RETHINKING SHORTCUTS IN AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We identify a *computational shortcut* in AI-generated image detection: linear classifiers on the frozen visual encoders of two trained multimodal large language model (MLLM) detectors recover much of their accuracy without language decoding. Further analysis reveals that these detectors are sensitive to input resolution, suggesting potential *dataset shortcuts*. For example, matching fake-image aspect ratios to the real-image distribution lowers Deep-VRM's AIGI-Now Semantic accuracy from 93.8% to 63.8%. This sensitivity motivates a systematic metadata-only audit of 12 datasets and training resources, revealing label-correlated differences in file format, encoding, and image geometry. Guided by these findings, we propose an aligned visual learning framework based on image editing. Pairing real photographs with complete text-guided editing outputs under consistent preprocessing suppresses dataset shortcuts and enables joint learning of pixel-level and semantic forgery features in a single visual representation space. Our DINOv3-L detector achieves 91.3% mean accuracy across 12 in-the-wild domains and 92.4% on AIGI-Now, retaining over 90% on both subsets after resolution alignment. It also transfers to full-image synthesis and local manipulation on DailyBench.

1 INTRODUCTION

AI-generated image detection aims to distinguish synthetic images from real photographs in the wild. Reliable decisions require transferable visual evidence, ranging from pixel-level generation artifacts to semantic inconsistencies in object appearance and scene content (Yan et al., 2025a; Cheng et al., 2025). Recent multimodal large language model (MLLM) detectors incorporate visual localization, reasoning, and explanation into authenticity prediction (Chen et al., 2025a). These capabilities raise a fundamental question: *what visual evidence supports their final decisions, and how can training make this evidence transferable?*

We identify a *computational shortcut* through direct visual readout (Figure 1(a)). We freeze the vision transformers (ViTs) of the trained Deep-VRM (Lin et al., 2026) and Veritas++ (Tan et al., 2026b) detectors and train binary linear classifiers on their original training data. These readouts approach or exceed full-model accuracy on several benchmarks (Table 1), indicating that substantial detection information is already linearly accessible in the ViT without language decoding. This finding motivates directly optimizing visual representations and examining their learning signals.

However, strong visual readout does not establish which cues make these representations discriminative. File format, JPEG encoding, image geometry, and image-source differences can provide easier label predictors than subtle generation traces (Grommelt et al., 2024; Rajan et al., 2025; Guillaro et al., 2025). Learning these incidental correlations yields *dataset shortcuts*: spurious cues that models readily learn and that can inflate evaluation scores when training and test data share the same biases. In MLLMs, preprocessing that preserves input geometry can expose resolution and aspect ratio through visual-token counts and grid shapes (Bai et al., 2025), providing an additional route for dataset shortcuts (Figure 1(b)). Both studied detectors exhibit substantial sensitivity to geometric processing: center cropping or real-fake resolution alignment reduces Veritas++ accuracy by more than 20 percentage points on two benchmarks (Section 3). Moreover, these latent distributional biases are also present in the detectors' training data: a linear probe on the pretrained Qwen-VL visual encoder, without any detector training, already attains seemingly strong accuracy on benchmarks that share these biases, such as AIGI-Now.

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.

This sensitivity calls for explicit control over the processing pipeline when assessing detection performance. To ground these controls, we audit file format, JPEG encoding, and image geometry across 12 datasets and training resources using metadata-only probes. The audit reveals correlations between processing properties and authenticity labels, which in turn guide training-data construction and evaluation under controlled processing conditions. Motivated by aligned-data approaches (Rajan et al., 2025; Chen et al., 2025b), we construct aligned editing data from source-paired object additions and replacements (Figure 1(c)): each fake image is produced by editing a real source photograph with a text instruction. Source pairing aligns content, scene, and style, while matching image size and encoding reduces processing-related label cues. Retaining complete edited outputs preserves two complementary forms of forgery evidence: (i) semantic evidence within the edited regions, such as implausible object appearance and scene incompatibility; and (ii) pixel-level generation and reconstruction traces across the full output, including regions untouched by the semantic edit. Binary authenticity supervision on the real and edited images jointly adapts a single ViT with a linear classifier to learn both families of cues under controlled source and processing variation.

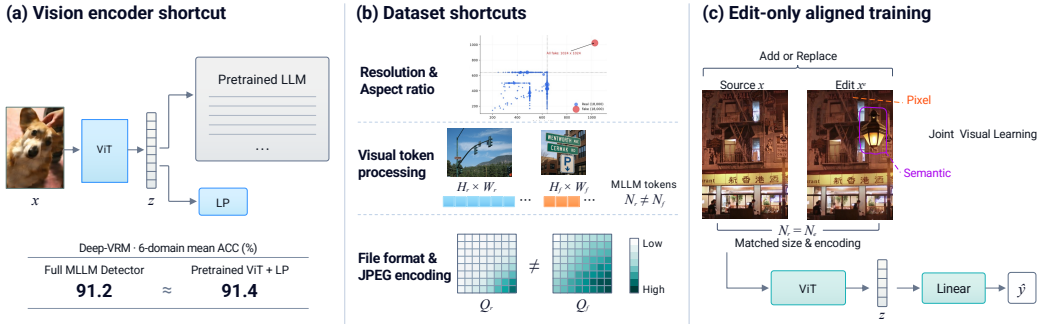


Figure 1: **From shortcut diagnosis to aligned visual learning.** (a) A linear probe on the frozen ViT of trained Deep-VRM matches the full detector's six-domain mean accuracy (91.4% vs. 91.2%), revealing linearly readable detection cues without language decoding. (b) Resolution and aspect-ratio biases can become label cues through visual-token counts; file format and JPEG encoding provide additional dataset shortcuts. (c) Source-paired object additions and replacements with matched size and encoding reduce these shortcuts, while a single ViT and linear head jointly learn image-wide pixel traces and local semantic cues.

We instantiate this training design as a DINOv3 (Siméoni et al., 2025) encoder equipped with a binary prediction head, trained on the source-controlled Mani-X editing pairs (Li et al., 2025b). Evaluation spans four established benchmarks; five in-the-wild benchmarks comprising 12 domains; AIGI-Now; and DailyBench (Jiang et al., 2026). The detector attains 95.7% mean accuracy on the established benchmarks, 91.3% across the 12 in-the-wild domains, and 92.4% on AIGI-Now, and retains strong accuracy under JPEG compression, platform transmission, and re-digitization. On DailyBench, it achieves state-of-the-art performance on both full-image generation and image manipulation detection. Under input-resolution alignment on AIGI-Now, the Qwen3.5-4B model adapted on the visual side with the same data exhibits smaller accuracy drops than all evaluated MLLM baselines.

Our contributions are:

- A systematic audit of format, encoding, and geometric shortcuts across 12 datasets and training resources, with quantitative metadata-only probes and explicit processing controls for evaluating detector robustness.
- A computational shortcut through direct visual readout in MLLM detectors: linear classifiers on frozen detector ViTs approach or exceed full-model accuracy on several benchmarks, revealing accessible authenticity information in visual representations.
- A source-paired editing-data framework that controls source and processing biases while preserving pixel and semantic forgery evidence for visual encoder adaptation, supported by extensive cross-domain evaluations.

2 Related Work

2.1 Generalizable AI-Generated Image Detection

Earlier detectors rely on low-level artifacts (Tan et al., 2024) or pretrained semantic features (Ojha et al., 2023; Yan et al., 2025b), while more recent methods combine both types of evidence through dedicated artifact and semantic modules (Yan et al., 2025a; Cheng et al., 2025; Chen et al., 2025a; Lin et al., 2026; Guo et al., 2026; Qin et al., 2026). Recent MLLM-based image and video detectors further enhance visual perception and reasoning for authenticity assessment (Wang et al., 2026; Tan et al., 2026a). Our experiments reveal a computational shortcut in the vision encoders of representative MLLM detectors, underscoring the central role of visual representations in this task. Unlike approaches that model artifact and semantic evidence in separate modules, we leverage source-paired image-editing data to expose a single visual representation to both types of forgery evidence under controlled variations in source and processing.

2.2 Dataset Shortcuts and Aligned Training

Recent studies attribute generalization failures to dataset biases in file format, resolution, and semantic content, which can induce shortcut learning (Grommelt et al., 2024; Rajan et al., 2025; Guillaro et al., 2025). To reduce these biases, existing approaches align real and fake data through autoencoder reconstruction (Rajan et al., 2025), pixel- and frequency-domain consistency (Chen et al., 2025b), or source-conditioned generation with inpainting augmentation (Guillaro et al., 2025; Corvi et al., 2025). Building on these efforts, we use metadata-only probes to systematically characterize common dataset shortcuts, and adopt source-paired image-editing data with consistent preprocessing to mitigate them, thereby improving generalization and robustness.

