REALIS: A CURATED DATASET FOR STUDYING THE CHALLENGES OF AI IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image detectors are often evaluated on benchmarks where real and synthetic images differ in content, quality, or generation artifacts, allowing models to rely on dataset-specific cues and fail on unfamiliar generators or processed images. Existing datasets provide limited support for evaluating these challenges jointly across diverse visual content. We introduce REALIS, a dataset of 1.43 million real and synthetic images generated by 42 modern text-to-image models, including the latest proprietary systems such as Nano Banana 2. REALIS combines prompts derived from real images, quality filtering, and stratified sampling to reduce class-specific shortcuts while preserving content diversity. We further introduce REALIS-Expert, a stress-test subset for high-quality synthetic images, where real and generated samples are selected with closely matched semantic and visual characteristics. We also propose a robustness protocol covering 35 transformations at five severity levels to analyze detector behavior under image processing. Based on REALIS, our benchmark evaluates pretrained detectors, fine-tuned models, and zero-shot vision-language models under generator and post-processing shifts. On the hardest processed split, the best pretrained conventional detector achieves 0.550 ROC-AUC, compared with 0.752 for the best REALIS-trained detector. REALIS provides a unified framework for measuring and improving the reliability of AI-image detectors under conditions that better reflect real-world use.

1 INTRODUCTION

AI-generated images are becoming increasingly difficult to distinguish from photographs as text-to-image (T2I) systems improve. Reliably detecting AI-generated images remains crucial for critical applications like media forensics and content moderation. A robust detector must remain reliable when the generator is unfamiliar, the image content differs from its training distribution, and the image has undergone the processing common to online media.

Current benchmarks often make this reliability hard to measure. Their structure may be unsuitable for evaluating detector performance because of inherited biases and limited distortion methods. Those experiments may reflect recognition of dataset-specific cues rather than generation traces (Grommelt et al. (2025); Rajan et al. (2025)), and they may degrade performance on generators not encountered during training (Ojha et al. (2023); Ren et al. (2026)). Finally, these benchmarks test robustness only for simple processing techniques like resizing, cropping, and JPEG compression. More modern postprocessing applications, like neural image compression or watermarking, could suppress forensic traces or create new artifacts that trigger false positives (Wang et al. (2020a); Corvi et al. (2023)). Strong performance on a clean, content-misaligned benchmark therefore need not translate into reliable real-world behavior.

We believe that these problems should be solved jointly in one dataset. Recent datasets and benchmarks have expanded coverage along several of the discussed axes (Zhu et al. (2023); Hong et al. (2025); Pellegrini et al. (2025); Gushchin et al. (2026)), but they lack support for content alignment and generator biases, and they use only simple post-processing strategies within a single controlled framework. To address these issues, we introduce REALIS, a dataset and benchmark designed around these three requirements. It contains 1.43 million real and fake images spanning 42 modern T2I generators. Each fake image is generated from a prompt derived from a real image, aligning data distributions of real and fake parts of the dataset. We combine strict real-image curation, symmetric

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.

filtering of both parts, and prompt-disjoint splits to reduce class-specific biases. We sample validation and test sets to ensure progressively increasing difficulty. We also introduce a post-processing protocol comprising 35 different transformations in eleven groups and five severity levels.

Overall, our contributions are:

- We introduce REALIS, a 1.43M-image dataset covering 42 contemporary T2I generators. We use a specific filtering pipeline including content-derived prompting, symmetric filtering, stratified sampling to reduce dataset-specific biases while preserving semantic diversity. We provide a test part of the dataset in the supplementary material `https://anonymous-hf.com/a/89q3tdboxa2m/`.
- We present a protocol for testing the robustness of the evaluations to 35 modern transformations, including neural network-based ones, compound processing chains, and transformations held out from detector training.
- We benchmark 35 different detectors, including vision-language models. Controlled experiments show that even mild processing causes a sharp performance loss and that degradation is driven largely by increased false positives on real images.

2 RELATED WORK

Datasets and Cross-Generator Generalization. ForenSynths Wang et al. (2020b) established an early cross-generator evaluation protocol using GAN-based synthesis methods. GenImage Zhu et al. (2023) extended this setting to approximately 2.7M images from eight generators and introduced cross-generator and degraded-image evaluation. WildFake Hong et al. (2025) broadened the content and generator distributions, while AI-GenBench Pellegrini et al. (2025) evaluates generalization to generators released after those used for training. Community Forensics Park & Owens (2025) instead maximizes generator diversity, collecting 2.7M images from 4,803 models and demonstrating that detector generalization improves with the number and diversity of training generators. However, raw generator count alone does not provide a controlled evaluation across modern architectural families, generation-quality levels, and proprietary systems. REALIS addresses this gap by curating 42 contemporary open and proprietary generators and assigning them to generator-aware splits that measure generalization to both seen and unseen generator families.

Robustness to Image Processing. Several benchmarks evaluate detectors beyond clean generator outputs. GenImage studies basic blur, compression, and resolution changes, while TrueFake Dell'Anna et al. (2025) and RRDataset Li et al. (2025a) consider social-platform transmission, challenging content, and re-digitization. The NTIRE 2026 challenge Gushchin et al. (2026) provides the closest setting to ours, combining 42 generators with 36 transformations and evaluating detectors on a mixture of transformed and untransformed images. However, these protocols provide limited support for isolating distortion severity and the cumulative effects of multi-stage processing on content-aligned classes. REALIS addresses this limitation using 35 transformations organized into eleven groups, stochastic transformation chains at five severity levels, and paired evaluation of clean and processed versions of the same source images.

Table 3 compares our proposed dataset with the previous AI-generated image detection datasets and benchmarks. Our dataset combines general-content coverage, recent generators, easy access, and extensive robustness evaluation. Existing benchmarks emphasize particular aspects, such as generator breadth, or temporal generalization: our benchmark combines these dimensions together.

3 METHODOLOGY

3.1 DATASET CONSTRUCTION

Figure 1 provides an overview of the construction pipeline. It proceeds in four stages: sourcing and filtering a large pool of real images; synthesizing a generation prompt for every retained real image and rendering it with several T2I models; sampling the Train/Val/Test/Expert splits from the resulting real–generated pool under two complementary criteria; and finally passing the evaluation splits through a chained image degradation pipeline. More details in App.B

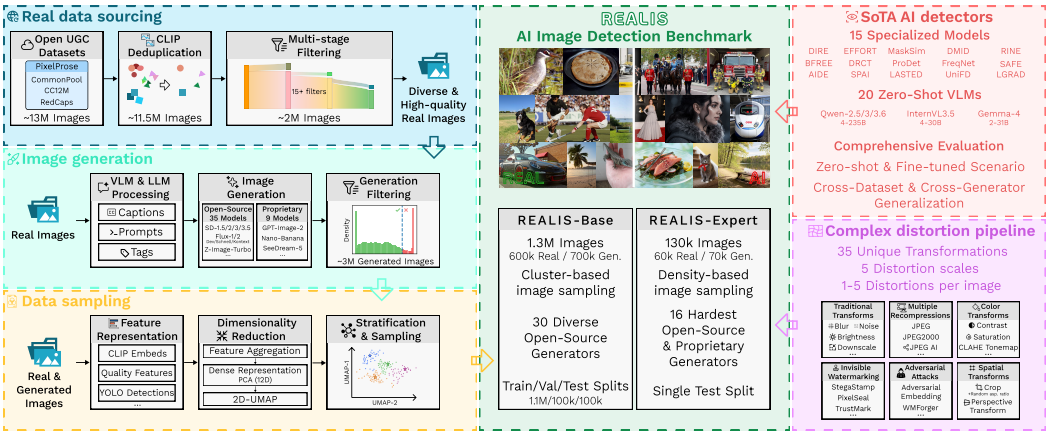

Figure 1: Overview of REALIS construction and evaluation.

3.1.1 Real Image Sourcing

Source corpus. We source real images from PixelProse, Singla et al. (2024) which aggregates CommonPool Gadre et al. (2023), CC12M Changpinyo et al. (2021), and RedCaps Desai et al. (2021) and provides dense captions and initial safety filtering. We apply additional filtering to remove low-resolution, watermarked, and duplicated images that could introduce class-specific shortcuts.

Filtering cascade. Starting from 13M candidate images, we apply a large collection of filters organized into five families. *(i) Deduplication:* exact-duplicate removal on the joint feature signature, followed by near-duplicate removal in both the CLIP image and text embedding spaces. *(ii) Low-level geometry and quality:* images must have reasonable resolution and aspect ratio (specifically, ≥ 256 px at a short side and ≥ 512 px at a long side, with an aspect ratio in $[0.33, 3]$), and must pass minimal image quality and aesthetic thresholds (as measured by TOPIQ Chen et al. (2024b) and LAION-Aesthetics models), set based on the distributions of the generated part of other tested AI-detection benchmarks, so that the real class is never systematically sharper or more attractive than the generated one — a bias that would give detectors a shortcut unrelated to synthesis artifacts. *(iii) Content composition:* we limit the acceptable number of YOLO detections, discarding both trivially empty and pathologically cluttered scenes, and we drop images dominated by rendered text (OCR-derived text length and area share), collages, multi-panel layouts, and uniform synthetic backgrounds. *(iv) Provenance artifacts:* a dedicated watermark detector and an EXIF scan remove stock-photography imagery. *(v) VLM filtering:* a final VLM pass re-labels every surviving image for several categories, catching the missed cases. The full cascade retains 2.9M images, i.e. $\sim 22\%$ of the input. Figure 2 shows the filtering pipeline in detail, along with distributions of several filtering statistics and example images drawn from the rejected regions of each distribution.

3.1.2 Fake Image Generation

Prompt synthesis. Each generated image originates from a prompt derived from a specific real image. This reduces the content gap between web photographs and images generated from independently written prompts, which could otherwise let detectors distinguish subject matter rather than synthesis artifacts.

We first obtain a dense image description with a VLM, then use an instruction-tuned LLM to rewrite it as a generation prompt. The instructions prioritize content type, subjects, object counts and relations, followed by scene, composition and lighting; preserve visible text; and prohibit invented details and quality adjectives. When shortening a prompt, the LLM drops lower-priority details first. We target $\sim$ 35–50 words so that most prompts fit within the 77-token context of older CLIP-based text encoders. The full instructions are in Appendix C. Each prompt is assigned to 3–4 randomly selected generators. We use the same prompt across generators, without model-specific rewriting, negative prompts or hyperparameter tuning, to avoid confounding generator identity with prompt style. The unfiltered pool contains 2.96M prompts and $\sim$ 8.4M generated images.

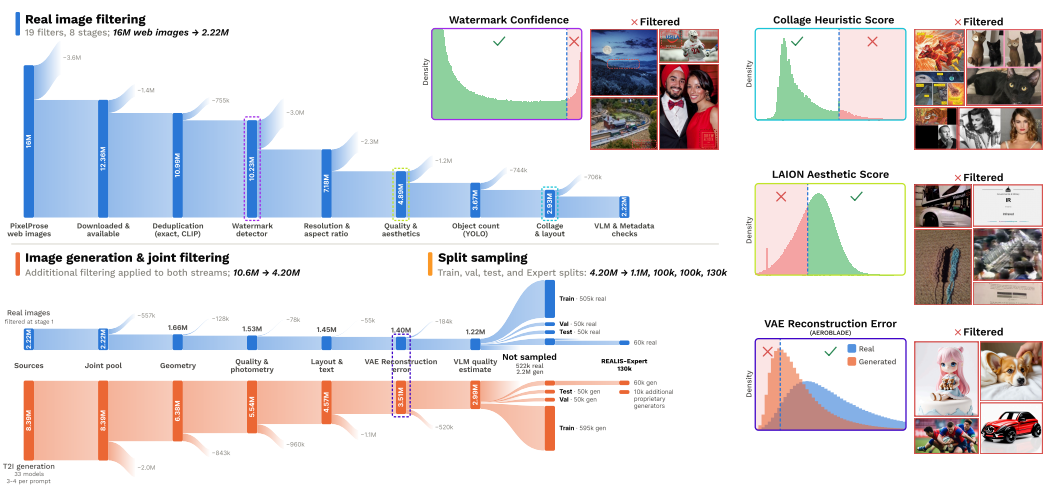

Figure 2: **Left**: Filtering and sampling pipeline for REALIS. Flow widths represent image counts. **Right**: Highlighted distributions illustrate selected filtering criteria, with dashed lines marking thresholds and adjacent examples showing rejected images.

Model selection. We select 33 open-source and 9 proprietary generators along three axes: generative paradigm, model scale and release recency, and inference-time compute regime. The set covers successive latent-diffusion models from the Stable Diffusion lineage; DiT-based PixArt models; cascaded pixel-space diffusion (DeepFloyd IF); flow-matching models from the FLUX family and other lineages; and autoregressive or unified multimodal generators. Including older and newer releases within the Stable Diffusion and Kandinsky lineages, alongside models from other research ecosystems and proprietary providers, broadens coverage across architectures and training pipelines. Appendix B and D provide the full model list and split assignments. We exclude older GAN-based generators: in preliminary experiments, DF-GAN and GALIP followed our image-derived prompts poorly and were detected nearly perfectly. Including them would reintroduce a content gap and inflate aggregate detection performance. We also include distilled and multi-step variants to examine how inference-time acceleration relates to detectability. These comparisons let us measure detectability along the quality–latency frontier directly rather than infer it from proxy metrics.

Assignment of generators to splits. Generators are partitioned across splits by architectural family and by generation quality, so that detection difficulty increases monotonically along train → val → test → Expert. The generator sets are deliberately *not* disjoint: a controlled overlap between splits (e.g. FLUX.1-Kontext-dev, Infinity-8B, Ovis-Image and DeepFloyd IF appear in all three base splits) makes it possible to separate seen-generator from unseen-generator performance for the same detector. Proprietary generators appear only in the Expert subset. Appendix D lists all assignments.

3.1.3 Dataset Sampling and Split Construction

Before sampling, the merged real–generated pool of 10.60M images passes through a second filtering stage applied *symmetrically* to both classes. Any criterion that real and generated images satisfy at systematically different rates is a potential shortcut. We therefore derive thresholds on image characteristics from the *real* distribution and apply the same bounds to both classes. Generated images additionally pass a VLM quality gate that removes generation failures. This leaves 4.2M images (1.3M real, 2.9M generated). We use two sampling strategies: a density-based criterion for Expert, which targets images that are statistically hardest to attribute, and stratified cluster-based sampling for Base, which preserves the diversity of the source distribution.

