

Explainable AI-Generated Image Detection RewardBench

Michael Yang¹, Shijian Deng², William T. Doan²
Kai Wang³, Tianyu Yang⁴, Harsh Singh⁵, Yapeng Tian²

¹Columbia University ²The University of Texas at Dallas

³University of Toronto ⁴University of Notre Dame ⁵Stony Brook University

mly2123@columbia.edu {shijian.deng,doan,yapeng.tian}@utdallas.edu

kaikai.wang@mail.utoronto.ca tyang4@nd.edu singh17@cs.stonybrook.edu

Abstract

Conventional, classification-based AI-generated image detection methods cannot explain why an image is considered real or AI-generated in a way a human expert would, which reduces the trustworthiness and persuasiveness of these detection tools for real-world applications. Leveraging Multimodal Large Language Models (MLLMs) has recently become a trending solution to this issue. Further, to evaluate the quality of generated explanations, a common approach is to adopt an "MLLM as a judge" methodology to evaluate explanations generated by other MLLMs. However, how well those MLLMs perform when judging explanations for AI-generated image detection generated by themselves or other MLLMs has not been well studied. We therefore propose **XAIGID-RewardBench**, the first benchmark designed to evaluate the ability of current MLLMs to judge the quality of explanations about whether an image is real or AI-generated. The benchmark consists of approximately 4,000 annotated triplets sourced from various image generation models and MLLMs as policy models (detectors) to assess the capabilities of current MLLMs as reward models (judges). Gemini 2.5 Flash achieves the highest four-way accuracy at 68.8%, while Qwen2.5-VL-72B-Instruct achieves the highest Both-Clear accuracy at 86.4%; these results show that current judges remain imperfect. Our benchmark also features a 59.8% raw four-way agreement and 87.4% clear-winner conditional agreement between human annotators. In addition, we provide an analysis of common pitfalls that these models frequently encounter. Code and benchmark are available at <https://github.com/RewardBench/XAIGID-RewardBench>.

1 Introduction

AI-generated and modified images are becoming increasingly prevalent, causing widespread damage to information ecosystems and public trust. Thus, the development of autonomous systems to detect such images is crucial. However, current methods typically provide a classification decision without the kind of explanation or justification that a human expert can offer, which limits the persuasiveness and trustworthiness of their outputs. To address this, an ideal model should not only perform binary classification but also analyze the image comprehensively, leveraging logic, fine-grained details, real-world knowledge, and even tool use to determine authenticity and articulate its reasoning with supporting evidence.

Multimodal Large Language Models (MLLMs) offer a promising approach to generating such nuanced explanations. By processing visual details, an MLLM can investigate an image's authenticity through a deductive process, much like a detective examining a crime scene. However, these explanations generated by models are not always accurate; challenges such as hallucinations and faulty reasoning remain significant obstacles. To mitigate these issues in a scalable way, a reward model "judge" can be implemented to evaluate the validity of a generated explanation, either during training or inference.

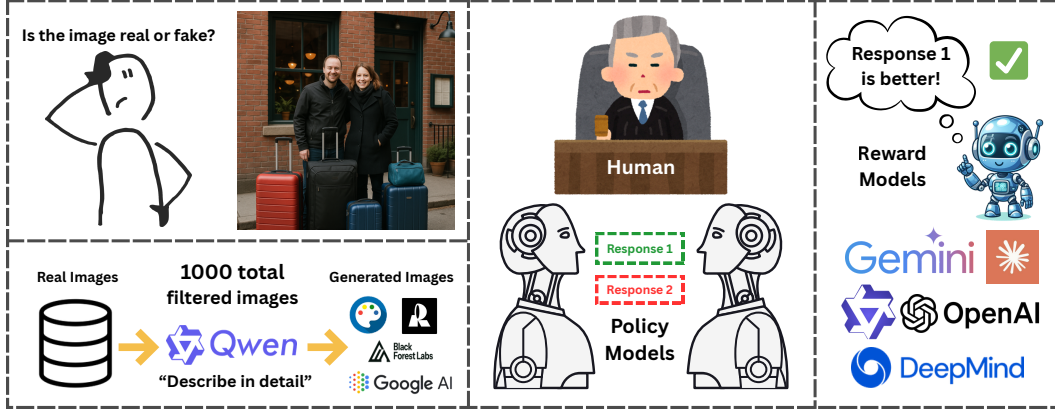


Figure 1: Overview of the XAIGID-RewardBench. With the capabilities of today’s MLLMs, an AI-generated image detection response can include both a verdict and an explanation, providing much more insight into the reasoning process. Furthermore, MLLMs are also now being used to judge the quality of these explanations, but little work has been done to assess the quality of the MLLM judges themselves. Therefore, we present a human-annotated benchmark to evaluate how well these reward models (i.e., MLLM judges) score the explanations provided by MLLMs for AI-generated image detection.

This raises a critical research question: *how is the judge itself validated?* If such a judge is to be reliable, its own performance must be rigorously benchmarked. To that end, we present the **Explainable AI-Generated Image Detection (XAIGID) RewardBench**. This benchmark contains approximately 4,000 triplets, each consisting of an image and two competing explanations of its authenticity (Figure 1). It is designed to evaluate reward models (i.e., “judges”) on their ability to determine the better response. These annotations were carefully curated and manually verified, adhering to a strict rubric to minimize bias. This benchmark can help identify which MLLMs are most suited to act as reward models for judging explanations from AI-generated image detectors.

Our primary contributions are as follows: (1) We are the first to conduct a systematic study of current MLLMs as both policy models and reward models within the context of explainable AI-generated image detection. (2) We introduce the first reward model benchmark designed to gauge MLLMs’ performance on the novel task: Judging Explainable AI-Generated Image Detection. (3) We provide meaningful insights into the capabilities and common failure modes in explainable AI-generated image detection.

2 Related works

2.1 MLLMs for explainable AIGC detection

Early AI-generated content (AIGC) detectors emphasized binary decisions with limited interpretability. Recent work has turned to MLLMs to couple detection with explanation and, in some cases, grounding. LEGION (Kang et al., 2025) introduces SynthScars, pixel-level artifact masks with textual rationales, and an MLLM pipeline that localizes artifacts and verbalizes why they indicate synthesis. FakeXplained (Ji et al., 2025b) fine-tunes MLLMs to align rationales with evidence via detection, box-level grounding, and coherent textual explanations. MLLM-DEFAKE (Ji et al., 2025a) benchmarks off-the-shelf MLLMs against dedicated detectors and humans, proposing a reasoning-driven pipeline for explainable decisions. ForenX (Tan et al., 2025) adds a “forensic prompt” and an explanation resource (ForgReason) to steer MLLMs toward human-aligned forensic cues. FakeShield (Xu et al., 2025) further formalizes explainable image forgery detection and localization for MLLMs, coupling authenticity decisions with pixel-level masks and textual justifications. In parallel, RADAR (Wang et al., 2025a) targets semantic fakes that violate world knowledge, pairing bounding boxes with textual descriptions. For the video modality, ExDDV (Hon-

dru et al., 2026) provides 5.4K real/deepfake clips with human text rationales and click supervision, showing that both textual and spatial supervision benefit explainable VLMs. Complementing these advances, Forensics-Bench (Wang et al., 2025b) offers a comprehensive LVLM evaluation suite spanning multiple modalities and tasks (binary decisions, localization, attribution), highlighting the performance gap that remains for current models. Orthogonally, LVLM-DFD (Yu et al., 2025) improves generalization and explanation via a Knowledge-guided Forgery Detector and a Forgery Prompt Learner that align forensic patterns with LVLM reasoning. Altogether, these works establish that MLLMs can produce human-readable justifications and sometimes explicit evidence (boxes/masks); however, they only evaluate policy models, not the reliability of reward models (judges) for evaluating explainable AIGC detection.

