How Do MLLMs Detect AI-Generated Facial Images? A Study of Cross-Source Detection, Provenance, and Explanations

Anonymous authors
Paper under double-blind review

Abstract

Multimodal large language models (MLLMs) are increasingly used to detect AI-generated facial images and provide natural-language explanations. However, it remains unclear which signals influence these judgments and whether the visual evidence reported is reliable across repeated judgments. In this paper, we construct a 4,000-image benchmark covering AI-generated facial images from seven generation sources and real photographs, then evaluate five leading MLLMs in a cross-source setting. We first use full cross-source evaluation and fixed-effects analysis to examine whether detectors more readily identify images produced by models from the same developer. Provenance effects are further investigated by removing C2PA records and applying MarkNull for SynthID removal, while measuring the pixel-level changes. Finally, we assess judgment repeatability and the image support of model explanations. Across the evaluated systems, we find no general detection advantage for same-developer pairs, although effects vary across model combinations. Our provenance interventions further show that C2PA removal does not significantly change overall judgments, whereas SynthID removal produces detector- and source-dependent changes in detection behaviors. Notably, Gemini's detection rate decreases for Gemini Image outputs but increases for GPT Image 2 outputs. For explanations from correct detections, the corresponding images support most reviewed observations, although classification outcomes vary across repeated judgments. Our findings provide a clearer basis for evaluating cross-source MLLM detection and interpreting the visual evidence reported in model explanations.

1 Introduction

Text-guided image generation has made image synthesis more accessible, increasing the need to distinguish AI-generated images from real images (Yan et al., 2025). This need is particularly pronounced for facial images, whose close connection to personal identity makes reliable authenticity assessment especially important. Recent studies have used multimodal large language models (MLLMs) to assess image authenticity and provide explanations alongside their classification decisions (Li et al., 2025; Guo et al., 2025). These explanations provide additional information beyond the classification results and offer a way to examine how models justify their authenticity judgments.

Prior work has begun to evaluate these explanations by examining whether the reported visual cues correspond to observable image evidence (Fei et al., 2026; Jing et al., 2024). Related evaluations show that models can identify some localized visual details while detecting generated images. This extends detection evaluation beyond classification accuracy to the factual support of model explanations (Wen et al., 2026; Ji et al., 2026). These advances provide a basis for understanding model judgments, but they also raise the question of whether such performance remains consistent across generation sources.

Differences across generation sources are reflected not only in visible image content but also in associated signals from the generation process. As Figure 1 illustrates, AI-generated facial images may contain provenance signals, such as C2PA credentials that record creation history and invisible SynthID watermarks (C2PA, n.d.; Gowal et al., 2025). As different services use different prove-

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

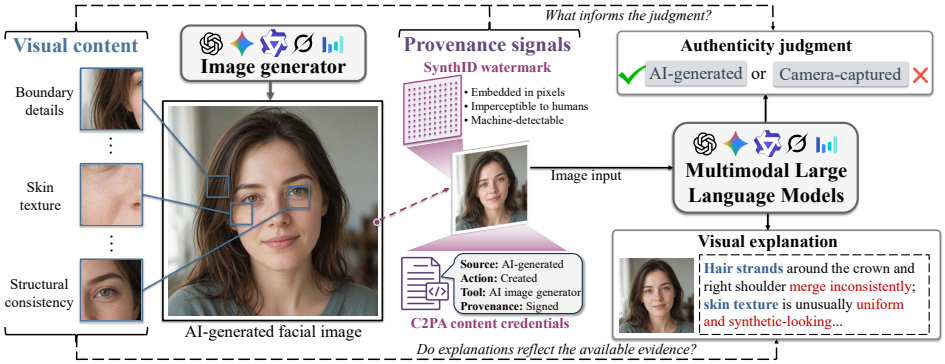

Figure 1: Multiple cues in AI-generated facial-image detection. An image may contain observable visual details, invisible watermarks, and file-level provenance credentials. MLLMs produce authenticity judgments and textual explanations, but how these information sources influence judgments and whether the reported observations are supported by image evidence remain to be examined.

nance mechanisms, cross-source detection may depend on both image content and source-specific signals. More importantly, many developers provide both image-generation and multimodal detection models. Whether detectors are more likely to identify images generated by models from the same developer remains unclear in cross-source evaluation.

Determining whether same-developer pairs show a detection advantage requires separating the developer relationship from detector capability and source difficulty. A high detection rate for a same-developer pair may reflect a stronger detector or an easier-to-detect source, so it does not by itself establish a developer-specific advantage (Rajan et al., 2025; Roytburg et al., 2026). We evaluate every detector on every generation source and compare the balanced observed contrast with fixed-effects estimates of the same-developer association. MLLMs may also use provenance markers, leaving it unclear whether their judgments rely on visual features or external signals. Moreover, natural-language explanations may sound plausible without image support, making them insufficient as direct evidence of the cues driving authenticity judgments (Jing et al., 2024). These challenges motivate a unified evaluation framework that examines detection behavior, provenance effects, and explanation reliability in MLLMs. We organize our study around the following research questions:
RQ1: *Do MLLMs show a detection advantage when the generator and detector come from the same developer?*
RQ2: *How do provenance interventions affect model judgments of AI-generated facial images?*
RQ3: *How repeatable are model judgments, and are the observations in their explanations supported by the input images?*

To answer these questions, we evaluate five MLLMs from the GPT, Gemini, Qwen, Seed, and Grok families. Using shared prompts, we construct a benchmark of facial images from seven generation sources, including sources provided by the developers of the MLLMs and two additional sources from FLUX and HiDream (Table 1), together with real photographs for measuring false positives. Each MLLM evaluates all seven sources, and the analysis compares observed rates with fixed-effects estimates across the five matched sources. We then perform provenance interventions by removing C2PA records and applying the MarkNull pipeline for SynthID removal (Cao et al., 2026), while repeated judgments measure changes in model decisions across these conditions. Finally, we measure classification agreement across repeated judgments and conduct a human image-evidence evaluation of explanations for correct detections. Our main contributions include:

- We construct a 4,000-image benchmark spanning AI-generated facial images from seven generation sources. The benchmark also provides a unified resource for evaluating MLLM-based detection behavior across generation sources.
- Using full cross-source evaluation and fixed-effects analysis, we find no general detection advantage when the generator and detector come from the same developer. After controlling for detector capability and generation-source difficulty, some same-developer pairs even show lower adjusted detection probabilities.

- We show that provenance changes do not have a simple directional effect on MLLM detection judgments. C2PA removal produces no statistically significant overall change after Holm correction, whereas the SynthID-removal procedure changes detection rates for some generation sources, with the direction varying across detectors.
- We analyze explanations for the same images and conduct human evaluation of their visual evidence support. Models often change the visual cues they emphasize across repeated judgments, while most observations in explanations for correct detections are supported by the images.

2 Related Work

Provenance and Developer Relationships. AI-generated image provenance is commonly represented through file-level credentials and embedded watermarks. C2PA uses digital signatures to record provenance and editing history (C2PA, n.d.), whereas SynthID embeds invisible watermarks in image pixels (Gowal et al., 2025). Prior work has examined these signals from both human-perception and technical perspectives. User studies showed that provenance information and its integrity can affect judgments of media trustworthiness and accuracy (Feng et al., 2023), while watermark evaluations examined whether embedded signals survive processing and how processing affects image quality (An et al., 2024). However, whether general-purpose MLLMs are influenced by provenance information when judging image authenticity remains underexplored. Beyond provenance signals, the relationship between image generators and detectors may also affect cross-model evaluation. Studies in text evaluation have shown that LLMs can exhibit preferences toward their own outputs and that evaluator capability affects the measurement of such effects (Panickssery et al., 2024; Roytburg et al., 2026; Ji et al., 2025). Therefore, further investigation is needed to determine whether similar same-developer effects arise in MLLM-based image detection. Such evaluations require separating detector capability, generation-source difficulty, and same-developer effects while considering provenance-related changes in model judgments.

Detection and Explanations. A central challenge in AI-generated image detection is cross-source generalization to generators unseen during training. Prior work has explored cross-source detection by leveraging pretrained vision-language representations to improve generalization across generative models (Ojha et al., 2023). As MLLMs are increasingly used for authenticity judgments, recent evaluations also considered whether they can identify visual cues and explain their decisions. Related benchmarks jointly evaluated classification accuracy and artifact-based explanations, while other approaches use explanation supervision or human preferences to improve detection outputs (Ye et al., 2025; Zhou et al., 2025). Region annotations and textual descriptions further connect reported explanation cues to specific image regions (Ji et al., 2026). Meanwhile, vision-language factuality studies evaluate whether generated descriptions are grounded in visual evidence and whether specific claims are supported by images (Li et al., 2023; Jing et al., 2024). Together, these studies highlight the importance of evaluating both detection behavior across generation sources and the reliability of explanation cues. In this paper, we extend these directions by conducting a cross-model evaluation with multiple generators and detectors, while examining the consistency of reported cues across repeated judgments and their support from image evidence for correct detections.

3 Dataset and Evaluation Protocol

We define a unified evaluation protocol for detecting AI-generated facial images using real photographs and AI-generated images from seven generation sources. The protocol covers cross-source detection, provenance interventions, repeated judgments, and image-evidence evaluation.

3.1 Dataset and Model Selection

Real image selection. We select 500 real photographs from Flickr-Faces-HQ (FFHQ) (Karras et al., 2019), stratified by age, gender, and apparent skin tone. Five age groups (20s, 30s, 40s, 50s, and 60 or older), two gender-presentation groups, and five skin-tone groups form 50 attribute combinations. We select ten photographs per combination. Each photograph shows a single adult with a frontal face and meets basic clarity and exposure requirements. Age and gender presentation are estimated from the images using fixed models. Selection criteria are detailed in Appendix B.1. Apparent skin

Table 1: Models, generator–detector pairings, and provenance information used in the evaluation. The first five rows define same-developer pairs; FLUX and HiDream provide additional open-weight generation sources without paired detectors. The `provenance` role identifies sources included in the C2PA/SynthID processing analysis.

Generator	Detector	Provenance	Study Role
GPT Image 2 (OpenAI, 2026c)	GPT-5.6 Sol (OpenAI, 2026b)	C2PA; SynthID (OpenAI, 2026a)	Detection; provenance
Gemini 3.1 Flash Image (Google DeepMind, 2026a)	Gemini 3.8 Flash (Google DeepMind, 2026b)	C2PA; SynthID (Google, 2026)	Detection; provenance
Qwen-Image-3.0-Pro (Alibaba Cloud, 2026b)	Qwen3.8-Max (Alibaba Cloud, 2026a)	Optional visible watermark	Detection
Seedream 5.0 Pro (ByteDance Seed, 2026b)	Seed2.1 Turbo (ByteDance Seed, 2026a)	Optional visible watermark	Detection
Grok Imagine Image 2.0 (SpaceXAI, 2026b)	Grok 4.6 (SpaceXAI, 2026a)	Provenance unspecified	Detection
FLUX.2 [dev] (Black Forest Labs, 2025)	—	Not assessed for local outputs	Detection
HiDream-O1-Image-Dev-2604 (Cai et al., 2026)	—	Not assessed for local outputs	Detection

tone is divided into five equal-frequency groups among eligible candidates, ordered from lighter to darker. This sampling scheme provides attribute-balanced real images and a shared basis for designing generation prompts.