3 From Shortcut Diagnosis to Unified Visual Learning

We investigate two forms of shortcuts in AI-generated image detection. First, a *computational shortcut*: linear readouts of the ViT features of trained MLLM detectors recover much of their detection performance without language decoding. Second, *dataset shortcuts*: file format, encoding, and geometry metadata correlate with authenticity labels, and resolution-dependent visual processing in MLLMs can turn these biases into prediction cues. Together, these findings motivate a bias-controlled training framework that learns authenticity directly in one visual feature space.

3.1 A Computational Shortcut: Direct Visual Readout

We first ask how much detection information is accessible without language decoding. A frozen visual encoder f_ϕ, a pooling operator P, and a trainable linear head (w, b) define a probing score

$$s_{\mathrm{LP}}(x) = \sigma\big(w^\top P(f_\phi(x)) + b\big)\,. \tag{1}$$

We freeze the visual encoders from the official Deep-VRM (Lin et al., 2026) and Veritas++ (Tan et al., 2026b) checkpoints and train linear readouts on their respective training data sources. For Deep-VRM, each image passes through both the semantic and low-level visual encoders. We mean-pool the post-merger tokens from each encoder into a d_1-dimensional vector, concatenate the two vectors into $2d_1$-dimensional features, and train only a linear classifier. For Veritas++, we sample 25k real and 25k fake images from the data sources reported in the original paper, and mean-pool the final post-merger visual tokens into a d_2-dimensional vector for linear classification. We also probe the corresponding base Qwen-VL visual encoders with the same training data, to measure how much detection information detector training adds.

Table 1 shows that the Deep-VRM visual readout reaches 91.4% mean accuracy across six domains, close to 91.2% for the full detector and well above 76.8% for the base Qwen2.5 visual probe. The Veritas++ readout closely matches the full model on AIGI-Now, although its six-domain mean is lower (89.1% vs. 94.1%). The base Qwen3 visual probe also approaches full-model performance on AIGI-Now (96.1%/95.3% on Pixel/Semantic), but remains near chance on the other four domains

Table 1: **Full detectors and frozen visual readouts: ACC (%).** We compare Deep-VRM (Lin et al., 2026) and Veritas++ (Tan et al., 2026b) with linear probes (LPs) on their frozen ViTs and the corresponding base Qwen-VL encoders. LPs use the respective detectors' training data. LPs on the detector ViTs recover much of the full models' detection performance without language decoding, indicating that these ViTs already encode substantial linearly readable forgery cues.

Model	SynthWildX	WildRF	ComAI	SocRF	Pixel	Semantic	AVG
Deep-VRM	88.6	87.5	97.6	81.7	97.7	93.8	91.2
Qwen2.5-VL ViT LP	66.7	69.5	77.6	63.9	90.7	92.1	76.8
Deep-VRM ViT LP	83.2	88.4	98.8	85.8	97.6	94.5	91.4
Veritas++	92.4	95.9	90.5	91.5	97.8	96.7	94.1
Qwen3-VL ViT LP	54.3	58.0	58.2	52.2	96.1	95.3	69.0
Veritas++ ViT LP	85.2	89.3	97.0	69.5	97.5	96.3	89.1

Table 2: **Metadata-only shortcut probes: bACC / AP (%).** We report five-fold cross-validation results for depth-3 decision trees using only filename extension (Format), aspect ratio, color mode (Color), or JPEG encoding features. Higher scores relative to baseline indicate stronger metadata shortcuts between real and fake images. Baselines: 50% bACC; fake-image proportion for AP.

Dataset	AP baseline	Format	Aspect ratio	Color	JPEG encoding
Established datasets and training resource					
GenImage	50.0	**100 / 100**	**98.0 / 95.9**	56.0 / 58.9	**100 / 100**
UniFD	50.0	**100 / 100**	54.5 / 51.4	51.5 / 51.3	**100 / 100**
ARForensics	50.0	**100 / 100**	**97.9 / 95.8**	50.1 / 50.0	**100 / 100**
AIGI-Now	50.0	66.7 / 71.4	**98.5 / 96.9**	49.8 / 49.9	74.2 / **84.4**
DiTFake	50.0	**100 / 100**	**98.1 / 96.1**	50.3 / 50.4	**100 / 100**
Holmes-Set	50.0	**85.0 / 87.8**	**98.1 / 96.3**	50.0 / 50.0	**99.9 / 100**
Mani-X	49.9	50.0 / 49.9	59.8 / 67.9	50.0 / 49.9	50.0 / 49.9
In-the-wild datasets					
AIGIBench	50.0	**83.3** / 74.6	62.9 / 67.2	50.5 / 50.4	**83.3** / 74.6
RRDataset	50.0	65.0 / 69.7	78.3 / 77.8	50.0 / 50.0	66.6 / 76.0
WildRF	53.0	69.5 / 73.6	79.5 / **86.1**	51.0 / 53.8	74.8 / **81.7**
SynthWildX	76.0	49.1 / 75.8	**82.4 / 94.0**	49.1 / 75.8	63.1 / **87.5**
Chameleon	42.9	49.7 / 43.0	70.3 / 63.6	49.7 / 43.0	**100 / 100**

(52.2–58.2%). This contrast suggests that shared biases between the training data and AIGI-Now may support shortcut learning, which motivates the dataset diagnosis below.

3.2 Dataset Shortcuts in Existing Data

File format, encoding, and image geometry can provide label cues unrelated to generation artifacts (Grommelt et al., 2024), and pretrained encoders can retain such acquisition and processing information (Stojnić et al., 2026). To uncover and understand such shortcuts in existing datasets, we audit 12 datasets and training resources with metadata-only probes: for each attribute (file format, aspect ratio, color mode, JPEG encoding), we train a depth-3 decision tree to predict the authenticity label, which quantifies these associations without accessing image pixels. Table 2 reports bACC and AP, with baselines of 50% and the fake-image proportion, respectively.

File format alone separates real from fake images perfectly in GenImage, UniFD, ARForensics, and DiTFake, and strongly in Holmes-Set and AIGIBench, because the two classes are stored in different formats (e.g., real images as JPEG, fake images as PNG). Resolution-related biases are pronounced in GenImage, ARForensics, DiTFake, Holmes-Set, and AIGI-Now, whose aspect-ratio probes reach about 98% bACC; absolute dimensions can reveal additional biases, as in Chameleon's fixed-width pattern (Figure 2). Figure 6 provides the full resolution-distribution audit. Implicit JPEG encoding can leak labels even under a shared file format: Chameleon's format probe is near chance (49.7% bACC), yet its JPEG attributes separate the labels perfectly.

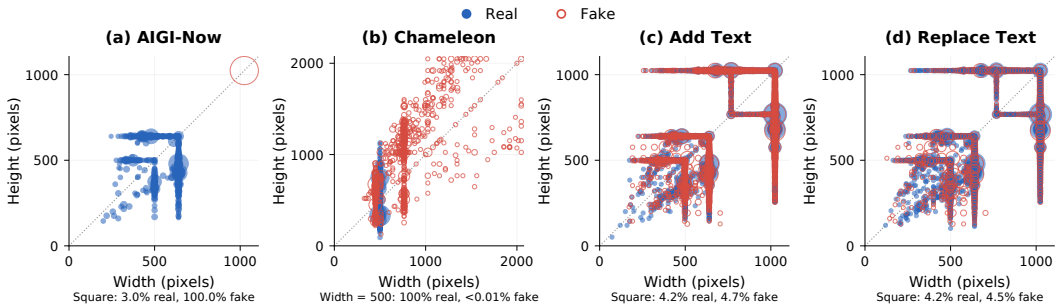


Figure 2: **Geometry reveals labels, including within the editing resource.** (a) AIGI-Now fake images are all 1024×1024, whereas only 3.0% of real images are square. (b) Chameleon has a less obvious fixed-width pattern: all real images have width 500, compared with only one fake image. (c,d) Mani-X matches real-fake resolution distribution.

Whether these biases become learned shortcuts depends on which features the detector can capture and whether preprocessing preserves or amplifies them. In MLLMs with geometry-preserving visual processing, for example, visual-token counts and grid shapes expose the input resolution and aspect ratio (Bai et al., 2025). Shared resolution biases between the training data and AIGI-Now can therefore act as resolution shortcuts, inflating benchmark accuracy beyond genuine generalization. As shown in Figure 3, we randomly crop 5–25% of the side pixels so that fake images match the real aspect-ratio distribution; this reduces the AIGI-Now accuracy of Deep-VRM and Veritas++ by up to 30 percentage points. These drops show that both detectors are sensitive to input processing and likely rely on the shared resolution bias: once this cue is removed, accuracy falls sharply.

3.3 Source-Paired Editing Data for Unified Visual Learning

Source-Paired Data Construction. The above diagnoses shape our training design: source pairing and consistent preprocessing suppress dataset shortcuts, while complete edited outputs expose both local semantic changes and image-wide pixel traces to a single visual representation.