2.2 Reward models and benchmarks

Reinforcement learning has recently been explored for MLLMs in AIGC detection, but primarily with rule-based reward functions. Veritas (Tan et al., 2026) trains an MLLM detector with Mixed Preference Optimization and pattern-aware GRPO to shape planning/self-reflection during reasoning about forgeries, improving OOD generalization on the HydraFake dataset. FakeXplained (Ji et al., 2025b) uses GRPO-style rewards on verdict correctness, grounding quality, and rationale format after SFT. For judging explanations, previous work has either relied on direct human judgment or used a popular MLLM as a judge, without systematically evaluating the MLLM judges themselves with a standardized reward-model benchmark in the AI-generated image domain. General reward-model benchmarks exist: RewardBench (Lambert et al., 2025) and its update RewardBench 2 (Malik et al., 2026) assess textual reward models across instruction following, reasoning, and safety; VL-RewardBench (Li et al., 2025) and Multimodal RewardBench (Yasunaga et al., 2025) extend this evaluation to VLM judges across perception, hallucination detection, reasoning, knowledge, and safety. These resources demonstrate broad trends for reward models but do not probe forensic explanation quality or artifact-grounded reasoning on AI-generated images. To our knowledge, no prior work provides a domain-specific reward-model benchmark for judging explainable AI-generated image detection. XAIGID-RewardBench fills this gap by supplying human-verified, 4-way-annotated triplets (image, two explanations) that directly measure how well MLLMs act as judges for explainable AI-generated image detection.

3 XAIGID-RewardBench

We first document the entire creation process of our XAIGID-RewardBench, including image dataset synthesis, response generation with MLLMs as AI-generated image detectors, and RewardBench annotation. Then we formalize the novel task: Judging Explainable AI-Generated Image Detection.

3.1 Image dataset

For real images, we used COCO-2017 (Lin et al., 2014). We assume "real" images do not change dramatically over time, so using a classic dataset is sufficient for providing non-AI-generated images. We sampled 500 random images from COCO-2017 and then cropped them to a square resolution.

To cover a diverse range of AI-generated images, we selected 10 representative image generation models to generate 500 images as the other half of our image dataset. We selected 10 representative image generation models, listed below, emphasizing recent SOTA models while also including several classic ones.

Image generation models. The generation models used in this study include Stable Diffusion 1.5 (Rombach et al., 2022), Stable Diffusion 3.5 (Esser et al., 2024), Imagen 3 (Baldridge et al., 2024), Imagen 4 (Google DeepMind, 2025), Glide (Nichol et al., 2022), GPT-Image-



Figure 2: A visual depiction of the input (a triplet) for MLLM judges. Both responses come from SOTA MLLMs as AI-generated image detectors, yet they disagree with each other. An MLLM judge must deduce which one is more convincing given all 3 pieces of information (One image and two detection responses from detectors). MLLM judges currently still cannot achieve human performance on the stated task.

1 (OpenAI, 2025a), Recraft V3 (Recraft, Inc., 2024), Flux.1 Schnell, Flux.1 Dev (Labs, 2024), and Bagel (Deng et al., 2025).

We generated 500 images in total, 50 per model. To ensure the topic distribution of generated images was similar to that of the real images, we used Qwen2.5-VL-3B-Instruct to generate basic, unbiased, and detailed captions of the real images. From there, we partitioned the 500 captions into ten groups of 50. Each group of captions was then used to generate 50 images with one of the ten generation models. This approach allows for pairing real images with generated images derived from similar descriptions, making the detection task more diverse and challenging.

3.2 Policy models’ detection response with explanation

We utilized several state-of-the-art MLLMs as AI-generated image detectors to generate responses. In total, we employed eight different models to provide a diverse pool of answers: Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct (Bai et al., 2025), Gemini 2.5 Flash, Gemini 2.5 Pro (Comanici et al., 2025), OpenAI o3 (OpenAI, 2025b), GPT-4o (Hurst et al., 2024), GPT-5.5 (OpenAI, 2026), and gemma-3n-E4B-it (Team, 2025).

Each model was given the image and asked to determine the image’s authenticity, followed by an explanation. We used the following prompt: *Start with one word, "real" or "fake", then explain your judgment. Use only visible image evidence: lighting, shadows, texture, geometry, anatomy, text, perspective, reflections, and compression or generation artifacts. Do not use filenames, paths, metadata, labels, or outside knowledge.*¹

Results. We calculated the accuracy of these models against our dataset to gauge their raw competence at detecting AI-generated images (Table 1). A source-specific breakdown is provided in Appendix Table 4.

¹Due to API safety policies, some responses from OpenAI were refused, but the vast majority of responses yielded long, rich answers, all at least one paragraph long and often multiple paragraphs.

<i>Proprietary</i>				<i>Open source (Qwen rows: Instruct)</i>			
Policy model	Overall	Real	Fake	Policy model	Overall	Real	Fake
GPT-5.5	92.8%	92.2%	93.4%	Qwen2.5-VL-3B	64.7%	91.0%	38.4%
o3	84.5%	97.4%	71.6%	gemma-3n-E4B-it	57.8%	97.6%	18.0%
Gemini 2.5 Pro	77.4%	99.0%	55.8%	Qwen2.5-VL-7B	51.7%	97.4%	6.0%
GPT-4o	73.5%	61.6%	85.4%				
Gemini 2.5 Flash	71.8%	96.8%	46.8%				

Table 1: Accuracy of MLLMs as policy models (AI-generated image detectors). Performance on detecting AI-generated images is high among proprietary models, and both o3 and GPT-5.5 exhibit high accuracy and capability on this task. A stronger MLLM generally performs better on AI-generated image detection tasks. However, Qwen2.5-VL-3B-Instruct outperforms its 7B counterpart; we hypothesize this is because the 7B model’s increased capability leads to overconfidence. While the 3B model correctly identifies real images, its performance on fake images is close to random chance. Conversely, the 7B model exhibits more complex reasoning, but this often leads it to be confidently incorrect.

Pairing. We then created 4 triplets per image for a total of 4,000 initial triplets. Pairing was conducted randomly in a round-robin fashion to ensure equal distribution. The triplets were then scrambled and sent for annotation (Figure 2).

3.3 RewardBench annotation

Objectively deciding which response is better can be challenging for annotators. They may be affected by many personal biases, even subtle hidden ones. To combat this, we established a general rubric derived from our own qualitative analysis of MLLM responses. More details are provided in Section A.

Annotation. To perform annotation, annotators were given the target image and the two responses. They were also presented with the original prompt to understand what the model was responding to. Lastly, they were allowed to make one of four selections to annotate:

- *"First": To denote that the first response was superior.*
- *"Second": To denote that the second response was superior.*
- *"Tie": Both responses were equally good in quality and not completely bad.*
- *"Both Bad": Both responses provided unsatisfactory explanations.*

Filtering. When multiple annotators label the same triplet, conflicts can arise. Conflicts can be resolved by majority vote. If ties are encountered, they can be resolved manually. Additionally, any annotations where too little time was spent (e.g., annotations completed in an unusually short time) can also be filtered out.

In a side experiment, we also evaluated which models generated the most favorable responses using an Elo system. GPT-5.5 achieved the highest rating, followed by Gemini 2.5 Pro and Gemini 2.5 Flash; the full ranking is provided in Appendix Table 5.²

We also analyzed the choice distribution to check for potential annotation bias, such as choosing Response 1 solely because it appeared first. Our data shows that our annotations were unbiased by the relative positions of the responses (Figure 3).

Inter-human agreement. Disparities in human preferences are natural. To maximize the effectiveness of the rubric, we conducted an experiment where two annotators were given a previously unseen batch and asked to annotate all 100 triplets. After doing so, we compared their annotations to evaluate the rubric’s effectiveness in aligning human preferences.