Attribute-aligned generation prompts. We design shared prompts around the same 50 attribute combinations used for real-image selection. Each prompt specifies age, gender presentation, and skin tone, while the remaining description stays constant. The template requests a close-up portrait of one adult facing the camera, with a clearly visible face, natural expression, lighting, and skin texture. Skin tone is described using five predefined levels. Each of the seven generators produces ten images per prompt, yielding 500 images per source and 3,500 generated images in total. Real-image selection and generation prompts therefore use the same attribute categories and sample quotas. The complete generation prompt template and attribute substitutions are provided in Appendix B.2.

All five MLLMs evaluated the 3,500 AI-generated images and the same 500 real photographs, yielding 20,000 responses. This cross-model design compares each MLLM across sources and enables comparisons among multiple MLLMs on the same images. The shared real photographs provide a common reference for measuring false positives.

Model selection. We select five generator–detector pairs from the same developers and two additional open-weight generation sources, FLUX and HiDream (Table 1). The paired systems support same-/cross-developer comparisons, while the unpaired sources extend the evaluation beyond such generator–detector pairs. For provenance analysis, we distinguish file-level C2PA records, pixel-level invisible watermarks, and visible marks rendered in the image (C2PA, n.d.; Gowal et al., 2025).

3.2 Unified Detection Protocol

Each MLLM received one image per independent request with the same prompt, summarized below. The prompt omits the ground-truth label, generation source, and processing condition, and does not direct attention to any specific watermark or artifact. We analyze the retained valid response for each completed task; Appendix C.5 reports the response-history audit.

> **Detection prompt (summary):** Inspect this portrait image and determine whether it was captured by a camera or generated by an AI image model. Return an authenticity judgment, an estimated probability that the image is AI-generated, and a brief explanation. When citing visual evidence, identify the corresponding image region. Output `FAKE` if the generation probability is at least 0.5; otherwise output `REAL`.

Detection rate is the fraction of generated images classified as `FAKE`, while false-positive rate is the fraction of real photographs incorrectly classified as `FAKE`. Balanced accuracy averages sensitivity

and specificity, while AUROC uses the reported generation probabilities as ranking scores. We compare judgments across provenance conditions and evaluate explanations for judgment repeatability and image support. The detection prompt and request settings are provided in Appendix C.

3.3 Evaluation Design

Same-developer detection association (RQ1). We first report detection rates for all seven generation sources and false-positive rates on the shared real photographs. The primary same-developer analysis uses the 2,500 generated images from the five paired sources and 12,500 MLLM responses. Let $Y_{d,i}$ indicate whether MLLM d classifies image i as `FAKE`, and $g(i)$ denote its source. The indicator $S_{d,g(i)}$ equals one when the detector and generator share a developer and zero otherwise. We fit

$$\operatorname{logit} \Pr(Y_{d,i} = 1) = \mu + \alpha_d + \gamma_{g(i)} + \beta S_{d,g(i)}, \tag{1}$$

where α_d and $\gamma_{g(i)}$ capture MLLM-specific and generation-source-specific effects, respectively. We estimated a standardized association by averaging the predicted probability difference between indicator values 1 and 0 over the 5×5 combinations. This model-based summary complements the observed diagonal–off-diagonal contrast. Post hoc linear-probability and probit fits assess specification sensitivity, with cell-level diagnostics in Appendix E.3.

We used post hoc Firth logistic regression to estimate pooled and developer-specific associations because the unpenalized developer-specific model exhibits separation. To assess sensitivity to selected developers, we excluded each developer's detector and generation source in turn and re-estimated the overall association. More details are provided in Appendices E.2 and E.4.

Paired provenance interventions (RQ2). We selected 100 images each from GPT Image 2 and Gemini 3.1 Flash Image, covering all 50 prompts with two images per prompt. Each selected image defines a three-condition provenance intervention set: A is the original image, B is obtained by selectively removing C2PA records, and C is generated from B using the MarkNull pipeline for SynthID removal (Cao et al., 2026). The A–B comparison isolates the effect of C2PA record removal, while the B–C comparison captures the effect of the MarkNull-based SynthID removal procedure.

We verified provenance credentials and decoded pixels before and after processing for every image. All 200 original images contain a C2PA manifest, which is absent after removal, and every A/B pair has identical pixel values. We applied OpenAI's official content-provenance checker and Google's official Gemini Apps verification function to the complete RQ2 provenance image set. Full-set verification confirms successful SynthID removal in all 200 processed images. Representative verification outputs for both generation sources across A, B, and C are shown in Figure 5.

MarkNull processing introduces visual changes. Across the 200 B/C pairs, the mean PSNR, SSIM, and LPIPS values are 29.94 dB, 0.814, and 0.145, respectively. Figure 2 shows representative pairs and means. The B–C comparison therefore captures successful SynthID removal together with the visual changes introduced by the MarkNull pipeline. Metric settings and source-level distribution summaries are provided in Appendix D.2 and Table 12.

To account for response variability, each MLLM evaluated every image under each condition three times, yielding 9,000 responses. Condition A is repeated alongside B and C for paired comparisons. For each MLLM and parent image, we computed the fraction of `FAKE` predictions under each condition. We averaged the changes $B - A$ and $C - B$ over parent images, both overall and by generation source. The repeated evaluations of condition A also support RQ3 judgment-repeatability analysis and supplementary cue-overlap analysis. We computed two-sided p-values using a normal approximation with hierarchical-bootstrap standard errors and applied Holm correction across ten overall comparisons. A paired sign-flip analysis at the parent-image level provides a sensitivity check.

To examine processing sensitivity on camera-origin inputs, we applied MarkNull to a supplementary set of 100 real photographs, with two from each of the same 50 attribute combinations. Each MLLM evaluated every photograph three times before and after processing. We compared FAKE rates before and after processing to examine sensitivity to the same pipeline on real-source inputs.

Repeated judgments and image-supported explanations (RQ3). The repeated-judgment analysis used the 3,000 responses for the 200 original images in condition A. These responses form 1,000 groups, each containing three outputs from the same MLLM for the same image. We retain both



B–C visual similarity
Mean over 100 pairs per source

GPT Image 2	
PSNR ↑	30.95 dB
SSIM ↑	0.826
LPIPS ↓	0.150
Gemini 3.1 Flash Image	
PSNR ↑	28.93 dB
SSIM ↑	0.803
LPIPS ↓	0.140

Figure 2: Paired provenance-intervention examples and B–C visual similarity. Rows show GPT Image 2 and Gemini 3.1 Flash Image outputs. Columns show the original image (A), C2PA removal with identical decoded pixels (B), and MarkNull processing for SynthID removal (C). The table reports mean PSNR, SSIM, and LPIPS over 100 B/C pairs per generation source.

`REAL` and `FAKE` responses. We classify each group as consistently correct, consistently incorrect, or containing a label change.

We also summarize within-group variation in the reported AI probability. For an exploratory description of explanation overlap, we extracted visual cues using a rule-based representation with three components: object, attribute, and state. For coarse-grained analysis, we grouped related attributes of the same object into categories, such as grouping skin pores and wrinkles into skin-surface texture. Category-level Jaccard similarity measures overlap in these categories while ignoring states. We also report the proportions of distinct categories appearing in all three explanations or only one. For each group, the union of categories across the three explanations is the denominator; we then averaged across groups. Strict object–attribute–state matching provides a complementary sensitivity analysis, while predefined opposing states were examined separately. Extraction coverage and empty-set handling are reported in Appendix F.3.

As a preliminary rule check, two reviewers independently inspected 20 groups of three explanations: four groups per detector, with two from each generation source. They assessed missed cues, extraction errors, and opposing-description matches, then adjudicated disagreements. This development review informed rule revisions before the full analysis. For the image-evidence assessment, we sampled 100 explanations from correct detections in the main detection evaluation, with 20 per detector. Sampling precedes explanation-quality review and seeks coverage across generation sources and prompts. The two reviewers independently compared each explanation with the original image and labeled its specific observations as supported, contradicted, or indeterminate. They then adjudicated disagreements. We report observation-level label proportions, the proportions of explanations containing each label, and reviewer agreement before adjudication. This assessment measures image support for observations in explanations accompanying correct detections.

For AI-generated image analyses, we estimated 95% confidence intervals using 2,000 hierarchical bootstrap replicates. We resampled prompts, then parent images within each prompt and generation source, retaining all detector responses and repeated judgments for each image. For the human image-evidence assessment, sensitivity analyses resampled explanations within detector–source strata or clustered by shared parent image. Results from the real-photograph control are reported descriptively. Full statistical details are provided in Appendix E.

4 Results

4.1 Cross-Model Detection Performance

Overall detection performance. Table 2 summarizes the overall performance of the five MLLMs with 95% confidence intervals. GPT shows the highest generated-image detection rate and balanced

accuracy among the evaluated MLLMs. Seed has the lowest false-positive rate but the lowest AI-generated image detection rate. Gemini and Qwen have similar overall detection rates, with Gemini showing lower balanced accuracy due to more frequent false positives on real photographs.

Table 2: Overall detection performance. Brackets indicate 95% bootstrap confidence intervals.

Detector	**Detection rate**(%) ↑	**False-positive rate**(%) ↓	**Balanced accuracy**(%) ↑	**AUROC** ↑
GPT	**84.34** [82.57, 86.14]	4.40 [2.60, 6.20]	**89.97** [88.64, 91.20]	**0.970**
Gemini	62.26 [60.31, 64.17]	27.60 [23.60, 31.40]	67.33 [65.10, 69.60]	0.714
Qwen	65.17 [63.20, 67.17]	7.20 [5.00, 9.40]	78.99 [77.57, 80.50]	0.915
Seed	29.71 [28.11, 31.34]	**0.80** [0.20, 1.60]	64.46 [63.54, 65.36]	0.838
Grok	46.83 [44.31, 49.34]	2.40 [1.20, 3.80]	72.21 [70.79, 73.61]	0.897

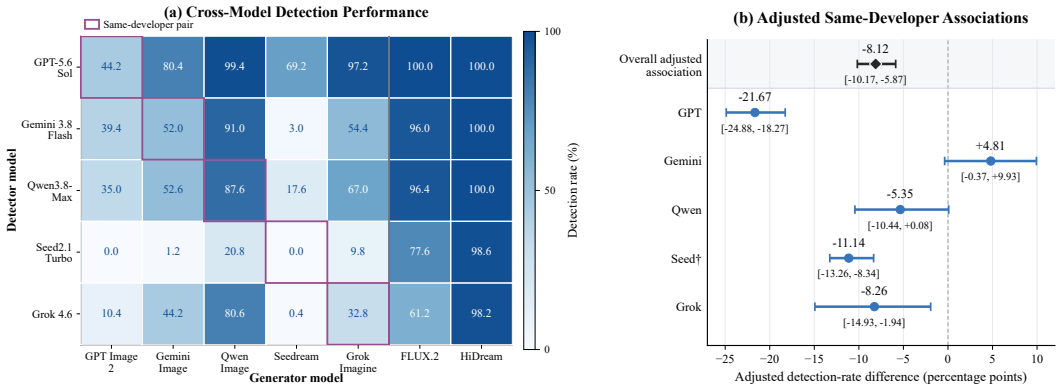

Figure 3: Cross-model detection and same-developer associations. (a) Detection rates (%); purple outlines mark same-developer pairs. (b) Prespecified pooled and post hoc pair-specific model-based associations, with 95% bootstrap intervals. † indicates an estimate affected by zero-detection cells.

