

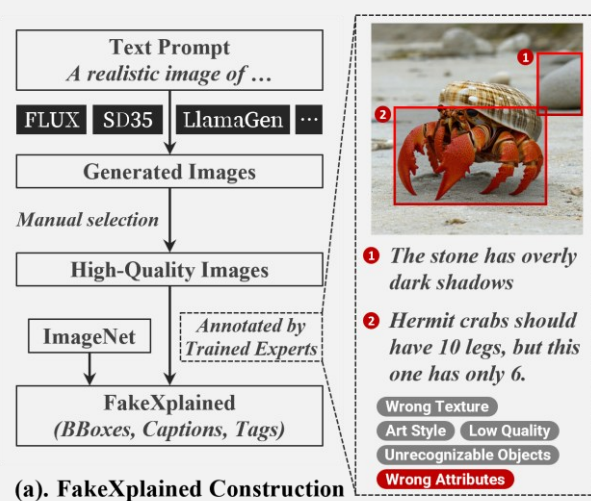
Dataset: FakeXplained

	Textual Localization	Tagged Reason	Image Sources	Human Alignment	Human-annotated
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.

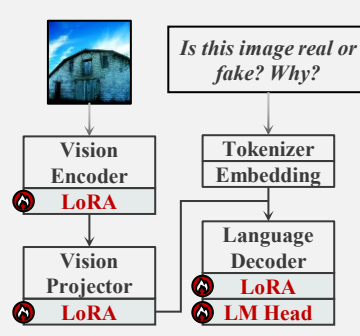
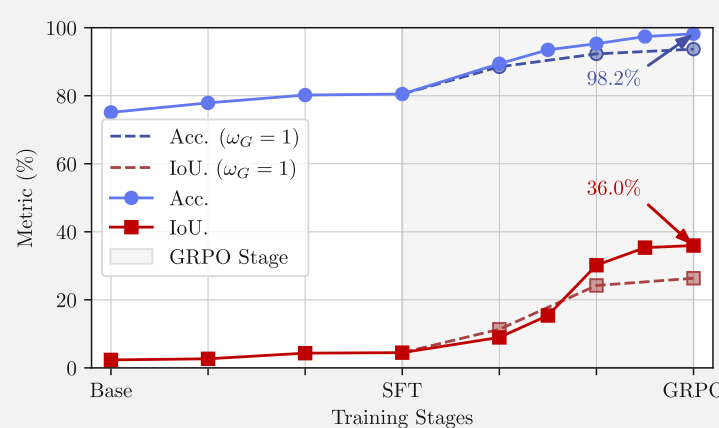


Method: FakeXplainer

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!

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

- 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.



FakeXplainer

Grounding

- Rearview mirror not connected to vehicle
- Tire broken (left)
- Tire too thin (right)
- Text distorted and not identifiable
- Handrail misplaced

Tags

- Wrong Perspective
- Art Style
- Unrecognizable
- Wrong Attributes
- Wrong Texture
- Other Anomalies

AI-Generated

Traditional Method

Generated

0 Real

Other MLLM-based Methods

Upon examining this image, I found the following artifacts:
The left rear wheel of the golf cart is **missing a tire** and has an **unnatural appearance**.

Hallucination (False claims)

This golf cart image has fairly **consistent sunlight** and a reasonably **natural shadow cast**, ...

Non-specific reasons

AI-Generated

FakeXplained by human annotators

FakeXplainer by fine-tuned Qwen-2.5-VL-32B-Instruct model

Sample 1: 80.0% accuracy. Issues: 1. The animal has distorted beak, 2. The animal has distorted legs.

Sample 2: 71.4% accuracy. Issues: 1. Mountain body partially missing, 2. Lava flow lacks connection, broken everywhere, lava flow texture incorrect, layering effect too pronounced, 3. Lava flow lacks connection, 4. Lava flow texture incorrect, layering effect too pronounced.

Sample 3: 46.2% accuracy. Issues: 1. Text blurry, 2. Screen lines broken and disconnected, 3. Knob twisted and deformed, 4. Knob twisted and deformed, 5. Three knobs of different sizes.

Sample 4: 100.0% accuracy. Issues: 1. Socks different in size and color, two socks asymmetrical, socks merged with floor, 2. Sock has different colors at both ends, 3. Wrinkles in toe area of sock, 4. Crack appearing in the floor.

Sample 5: 85.0% accuracy. Issues: 1. Font blurry, 2. Object twisted and deformed, 3. Object twisted and deformed, 4. Inner circle width inconsistent between top and bottom, 5. Shape asymmetrical.

Legend: 1) Perspective errors; 2) Artistic styles; 3) Unknown objects; 4) Structure/attribute errors; 5) Texture errors; 6) Other anomalies.

Project Website, Project GitHub, Get In Touch