²Tie and Both Bad outcomes were excluded when calculating Elo ratings.

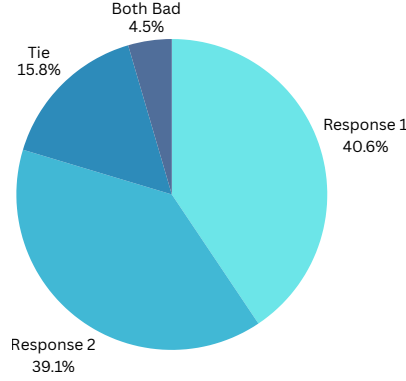


Figure 3: Distribution of annotator choices. Response 1 was selected 1.5 percentage points more often than Response 2, indicating little evidence of position bias in the human annotations.

In this separate 100-triplet calibration experiment, the two human annotators agreed on 68.0% of all annotations and 98.3% on cases where both annotators selected a clear winner (considering only annotations without *Tie* or *Both Bad*).

Overall, these statistics show that our rubric was effective at aligning the annotations in a vast majority of cases. The reason for the large disparity between 4-way agreement and clear-winner agreement could be that annotators have differing thresholds for "high-quality" vs. "low-quality." In a case where both responses in a triplet may be "low-quality," one annotator might mark the triplet as "Both Bad," while another might decide that one is slightly better than the other. Regardless, a 98.3% agreement for annotations with a hard preference shows that annotators were still able to agree in most cases on which response was better.

Benchmark statistics. After automated validation, we retain 3,988 clean preference triplets from 4,000 candidates, removing 12 triplets with malformed or failed policy-model generations. The release covers 500 COCO-2017 real images and 500 synthetic images from ten generators. We collect 4,412 annotations from 11 annotators, covering 99.27% of the clean triplets. Among multiply annotated items, pairwise agreement is 59.78% under the full four-way label space and 87.39% in the clear-winner conditional view. We report both figures because the conditional view does not measure disagreements involving *Tie* or *Both Bad*. We’ve also provided a table in the appendix comparing our benchmark size to other related benchmarks (Table 22).

3.4 Task: Judging Explainable AI-Generated Image Detection

The goal of the task is to evaluate how well MLLMs, acting as reward models, can judge explanations generated by MLLMs as detectors in AI-generated image detection.

Let $\mathcal{G}$ denote a set of image generation models. Each $G \in \mathcal{G}$ induces a distribution over images given a caption c and randomness z , written $I = G(c, z)$. Let $\mathcal{I}_{\text{real}}$ be a set of photographs sampled from a real-image distribution, and let $\mathcal{I}_{\text{gen}} = \{G(c, z) : G \in \mathcal{G}\}$ be generated images. We define $\mathcal{I} = \mathcal{I}_{\text{real}} \cup \mathcal{I}_{\text{gen}}$ and a ground-truth authenticity label $l : \mathcal{I} \rightarrow \{0, 1\}$ where $l(I) = 1$ for real and $l(I) = 0$ for AI-generated.

A policy model (MLLM detector) $\pi \in \Pi$, given a policy prompt p_p maps an image to a structured response

$$r = \pi(I, p_p) = (y, e), \quad y \in \{0, 1\}, e \in \mathcal{E},$$

where y is the one-word verdict and e is a free-form explanation. Detector (policy) accuracy is

$$\text{Acc}_{\text{det}}(\pi) = \frac{1}{|\mathcal{I}|} |\{I \in \mathcal{I} : y(\pi(I, p_p)) = l(I)\}|.$$

For reward modeling (judging), a triplet t consists of a single image and two competing policy responses to the same prompt p_p : $t = (I, r_a, r_b)$ with $r_a = (y_a, e_a)$ and $r_b = (y_b, e_b)$. A judge (MLLM as a reward model) $J \in \mathcal{J}$ prompted with p_j : *Between the two responses, the original prompt, and the image, which response is better?* outputs a 4-way preference

$$J(t, p_p, p_j) \in \mathcal{Y}, \quad \mathcal{Y} = \{\text{First, Second, Tie, BothBad}\},$$

indicating whether the first response is better, the second is better, both are comparably good, or both are unsatisfactory.

A human judge H provides labels in the same space $\mathcal{Y}$ and produces the ground-truth preference

$$H(t, p_p, p_j) \in \mathcal{Y}.$$

Let $\mathcal{T}$ be the set of annotated triplets with ground-truth labels H_t . The judge’s 4-way accuracy is

$$\text{Acc}^{(4)}(J) = \frac{1}{|\mathcal{T}|} |\{t \in \mathcal{T} : J_t = H_t\}|.$$

Let $\mathcal{C} = \{\text{First, Second}\}$ denote the clear-winner labels. Human-Clear accuracy is evaluated on $\mathcal{T}_{\text{HC}} = \{t \in \mathcal{T} : H_t \in \mathcal{C}\}$, retaining non-clear judge predictions as errors:

$$\text{Acc}_{\text{HC}}(J) = \frac{1}{|\mathcal{T}_{\text{HC}}|} |\{t \in \mathcal{T}_{\text{HC}} : J_t = H_t\}|.$$

Both-Clear accuracy is conditioned on both the human and judge selecting a clear winner. Define $\mathcal{T}_{\text{BC}}(J) = \{t \in \mathcal{T} : H_t \in \mathcal{C}, J_t \in \mathcal{C}\}$; then

$$\text{Acc}_{\text{BC}}(J) = \frac{1}{|\mathcal{T}_{\text{BC}}(J)|} |\{t \in \mathcal{T}_{\text{BC}}(J) : J_t = H_t\}|.$$

4 Experiment and analysis

We performed experiments based on our newly created benchmark. This allowed us to evaluate how current state-of-the-art MLLMs perform as judges when evaluating explainable AI-generated image detection.

4.1 Experimental setup

We reused all seven policy models that we used to generate initial responses as judges along with 2 additional models: Qwen 2.5-VL-72B-Instruct and Claude Opus 4.7 (Anthropic, 2026). We selected the 72B model to compare open-source scaling while we used Opus 4.7 as it is a recently released and highly powerful proprietary model. They were utilized as described in Section 3.4. Each model was presented with all annotated triplets, asked to respond with their preference among the four options followed by an elaboration, and then scored based on their alignment with human preferences.

4.2 Results

To evaluate each model’s performance as a judge, we calculated the accuracy it achieved when compared with human annotations. The results are presented in Table 3. These accuracies shine a light on the ability of today’s SOTA models to act as reward models for AI-generated images. While these models are able to identify high-quality responses to an extent, their ability to perform this judge role still has room for improvement. No model achieved an accuracy over 90% on this benchmark, which reveals the gap between human and current MLLM judges (Table 3).

4.3 Analysis

Beyond just analyzing qualitative results, we found multiple factors that can play a role in an explanation’s quality. This is crucial for the development of both policy and reward models, and these findings can also provide better insight on how to construct a more thorough rubric or standard for judging explainable AI-generated image detection.

Metric (evaluated model)	What It Measures
Detector Accuracy (policy model)	Correct real/fake classification of an image
Four-Way Judge Accuracy (reward model)	Agreement with human labels: First, Second, Tie, or Both Bad
Human-Clear Accuracy (reward model)	Subset of the benchmark where humans selected a clear-winner
Both-Clear Accuracy (reward model)	Subset where both humans and judge model selected a clear-winner
DPO Win Rate (fine-tuned policy model)	Preference for the fine-tuned model over its base model

Table 2: Clarification of the evaluation metrics used in this work. Each metric evaluates a distinct model capability and should not be compared directly.