Figure 3(a) shows detection rates across generation sources. All five MLLMs identify most HiDream images, whereas detection rates on Seedream images vary substantially. Even for the same detector, detection rates differ markedly across generation sources. The same-developer pairs outlined in purple show no consistent detection advantage.

Same-developer detection association. For the five matched generator–MLLM pairs, the mean detection rates for same-developer and cross-developer combinations are 43.32% and 43.68%, respectively. The raw difference is -0.36 percentage points (95% CI: $[-2.74, 1.92]$). Appendix Figure 6 displays this balanced observed comparison. The prespecified logit model gives a standardized difference of -8.12 points, and a supplementary probit model gives -7.96 points. Figure 3(b) presents the original model-based estimates. The linear-probability model reproduces the raw difference. Appendices E.2–E.4 report the estimates, intervals, and specification diagnostics.

An output-history sensitivity check excludes every parent with an output-validation error for any detector, preserving complete detector panels. The observed difference is -0.40 points and the logit difference is -8.12 points after exclusion (Appendix C.5).

Finding 1 (RQ1). Models do not show a general tendency to detect AI-generated facial images more readily when the generator and detector share a developer. The balanced observed contrast is close to zero. The prespecified logit and supplementary probit models yield negative standardized associations, but their magnitudes depend on the model specification and pair-specific estimates vary across systems.

4.2 Effects of Provenance Processing on Detection

Overall and source-specific changes. Table 3 reports FAKE-rate changes after C2PA removal ($B - A$) and SynthID removal ($C - B$). None of the five C2PA comparisons remains significant

Table 3: Overall changes in FAKE rate (percentage points) after C2PA removal ($B - A$) and the subsequent SynthID removal ($C - B$), with 95% bootstrap confidence intervals and Holm-adjusted p-values.

	C2PA removal ($B - A$)		SynthID removal ($C - B$)	
Detector	**Change [95% CI]**	**Adj. p**	**Change [95% CI]**	**Adj. p**
GPT	0.00 [-3.17, 3.17]	1.000	−13.00 [-19.00, -7.00]	< 0.001
Gemini	−3.50 [-9.17, 2.00]	0.918	+4.00 [-3.83, 12.33]	0.997
Qwen	−4.83 [-10.33, 1.00]	0.565	+30.50 [24.50, 36.83]	< 0.001
Seed	+1.00 [-0.33, 2.67]	0.918	+15.67 [11.00, 20.83]	< 0.001
Grok	+0.17 [-4.00, 4.33]	1.000	−13.50 [-17.83, -9.17]	< 0.001

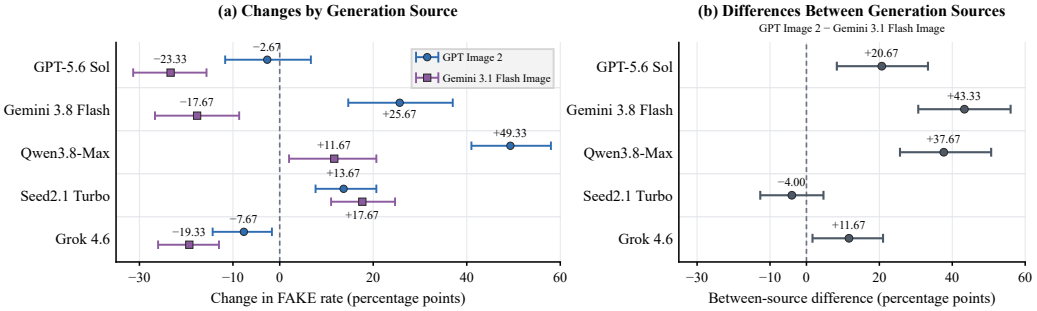

Figure 4: Changes in FAKE rate after SynthID removal ($C - B$). (a) Changes by generation source. (b) Between-source differences, calculated as GPT Image 2 minus Gemini 3.1 Flash Image. Bars indicate 95% hierarchical-bootstrap intervals.

after Holm correction. After SynthID removal, FAKE rates decrease for GPT and Grok and increase for Qwen and Seed, while Gemini shows no significant change after correction. A parent-level sign-flip analysis yields the same Holm-corrected rejection decisions (Table 19).

The source-stratified results reveal differences hidden by overall averages. As shown in Figure 4(a), after SynthID removal, Gemini shows opposite changes across generation sources: its FAKE rate increases for GPT Image 2 outputs but decreases for Gemini Image outputs. Figure 4(b) compares these source differences, with a between-source difference of 43.33 percentage points (95% CI: [30.67, 56.00]). Qwen's FAKE rate increases for both sources, with a larger increase for GPT Image 2 images, whereas Seed shows similar increases across sources. The Gemini source difference also passes a post hoc Holm correction across ten source-effect contrasts (Appendix E.6). These results show that aggregated changes can hide source-specific effects; estimates are provided in Table 13.

Control analysis on real photographs. Applying the same procedure to 100 real photographs increases the FAKE rate for all five detectors, by 4.33–56.67 percentage points. Thus, processing can make even images originating from real photographs more likely to be classified as AI-generated. Together with the measured visual changes from MarkNull processing, these results show that the intervention introduces both provenance-related changes and visual changes. On real photographs, the processing-induced changes shift judgments toward AI-generated classifications. This supplementary analysis is reported separately and is not included in the overall comparisons on the original 200 generated parent images. Before-and-after FAKE rates are reported in Table 14.

Finding 2 (RQ2). C2PA removal produces no statistically significant overall change in judgments after Holm correction. Full-set verification confirms successful SynthID removal across the intervention set. The verified SynthID-removal procedure changes detection outcomes in detector- and generation-source-dependent ways. In source-stratified analyses, Gemini's detection rate decreases for Gemini Image outputs but increases for GPT Image 2 outputs, while Qwen and Seed show increases across both sources. The SynthID-removal procedure also increases FAKE rates on real photographs across detectors, showing that its accompanying visual changes can shift judgments to-

ward AI-generated classifications. Together, these results show source-dependent detector responses to a verified SynthID-removal procedure and its associated image changes.

4.3 Judgment Repeatability and Explanation Support

Repeatability across judgments. Table 4 separates repeatability by correctness on the 200 original generated images. Across all detectors, 21.7% of groups are consistently correct, 48.0% are consistently incorrect, and 30.3% change labels. Seed repeatedly labels 198 of 200 images REAL, giving 99.0% agreement and 99.0% stable errors.

Table 4: Repeated judgments on original generated images (%). Each detector contributes 200 groups of three responses. The three columns partition all groups.

Detector	Stable correct	Stable incorrect	Label change
GPT	49.0	22.5	28.5
Gemini	24.0	32.0	44.0
Qwen	22.0	27.0	51.0
Seed	0.0	99.0	1.0
Grok	13.5	59.5	27.0
Overall	21.7	48.0	30.3

Within-group probability standard deviations average 0.100 across all detectors. Appendix F.1 reports intervals and score ranges. Exploratory coded cue overlap remains in Appendices F.3–F.6.

Visual evidence support for correct detections. Two reviewers evaluated image evidence in 100 explanations accompanying correct detections, with 20 sampled per MLLM. After adjudication, 259 of 288 specific observations were supported by the images, one was contradicted, and 28 were indeterminate. Thus, most sampled explanations contained observations supported by the image, while explicit contradictions were rare and some descriptions lacked sufficient evidence for determination.

At the explanation level, 97 explanations contain supported observations and 23 contain indeterminate observations; an explanation can contain both. Before adjudication, observation-level agreement was 89.9%, with Cohen's $\kappa = 0.253$. Agreement is concentrated in supported observations; Appendix F.8 reports the full label matrix and disagreement examples. Two resampling analyses yielded similar results. This review covered sampled explanations accompanying correct detections. This assessment measures whether stated observations match the input images. Appendix F.9 reports reviewer agreement and results by detector, while Table 21 gives the resampling intervals.

Finding 3 (RQ3). Across repeated judgments, groups show three patterns: stable correct classifications, stable incorrect classifications, and label changes. Among sampled explanations accompanying correct detections, most specific observations are supported by the image, and explicit contradictions are rare. These results show that judgment repeatability and image support capture different aspects of detection behavior: models can repeat an incorrect judgment, while most reviewed observations still match visible image content.

5 Conclusion

Understanding how MLLMs detect AI-generated facial images requires examining both visual evidence and provenance-related signals that may influence their judgments. In our evaluation, same-developer pairings show no consistent detection advantage, while the effects of provenance processing vary across detectors and generation sources. These findings suggest that future evaluations should jointly consider visual content and provenance information, distinguish file-level record changes from pixel-level processing effects, and evaluate classification behavior together with explanation consistency and image support. The facial images used in this study do not cover all types of generated content, but their controlled attributes provide a basis for comparing detection behavior across models. As models and provenance technologies continue to evolve, future work should combine more content-preserving processing with independent provenance verification. This would help disentangle the roles of visual features, file-level credentials, and pixel-level signals in model judgments across broader image types and out-of-distribution settings.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with research ideation or execution, drafting sections of the paper, polishing the manuscript, and generating synthetic datasets used for experimental evaluation. We have not used generative AI tools for retrieval and discovery of related work or for proving mathematical claims. Other required disclosure categories are not applicable to this work. All AI-assisted materials and outputs were reviewed and verified by the authors. The authors take responsibility for the final content of this work, including the methodology, experimental design, analyses, claims, and artifacts presented in the paper.

ETHICS STATEMENT

This study involved a minimal-risk human annotation task in which two annotators assessed whether observations in MLLM explanations were supported by AI-generated facial images. The task collected only anonymized judgments and no identifying, sensitive, or biometric information; under the research ethics policy of the lead author's institution, it did not require institutional ethics review. Real photographs were used only for aggregate false-positive analysis, without identifying the depicted individuals or redistributing the raw images. Age, gender presentation, and apparent skin tone were used solely as coarse image-level sampling strata, not as identity labels or protected-attribute assessments.

REFERENCES

Alibaba Cloud. qwen3.8-max model info. Alibaba Cloud Model Studio documentation, 2026a. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-8-max.

Alibaba Cloud. qwen-image-3.0-pro model info. Alibaba Cloud Model Studio documentation, 2026b. URL https://www.alibabacloud.com/help/en/model-studio/qwen-image-3-0-pro.

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, pp. 1456–1492, 2024. URL https://proceedings.mlr.press/v235/an24a.html.

Black Forest Labs. FLUX.2 [dev]. Hugging Face model card, 2025. URL https://huggingface.co/black-forest-labs/FLUX.2-dev.

ByteDance Seed. Seed2.1. Official model page, 2026a. URL https://seed.bytedance.com/en/seed2_1.

ByteDance Seed. Seedream 5.0 Pro. Official model page, 2026b. URL https://seed.bytedance.com/en/seedream5_0_pro.

C2PA. *C2PA and Content Credentials Explainer*. Coalition for Content Provenance and Authenticity (C2PA), n.d. URL https://spec.c2pa.org/specifications/specifications/2.2/explainer/Explainer.html. Specification version 2.2.

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, et al. HiDream-O1-Image: A natively unified image generative foundation model with pixel-level unified transformer. arXiv preprint arXiv:2605.11061 [cs.CV], 2026. URL https://arxiv.org/abs/2605.11061.

Jie Cao, Qi Li, Zelin Zhang, Xiaodong Wu, Lingshuang Liu, Xiangman Li, and Jianbing Ni. MarkNull: Model-agnostic watermark removal in AI-generated images via on-manifold latent manipulation. In *Proceedings of the 35th USENIX Security Symposium*, pp. 4721–4740, 2026. URL https://www.usenix.org/conference/usenixsecurity26/presentation/cao.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019. doi: 10.1109/CVPR.2019.00482.