Feature representation. Both criteria operate in a shared space combining semantic content, object detections, content-type tags, and image quality and structural features. Directly estimating densities in this heterogeneous space would let high-dimensional blocks such as CLIP dominate distances. We therefore reduce and rescale the feature blocks before projecting them to 12 dimensions. The

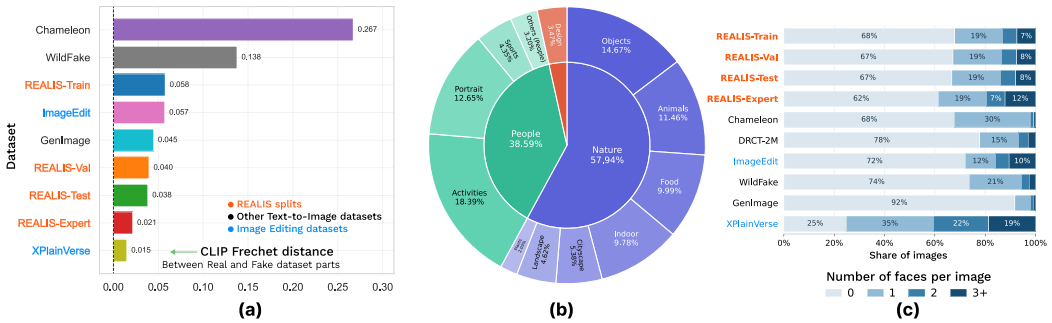

Figure 3: Class alignment and content composition of REALIS splits and other datasets. **(a)** Fréchet distance between real and generated CLIP-embedding distributions; lower values indicate closer class distributions. **(b)** Hierarchical content taxonomy of the pooled REALIS evaluation splits, assigned by a VLM; inner and outer sectors show broad categories and their subdivisions. **(c)** Proportions of images containing different numbers of faces, as detected with the MTCNN model. REALIS combines closely aligned real and generated distributions with varied semantic and scene composition.

projection is fit on a class-balanced subsample; otherwise, the larger generated class would tilt the principal axes towards its distribution. Details are in Appendix B.3.1.

REALIS-Expert. We select hard images using the local log density ratio $\log(p_{\text{real}}(x)/p_{\text{gen}}(x))$. Images with ratios near zero lie in regions where the two classes occur at similar density in the chosen feature space, removing much of the semantically and photometrically easy portion of the distribution. This guarantee is feature-space specific: low-level synthesis fingerprints may still separate the classes in pixel space. The criterion is therefore generator- and detector-agnostic.

We build separate HNSW indices over the 12-D projections of the real and generated pools and estimate both densities with the Loftsgaarden–Quesenberry k-NN estimator. Retaining images with $|\log \hat{r}(x)| \leq 0.5$ keeps $\sim 12\%$ of the pool ($\sim$ 500k images) and lowers mean ROC-AUC across pretrained detectors from 0.826 to 0.809. Restricting to the hardest generators lowers it to 0.723, and requiring a VLM quality score of at least 8 on every axis yields 0.716, compared with 0.871 on the training split. REALIS-Expert is sampled uniformly from this pool, giving 130k images (60k real, 70k generated) across 16 open-source and proprietary generators.

We deliberately do *not* use detector predictions to select Expert images, since this would encode the blind spots of current detectors rather than properties of the data. As a sanity check, $|\log \hat{r}(x)|$ is negatively correlated with per-image detection error on generated images across the tested architectures, with PLCC reaching -0.15.

Base splits. For train, validation, and test, the objective is the opposite: to preserve diversity across the source distribution. We partition the same 12-D representation into 20 K-means clusters, with $k = 20$ selected by standard cluster-quality criteria, and sample with equal quotas per cluster. We use 55k images per cluster for training and 2.5k real plus 2.5k generated images per cluster for validation and test, preventing popular content types from crowding out sparser regions. Sampling respects each split's assigned generators. REALIS-Base contains 1.1M training images and balanced 100k validation and 100k test sets over 30 open-source generators.

Split disjointness. Splits are made disjoint at the level of *prompts*, not images. Each real image induces one prompt and 3–4 near-isosemantic generated counterparts; splitting them across train and test would inflate reported generalization. A real image and all its generated counterparts therefore always belong to exactly one split.

3.1.4 Image Degradation Pipeline

Images encountered in the wild are almost never bit-exact generator output: they have been recompressed, resized, color-graded, watermarked, screenshotted or deliberately perturbed. To evaluate detectors under these conditions, we build a degradation pipeline and apply it to the validation, test

and Expert splits, producing a distorted variant of each; detection performance is reported on both the clean and the distorted version of every split.

The pipeline covers 35 distinct transformations organized into eleven groups: blur, color distortion, algorithmic compression, neural compression, multiple and mixed recompression chains, noise, brightness changes, geometric transforms, spatial distortions, invisible watermarks and some adversarial attacks. The Appendix E provides a detailed list of distortions. Each transformation is parameterized by five strength levels, calibrated so that level 1 is barely perceptible, while level 5 is severe but still leaves the image recognizable.

Rather than applying a single transformation, we simulate realistic processing histories by chaining several. For each image we draw 1–5 groups without replacement, pick one transformation from each and apply them in random order, with the strength level drawn from a discretized Gaussian over the five levels so that moderate degradations dominate while both extremes remain rare. The exact chain and the strength of each element are recorded per image, so robustness can be analyzed per transformation, per group, and per strength level in addition to being reported in aggregate.

For detectors trained on the REALIS-Train split, we employ a restricted augmentation set — essentially the classical group (JPEG, blur, noise, crop, color) — while the neural codecs, recompression chains, adversarial attacks and watermarking methods are held out entirely. A detector trained on our training split therefore still faces genuinely unseen degradations at evaluation time, which is what makes the robustness numbers informative rather than a measure of augmentation coverage.

3.2 Benchmark

We evaluate three groups of detectors: off-the-shelf specialized AI-image detectors, the same detector family fine-tuned on REALIS-Train, and zero-shot vision-language models. We additionally evaluate non-tuned detectors on GenImage, Chameleon, and WildFake to measure how rankings transfer from existing benchmarks to REALIS. Our primary metric is ROC-AUC, which avoids committing to detector-specific score thresholds. For analyses spanning multiple generators, we report generator-macro ROC-AUC by computing AUC separately for each generator against a shared real-image pool and averaging uniformly across generators. We additionally report false-positive and false-negative rates, as well as recall and F1 metrics, for analyses of operating-point behavior.

4 Results

4.1 Dataset Content, Diversity, and Complexity

REALIS combines broad semantic and structural diversity with increasingly aligned real and synthetic distributions. Its evaluation splits cover diverse content types, object and face counts, scene complexity, and aspect ratios (Figure 3c), while all splits occupy a broad region of the pooled CLIP embedding space (Figure 7). Vendi scores and Participation Ratios further show that all REALIS splits exceed the compared T2I benchmarks on both diversity measures, with REALIS-Expert achieving the highest Vendi score. At the same time, the Fréchet distance between real and generated CLIP embeddings decreases from train to Expert (Figure 3b). REALIS-Expert has the smallest gap among the compared text-to-image detection datasets, second only to XPlainVerse, whose generated images are produced by editing real images. Thus, REALIS-Expert preserves broad content coverage while reducing coarse real–synthetic distribution differences, supporting evaluation beyond simple semantic shortcuts. A small Fréchet distance alone, however, does not eliminate residual shortcuts.

4.2 Overall Performance

Table 1 compares off-the-shelf detectors and their REALIS-tuned variants on three external benchmarks and six REALIS evaluation settings. Table 2 compared results for off-the-shelf VLM models on REALIS-Expert. Off-the-shelf results on clean Test are unavailable and are marked with dashes. Results for Expert and Expert-distorted cover the open-source component of REALIS-Expert. Strong performance on external benchmarks does not necessarily transfer: off-the-shelf RINE achieves 97.34% ROC-AUC on GenImage but 44.3% on Expert-distorted. REALIS-tuned DMID reaches 98.5% ROC-AUC on clean Test, while Expert-distorted remains challenging. The REALIS-tuned

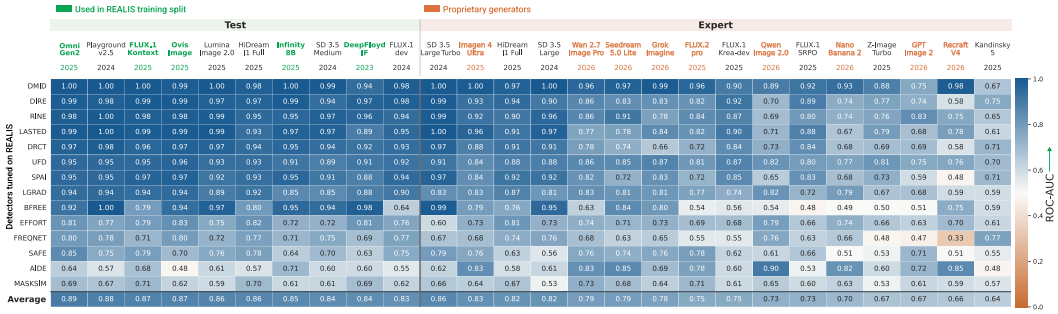

Figure 4: Per-generator ROC-AUC of 14 REALIS-tuned detectors against a shared real-image pool from Test and Expert. Green/orange labels mark training/proprietary generators; release years appear below. Generators are sorted by mean AUC within each split; the bottom row averages detectors.

Table 1: Comparison of off-the-shelf (OFT) and REALIS-tuned (FT) detector performance across datasets. All values are ROC-AUC in percent.

Detector	GenImage		Chameleon		WildFake		REALIS											
							Val		Val-distorted		Test		Test-distorted		Expert		Expert-distorted	
	OTS	FT	OTS	FT	OTS	FT	OTS	FT	OTS	FT	OTS	FT	OTS	FT	OTS	FT	OTS	FT
AIDE	8.12	73.12	47.99	48.80	15.83	74.43	49.9	64.1	50.5	65.5	—	60.1	50.9	64.0	50.1	57.8	51.7	58.6
DIRE	41.21	82.71	54.96	68.63	65.51	40.67	39.5	98.4	40.6	80.1	—	97.3	41.4	77.9	47.9	86.9	48.7	65.7
DRCT	67.04	80.29	68.64	71.01	—	64.55	74.1	96.3	59.3	80.8	—	94.9	57.5	78.4	65.6	82.6	53.5	68.5
EFFORT	56.06	62.68	65.87	58.44	**89.86**	48.64	67.7	80.4	59.0	75.9	—	79.1	56.3	74.3	55.3	66.7	53.1	63.7
FreqNet	64.69	66.93	31.06	57.31	89.04	47.46	49.6	75.9	51.7	58.7	—	75.8	51.5	59.0	41.8	63.3	50.3	52.7
MaskSim	61.94	67.28	61.87	64.95	46.07	50.98	62.1	62.9	52.7	59.6	—	65.5	52.7	61.7	53.4	58.0	51.2	55.3
ProDet	56.43	58.70	47.62	39.50	53.42	0.00	58.2	48.3	54.6	47.6	—	48.3	54.5	47.8	53.9	50.4	52.0	49.9
RINE	**97.34**	88.16	51.40	54.35	87.44	79.50	53.6	97.5	45.7	75.1	—	96.3	45.3	72.2	54.7	85.1	44.3	63.4
SAFE	89.12	78.05	54.63	72.17	86.29	49.78	44.6	68.9	46.3	57.5	—	69.4	46.4	57.7	46.9	62.8	48.9	55.5
SPAI	85.39	84.70	57.94	64.26	73.0	30.11	79.4	94.8	58.0	70.3	—	93.7	57.1	68.8	73.7	82.7	54.1	59.4
UFD	80.09	71.22	43.25	**74.22**	80.83	55.01	54.6	94.0	49.1	**87.6**	—	92.8	47.6	**86.3**	44.0	82.3	42.2	**75.2**

ProDet run failed to converge, and its near-chance REALIS AUCs should not be interpreted as successful fine-tuning.

4.3 Generalization Across Generators

Transfer is strongly generator-dependent. Figure 4 compares REALIS-tuned detectors using a real-image reference pooled from Test and Expert. Many approaches saturate on Test, but Expert results vary widely: SD 3.5 Large-Turbo and Imagen 4 Ultra remain relatively easy, while Kandinsky 5 and Z-Image-Turbo challenge most detectors. Training exposure does not explain this ordering; held-out Playground v2.5 is among the easiest Test generators, with higher AUC than several generators represented in training. REALIS thus reveals both transfer to unseen generators and failures obscured by aggregate scores.

Shared components do not imply shared detectability. Kandinsky 5 and Z-Image-Turbo share the FLUX autoencoder but use cross-attention and single-stream diffusion transformers, respectively Arkhipkin et al. (2025); Z-Image Team et al. (2025). Their lower average AUCs than FLUX.1-dev suggest that autoencoder sharing alone does not predict transfer. Acceleration is likewise not a universal indicator of difficulty: distilled SD 3.5 Large-Turbo is easier to detect than its Large counterpart, while Z-Image-Turbo remains challenging. These comparisons motivate evaluation of individual synthesis pipelines but cannot isolate the effects of architecture, training setup, or sample selection.

Detector rankings and operating thresholds also matter. FREQNET is weak on many generators but achieves the highest plotted AUC on Kandinsky 5. High AUC need not imply high recall at a fixed threshold: on FLUX.1-dev SRPO, DMID reaches about 0.92 AUC but detects only 35% of images at the default 0.5 threshold (Appendix F.2, Figure 10). Operating-point reliability therefore requires interpreting recall alongside false-positive rates.

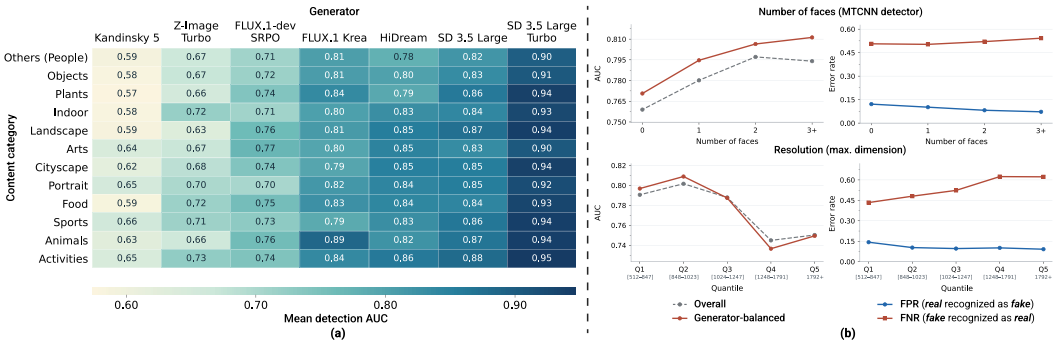

Figure 5: Relationship between image content and characteristics and average detector efficiency (across 16 best-performing models), evaluated on REALIS-Expert split. Proprietary generators are excluded due to lower number of samples per model. (a) Detection quality across various VLM-classified content types for different generators. (b) ROC-AUC, FPR and FNR variation across different image characteristics. Resolution is binned into 5 groups using quantiles (20th, 40th, etc.). For ROC-AUC, both per-generator-balanced and overall generator-agnostic values are reported.

4.4 Performance Across Content and Image Characteristics

Generator differences exceed content variations. Figure 5 analyzes the seven open-source Expert generators, averaging over 16 reference detectors spanning tuned methods, off-the-shelf methods, and zero-shot VLMs. Both real and generated images are restricted to the same content category or statistic bin; generator-balanced AUC then weights eligible generators equally. In Figure 5(a), the range of category-averaged AUC across generators is approximately 0.31, compared with 0.05 for generator-averaged AUC across content categories. Kandinsky 5 remains difficult, and SD 3.5 Large-Turbo is comparatively easy throughout the taxonomy. Thus, generator identity is the stronger correlate of detection difficulty in this analysis, although content still introduces smaller yet measurable variation.