We construct training data from Mani-X (Li et al., 2025b), which pairs source photographs with outputs from mask-guided inpainting and instruction-based editing. Our main configuration uses *Add Text* and *Replace Text* (text-conditioned object addition and replacement); each training pair (x_i, x_i^e) links a real source photograph to its edited output. As shown in Figure 2(c,d), the real and fake images in our training subsets have closely matched original resolution distributions, which reduces potential resolution shortcuts. Appendix A lists the six editors and their underlying generative models, including SD 1.5 and FLUX.1-dev.

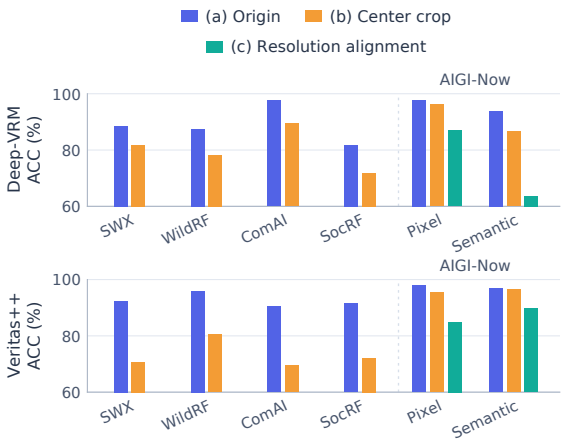


Figure 3: **Accuracy under geometric processing.** (a) Origin; (b) Center crop; (c) resolution alignment on AIGI-Now (Table 3) by altering the aspect ratio of fake images to match the distribution of real images.

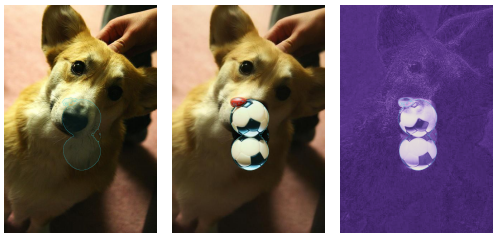
Pixel and Semantic Evidence in One Representation. We retain the full edited output, $x_i^{\text{full}} = x_i^e$, including changes outside the intended edit region (Figure 4). *Pixel evidence* covers texture, frequency, and reconstruction changes introduced by the editing pipeline; *semantic evidence* concerns the edited object's appearance and its compatibility with the scene. The two evidence types coexist within and beyond the edited region. Source-conditioned and aligned training already offer ways to reduce content discrepancies (Rajan

et al., 2025; Guillaro et al., 2025; Chen et al., 2025b). Keeping the complete output goes further: a shared encoder sees local content changes together with image-wide processing traces, which can arise from autoencoder reconstruction (Nebioglu et al., 2026). Unlike detectors with separate pixel and semantic branches (Chen et al., 2025a; Cheng et al., 2025; Lin et al., 2026), ours receives one image and learns a single representation from both forms of evidence.

Direct Visual Detector. Our primary detector fully fine-tunes DINOv3-Large (Siméoni et al., 2025). The [CLS] token of its final layer serves as the image representation $z_\theta(x)$, and a linear binary head produces the score

$$s_\theta(x) = \sigma\big(w^\top z_\theta(x) + b\big)\,, \tag{2}$$

We train the detector with binary cross-entropy (BCE) on the real/edited labels. Unlike the frozen probes in Equation 1, the encoder itself adapts to the detection task, and inference maps a single image directly to an authenticity score.



Source Edited Difference

Figure 4: **Local edits and broader changes coexist.** The outlined region marks the intended edit; the difference map shows changes across the full output.

We adopt Qwen3.5-4B as a representative MLLM. We prompt it with a fixed real/fake question and supervise the answer with token-level cross-entropy on the binary response (SFT). By default, we freeze the LLM and fully fine-tune the ViT, which directs adaptation toward visual perception while retaining the language-based prediction interface.

4 Experiments

4.1 Experimental Setup

Benchmarks and metrics. AIGI-Now tests nine generators in each of its Pixel and Semantic subsets (Chen et al., 2025a). We also evaluate on 12 domains from Chameleon (Yan et al., 2025a), SynthWildX (Cozzolino et al., 2023), WildRF (Cavia et al., 2024), AIGIBench (Li et al., 2025c), and RRDataset (Li et al., 2025a). Established benchmarks comprise GenImage, UniFD, ARForensics, and HolmesSet (Zhu et al., 2023; Ojha et al., 2023; Zhang et al., 2025; Zhou et al., 2025). DailyBench adds modern full-image synthesis and local object editing (Jiang et al., 2026). We report accuracy (ACC, %) and, for the class-imbalanced DailyBench manipulation subsets, balanced accuracy (bACC). All means weight each domain or generator subset equally.

Comparisons and configurations. We compare visual detectors NPR (Tan et al., 2024), FatFormer (Liu et al., 2024), RINE (Koutlis & Papadopoulos, 2024), Effort (Yan et al., 2025b), CO-SPY (Cheng et al., 2025), SPAI (Karageorgiou et al., 2025), GAPL (Qin et al., 2026), and MIRROR-H+ (Liu et al., 2026); aligned-data detectors DDA (Chen et al., 2025b) and B-Free (Guillaro et al., 2025); and MLLM detectors Deep-VRM (Lin et al., 2026) and Veritas++ (Tan et al., 2026b). If the original paper reports results on the corresponding benchmark, we directly cite the reported numbers. Otherwise, we use the authors' officially released pretrained weights and test-time preprocessing, and evaluate the detectors under the same environment as our method. DailyBench additionally uses published official-checkpoint results from Jiang et al. (2026). Appendices B and C provide protocol details and complete results.

Implementation details. Unless otherwise specified, we use DINOv3-Large. Each image is resized to a 256-pixel short side with the aspect ratio preserved, followed by a random square crop for training and a center crop for validation. We fully fine-tune the detector for five epochs with the Adam optimizer, a batch size of 64, and a learning rate of 10^{-6}. Table 6 also evaluates larger inputs and the H+ backbone. For Qwen3.5-4B, we fully fine-tune the ViT with the LLM frozen, using token-level cross-entropy on real/fake answers (Section 3.3).

Table 3: **AIGI-Now accuracy (%).** Pix/Sem: Pixel/Semantic. AIGI-Now exhibits resolution bias (Section 3); we report results under the original settings and after aligning the real and fake resolution distributions through random edge cropping. Δ: aligned minus original ACC (percentage points).