Hongyan Fei, Zexi Jia, Chuanwei Huang, Jinchao Zhang, and Jie Zhou. Exploring specular reflection inconsistency for generalizable face forgery detection. In *Proceedings of the 2026 International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1de15773f510ec4edd1b3a105d979667-Abstract-Conference.html.

K. J. Kevin Feng, Nick Ritchie, Pia Blumenthal, Andy Parsons, and Amy X. Zhang. Examining the impact of provenance-enabled media on trust and accuracy perceptions. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):270:1–270:42, 2023. doi: 10.1145/3610061.

Google. Nano Banana 2: Google's latest AI image generation model. Blog article, 2026. URL https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/.

Google DeepMind. Gemini 3.1 Flash Image: Model card. Google DeepMind, 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-flash-image/.

Google DeepMind. Gemini 3.8 Flash: Model card. Google DeepMind, 2026b. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/.

Sven Gowal, Rudy Bunel, Florian Stimberg, David Stutz, Guillermo Ortiz-Jimenez, Christina Kouridi, Mel Vecerik, Jamie Hayes, Sylvestre-Alvise Rebuffi, Paul Bernard, Chris Gamble, Miklós Z. Horváth, Fabian Kaczmarczyck, Alex Kaskasoli, Aleksandar Petrov, Ilia Shumailov, Meghana Thotakuri, Olivia Wiles, Jessica Yung, Zahra Ahmed, Victor Martin, Simon Rosen, Christopher Savčak, Armin Senoner, Nidhi Vyas, and Pushmeet Kohli. SynthID-Image: Image watermarking at internet scale. arXiv preprint arXiv:2510.09263 [cs.CR], 2025. URL https://arxiv.org/abs/2510.09263v1. Version 1.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. In *Proceedings of the 2022 International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RhB1AdoFfGE.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 105–116, 2025. doi: 10.1109/CVPR52734.2025.00019.

ISO/CIE. Colorimetry—Part 4: CIE 1976 L*a*b* colour space. Technical Report ISO/CIE 11664-4:2019, International Organization for Standardization, 2019.

Jiazhou Ji, Jie Guo, Weidong Qiu, Zheng Huang, Yang Xu, Xinru Lu, Xiaoyu Jiang, Ruizhe Li, and Shujun Li. "I know myself better, but not really greatly": How well can LLMs detect and explain LLM-generated texts? arXiv preprint arXiv:2502.12743 [cs.CL], 2025. URL https://arxiv.org/abs/2502.12743v2. Version 2.

Yikun Ji, Yan Hong, Qi Fan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, Jianfu Zhang, et al. FakeXplain: AI-generated image detection via human-aligned grounded reasoning. In *Proceedings of the 2026 International Conference on Learning Representations*, volume 2026, pp. 68549–68570, 2026. URL https://openreview.net/forum?id=UcpTOa8OnG.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063, 2024. doi: 10.18653/v1/2024.findings-emnlp.290. URL https://aclanthology.org/2024.findings-emnlp.290/.

Kimmo Karkkainen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the 2021 IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1548–1558, 2021. doi: 10.1109/WACV48630.2021.00159.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019. doi: 10.1109/CVPR.2019.00453.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Yixuan Li, Xuelin Liu, Xiaoyang Wang, Bu Sung Lee, Shiqi Wang, Anderson Rocha, and Weisi Lin. FakeBench: Probing explainable fake image detection via large multimodal models. *IEEE Transactions on Information Forensics and Security*, 20:8730–8745, 2025. doi: 10.1109/TIFS.2025.3597211.

Ellis Monk. Monk Skin Tone Scale™. Official scale website, 2019. URL https://monkscale.com/.

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

OpenAI. Advancing content provenance for a safer, more transparent AI ecosystem. Official announcement, 2026a. URL https://openai.com/index/advancing-content-provenance/.

OpenAI. GPT-5.6 system card. OpenAI Deployment Safety Hub, 2026b. URL https://deploymentsafety.openai.com/gpt-5-6.

OpenAI. GPT-Image-2. OpenAI API documentation, 2026c. URL https://developers.openai.com/api/docs/models/gpt-image-2.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2197. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html.

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

Dani Roytburg, Matthew Bozoukov, Matthew Nguyen, Jou Barzdukas, Mackenzie Puig-Hall, and Narmeen Fatimah Oozeer. Are LLM evaluators really narcissists? sanity checking self-preference evaluations. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=sRPyBOSZQk.

SpaceXAI. Model card: Grok 4.6, 2026a. URL https://media.x.ai/v1/website/card-4p6-4cd2dc57.pdf. Revision: 2026-08-17.

SpaceXAI. Grok Imagine Image 2.0. API documentation, 2026b. URL https://docs.x.ai/developers/models/grok-imagine-image-2.0.

Siwei Wen, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Conghui He, Weijia Li, et al. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *Advances in Neural Information Processing Systems*, 38:58972–59005, 2026. URL https://openreview.net/forum?id=xLFYd1owiP.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *Proceedings of the 2025 International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025. URL https://openreview.net/forum?id=ODRHZrkOQM.

Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, et al. LOKI: A comprehensive synthetic data detection benchmark using large multimodal models. In *Proceedings of the 2025 International Conference on Learning Representations*, volume 2025, pp. 70440–70522, 2025. URL https://openreview.net/forum?id=z8sxoCYgmd.

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 *Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision*, pp. 18746–18758. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01742.

A Appendix Overview

The appendix provides the following details:

- B. **Dataset Construction and Sampling:** Real-image selection, generation prompts, image collection, and sampling for the provenance and explanation analyses.
- C. **Detection Protocol and Model Configurations:** The complete detection prompt, model and request configurations, image submission procedures, and response handling.
- D. **Provenance Processing and Verification:** C2PA removal, MarkNull processing, image-quality measurements, the scope of watermark verification, and supplementary results on generated and real-source images.
- E. **Statistical Analysis and Sensitivity Checks:** Estimand definitions, hierarchical bootstrap procedures, multiple-comparison correction, and supplementary analyses of same-developer associations, paired judgment changes, and uncertainty in human evaluation.
- F. **Explanation Analysis and Human Evaluation:** Cue extraction and consistency metrics, preliminary rule checks, sampling and annotation procedures, reviewer agreement, and image-evidence assessments.
- G. **Limitations and Future Work:** Dataset and model coverage, interpretation of provenance-processing results, and future directions for explanation evaluation.

B Dataset Construction and Sampling

B.1 Real-Image Selection

We screen the 70,000 images in Flickr-Faces-HQ (FFHQ) (Karras et al., 2019). We use SCRFD (Guo et al., 2022) for face detection and landmarks, FairFace (Karkkainen & Joo, 2021) for apparent age and gender presentation, and ArcFace embeddings (Deng et al., 2019) for identity-similarity checks. Table 5 lists the selection thresholds. Candidates must have a resolution of 1024×1024 pixels and contain one qualifying face. The selection retains ten images in each attribute combination defined in the main text. Expression, hair, accessories, background, and lighting have no sampling quotas.

Apparent skin tone. We estimate skin color from forehead and cheek regions in the face crop. Each region contributes its median RGB value after excluding pixels with mean intensity outside $[25, 240]$. We convert these colors to the CIELAB color space (ISO/CIE, 2019) and compare their median with the ten Monk Skin Tone (MST) palette colors (Monk, 2019). A softmax over negative color distances, with temperature 10, produces weights for the expected MST score. The confidence threshold uses the largest summed weight among adjacent palette pairs; the sampling groups use quintiles of the continuous score. The four quintile boundaries are 4.8184, 5.5341, 6.1191, and 6.7064, estimated from adult candidates that meet the image-quality and frontal-face criteria. These groups describe relative apparent skin tone within FFHQ.

Skin-patch dispersion is the mean CIELAB distance from the three patch colors to their median. Extreme pixels have at least one RGB channel at or below 5 or at or above 250. These checks reduce poorly exposed or unevenly colored candidates. Age and gender-presentation labels likewise describe image-based estimates, without establishing demographic ground truth.

Table 5: Thresholds for real-image selection. Quality measurements use the face crop.

Criterion	Threshold
Face-detection confidence	≥ 0.25
Adult probability / age-group confidence	≥ 0.90 / ≥ 0.40
Gender-presentation confidence	≥ 0.90
Skin-tone grouping confidence	≥ 0.40
Face area / image area	≥ 0.08
Absolute yaw	$\leq 15^\circ$
Laplacian variance	≥ 50
Mean grayscale intensity	$[35, 220]$
Extreme-pixel proportion	≤ 0.15
Skin-patch color dispersion	≤ 18
Identity cosine similarity	< 0.60

Selection order and checks. We use a fixed ordering of FFHQ filenames and retain the first eligible images that fill each attribute quota. File-duplicate checks and ArcFace similarity screening exclude repeated files and suspected repeated identities. The latter uses the threshold in Table 5; independent calibration of that threshold was not completed. The selected set contains 500 images with unchanged source pixels. We manually inspect all 500 selected photographs against the image-selection criteria.

B.2 Generation Prompts

All generation sources receive the same prompt text for each attribute combination. The following template and substitutions specify all 50 prompts; each source produces ten images per prompt.

Generation prompt (complete template): A realistic square portrait photograph of a single adult **{gender}** **{age}** with **{skin}**, wearing everyday clothing. Use a close-up portrait composition with the face clearly visible and generally oriented toward the camera. Capture a natural expression in a natural-looking setting, with lifelike skin texture, natural-looking photographic lighting, and realistic levels of sharpness and detail.

Gender	woman; man.
Age	in their twenties; in their thirties; in their forties; in their fifties; in their sixties or older.
Skin	very light skin; light skin; medium skin; medium-dark skin; dark skin.

The skin descriptions correspond in order to the five real-image sampling groups, from lighter to darker. This alignment specifies attribute categories and quotas; it does not establish exact skin-color matching between generated and real images. We leave the remaining prompt text unchanged across sources and repetitions. Generated images retain their assigned prompt categories without post-generation relabeling to enforce visual agreement.

B.3 Image Collection and File Handling

Table 6 lists the generation models and image dimensions. Each source contributes 500 images, giving 3,500 generated images in total. We preserve provider-returned files for the original-image evaluation and create separate copies for the provenance-processing experiment. The common prompt specifies a square portrait.

All models use the shared input prompts above, with generation settings fixed within each source.

We preserve the original generated images and their prompt assignments. Image selection is independent of detector outputs.

B.4 Sampling for Subsequent Analyses

The provenance panel uses repetitions 1 and 2 of each prompt from GPT Image 2 and Gemini Image, selected before verification or detection. The explanation-consistency analysis reuses its original

Table 6: Generation models and image dimensions. Each model contributes 500 images.

Generation model	Resolution
GPT Image 2	1024×1024
Gemini 3.1 Flash Image	1024×1024
Qwen-Image-3.0-Pro	1024×1024
Seedream 5.0 Pro	1024×1024
Grok Imagine Image 2.0	1024×1024
FLUX.2 [dev]	1024×1024
HiDream-O1-Image-Dev-2604	1024×1024

A images with both REAL and FAKE judgments. The real-source supplement uses positions 1 and 2 within each real-image cell. The human image-evidence evaluation samples explanations accompanying correct RQ1 detections.

