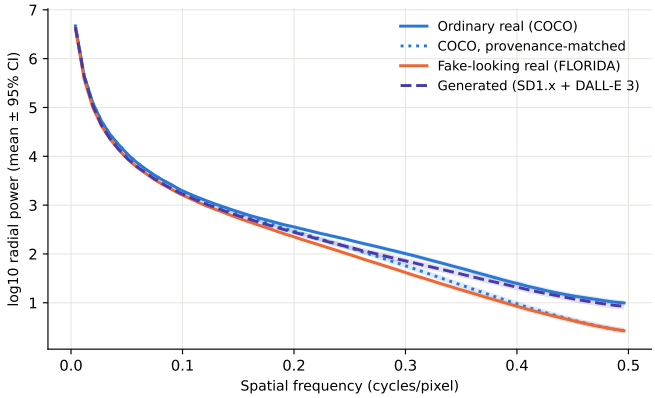


Figure 3: **Radial power spectra.** Azimuthally averaged log power (center square crop resized to 512 px, grayscale, Hann-windowed FFT, 64 radial bins), mean ± 95% CI per group. Modern generated images track ordinary real photographs closely, while FLORIDA images carry markedly less high-frequency power; provenance-matched COCO (dotted) closes most, but not all, of that gap.

Two facts stand out. First, **no category is safe**: the union false-positive rate is 91–98% in every category with $n \geq 20$, and the mean number of detectors fooled per image varies only from 2.3 to 2.8 across categories. Second, individual detectors are nonetheless **category-sensitive in different ways** (Figure 8, Appendix): flag rate depends significantly on category for five of seven detectors (χ^2 over $n \geq 20$ categories, p from $4{\times}10^{-5}$ to 0.04), but the weak spots differ — CLIPDet peaks on color-graded (95%) and atypical-subject (87%) images, AEROBLADE and Corvi-2023 on rare coincidences (42% and 28%, their respective maxima), Ateeqq on atypical subjects (43%), while UnivFD and Organika are category-indifferent ($p = 0.99, 0.67$). Even the robust Haywoodsloan is structured: its few errors concentrate on atypical subjects (8.8%) and it never fires on a coincidence image (0/178). This refines the independence result of §4.4: detectors are not indifferent to *why* an image looks fake — they are sensitive to *different whys*, which is precisely what makes their failures union so completely.

4.7 Fake-looking real images are spectrally more atypical than fakes

A common account of detector generalization is that generated images carry low-level spectral artifacts. Figure 3 shows the opposite ordering on our data: the mean radial power spectrum of our diffusion-era fakes is nearly indistinguishable from that of ordinary real photographs, whereas FLORIDA images have systematically *depressed* high-frequency power. A logistic probe on the 64-bin radial profile separates FLORIDA from ordinary real photographs at AUROC $0.74{\pm}0.05$ (5-fold CV), *better* than it separates actual generated images from real ones ($0.66{\pm}0.04$): in the very feature space that artifact-based detection is assumed to exploit, the hard tail of real photography is more anomalous than the fakes are. Consistent with §4.3, this spectral gap itself decomposes: provenance-matched COCO reproduces most of it (dotted line; probe AUROC 0.84 vs. original COCO), and FLORIDA-vs-matched separability drops to 0.63 ± 0.04 — a residual carried by content (long exposures, smooth gradients, fog, and minimal-texture scenes are exactly the photographs that look rendered). Either way the mechanism stands: any detector whose decision leans on smoothness or missing high-frequency texture — and AEROBLADE's reconstruction error is one by construction — will systematically misread this tail, whether the smoothness comes from the web pipeline or from the scene itself.

4.8 Detectors do not look at the anomaly

If detectors misread fake-looking photographs because of the anomaly a human reacts to, their decision evidence should be spatially concentrated on it. We test this for three mechanistically distinct detectors, pairing each with the attribution native to its architecture: Grad-CAM at the last convolutional block for Corvi-2023; occlusion maps for CLIPDet (a $16{\times}16$ grid of gray patches on its 224px input — its frozen CLIP backbone precludes gradient attribution); and the spatial LPIPS reconstruction-error map for AEROBLADE, whose score *is* that map's mean. Each map is compared against bottom-up visual saliency (spectral residual; Hou & Zhang, 2007) computed on the identical input frame, by rank correlation over a common $64{\times}64$ grid, with a shuffled-pair null.

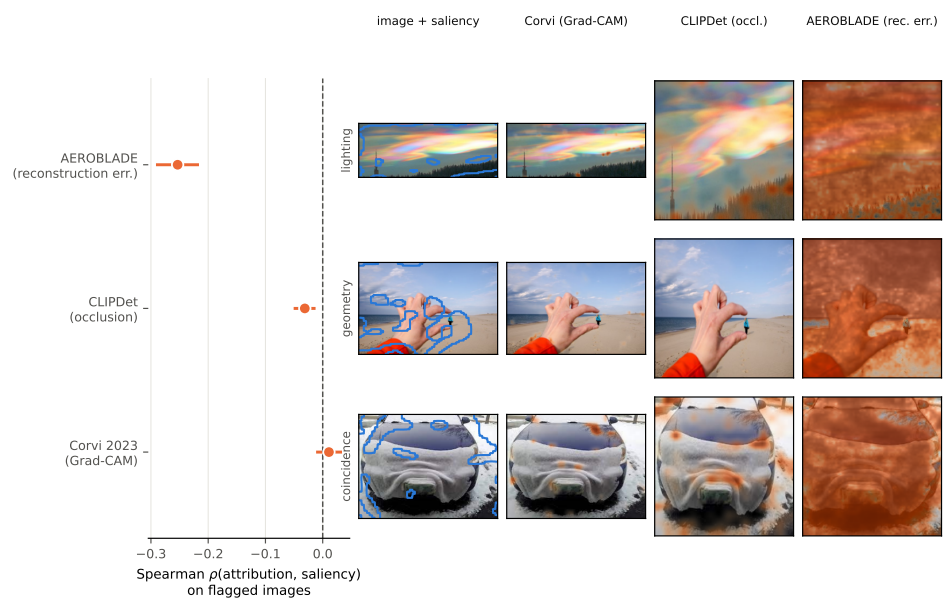
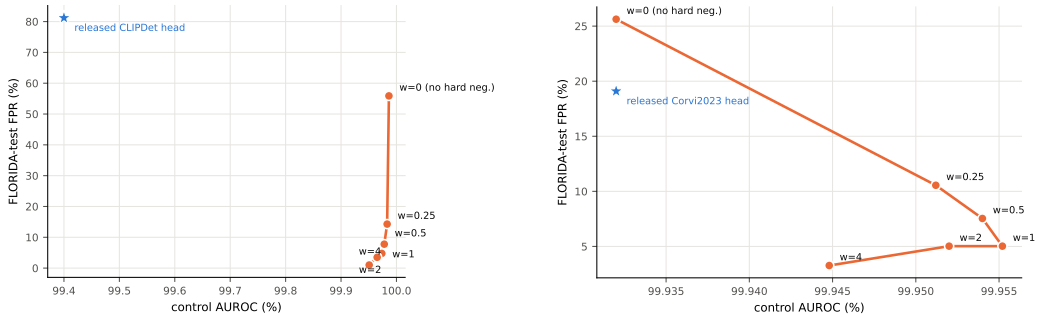


