FakeXplain Trains an LLM to Detect AI-Generated Images & Answer *Where* and *Why*

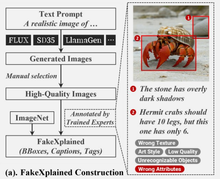
Dataset: FakeXplained

	Localization	Textual Reason	Tagged Reason	Image Sources	Human-annotated	Human Alignment
GenImage	✗	✗	✗	8	✗	✗
FakeBench	✓	–	–	10	✗	✗
MMFR-Dataset	✗	✓	✗	10*	✗	✗
SynthScars	✓	✓	✗	–	✓	✗
MMTD-Set	✓	✓	✓	–	✗	✗
Ours	✓	✓	✓	29	✓	✓

FakeXplained is a new dataset of 8,000+ AI-generated images paired with human-annotated explanations of the visual cues that reveal their synthetic origin.

For each image, expert annotators provide **one or more bounding boxes**, a **textual descriptor for each box**, and a set of predefined global tags characterizing the image-level cues.

- Rich image origin: 28 text-to-image generators across 4 architecture families: Diffusion (Midjourney, SD, DALL-E, etc.), GAN (StyleGAN, VQGAN, etc.), DiT (PixArt, DiT, etc.), and Others (VAR, Infinity, MaskGIT, LlamaGen).
- Semantic sources: ImageNet-1K classes + MS COCO captions. Real images are also sampled from these datasets.
- Manual + Automatic Quality Control.



(a). FakeXplained Construction

Method: FakeXplainer

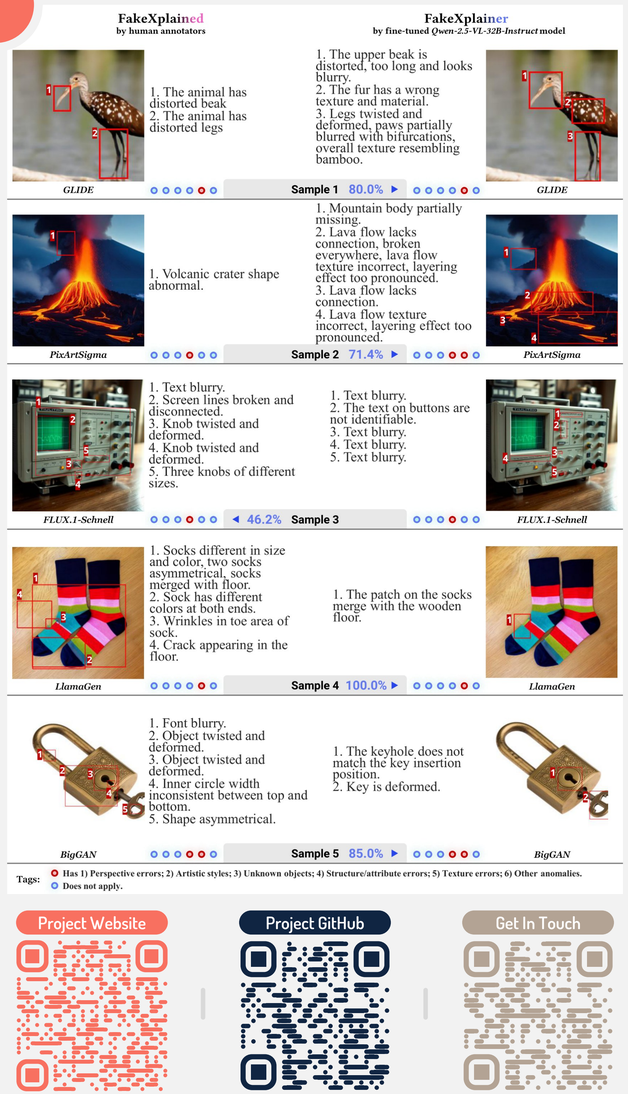
FakeXplainer is a two-stage fine-tuning pipeline that teaches a multimodal LLM to **detect** AI-generated images, **localize** artifact regions with bounding boxes, and **articulate** human-aligned explanation all in a single forward pass.

The training starts from supervised fine-tuning (SFT) for structured **chain-of-thought output**, followed by a progressive Group Relative Policy Optimization (**pGRPO**) strategy with three reward signals: **classification accuracy, grounding IoU, and output format validity**.

pGRPO trains FakeXplainer to sharpen **one skill at a time**. The grounding accuracy reward item starts off with 0.5 in weight and gradually grows to 1.0 as training goes. Ablation studies prove its necessity and effectiveness.

Irregular Handles!

- *Black-box classifiers achieve high accuracy but cannot explain where or why an image appears fake.*
- *MLLM-based detectors can reason but frequently hallucinate false claims without spatial grounding.*
- *We need a forensic tool that can accurately classify and explain why an image is AI-generated.*



Project Website | Project GitHub | Get In Touch

FakeXplain: AI-Generated Image Detection via Human-aligned Grounded Reasoning. Yikun Ji[1,2], Yan Hong[2], Qi Fan[1], Jun Lan[2], Huijia Zhu[2], Weiqiang Wang[2], Liqing Zhang[1], Jianfu Zhang[1]