Method	Nano Banana		GPT-4o		Jimeng		Kling		Minimax		FLUX Pro		FLUX Krea		FLUX Dev		FLUX Kontext		**AVG**	
	Pix	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*
Vision-only Detectors																				
NPR	76.3	47.9	77.7	47.5	29.0	46.3	77.5	48.7	71.8	46.9	30.8	47.1	55.4	46.7	76.8	46.6	76.6	49.4	63.5	47.4
FatFormer	56.1	50.0	51.5	49.9	48.7	50.0	56.8	50.0	49.1	50.0	48.7	50.0	50.3	50.0	53.5	50.0	64.8	50.0	53.3	50.0
Effort	80.8	66.0	76.9	58.6	54.1	56.0	80.2	68.3	79.2	69.2	70.6	69.1	82.4	67.6	81.1	68.3	75.0	61.5	75.6	64.9
CO-SPY	75.6	58.3	80.5	62.6	83.5	75.7	88.6	78.5	78.4	65.5	77.7	71.8	89.7	74.1	88.3	74.5	65.3	64.3	80.8	69.5
DDA	81.6	56.2	**92.3**	65.4	87.0	55.7	96.0	64.6	83.3	50.5	76.6	55.0	59.4	50.0	91.6	51.2	82.7	52.9	83.4	55.7
B-Free	68.1	51.1	57.2	54.3	81.2	55.0	89.0	60.3	54.2	49.2	73.4	53.5	58.0	50.5	76.5	52.2	82.0	53.0	71.0	53.2
SPAI	85.8	62.7	51.4	65.2	86.3	64.4	88.6	65.0	82.2	63.4	79.1	62.7	91.9	62.0	83.4	64.5	90.8	61.5	82.1	63.5
GAPL	84.1	65.3	90.6	55.8	77.3	55.5	**98.3**	57.4	79.6	53.6	76.4	60.5	81.0	60.6	95.9	66.7	**92.3**	64.2	86.1	59.9
MIRROR-H+	80.1	84.4	87.1	83.3	**94.3**	90.2	89.2	87.4	83.0	71.6	93.1	90.9	81.0	70.3	97.4	93.7	78.4	71.8	87.0	82.6
Ours DINOv3	97.3	94.5	84.6	90.9	90.8	95.4	89.6	92.8	90.7	89.5	98.1	95.9	96.8	93.1	99.0	97.5	82.0	85.8	92.1	92.8
Resolution-aligned	**97.3**	**94.5**	80.8	**91.6**	86.7	**94.2**	88.4	**92.6**	**88.8**	**85.4**	**97.9**	**95.4**	**93.9**	**89.5**	**98.8**	**97.7**	82.0	**83.6**	**90.5**	**91.6**
Δ (pp)	0.0	0.0	−3.8	+0.7	−4.2	−1.2	−1.3	−0.2	−1.9	−4.1	−0.2	−0.5	−2.9	−3.6	−0.2	+0.2	0.0	−2.2	−1.6	−1.2
MLLM-based detectors																				
Deep-VRM	99.9	99.0	100.0	97.0	99.4	96.5	99.9	97.3	93.6	92.4	96.2	97.3	94.4	88.5	99.1	96.3	96.7	79.6	97.7	93.8
Resolution-aligned	**98.1**	79.4	**97.9**	58.2	75.2	66.3	**97.9**	72.8	87.3	55.7	65.8	67.9	71.0	54.1	96.5	67.4	**93.3**	52.3	87.0	63.8
Δ (pp)	−1.9	−19.7	−2.2	−38.8	−24.2	−30.3	−2.0	−24.5	−6.3	−36.8	−30.5	−29.5	−23.5	−34.4	−2.7	−29.0	−3.5	−27.4	−10.7	−30.0
Veritas++	99.2	99.4	99.5	99.2	99.4	98.6	99.2	97.8	97.4	95.0	97.5	98.2	99.0	95.4	99.1	98.8	90.3	87.9	97.8	96.7
Resolution-aligned	89.4	**95.6**	91.2	**94.2**	86.0	94.0	86.2	91.7	77.8	83.0	79.7	91.6	**85.6**	84.5	94.2	94.2	74.7	78.5	85.0	89.7
Δ (pp)	−9.8	−3.9	−8.3	−5.1	−13.4	−4.6	−13.0	−6.2	−19.6	−12.1	−17.9	−6.6	−13.4	−10.9	−5.0	−4.6	−15.6	−9.4	−12.9	−7.0
Ours Qwen3.5-4B	96.0	95.8	90.8	95.3	97.0	95.1	89.3	95.1	94.9	91.8	98.1	97.0	88.4	91.0	98.7	97.9	76.7	86.5	92.2	93.9
Resolution-aligned	92.1	94.2	84.2	93.7	**95.1**	**94.1**	84.9	**93.5**	**90.8**	**89.4**	**97.1**	**95.8**	83.6	**88.0**	**97.8**	**96.6**	75.9	**85.2**	**89.0**	**92.3**
Δ (pp)	−3.9	−1.6	−6.6	−1.6	−1.9	−1.0	−4.4	−1.6	−4.1	−2.4	−1.0	−1.2	−4.8	−3.0	−0.9	−1.3	−0.8	−1.3	−3.2	−1.6

Table 4: **In-the-wild accuracy (%).** †: Deep-VRM training source overlap on Chameleon.

Method	Chame-leon	SynthWildX			WildRF			AIGIBench		RRDataset			AVG
		DALL-E 3	Firefly	Midj.	FB	Reddit	Twitter	ComAI	SocRF	Orig.	Trans.	Re-digit.	
NPR	59.5	40.6	63.5	46.5	74.1	63.8	48.5	55.1	53.3	76.4	49.8	56.5	57.3
FatFormer	58.0	47.4	59.7	49.8	47.4	59.7	49.8	51.9	56.9	57.1	50.1	51.8	53.3
RINE	57.4	45.7	68.6	48.0	52.5	67.8	45.8	51.2	56.0	58.0	48.5	49.7	54.1
Effort	63.9	51.6	55.7	47.9	57.8	58.3	67.9	56.8	51.4	63.1	50.0	56.3	56.7
CO-SPY	71.1	72.0	79.7	70.8	73.1	77.3	75.3	66.7	68.2	72.0	74.3	66.7	72.3
SPAI	60.3	74.4	79.8	67.5	80.9	77.1	83.9	61.9	71.4	72.9	62.0	54.4	70.5
DDA	84.1	92.4	87.2	93.1	93.1	86.4	91.2	84.6	81.7	91.8	73.3	62.9	85.1
B-Free	78.7	**95.7**	**95.8**	**95.6**	**95.3**	86.5	**97.9**	81.5	85.2	81.6	73.8	61.5	85.8
GAPL	70.2	87.3	87.9	84.7	88.1	82.1	90.0	76.0	80.1	88.0	78.1	63.7	81.3
Deep-VRM	95.1†	86.6	79.2	85.1	74.7	90.1	88.1	**97.6**	81.7	84.3	81.2	60.3	83.7
Veritas++	90.6	94.2	85.9	93.6	95.0	**95.9**	96.7	90.5	91.5	88.8	85.5	71.6	90.0
Ours	90.6	94.0	82.7	94.1	92.5	94.3	95.7	94.0	**92.5**	**94.1**	**88.4**	**82.4**	**91.3**

4.2 Generalization and Robustness

AIGI-Now Pixel and Semantic detection. Many conventional vision detectors struggle on AIGI-Now, especially on its Semantic subset (Table 3), where heavy downsampling, noise, blur, and JPEG compression suppress pixel traces (Chen et al., 2025a). This setting exposes the difficulty of transferring detectors that rely on fragile generation artifacts to heavily processed images. The strong MLLM results before resolution alignment may reflect both semantic understanding and dataset shortcuts (Section 3): all fake images are square, whereas real images have diverse aspect ratios. Geometry-preserving MLLM inputs expose this difference through visual-token grids even after pixel degradation, leaving a label cue that survives the Semantic subset's processing. Resolution alignment substantially lowers the accuracy of both baseline MLLMs; Deep-VRM loses 30.0 points on Semantic. Our models, trained with paired and aligned editing data, retain 90.5%/91.6% Pixel/Semantic ACC with DINOv3 and 89.0%/92.3% with Qwen3.5-4B. Their smaller drops show that this training paradigm preserves useful detection evidence even when the label-related geometry cue is removed. DINOv3's strong Semantic result further shows that a visual encoder can handle this degraded setting without language decoding.

In-the-wild benchmarking results. Our detector achieves the highest mean ACC across the 12 domains (91.3%), exceeding both visual and MLLM baselines (Table 4). Its lead on RRDataset persists after platform transmission and re-digitization, where it retains 88.4% and 82.4% ACC. These conditions are particularly informative because they change the original encoding and acquisition

characteristics of the images. The sustained advantage shows that the learned evidence transfers beyond untouched generator outputs to images that have passed through real distribution and acquisition pipelines, which supports the practical value of training with controlled source and processing differences.

Established benchmarks. JPEG alignment reduces the format-related label cues identified in Section 3, while also altering forensic traces. Deep-VRM and Veritas++ fall from 91.0%/81.9% to 75.0%/77.4% mean ACC, below DDA (88.6%), B-Free (92.5%), and ours (95.7%) under JPEG alignment (Table 5). Our detector leads under both the Original and JPEG conditions (94.8%/95.7%). Together with the in-the-wild results, these comparisons show that the two evaluated MLLM detectors still trail leading purely vision-based methods in overall binary detection, especially after encoding alignment.

Table 5: **Established-benchmark accuracy (%).** Cells: Original / JPEG (Q=96). Detailed results are in Appendix.

Method	GenImage	UniFD	ARForensics	HolmesSet	AVG
NPR	77.2 / 48.5	**94.6** / 48.8	76.8 / 39.8	64.5 / 41.6	78.3 / 44.7
Effort	81.0 / 69.0	92.0 / 73.2	81.7 / 66.3	74.5 / 71.0	82.3 / 69.9
CO-SPY	78.0 / 77.4	76.9 / 76.3	69.6 / 68.4	83.8 / 83.1	77.1 / 76.3
DDA	88.1 / 88.7	73.6 / 81.6	83.3 / 91.1	94.8 / 93.1	85.0 / 88.6
B-Free	88.8 / 89.7	89.1 / 89.9	91.9 / 93.1	96.9 / **97.1**	91.7 / 92.5
Deep-VRM	94.1 / 77.8	85.2 / 65.8	92.7 / 67.2	92.1 / 89.2	91.0 / 75.0
Veritas++	82.9 / 78.5	76.0 / 68.8	76.0 / 69.8	92.8 / 92.4	81.9 / 77.4
Ours	**95.6** / **97.0**	89.0 / **90.6**	**97.6** / **98.4**	**97.0** / 96.7	**94.8** / **95.7**

Modern synthesis and local manipulation. DailyBench tests transfer to recent generators, including GPT-Image 2 and Nano Banana 2, on both full-image synthesis and local manipulation (Jiang et al., 2026). DINOv3-H+ 512p exceeds the strongest published baselines by 5.9 ACC points on synthesis and 5.2 bACC points on manipulation (Table 6), consistent with training on complete editing outputs that expose both local content changes and image-wide generation traces. With this recipe fixed, larger DINOv3 backbones and higher input resolutions improve both task averages. Table 10 in Appendix B.1 extends this analysis across encoders and scales.