Figure 4: **Attribution vs. visual saliency.** Left: mean Spearman correlation between each detector's spatial attribution and bottom-up saliency on the FLORIDA images it flags (bootstrap 95% CIs; dashed line = shuffled-pair null). Right: examples — original with the top-quintile saliency contour (blue), then each detector's attribution overlaid on the input frame it actually sees. AEROBLADE's "fake" evidence concentrates in smooth, low-salience regions.

The answer is no (Figure 4). On flagged images, Corvi-2023 ($\rho = 0.01$ vs. null -0.00) and CLIPDet ($\rho = -0.03$ vs. null -0.02) are statistically indistinguishable from chance: whatever drives their verdicts, it is not spatially co-located with the image's salient anomaly. AEROBLADE is significantly *anti*-correlated ($\rho = -0.25$; top-quintile IoU 0.079 vs. 0.111 chance), consistently in every taxonomy category (ρ from -0.13 to -0.42): its evidence for "fake" concentrates precisely in smooth, low-salience regions — sky, sand, fog — which reconstruct well through diffusion autoencoders, the spatial counterpart of the depressed high-frequency spectra of §4.7. Together with the near-independence of failures, this paints a coherent picture: fake-looking real photographs are not misclassified because detectors "see" the impossible hand or the too-perfect timing. They are misclassified on global, low-level statistics that happen to co-occur with looking fake — statistics humans do not attend to at all.

4.9 The content failure is fixable with data alone

Haywoodsloan proves robustness is *attainable*; we now show it is *cheap*. We split FLORIDA 50/50 (stratified by taxonomy category, seeded; 396 train / 399 test) and train logistic probes on CLIPDet's frozen backbone using fresh data disjoint from the evaluation benchmark (1,000 COCO reals, 500 SD/DALL-E fakes). The base probe (w=0) reproduces the phenomenon — control AUROC 0.9999 yet 55.9% FPR on FLORIDA-test — confirming the failure is a property of the feature space and training distribution, not of the released head's particular data. Adding the 396 FLORIDA-train images as real hard negatives collapses the failure: 14.3% FPR at w=0.25, 4.8% at w=1, and **1.0%** at w=4, while control AUROC stays above 0.9995 and control accuracy at or above 98.6% (Figure 5a; the released head scores 81.2% on the same test images). Four hundred unusual real photographs and a linear probe suffice: the fake-looking tail is not intrinsically hard — current detectors have simply never seen it labeled *real*. The recipe is not CLIP-specific either: repeating it on Corvi et al. 2023's frozen CNN backbone (an architecturally unrelated detector family) cuts FLORIDA-test FPR from 25.6% to 3.3% at the same negligible cost (Figure 5b; full protocol in Appendix A.8). Nor does the fix require having seen the specific *kind* of fake-looking image: leaving an entire taxonomy category out of training entirely, it still collapses FPR on that unseen category to near zero (Appendix A.9).



(a) CLIPDet (CLIP ViT-L/14 probe)

(b) Corvi et al. 2023 (CNN backbone)

Figure 5: **Hard-negative training erases the failure at negligible cost, in two unrelated detector families.** Linear probes trained on data disjoint from the evaluation benchmark, with FLORIDA-train reals added as hard negatives at weight w. Thresholds calibrated on the untouched control benchmark; FPR on held-out FLORIDA-test. **(a)** CLIPDet: a linear probe on frozen CLIP ViT-L/14 features; w=0 reproduces the phenomenon (55.9% FPR at 0.9999 control AUROC), w=4 collapses it to 1.0%. **(b)** Corvi et al. 2023's frozen ResNet-50 backbone; w=0 gives 25.6% FPR at 0.999 control AUROC, w=4 collapses it to 3.3% — the same recipe, on an architecturally unrelated CNN-artifact detector.

5 Discussion

What do detectors actually measure? The near-perfect control AUROC of CLIPDet and Corvi-2023 alongside 33×/10× FPR elevation suggests these models measure something correlated with, but not identical to, generation: a mixture of true generator artifacts, web-provenance statistics (§4.3), and image *typicality*. The independence structure indicates these shortcuts are not a single direction in feature space shared across models; each is induced separately by each training distribution — a distributed form of shortcut learning (Geirhos et al., 2020).

Deployment implications. Detection services return verdicts on exactly the images users find unbelievable, and there a randomly chosen well-functioning detector returns a false "AI-generated" verdict with probability 5–80%; any-fires aggregation makes it worse (95.5% union FPR). Verdicts on extraordinary real photographs should be treated as near-uninformative by current tools, and interfaces should communicate uncertainty on atypical inputs rather than binary verdicts. Nor is the difficulty confined to detection: segmentation and face detection qualitatively degrade on the same tail (Appendix A.6).

Limitations. Our control fakes are built from two diffusion generators (SD 1.x, DALL-E 3); detectors tuned to other generators may be under-rated by the sanity check, though adding a newer generator (SDXL) as a robustness check leaves both the sanity check and the main result intact (Appendix A.7). The provenance-matched control (§4.3) approximates FLORIDA's compression via quality estimated from quantization tables and introduces a second JPEG generation on COCO images, so it bounds rather than perfectly reproduces the provenance effect. FLORIDA is modest in size (795 images), and our "why it looks fake" category labels (§4.6) were produced by a vision–language model under a fixed rubric rather than by human annotators; they support the coarse category-level analysis reported here and agree with independent human raters at the human–human ceiling on a validation subsample (Appendix A.5). We evaluate open detectors; commercial APIs remain untested. Finally, the attribution analysis of §4.8 uses a single classical bottom-up saliency reference; gaze-prediction models and human annotations of *what* looks wrong per image are natural refinements.