Proprietary (%)				Open source (%; Qwen rows: Instruct)			
Reward model	Four-Way	Human-Clear	Both-Clear	Reward model	Four-Way	Human-Clear	Both-Clear
GPT-4o	67.99	83.97	85.01	Gemma 3n	66.25	80.88	81.81
o3	63.92	75.15	79.81	Qwen2.5-VL-3B	37.51	32.69	75.52
Gemini 2.5 Flash	68.78	84.03	85.37	Qwen2.5-VL-7B	33.19	25.13	71.67
Gemini 2.5 Pro	68.02	78.79	86.29	Qwen2.5-VL-72B	61.58	66.09	86.43
Claude Opus 4.7	65.24	79.22	84.41				

Table 3: Performance of MLLMs as reward models on the clean benchmark. *Four-Way* measures exact agreement with the human label among *First*, *Second*, *Tie*, and *Both Bad*. *Human-Clear* evaluates only triplets where the human selected a clear winner, while retaining neutral model predictions as errors. *Both-Clear* is conditioned on both the human and judge selecting a clear winner and corresponds to the legacy two-way evaluation. Four-way accuracy is our primary metric.

Even one mistake should qualify an image as AI-generated. Frequently, a model correctly identifies anomalies within an image (Figure 7). However, the model assumes that because the rest of the image looks fine, the image is probably real and that the anomaly was just a "hiccup." In reality, authentic images do not have such "hiccups." It only takes one piece of evidence to show that an image is fake.

Simply stating a detail doesn't provide any convincing argument. Models occasionally state something true about the image (Figure 4a). While their statement about the image is true, it does not prove anything. Something needs to be stated about the mentioned detail, or why the existence of that detail proves that an image is real or fake. This shortcoming highlights the gap between the ability of current SOTA models to identify and then reason.

Models tend to generalize details in media. Instead of looking into an area that might cause suspicion, a model might just skip over it and say it "looks realistic" (Figure 4b). By doing so, they miss the key anomaly that proves why an image is fake. Humans still hold the ability to look at an image, identify which areas might be suspicious, and then further scrutinize that region for clues. A current known issue is that models commonly hallucinate details. A more pertinent issue is that models are not identifying everything that they should be.

Longer responses are not equivalent to better answers. During annotation, we instructed annotators to look beyond just length and pursue responses that held real, meaningful arguments. For example, Gemini is known to present larger, more detailed arguments, and while these responses are usually better, this is not always the case. Hallucination, meaningless points (Figure 2), and poor analysis of details can sometimes occur in long responses. In fact, a shorter response that correctly spots all the anomalies is better. A

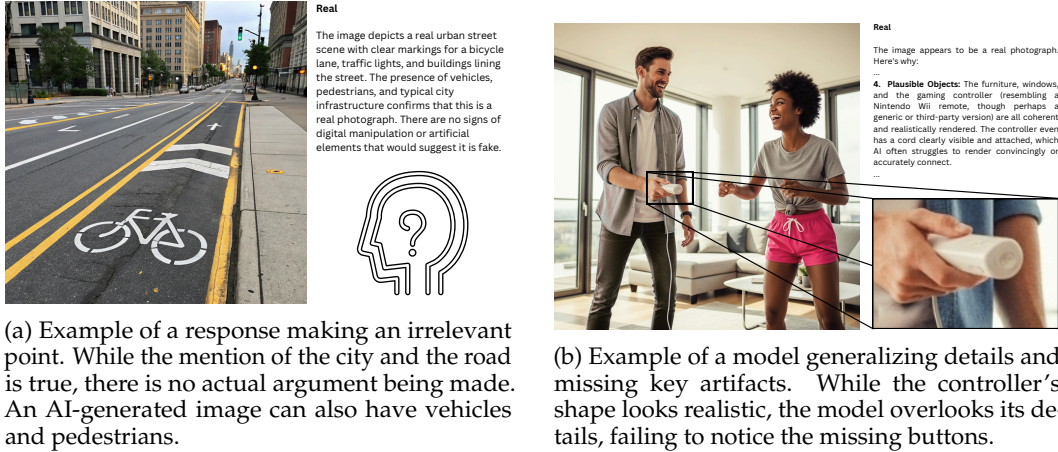


Figure 4: Some failure cases of MLLM policy models.

shorter response that accomplishes more is preferable to a long response that accomplishes the same level of reasoning.

Response-length effects. To quantify the association between response length and judge preference, we measured how often each judge selected the longer response among decisive predictions for which the two responses had unequal word counts. The longer-response selection rate ranges from 71.71% to 84.36% across judges. This indicates a strong association between response length and model preference. However, it does not establish that judges rely on length alone, because longer responses may also contain more relevant evidence, reasoning, or visual details. Detailed per-judge rates are reported in Appendix Table 6.

Length-controlled policy ranking. We additionally recomputed the policy-model ranking using a length-controlled Bradley–Terry formulation. The adjustment preserves the same three-model top tier but changes its ordering: GPT-5.5 ranks first after controlling for response length, while Gemini 2.5 Pro and Gemini 2.5 Flash receive decreases of 49.6 and 36.6 Elo points, respectively. Thus, response length explains part of Gemini’s raw preference advantage, but does not fully account for the overall performance differences among policy models (Appendix Table 7).

4.4 Ablation study

Modern real-world image generalization. As stated earlier, our real image dataset is compiled from COCO 2017. The motivation behind this decision was that it was a readily available dataset with a large variety of "real" images. However, a concern was that the dataset was outdated. To examine this concern, we collected a 50-image researcher-captured subset of modern real-world images spanning varied backgrounds, scenes, and angles. We then had the models respond to the same original prompt and compared their accuracy when judging image authenticity. Mean detector accuracy was 98.99% on COCO-2017 and 98.5% on the researcher-captured images. This small 0.49-point difference indicates that the evaluated detectors generalize well from COCO-2017 to contemporary researcher-captured images.

Performance on the Imagen subset. We additionally recomputed judge performance on triplets associated with Imagen-generated images, which represent a particularly photo-realistic subset of the benchmark. The effect is not uniform across every judge. However, several strong judges exhibit substantial degradation: Claude Opus 4.7 decreases by 8.9 percentage points in four-way accuracy and o3 decreases by 7.8 points. Both models also exhibit larger declines under the Both-Clear metric.

Other judges, including GPT-4o and Gemma 3n, maintain or improve their four-way accuracy on this subset. Therefore, the Imagen subset should not be characterized as universally harder for every model. Instead, it reveals model-specific robustness differences that are obscured by aggregate benchmark results. Full subset results are reported in Appendix Table 8.

Human against policy models. In addition to our Elo calculations among MLLMs as policy models, we wanted to see how a human response would fare against MLLMs in providing explanations in detecting AI-generated images. We had annotators write 100 new human responses. We then paired each human response with responses from three other random AI models to create a new batch of 336 triplets. The human baseline ranked sixth in this comparison, below Gemini 2.5 Pro, Gemini 2.5 Flash, GPT-5.5, o3, and GPT-4o; the full Elo results are provided in Appendix Table 9.

Judge position sensitivity. Human annotations showed little aggregate position imbalance, but MLLM judges may still be sensitive to response ordering. To evaluate this effect, we selected a 500-triplet subset and repeated judge inference after swapping the positions of the two candidate responses. Under a perfectly order-invariant judge, *First* and *Second* predictions should reverse after the swap, while *Tie* and *Both Bad* predictions should remain unchanged.

We report exact swap consistency as the proportion of examples whose four-way prediction transforms correctly after swapping. Clear-choice consistency measures whether the judge selects the same underlying response when it makes a decisive choice in both orders. Stronger judges generally exhibit greater consistency, although residual ordering effects remain. GPT-4o achieves the highest exact swap consistency at 88.38%, while Gemini 2.5 Pro achieves the highest clear-choice consistency at 95.09%. Detailed swap-consistency results are provided in Appendix Table 10.

