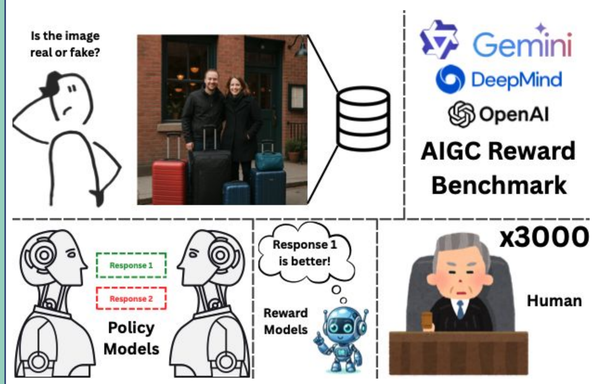
Explainable AI-Generated Image Detection RewardBench

Michael Yang, Shijian Deng, William T. Doan, Kai Wang, Tianyu Yang, Harsh Singh, Yapeng Tian

THE UNIVERSITY OF TEXAS AT DALLAS

ERIK JONSSON SCHOOL OF ENGINEERING AND COMPUTER SCIENCE

OVERVIEW OF THE XAIGID-RewardBench



SOME FAILURE CASES OF MLLM POLICY MODELS

Real

The image depicts a real urban street scene with clear markings for a bicycle lane, traffic lights, and buildings lining the street. The presence of vehicles, pedestrians, and typical city infrastructure confirms that this is a real photograph. There are no signs of digital manipulation or artificial elements that would suggest it is fake.

(a) Example of a response making an irrelevant point. While the mention of the city and the road is true, there is no actual argument being made. An AI-generated image can also have vehicles and pedestrians.

Real

The image appears to be a real photograph. Here's why:

...

4. **Plausible Objects:** The furniture, windows, and the gaming controller (resembling a Nintendo Wii remote, though perhaps a generic or third-party version) are all coherent and realistically rendered. The controller even has a cord clearly visible and attached, which AI often struggles to render convincingly or accurately connect.

...

(b) Example of a model generalizing details and missing key artifacts. While the controller's shape looks realistic, the model overlooks its details, failing to notice the missing buttons.

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 persuasive- ness of these detection tools for real-world applications. Leveraging Multi-4 modal Large Language Models (MLLMs) has recently become a trending solution to this issue. Further, to evaluate the quality of generated explana- tions, a common approach is to adopt an "MLLM as a judge" methodology7 to evaluate explanations generated by other MLLMs. However, how well those MLLMs perform when judging explanations for AI-generated im- age 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 bench- mark consists of approximately 3,000 annotated triplets sourced from var- ious image generation models and MLLMs as policy models (detectors) to assess the capabilities of current MLLMs as reward models (judges). Our results show that the current best reward model scored 88.76% on this benchmark (while human inter-annotator agreement reaches 98.30%), demonstrating that a visible gap remains between the reasoning abilities of today's MLLMs and human-level performance. In addition, we provide an analysis of common pitfalls that these models frequently encounter.

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 4. 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 4). Table 4: Performance of MLLMs as Reward Models (Judges). All proprietary models and Gemma performed reasonably well on this benchmark. Gemma's strong performance may be related to its origin as a model from the Gemini family. The two Qwen models are much weaker and thus performed much worse on the benchmark. The '2-Way Acc' column considers only cases where the human annotator marked a clear winner. The goal was to see if models could identify the correct winner in unambiguous cases. This is because ties create "gray areas" which make it harder to classify, but the task is more straightforward when a clear winner exists.

Reward Model (MLLM as a Judge)	4-Way Acc (%)	2-Way Acc (%)
Proprietary		
Gemini 2.5 Pro	68.35	**88.76**
Gemini 2.5 Flash	68.92	87.26
GPT-4o	60.62	85.59
o3	64.74	82.17
Open Source		
gemma-3n-E4B-it	66.88	84.54
Qwen2.5-VL-7B-Instruct	29.37	82.28
Qwen2.5-VL-3B-Instruct	31.35	77.12

CONTRIBUTIONS

(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.

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 in- troduce 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 3,000 human-annotated preference triplets. Second, through this benchmark, we quantitatively demonstrate a visible gap in current model capabilities, with our top-performing reward model reaching only 88.76% accuracy (while human inter-annotator agreement reaches 98.30%). Third, we provide meaningful insights into common failure modes, identifying critical flaws in model reasoning such as generalizing over clear artifacts and failing to connect evidence to a logical conclusion. Collectively, our contribu- tions 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.

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 did our best to make this process rigorous by using a detailed rubric, and our high inter-annotator agreement confirms a consistent signal. Future work, however, could achieve even greater robustness by dramatically 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.