Generator composition explains many apparent statistic effects. The extended sweeps in the Appendix G.1 show that associations with estimated noise variance, high-frequency energy, and aesthetic scores shrink markedly when generator contributions are balanced. Consequently, pooled trends can largely reflect which generators populate each bin. Other associations persist: images with more detected objects have higher generator-balanced AUC, and the face-count analysis in Figure 5(b) also shows a modest increase. At the default threshold of 0.5, however, additional faces accompany fewer false positives on real images, while the false-negative rate stays similar or increases. Better AUC therefore does not imply that generated images containing more faces are more often flagged.

Resolution exposes a different error pattern. In Figure 5(b), the two highest resolution bins have lower AUC and substantially more missed synthetic images, despite relatively low false-positive rates. This association remains visible after generator balancing, yet it can also be affected by intrinsic relationships between image resolution and depicted content, limiting its interpretation to the evaluated distributions. Together, these results show why content and image statistics remain useful diagnostic axes even when generator differences dominate: they reveal distinct false-alarm and missed-detection patterns that a single aggregate score cannot describe.

4.5 Robustness to Post-Processing

Clean performance predicts only part of post-processing robustness. We compare clean images with their post-processed counterparts in the open-source component of REALIS-Expert. Table 1 shows sharp detection losses even for detectors with strong clean-image performance, with many approaching random guessing. Using generator-balanced AUC, clean and processed performance correlate across all detectors (Spearman $\rho = 0.81$; Figure 6b), but rankings can shift: REALIS-tuned UFD trails DMID on clean images yet retains about 0.76 AUC after processing versus DMID's 0.63. Near-chance detectors performance may change slightly without becoming reliable. Clean performance and retained performance therefore offer complementary measures of robustness.

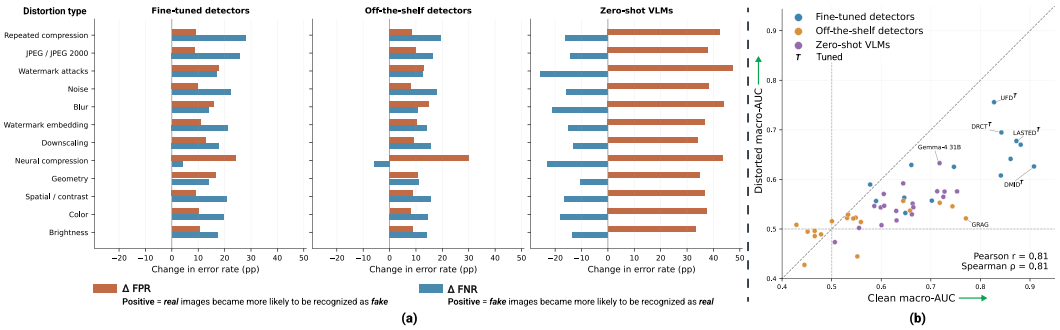

Figure 6: Post-processing robustness on REALIS-Expert. **(a)** Changes in FPR and FNR at threshold 0.5 relative to matched clean images, averaged over 14 representative detectors for subsets containing each distortion group. Positive values indicate more errors; transformations may co-occur. **(b)** Clean versus distorted generator-macro ROC-AUC for all evaluated detectors, using paired predictions.

Table 2: Zero-shot VLM ROC-AUC on REALIS-Expert and its distorted variant. Median AUC across models is 0.625 on Expert and 0.538 on Expert-Distorted.

Model	Expert	Expert-Dist.	Model	Expert	Expert-Dist.	Model	Expert	Expert-Dist.
Qwen3-VL 235B-A22B	**0.734**	0.562	Qwen2.5-VL 32B	0.645	0.539	InternVL3.5 30B-A3B	0.597	0.543
Qwen3-VL 32B	0.712	0.566	Qwen2.5-VL 7B	0.617	0.508	InternVL3.5 14B	0.590	0.537
Qwen3-VL 30B-A3B	0.688	0.560	Qwen2.5-VL 3B	0.584	0.499	InternVL3.5 8B	0.536	0.490
Qwen3-VL 30B-A3B (think)	0.636	0.533	Gemma-4 31B	0.705	**0.622**	InternVL3.5 4B	0.492	0.464
Qwen3-VL 8B	0.706	0.555	Gemma-4 12B	0.633	0.584			
Qwen3-VL 8B (think)	0.613	0.522	Gemma-4 4B	0.570	0.536			
Qwen3-VL 4B	0.639	0.516	Gemma-4 2B	0.594	0.565			

Post-processing induces different error biases across detector groups. Figure 6a shows error changes at the default 0.5 threshold for 14 reference detectors selected by clean-image performance. Conventional detectors generally incur more false positives and false negatives; repeated compression and JPEG degradation especially increase misses. Across every distortion group, the VLM instead produces more false positives and fewer false negatives, indicating a shift toward the generated label despite poorer class separation. Neural compression shows a similar asymmetry among conventional detectors: false-positive increases dominate, while off-the-shelf detectors also miss fewer generated images. Recall or aggregate accuracy alone would obscure these patterns.

Transformation type and strength expose complementary vulnerabilities. Appendix H identifies mixed recompression, glass blur, and shot noise among the largest AUC losses. Increasing distortion severity generally worsens discrimination, especially for blur and JPEG compression, while repeated recompression is damaging even at its lowest setting. A separate validation-sample ablation shows a large drop when the distortion pipeline is enabled and further losses as its strength increases. Expert-split comparisons use clean versions of the same images; because transformations can co-occur, they characterize complete processing chains rather than individual stages' causal effects.

5 Conclusion

We introduce REALIS, a dataset of 1.43 million images spanning 42 text-to-image generators, designed to study detection under generator and post-processing shifts. Its construction reduces semantic and quality differences between real and synthetic images while preserving content diversity. REALIS-Expert further targets closely matched class distributions without using detector predictions for selection, and paired clean and processed images support controlled robustness evaluation. Our experiments show that strong performance on existing benchmarks or clean images does not ensure reliable detection after processing, with substantial degradation affecting both specialized detectors and zero-shot VLMs. Fine-tuning on REALIS improves performance, yet the best tuned detector reaches only 0.752 ROC-AUC on the processed open-source component of REALIS-Expert. Processing also changes error patterns, including increased false positives on real images, highlighting the need to assess false alarms alongside missed detections. Together, these findings motivate joint evaluation of generator generalization and processing robustness, with REALIS providing a common resource for developing and diagnosing detectors under both challenges.

AI USE STATEMENT

In this work, generative AI is part of the methodology. The synthetic part of REALIS was produced by the text-to-image models listed in Appendix C; dense VLM captions of real images were rewritten into generation prompts with Qwen3-30B-A3B (Appendix B); VLMs, including InternVL3.5-38B, were used to filter, score and tag images during dataset construction (Section 3.1); and VLMs are evaluated as zero-shot detectors (Section 3.2). We additionally used generative AI tools to polish the writing and to assist literature search. All AI-assisted outputs, including references, were checked manually, and the authors take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

A subset of REALIS is available for peer review at `https://anonymous-hf.com/a/89q3tdboxa2m/`. Due to the practical challenges of fully anonymizing the dataset for double-blind review, we provide only a partial release during the review period. The full dataset will be released upon acceptance. The dataset construction procedure is described in Section 3.1 and further detailed in the appendices. Code for dataset generation and detector evaluation is available at `https://anonymous-hf.com/a/jshtpiysv3ua/`.

REFERENCES

Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, and Furong Huang. WAVES: Benchmarking the robustness of image watermarks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1456–1492. PMLR, 2024. URL `https://proceedings.mlr.press/v235/an24a.html`.

Vladimir Arkhipkin, Andrei Filatov, Viacheslav Vasilev, Anastasia Maltseva, Said Azizov, Igor Pavlov, Julia Agafonova, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky 3.0 technical report. *arXiv preprint arXiv:2312.03511*, 2023. doi: 10.48550/arXiv.2312.03511.

Vladimir Arkhipkin, Vladimir Korviakov, Nikolai Gerasimenko, Denis Parkhomenko, Viacheslav Vasilev, Alexey Letunovskiy, Nikolai Vaulin, Maria Kovaleva, Ivan Kirillov, Lev Novitskiy, et al. Kandinsky 5.0: A family of foundation models for image and video generation. *arXiv preprint arXiv:2511.14993*, 2025. doi: 10.48550/arXiv.2511.14993.

João Ascenso, Elena Alshina, and Touradj Ebrahimi. The jpeg ai standard: Providing efficient human and machine visual data consumption. *IEEE MultiMedia*, 30(1):100–111, 2023. doi: 10.1109/MMUL.2023.3245919.

Shuai Bai et al. Qwen2.5-VL Technical Report, 2025a.

Shuai Bai et al. Qwen3-VL Technical Report, 2025b.

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.

Jordan J. Bird and Ahmad Lotfi. Cifake: Image classification and explainable identification of ai-generated synthetic images. *IEEE Access*, 12:15642–15650, 2024. doi: 10.1109/ACCESS.2024.3356122.

Black Forest Labs. FLUX.1. `https://github.com/black-forest-labs/flux`, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL `https://arxiv.org/abs/2506.15742`.

Tu Bui, Shruti Agarwal, and John Collomosse. TrustMark: Robust watermarking and watermark removal for arbitrary resolution images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18629–18639, 2025. doi: 10.1109/ICCV51701.2025.01731.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025. doi: 10.48550/arXiv.2505.22705.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training to Recognize Long-Tail Visual Concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3557–3567. IEEE, 2021. doi: 10.1109/CVPR46437.2021.00356.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7621–7639, 2024a. URL https://proceedings.mlr.press/v235/chen24ay.html.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A Top-Down Approach from Semantics to Distortions for Image Quality Assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024b. doi: 10.1109/TIP.2024.3378466.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-σ: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024c. doi: 10.1007/978-3-031-73411-3_5.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, 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 *International conference on learning representations*, volume 2024, pp. 57611–57640, 2024d. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fe989bb038b5dcc44181255dd6913e43-Abstract-Conference.html.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. *arXiv preprint arXiv:2501.17811*, 2025. doi: 10.48550/arXiv.2501.17811.

Zhongxi Chen, Ke Sun, Ziyin Zhou, Xianming Lin, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. Diffusionface: Towards a comprehensive dataset for diffusion-based face forgery analysis. *arXiv preprint arXiv:2403.18471*, 2024e. doi: 10.48550/arXiv.2403.18471.

Harry Cheng, Yangyang Guo, Tianyi Wang, Liqiang Nie, and Mohan Kankanhalli. Diffusion facial forgery detection. In *Proceedings of the ACM International Conference on Multimedia*, pp. 5939–5948, 2024a. doi: 10.1145/3664647.3680797.

Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can We Leave Deepfake Data Behind in Training Deepfake Detector? In *Advances in Neural Information Processing Systems*, volume 37, pp. 21979–21998, 2024b. doi: 10.52202/079017-0691.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules. In *Proceedings of CVPR*, pp. 7939–7948, 2020. doi: 10.1109/CVPR42600.2020.00796.

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 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10095167.

DeepFloyd Lab. DeepFloyd IF. https://github.com/deep-floyd/IF, 2023. Text-to-image diffusion model.

Stefano Dell'Anna, Andrea Montibeller, and Giulia Boato. Truefake: A real world case dataset of last generation fake images also shared on social networks. In *2025 International Joint Conference on Neural Networks*, 2025. doi: 10.1109/IJCNN64981.2025.11228911.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/e00da03b685a0dd18fb6a08af0923de0-Abstract-round1.html.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 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.

Han Fang, Zhaoyang Jia, Zehua Ma, Ee-Chien Chang, and Weiming Zhang. Pimog: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 2267–2275, 2022. doi: 10.1145/3503161.3548049.

Han Fang, Yupeng Qiu, Kejiang Chen, Jiyi Zhang, Weiming Zhang, and Ee-Chien Chang. Flow-based robust watermarking with invertible noise layer for black-box distortions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 5054–5061, 2023. doi: 10.1609/aaai.v37i4.25633.

Tinglei Feng, Yingjie Zhai, Jufeng Yang, Jie Liang, Deng-Ping Fan, Jing Zhang, Ling Shao, and Dacheng Tao. Ic9600: A benchmark dataset for automatic image complexity assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8577–8593, 2023. doi: 10.1109/TPAMI.2022.3232328.

Pierre Fernandez, Tomáš Souček, Nikola Jovanović, Hady Elsahar, Sylvestre-Alvise Rebuffi, Valeriu Lacatusu, Tuan Tran, and Alexandre Mourachko. Geometric image synchronization with deep watermarking. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9522–9526, 2026. doi: 10.1109/ICASSP55912.2026.11464560.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=g97OHbQyk1.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. DataComp: In Search of the Next Generation of Multimodal Datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023. doi: 10.52202/075280-1179.

Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, 2017. doi: 10.1101/214262. URL https://www.biorxiv.org/content/early/2017/11/12/214262.

Gemma Team. Gemma 4 Technical Report, 2026.

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 *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2021. doi: 10.1109/ICME51207.2021.9428429.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or JPEG? revealing common biases in generated image detection datasets. In *Computer Vision – ECCV 2024 Workshops*, pp. 80–95. Springer, 2025. doi: 10.1007/978-3-031-92089-9_6.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general AI-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. doi: 10.1109/CVPR52734.2025.01741.

Hengchang Guo, Qilong Zhang, Junwei Luo, Feng Guo, Wenbin Zhang, Xiaodong Su, and Minglei Li. Practical deep dispersed watermarking with synchronization and fusion. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 7922–7932, 2023. doi: 10.1145/3581783.3612015.

Aleksandr Gushchin, Khaled Abud, Ekaterina Shumitskaya, Artem Filippov, Georgii Bychkov, Sergey Lavrushkin, Mikhail Erofeev, Anastasia Antsiferova, Changsheng Chen, Shunquan Tan, et al. NTIRE 2026 challenge on robust AI-generated image detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1895–1913, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/NTIRE/html/Gushchin_NTIRE_2026_Challenge_on_Robust_AI-Generated_Image_Detection_in_the_CVPRW_2026_paper.html.

Piotr Guzik, Andrzej Matiolanski, and Andrzej Dziech. Real data performance evaluation of caiss watermarking scheme. *Multimedia Tools and Applications*, 74(12):4437–4451, 2015. doi: 10.1007/s11042-013-1544-3.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling Bitwise Autoregressive Modeling for High-Resolution Image Synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15733–15744, 2025. doi: 10.1109/CVPR52734.2025.01467.

David Hasler and Sabine E. Suesstrunk. Measuring colorfulness in natural images. In Bernice E. Rogowitz and Thrasyvoulos N. Pappas (eds.), *Human Vision and Electronic Imaging VIII*, volume 5007, pp. 87 – 95. International Society for Optics and Photonics, SPIE, 2003. doi: 10.1117/12.477378. URL https://doi.org/10.1117/12.477378.