Table 6: **DailyBench results (%).** Means over seven subsets. Baselines: Jiang et al. (2026).

Method	Fake ACC	Manipulation ACC	Manipulation BACC
NPR	67.5	51.7	54.4
Effort	60.2	57.3	54.6
B-Free	52.7	55.8	61.7
DDA	67.7	62.8	66.9
DINO-Det	76.5	50.3	58.2
FPD	79.8	58.8	63.4
Our models			
DINOv3-L 256p	82.6	63.1	67.0
DINOv3-L 512p	85.0	71.6	73.8
DINOv3-H+ 256p	84.4	67.1	71.2
DINOv3-H+ 512p	**85.7**	**71.7**	**76.0**

Table 7: **Training-data ablations (ACC, %).** Estab.: Original condition. Add/Replace: Text variants.

Training data	Estab.	Wild	AIGI-Now Pixel	AIGI-Now Sem.
(a) Training-data recipe				
pixel-aligned (DDA)	**95.0**	85.7	68.6	62.8
semantic-aligned (B-Free)	93.4	82.5	60.1	72.4
pixel+semantic (Ours)	94.8	**91.3**	**92.1**	**92.8**
(b) Mani-X training subset				
Add	91.1	90.1	91.5	92.2
Replace	91.3	**91.3**	90.3	92.3
All subsets	94.6	90.1	89.6	86.7
Add + Replace (Ours)	**94.8**	**91.3**	**92.1**	**92.8**
(c) Training-generator family				
SD 1.5	91.0	74.3	50.9	68.7
FLUX	80.9	90.7	**93.8**	91.8
SD 1.5 + FLUX (Ours)	**94.8**	**91.3**	92.1	**92.8**

4.3 Training Ablations

Table 7 varies training-data composition with the backbone, input resolution, training-set size, and update count fixed. DDA and B-Free denote data recipes under this shared training setup.

Training-data recipe. Mani-X and DDA reach nearly matched accuracy on established benchmarks, but this similarity masks a large transfer gap: Mani-X performs better in the wild and on both AIGI-Now subsets. Established-benchmark accuracy alone therefore gives an incomplete measure of how broadly useful the learned evidence is. The same transfer advantage over DDA and B-Free appears on CLIP-L and DINOv2-L (Appendix B.1), which shows that aligned editing data contributes to generalization across pretrained encoders. *Add Text + Replace Text* gives the best or tied-best results across all four evaluations, outperforming the full seven-operation mixture. Broader operation coverage alone is therefore not enough for better transfer. These results favor selecting object addition and replacement, which expose local semantic changes together with full-image pixel traces.

Generator-family composition. SD 1.5 and FLUX have complementary transfer profiles: the former is stronger on Established, whereas the latter leads elsewhere. Combining them improves three of the four evaluations, at the cost of some FLUX-only Pixel accuracy. Mixing generator families thus broadens coverage, though the gains are not uniform. This SD+FLUX coverage is consistent with the SD/FLUX expert extension in AlignGemini (Chen et al., 2025a) and the training sources of Deep-VRM (Lin et al., 2026).

4.4 Qualitative Analysis of Editing Evidence

In Figure 5, both *local* and *outside* receive higher fake probabilities than the sources. The strong *outside* predictions, together with the distributed Grad-CAM responses and pixel differences, suggest that the detector uses evidence beyond the intended edit. These examples support training on complete outputs, which preserve both localized changes and broader editing traces.

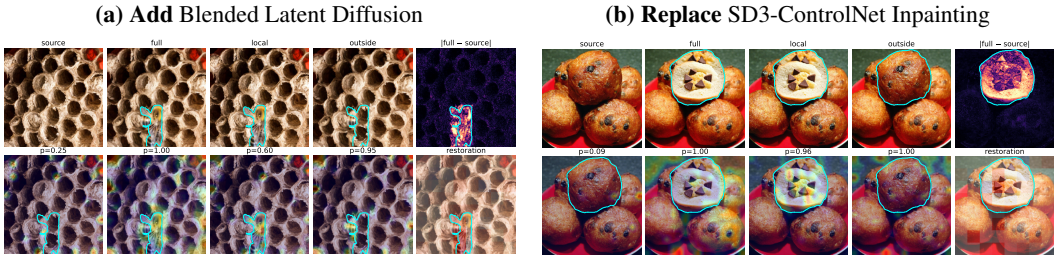


Figure 5: **Cross-domain Grad-CAM and difference analysis on Mani-X.** After training on Add Text and Replace Text, we test mask-guided Add and Replace samples. Columns 1–4 show the source, complete edited output (*full*), and composites retaining edited pixels only inside (*local*) or outside (*outside*) the mask, with source pixels elsewhere. Top: input images; bottom: Grad-CAM heatmaps with fake probability p.

5 Conclusion

Our study links two shortcuts in AI-generated image detection: a computational shortcut, where linear probes recover much of the performance of trained MLLM detectors, and dataset shortcuts, where metadata and resolution cues can inflate accuracy. The first motivates adapting a visual encoder directly for detection; the second shows why high accuracy alone does not establish generalization. We therefore pair real source photographs with edited outputs through text-guided image editing. Source pairing and consistent preprocessing control content and processing biases, while the edits inject pixel-level generation and reconstruction traces alongside semantic-level forgery cues. Keeping the complete outputs lets a single visual encoder learn both forms of evidence in one representation. Evaluations on full-image synthesis, local manipulation, and real-world processing support this design as a practical route to robust detection that holds up under compression, transmission, and re-digitization.

Limitations. Our evaluation covers a broad range of existing generators, but a strictly controlled quantitative comparison of transfer across generator families remains for future work. Resolution-alignment experiments reveal the studied MLLM detectors' sensitivity to input geometry. A small part of this sensitivity may stem from the loss of original image information during processing, and the present interventions do not separately quantify the contributions of biased learning and information loss.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923, 2025. URL https://arxiv.org/abs/2502.13923.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. arXiv preprint arXiv:2406.09398, 2024. URL https://arxiv.org/abs/2406.09398.

Ruoxin Chen, Jiahui Gao, Kaiqing Lin, Keyue Zhang, Yandan Zhao, Isabel Guan, Taiping Yao, and Shouhong Ding. AlignGemini: Generalizable ai-generated image detection through task-model alignment, 2025a.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes ai-generated image detector easier generalizable. In *NeurIPS*, 2025b.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai. In *CVPR*, 2025.

Riccardo Corvi, Davide Cozzolino, Ekta Prashnani, Shalini De Mello, Koki Nagano, and Luisa Verdoliva. Seeing what matters: Generalizable ai-generated video detection with forensic-oriented augmentation. In *NeurIPS*, 2025.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *CVPRW*, 2023.

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

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

Yuncheng Guo, Junyan Ye, Chenjue Zhang, Hengrui Kang, Haohuan Fu, Conghui He, and Weijia Li. Omniaid: Decoupling semantics and artifacts for universal ai-generated image detection in the wild. In *ICML*, 2026.

Xin Jiang, Hao Tang, Junyao Gao, Zijie Yang, Meiqi Cao, Fei Shen, Dongming Zhang, Zechao Li, and Yongdong Zhang. Dailybench: A unified benchmark for ai-generated and manipulated images from modern generative models. arXiv preprint arXiv:2607.24016, 2026. URL https://arxiv.org/abs/2607.24016.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution ai-generated image detection by spectral learning. In *CVPR*, 2025.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *ECCV*, 2024.

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 *ICCV*, 2025a.

Yifei Li, Haoyuan He, Yu Zheng, Bingyao Yu, Wenzhao Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. Next-imdl: Build benchmark for next-generation image manipulation detection & localization, 2025b.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? In *NeurIPS*, 2025c.

Kaiqing Lin, Zhiyuan Yan, Ruoxin Chen, Ke-Yue Zhang, Yue Zhou, Caiyong Piao, Bin Li, Taiping Yao, Bo Wang, Youchang Xiao, and Shouhong Ding. Deep residual injection for full-spectrum forensic signal perception in multimodal large language models. In *ICML*, 2026.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *CVPR*, pp. 10770–10780, 2024.

Ruiqi Liu, Manni Cui, Ziheng Qin, Zhiyuan Yan, Ruoxin Chen, Yi Han, Zhiheng Li, Junkai Chen, ZhiJin Chen, Kaiqing Lin, Jialiang Shen, Lubin Weng, Jing Dong, Yan Wang, and Shu Wu. Mirror: Manifold ideal reference reconstructor for generalizable ai-generated image detection. arXiv preprint arXiv:2602.02222, 2026. URL https://arxiv.org/abs/2602.02222.

Elif Nebioglu, Emirhan Bilgiç, and Adrian Popescu. Ai-generated image detectors overrely on global artifacts: Evidence from inpainting exchange. arXiv preprint arXiv:2602.00192, 2026. URL https://arxiv.org/abs/2602.00192.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, pp. 24480–24489, 2023.