6 Conclusion

Evaluated under a calibrated, sanity-checked protocol, the modern synthetic-image detection ecosystem systematically misclassifies fake-looking real photographs, at rates up to 33× baseline, with failures distributed nearly independently across detectors rather than concentrated on a shared cue; one robust detector and a 396-image hard-negative fix show the tail is learnable. We release code, per-image scores for ten detectors, and the control benchmark to make the real-image tail a standard axis of detector evaluation.

REPRODUCIBILITY STATEMENT

All detectors are publicly released checkpoints run with their original test-time pipelines; the control benchmark is built by deterministic seeded sampling from public datasets (COCO val2017, DiffusionDB, DALL-E 3 images); we release FLORIDA with its taxonomy labels. Our evaluation harness, per-image score files, analysis scripts (bootstrap, Fisher, independence), and exact package versions are included in the supplementary material.

ETHICS STATEMENT

This work measures false accusations of fakery against real photographs, an error mode that damages photojournalists and ordinary users. Publishing detector failure rates could conceivably help adversaries choose detectors to evade; we judge this risk low (the failure concerns real images, not generated ones) and outweighed by the deployment-reliability findings. FLORIDA consists of photographs publicly posted online; the dataset is released for research use with the underlying images remaining the property of their photographers, and we honor removal requests.

AI USE STATEMENT

We used generative AI tools in three capacities. First, as a component of the research methodology: a vision–language model (Claude) produced the FLORIDA "why it looks fake" taxonomy labels (§4.6, Appendix A.4) and served as one zero-shot detector baseline (Appendix A.12); both uses are described in the main text and validated against independent human raters (Appendix A.5). Second, for code and analysis: an AI coding assistant (Claude Code) was used to write and edit analysis and plotting scripts, and to typeset and format the manuscript (LaTeX figures, tables, and layout). Third, for writing: an AI assistant was used to help draft and edit portions of the manuscript text. All AI-assisted code, figures, and text were reviewed and verified by the authors, who take full responsibility for the final content of this work, including all claims, analyses, and artifacts.

REFERENCES

Evgeniy Christoforu. DALL-E 3 images dataset. `https://huggingface.co/datasets/ehristoforu/dalle-3-images`, 2023.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of AI-generated image detection with CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or JPEG? revealing common biases in generated image detection datasets. *arXiv preprint arXiv:2403.17608*, 2024.

Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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 *European Conference on Computer Vision (ECCV)*, 2014.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024a.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023a.

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023b.

Wei Wu, Hanyang Peng, and Shiqi Yu. YuNet: A tiny millisecond-level face detector. *Machine Intelligence Research*, 20:656–665, 2023.

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

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2019.

A Appendix

A.1 Native-threshold operating points

§3.3 calibrates each detector's decision threshold on our control benchmark; Table 4 instead reports the same control/FLORIDA FPR comparison at each detector's own *native* threshold ($t = 0.5$ unless the author specifies otherwise, AEROBLADE excluded since it has no native threshold). The elevation pattern is qualitatively similar to the calibrated result (§4), confirming the main finding is not an artifact of our particular choice of operating point.

A.2 Main result, graphically

Table 1 (§4) reports the control-real-to-FLORIDA FPR shift numerically for all ten detectors; Figure 6 plots the same shift for the seven sanity-passing ones, making the size and significance of each detector's elevation visually comparable at a glance.

Table 4: Native-threshold ($t = 0.5$) operating points for all detectors with a native threshold. AEROBLADE has no native threshold (training-free distance) and appears only in the calibrated analysis.

Detector	Control AUROC	Control acc.	Control real FPR (%)	FLORIDA FPR (%)
clipdet_latent10k_plus	0.994	0.834	0.4	36.9
ateeqq	0.957	0.879	10.4	36.6
organika_sdxl	0.903	0.808	6.8	31.7
ummmaybe	0.551	0.554	29.6	30.4
npr	0.570	0.547	46.8	7.5
univfd	0.776	0.632	2.0	7.3
haywoodsloan	0.983	0.953	3.2	4.9
corvi2023	0.994	0.858	0.0	0.6
freqnet	0.540	0.528	19.2	0.5

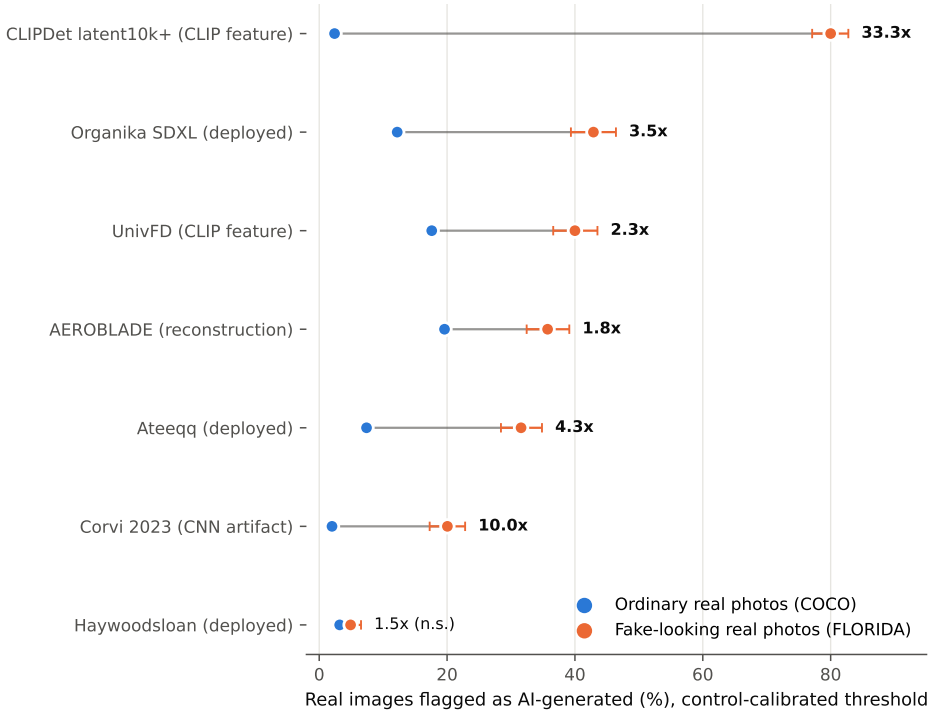


Figure 6: **False-positive rate shift from ordinary to fake-looking real photographs** (the content of Table 1, graphically), for the seven detectors passing the control sanity check (AUROC ≥ 0.75). Each detector's decision threshold is calibrated once on the control benchmark and held fixed. Whiskers: bootstrap 95% CIs. Bold ratios: significant elevation (one-sided Fisher exact, $p < 10^{-4}$).