C Detection Protocol and Model Configurations

C.1 Complete Detection Prompt

We hold the prompt and single-image request format constant across detectors. Each image is evaluated in a fresh context, with study labels and verification results withheld to limit contextual and source-related cues. The prompt requests the generation probability separately from the classification and explanation.

Detection prompt (complete):

Examine this single portrait image and determine whether it was captured by a camera or generated by an AI image model.

Return only a JSON object with the following fields:

- "**classification**": "`FAKE`" if the image is AI-generated, or "`REAL`" if it is camera-captured.
- "**probability_ai**": a number between 0 and 1 estimating the probability that the image is AI-generated. This is an estimated probability, not confidence in the selected label.
- "**justification**": a brief justification in English containing at most three supporting observations and no more than 40 words. When citing visual evidence, indicate the relevant image region.

Set "**classification**" to "`FAKE`" when "**probability_ai**" is at least 0.5, and to "`REAL`" otherwise. Do not include markdown or any text outside the JSON object.

C.2 Model Settings and Image Submission

We evaluate GPT-5.6 Sol, Gemini 3.8 Flash, Qwen3.8-Max, Seed2.1 Turbo, and Grok 4.6. Each detector uses a fixed request configuration throughout the comparisons, with low reasoning effort. GPT uses the endpoint's default temperature; the other four detectors use temperature zero. Qwen and Seed requests specify a maximum output budget of 4,096 tokens. We hold these settings fixed across repeated requests.

The client encodes the input file bytes as a base64 data URL and supplies the image and prompt in a single user message. It performs no additional resizing or re-encoding at submission and requests the three-field JSON schema. We use the same submission procedure across conditions, so the paired C2PA comparison measures downstream judgments for the submitted files. The recorded client payload establishes file submission; server-side credential exposure to the answering model remains unobserved. The dedicated verification interfaces provide separate evidence about the files.

C.3 Independent Requests and Repeated Judgments

RQ1 assigns one evaluation task to each detector–image pair. RQ2 assigns three independent tasks to each detector–parent–condition combination. It reevaluates condition A alongside B and C using

new requests. The real-source supplement likewise uses three independent requests per detector and condition. We randomize image or task order and evaluate each judgment in a fresh context.

C.4 Response Validation and Completion

We parse the response as JSON, accepting an enclosing Markdown code fence if present. A valid response must contain exactly the three requested fields: a REAL or FAKE label, a numeric probability in $[0, 1]$, and a nonempty justification. We reject label–probability conflicts at the 0.5 threshold, malformed responses, and refusals without substituting labels or scores. The word limit guides generation but does not determine response validity. We retain and flag overlength explanations; their length does not alter the classification, probability, or inclusion rule.

The initial retry rule permits one identical retry for transient request failures. Separate completion passes fill unresolved tasks while preserving prior attempts and already valid responses. Each completed task contributes one retained valid response. Completion passes revisit unresolved tasks, including tasks with earlier output-validation errors. Appendix C.5 distinguishes transmission failures, returned-answer validity, and subsequent completion. The final panels contain 20,000 valid RQ1 responses, 9,000 generated-image RQ2 responses, and 3,000 real-source supplementary responses.

C.5 Response-History Audit and Completion Sensitivity

We link all 40,078 RQ1 log events to their detector–image tasks and order events by request start time. The 20,078 error events include 19,667 SSL EOF failures (97.95%), alongside other transmission and output-validation failures (Table 7). Table 8 separates event counts from affected tasks and first-answer validity. For this audit, the first API answer is the earliest event with an archived completion or a retained valid response. Transmission failures precede that answer and contribute to the event count.

Of the 20,000 tasks, 19,768 (98.84%) return a valid first recorded API answer. Output-validation errors affect 232 tasks: 194 for Seed and 38 for Qwen. Among 262 nontext-content errors, 255 have a `length` finish reason and seven have `stop`; 11 of 20 invalid-JSON errors also have `length`. All 322 error events containing archived API choices have an empty refusal field. This field audit records the API metadata available for those errors.

The runner allows one within-run retry for transient failures; later completion passes revisit unresolved tasks, including prior output-validation failures. The single label–probability conflict occurs in a task that later returns a valid REAL judgment. The final performance estimates describe the retained valid outputs from this completion procedure.

Table 7: RQ1 error events and distinct affected tasks. Tasks can appear in more than one category.

Error category	Events	Affected tasks
SSL EOF	19667	11083
Nontext assistant content	262	176
Invalid field set	39	38
Timeout	37	37
Invalid JSON	20	20
Remote disconnect	18	15
HTTP 400	16	4
HTTP 502	10	10
HTTP 429	2	2
HTTP 503	2	2
Incomplete read	2	2
HTTP 520	1	1
Connection reset	1	1
Label–probability conflict	1	1

Source coverage and sensitivity. We exclude every parent image with an output-validation error for any detector, retaining all five responses for each remaining parent. This removes 229 parents:

Table 8: RQ1 task-level response coverage. Each detector has 4,000 tasks and 4,000 retained valid responses. First-answer validity excludes preceding transmission failures.

Detector	Error events	Tasks with errors	Valid first answer	Output-error tasks
GPT	3403	1943	4000	0
Gemini	2913	1690	4000	0
Qwen	4190	2348	3962	38
Seed	5666	2953	3806	194
Grok	3906	2191	4000	0

213 generated images and 16 real photographs. Table 9 reports the retained source composition. The matched-source analysis retains 2,351 parents, with at least four retained images in each prompt–source stratum. For the observed contrast, we retain equal weights over the five diagonal and 20 off-diagonal cell means. The logit fit uses the retained binomial counts and the original 25-cell standardization.

We generate 2,000 shared hierarchical resamples with fixed random-number initialization. Each replicate resamples prompts and original parents, applies the fixed exclusion flags, and recomputes retained denominators. All fits are valid in both the full and excluded panels. This post hoc sensitivity describes the retained panel alongside the original data (Table 10). Across the seven sources and real photographs, balanced accuracy changes by at most 1.08 percentage points after exclusion. The paired raw difference changes from -0.36 to -0.40 points, while both logit differences round to -8.12 points. The exclusion changes source composition and provides a descriptive check on the observed completion history.

Table 9: Source-level errors and retained parents. Each source originally contains 500 parents and 2,500 detector–image tasks. Parent exclusion retains complete five-detector panels.

Source	Error events	Output-error tasks	Retained parents
GPT	2545	11	489
Gemini	2557	26	474
Qwen	2586	88	413
Seed	2383	7	493
Grok	2598	18	482
FLUX	2466	41	459
HiDream	2428	24	477
FFHQ	2515	17	484

Table 10: Same-developer contrasts under the response-history sensitivity. Both panels use the same new bootstrap run; the original primary intervals remain in Table 17. Differences are percentage points.

Panel	Contrast	Difference [95% CI]
Full	Observed	-0.36 $[-2.70, 1.98]$
Full	Logit standardized	-8.12 $[-10.21, -5.88]$
Parent exclusion	Observed	-0.40 $[-2.75, 2.11]$
Parent exclusion	Logit standardized	-8.12 $[-10.29, -5.78]$

The accompanying event index also provides detector–source–error-type counts and logical task histories for offline inspection.

D Provenance Processing and Verification

D.1 C2PA Removal and File Checks

We inspect the original files with ExifTool 13.59 and c2patool 0.27.22. Before removal, the script checks that the JUMBF content contains only the target C2PA label; ambiguous or non-target labels

pause processing. It then copies each original file and removes the target content using ExifTool's `-jumbf:all=` option. The checks compare decoded pixels, dimensions, EXIF orientation, and ICC profiles before and after removal. All 200 files pass these checks, and c2patool finds no remaining target manifest.

The later image-quality audit independently confirms pixel equality for all A/B pairs using the same decoding procedure.

D.2 MarkNull Configuration and Image Quality

We use the official MarkNull implementation with Stable Diffusion v1.5 as the proxy model. Table 11 lists the processing settings. The run uses Python 3.10.21, PyTorch 2.2.2 with CUDA 12.1, Diffusers 0.21.3, and Transformers 4.40.1 on an NVIDIA A40. For this run, we use an empty text prompt, apply the same random-number initialization to each image, and select the math attention backend. All 200 images complete processing. The $C - B$ comparison therefore includes changes from the complete processing and output pipeline.

Table 11: MarkNull settings shared by the generated-image experiment and the real-source supplement. Parameter names follow the implementation.

Setting	Value
Proxy model	Stable Diffusion v1.5
Optimization steps / inference steps	50 / 50
Learning rate / ϵ	0.05 / 1.0
Guidance scale	7.5
λ_{sim}, λ_{LPIPS}, λ_{SSIM}, λ_{MSE}	3, 1.0, 1.5, 20
Super-resolution	Disabled
Attention backend	math

We compute quality metrics on decoded RGB images without resizing or applying EXIF rotation. PSNR uses an intensity range of 255. SSIM uses an 11×11 Gaussian window with standard deviation 1.5, population covariance, and a five-pixel border exclusion. LPIPS uses the VGG network, version 0.1, from `lpips 0.1.4`, with inputs scaled to $[-1, 1]$. Table 12 reports the B/C distributions over all selected pairs, without filtering by quality or detector output.

Table 12: Summary of B/C image-quality distributions. Each source contains 100 pairs. The ranges are empirical 5th–95th percentiles, not confidence intervals.

Source	Metric	Mean	Median	5th–95th percentile
GPT Image 2	PSNR (dB)	30.95	31.07	[27.93, 34.08]
	SSIM	0.826	0.828	[0.738, 0.893]
	LPIPS	0.150	0.147	[0.124, 0.186]
Gemini Image	PSNR (dB)	28.93	29.15	[26.00, 30.98]
	SSIM	0.803	0.806	[0.745, 0.851]
	LPIPS	0.140	0.139	[0.119, 0.167]

D.3 Watermark Verification

We verify all images used in the provenance experiment through OpenAI's content-provenance checker and Gemini Apps' verification function. Full-set verification confirms C2PA removal in B and successful SynthID removal in C for all 200 parent images. The manual verification ledger records full-set coverage; the retained screenshots document representative interface outputs. Figure 5 illustrates these checks with one example from each generation source across A, B, and C. For the GPT Image 2 example, the interface detects SynthID in A and B, detects C2PA credentials only in A, and detects neither signal in C. For the Gemini Image example, the response cites C2PA

records in A, attributes B to Google AI, and reports no SynthID or content credentials in C. The displayed response for B attributes the image to Google AI without naming the signal. The verification interfaces report SynthID in the corresponding inputs and no SynthID detection after the removal procedure.