Ziheng Qin, Yuheng Ji, Renshuai Tao, Yuxuan Tian, Yuyang Liu, Yipu Wang, and Xiaolong Zheng. Scaling up ai-generated image detection with generator-aware prototypes. In *CVPR*, 2026.

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

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025.

Vladan Stojnić, Ryan Ramos, Giorgos Kordopatis-Zilos, Noa Garcia, and Giorgos Tolias. Invisible shortcuts: Why vision encoders know your camera. In *ECCV*, 2026.

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 *CVPR*, pp. 28130–28139, 2024.

Hao Tan, Jun Lan, Senyuan Shi, Zichang Tan, Zijian Yu, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Videoveritas: Ai-generated video detection via perception pretext reinforcement learning, 2026a.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Zijian Yu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas++: Value-aware on-policy distillation for perception-enhanced aigi detection. arXiv preprint arXiv:2607.27113, 2026b. URL https://arxiv.org/abs/2607.27113.

Youqi Wang, Shen Chen, Haowei Wang, Rongxuan Peng, Taiping Yao, Shunquan Tan, Changsheng Chen, Bin Li, and Shouhong Ding. Forgeryvcr: Visual-centric reasoning via efficient forensic tools in mllms for image forgery detection and localization. In *ACM MM*, 2026.

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

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 *ICML*, 2025b.

Yanran Zhang, Bingyao Yu, Yu Zheng, Wenzhao Zheng, Yueqi Duan, Lei Chen, Jie Zhou, and Jiwen Lu. $\bf{d^3}$qe: Learning discrete distribution discrepancy-aware quantization error for autoregressive-generated image detection. In *ICCV*, 2025.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *ICCV*, 2025.

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

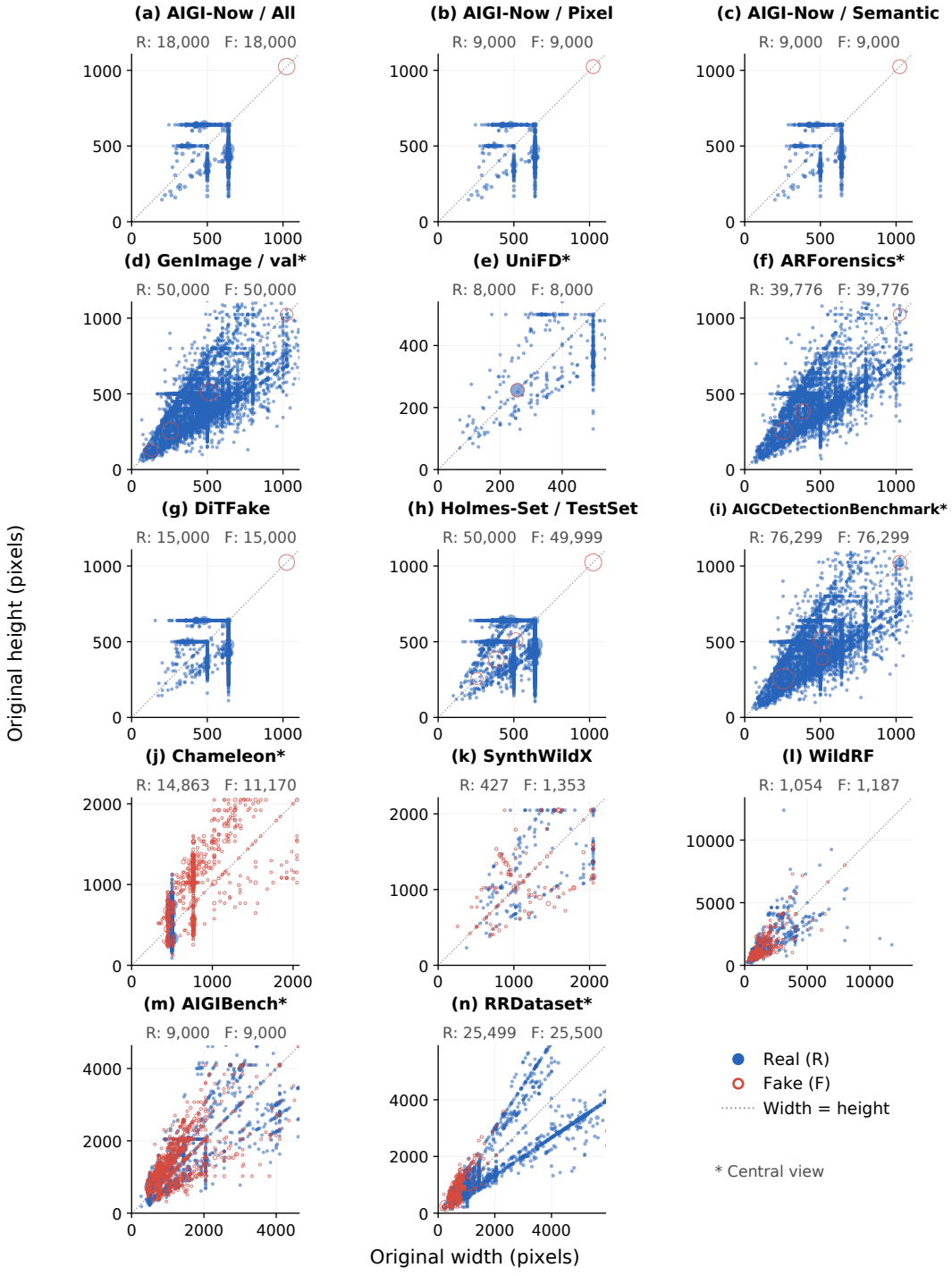


Figure 6: **Resolution biases across datasets.** Real and fake images often exhibit distinct resolution and aspect-ratio distributions, with fake images concentrated at a few fixed resolutions in several datasets. These label-correlated patterns reveal potential shortcuts unrelated to generation artifacts, highlighting the need to control image geometry during training and evaluation.

A MANI-X RESOURCE AND TRAINING OPERATIONS

Editors. Table 8 lists the editing methods and supported base models in Add Text and Replace Text, following the public implementations of FireFlow, RF-Solver-Edit, InstructDiffusion, Paint-by-Inpaint, HIVE, and FlowEdit. These methods are built upon two major open-source model families, Stable Diffusion (SD) and FLUX.1-dev, and cover the two editing operations considered in this work, i.e., adding and replacing content. Rather than relying on user-provided local masks, they adopt a full-image mask and control the editing scope solely through language instructions. Compared with local-mask editing, this paradigm injects pixel-level and semantic-level forgery cues more uniformly across the image.

Table 8: **Mani-X editors and supported base models.** These training editors are based on two major backbone families: FLUX.1 and SD1.5.

Editor	Add Text	Replace Text	Base-model family
FireFlow	14,276	9,502	FLUX.1-dev
RF-Solver-Edit	14,236	9,398	FLUX.1-dev
Paint-by-Inpaint	14,318	–	SD 1.5
InstructDiffusion	–	9,431	SD 1.5
HIVE	14,305	9,472	SD 1.5
FlowEdit	–	9,475	FLUX.1-dev
Total	57,135	47,278	

A.1 INDIVIDUAL TRAINING SUBSETS

Table 9 compares the individual Mani-X training subsets. Subsets with *add* and *replace* edits tend to yield higher full-image classification accuracy: built on aligned source pairs, these edits inject salient semantic anomalies, which give the detector clear semantic evidence to learn from. Among all subsets, *Add Text* and *Replace Text* achieve the best generalization, which motivates our choice of them as the main configuration.

Table 9: **Individual Mani-X training subsets: ACC (%).** AVG equally weights Established, In-the-wild, and AIGI-Now (the mean of Pixel and Semantic). Bold marks the best result in each column.

Subset	Samples (k)	Established	In-the-wild	AIGI-Now		AVG
				Pixel	Semantic	
Add	49	84.9	88.0	63.7	55.3	77.5
Add Text	81	**92.6**	90.1	**89.1**	86.9	90.2
Replace	171	86.8	83.2	54.4	62.1	76.1
Replace Text	66	91.5	**91.3**	88.0	**90.5**	**90.7**
Null	122	88.1	83.1	53.3	72.5	78.0
Removal	143	85.2	77.8	64.8	51.7	73.8
Removal Text	72	81.9	62.4	50.0	48.0	64.4

B SUPPLEMENTARY RESULTS AND CONFIGURATION DETAILS

Comparison methods. We compare pixel and visual-feature detectors NPR, FatFormer, RINE, Effort, CO-SPY, SPAI, GAPL, WaveRep, and MIRROR (Tan et al., 2024; Liu et al., 2024; Koutlis & Papadopoulos, 2024; Yan et al., 2025b; Cheng et al., 2025; Karageorgiou et al., 2025; Qin et al., 2026; Corvi et al., 2025; Liu et al., 2026); aligned-data methods DDA and B-Free (Chen et al., 2025b; Guillaro et al., 2025); and MLLM detectors Deep-VRM and Veritas++ (Lin et al., 2026; Tan et al., 2026b).