A.3 Most-flagged images

As a qualitative complement to Figure 6's aggregate rates, Figure 7 shows the specific FLORIDA images that fooled the largest number of sanity-passing detectors at once — a direct look at what "elevated false-positive rate" means for individual photographs.

A.4 "Why it looks fake" taxonomy and labeling protocol

Each FLORIDA image received one primary category (the dominant reason a viewer would doubt the photograph is real), an optional secondary category, a confidence level, and a $\leq$15-word ra-

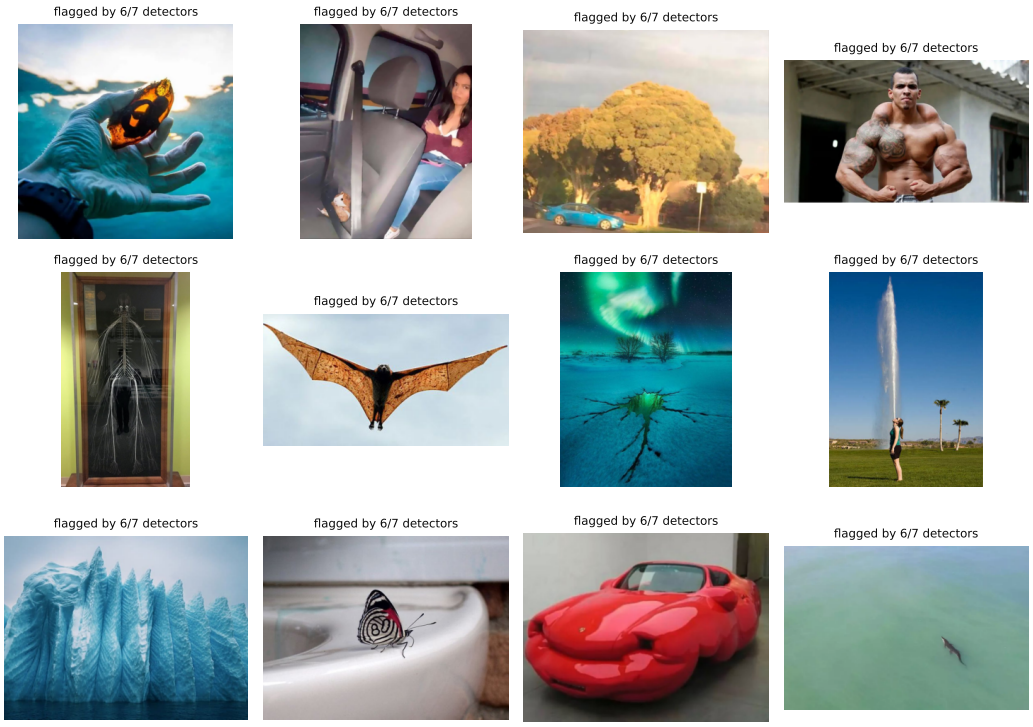


Figure 7: The twelve FLORIDA images flagged as AI-generated by the most sanity-passing detectors. All are real photographs.

tionale. Labels were produced by a vision–language model (Claude) viewing each image under the fixed rubric in Table 5 below, in twelve independent batches; the full label file is released as `results/tags_merged.csv`.

We validate these labels against independent human raters in Appendix A.5.

Table 5: Primary-category rubric.

Category	Definition
Lighting / sky	extreme or unusual light: auroras, lightning, god rays, harsh neon, impossible-looking sun or sky
Geometry / scale	improbable geometry, perspective, or scale: forced perspective, tilted horizons, objects appearing to defy physics
Coincidence / timing	perfect timing or rare juxtaposition: split-second capture, aligned objects, one-in-a-million framing
Color grading	unusual color grading or saturation: tone-mapped, HDR-overcooked, monochrome or candy-colored scenes
Atypical subject	the subject itself is rare: bizarre animals or plants, extreme body modification, unusual vehicles, architecture, food
Reflection / transparency	mirrors, reflections, transparency, or camouflage reading as compositing errors
Motion / long exposure	long exposure or motion blur; frozen liquids or smoke that look rendered
Texture too perfect	surfaces too smooth, regular, or perfect; reads as a CGI render
Other	none of the above (used sparingly)

§4.6 reports, per detector, which one or two categories draw its worst flag rate; Figure 8 shows the full per-category breakdown behind that summary, for every sanity-passing detector and every category with $n \geq 20$.

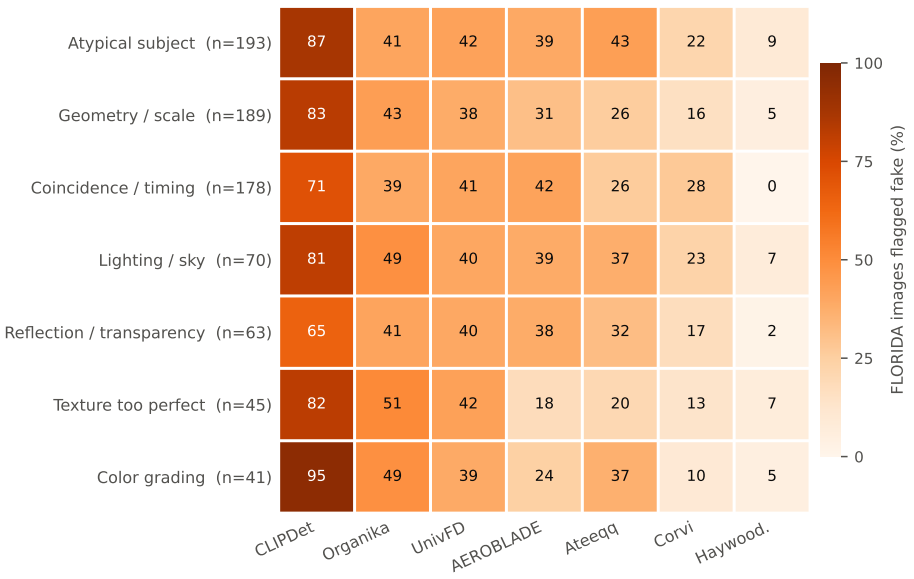


Figure 8: **Per-category false-positive rates** (% of category flagged fake at calibrated thresholds), for categories with $n \geq 20$. Detectors ordered by overall FLORIDA FPR. Flag rates depend significantly on category for five of seven detectors, but the sensitive categories differ per detector.