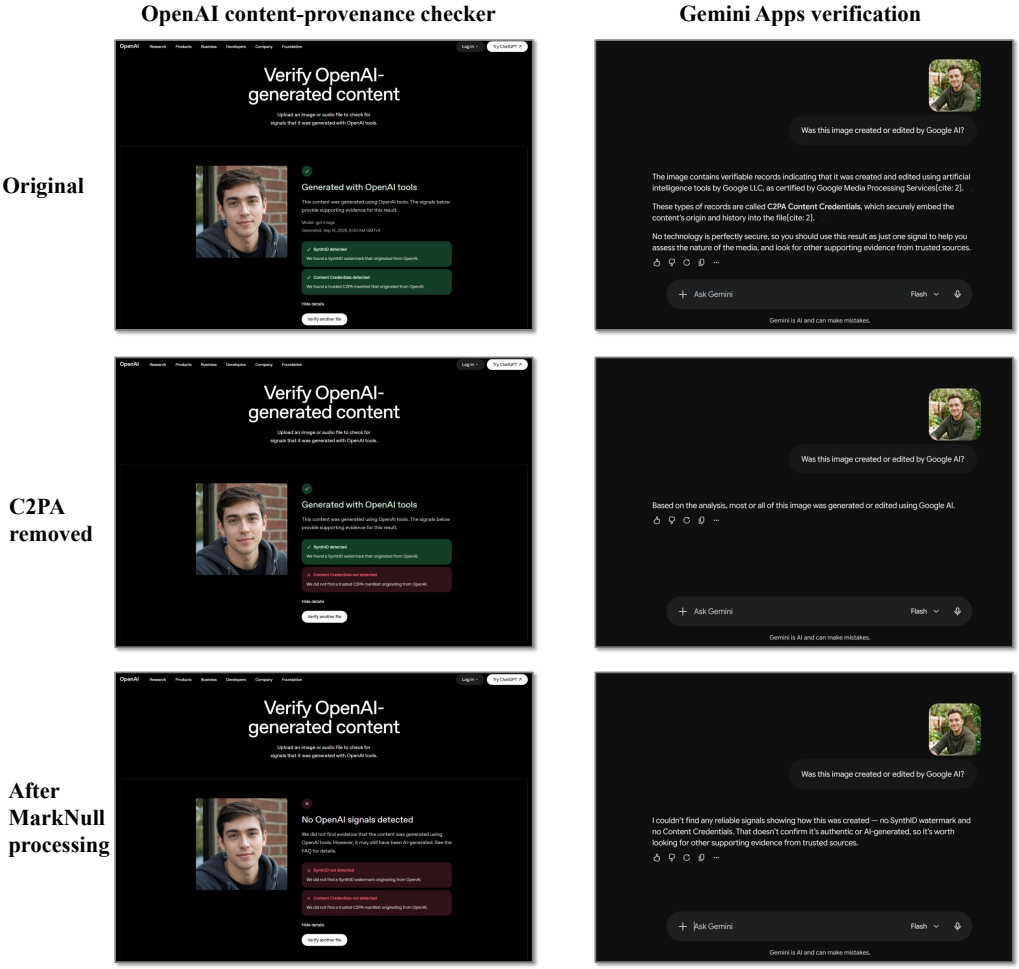

Figure 5: Representative provenance-verification outputs. The left column shows a GPT Image 2 example checked with OpenAI's content-provenance checker; the right column shows a Gemini Image example checked through Gemini Apps. Rows show the original image (A), C2PA removal (B), and MarkNull processing (C). The screenshots report the interfaces' responses for these two examples.

D.4 Results by Generation Source

Table 13 gives the numerical source-specific contrasts accompanying the main-text results. These marginal 95% intervals use hierarchical bootstrap resampling; the ten-test Holm correction applies to the overall comparisons. The opposite Gemini $C - B$ changes illustrate why the overall average alone is insufficient to describe processing sensitivity.

D.5 Control Analysis on Real Photographs

We select images with within-cell indices 1 and 2 from each of the 50 attribute combinations in the frozen real-image pool. The selection uses no detector outcomes. All 100 inputs have a resolution of 1024×1024 pixels, and the C2PA checks find no active manifest or C2PA claim. We apply the

Table 13: Source-specific changes in FAKE rate (percentage points). Brackets indicate marginal 95% bootstrap intervals; each source contains 100 parent images.

Source	Detector	C2PA removal ($B - A$)	MarkNull processing ($C - B$)
GPT Image 2	GPT	1.00 [−4.00, 5.67]	−2.67 [−11.67, 6.67]
	Gemini	−7.67 [−13.67, −1.33]	25.67 [14.66, 37.01]
	Qwen	−9.67 [−16.67, −2.00]	49.33 [41.00, 58.00]
	Seed	0.33 [−1.67, 2.67]	13.67 [7.67, 20.67]
	Grok	2.00 [−3.67, 8.00]	−7.67 [−14.34, −1.67]
Gemini Image	GPT	−1.00 [−5.67, 3.67]	−23.33 [−31.33, −15.66]
	Gemini	0.67 [−7.33, 8.67]	−17.67 [−26.68, −8.67]
	Qwen	0.00 [−7.67, 8.33]	11.67 [2.00, 20.67]
	Seed	1.67 [0.00, 4.00]	17.67 [10.99, 24.67]
	Grok	−1.67 [−7.67, 4.33]	−19.33 [−26.01, −12.99]

same MarkNull settings in Table 11 and obtain all 100 outputs. Each detector evaluates each image three times before and after processing. Table 14 reports the resulting descriptive FAKE rates.

Table 14: FAKE judgments on real-source images before and after MarkNull processing. Rates are percentages; changes are percentage points computed before rounding. Each detector evaluates 100 parents three times per condition.

Detector	Before	After	Change
GPT	4.33	15.67	+11.33
Gemini	24.67	60.00	+35.33
Qwen	5.67	62.33	+56.67
Seed	0.00	23.00	+23.00
Grok	0.67	5.00	+4.33

On real photographs, the processing-induced changes shift judgments toward AI-generated classifications across all five detectors.

E Statistical Analysis and Sensitivity Checks

E.1 Analysis Units and Resampling

We use 2,000 hierarchical bootstrap replicates for generated-image analyses. Each replicate samples the 50 prompts with replacement, then samples parent images within each prompt–source stratum with replacement, preserving its sample size. Each parent retains all detector responses, processing conditions, and repeated judgments available for that analysis. We report the 2.5th and 97.5th percentiles of the valid bootstrap estimates as 95% confidence intervals.

For the RQ1 false-positive rate, we resample real photographs and retain all five detector responses to each photograph. Balanced-accuracy intervals combine the resampled sensitivity and specificity. Retries contribute one valid response per evaluation task.

E.2 Same-Developer Estimands

Observed detection rates. The raw contrast subtracts the mean of the 20 off-diagonal detection rates from the mean of the five diagonal rates. Each cell receives equal weight within its group.

Adjusted association. We use the fixed-effects model in Equation 1, with GPT and GPT Image 2 as the reference levels. Let $\widehat{p}_{dg}(s)$ denote its predicted FAKE probability for detector d and source

g when the same-developer indicator equals s. The standardized difference is

$$\widehat{\Delta}_{\text{adj}} = \frac{1}{25}\sum_{d=1}^{5}\sum_{g=1}^{5}\left[\widehat{p}_{dg}(1) - \widehat{p}_{dg}(0)\right], \tag{2}$$

We multiply this difference by 100 to report percentage points; the odds ratio is $\exp(\widehat{\beta})$.

In this balanced design, both groups have the same marginal detector and source weights. The adjusted contrast uses model predictions for both indicator settings at each fixed detector–source combination, including settings absent from the observed pairings. It therefore describes a model-dependent association. Table 17 compares the raw and adjusted estimates.

E.3 Model Specification and Cell-Level Fit

We compare three fixed-effects specifications using the same 12,500 responses from the five matched sources. All models contain an intercept, detector effects, source effects, and a common same-developer term. The linear-probability model uses an identity link; the binomial models use logit and probit links. For the nonlinear models, we use the equal-cell standardization in Equation 2. The linear coefficient equals its standardized difference.

In the balanced design, each detector and source contributes equally to the diagonal and off-diagonal groups. After projecting out the intercept and row and column effects, the same-developer indicator has residual $S_{dg} - 1/5$. The least-squares coefficient therefore equals

$$\widehat{\beta}_{\text{LPM}} = \frac{1}{5}\sum_{d}\bar{Y}_{dd} - \frac{1}{20}\sum_{d\neq g}\bar{Y}_{dg}. \tag{3}$$

The nonlinear links impose different probability structures and produce similar negative standardized associations (Table 15). The linear fit yields probabilities from -0.195 to 1.104; we retain these unconstrained fitted values for the diagnostic comparison.

These checks are post hoc and retain the frozen original analysis in Table 17. We use 2,000 shared hierarchical resamples with fixed random-number initialization, retaining all five detector outcomes for each parent. All three models yield 2,000 valid fits. The new intervals use this common resampling run, while the original table retains its original bootstrap intervals.

Table 15: Specification sensitivity. Differences and cell errors are in percentage points. RMSE and maximum error summarize the 25 observed cells.

Model	Difference [95% CI]	Cell RMSE	Maximum error
Linear probability	-0.36 $[-2.67, 2.03]$	10.67	19.50
Logit	-8.12 $[-10.28, -5.93]$	7.20	16.63
Probit	-7.96 $[-10.12, -5.76]$	7.15	18.06

Cell-level residuals identify pairing structure beyond the fitted main effects and common diagonal term (Table 16). For example, GPT detects 69.2% of Seedream images, while the logit model predicts 52.6%. The maximum absolute residual is 16.63 percentage points under logit and 18.06 under probit. A saturated detector–source model fits the observed cells directly. Its 25 cell indicators also span the same-developer indicator, so an additional same-developer coefficient requires further constraints. Together, these diagnostics describe how the pooled summary depends on the assumed probability structure across the selected systems.

E.4 Firth, Developer Exclusion, and Boundary Checks

Firth analyses. We fit post hoc Firth models because the unpenalized developer-specific model exhibits separation. The pooled model retains the main-text specification; the developer-specific model replaces the common same-developer coefficient β with a source-specific coefficient β_g. For each source, we average its predicted probability differences equally over the five detectors and report $\exp(\widehat{\beta}_g)$ as its odds ratio.

Table 16: Observed detection rates and fitted values (%) for all 25 matched-source cells. The final column is observed minus logit-fitted rate in percentage points. Source labels identify the corresponding generator families.

Detector	Source	Observed	Linear	Logit	Probit	Residual
GPT	GPT	44.2	60.0	57.4	56.8	-13.16
GPT	Gemini	80.4	80.6	89.5	89.3	-9.11
GPT	Qwen	99.4	110.4	98.5	99.3	+0.88
GPT	Seed	69.2	52.6	52.6	51.1	+16.63
GPT	Grok	97.2	86.8	92.4	93.3	+4.76
Gemini	GPT	39.4	30.2	27.2	27.4	+12.24
Gemini	Gemini	52.0	50.1	41.9	41.6	+10.11
Gemini	Qwen	91.0	80.3	91.1	90.8	-0.06
Gemini	Seed	3.0	22.5	14.5	13.9	-11.50
Gemini	Grok	54.4	56.7	65.2	64.9	-10.79
Qwen	GPT	35.0	34.2	29.7	30.7	+5.32
Qwen	Gemini	52.6	54.5	59.6	58.9	-7.05
Qwen	Qwen	87.6	83.9	86.4	86.1	+1.18
Qwen	Seed	17.6	26.5	16.1	16.2	+1.49
Qwen	Grok	67.0	60.7	67.9	68.5	-0.95
Seed	GPT	0.0	-11.4	1.0	0.4	-1.03
Seed	Gemini	1.2	8.9	3.5	2.8	-2.30
Seed	Qwen	20.8	38.7	22.1	24.0	-1.26
Seed	Seed	0.0	-19.5	0.3	0.0	-0.26
Seed	Grok	9.8	15.1	4.9	4.9	+4.85
Grok	GPT	10.4	15.9	13.8	13.5	-3.37
Grok	Gemini	44.2	36.2	35.9	35.4	+8.35
Grok	Qwen	80.6	66.0	81.3	79.6	-0.74
Grok	Seed	0.4	8.2	6.8	5.6	-6.37
Grok	Grok	32.8	42.0	30.7	32.3	+2.13

We obtain point estimates with `pyfirth 0.0.6` and use a frequency-weighted implementation for bootstrap fits. An independent optimization of the same penalized likelihood agrees within 2×10^{-7} for every coefficient in both models. Both Firth models and the primary unpenalized pooled model yield 2,000 valid bootstrap fits. The unpenalized developer-specific model yields no valid bootstrap fits.