Yan Hong, Jianming Feng, Haoxing Chen, Jun Lan, Huijia Zhu, Weiqiang Wang, and Jianfu Zhang. Wildfake: A large-scale and hierarchical dataset for ai-generated images detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. doi: 10.1609/aaai.v39i4.32363.

Runyi Hu, Jie Zhang, Shiqian Zhao, Nils Lukas, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang. Mask image watermarking. In *Advances in Neural Information Processing Systems*, volume 38, pp. 162403–162436, 2025. doi: 10.52202/085713-4895.

Jiangtao Huang, Ting Luo, Li Li, Gaobo Yang, Haiyong Xu, and Chin-Chen Chang. Arwgan: Attention-guided robust image watermarking model based on gan. *IEEE Transactions on Instrumentation and Measurement*, 72:1–17, 2023. doi: 10.1109/TIM.2023.3285981.

Zhaoyang Jia, Han Fang, and Weiming Zhang. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 41–49, 2021. doi: 10.1145/3474085.3475324.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-Resolution AI-Generated Image Detection by Spectral Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18706–18717, 2025. doi: 10.1109/CVPR52734.2025.01743.

Christos Koutlis and Symeon Papadopoulos. Leveraging Representations from Intermediate Encoder-Blocks for Synthetic Image Detection. In *Computer Vision – ECCV 2024*, pp. 394–411. Springer, 2025. doi: 10.1007/978-3-031-73220-1_23.

Kwai-Kolors. Kolors. https://github.com/Kwai-Kolors/Kolors, 2024. Code and model repository.

Sangwu Lee, Titus Ebbecke, Erwann Millon, Will Beddow, Le Zhuo, Iker García-Ferrero, Liam Esparraguera, Mihai Petrescu, Gian Saß, Gabriel Menezes, and Victor Perez. Flux.1 krea [dev]. https://github.com/krea-ai/flux-krea, 2025.

Chunxiao Li, Xiaoxiao Wang, Meiling Li, Boming Miao, Peng Sun, Yunjian Zhang, Xiangyang Ji, and Yao Zhu. Bridging the gap between ideal and real-world evaluation: Benchmarking ai-generated image detection in challenging scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a. doi: 10.1109/ICCV51701.2025.01895.

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*, 2024a. doi: 10.48550/arXiv.2402.17245.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving Synthetic Image Detection Towards Generalization: An Image Transformation Perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2405–2414, 2025b. doi: 10.1145/3690624.3709392.

Yanhao Li, Quentin Bammey, Marina Gardella, Tina Nikoukhah, Jean-Michel Morel, Miguel Colom, and Rafael Grompone Von Gioi. MaskSim: Detection of Synthetic Images by Masked Spectrum Similarity Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3855–3865, 2024b. doi: 10.1109/CVPRW63382.2024.00390.

Yewon Lim, Changyeon Lee, Aerin Kim, and Oren Etzioni. DistilDIRE: A Small, Fast, Cheap and Lightweight Diffusion Synthesized Deepfake Detection, 2024.

Shanchuan Lin, Anran Wang, and Xiao Yang. SDXL-Lightning: Progressive Adversarial Diffusion Distillation. *arXiv preprint arXiv:2402.13929*, 2024. doi: 10.48550/arXiv.2402.13929.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware Adaptive Transformer for Generalizable Synthetic Image Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10770–10780, 2024. doi: 10.1109/CVPR52733.2024.01024.

Zhengzhe Liu, Xiaojuan Qi, and Philip H. S. Torr. Global Texture Enhancement for Fake Face Detection in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8060–8069, 2020. doi: 10.1109/CVPR42600.2020.00808.

Victor Livernoche, Akshatha Arodi, Andreea Musulan, Zachary Yang, Adam Salvail, Gaétan Marceau Caron, Jean-François Godbout, and Reihaneh Rabbany. Openfake: An open dataset and platform toward large-scale deepfake detection, 2025. URL https://arxiv.org/abs/2509.09495v1.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1105.

Yihong Luo, Xiaolong Chen, Xinghua Qu, Tianyang Hu, and Jing Tang. You Only Sample Once: Taming One-Step Text-to-Image Synthesis by Self-Cooperative Diffusion GANs. In *International Conference on Learning Representations*, volume 2025, pp. 89196–89217, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ddb0a18cc21b98fffb4b69c43f9b56f5-Abstract-Conference.html.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023. doi: 10.1109/CVPR52729.2023.02345.

Anisha Pal, Julia Kruk, Mansi Phute, Manognya Bhattaram, Diyi Yang, Duen Horng Chau, and Judy Hoffman. Semi-truths: A large-scale dataset of ai-augmented images for evaluating robustness of ai-generated image detectors. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-3748.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8245–8257, 2025. doi: 10.1109/CVPR52734.2025.00772.

Lorenzo Pellegrini, Davide Cozzolino, Serafino Pandolfini, Davide Maltoni, Matteo Ferrara, Luisa Verdoliva, Marco Prati, and Marco Ramilli. AI-GenBench: A New Ongoing Benchmark for AI-Generated Image Detection. In *2025 International Joint Conference on Neural Networks*, pp. 1–9, 2025. doi: 10.1109/IJCNN64981.2025.11228377.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/081b08068e4733ae3e7ad019fe8d172f-Abstract-Conference.html.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Xinyue Li, Dongyang Liu, Xiangyang Zhu, et al. Lumina-Image 2.0: A Unified and Efficient Image Generative Framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20031–20042, 2025. doi: 10.1109/ICCV51701.2025.01863.

Md Awsafur Rahman, Bishmoy Paul, Najibul Haque Sarker, Zaber Ibn Abdul Hakim, and Shaikh Anowarul Fattah. Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection. In *2023 IEEE International Conference on Image Processing*, pp. 2200–2204. IEEE, 2023. doi: 10.1109/ICIP49359.2023.10222083.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=doBkiqESYq.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 286–295, 2023. doi: 10.18653/v1/2023.emnlp-demo.25.

Simiao Ren, Yuchen Zhou, Xingyu Shen, Kidus Zewde, Tommy Duong, George Huang, Hatsanai (Neo) Tiangratanakul, Tsang (Dennis) Ng, En Wei, and Jiayu Xue. How well are open sourced AI-generated image detection models out-of-the-box: A comprehensive benchmark study, 2026.

Lodestone Rock. Chroma1-HD. https://huggingface.co/lodestones/Chroma1-HD, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042.

Rajarshi Roy, Nasrin Imanpour, Ashhar Aziz, Shashwat Bajpai, Gurpreet Singh, Shwetangshu Biswas, Kapil Wanaskar, Parth Patwa, Subhankar Ghosh, Shreyas Dixit, Nilesh Ranjan Pal, Vipula Rawte, Ritvik Garimella, Gaytri Jena, Vasu Sharma, Vinija Jain, Aman Chadha, Aishwarya Naresh Reganti, and Amitava Das. A comprehensive dataset for human vs. ai generated image detection, 2026. URL https://arxiv.org/abs/2601.00553v1.

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3418–3432, 2023. doi: 10.1145/3576915.3616588.

Xiangwei Shen, Zhimin Li, Zhantao Yang, Shiyi Zhang, Yingfang Zhang, Donghao Li, Chunyu Wang, Qinglin Lu, and Yansong Tang. Directly aligning the full diffusion trajectory with fine-grained human preference. *arXiv preprint arXiv:2509.06942*, 2025. doi: 10.48550/arXiv.2509.06942.

Wissam Siblini, Jordan Fréry, Liyun He-Guelton, Frédéric Oblé, and Yi-Qing Wang. *Master Your Metrics with Calibration*, pp. 457–469. Springer International Publishing, 2020. ISBN 9783030445843. doi: 10.1007/978-3-030-44584-3_36. URL http://dx.doi.org/10.1007/978-3-030-44584-3_36.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions. *arXiv preprint arXiv:2406.10328*, 2024. doi: 10.48550/arXiv.2406.10328.

Tomáš Souček, Pierre Fernandez, Hady Elsahar, Sylvestre-Alvise Rebuffi, Valeriu Lacatusu, Tuan Tran, Tom Sander, and Alexandre Mourachko. Pixel Seal: Adversarial-Only Training for Invisible Image and Video Watermarking. *arXiv preprint arXiv:2512.16874*, 2025. doi: 10.48550/arXiv.2512.16874.

Tomáš Souček, Sylvestre-Alvise Rebuffi, Pierre Fernandez, Nikola Jovanović, Hady Elsahar, Valeriu Lacatusu, Tuan Tran, and Alexandre Mourachko. Transferable black-box one-shot forging of watermarks via image preference models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1675.

Stability AI. SDXL-Turbo. `https://huggingface.co/stabilityai/sdxl-turbo`, 2023. Hugging Face model card.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12105–12114, 2023. doi: 10.1109/CVPR52729.2023.01165.

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*, volume 38, pp. 5052–5060, 2024a. doi: 10.1609/aaai.v38i5.28310.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the Up-Sampling Operations in CNN-based Generative Network for Generalizable Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024b. doi: 10.1109/CVPR52733.2024.02657.

Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2114–2123, 2020. doi: 10.1109/CVPR42600.2020.00219.

Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16515–16525, 2022. doi: 10.1109/CVPR52688.2022.01602.

Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14214–14223, 2023. doi: 10.1109/CVPR52729.2023.01366.

Guo-Hua Wang, Liangfu Cao, Tianyu Cui, Minghao Fu, Xiaohao Chen, Pengxin Zhan, Jianshan Zhao, Lan Li, Bowen Fu, Jiaqi Liu, et al. Ovis-image technical report. *arXiv preprint arXiv:2511.22982*, 2025a. doi: 10.48550/arXiv.2511.22982.

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*, pp. 8695–8704, 2020a. doi: 10.1109/CVPR42600.2020.00872.

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*, pp. 8695–8704, 2020b. doi: 10.1109/CVPR42600.2020.00872.

Weiyun Wang et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, 2025b.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023. doi: 10.1109/ICCV51070.2023.02051.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image Technical Report. *arXiv preprint arXiv:2508.02324*, 2025a. doi: 10.48550/arXiv.2508.02324.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Exploration to Advanced Multimodal Generation. *arXiv preprint arXiv:2506.18871*, 2025b. doi: 10.48550/arXiv.2506.18871. URL https://arxiv.org/abs/2506.18871v1.

Haiwei Wu, Jiantao Zhou, and Shile Zhang. Generalizable Synthetic Image Detection via Language-Guided Contrastive Learning. *IEEE Transactions on Artificial Intelligence*, 7(6):3485–3496, 2026. doi: 10.1109/TAI.2025.3641104.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025. doi: 10.1109/CVPR52734.2025.01241.

A Yakushev, A Akimenkov, K Abud, D Obydenkov, I Serzhenko, K Aistov, E Kovalev, S Fomin, A Antsiferova, K Lukianov, et al. Wibe: Watermarks for generated images-benchmarking & evaluation. In *Proceedings of the 40th IEEE/ACM International Conference on Automated Software Engineering*, 2025. doi: 10.1109/ASE63991.2025.00368.

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*, 2025a. URL https://openreview.net/forum?id=ODRHZrkOQM.

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*, 2025b. URL https://openreview.net/forum?id=ODRHZrkOQM.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal Subspace Decomposition for Generalizable AI-Generated Image Detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70268–70288. PMLR, 2025c. URL https://proceedings.mlr.press/v267/yan25b.html.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Shijie Huang, Zhaohui Hou, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Feng Yu, Chi Zhang, Shifeng Zhang, and Shilin Zhou. Z-Image: An Efficient Image Generation Foundation Model with Single-Stream Diffusion Transformer. *arXiv preprint arXiv:2511.22699*, 2025. doi: 10.48550/arXiv.2511.22699. URL https://arxiv.org/abs/2511.22699v1.

Nan Zhong, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection, 2023. URL https://arxiv.org/abs/2311.12397v1.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, pp. 682–697, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-01266-3. doi: 10.1007/978-3-030-01267-0_40. URL https://doi.org/10.1007/978-3-030-01267-0_40.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3398.

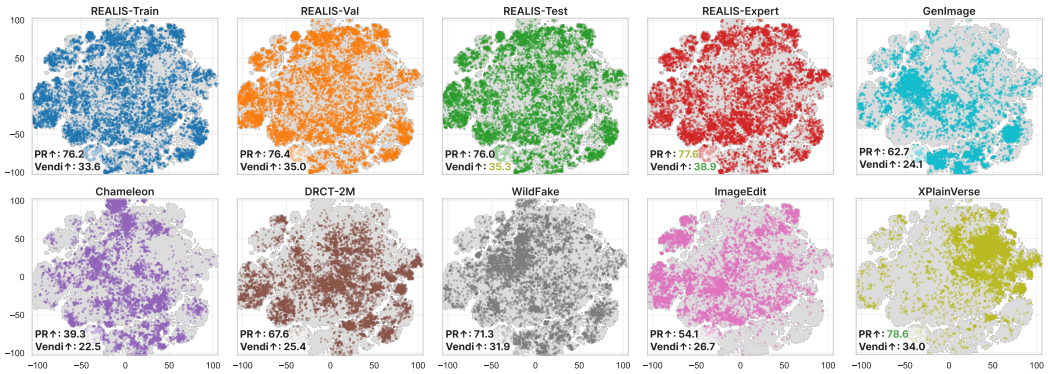

Figure 7: 2D t-SNE representation of CLIP-space coverage for different deepfake detection datasets, alongside their Vendi Scores (↑) Friedman & Dieng (2023) and Participation Ratios (↑) Gao et al. (2017) to quantify the diversity. Scores are calculated over original CLIP embeddings on full datasets; t-SNE projection is used only for visualization. Grey dots represent the concatenation of all datasets, with random subsamples of similar size (10k) used for each entry.

A Contents

Appendix B details relevant dataset comparison B.1, prompt synthesis B.2, generator selection B.3, data filtering, and split construction B.3.1.

Appendix C gives the prompt-generation instructions.

Appendix D lists generators by split.

Appendix E specifies the post-processing pipeline.

Appendix F extends the detector results with per-generator F.1 and fixed-threshold evaluations F.2.

Appendix G examines image characteristics G.1, generation quality, and generator profiles G.2.

Appendix H analyzes sensitivity to post-processing transformations and their strength, including a separate validation-sample ablation.

B Dataset details

B.1 Comparison with Existing Datasets

Table 3 summarizes representative AI-generated image detection datasets by size, generator coverage, content scope, access, and robustness evaluation. These attributes describe different design choices rather than a single ranking: for example, some benchmarks prioritize the number of generation sources, while others focus on temporal generalization or transformations encountered during image sharing. Our dataset combines 2.2 million real and 8.8 million synthetic images from 42 generation sources, including 12 closed models, with general-content coverage and extensive post-processing evaluation. Reported counts and qualitative categories follow the respective dataset descriptions and may reflect different counting conventions.

B.2 Prompt synthesis

Every generated image in REALIS originates from a prompt that describes a specific real image from the sourced pool. This design eliminates the content gap that otherwise separates the two classes in detection datasets, where real images are web photographs while generated images follow the very different distribution of user-written prompts. Under such a gap, a detector can reach high accuracy by recognizing *subject matter* rather than *synthesis artifacts*, and its reported accuracy says little about what it will do in deployment.