4.5 Downstream task validation

To validate the practical utility of XAIGID-RewardBench, we used preferences produced by a high-performing reward model to optimize policy models through Direct Preference Optimization (DPO) (Rafailov et al., 2023). An independent GPT-4o judge preferred the DPO-optimized Qwen2.5-VL-3B model in 56.5% of comparisons. For InternVL3.5-4B (Wang et al., 2025c), GPT-4o and human evaluators preferred the DPO-optimized model in 78.0% and 88.5% of comparisons, respectively. Together, these results show that reward-model preferences can improve policy models through DPO across both model families. Full methodology and results are provided in Appendix D.

5 Conclusion

In this paper, we presented the first systematic study of MLLMs as both policy (AI-generated image detectors) and especially reward models (judges) for the tasks related to explainable AI-generated image detection, with an emphasis on judging explanations for AI-generated image detection. Our work delivers three key contributions. First and foremost, we introduce **XAIGID-RewardBench**, the first benchmark designed to evaluate MLLM reward models (judges) in this novel task: Judging Explainable AI-Generated Image Detection, composed of approximately 4,000 human-annotated preference triplets. Second, using this benchmark, we quantify the limitations of current reward models: Gemini 2.5 Flash achieves the highest four-way accuracy (68.8%), while Qwen2.5-VL-72B-Instruct achieves the highest Both-Clear accuracy (86.4%). Third, we provide meaningful insights into common failure modes, identifying critical flaws in model reasoning, such as overgeneralizing from clear artifacts and failing to connect evidence to a logical conclusion. Collectively, our contributions provide a foundational benchmark and a clear analysis to guide the community in developing more accurate, reliable, and trustworthy systems for detecting and explaining AI-generated images, including judging explainable AI-generated image detection.

References

- Anthropic. Introducing Claude Opus 4.7, April 2026. URL <https://www.anthropic.com/news/claude-opus-4-7>. Published April 16, 2026.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuezhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluís Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.
- Zhaorun Chen, Zichen Wen, Yichao Du, Yiyang Zhou, Chenhang Cui, Siwei Han, Jen Weng, Chaoqi Wang, Zhengwei Tong, Leria HUANG, et al. MJ-Bench: Is your multimodal reward model really a good judge for text-to-image generation? *Advances in Neural Information Processing Systems*, 38, 2026.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- Google DeepMind. Imagen 4 model card, May 2025. URL <https://storage.googleapis.com/deepmind-media/Model-Cards/Imagen-4-Model-Card.pdf>. Model card published May 20, 2025.
- Vlad Hondru, Eduard Hoge, Darian Onchis, and Radu Tudor Ionescu. ExDDV: A new dataset for explainable deepfake detection in video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4273–4284, 2026.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuezhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Yikun Ji, Yan Hong, Jiahui Zhan, Haoxing Chen, Jun Lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang. Towards explainable fake image detection with multimodal large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4398–4407, 2025a.
- Yikun Ji, Hong Yan, Jun Lan, Huijia Zhu, Weiqiang Wang, Qi Fan, Liqing Zhang, and Jianfu Zhang. Interpretable and reliable detection of ai-generated images via grounded reasoning in mllms. *arXiv preprint arXiv:2506.07045*, 2025b.
- Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, et al. LEGION: Learning to ground and explain for synthetic image detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18937–18947. IEEE, 2025.

- Black Forest Labs. FLUX. <https://github.com/black-forest-labs/flux>, 2024.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025.
- Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. VL-RewardBench: A challenging benchmark for vision-language generative reward models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24657–24668. IEEE, 2025.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.
- Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah Smith, Hanna Hajishirzi, and Nathan Lambert. RewardBench 2: Advancing reward model evaluation. In *International Conference on Learning Representations*, volume 2026, pp. 144839–144866, 2026.
- Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pp. 16784–16804. PMLR, 2022.
- OpenAI. Introducing our latest image generation model in the API, April 2025a. URL <https://openai.com/index/image-generation-api/>. Published April 23, 2025.
- OpenAI. Introducing OpenAI o3 and o4-mini, April 2025b. URL <https://openai.com/index/introducing-o3-and-o4-mini/>. Published April 16, 2025.
- OpenAI. GPT-5.5 system card, April 2026. URL <https://openai.com/index/gpt-5-5-system-card/>. Published April 23, 2026.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.
- Recraft, Inc. Recraft introduces a revolutionary AI model that thinks in design language, October 2024. URL <https://www.recraft.ai/blog/recraft-introduces-a-revolutionary-ai-model-that-thinks-in-design-language>. Blog post published October 30, 2024.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. IEEE, 2022.
- Jiacheng Ruan, Wenzhen Yuan, Xian Gao, Ye Guo, Daoxin Zhang, Zhe Xu, Yao Hu, Ting Liu, and Yuzhuo Fu. VLRMBench: A comprehensive and challenging benchmark for vision-language reward models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3163–3173. IEEE, 2025.
- Chuangchuang Tan, Jinglu Wang, Xiang Ming, Renshuai Tao, Yunchao Wei, Yao Zhao, and Yan Lu. ForenX: Towards explainable AI-generated image detection with multimodal large language models. *arXiv preprint arXiv:2508.01402*, 2025.
- Hao Tan, Zichang Tan, Senyuan Shi, Ajian Liu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, Zhen Lei, et al. Veritas: Generalizable deepfake detection via pattern-aware reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 66420–66477, 2026.
- Gemma Team. Gemma 3n, 2025. URL <https://ai.google.dev/gemma/docs/gemma-3n>.

- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers Reinforcement Learning, 2020. URL <https://github.com/huggingface/trl>.
- Haochen Wang, Xuhui Liu, Ziqian Lu, Cilin Yan, Xiaolong Jiang, Runqi Wang, and Efstratios Gavves. RADAR: Reasoning AI-generated image detection for semantic fakes. *Technologies*, 13(7):280, 2025a.
- Jin Wang, Chenghui Lv, Xian Li, Shichao Dong, Huadong Li, Kelu Yao, Chao Li, Wenqi Shao, and Ping Luo. Forensics-bench: A comprehensive forgery detection benchmark suite for large vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4233–4245. IEEE, 2025b.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c.
- Zhipeng Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. FakeShield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations*, volume 2025, pp. 31186–31216, 2025.
- Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal Reward-Bench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*, 2025.
- Peipeng Yu, Jianwei Fei, Hui Gao, Xuan Feng, Zhihua Xia, and Chip Hong Chang. Unlocking the capabilities of large vision-language models for generalizable and explainable deepfake detection. In *Forty-second International Conference on Machine Learning*, 2025.

A Additional information for benchmark annotation

Rubric:

- *Hallucination: Any extraneous details?*
- *Details: Did it get all the important details(distortions)?*
- *Logic: Did it correctly formulate an argument(regardless of wrong/right)?*
- *Relevance: Do the points that the LLM make actually help its argument?*
- *Counterargument: Did it shut down any opposing points?*
- *Weighing: Did it correctly weigh why the image might be real or fake(Lots of points may be thrown out, but there needs to be a tying thread to unify all the points in favor of its argument)*
- *Contradiction: Does the model contradict itself?*
- *Insight: Are the arguments shallow or unrefined?*

B Setup of annotation tool

Human annotators might be prone to bias, such as choosing longer answers simply because of their length or judging a response purely based on its final answer rather than its thought process. Our goal is to maximize the MLLM’s useful reasoning steps and minimize nonsensical output. To ensure minimal bias in annotation, several countermeasures were taken. First, a hidden stopwatch recorded the time spent on each annotation as metadata to record any overly quick annotations. Second, cropping the images helps hide the image’s true nature. Some annotators may be inclined to select the response that agrees with their own initial assessment. Thus, cropping the images helps encourage annotators to truly think about the model’s response (Figure 5).