Table 17: Raw and adjusted same-developer associations. Differences are in percentage points; brackets indicate 95% hierarchical-bootstrap intervals. Firth analyses are post hoc.

Analysis	Contrast	Difference [95% CI]	Odds ratio [95% CI]
Observed	Diagonal vs. off-diagonal	-0.36 $[-2.74, 1.92]$	—
Primary logistic	Overall	-8.12 $[-10.17, -5.87]$	0.552 $[0.469, 0.652]$
Pooled Firth	Overall	-8.11 $[-10.19, -6.05]$	0.552 $[0.469, 0.647]$
Developer-specific Firth	GPT	-21.67 $[-24.88, -18.27]$	0.147 $[0.099, 0.208]$
	Gemini	4.81 $[-0.37, 9.93]$	1.342 $[0.978, 1.840]$
	Qwen	-5.35 $[-10.44, 0.08]$	0.614 $[0.393, 1.008]$
	Seed[†]	-11.14 $[-13.26, -8.34]$	0.243 $[0.180, 0.357]$
	Grok	-8.26 $[-14.93, -1.94]$	0.602 $[0.400, 0.887]$

[†] Zero-detection cells affect the Seed estimate and its interval. Developer-specific results are exploratory.

Developer exclusion. We remove each developer's detector and matching generation source in turn, refit the unpenalized pooled model, and standardize over the remaining 4×4 combinations. Each analysis yields 2,000 valid bootstrap fits. All estimates remain negative (Table 18), although each exclusion changes the set of systems that the contrast describes.

Table 18: Leave-one-developer-out estimates over the remaining 4×4 combinations. Differences are in percentage points; brackets indicate 95% hierarchical-bootstrap intervals.

Excluded pair	Adjusted difference [95% CI]	Odds ratio [95% CI]
GPT	-3.50 $[-6.22, -1.03]$	0.778 $[0.636, 0.929]$
Gemini	-10.79 $[-13.30, -8.33]$	0.388 $[0.302, 0.486]$
Qwen	-8.35 $[-10.80, -5.96]$	0.525 $[0.428, 0.635]$
Seed	-9.21 $[-12.20, -6.36]$	0.594 $[0.500, 0.696]$
Grok	-8.41 $[-11.19, -5.75]$	0.521 $[0.407, 0.647]$

Response recovery. We also exclude the 17 generated-image response slots that required additional recovery for the Seed detector, retaining all other responses. The supplementary pooled Firth estimate is -8.11 percentage points, with a 95% interval of $[-10.24, -6.07]$ and an odds ratio of 0.552 $[0.467, 0.646]$.

Boundary behavior. Seed records no detections for GPT Image 2 and Seedream images; both cells remain at zero in every bootstrap replicate. Firth yields finite coefficients, but empirical resampling preserves these zero-event cells. Intervals involving boundary cells therefore require caution; a zero-width interval does not establish an exact population rate.

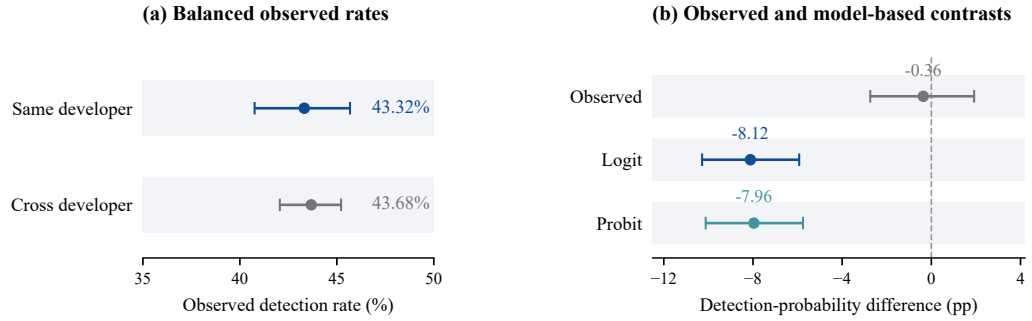

Figure 6: Observed and model-based same-developer comparisons over the five matched sources. (a) Mean detection rates for same-developer and cross-developer cells. (b) The observed diagonal–off-diagonal difference and standardized logit and probit differences. Bars indicate 95% hierarchical-bootstrap intervals: original intervals for the observed quantities and the shared supplementary run for logit and probit (Table 15).

E.5 Paired Contrasts and Multiple Comparisons

We use the parent-level $B - A$ and $C - B$ contrasts defined in the main text, weighting the two generation sources equally in the overall estimate. For source differences, we subtract the Gemini Image contrast from the GPT Image 2 contrast within each shared bootstrap replicate.

For each primary test, we divide the observed contrast by its bootstrap standard deviation and use a two-sided standard-normal test. Holm correction covers the ten overall comparisons: five detectors and two contrasts. The 95% percentile intervals are marginal, without simultaneous-coverage adjustment. Source-specific results are supplementary; the control analysis on real photographs reports descriptive changes separately.

Sign-flip sensitivity. We independently multiply each parent's paired difference by $+1$ or -1 with equal probability and recompute its mean over parents. Across 50,000 draws, we count b means at least as extreme in absolute value as the observed mean and calculate $p = (1 + b)/50{,}001$. This check assumes exchangeable parent-level signs and omits prompt-level dependence. After Holm correction over the same ten comparisons, it yields the same rejection decisions as the primary tests (Table 19).

Table 19: Holm-adjusted p values under the primary normal approximation and the paired sign-flip sensitivity analysis. Both methods use a ten-comparison family.

	C2PA removal ($B - A$)		MarkNull processing ($C - B$)	
Detector	Primary	Sign-flip	Primary	Sign-flip
GPT	1.000	1.000	< 0.001	< 0.001
Gemini	0.918	0.470	0.997	0.704
Qwen	0.565	0.212	< 0.001	< 0.001
Seed	0.918	0.470	< 0.001	< 0.001
Grok	1.000	1.000	< 0.001	< 0.001

We did not prespecify an equivalence margin, so nonsignificant C2PA results do not establish equivalence.

E.6 Source-Effect Contrasts and Multiplicity

The prespecified primary family contains the ten overall detector–processing comparisons in Table 3. The source-effect differences form a separate supplementary family: five detectors across $B - A$ and $C - B$. We apply a post hoc Holm correction to all ten bootstrap-normal two-sided p values in that family (Table 20). For each contrast, the source difference is the GPT Image 2 effect minus the Gemini Image effect. The original source-specific and source-difference intervals remain marginal 95% intervals. This additional correction yields adjusted $p < 0.05$ for the MarkNull source differences of GPT, Gemini, and Qwen.

Table 20: Post hoc source-effect multiplicity check. Differences are percentage points; Holm correction covers all ten rows. The contrasts concern between-source differences in the paired processing effects.

Detector	Processing	Difference	Raw p	Holm p
GPT	$B - A$	$+2.00$	0.582	1.000
GPT	$C - B$	$+20.67$	0.001	0.011
Gemini	$B - A$	-8.33	0.072	0.361
Gemini	$C - B$	$+43.33$	< 0.001	< 0.001
Qwen	$B - A$	-9.67	0.056	0.335
Qwen	$C - B$	$+37.67$	< 0.001	< 0.001
Seed	$B - A$	-1.33	0.383	1.000
Seed	$C - B$	-4.00	0.378	1.000
Grok	$B - A$	$+3.67$	0.391	1.000
Grok	$C - B$	$+11.67$	0.021	0.146

E.7 Uncertainty in Explanation Evaluation

For repeated explanations, resampling retains all three responses and all detector groups for each parent, preserving the dependence between within-group explanation pairs.

The 100 human-reviewed explanations cover 98 parent images; two parents each contribute explanations from two detectors. We keep all observations within an explanation together in both sensitivity analyses below, each using 2,000 replicates.

Stratified resampling samples explanations within detector–source strata, preserving detector totals and source quotas but sampling shared parents separately across detectors. **Parent-cluster resampling** samples parent identifiers and retains their associated explanations, preserving shared-parent dependence while allowing detector and source totals to vary. Table 21 reports observation-level intervals under both schemes. They describe uncertainty within the reviewed correct-detection sample.

Table 21: Human image-evidence proportions and 95% sensitivity intervals (%). Each resampling scheme retains all observations within a sampled explanation.

Label	Observed	Detector–source strata	Parent clusters
Supported	89.93	[86.71, 93.01]	[85.91, 93.73]
Contradicted	0.35	[0.00, 1.06]	[0.00, 1.09]
Indeterminate	9.72	[6.73, 12.80]	[6.03, 13.68]

F Explanation Analysis and Human Evaluation

F.1 Correctness and Probability Variation Across Repeats

Each group contains three responses from one detector to one of the 200 original generated images. Stable correct groups contain three FAKE labels, stable incorrect groups contain three REAL labels, and label-change groups contain both. These mutually exclusive outcomes partition the 1,000 groups into 217, 480, and 303 groups, respectively. Table 22 reports uncertainty for these proportions.

For each group, we calculate the sample standard deviation of its three reported probabilities with denominator $3-1$, and their maximum-minus-minimum range. We then average these measures over groups (Table 23). Intervals use 2,000 hierarchical resamples, retaining all detectors and repeated responses per sampled parent. The results characterize the fixed service configurations described in Appendix C.2.

Table 22: Repeated classification outcomes and marginal 95% hierarchical-bootstrap intervals (%). For Seed's zero observed stable-correct groups, we report the point estimate.

Detector	Stable correct	Stable incorrect	Label change
GPT	49.0 [40.0, 57.5]	22.5 [16.0, 29.5]	28.5 [21.5, 36.0]
Gemini	24.0 [17.0, 32.0]	32.0 [24.0, 40.0]	44.0 [36.0, 52.5]
Qwen	22.0 [15.5, 29.0]	27.0 [19.5, 35.0]	51.0 [42.5, 59.5]
Seed	0.0	99.0 [97.0, 100.0]	1.0 [0.0, 3.0]
Grok	13.5 [8.0, 19.0]	59.5 [51.5, 67.5]	27.0 [19.5, 35.5]
Overall	21.7 [18.2, 25.4]	48.0 [43.5, 52.5]	30.3 [26.9, 33.8]

Table 23: Probability variation within three-response groups. The table reports mean group-level SD and range with marginal 95% bootstrap intervals, and median group-level SD. Scores are on the $[0, 1]$ scale.

Detector	Mean SD [95% CI]	Median SD	Mean range [95% CI]
GPT	0.097 [0.084, 0.112]	0.063	0.180 [0.156, 0.206]
Gemini	0.105 [0.088, 0.124]	0.025	0.187 [0.156, 0.219]
Qwen	0.164 [0.140, 0.186]	0.221	0.296 [0.254, 0.336]
Seed	0.033 [0.028, 0.038]	0.029	0.060 [0.050, 0.069]
Grok	0.102 [0.083, 0.125]	0.040	0.183 [0.148, 0.223]
Overall	0.100 [0.092, 0.109]	0.050	0.181 [0.166, 0.197]

F.2 Cue Extraction and Category Definitions

We extract visual cues through predefined rules that identify objects, attributes, and descriptive states in each explanation. The rules normalize word forms and synonyms, attach descriptions to local objects, and retain the original text spans for inspection. Negation applies to the associated attribute or state; for example, "without AI artifacts" negates the artifact description while preserving preceding texture observations. General authenticity judgments, such as "looks fake," do not count as specific visual observations.