Table 3: Comparison with representative AI-generated image detection datasets and benchmarks. #Gen. denotes the number of fake-image generators or generation sources when reported. #Closed denotes proprietary/API/community-service generators whose weights are not publicly released. "—" indicates that the paper or dataset page does not report a clean value. Robust. stands for robustness to processing techniques.

Dataset / Benchmark	Year	Real	Fake	#Gen.	#Closed	Scope	Access	Latest Generator	Robust.
ForenSynths Wang et al. (2020b)	2020	362K	362K	11	0	GAN-era	Hard	StyleGAN2 (2019)	Basic
Diff.Forensics Wang et al. (2023)	2023	140K	570K	8	0	Scene	Hard	SD v2 (2022)	No
DMDetection Corvi et al. (2023)	2023	200K	200K	3	1	Scene	Hard	DALL-E 2 (2022)	Basic
DE-FAKE Sha et al. (2023)	2023	—	—	4	1	T2I	Hard	DALL-E 2 (2022)	No
GenImage Zhu et al. (2023)	2023	1330K	1350K	8	1	Object	Hard	Midjourney (2022)	Basic
ArtiFact Rahman et al. (2023)	2023	965K	1530K	25	0	Multi-cat.	Easy	SD v2 (2022)	Basic
Synthbuster Bammey (2024)	2023	1K	9K	9	4	Prompt	Easy	DALL-E 3 (2023)	No
CiFAKE Bird & Lotfi (2024)	2023	60K	60K	1	0	CIFAR-10	Easy	LD (2022)	No
Fake2M Lu et al. (2023)	2023	1090K	2080K	3	0	Perception	Hard	IF (2023)	Basic
SIDBench Zhong et al. (2023)	2024	406K	412K	17	2	General	Easy	DALL-E 2 (2022)	Basic
DRCT-2M Chen et al. (2024a)	2024	—	2000K	16	0	Diffusion	Easy	SDXL (2023)	Basic
Semi-Truths Pal et al. (2024)	2024	26K	1340K	8	—	General	Easy	SDXL (2023)	No
DiffusionFace Chen et al. (2024e)	2024	30K	600K	11	0	Face	Hard	SD v2.1 (2022)	No
DiFF Cheng et al. (2024a)	2024	23K	500K	13	1	Face	Hard	Midjourney 5 (2023)	No
WildFake Hong et al. (2025)	2025	1010K	2560K	23	—	Wild	Hard	Community (—)	Basic
Chameleon Yan et al. (2025a)	2025	15K	11K	—	—	Wild	Gated	DALL-E 3 (2023)	Hard set
AI-GenBench Pellegrini et al. (2025)	2025	180K	180K	36	3	General	Easy	FLUX.1 (2024)	Temporal
TrueFake Dell'Anna et al. (2025)	2025	160K	440K	8	2	Social	Easy	FLUX.1 (2024)	Sharing
C.Forensics Park & Owens (2025)	2025	2700K	2700K	4803	11	Gen.-rich	Easy	Imagen 3 (2024)	Gen. stress
RRDataset Li et al. (2025a)	2025	10K	10K	—	—	Scenario	Easy	FLUX.1 (2024)	Re-digit.
OpenFake Livernoche et al. (2025)	2025	3000K	963K	18	7	Political	Easy	Imagen 4 (2025)	Arena
MS COCOAI Roy et al. (2026)	2026	16K	80K	5	2	COCO	Easy	Midj. v6 (2023)	Basic
Ours	**2026**	**2200K**	**8800K**	**42**	**12**	**General**	**Easy**	**GPT Image 2 (2026)**	**Extensive**

Prompts are produced in two steps. First, a dense description of the real image is obtained with a VLM. Second, an instruction-tuned LLM rewrites that description into a generation prompt. The LLM is instructed to follow standard prompt-construction practice— lead with content type, exact object counts, subjects and their main relation, then add scene, composition and lighting; reproduce any on-image text verbatim together with its location; invent no details and no quality adjectives — and to drop material in a fixed tail-priority order when the length budget is exceeded, never dropping content type, entities, counts, or relations. We target a compact prompt of $\sim 35 - 50$ words: long enough to pin down the scene, short enough that most of the prompts fit inside the 77-token context of CLIP-based text encoders and are therefore rendered without truncation even by the older models in our pool. The full instruction template is given in Appendix C.

Each prompt is then assigned 3 to 4 randomly chosen generators, so that in the unfiltered pool every real image has 3–4 content-matched synthetic counterparts. We deliberately issue the same prompt to every generator, with no model-specific rewriting, negative prompts or per-model hyperparameter tuning: such tuning would confound generator identity with prompt style and make cross-generator comparison unreliable. The full pool comprises 2.96M prompts and $\sim$8.4M generated images.

B.3 T2I MODELS SELECTION

To ensure that our evaluation is representative of the contemporary text-to-image landscape rather than tied to a single architectural paradigm, training regime or organizational lineage, we assemble 33 open-source and 9 proprietary generators. Our choice is guided by three criteria: generative paradigm, model scale and release recency, and inference-time compute regime.

Coverage of generative paradigms. The latent diffusion paradigm is covered exhaustively through the Stable Diffusion lineage (SD-1.4, SD-1.5, SD-2.1 Rombach et al. (2022), SDXL Podell et al. (2024), SD-3 Medium, SD-3.5 Medium and SD-3.5 Large Esser et al. (2024)), through PixArt-α Chen et al. (2024d) and PixArt-Σ Chen et al. (2024c), which isolate the effect of DiT-style transformer backbones, and through Playground v2.5 Li et al. (2024a), Kolors Kwai-Kolors (2024) and Lumina-Image-2.0 Qin et al. (2025), which probe alternative training recipes on comparable architectures. Cascaded pixel-space diffusion is represented by DeepFloyd IF DeepFloyd Lab (2023), a counterpoint to latent approaches. Flow-matching models are covered by the FLUX family (FLUX.1-dev, FLUX.1-schnell Black Forest Labs (2024), FLUX.1 Krea [dev] Lee et al. (2025), FLUX.1-Kontext-dev Black Forest Labs et al. (2025)), HiDream-I1-Full Cai et al. (2025), Qwen-Image Wu et al. (2025a),

Chroma1-HD Rock (2025) and SRPO Shen et al. (2025), which together span the current open state of the art. Autoregressive and visual-autoregressive synthesis is represented by Janus-Pro-7B Chen et al. (2025), Infinity-2B and Infinity-8B Han et al. (2025), and by the unified multimodal models OmniGen Xiao et al. (2025), OmniGen2 Wu et al. (2025b) and Ovis-Image Wang et al. (2025a).

We deliberately omit older GAN-based generators entirely. In preliminary experiments, models such as DF-GAN Tao et al. (2022) and GALIP Tao et al. (2023) proved unable to follow prompts of the specificity our pipeline produces, and the images they return are separated near-perfectly by most detectors in our pool. Including them would inflate reported accuracy without measuring anything about modern synthesis, and would reintroduce the content gap, which the prompt design is meant to remove.

Temporal coverage and lineage diversity. Including SD-1.4 alongside SD-3.5 Large and the FLUX family lets us trace the trajectory of a single open lineage across several generations under a controlled architectural family, while the Kandinsky line (Kandinsky-2 Razzhigaev et al. (2023), Kandinsky-3.1 Arkhipkin et al. (2023), Kandinsky-5 Arkhipkin et al. (2025)), Ovis-Image Wang et al. (2025a), Z-Image-Turbo Z-Image Team et al. (2025) and Qwen-Image Wu et al. (2025a) ensure that the benchmark is not biased towards any single research ecosystem or pre-training corpus. This organizational and geographic diversity matters because text-to-image models inherit cultural, linguistic, and visual priors from their data pipelines, and a benchmark drawn from a single lineage would systematically under-report that variability. The proprietary tier — GPT-Image-2, Nano-Banana-2, Nano-Banana-Pro, Seedream-5-Lite, Imagen-4-Ultra, FLUX.2, Grok-Imagine and Ideogram-v3 — captures the closed-source frontier that, in practice, most misuse in the wild is produced with.

Inference-time compute regimes. A growing fraction of practical deployments relies on few-step or single-step samplers rather than the dozens of denoising steps assumed in the original diffusion formulations. To make our evaluation robust to this shift, we include distilled variants alongside their multi-step parents: SDXL-Turbo Stability AI (2023) and SDXL-Lightning Lin et al. (2024) against SDXL Podell et al. (2024), SD-3.5 Large-Turbo against SD-3.5 Large Esser et al. (2024), FLUX.1-schnell against FLUX.1-dev Black Forest Labs (2024), as well as YOSO Luo et al. (2025) and Z-Image-Turbo Z-Image Team et al. (2025) as recent few-step approaches. This lets us measure detectability along the quality–latency frontier directly, rather than inferring it through proxy metrics.

B.3.1 Dataset Sampling and Split Construction

Before sampling, the merged real–generated pool of 10.60M images passes through a second filtering stage applied *symmetrically* to both classes. Any criterion that real and generated images satisfy at systematically different rates is a potential shortcut, so the thresholds on brightness, image complexity, spectral characteristics, and no-reference quality are taken from percentiles of the *real* distribution, and the geometry bounds apply identically to both. Generated images additionally pass a quality gate: a VLM (InternVL3.5-38B) scores each of them for prompt relevancy, semantic coherence, technical quality and scene complexity on a 1–10 scale, and images below 6.5 on any axis are discarded as generation failures. This stage leaves 4.2M images (1.3M real, 2.9M generated), the pool from which all splits are drawn.

We use two different sampling strategies: (i) a density-based criterion for the Expert subset, which is meant to hold the images that are statistically hardest to attribute; and (ii) a stratified cluster-based criterion for the base splits, which preserve the diversity of the source distribution.

Feature representation. Both criteria operate in a common low-dimensional space built from a heterogeneous per-image description. We collect 512-D CLIP image embeddings; an 80-D vector of per-class YOLO detection counts; a 21-D multi-hot vector of VLM content-type tags (e.g., *outdoor photo*, *2d art*, *night scene*, *portrait*, etc.); and various scalar features spanning quality and aesthetics metrics, image structural complexity estimate, geometry, colorfulness, brightness, spatial information, face and OCR statistics, and frequency-, noise-, and forensics characteristics.

Estimating class densities directly in this space is impractical: the dimensionality is high, the blocks live on wildly different scales, and naive concatenation would let the CLIP embeddings dominate distance. We therefore reduce the features in two steps. Each block is first mapped to a comparable dimensionality using applicable transformations (25-D PCA for CLIP, 12-D KL-NMF for YOLO detections, TF-IDF + SVD-decomposition for VLM tags) — after which each block g is rescaled

by $w_g = 1/\sqrt{d_g}$ so that all four contribute equally to the concatenated distance. A single linear PCA then projects the concatenation to 12 dimensions, retaining 97.3% of the variance. The final projection is deliberately linear and orthogonal: unlike a nonlinear embedding, it preserves Euclidean distances up to the discarded variance, which is what makes the subsequent nearest-neighbor density estimates meaningful. It is fit on a class-balanced subsample, since fitting on the natural $1{:}2.5$ real-to-generated mix would tilt the principal axes towards the generated distribution.

REALIS-Expert. To select the hardest images, we use a criterion that directly approximates the local log density ratio $\log\big(p_{\text{real}}(x)/p_{\text{gen}}(x)\big)$ in the neighborhood of each image. The motivation comes from statistical decision theory: for a two-class problem the Bayes-optimal decision boundary is exactly the level set on which the density ratio equals its prior-corrected threshold, and classification error is maximal there. Images whose local log ratio is close to 0 therefore lie, in this feature space, precisely in the region where *any* classifier operating on these features must be maximally uncertain. We note that the guarantee is relative to the chosen feature space and not to pixel space: a detector reading low-level synthesis fingerprints can in principle still separate these images. What the criterion does provide is a generator-agnostic and detector-agnostic way of removing the semantically and photometrically easy part of the distribution.

Specifically, we build two HNSW indices over the 12-D projections of the real and generated pools and estimate both log densities at every point with the Loftsgaarden–Quesenberry k-nearest-neighbour estimator . Retaining only images with $|\log \hat{r}(x)| \leq 0.5$ keeps $\sim 12\%$ of the pool ($\sim$500k images) and by itself lowers the average ROC-AUC of multiple tested pretrained detectors from 0.826 to 0.809. Restricting to the strongest generators drops it further to 0.723, and a final VLM gate requiring a score of at least 8 on every quality axis yields 0.716, against 0.871 on the training split. REALIS-Expert is sampled uniformly from what remains, giving 130k images (60k real, 70k generated) across the 16 hardest open-source and proprietary generators.

We further verified that the density signal is predictive of detector behavior, and not merely of our own feature construction. Across the tested detectors, $|\log \hat{r}(x)|$ correlates *negatively* with per-image detection error on generated images (PLCC reaching -0.15 for UFD and -0.13 for SAFE and RINE): the closer an image lies to the decision boundary in our feature space, the more likely existing detectors are to misclassify it. The effect is small in absolute terms — as it must be, since our features contain no low-level forensic information — but it is consistent in sign across architectures and highly significant at this sample size. We deliberately do *not* use pretrained detector predictions as the selection criterion for the hard subset, even though this would yield a subset on which those detectors score far worse. Such a subset would encode the specific blind spots of the current model generation and of the prior detection datasets they were trained on, and a future detector that happens to fail differently would be scored against difficulty defined by its predecessors. The density criterion depends only on the data.

Base splits. For the train, validation, and test splits, the objective is the opposite: to preserve the diversity of the source distribution across multiple feature scales. We reuse the same 12-D representation and partition it with the K-means clusterer. We chose $k = 20$ based on cluster-quality evaluations with silhouette, Davies–Bouldin and Calinski–Harabasz scores. Sampling then proceeds with equal quotas per cluster — 55k images per cluster for training, and 2.5k real plus 2.5k generated per cluster for validation and test — which prevents the most popular content types of the web-image distribution (e.g., outdoor photographs of people) from crowding out sparser regions such as aerial or night scenes. Within each split, only images from that split's assigned generators are eligible. The resulting REALIS-Base contains 1.3M images: 1.1M train (505k real, 595k generated), 100k validation and 100k test, both balanced $50/50$, over 30 open-source generators.

Split disjointness. Splits are made disjoint at the level of *prompts*, not images. Because each real image induces one prompt and 3–4 near-isosemantic generated counterparts, splitting at the image level would place visually near-identical content on both sides of a train/test boundary and inflate reported generalization. Therefore, a real image and all of its generated counterparts always live in exactly one split. The difficulty gradient the design targets is borne out empirically: averaged across pretrained detectors, mean ROC-AUC falls monotonically across the splits—from 0.871 on Train and 0.854 on Val splits, down to 0.836/0.716 on the Test and Expert sets, respectively.