A.5 Human validation of the taxonomy

The "why it looks fake" category labels (§4.6) are produced by a vision–language model rather than by human annotators. To check that these categories track human judgement, we ran a small blind human study.

Protocol. We drew a stratified random subsample of 100 FLORIDA images (seeded; proportional to primary category, minimum two per category) and rendered them into a single self-contained labeling page (Figure 9). For each image the rater picks the *dominant* reason the real photograph could be doubted, choosing among the nine taxonomy categories (Table 5); the model's own label is never shown, so the rater is blind. Two independent raters each labeled all 100 images. The labeling page, the answer key, and the scorer are released under `tagging/validation/` and `scripts/score_validation.py`.

Results. Exact agreement between a rater and the model's primary label was 47% and 37% for the two raters (Cohen's $\kappa = 0.35$ and 0.26, bootstrap 95% CI $[0.23, 0.47]$ and $[0.16, 0.37]$). This is on par with the human–human ceiling itself — the two raters agreed with *each other* only 38% of the time ($\kappa = 0.27$) — which indicates that the residual disagreement reflects the genuine subjectivity of a fine-grained nine-way scheme rather than model error. The confusion is concentrated in semantically adjacent categories (atypical subject vs. geometry vs. coincidence), not spread uniformly. Accordingly we also report top-2 agreement, counting a match when a rater's category is the model's primary *or* secondary label: this rises to 67% and 58%. We therefore read the model labels as a reliable coarse index of *why* an image looks fake, adequate for the category-level analyses in this paper, while acknowledging that the exact primary category is inherently ambiguous for a minority of images.

A.6 Beyond detection: other vision tasks on the fake-looking tail

The difficulty of FLORIDA is not specific to synthetic-image detection: segmentation and face detection also degrade on this material, which we demonstrate with current open models under a documented procedure. Candidate images were selected by searching the released taxonomy ratio-

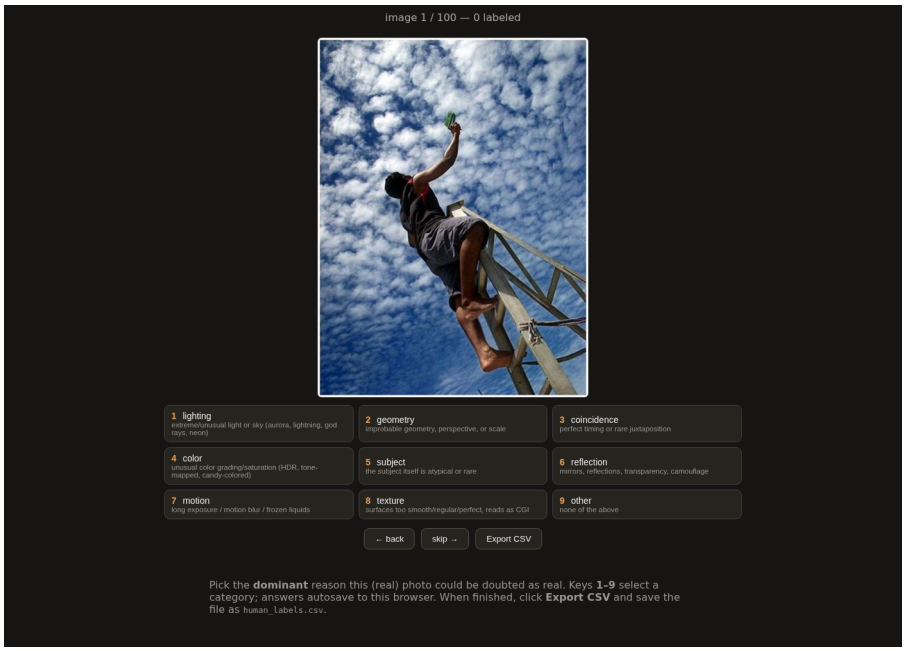


Figure 9: **Blind human-validation interface.** Each rater sees one image at a time and selects the dominant reason it could be doubted as real, from the nine taxonomy categories (with their definitions); the model's label is never displayed. Answers autosave in the browser and are exported to a CSV scored against the held-out model labels.

nales for descriptions of merged or overlapping subjects and occluded heads; the shown examples were then curated by inspecting model outputs, so this is a qualitative demonstration, not an evaluation (FLORIDA carries no segmentation or face ground truth). Figure 10 shows the failures. SAM's automatic mask generation (Kirillov et al., 2023) segments appearance rather than objects on these images: two hugging players become shirt, hair, and leg fragments with their merged torsos left as a single unlabeled region; a child riding on an adult's shoulders is segmented as *one* person; two overlapping dogs are treated as a single two-headed object. YuNet (Wu et al., 2023), a widely deployed face detector, finds no faces in images containing several clearly visible ones (partially occluded by other people or unusual poses), and in a vintage photograph detects two of five frontal faces. The theme is consistent with our main results: models tuned on typical imagery inherit its regularities, and the fake-looking tail of real photography violates those regularities across every task we examined.

A.7 Robustness to a newer generator (SDXL)

The control fakes in §3 come from two diffusion generators current at the time of dataset construction (SD 1.x, DALL-E 3); a natural concern is that the sanity check and main result are specific to those generators and would not hold against a newer, higher-fidelity model. We add 250 SDXL-base-1.0 images to the control fake set, generated from fresh prompts sampled from DiffusionDB's prompt metadata (disjoint from the prompts underlying the original SD 1.x images), 20 denoising steps, one seeded generation per prompt.