For category overlap, we combine the object or region with an attribute family. The object remains part of the category: skin texture and clothing texture have different keys. We remove duplicate

category keys within an explanation. The category metric ignores descriptive state, while strict matching retains the full object–attribute–state combination. For otherwise unspecified visual details, state terms determine the family where possible, including texture, naturalness, geometry, optical properties, and presence. Unmapped details retain a general visual-detail category.

Attribute families.

Surface texture: texture, pores, wrinkles, and regularity.
Detail structure: fine detail and hair strands.
Geometry: shape, proportion, and edge boundaries.
Optical properties: lighting and focus or blur.
Color: color tone.
Naturalness: realism and provenance descriptions.
Separate families: coherence, artifacts, quality, and text legibility each retain their own category family.

F.3 Metric Details and Extraction Coverage

The category Jaccard for a repeat group averages the three pairwise overlaps. For each group, the union of categories across all three explanations is the denominator for the proportions appearing once or in all three responses. We then average these proportions across groups, giving each group equal weight. A comparison between two empty sets is undefined; one empty and one nonempty set gives zero overlap. All explanations in the final analysis contain at least one extracted cue, so empty-set handling does not affect the reported results.

The parser identifies cues in 8,402 of 8,466 candidate text segments (99.2%) and in all 3,000 explanations. This coverage statistic records the proportion of candidate segments with a recognized cue. Strict object–attribute–state matching gives a mean Jaccard of 0.188, compared with 0.339 for category matching. Strict matching distinguishes attributes and states that category matching merges.

We check opposing descriptions only within shared categories. The predefined pairs cover presence, naturalness, consistency, texture, geometry, focus, symmetry, and plausibility. For these pair checks, both descriptions must have the same negation status. A matched opposing pair takes precedence over an unspecified state; otherwise, an unspecified state makes the comparison uncertain. Across 8,906 shared-category comparisons, the rules flag 259 opposing descriptions (2.9%) and 660 uncertain comparisons (7.4%). These counts summarize rule-flagged opposing and uncertain descriptions across repeated explanations. The brief prompt allows at most three observations, so category overlap also reflects which cues each response selects. A cue absent from one explanation contributes to lower overlap; opposing states require a separate comparison within a shared category.

F.4 Exploratory Coded Cue Overlap

Table 24: Exploratory rule-based cue overlap for 1,000 groups of three judgments. Category Jaccard and recurrence summarize the frozen coded representation.

Detector	Identical classifications (%)	Category Jaccard [95% CI]	In all three (%)
GPT	71.5	0.273 [0.248, 0.300]	13.5
Gemini	56.0	0.244 [0.222, 0.266]	12.0
Qwen	49.0	0.397 [0.369, 0.425]	25.8
Seed	99.0	0.375 [0.348, 0.401]	22.1
Grok	73.0	0.405 [0.385, 0.425]	26.3
Overall	69.7	0.339 [0.324, 0.354]	20.0

Under the frozen coding rules, category Jaccard averages 0.339. Across the union of categories in each group, 20.0% appear in all three responses and 57.9% appear once, averaged over groups. These statistics describe recurrence within the selected object–attribute representation.

F.5 Coded Cue Overlap by Classification Pattern

We stratify the 1,000 groups by their three labels, retaining the frozen rules and category mapping. Stable groups contain identical labels; flipped groups contain both labels. Table 25 separates stable groups into all-REAL and all-FAKE patterns. One observation can contain several coded categories under the three-observation, 40-word prompt.

Intervals use 2,000 hierarchical resamples, preserving all detector responses per parent and recomputing stratum denominators. These descriptive strata follow observed classification patterns and reflect their detector composition.

Table 25: Cue overlap by classification pattern. Words and categories are mean counts per response. All-stable combines the all-REAL and all-FAKE rows.

Pattern	Groups	Jaccard [95% CI]	Words	Categories
All-stable	697	0.371 [0.353, 0.389]	24.55	6.02
All-REAL	480	0.415 [0.396, 0.432]	22.75	6.59
All-FAKE	217	0.276 [0.250, 0.304]	28.54	4.76
Flipped	303	0.263 [0.248, 0.279]	27.66	6.06

Stable groups have higher pooled category overlap (0.371) than flipped groups (0.263). Within stable groups, overlap is 0.415 for all-REAL judgments and 0.276 for all-FAKE judgments. These descriptive differences compare label patterns with overlap in the coded categories. For Seed, 198 groups contain three REAL labels and two groups contain one FAKE label each. Its FAKE share across the 600 original-image responses is 0.33%. Table 26 reports the detector-level breakdown.

Table 26: Within-detector classification patterns and category overlap. Parenthetical counts are repeat groups. For Seed's two flipped groups, we report the point estimate.

Detector	Stable Jaccard [95% CI]	Flipped Jaccard [95% CI]	FAKE share (%)
GPT	0.305 [0.272, 0.339] (143)	0.193 [0.169, 0.217] (57)	63.00
Gemini	0.278 [0.249, 0.307] (112)	0.200 [0.178, 0.224] (88)	45.33
Qwen	0.480 [0.438, 0.522] (98)	0.317 [0.293, 0.341] (102)	48.00
Seed	0.376 [0.351, 0.402] (198)	0.183 (2)	0.33
Grok	0.428 [0.406, 0.453] (146)	0.341 [0.308, 0.377] (54)	25.50

F.6 Rule Development and Checks

Two reviewers inspect four repeat groups per detector, with two groups from each generation source. They check extraction errors, missed observations, and the proposed comparisons across repeats. The adjudicated extraction labels comprise 127 correct, 54 incorrect, and three uncertain cases out of 184 observations. For the 20 group-level comparisons, the adjudicated labels comprise five correct, five false positives, and ten uncertain cases.

The checks inform revisions to clause boundaries, object assignment, word-form matching, and negation scope. We then recompute the full analysis with the revised rules. These results document the development check. The final rules have no independent human test set. We use their overlap and opposing-description counts as exploratory summaries of the coded representation.

F.7 Human-Evaluation Sampling and Annotation

We sample explanations accompanying correct RQ1 detections before reviewing their evidence quality. Sampling allocates 20 explanations to each detector and seeks coverage across available sources and prompts. GPT, Gemini, Qwen, and Grok each contribute two or three explanations per source across all seven sources. Seed has no eligible correct detections for GPT Image 2 or Seedream images, so we sample four explanations from each of the remaining five sources.

Reviewers inspect the original image and explanation, then label each specific observation as supported, contradicted, or indeterminate. The interface presents complete observations and allows

reviewers to add observations that the parser missed. It hides detector and source identities, truth and prediction labels, and the other reviewer's answers. The annotation unit preserves enough text to interpret the claim; it can contain several normalized components used by the automatic analysis. The study lead confirms the adjudicated labels after independent review, while retaining both reviewers' original labels.

Annotation criteria.

Supported. Visible image evidence supports the observation.

Contradicted. Visible image evidence clearly conflicts with the observation.

Indeterminate. The detail is unclear, the description is too subjective, or the image provides insufficient evidence.

Provenance or metadata claims require corresponding evidence visible in the review interface; otherwise reviewers mark them indeterminate. The task assesses image support for stated observations, without determining whether those observations drove the authenticity judgment.

F.8 Label-Specific Agreement and Existing Disagreements

We recover the paired pre-adjudication labels for all 288 observations from the existing review records. Table 27 reports their full cross-tabulation. For label k, category-specific agreement is $2n_{kk}/(n_{k\cdot} + n_{\cdot k})$, where the row and column totals count each reviewer's uses of that label. The values are 95.0% for supported, 15.4% for contradicted, and 28.6% for indeterminate observations. The latter two labels have small marginal counts, as the matrix shows. Of the 29 disagreements, 27 involve a supported label from one reviewer.

Table 27: Pre-adjudication labels for the existing 288 observations. Rows are reviewer A; columns are reviewer B.

	Supported	Contradicted	Indeterminate
Supported	254	1	9
Contradicted	8	1	1
Indeterminate	9	1	4

Examples from the existing adjudication. We select the first task–observation entry in each of three disagreement label pairs to illustrate the recorded boundaries. We retain the original observations, reviewer labels, and adjudications.

Example 1. "The face is unusually symmetric and uniformly polished." Reviewer A labels this indeterminate, and reviewer B labels it supported. The existing adjudication assigns supported. The recorded rationale states: "The face is visibly near-symmetric and evenly polished enough to support the reported visual cue."

Example 2. "and the blurred shelving and plant in the background appear amorphous rather than optically natural." Reviewer A labels this contradicted, and reviewer B labels it supported. The existing adjudication assigns indeterminate. The recorded rationale states: "Shelving and plant blur is visible, but the distinction between natural optical blur and amorphous rendering is unresolved."

Example 3. "Facial texture is unusually uniform and hyper-detailed, especially across the cheeks and forehead." Reviewer A labels this supported, and reviewer B labels it indeterminate. The existing adjudication assigns indeterminate. The recorded rationale states: "The face has visible natural detail, but the degree of unusual uniformity is subjective at this resolution."

These examples locate disagreement around descriptive degree and the interpretation of visible texture or blur. The agreement analysis reuses the original review records and preserves the adjudicated image-support totals.

F.9 Reviewer Agreement and Adjudicated Results

The paired labels in Table 27 yield 89.9% agreement, 86.5% expected agreement, and Cohen's $\kappa = 0.253$.

Table 28 provides the adjudicated counts by detector. At the explanation level, 97 explanations contain a supported observation, one contains a contradicted observation, and 23 contain an indeterminate observation. These explanation-level categories overlap. Detector-level results describe the sampled source composition and are not used to rank detectors.

Table 28: Adjudicated observations by detector. Each detector contributes 20 explanations. Percentages use its observation count as the denominator.

Detector	Observations	Supported	Contradicted	Indeterminate
GPT	59	48 (81.4%)	0 (0.0%)	11 (18.6%)
Gemini	45	38 (84.4%)	0 (0.0%)	7 (15.6%)
Qwen	76	71 (93.4%)	0 (0.0%)	5 (6.6%)
Seed	51	49 (96.1%)	1 (2.0%)	1 (2.0%)
Grok	57	53 (93.0%)	0 (0.0%)	4 (7.0%)
Overall	288	259 (89.9%)	1 (0.3%)	28 (9.7%)

G Limitations and Future Work

Our evaluation focuses on facial images with controlled attributes and the selected generation sources and detectors. Extending the evaluation to broader image types and out-of-distribution settings would help assess whether the observed patterns persist. Continued evaluation across model versions would also track changes in detection behavior as models and provenance technologies evolve.

MarkNull changes visual content alongside provenance-related signals. The current comparisons leave the relative contributions of watermark changes and processing-induced visual changes unresolved. Future work should combine more content-preserving processing with independent provenance verification to better distinguish the roles of visual features, file-level credentials, and pixel-level signals.

Our RQ3 analysis examines classification repeatability and image support for observations accompanying correct detections, alongside exploratory rule-based cue summaries. Human evaluation of incorrect judgments would extend the explanation evidence.