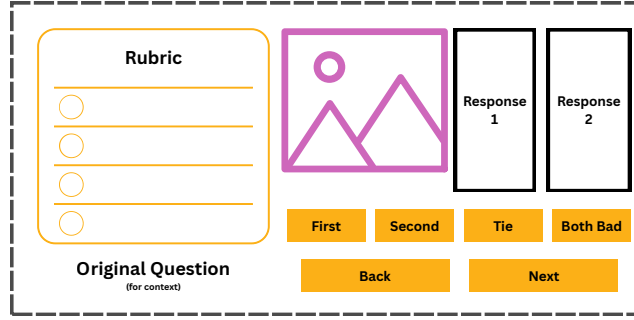


Figure 5: A conceptual illustration of our annotation tool. The image and response sources are hidden to remove bias. Annotators are also encouraged to look past the initial one-word verdict and focus more on the explanation given by MLLM detectors.

C Additional experiment tables

Policy Model	SD 1.5	Imagen 3	Bagel	Dev	Imagen 4
<i>Proprietary</i>					
GPT-5.5	98.0%	80.0%	96.0%	96.0%	94.0%
GPT-4o	100.0%	48.0%	98.0%	96.0%	82.0%
o3	98.0%	28.0%	98.0%	90.0%	58.0%
Gemini 2.5 Pro	94.0%	22.0%	86.0%	72.0%	24.0%
Gemini 2.5 Flash	88.0%	14.0%	84.0%	54.0%	14.0%
<i>Open Source</i>					
Qwen2.5-VL-3B-Instruct	32.0%	20.0%	70.0%	52.0%	40.0%
gemma-3n-E4B-it	16.0%	12.0%	28.0%	10.0%	14.0%
Qwen2.5-VL-7B-Instruct	10.0%	2.0%	14.0%	4.0%	2.0%

Policy Model	Recraft	Schnell	SD 3.5	GPT-Image-1	Glide
<i>Proprietary</i>					
GPT-5.5	88.0%	96.0%	98.0%	98.0%	90.0%
GPT-4o	44.0%	96.0%	92.0%	98.0%	100.0%
o3	18.0%	88.0%	70.0%	70.0%	98.0%
Gemini 2.5 Pro	10.0%	52.0%	56.0%	64.0%	78.0%
Gemini 2.5 Flash	8.0%	40.0%	42.0%	46.0%	78.0%
<i>Open Source</i>					
Qwen2.5-VL-3B-Instruct	2.0%	36.0%	20.0%	46.0%	66.0%
gemma-3n-E4B-it	2.0%	4.0%	6.0%	26.0%	62.0%
Qwen2.5-VL-7B-Instruct	0.0%	2.0%	4.0%	4.0%	18.0%

Generator provenance was obtained from the source metadata associated with each synthetic image.

Table 4: Policy model accuracy on detecting images from various generation models. The accuracies show that images from the Recraft and Imagen models were the most difficult to detect.

Rank	Policy Model	Elo Rating	Decisive Games	Wins	Losses	Win Rate (%)
1	GPT-5.5	1520.2	895	765	130	85.47
2	Gemini 2.5 Pro	1503.0	963	866	97	89.93
3	Gemini 2.5 Flash	1490.6	965	807	158	83.63
4	GPT-4o	1095.7	857	248	609	28.94
5	o3	1085.2	880	399	481	45.34
6	Gemma 3n	1016.0	822	131	691	15.94
7	Qwen2.5-VL-3B-Instruct	949.6	834	147	687	17.63
8	Qwen2.5-VL-7B-Instruct	939.7	812	151	661	18.60

Table 5: Human-preference Elo rankings of the policy models on the clean benchmark. Sequential Elo is calculated using 3,514 decisive human comparisons; 898 neutral annotations labeled *Tie* or *Both Bad* are excluded. GPT-5.5 achieves the highest Elo rating, followed by Gemini 2.5 Pro and Gemini 2.5 Flash. Elo starts at 1200 with a K-factor of 32.

Judge Model	Unequal-Length Pairs	Longer Selected	Shorter Selected	Longer Rate (%)
Claude Opus 4.7	3,643	2,763	880	75.84
Gemini 2.5 Flash	3,821	3,000	821	78.51
Gemini 2.5 Pro	3,428	2,779	649	81.07
GPT-4o	3,888	3,083	805	79.30
o3	3,568	2,618	950	73.37
Gemma 3n	3,851	3,060	791	79.46
Qwen2.5-VL-3B-Instruct	1,604	1,210	394	75.44
Qwen2.5-VL-7B-Instruct	1,255	900	355	71.71
Qwen2.5-VL-72B-Instruct	2,787	2,351	436	84.36

Table 6: Response-length preference among MLLM judges. Results are conditioned on examples where the judge selected a clear winner and the two candidate responses had unequal word counts. A high longer-response selection rate indicates an association between length and preference, but does not by itself establish a causal length bias.

LC Rank	Policy Model	Raw BT Elo	LC-Elo	Difference
1	GPT-5.5	1598.7	1602.9	+4.2
2	Gemini 2.5 Pro	1645.4	1595.8	-49.6
3	Gemini 2.5 Flash	1547.1	1510.4	-36.6
4	o3	1162.6	1185.6	+23.0
5	GPT-4o	982.6	997.2	+14.6
6	Qwen2.5-VL-7B-Instruct	904.5	912.0	+7.5
7	Qwen2.5-VL-3B-Instruct	889.6	908.0	+18.5
8	Gemma 3n	869.5	888.0	+18.5

Table 7: Length-controlled Elo ranking of policy models based on human preferences. Raw BT Elo and LC-Elo are estimated under comparable Bradley–Terry formulations. The final column reports LC-Elo minus raw BT Elo.

Judge Model	Four-Way			Both-Clear		
	Overall	Imagen	Δ	Overall	Imagen	Δ
Claude Opus 4.7	65.2	56.4	-8.9	84.4	78.1	-6.3
o3	63.9	56.1	-7.8	79.8	68.0	-11.8
Gemini 2.5 Pro	68.0	65.6	-2.5	86.3	80.0	-6.3
Gemini 2.5 Flash	68.8	67.9	-0.9	85.4	79.8	-5.6
GPT-4o	68.0	70.9	+2.9	85.0	83.9	-1.1
Gemma 3n	66.3	71.2	+4.9	81.8	82.9	+1.1
Qwen2.5-VL-3B-Instruct	37.5	38.0	+0.5	75.5	72.5	-3.1
Qwen2.5-VL-7B-Instruct	33.2	30.1	-3.1	71.7	74.1	+2.4
Qwen2.5-VL-72B-Instruct	61.6	62.0	+0.4	86.4	85.1	-1.3

Table 8: Judge performance on the Imagen-generated subset compared with the complete benchmark. Differences are calculated as Imagen performance minus overall performance. The subset produces substantial drops for some leading judges, although the effect is not uniform across all models.

Rank	Policy Model	Elo Rating	Games	Wins	Losses	Win Rate
<i>Proprietary Models</i>						
1	Gemini 2.5 Pro	1474.7	36	34	2	0.944
2	Gemini 2.5 Flash	1386.2	38	31	7	0.816
3	GPT-5.5	1341.2	32	26	6	0.813
4	o3	1218.8	16	9	7	0.562
5	GPT-4o	1191.9	26	12	14	0.462
<i>Open Source Models</i>						
7	Qwen2.5-VL-3B-Instruct	1083.0	19	4	15	0.211
8	Qwen2.5-VL-7B-Instruct	1002.5	27	3	24	0.111
9	gemma-3n-E4B-it	1002.1	26	3	23	0.115
<i>Human Reference</i>						
6	Human	1099.7	220	98	122	0.445

Table 9: Elo ratings of policy models including a human baseline. As shown, the human is able to compete with other SOTA MLLMs in response generation but surprisingly loses to both Gemini models. This may be attributed to the fact that Gemini can generate more comprehensive responses while human-generated responses are shorter and with fewer details. This indicates that MLLMs can be more practical as policy models (not as reward models yet) than humans to generate explanations in AI-generated image detection, as they are good at generating detailed and comprehensive explanations which are more appealing to human judges.