C Prompt Generation Instructions

The prompts for image generation were produced using a Qwen3-30B-A3B LLM model prompted with the following instructions:

> You rewrite long image captions into ONE short prompt with strict importance ordering.
>
> INPUT CAPTION between « ».
>
> OUTPUT
>
> - From one to three sentences of 50 words in total. Output ONLY the sentences.
>
> ORDER & RULES
>
> 1. Begin with content type (if given/obvious) + exact counts + subjects + main action + main relation.
> 2. Then add, as relevant:
>   (a) For photos: scene/place/time →composition → lighting.
>   (b) For screenshots/documents/charts/abstract/CGI: domain-specific layout/detail (e.g., centered layout, dark mode, A4 portrait, legend, north-up, scale bar) → optional style (vector/schematic/CGI/photoreal if explicitly stated).
> 3. If on-image text is mentioned in the caption, state it EXACTLY as it is, with its location.
> 4. No invented details or quality words. If >50 words, drop from the tail in this order: domain-neutral palette → camera → extra style → secondary layout. Never drop content type, entities, counts, relations, or required text.
> 5. Output sentences MUST be in English.
>
> Now produce the sentence.
>
> «CAPTION_START»
> *caption*
> «CAPTION_END»

D Generator Inventory Across Splits

Table 4 lists the generators of every split, grouped by release year. Older models are concentrated in the training split, validation and test add newer ones, and REALIS-Expert is formed by models released from late 2024 onwards together with all proprietary generators; of its open-source models, only HiDream-I1-Full also appears in a base split (Test). Five generators (DeepFloyd IF, FLUX.1-Kontext-dev, Infinity-8B, OmniGen2 and Ovis-Image) are present in all three base splits, so that performance on seen and unseen generators can be compared within the same evaluation split, while Playground v2.5 and Lumina-Image-2.0 are shared by validation and test only.

E Complete List of Post-Processing Transformations

Tables 5 and 6 list all transformations of the degradation pipeline used to produce the distorted variants of the validation, test and Expert splits (Section 3.1). The 34 core transformations are organized into ten groups. For every image, 1-5 groups are drawn without replacement (each count with equal probability), one transformation is selected uniformly within each group, and the selected transformations are applied in sequence. Every transformation has five strength levels, and the level of each selected transformation is drawn independently from a discretized Gaussian over the level index; in the released splits, levels 1–5 occur with frequencies of about 13, 17, 22, 25 and 23%, respectively. After the chain, an invisible watermark produced by one of 12 methods is embedded with probability 0.35, the image is further downscaled by a random factor in $[0.3, 0.8]$ with probability 0.75, and the result is saved as near-lossless JPEG at quality 95, with the full distortion chain logged for every image.

Table 4: Generators in each REALIS split, grouped by public release year of the evaluated model. Train, validation and test form REALIS-Base. *Also in the training split, which allows seen and unseen generators to be compared on the same detector; †proprietary, accessed through public APIs and used only in REALIS-Expert. The last two rows give the number of generators and of images (real / generated) per split.

Year	**Train**	**Validation**	**Test**	**Expert**
2022	SD-1.4 SD-1.5 SD-2.1	—	—	—
2023	DeepFloyd IF Kandinsky-2 PixArt-α SDXL SDXL-Turbo	DeepFloyd IF*	DeepFloyd IF*	—
2024	Infinity-2B Kandinsky-3.1 Kolors OmniGen PixArt-Σ SDXL-Lightning YOSO-PixArt-512	FLUX.1-schnell Playground v2.5 SD-3 Medium Switti-1024	FLUX.1-dev Playground v2.5 SD-3.5 Medium	SD-3.5 Large SD-3.5 Large-Turbo
2025	FLUX.1-Kontext-dev Infinity-8B OmniGen2 Ovis-Image	FLUX.1-Kontext-dev* Infinity-8B* OmniGen2* Ovis-Image* Lumina-Image-2.0	FLUX.1-Kontext-dev* Infinity-8B* OmniGen2* Ovis-Image* HiDream-I1-Full Lumina-Image-2.0	FLUX.1 Krea [dev] FLUX.1-dev SRPO HiDream-I1-Full Kandinsky-5 Z-Image-Turbo FLUX.2 [pro]† Imagen-4-Ultra†
2026	—	—	—	GPT-Image-2† Grok-Imagine† Nano-Banana-2† Qwen-Image-2.0† Recraft-V4† Seedream-5.0-Lite† Wan-2.7-Image-Pro†
Generators	19	10	10	7 + 9†
Images	504.6k / 595.4k	50k / 50k	50k / 50k	60k / 69.1k

To indicate the severity of each setting, the last two columns of Table 5 report PSNR and SSIM for every transformation applied in isolation at its weakest and strongest level. The two metrics are complementary: global photometric changes such as RGB shift, tone curves and saturation changes lower PSNR to 15–20 dB at level 5 while SSIM stays at or above 0.90, whereas noise, glass blur and strong darkening degrade structure the most (SSIM down to 0.13 for shot noise). Neural codecs, adversarial embedding attacks and WMForger remain at or above 29 dB and 0.85 SSIM even at level 5. Because transformations are chained and followed by watermarking and downscaling, the cumulative degradation of a distorted image is generally larger than these single-transformation values.

F Additional Detector Results

- **Conventional detectors:** AIDE Yan et al. (2025b), BFREE Guillaro et al. (2025), DIRE Wang et al. (2023), DistilDIRE Lim et al. (2024), DRCT Chen et al. (2024a), EFFORT Yan et al. (2025c), FatFormer Liu et al. (2024), FreqNet Tan et al. (2024a), GRAG Gragnaniello et al. (2021), GramNet Liu et al. (2020), LASTED Wu et al. (2026), LGrad Tan et al. (2023), MaskSim Li et al. (2024b), NPR Tan et al. (2024b), ProDet Cheng

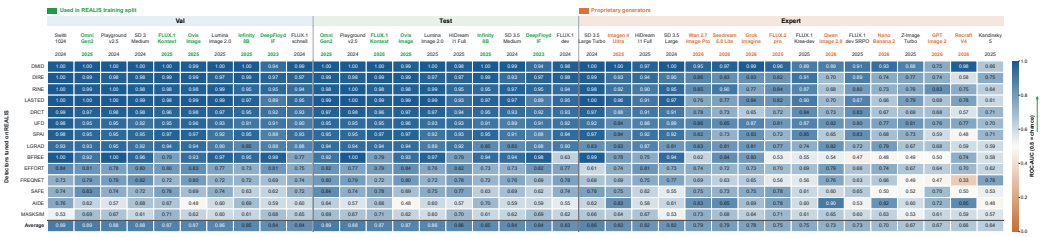

Figure 8: ROC-AUC of the REALIS-tuned detectors on every generator of the validation, test and Expert splits. For each detector, the real images of all three splits are pooled into a single reference set (160k images), and each cell compares the images of one generator with this set. Green labels mark generators used in the REALIS training split, orange labels mark proprietary generators, and release years are given below the names. Columns are ordered within each split by decreasing average AUC, and the last row averages the detectors.

et al. (2024b), RINE Koutlis & Papadopoulos (2025), SAFE Li et al. (2025b), SPAI Karageorgiou et al. (2025), UFD (UniFD) Ojha et al. (2023), DMID and SelfCon.

- **Zero-shot VLM detectors:** Qwen3-VL Bai et al. (2025b) (235B-A22B, 32B, 30B-A3B, 30B-A3B think, 8B, 8B think, 4B); Qwen2.5-VL Bai et al. (2025a) (32B, 7B, 3B); Gemma-4 Gemma Team (2026) (31B, 12B, 4B, 2B); InternVL3.5 Wang et al. (2025b) (30B-A3B, 14B, 8B, 4B).

This section extends the per-generator analysis of Sections 4.3 and 4.4.

F.1 Per-Generator Results

Pooled real reference. Figure 8 extends Figure 4 to the validation split. Since the real images of Val, Test and Expert form one reference set for each detector, differences between columns reflect only the generated images. Averaged over detectors and generators, the AUC is 0.867 on Val and 0.859 on Test, but 0.754 for the open-source and 0.744 for the proprietary Expert generators. Generators used in training are not detected consistently better than held-out ones: their mean AUC is 0.864 versus 0.870 on Val and 0.865 versus 0.853 on Test. On Expert, generator averages range from 0.86 for SD-3.5 Large-Turbo to 0.64 for Kandinsky-5. The proprietary models are spread over this range, from Imagen-4-Ultra (0.83) to Recraft-V4 (0.66).

Clean and distorted images. Figure 9 repeats the per-generator comparison on the distorted Expert images of Section 4.5. Distortion lowers the AUC in every detector–generator cell with both runs, from 0.79 to 0.60 on average. The losses are largest for the generators that are easiest on clean images (SD-3.5 Large-Turbo 0.93 $\rightarrow$ 0.69, SD-3.5 Large 0.87 $\rightarrow$ 0.65) and smallest for Kandinsky-5 (0.62 $\rightarrow$ 0.53), so the gap between the easiest and the hardest generator halves, from 0.31 to 0.16. Averaged over generators, the fine-tuned reference detectors fall from 0.87 to 0.65, the zero-shot VLMs from 0.73 to 0.59, and the off-the-shelf detectors from 0.72 to 0.54, close to chance. Across all detectors with both runs, including weaker ones, 82% of the cells decrease and 24% of the distorted cells fall below 0.5.

F.2 Detection at a Fixed Operating Threshold

AUC measures how well scores rank generated above real images, independently of a decision threshold. Figures 10 and 11 instead evaluate the detectors at the default threshold of 0.5. Averaged over detectors and generators, recall falls from 72.1% on Val and 69.3% on Test to 44.7% on Expert (48.3% for open-source and 41.9% for proprietary generators). The false-positive rate on real images averages 16% in every split, but differs strongly between detectors: DMID, AIDE and LASTED flag less than 1% of real images (LASTED less than 0.05%), whereas MASKSIM flags 41–43% and EFFORT 52–59%. A high AUC can therefore coincide with low recall. LASTED, for example, reaches a pooled Expert AUC of 0.81 but a recall of only 12%, because it scores almost all images of both classes below the threshold.

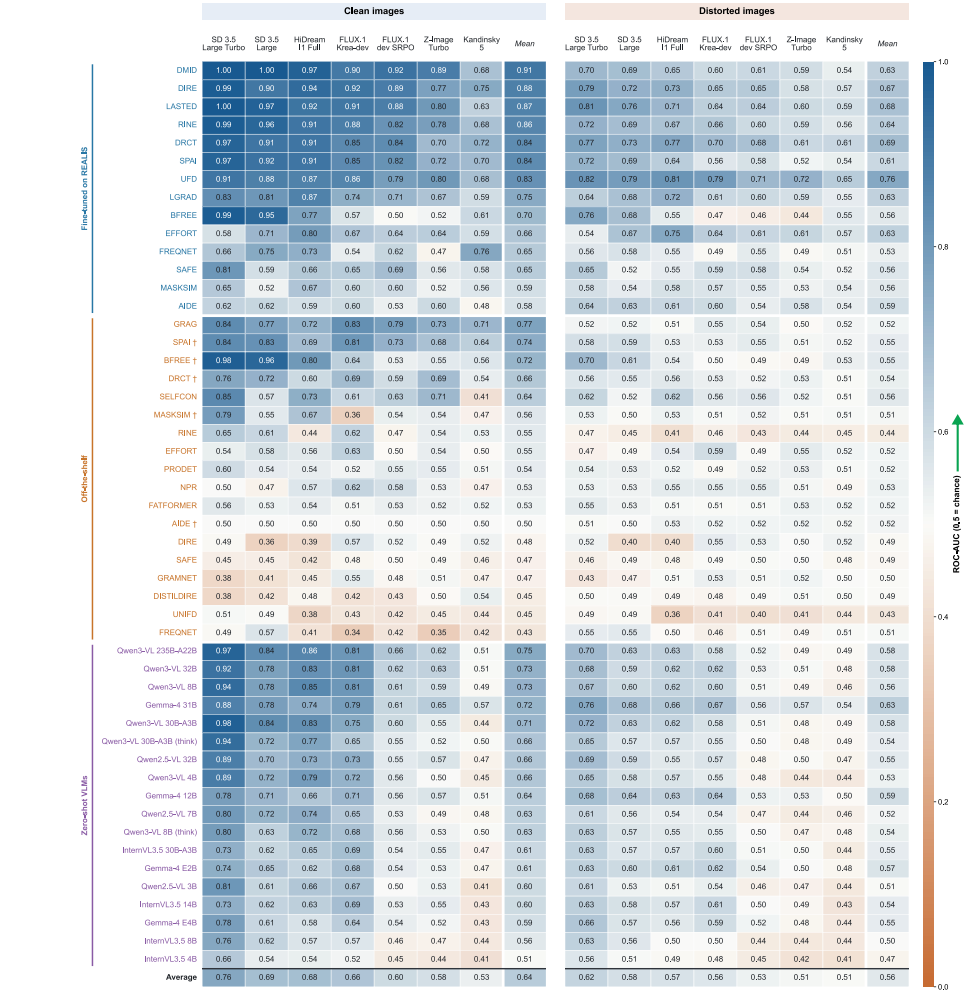

Figure 9: Per-generator ROC-AUC for all tested detectors on the open-source part of REALIS-Expert, for clean images (left) and their distorted versions (right). Each cell compares the images of one generator with all real Expert images of the same version, and for each detector both halves use the same images; *Mean* averages the seven generators. Rows are grouped into detectors fine-tuned on REALIS, off-the-shelf detectors and zero-shot VLMs, and are ordered by clean AUC within each group. †: the detector's output contains only binary decisions, so its AUC equals balanced accuracy.

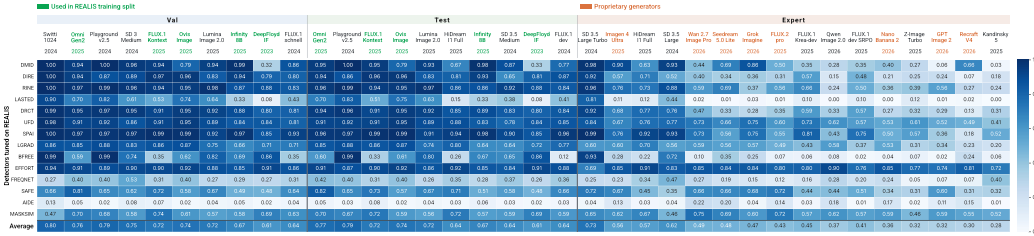

Figure 10: Per-generator recall of REALIS-tuned detectors on the validation, Test, and Expert splits. Recall is the fraction of generated images assigned a score greater than 0.5. Green labels identify generators represented in training, orange labels identify proprietary generators, and the bottom row averages the displayed detectors equally. Low recall on several Expert generators reveals limitations of the default operating threshold that are not fully captured by ROC-AUC; recall should be interpreted together with false-positive rates.