Training-data ablation protocol. Table 7 compares data recipes using a common backbone, input resolution, training-set size, and update count. DDA and B-Free denote training recipes under this

shared setup. Mani-X (ours), Add Text + Replace Text, and SD 1.5 + FLUX denote the same configuration across the three blocks. "All Mani-X subsets" uses all seven operations.

B.1 Training Data across Visual Backbones and Scales

Table 10 extends the training-data comparison to CLIP-L and DINOv2-L and evaluates larger DINOv3 backbones and inputs.

Table 10: **Training data, visual backbone, and input resolution: ACC (%).** Estab. denotes established benchmarks; Wild denotes in-the-wild evaluation. AVG is the mean of the four evaluation columns. Bold marks the best score within each block.

Backbone	Input	Training data	Estab.	Wild	AIGI-Now		AVG
					Pixel	Semantic	
CLIP-L	224p	Pixel (DDA)	**88.7**	59.8	70.5	50.0	67.3
CLIP-L	224p	Semantic (B-Free)	84.6	61.1	57.2	57.9	65.2
CLIP-L	224p	Mani-X (ours)	84.2	**85.5**	**89.8**	**84.2**	**85.9**
DINOv2-L	224p	Pixel (DDA)	**91.4**	68.3	59.6	54.5	68.5
DINOv2-L	224p	Semantic (B-Free)	89.8	65.4	82.2	47.0	71.1
DINOv2-L	224p	Mani-X (ours)	88.5	**84.8**	**90.0**	**87.4**	**87.7**
DINOv3-L	256p	Mani-X (ours)	94.8	91.3	92.1	92.8	92.8
DINOv3-H+	256p	Mani-X (ours)	94.4	93.9	91.5	95.6	93.8
DINOv3-7B	256p	Mani-X (ours)	94.9	93.0	**95.2**	**96.8**	**95.0**
DINOv3-L	512p	Mani-X (ours)	**96.0**	92.0	94.7	91.8	93.6
DINOv3-H+	512p	Mani-X (ours)	95.7	**95.8**	92.9	90.8	93.8

Established accuracy does not guarantee broad transfer. Pixel-aligned training with DDA performs well on established benchmarks, yet this strength does not consistently carry over to in-the-wild images or AIGI-Now. Training on our source-matched image-editing pairs changes this generalization profile: the detector retains strong established benchmark performance while becoming substantially more effective on both real-world imagery and the Pixel and Semantic subsets of AIGI-Now. This is a broader balance of detection capabilities, rather than an improvement confined to one evaluation setting. Retention does not imply identical accuracy: CLIP-L and DINOv2-L incur modest established-benchmark reductions relative to DDA, whereas DINOv3-L nearly matches DDA in the shared-configuration comparison (Table 7).

The data effect extends across encoder families. Within both CLIP-L and DINOv2-L, Mani-X outperforms the DDA and B-Free recipes on in-the-wild evaluation and both AIGI-Now subsets. Together with the DINOv3 comparison in Table 7, this consistency indicates that the benefit is not specific to DINOv3: source-matched editing data improves transfer for both a CLIP visual encoder and different generations of DINO encoders. The repeated within-backbone pattern supports the training-data design as a contributor to generalization, beyond the choice of pretrained representation.

Scaling visual capacity offers further gains. With Mani-X fixed, increasing DINOv3 capacity at 256p improves aggregate performance from L through H+ to 7B. Larger inputs provide another route to improvement: 512p benefits L overall and strengthens established and in-the-wild detection for H+, although H+'s Semantic decline leaves its aggregate unchanged. Thus, visual scaling can improve performance, but its benefits depend on the evaluation regime; higher resolution is not a uniform substitute for a larger backbone. Together with the direct visual-readout results (Table 1), these findings suggest that scaling visual capacity may offer clearer detection gains than enlarging an MLLM's language head. This remains a hypothesis about where to allocate capacity, since these experiments do not include a controlled language-head size sweep or a compute-matched scaling comparison.

C DailyBench Results and Aggregation

Tables 11 and 12 report per-generator results on DailyBench's synthesis and manipulation tasks (Jiang et al., 2026). The 17 baseline results are from Tables 3 and 13 of the benchmark paper under its official-checkpoint protocol; FPD is the benchmark authors' baseline. Model size and training data vary across methods. Each of our DINOv3 configurations is evaluated on both tasks.

Means weight the seven generator/operation subsets equally. ACC is the fraction of correct predictions; bACC is $(\mathrm{TPR} + \mathrm{TNR})/2$. FakeBench is class-balanced within each generator. Manipulation subsets have unequal class proportions, so Table 6 also reports bACC.

Table 11: **DailyBench FakeBench: per-generator ACC (%).**

Model	Stable Diffusion 3.5	FLUX.1	FLUX.2	Z-Image	Qwen-Image-2512	Nano Banana 2	GPT-Image 2	Mean
UniFD	50.8	49.5	49.7	49.9	49.6	49.8	49.3	49.8
LOTA	50.2	49.8	48.5	49.7	48.6	49.4	52.2	49.8
DRCT	57.1	53.4	56.2	63.5	57.7	52.7	54.4	56.4
FatFormer	50.9	49.8	49.9	50.3	50.0	50.0	55.8	51.0
NPR	62.3	70.3	63.3	67.8	70.9	61.4	76.6	67.5
SAFE	48.7	49.1	48.8	48.3	49.4	51.2	86.0	54.5
AIDE	53.6	57.4	53.0	58.1	58.4	50.8	55.1	55.2
Effort	54.4	57.0	53.7	61.0	63.9	62.7	68.8	60.2
ForgeLens	50.1	50.3	62.1	50.1	53.6	50.0	99.5	59.4
LTD	51.2	50.6	51.1	52.4	52.3	50.2	84.5	56.0
Aligned	74.6	58.0	51.4	56.0	60.6	71.0	54.9	60.9
B-Free	72.5	48.7	47.7	63.2	46.1	45.9	44.9	52.7
PROBE	92.2	53.5	58.6	79.6	82.8	66.1	61.5	70.6
OmniAID	69.0	62.9	52.7	63.8	67.1	72.4	59.1	63.8
DDA	81.3	58.4	54.8	74.6	71.5	65.7	67.7	67.7
DINO-Det	84.0	90.4	72.9	75.6	76.6	62.0	74.1	76.5
FPD	86.3	91.1	80.0	83.2	83.3	60.8	74.0	79.8
DINOv3-L 256p	85.3	93.2	85.0	89.1	82.0	73.1	70.6	82.6
DINOv3-L 512p	88.5	92.4	84.7	90.8	87.6	75.3	75.7	85.0
DINOv3-H+ 256p	91.5	94.7	87.8	91.0	85.6	71.8	68.4	84.4
DINOv3-H+ 512p	95.5	97.1	86.7	95.0	89.1	69.8	66.4	85.7

Table 12: **DailyBench ManipulationBench: per-generator ACC (%).**

Model	FLUX-Fill Random masks	FLUX-Fill Object masks	FLUX.2-klein-9B	Qwen-Image-Edit-2511	Step1X-Edit-v1p2	Nano Banana 2	GPT-Image 2	Mean
UniFD	50.2	39.0	38.4	39.6	39.1	37.6	37.5	40.2
LOTA	49.9	60.4	60.7	59.5	60.5	57.1	63.7	58.8
DRCT	51.0	42.5	41.6	44.8	41.3	49.9	57.8	47.0
FatFormer	49.9	37.9	37.6	39.1	38.4	37.3	45.9	40.9
NPR	45.1	41.0	45.0	52.4	44.0	52.7	81.9	51.7
SAFE	52.2	41.6	39.0	39.0	37.4	38.0	84.9	47.5
AIDE	51.1	39.9	38.2	50.5	48.3	43.5	51.7	46.2
Effort	45.6	46.8	52.6	60.0	56.1	64.2	75.9	57.3
ForgeLens	50.0	37.9	37.5	41.3	39.4	41.4	91.8	48.5
LTD	49.9	37.8	37.4	39.0	38.0	37.5	66.6	43.7
Aligned	50.5	38.5	37.9	67.3	38.5	55.2	45.9	47.7
B-Free	62.4	57.7	57.1	77.9	58.6	34.7	42.5	55.8
PROBE	61.7	56.6	59.6	92.1	55.1	65.5	80.3	67.3
OmniAID	52.7	44.3	47.8	88.1	54.6	67.3	56.6	58.8
DDA	54.5	46.5	55.5	89.4	56.8	56.6	80.2	62.8
DINO-Det	52.2	42.4	43.4	63.0	45.7	40.4	65.1	50.3
FPD	53.7	46.0	49.9	79.0	54.2	46.3	82.4	58.8
DINOv3-L 256p	75.4	68.6	56.9	75.5	56.8	49.6	58.7	63.1
DINOv3-L 512p	88.2	86.1	65.8	85.5	53.0	49.9	72.5	71.6
DINOv3-H+ 256p	81.6	76.5	60.7	82.7	58.3	50.7	59.3	67.1
DINOv3-H+ 512p	89.9	85.1	66.8	89.9	50.9	47.1	72.2	71.7

The two FLUX-Fill columns in Table 12 correspond to random and object masks, respectively.