Table 6 shows every sanity-passing detector still separates SDXL images from ordinary real photographs well above the sanity threshold on their own (sdxl-only AUROC ≥ 0.75 for all seven, up to 0.999 for Corvi-2023), and adding SDXL to the control fake set leaves the extended AUROC close to the original for every detector. Recalibrating the decision threshold on this extended (750-fake) control set changes the FLORIDA FPR only marginally (within 3 percentage points for six of seven detectors), and the elevation remains significant for all six detectors that were significant under the original protocol: broadening the control generator distribution does not explain away the main result. The seventh, Haywoodsloan, is the one qualitative change: its already-small FPR shifts

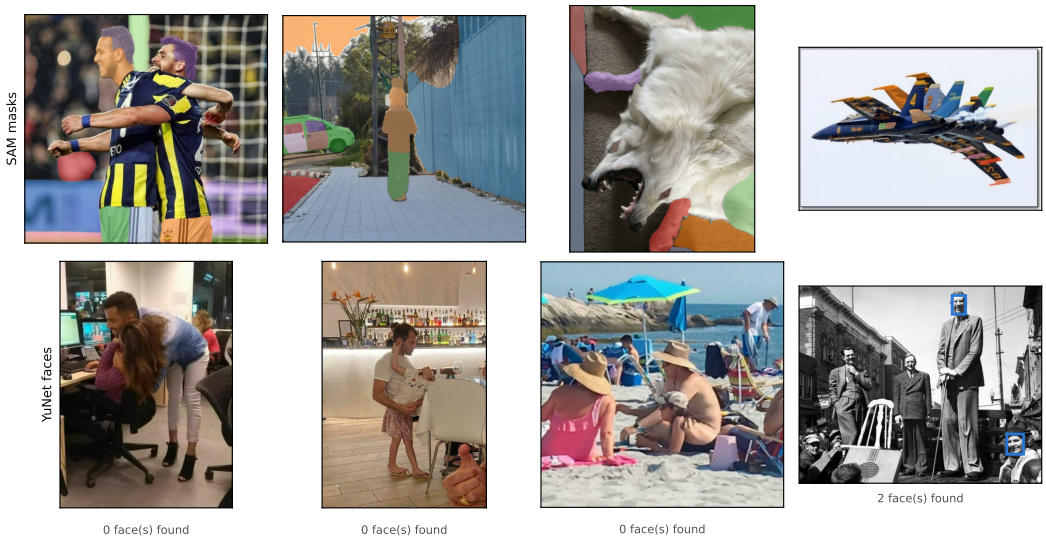


Figure 10: **Other tasks fail on the tail too (qualitative).** Top: SAM automatic mask generation (12 largest masks overlaid) merges or fragments mutually occluding subjects. Bottom: YuNet face detections (blue boxes) on images with multiple clearly visible faces.

Table 6: **Robustness to a newer generator (SDXL).** Adding 250 SDXL-base-1.0 images (prompts sampled fresh from DiffusionDB's metadata, disjoint from the main control set) as a third fake source. sdxl-only AUROC: control-real vs. the 250 SDXL images alone. Extended AUROC: control-real vs. all 750 fakes (original 500 + SDXL). FLORIDA FPR (extended) uses the threshold recalibrated on this extended control set; significance of elevation over control-real FPR by one-sided Fisher exact test: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

Detector	Control AUROC (orig.)	SDXL-only AUROC	Extended AUROC	FLORIDA FPR (orig., %)	FLORIDA FPR (extended, %)
CLIPDet (latent10k+)	0.994	0.992	0.993	80.0	80.0^{***}
Corvi et al. 2023	0.994	0.999	0.996	20.1	20.1^{***}
Ateeqq	0.957	0.975	0.963	31.6	32.7^{***}
Organika SDXL	0.903	0.997	0.935	42.9	40.0^{***}
UnivFD	0.776	0.825	0.792	40.0	41.5^{***}
AEROBLADE	0.859	0.835	0.851	35.7	36.1^{***}
Haywoodsloan	0.983	0.994	0.987	4.9	3.4^{*}

from 1.6% (control-real) to 3.4% (FLORIDA) under the recalibrated threshold, which crosses into nominal significance ($p = 0.036$, vs. $p = 0.09$ under the original threshold) — a reminder that "non-significant" at one calibration can flip with a different one, though the absolute rates involved remain an order of magnitude below every other detector's.

A.8 The hard-negative fix generalizes across detector architectures

This appendix is the full protocol behind Figure 5b in the main text (§4.9), applied to Corvi et al. 2023's frozen CNN backbone (global-average-pooled penultimate feature, i.e. everything before its released classification head).

§4.9 fixes CLIPDet, a linear probe on frozen CLIP ViT-L/14 features; here we check whether the same recipe helps a detector from an unrelated family. We repeat the identical protocol — same FLORIDA train/test split (seed 5, stratified by taxonomy category), same fresh finetuning data, same weight sweep w — but extract features from Corvi et al. 2023's frozen ResNet-50 backbone instead (global average pooling over its final convolutional feature map, dropping only the released

1×1-convolution classification head), of the 795 FLORIDA images, 792 are processed successfully at native resolution (three exceed available GPU memory when extracting raw backbone features at full resolution — one more image than, and not necessarily the same image as, the two excluded from Corvi-2023's classifier evaluation in the main text, Appendix A.10, since that pipeline runs the released classification head rather than the bare backbone), giving a 394/398 train/test split.

The base probe (w=0) reproduces the phenomenon: control AUROC 0.999 yet 25.6% FPR on FLORIDA-test, 21$\times$ its 1.2% control-real FPR — comparable to the released Corvi2023 head's 19.1% FPR on the identical held-out images. Adding the 394 FLORIDA-train images as hard negatives again collapses the failure: 10.6% at w=0.25, 5.0% at w=1, and 3.3% at w=4, with control AUROC essentially unchanged (0.9994–0.9995; Figure 5b, main text). The fix is therefore not a CLIP-feature peculiarity: a few hundred labeled hard negatives close most of the gap for an architecturally unrelated CNN-artifact detector too.

A.9 The hard-negative fix generalizes to unseen taxonomy categories

Both hard-negative checks above hold out *images*, but every taxonomy category is represented in both the training and test split — so they leave open whether the probe generalizes to a fundamentally new *kind* of fake-looking real photograph, or whether it has simply memorized per-category regularities of the exact 795 FLORIDA images we curated. We test this directly with leave-one-category-out cross-validation: for each of the seven taxonomy categories with $n \geq 20$ (§4.6), we train the CLIPDet hard-negative probe using FLORIDA images from only the *other six* categories as real hard negatives (w=1, same fresh fine-tuning data as §4.9), then evaluate FPR on the held-out category alone — which contributes *zero* training examples of its own.

Table 7: **Hard-negative fix, leave-one-taxonomy-category-out.** For each held-out category, the probe's hard negatives are drawn only from FLORIDA images in the *other* six categories ($n \geq 20$); the held-out category contributes zero training examples. n here is counted over the full 795-image taxonomy; Table 3's Atypical Subject/Geometry counts are one lower each because that table is joined against Corvi-2023's per-image scores, which are missing for 2 images. w=0: base probe (no hard negatives). w=1: hard-negative probe. Control AUROC stays ≥ 0.999 for every row (not shown).