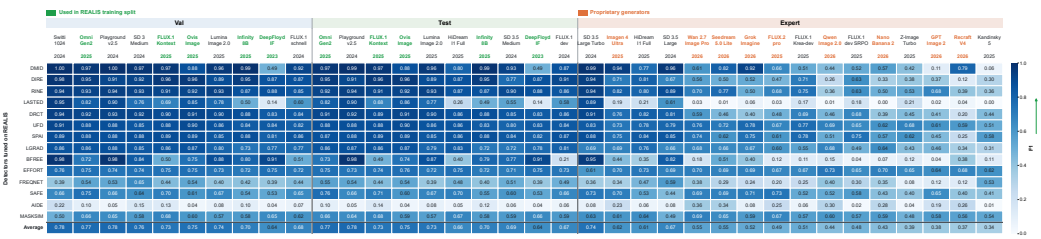

Figure 11: F1 of the REALIS-tuned detectors at the default threshold of 0.5, computed with real and generated images weighted equally (calibrated F1): $F_1 = 2R/(1 + R + \mathrm{FPR})$, where R is the recall on the generator's images and FPR is the share of the split's real images with a score above 0.5. Rows, columns and label colors as in Figure 8.

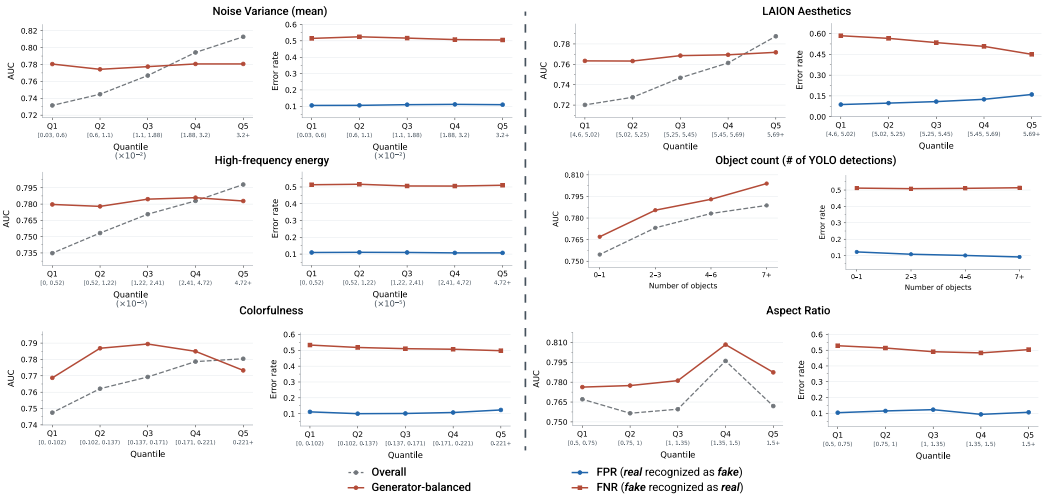

Figure 12: Detection performance across quantile bins of six per-image statistics on the open-source part of REALIS-Expert (seven generators, 120k images), averaged over the 16 reference detectors of Figure 5. For each statistic, the left panel shows the overall AUC, which pools all generated images of a bin, and the generator-balanced AUC, which averages per-generator AUCs against the real images of the same bin; the right panel shows the FPR on real images and the generator-balanced FNR at a score threshold of 0.5. Noise variance is the mean over 8×8 patches of the variance of the Laplacian response; high-frequency energy is the share of spectral power above half the Nyquist frequency; colorfulness follows Hasler and SüsstrunkHasler & Suesstrunk (2003); object count is the number of YOLO11x detections with confidence above 0.5.

The calibrated F1 score Siblini et al. (2020) in Figure 11 combines recall and false positives in one number per generator. Since a generator contributes between 1k and 12k images, whereas each split has 50k–60k real images, F1 is computed with both classes weighted equally; otherwise it would mainly reflect the number of images per generator. Mean F1 is 0.73 on Val, 0.71 on Test and 0.51 on Expert (0.54 for open-source and 0.48 for proprietary generators). The ranking of detectors by F1 differs markedly from the ranking by AUC (Spearman correlation 0.35 on Expert). DMID has the highest Expert AUC (0.92) but only the fifth-highest F1 (0.63), while UFD has the highest F1 (0.70). Because F1 ignores correctly rejected real images, it also favors detectors that flag many images: EFFORT ranks second (0.68) despite a false-positive rate of 59% on Expert real images. F1 should therefore be read together with the false-positive rate.

G Additional Analyses by Image Characteristics

G.1 Generator-Balanced Statistical Sweeps

Figure 12 extends the statistic sweeps of Figure 5(b) to six further image statistics, using the same protocol: quantile bins are computed on real and generated images together, and the real reference of each AUC comes from the same bin. Since every generator has its own characteristic noise, spectral and aesthetic profile, the mixture of generators changes from bin to bin, and the overall AUC inherits this change; the generator-balanced AUC removes it by weighting the generators in each bin equally.

Pooled trends explained by generator composition. For estimated noise variance, the overall AUC rises from 0.73 in the lowest to 0.81 in the highest quintile, whereas the generator-balanced AUC stays between 0.774 and 0.781. High-frequency energy behaves in the same way (0.74 to 0.80 overall, 0.778–0.786 balanced), and for the LAION Aesthetics score the range shrinks from 0.72–0.79 to 0.76–0.77. For these three statistics the balanced spread retains only 8–13% of the pooled spread, and for noise variance and high-frequency energy the error rates are flat as well. Aesthetics is the exception: although its balanced AUC barely changes, the FPR rises from 8.8% to 16.0% and the FNR falls from 58.4% to 45.0% between the lowest and highest quintile. More aesthetic images of both classes thus receive higher scores, a shift that changes the behavior at a fixed threshold without improving the ranking.

Associations that persist. The object count keeps its association after balancing: the generator-balanced AUC increases from 0.767 for images with at most one confident detection to 0.804 for images with seven or more objects. This increase is accompanied by fewer false positives (FPR from 12.2% to 9.2%), while the FNR stays near 51%. Colorfulness shows a weaker, non-monotone association: the balanced AUC is 0.769 for the least colorful and 0.773 for the most colorful quintile, but 0.785–0.789 in between, and the FPR increases in the most colorful bin.

A limitation of balancing. Both AUC curves for aspect ratio peak for landscape images with ratios in $[1.35, 1.5)$ (balanced AUC 0.808, compared with 0.776–0.787 elsewhere). Kandinsky 5, the most difficult Expert generator, produces no images in this range, so this bin's balanced AUC averages six rather than seven generators. Repeating the sweep without Kandinsky 5 removes the peak (balanced AUC 0.800–0.814 in all bins). Generator balancing equalizes the weight of the generators present in a bin, but cannot compensate for a generator that is absent from it; associations with geometry should therefore be read together with the output sizes each generator supports. In the other five statistics, every generator contributes at least 140 images to every bin.

G.2 Generation Quality and Generator Profiles

Figures 13 and 14 characterize the open-source generators themselves, without reference to any detector. Both are computed on all 8.39M images generated for REALIS before filtering, and therefore describe each generator with the resolution and inference settings used in our pipeline, not architectures compared under identical settings.

Prompt-matched quality. Because each prompt was rendered by up to three of these generators, quality can be compared on identical prompts. We score all images with three automated preference models, HPSv3, PickScore and ImageReward, and compute for every pair of generators the share of common prompts (about 13k per pair) on which one image scores higher. HiDream-I1-Full wins most often under all three metrics (79.5, 85.2 and 77.0%), whereas SD-1.4, SD-1.5, SD-2.1 and YOSO-PixArt-1024 win at most 27% of their comparisons under any metric. The three rankings agree closely overall (pairwise Spearman correlation 0.84–0.88), but individual generators can move far: Janus-Pro-7B ranks 9th under HPSv3 but 28th under PickScore, and Switti-1024 ranks 8th under PickScore but 20th under HPSv3. As the confidence intervals are this narrow, such differences reflect what the scoring models reward rather than prompt sampling, and we report all three metrics instead of a single ranking. Averaged over the three metrics and over the generators of each split, the win rate rises from 42.7% in the training split to 55.5% in validation, 59.3% in test and 67.8% for the seven open-source Expert generators, in line with the assignment of generators to splits by generation quality (Section 3.1). Within Expert, the lowest-ranked generator, SD-3.5 Large-Turbo (54.0%), is also the easiest to detect (category-averaged AUC 0.93 in Figure 5(a)). However, the most difficult one, Kandinsky-5 (AUC 0.61), ranks only fourth of seven (67.9%), and the top-ranked

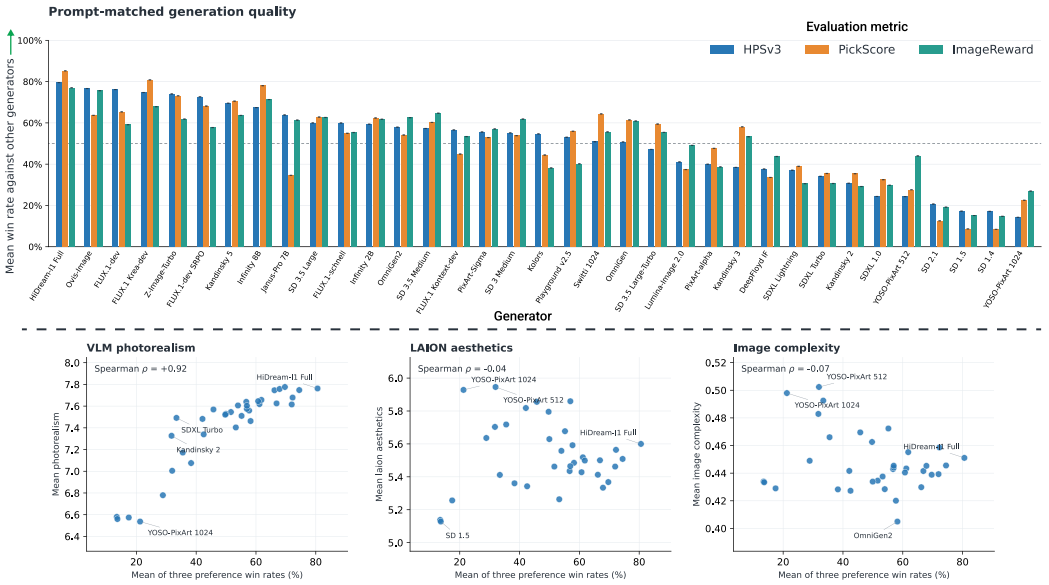

Figure 13: Generation quality of the open-source generators, measured on all 8.39M generated images before filtering. Top: mean prompt-matched win rate of each generator against the other 34 under three automated preference models. For a pair of generators, the win rate is the share of common prompts on which the generator's image receives the higher score, with ties counted as one half; each bar averages these pairwise win rates with equal weight over opponents. Generators are ordered by HPSv3, the dashed line marks 50%, and error bars show 95% prompt-cluster confidence intervals (at most ±0.16 percentage points). Bottom: generator means of VLM photorealism score, LAION Aesthetics and image complexity (as measured by IC9600 Feng et al. (2023) model) against the average of the three win rates, one point per generator.

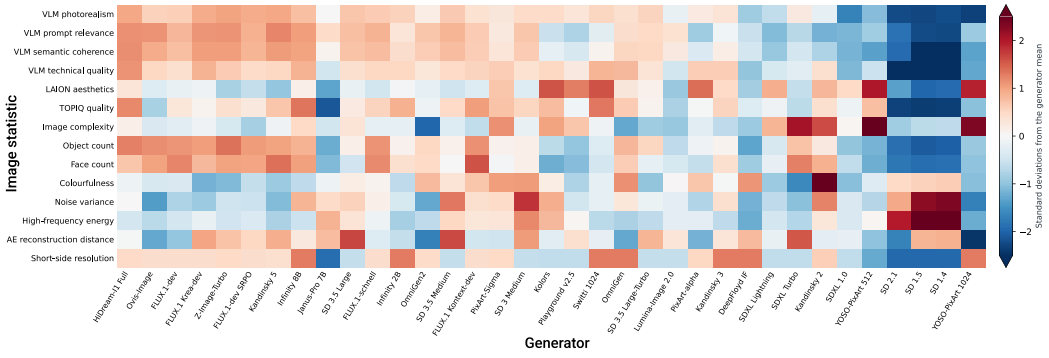

Figure 14: Image-statistic profiles of the open-source generators on the same unfiltered pool. Each cell is a generator's mean of a statistic, standardized across the generator means and clipped at ±2.5; color therefore compares generators within a row and does not indicate better or worse quality. Generators are ordered by HPSv3 win rate, as in Figure 13. VLM ratings are on a 1–10 scale and come from the scoring step of the VLM quality gate (Section 3.1); TOPIQ is a no-reference quality score; image complexity is the IC9600 estimate; AE reconstruction distance is the LPIPS distance used by the reconstruction filter; object count, colorfulness, noise variance and high-frequency energy are defined in Figure 12.

HiDream-I1-Full is comparatively easy to detect (AUC 0.83). Preference-based quality is therefore not a direct proxy for detection difficulty.

Generator profiles. Across generators, the mean win rate agrees with the VLM ratings (Spearman correlation 0.92 with photorealism and 0.93–0.95 with prompt relevance and semantic coherence) and, more weakly, with the numbers of detected objects (0.78) and faces (0.64). Aesthetic and complexity scores capture something else: their correlations with the win rate are -0.04 and -0.07, and the two YOSO-PixArt models, 29th and 32nd of 35 by win rate, have the highest mean LAION aesthetic and complexity scores of all generators. The oldest models, SD-1.4, SD-1.5 and SD-2.1, combine the lowest VLM ratings and the smallest outputs with the highest high-frequency energy (2.0 to 3.0 standard deviations above the generator average). Because generators differ this much, the filters that are applied identically to both classes still remove different shares of each generator's images. The reconstruction filter discards images that the autoencoder reconstructs unusually well, and it removes 67% of all YOSO-PixArt-1024 images, the generator with the lowest mean reconstruction distance, compared with at most 35% for any other generator.

H Post-Processing Sensitivity by Transformation and Strength

Evaluation details. The Expert-split analyses provided in Section 4.5 use real images and images from the seven open-source generators; proprietary generators are excluded. For each distortion-containing subset in Figure 6a, we compute clean and distorted AUC on identical images, average the seven generator-versus-real AUCs equally, and then average over the reference detectors. Figure 6a employs 14 representative detectors of various types selected by their performance on non-distorted splits: 6 fine-tuned detectors, 4 off-the-shelf detectors, and 4 VLMs. Figure 6b includes all 50 available paired detectors. Positive matched AUC loss in Figure 16 denotes clean minus distorted AUC. Error bars are pointwise 95% intervals obtained from paired AUC influence functions clustered by image stem, conditional on the chosen detectors and generators. Some off-the-shelf exports contain binary predictions, so their AUC reflects a single operating point rather than a continuous score ranking.

Individual transformations reveal variation within broad groups. Figure 15 shows nearly constant clean reference AUC across the transformation-containing subsets, but substantially different performance after processing. Mixed recompression, glass blur, and shot noise give the lowest mean distorted AUCs, whereas tone-curve, linear-contrast, and brightness changes are associated with smaller losses. The distinction also holds within broad categories: glass blur is more damaging than Gaussian or motion blur, and shot noise more damaging than Gaussian or multiplicative noise in these pipeline outputs. Broad labels such as "blur" or "noise" therefore conceal meaningful variation between operations. The detector-group heatmap shows a broadly shared ordering, with model-group differences in its magnitude. Even the lower-loss rows contain other processing stages, so they should not be interpreted as measurements of harmless individual transformations.