D ESTABLISHED BENCHMARKS: PER-GENERATOR RESULTS

Tables 13–16 give per-generator results under Original and JPEG conditions. AVG is the dataset mean used in Table 5. Bold and underlining mark the best and second-best distinct scores in each column and condition, including ties; ranking precedes rounding.

Table 13: **GenImage: per-generator ACC (%).**

Method	Mid-journey	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	AVG
Original									
NPR	77.9	78.6	78.9	69.7	78.4	76.1	78.1	80.1	77.2
Effort	76.2	87.8	87.6	70.9	80.8	86.8	82.7	75.5	81.0
CO-SPY	69.9	93.0	92.9	58.7	81.4	91.0	72.1	64.8	78.0
DDA	96.3	98.0	98.2	88.4	86.5	97.8	63.1	76.4	88.1
B-Free	94.9	98.8	98.9	78.1	83.8	98.8	88.3	68.6	88.8
Deep-VRM	**99.3**	**99.3**	99.1	**89.6**	92.5	98.8	92.4	82.1	94.1
Veritas++	91.8	96.2	95.9	68.9	75.3	93.8	72.6	68.5	82.9
Ours	94.6	**99.3**	**99.4**	77.9	**98.5**	**99.4**	**97.5**	**98.3**	**95.6**
JPEG									
NPR	50.0	52.4	52.3	34.1	33.1	52.5	37.0	76.8	48.5
Effort	74.6	87.5	87.4	51.1	67.1	86.2	59.7	38.3	69.0
CO-SPY	69.6	92.3	92.0	58.3	81.4	90.1	72.3	62.9	77.4
DDA	94.8	97.9	97.9	86.2	86.5	97.9	63.5	84.7	88.7
B-Free	95.2	98.8	98.8	79.3	86.8	98.8	89.7	70.4	89.7
Deep-VRM	**95.8**	96.6	96.5	62.0	66.6	89.3	58.1	57.2	77.8
Veritas++	92.7	95.3	95.6	60.2	68.5	92.3	63.2	60.5	78.5
Ours	94.8	**99.4**	**99.4**	**86.7**	**98.4**	**99.4**	**98.9**	**98.8**	**97.0**

Table 14: **UniFD: per-generator ACC (%).**

Method	Guided	LDM200	LDM200 -CFG	LDM100	GLIDE 100-27	GLIDE 50-27	GLIDE 100-10	DALLE	AVG
Original									
NPR	75.1	**98.5**	**98.3**	**98.7**	**97.1**	**97.6**	**97.6**	93.9	**94.6**
Effort	78.0	96.1	91.8	96.7	93.2	91.7	92.2	**96.8**	92.0
CO-SPY	62.6	83.2	85.3	82.8	73.6	69.2	76.6	81.8	76.9
DDA	87.2	76.6	77.7	76.0	70.6	70.3	74.7	56.1	73.6
B-Free	81.9	93.8	94.0	93.9	84.3	85.8	87.3	91.7	89.1
Deep-VRM	**94.1**	84.6	84.4	84.9	82.5	82.1	84.5	84.3	85.2
Veritas++	71.7	78.6	66.6	79.8	78.3	76.7	78.3	77.6	76.0
Ours	76.2	91.0	90.9	91.0	90.6	90.8	90.8	90.7	89.0
JPEG									
NPR	34.9	51.3	50.7	51.0	50.3	50.3	50.4	52.0	48.8
Effort	49.3	77.0	69.7	77.2	80.6	78.0	79.7	74.6	73.2
CO-SPY	62.5	82.6	84.6	81.9	73.6	69.4	75.8	80.5	76.3
DDA	**90.3**	84.2	84.5	84.2	78.1	76.8	80.1	75.1	81.6
B-Free	83.2	**93.8**	**94.0**	**93.9**	86.0	87.3	88.5	**92.4**	89.9
Deep-VRM	63.7	68.8	63.1	70.6	68.9	68.6	68.6	54.3	65.8
Veritas++	60.7	68.9	61.8	69.8	71.8	73.5	74.1	69.6	68.8
Ours	89.3	91.0	91.0	91.0	**90.5**	**90.7**	**90.7**	90.9	**90.6**

Table 15: **ARForensics: per-generator ACC (%).**

Method	Infinity	Janus Pro	LlamaGen	MAR	Open MAGVIT2	RAR	VAR	AVG
Original								
NPR	80.5	78.6	73.7	74.5	70.5	79.7	80.2	76.8
Effort	86.9	87.7	79.3	77.3	73.3	85.9	81.6	81.7
CO-SPY	93.1	80.0	60.4	66.2	59.2	65.1	63.2	69.6
DDA	97.6	98.0	90.9	86.9	59.1	81.2	69.6	83.3
B-Free	97.1	92.0	95.3	97.9	83.7	90.6	87.1	91.9
Deep-VRM	**99.4**	99.3	98.3	87.1	85.9	94.5	84.2	92.7
Veritas++	**99.4**	97.9	77.5	65.8	62.3	69.0	59.6	76.0
Ours	99.3	**99.4**	**98.4**	**98.8**	**90.5**	**98.9**	**97.7**	**97.6**
JPEG								
NPR	51.1	41.4	42.7	36.0	35.4	35.8	36.1	39.8
Effort	86.8	83.0	70.0	51.2	49.6	71.5	52.0	66.3
CO-SPY	92.8	74.0	57.6	66.2	59.7	63.9	64.4	68.4
DDA	97.6	98.7	95.2	96.5	78.5	91.2	80.3	91.1
B-Free	96.9	92.3	95.5	98.2	86.4	91.9	90.3	93.1
Deep-VRM	98.7	88.8	60.3	55.2	56.6	55.1	55.2	67.2
Veritas++	**99.3**	95.6	69.2	59.3	54.5	57.6	53.2	69.8
Ours	99.2	**99.4**	**99.0**	**99.4**	**94.2**	**99.3**	**98.3**	**98.4**

Table 16: **HolmesSet: per-generator ACC (%).**

Method	FLUX	Infinity	Janus	Janus-Pro 1B	Janus-Pro 7B	LlamaGen	PixArt-XL	SD35-L	Show-o	VAR	AVG
Original											
NPR	79.1	79.8	30.1	29.8	29.8	79.5	78.7	79.4	79.6	79.3	64.5
Effort	81.9	82.6	50.1	66.7	59.0	82.6	82.6	79.2	81.5	78.5	74.5
CO-SPY	91.0	96.5	56.9	85.3	77.3	90.8	96.0	84.6	94.3	64.9	83.8
DDA	98.5	99.0	97.3	98.8	98.2	99.2	99.2	98.0	87.0	72.9	94.8
B-Free	88.5	99.2	98.5	99.2	98.5	99.4	99.6	96.7	99.2	90.4	96.9
Deep-VRM	**99.9**	**99.9**	59.1	86.7	88.5	**99.9**	**99.9**	**99.7**	**99.8**	87.4	92.1
Veritas++	98.7	99.8	79.0	97.8	98.0	99.6	99.6	94.6	99.0	62.0	92.8
Ours	97.7	99.7	**99.6**	**99.6**	**99.6**	99.6	99.6	76.6	99.7	**98.2**	**97.0**
JPEG											
NPR	58.6	52.8	30.1	29.8	29.8	32.8	64.8	51.7	33.9	31.8	41.6
Effort	81.7	82.6	50.1	66.7	59.0	82.6	82.5	77.9	79.1	47.8	71.0
CO-SPY	90.3	96.0	56.9	85.3	77.3	87.5	95.3	83.2	93.3	66.2	83.1
DDA	97.3	98.9	97.3	98.8	98.2	99.1	99.1	92.7	82.3	67.6	93.1
B-Free	87.8	99.2	98.5	99.2	98.5	99.5	99.6	96.2	99.3	92.9	**97.1**
Deep-VRM	**99.2**	99.3	62.8	89.8	92.0	97.8	99.4	**98.0**	97.9	56.1	89.2
Veritas++	98.8	**99.7**	79.9	97.9	97.9	**99.7**	**99.7**	97.0	99.0	54.5	92.4
Ours	97.8	99.6	**99.6**	**99.6**	**99.6**	99.6	99.6	73.0	**99.7**	**99.1**	96.7

E AI USE STATEMENT

We used large language models (LLMs) to assist with organizing the paper's narrative and logical flow, polishing the writing, and preparing figures and tables. LLMs were not involved in preparing the raw experimental data or developing the core research ideas. The authors take full responsibility for the final content of this work, including all text, claims, figures, and tables prepared with LLM assistance.