Held-out category	n	FPR w=0 (%)	FPR w=1 (%)
Color	41	85.4	2.4
Texture	45	77.8	0.0
Lighting	70	74.3	2.9
Geometry	190	50.5	1.1
Coincidence	178	48.3	1.7
Subject	194	45.4	2.1
Reflection	63	44.4	0.0

Table 7 shows the fix transfers cleanly to every held-out category. The base probe fails badly on each one in isolation (44–85% FPR, worst for color-graded and too-perfect-texture images). Training hard negatives drawn exclusively from the *other* six categories collapses every one of them to 0–2.9% FPR, with control AUROC unchanged at ≥ 0.9996 throughout. Since geometry, coincidence, and subject — together 62% of FLORIDA — generalize just as cleanly as the smaller, more visually distinctive categories (color grading, texture), the fix is not exploiting a shortcut specific to any one look: a probe that has never seen a single labeled reflection photograph nonetheless correctly treats new reflection photographs as real once it has learned that unusual real photographs *in general* are hard negatives. This is the strongest evidence in the paper that §4.9's result is a genuine fix for the fake-looking tail rather than memorization of FLORIDA itself.

A.10 Per-image scores and error matrix

Released as `results/*.csv`; the binary error matrix over the 793 FLORIDA images scored by all seven sanity-passing detectors is `results/florida_errors.csv`. Two images exceeded GPU memory at full resolution for Corvi-2023 and are excluded from its rows. Figure 11 visualizes

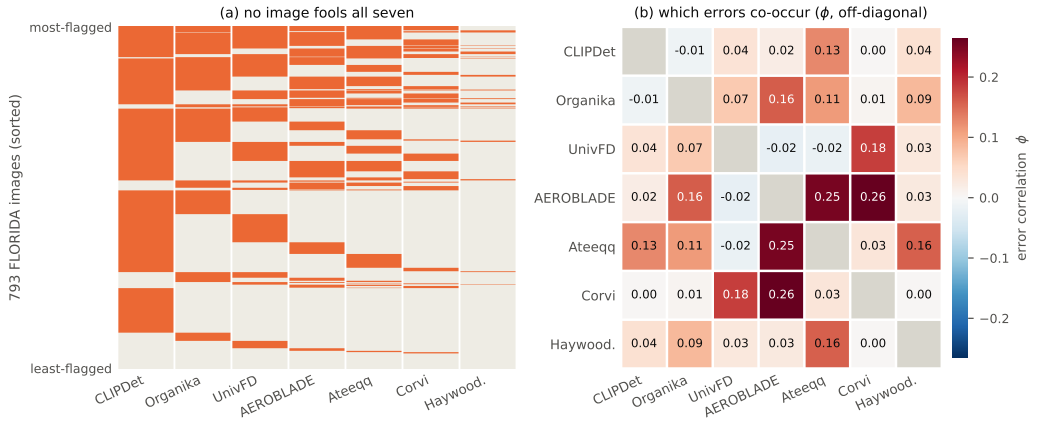


Figure 11: **(a)** Binary error matrix (793 FLORIDA images × 7 sanity-passing detectors; orange = flagged fake at the calibrated threshold), rows sorted by number of detectors flagged, then by pattern. **(b)** Pairwise ϕ-correlation between detectors' errors on the same images (diagonal masked).

this matrix directly: **(a)** the 793 rows sorted so similar flag-patterns cluster, making the "no image fools all seven" claim from §4.4 visible rather than only reported, and **(b)** the pairwise ϕ-correlation matrix between detectors' errors underlying the same section's median/maximum figures.

A.11 Preliminary check: two consumer-facing platforms

As a smaller-scale, earlier proof-of-concept preceding the systematic evaluation of §3.2, we manually uploaded a 510-image subset of FLORIDA (drawn from PART1 and PART2) through the ordinary user interfaces of two publicly accessible systems and recorded each verdict: a general-purpose multimodal chatbot, Bard[1], and a Hugging Face–hosted image classifier specialized for this task, accessed through its public inference widget rather than downloaded weights[2].

Neither system performed well at recognizing these images as real. The chatbot correctly labeled only 38.2% of the (entirely real) images as real, and separately refused to process 92 images containing people, plausibly for privacy reasons. The hosted classifier reached approximately 67% accuracy — surprisingly high given how atypical these photographs look to a human viewer, though still leaving roughly a third misclassified. Figure **??** shows example verdicts from both systems; Figure **??** shows the hosted classifier's confidence when it predicts real versus fake, which skews higher for its real verdicts.

Two caveats limit how much weight this pilot can bear. First, we cannot rule out that some of these images appeared in either system's training data, which would inflate the reported accuracy. Second, neither system exposes a rationale for its verdict, so unlike the taxonomy and attribution analyses in §4.6–§4.8 we cannot examine which visual property drove a given decision. We treat this pilot as motivating evidence for the systematic, weights-available evaluation that makes up the rest of the paper rather than as a controlled result in its own right.

A.12 A calibrated VLM zero-shot baseline

The two systems in Appendix A.11 were tested informally, with no control-set calibration. Here we ask the same question — what happens if you simply ask a vision–language model whether an

[1] `https://bard.google.com/`; since folded into Google's Gemini product line.

[2] `https://huggingface.co/umm-maybe/AI-image-detector`. The same model's downloaded weights are also one of the ten detectors in §3.2 (`umm-maybe`), where it fails the control sanity check on the full 793-image set; it is included here too, on a different subset and access path, because a hosted API and local weights do not always behave identically.

image is real or AI-generated — under the exact protocol of §3.3, treating the model's verdict as an eleventh detector.

Blinding protocol. We pool all 795 FLORIDA images with the full 1,000-image control set (500 COCO reals, 500 SD/DALL-E fakes), shuffle the 1,795 images with a fixed seed, and copy each into a flat directory under a sequential neutral filename (`img_00000`, `img_00001`, ...) that carries no information about source dataset or class. The mapping back to true labels is held out of view during classification. The pool is split into 20 batches of ∼90 images; for each batch, an independent instance of the vision–language model (Claude) is shown only the anonymized filenames and asked, per image, whether it is a real photograph or AI-generated, with a 0–100 confidence that it is AI-generated — no other context, no mention of FLORIDA, the taxonomy, or the study. Model confidence divided by 100 is treated as the detector score $s(x)$, entering the same AUROC/threshold-calibration/Fisher-test pipeline as every other detector (§3.3).