Sensitivity to severity depends on the transformation group. Figure 16 conditions on both group membership and the recorded nominal strength level. Blur and JPEG compression show pronounced increases in AUC loss with strength across all three detector groups. Color changes have flatter, less monotonic profiles, while repeated compression incurs a large loss throughout the range. Geometry is also detector-dependent: VLM losses increase with strength, whereas the off-the-shelf mean changes relatively little. The levels are operation-specific settings, not a common perceptual distance, and each point samples different images and co-occurring transformations. Only uniquely recoverable levels are used: repeated parameter values that cannot distinguish two nominal levels are excluded from these curves. These considerations are particularly relevant to neural and repeated compression, where the operation mix can vary between the plotted levels.

Enabling the pipeline causes a larger initial drop than subsequent strength increments. Figure 17 complements the Expert analysis with a separate validation sample of 5,000 real and 5,000 generated images spanning ten generators. The pipeline parameter m changes the sampling distribution over transformation strengths, while its other settings remain fixed; "base" denotes $m = 3$. Thus, $m = 1$ still enables a compound pipeline, rather than applying only level-1 transformations. For the configurations shown, mean pooled ROC-AUC falls from about 0.84 on clean images to 0.71 at the lowest pipeline setting, then to approximately 0.68 at the highest. Panel (b) normalizes each configuration by its clean AUC margin above chance, $(\mathrm{AUC}_m - 0.5)/(\mathrm{AUC}_{\mathrm{clean}} - 0.5)$; the mean

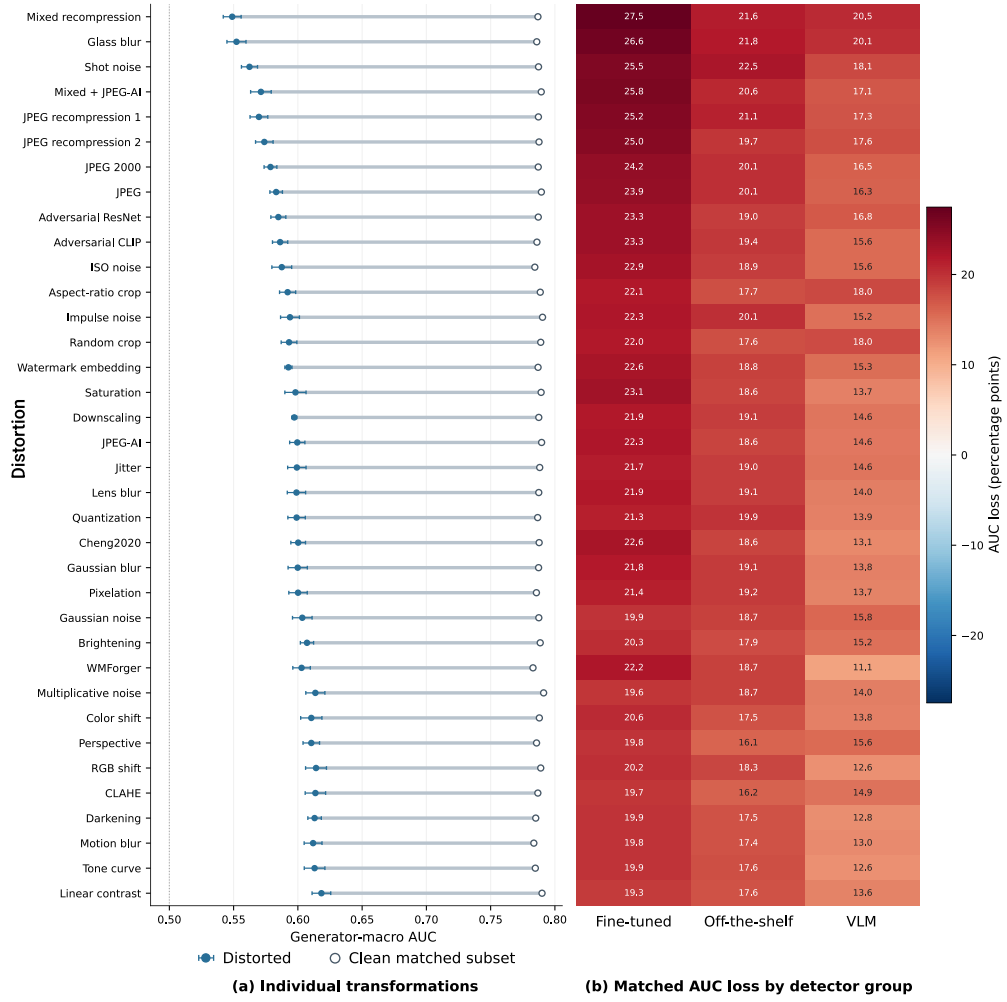

Figure 15: Individual transformations within the Expert post-processing pipeline. **(a)** Generator-macro AUC on each transformation-containing subset and its matched clean counterpart, averaged over the 14 reference detectors. **(b)** Matched AUC loss, in percentage points, averaged separately within each detector group. Rows are ordered by mean loss. Error bars in (a) are pointwise 95% prompt-clustered intervals for distorted AUC. The subsets overlap and can include additional transformations.

retained margin is roughly one half at the strongest setting. Gemma-4 31B retains more of its margin than the displayed Qwen3-VL configurations, consistent with their different relative losses on Expert. This auxiliary experiment uses pooled AUC and a different generator and detector cohort from the Expert analysis. Its strongest setting also has incomplete prediction coverage, so the endpoints describe available-case trends rather than a fully matched comparison across all settings.

Table 5: Transformations of the REALIS degradation pipeline, grouped as sampled (Appendix E); parameters are given from level 1 to level 5. PSNR (dB, RGB) and SSIM (luma): median fidelity to the unprocessed image when the transformation is applied *alone* at level 1 and at level 5, over 600 images (300 real, 300 generated).

Transformation	Description (level 1 to level 5)	PSNR, dB	SSIM
Blur			
Gaussian blur	Isotropic Gaussian filter, σ = 0.1–5 px.	∞–24.5	1.00–0.77
Lens blur	Uniform disk (defocus) kernel, radius 1–8 px.	38.6–24.0	0.99–0.76
Motion blur	Linear motion kernel at a random angle, length 1–10 px.	∞–28.4	1.00–0.88
Glass blur	Gaussian blur (σ=0.5) with local pixel swaps, offset 1–6 px.	25.2–19.5	0.78–0.50
Color distortion			
Color shift	Edge-weighted misregistration of the green channel, 1–12 px.	39.4–30.5	0.99–0.97
RGB shift	Random per-channel offset of up to $\pm$10–$\pm$50 (of 255).	34.0–19.7	1.00–0.98
Saturation	Saturation scaled by 0.4 to -0.4 (≤ 0: gray or inverted chroma).	23.2–15.4	0.99–0.90
Tone curve	Random per-channel spline tone curve, scale 0.05–0.4.	32.2–19.4	1.00–0.97
CLAHE	Contrast-limited adaptive histogram equalization, clip limit 1–6.5.	22.5–16.3	0.91–0.72
Algorithmic compression			
JPEG	JPEG compression, quality factor 45–7.	36.9–28.0	0.97–0.85
JPEG 2000	JPEG 2000 wavelet compression, ratio 16:1–170:1.	40.5–28.6	0.98–0.85
Neural compression			
JPEG AI	JPEG AI Ascenso et al. (2023) learned codec (v7), four decreasing rates (levels 4, 5 equal).	38.6–30.0	0.99–0.92
Cheng2020	Learned codec of Cheng et al. (2020) (anchor), quality 5–1.	37.1–31.7	0.96–0.91
Multiple and mixed recompression			
JPEG recompression 1	2–5 JPEG round trips, each with quality drawn from $[7, 43]$.	31.8–29.0	0.92–0.88
JPEG recompression 2	Three JPEG round trips, quality range $[15, 45]$ to $[7, 25]$.	32.4–29.1	0.93–0.88
Mixed recompression	2–5 round trips, each JPEG ($q \in [7, 43]$) or JPEG 2000 (16–170:1).	31.0–28.4	0.90–0.85
Mixed + JPEG AI	2–5 round trips, each JPEG ($q \in [7, 43]$) or JPEG AI (random rate).	30.9–27.5	0.93–0.89
Noise			
Gaussian noise	Additive Gaussian noise, $\sigma \approx$ 8–18 (of 255).	30.3–23.6	0.86–0.62
Multiplicative noise	Speckle noise $x(1 + n)$, $n \sim \mathcal{N}(0, v)$, v = 0.001–0.035.	36.5–21.6	0.96–0.59
Impulse noise	Salt-and-pepper noise on 0.1–2% of pixel values.	35.1–22.1	0.98–0.67
ISO noise	Camera-sensor luma and chroma noise, intensity 0.2–0.3 to 0.5–0.75.	29.2–17.5	0.78–0.36
Shot noise	Poisson (photon) noise, scale 0.025–0.075 to 0.2–0.3.	17.8–10.0	0.36–0.13
Brightness change			
Brightening	Spline curve raising mid-tones (RGB and CIELAB), amount 0.1–1.1.	29.1–9.3	0.99–0.66
Darkening	Spline curve lowering mid-tones, amount 0.05–0.8.	33.9–11.0	1.00–0.45
Geometric			
Random crop	Random crop keeping 80–40% of each side.	—	—
Aspect-ratio crop	Random crop, height 80–40%, aspect ratio drawn from $[0.5, 2]$.	—	—
Perspective	Random perspective warp, corner offset scale 2.5–7.5% to 30–40%.	—	—

List continues in Table 6.

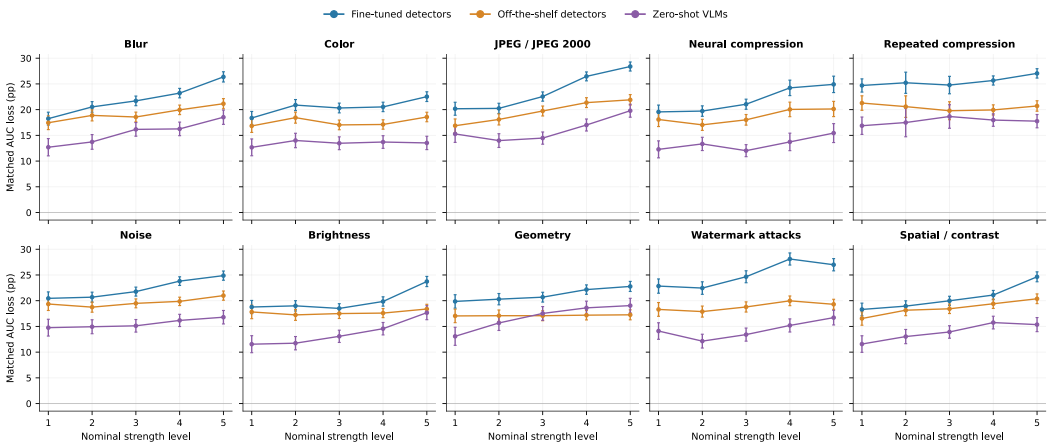

Figure 16: Matched AUC loss on Expert versus nominal transformation strength, conditioned on distortion-group membership. Curves average the fixed reference detectors within each detector group; error bars are pointwise 95% prompt-clustered intervals. Each clean reference uses the same images as its distorted counterpart. Only uniquely recoverable levels are included. Operation mixtures and other processing stages can differ across points.

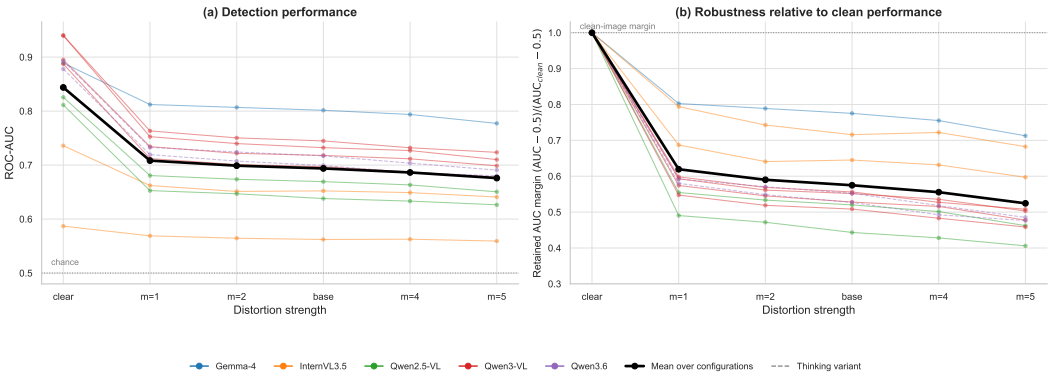

Figure 17: VLM sensitivity to the pipeline strength parameter on a separate validation sample. **(a)** Pooled ROC-AUC on clean images and pipeline outputs as m varies; "base" denotes $m = 3$. **(b)** Fraction of each configuration's clean AUC margin above chance retained after processing. Thin curves show individual configurations, dashed curves indicate thinking variants, and thick black curves average the configurations shown. The strongest setting has fewer available predictions; see the evaluation details in Appendix H.

Table 6: Continuation of Table 5 describing transformations used in REALIS.

Transformation	Description (level 1 to level 5)	PSNR, dB	SSIM
Adversarial and watermark-removal attacks			
Adversarial CLIP	PGD (100 steps) on the CLIP ViT-B/32 embedding, ℓ_∞ 4–12/255; method from An et al. (2024).	39.0–30.5	0.98–0.89
Adversarial ResNet	Same PGD attack on a ResNet-18 embedding, ℓ_∞ 4–12/255; method from An et al. (2024).	38.3–30.2	0.97–0.87
WMForger	Watermark removal method Souček et al. (2025) with a learned preference model, step 0.05–1.	36.4–29.1	0.97–0.85
Spatial distortion			
Pixelation	Nearest-neighbor down- and upsampling by 0.89–0.29.	25.7–21.8	0.85–0.72
Jitter	Random per-pixel displacement (5 iterations), std. 0.05–1 px.	43.0–25.3†	0.99–0.80
Quantization	Uniform quantization to 20–7 intensity levels.	31.5–22.0	0.97–0.88
Linear contrast	Spline contrast curve, offset 0, 0.15, -0.4, 0.3, -0.6 (<0 lowers contrast).	27.8–21.8†	0.97–0.93†
Applied after the chain			
Watermark embedding	Invisible watermark by one of 12 methods[§]; p=0.35.	—	—
Downscaling	Resize by a factor drawn from $[0.3, 0.8]$; p=0.75.	—	—

[§]Employed methods: ARWGAN (Huang et al. (2023)), DCT-Marker (Guzik et al. (2015)), DWSF (Guo et al. (2023)), FIN (Fang et al. (2023)), MaskWM (Hu et al. (2025)), MBRS (Jia et al. (2021)), PIMoG (Fang et al. (2022)), PixelSeal (Souček et al. (2025)), SSHidden (Zhu et al. (2018)), StegaStamp (Tancik et al. (2020)), SyncSeal (Fernandez et al. (2026)), TrustMark (Bui et al. (2025)). Implementations taken from WIBE framework (Yakushev et al. (2025)).