Judge Model	Exact Consistency (%)	Clear-Choice Consistency (%)	First Before (%)	First After (%)
Qwen2.5-VL-3B-Instruct	47.60	75.50	64.35	36.02
Qwen2.5-VL-7B-Instruct	66.00	78.57	59.88	39.53
Qwen2.5-VL-72B-Instruct	84.00	91.33	53.14	38.22
Gemma 3n	75.40	76.63	44.31	37.40
GPT-4o	88.38	89.27	49.70	44.98
o3	80.80	87.78	49.67	49.39
Gemini 2.5 Flash	87.40	90.61	48.64	47.08
Gemini 2.5 Pro	85.19	95.09	51.00	45.60
Claude Opus 4.7	85.62	88.41	47.54	40.93

Table 10: Position-swap consistency of MLLM judges on a 500-triplet subset. Exact consistency requires the complete four-way label to transform correctly after swapping the two responses. Clear-choice consistency measures whether the same underlying response is preferred when the judge makes a decisive choice in both response orders.

D Downstream task validation

Beyond evaluating reward models on XAIGID-RewardBench, we investigate whether preferences produced by a high-performing judge can improve policy models through Direct Preference Optimization (DPO). We conduct two downstream experiments using Qwen2.5-VL-3B-Instruct and InternVL3.5-4B as policy models. In both experiments, Gemini 2.5 Pro provides the training preferences, while evaluation is conducted using Gemini 2.5 Pro, an independent GPT-4o judge, and human annotators.

Importantly, the results in this section are pairwise preference win rates between a DPO-optimized model and its corresponding base model. They are not accuracy scores on XAIGID-RewardBench.

D.1 Image and preference data

Separate from the benchmark, we constructed a candidate pool containing 5,000 COCO-2017 real images and 5,000 synthetic images. The synthetic images were generated using Stable Diffusion 3.5, GPT-Image-1, Imagen 4, and Flux 2 Klein. As in the benchmark construction process, an MLLM first generated a caption describing a real image, and the caption was then provided to a text-to-image model to produce a semantically paired synthetic image.

For each image, we independently sampled two responses from the target base policy model, producing one image-response-response triplet. Responses were generated with a maximum of 512 new tokens, and generations containing error strings or malformed outputs were removed.

For the initial Qwen2.5-VL-3B-Instruct experiment, we processed 1,000 real images and 250 images from each of the four synthetic-image sources, yielding approximately 2,000 candidate preference triplets. After filtering, 1,890 triplets were retained for training. A separate 200 triplet test set was used to evaluate Qwen. We repeated the same response-generation and preference-annotation pipeline for the InternVL3.5-4B replication.

Gemini 2.5 Pro was asked to select the more convincing response according to criteria aligned with the benchmark annotation rubric. The annotation prompt was:

You are an expert image forensics annotator. You will be given an image and two analytical arguments about whether the image is AI-generated or a real photograph. Your job is to decide which argument is more convincing and better supported by visual evidence in the image.

The selected response was treated as the chosen response y_w , while the other response was treated as the rejected response y_l .

D.2 DPO training

We applied DPO using the DPOTrainer implementation from the TRL library (von Werra et al., 2020), together with LoRA (Hu et al., 2022). We separately fine-tuned Qwen2.5-VL-3B-Instruct and InternVL3.5-4B using the same training objective and hyperparameter configuration.

The DPO objective (Rafailov et al., 2023) is

$$\mathcal{L}_{\text{DPO}}(\pi_\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right],$$

where π_θ is the policy being optimized, π_{ref} is the frozen base policy, y_w and y_l are the chosen and rejected responses, respectively, and β controls the strength of deviation from the reference policy.

The Qwen2.5-VL-3B-Instruct experiment was trained on a single NVIDIA RTX A4000 with 16GB of VRAM and completed in approximately six hours. Training consisted of 711

Parameter	Value
Epochs	3
Per-device batch size	1
Gradient accumulation	8 steps
Effective batch size	8
Learning rate	5×10^{-5}
LR scheduler	Cosine
Warmup ratio	0.1
DPO beta	0.1
Maximum sequence length	2048
Gradient checkpointing	Enabled
Precision	bfloat16

Table 11: DPO training configuration used for the downstream policy-model experiments.

Parameter	Value
Rank (r)	16
LoRA alpha	32
Target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
LoRA dropout	0.05
Bias	None
Task type	CAUSAL_LM

Table 12: LoRA configuration used for downstream DPO training.

gradient steps over 1,890 retained preference examples. The InternVL3.5-4B replication used the same training configuration and environment.

D.3 Evaluation

For each policy model, we compared responses from the DPO-optimized model against responses from its corresponding frozen base model. Model identities were hidden from evaluators, and response order was randomized to reduce position and model-identity bias.

We used two automated judges. Gemini 2.5 Pro was included because it generated the training preferences. GPT-4o was used as an independent judge because it did not participate in preference annotation or training. We also conducted blind human evaluation using the same pairwise response format.

D.4 Results

Table 13 reports the percentage of comparisons in which the DPO-optimized policy was preferred over the corresponding base policy.

Policy Model	Gemini 2.5 Pro (%)	GPT-4o (%)	Human (%)	Automated-Judge Agreement (%)
Qwen2.5-VL-3B-Instruct	61.50	56.50	60.69	73.00
InternVL3.5-4B	73.15	78.00	88.51	81.88

Table 13: Downstream DPO evaluation across two policy models. Values report the pairwise win rate of the DPO-optimized policy over its corresponding base policy. GPT-4o did not provide the training preferences and therefore serves as an independent automated judge.

For Qwen2.5-VL-3B-Instruct, the DPO model was preferred in 61.50% of comparisons by Gemini 2.5 Pro, 56.50% by GPT-4o, and 60.69% by human annotators. These results indicate a modest but consistent improvement over the base model. Because Gemini 2.5 Pro provided

the training labels, the GPT-4o and human results offer the more independent measures of improvement.

The InternVL3.5-4B replication produced a stronger downstream signal. The DPO model achieved win rates of 73.15% under Gemini 2.5 Pro, 78.00% under GPT-4o, and 88.51% under human evaluation. Agreement between the two automated judges also increased from 73.00% in the Qwen experiment to 81.88% in the InternVL experiment.

We additionally compare each automated judge with the human preferences for the InternVL3.5-4B experiment in Table 14.

Judge	DPO Win Rate (%)	Agreement with Human (%)
Human	88.51	–
Gemini 2.5 Pro	73.15	76.12
GPT-4o	78.00	85.06

Table 14: Automated-judge agreement with human preferences in the InternVL3.5-4B downstream experiment.

D.5 Discussion

The Qwen2.5-VL-3B-Instruct experiment provides initial evidence that preferences produced by a benchmark-selected judge can improve policy responses. However, the improvement is modest, particularly under the independent GPT-4o evaluation. Qualitative inspection suggests that the fine-tuned and base responses often retain similar overall structures, but the DPO model is less likely to introduce unsupported or hallucinated visual details.

The stronger InternVL3.5-4B results show that the downstream effect is not restricted to a single policy architecture. In particular, its 78.00% win rate under GPT-4o and 88.51% human win rate provide stronger evidence than the initial Qwen experiment. The high agreement between GPT-4o and human annotators further suggests that this improvement is not solely an artifact of alignment with the Gemini 2.5 Pro training labeler.