Table 8: **A calibrated VLM zero-shot baseline.** A vision-language model (Claude) blindly classifying anonymized, shuffled images one at a time ("real photograph or AI-generated?", 0–100 confidence), evaluated under the identical protocol as Table 1: control AUROC gates interpretability, threshold calibrated on the control set, applied unchanged to FLORIDA. Included for comparison: Haywoodsloan, the only detector in the main study with a non-significant elevation, and CLIPDet, the detector with the largest. Significance of elevation over control-real FPR, one-sided Fisher exact test: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

Detector	Control AUROC	Control real FPR (%)	FLORIDA FPR (%)	Ratio
VLM zero-shot (Claude)	0.995	0.6	11.2^{***} [9.1, 13.5]	18.7×
Haywoodsloan (for reference)	0.983	3.2	4.9 [3.4, 6.5]	1.5×
CLIPDet (for reference)	0.994	2.4	80.0^{***} [77.1, 82.8]	33.3×

Results. The VLM's control AUROC (0.995) is the highest of all eleven detectors evaluated in this paper, comfortably passing the sanity check. Yet at its control-calibrated threshold, FLORIDA FPR is significantly elevated: 11.2% versus 0.6% on control-real photographs (18.7×, $p < 10^{-4}$, one-sided Fisher exact; Table 8). The failure therefore generalizes beyond trained classifiers, frequency detectors, and reconstruction-error scorers to a fourth, qualitatively different paradigm: in-context zero-shot judgment by a foundation model. At the same time, its *absolute* FLORIDA FPR is far below every other significantly-elevated detector in the main study — a fifth to an eighth of CLIPDet's, Organika's, UnivFD's, AEROBLADE's, Ateeqq's, or Corvi-2023's — second only to Haywoodsloan's non-significant 4.9%. Two detectors built on very different foundations (a SwinV2 classifier fine-tuned for this task, and a large vision–language model never fine-tuned for it at all) are both markedly more robust to this tail than purpose-built synthetic-image detectors, consistent with the hard-negative result of §4.9: broader exposure to atypical real photography, however it enters a model, appears to help.

Limitations specific to this check. We could not corroborate this result with a second model family (e.g. GPT-4V or Gemini): no external API access was available in our environment, so this measures one vision–language model, not VLMs in general. Separately, the same model (Claude) produced the FLORIDA taxonomy labels of §4.6 in an earlier, unrelated session; no weights are updated between sessions and the classification agents here were given only anonymized, shuffled images with no mention of FLORIDA, the taxonomy, or the study, so this is not training-data leakage in the conventional sense — but we flag the shared model family rather than treat it as a fully independent replication. Finally, the 0–100 confidence is a self-reported estimate, not a calibrated probability, the same caveat that applies to any detector score used only for ranking and thresholding.

A.13 A HUMAN BASELINE: DOES THIS LOOK FAKE?

The checks above establish that a chatbot (Appendix A.11) and a calibrated vision–language model (Appendix A.12) both misjudge FLORIDA images at elevated rates, and that VLM-produced taxonomy labels track human judgement of *why* an image looks fake (Appendix A.5). None of these

directly measures the more basic question the whole paper rests on: do ordinary human raters, blind to source, actually doubt these specific images are real, and at what rate compared to ordinary photographs? That base rate is assumed by FLORIDA's curation but was never directly measured against a control group.

Protocol. We reuse the same stratified 100-image FLORIDA sample as Appendix A.5 (for continuity with the taxonomy validation) and add a fresh seeded 50-image sample of ordinary real photographs from the control set (COCO). The 150 images are pooled and shuffled into a single blind labeling page with no group labels shown; for each image the rater picks one of five options (definitely real / probably real / not sure / probably fake / definitely fake, mapped to a confidence-of-fake score of 0/25/50/75/100). Two independent raters labeled all 150 images. We report, per rater and pooled, the fraction of each group judged fake-leaning ("not sure" or worse, confidence ≥ 50), with bootstrap 95% CIs and a one-sided Fisher exact test for elevation — identical statistics to Table 1 and Table 8. The labeling page, sample, and scorer are released under `tagging/realism/` and `scripts/score_realism_validation.py`.

Table 9: **Human baseline: does this look fake?** Independent raters, blind to source, judge a pooled and shuffled set of 100 FLORIDA images and 50 ordinary real photographs (control-real) on a 5-point real/fake scale; "doubted as fake" = "not sure" or worse (confidence ≥ 50). Significance of elevation over control-real rate, one-sided Fisher exact test: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

Rater	Control-real 'doubted' rate (%)	FLORIDA 'doubted' rate (%)	Ratio
Rater 1	2.0	40.0^{***} [30.0, 50.0]	20.0×
Rater 2	18.0	66.0^{***} [57.0, 75.0]	3.7×
Pooled (2 raters × 150 imgs)	10.0	53.0^{***} [46.0, 60.0]	5.3×

Results. Table 9 reports the doubted-as-fake rate for each rater and pooled. Both raters judge FLORIDA images as doubtful far more often than ordinary photographs, though the two differ substantially in overall skepticism. Rater 1 flags only 2.0% of control-real photos as doubtful but 40.0% of FLORIDA images (20.0×, $p = 5.98 \times 10^{-8}$). Rater 2 is a generally more skeptical judge overall, flagging 18.0% of ordinary photos as doubtful, yet still shows a highly significant 3.7× elevation (66.0% vs. 18.0%, $p = 1.75 \times 10^{-8}$). Pooling both raters' 300 judgments, FLORIDA images are judged fake-leaning at 53.0% versus 10.0% for control-real (5.3×, $p = 2.99 \times 10^{-14}$). Despite disagreeing on where to set their own threshold for "looks suspicious," both raters show a large, unambiguous gap between the two image groups. Together with the chatbot pilot (Appendix A.11) and the calibrated VLM baseline (Appendix A.12), this is now three independent kinds of judge — a general-purpose chatbot, a foundation model scored under the paper's own statistical protocol, and ordinary human raters — that agree FLORIDA images are mistaken for fakes at rates far above ordinary photographs, directly supporting the premise the paper's curation and detector evaluation rest on.

Limitations specific to this check. Only two raters and a small sample (100 FLORIDA, 50 control-real) were used, with no rater training beyond the five-option rubric shown on the labeling page. The two raters' very different control-real rates (2.0% vs. 18.0%) indicate substantial individual variation in overall strictness; we did not measure whether they agree on *which* specific images look suspicious, only that both judge the FLORIDA group as more doubtful in aggregate. A per-image inter-rater agreement analysis, as done for the taxonomy labels in Appendix A.5, is left for future work. Finally, "doubted as fake" collapses "not sure" together with "probably/definitely fake," a coarser threshold than a full analysis of the 5-point scale would give.