Nevertheless, these experiments do not establish that Gemini 2.5 Pro is an unbiased or universally reliable reward model. Both experiments rely on preferences from a single training-time judge, and the resulting policies may inherit that judge’s biases, including its preference for particular response styles or lengths. Future work should compare multiple preference-labeling judges, combine human and automated preferences, and evaluate additional policy architectures and out-of-distribution image sources.

E Chain-of-thought prompting ablation

We tested whether explicitly requesting chain-of-thought reasoning improves Qwen2.5-VL-7B-Instruct as a judge on a 500-triplet subset. Chain-of-thought prompting did not improve overall alignment with human preferences. Overall human agreement decreased from 43.6% to 28.8%, and Human-Clear agreement decreased from 41.1% to 20.7%. Although Both-Clear accuracy increased slightly from 70.6% to 72.6%, decisive predictions fell from 258 to 127 and invalid outputs increased from 0 to 8. These results suggest that chain-of-thought prompting changed the judge’s output calibration and increased neutral behavior rather than improving its overall judging ability.

Metric	No CoT	CoT	Difference
Invalid outputs	0 (0.0%)	8 (1.6%)	N/A
Decisive selections	258 (51.6%)	127 (25.4%)	-26.2 pp
Overall human agreement	217/498 (43.6%)	141/490 (28.8%)	-14.8 pp
Human-Clear agreement	166/404 (41.1%)	82/397 (20.7%)	-20.4 pp
Both-Clear accuracy	166/235 (70.6%)	82/113 (72.6%)	+2.0 pp
Longer-response selection rate	183/257 (71.2%)	94/126 (74.6%)	+3.4 pp
Mean judge-output length	52.2 words	55.7 words	+3.5 words

Table 15: Chain-of-thought prompting ablation for Qwen2.5-VL-7B-Instruct on 500 triplets. Although conditional Both-Clear accuracy increases slightly, chain-of-thought prompting substantially reduces decisive predictions and lowers overall agreement with human labels.

F AI-generated image examples

We show some examples of generated images from our benchmark in Figure 6.



Figure 6: A collage of several AI-generated images from our benchmark. Some of these images were challenging to identify, even for human eyes.

G More response example

We provide additional examples of policy model responses in Figure 7, Figure 4a, and Figure 4b.



Figure 7: Example of a model downplaying a critical artifact. The distorted text on the street sign is a dead giveaway that the image is generated. While the rest of the image is very consistent, a mistake like this is undeniable. Still, the model tries to argue that one slip-up is acceptable.

H Limitations and future work

While this benchmark provides a crucial tool for the community, our reliance on human preference as the "ground truth" for explanation quality has its own limitations. However, at present, no absolutely objective metric exists for evaluating explanations in AI-generated image detection, considering the complex nature of the task itself. Therefore, human judgment remains the only viable gold standard. We used a detailed rubric, but the 59.78% full four-way pairwise agreement indicates substantial ambiguity; the higher 87.39% clear-winner conditional agreement does not cover disagreements involving *Tie* or *Both Bad*. Future work could achieve greater robustness by increasing the scale of the dataset and employing a larger, more diverse pool of annotators to build a stronger consensus on challenging or ambiguous cases.

Similarly, our initial focus on static images was a pragmatic choice to ensure depth and quality before tackling more complex data types. A key direction for future work is to build upon this foundation to create benchmarks for more modalities, such as AI-generated videos, synthesized audio, and even generated 3D assets, each presenting unique forensic challenges. This expansion is critical for developing comprehensive and versatile detection systems.

Our selection of MLLMs was primarily constrained due to computational costs. In an ideal benchmark, we would evaluate a broader range of policy and reward models. However, we believe that the MLLMs we covered provide a solid overview of proprietary vs open-source, large vs small, and different architectures.

Our primary real-image benchmark uses COCO-2017. A 50-image researcher-captured out-of-distribution ablation found comparable detector accuracy on contemporary photographs (98.5%) and COCO-2017 (98.99%), supporting cross-domain generalization of authenticity classification. Future work can extend this analysis to explanation quality and judge agreement.

Finally, our failure-mode analysis is qualitative, anecdotal, and non-exhaustive. The selected examples illustrate plausible patterns but do not establish their prevalence; future work should quantify these categories across the full benchmark using independent human annotation or validated classifiers.

Ultimately, the goal is to detect AI-generated content and explain the detection process in a trustworthy way. The path forward lies in moving beyond using MLLMs on their own and toward building MLLM-powered agents that can actively reason and leverage external tools. These agents could dynamically use forensic tools and construct evidence-based narratives, mirroring a human expert's workflow. The central challenge will be to develop

reward models that can guide this complex reasoning and action process, a necessary step to close the gap between current models and the dynamic, trustworthy analysis required for real-world applications.

I Released benchmark statistics

This section summarizes benchmark construction and cleaning (Tables 16 and 17), composition (Table 18), annotation coverage and agreement (Tables 19 and 20), and GPT-5.5 detector accuracy (Table 21).

Slice	Candidate triplets	Retained	Removed
Original benchmark	3,000	2,990	10
GPT-5.5 extension	1,000	998	2
Combined	4,000	3,988	12

We remove triplets whenever either policy response contains a generation-error payload or lacks a usable verdict prefix.

Table 16: Construction of the released clean benchmark.

Flag	Affected triplets
Generation-error payload	12
Missing usable verdict prefix	12
API content filter	9
CUDA out of memory	3

Flags overlap: each removed triplet has an error payload and lacks a usable verdict prefix; the final number of removed triplets is 12.

Table 17: Reasons flagged during failed-generation removal.

Image source	Type	Unique images	Clean triplets
COCO-2017 real	Real	500	1,995
FLUX.1-dev	Synthetic	50	200
FLUX.1-schnell	Synthetic	50	200
Stable Diffusion 3.5	Synthetic	50	200
Stable Diffusion 1.5	Synthetic	50	200
Imagen 4	Synthetic	50	200
GPT Image	Synthetic	50	200
Recraft V3	Synthetic	50	197
GLIDE	Synthetic	50	200
BAGEL	Synthetic	50	200
Imagen 3	Synthetic	50	196
All synthetic	Synthetic	500	1,993
Full benchmark	All	1,000	3,988

Table 18: Image composition of the clean benchmark.

Statistic	Value
Clean benchmark triplets	3,988
Triplets with at least one annotation	3,959
Annotation coverage	99.27%
Total annotations	4,412
Unique annotators	11
Response 1 better	1,791
Response 2 better	1,723
Tie	698
Both bad	200

Table 19: Human annotation summary for the clean benchmark.

View	Eligible triplets	Unanimous (%)	Matching/total pairs	Agreement (%)
All four labels	349	55.59%	333/557	59.78%
Clear-winner conditional	245	84.08%	298/341	87.39%

The all-label view includes Response 1, Response 2, Tie, and Both Bad. The clear-winner conditional view includes only comparisons where the paired annotations both select a response; it is reported alongside, not in place of, the four-way result.

Table 20: Inter-annotator agreement under the four-way and clear-winner views.

Split	Correct	Total	Accuracy
Real	461	500	92.20%
Synthetic	467	500	93.40%
Overall	928	1,000	92.80%

Table 21: GPT-5.5 direct real-versus-synthetic detector accuracy.

Name	Task	Size	Human?
MJ-Bench	T2I judge	8K	Expert
VL-RewardBench	VL judge	1.25K	Verified
Multimodal RB	VLM RM eval.	5.21K	Expert
VLRMBench	VL RM eval.	12.6K	Reviewed
Ours	AIGC explanations	3.99K	Yes

Sizes use each paper’s reported unit. Our benchmark contains 3,988 clean triplets and 4,412 human annotations from 11 annotators.

Table 22: Comparison with multimodal reward-model benchmarks: MJ-Bench (Chen et al., 2026), VL-RewardBench (Li et al., 2025), Multimodal RewardBench (Yasunaga et al., 2025), and VLRMBench (Ruan et al., 2025).