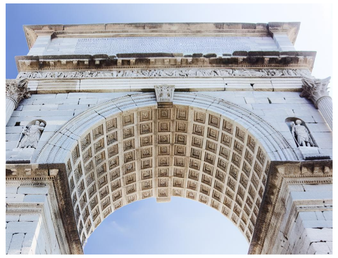


Vision Encoder

Please analyze the image and:
1. Provide an *initial verdict* on whether the image is real or AI-generated.
2. You will need to examine specific regions more closely. Output *bounding boxes* for areas that require *closer inspection*.

Locate THEN Examine

Locate — VLM Query 1 – Global Scan

Programmtic Zoom-In

[Bounding box coordinates]

Extra V-Tokens

Original Image

Vision Encoder

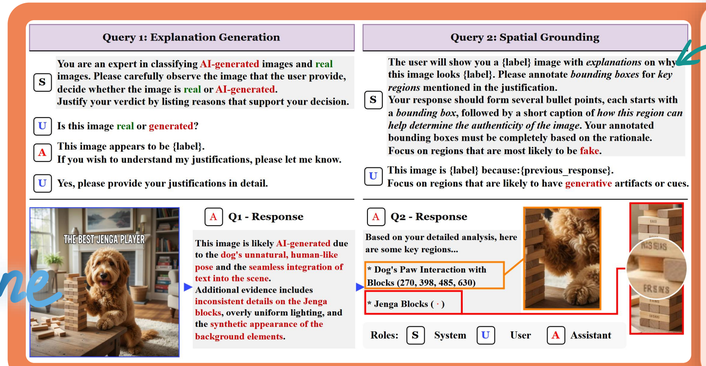
I.V. Initial Verdict

80%

...Initially, you claimed that this image is I.V.
Observe these regions and give a *final verdict*.

97%

Examine — VLM Query 2 – Local Evidence Check



fully automated pipeline

DATASET
TRACE

10K real images from COCO / ImageNet + 10K high-quality AI-generated images
Region-level annotations with bounding boxes highlighting telltale regions
Forensic explanations auto-generated via VLM pipeline with consistency checks
Multiple generator sources including FLUX and Stable Diffusion variants
Quality control through automated filtering and human review

10K+10K
Real + AI-Generated

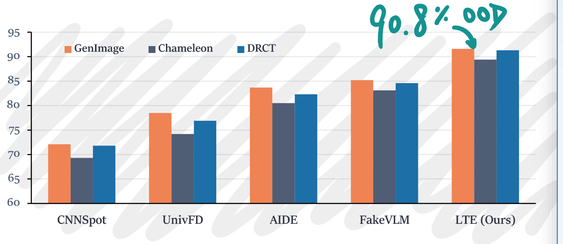
3 regions
Annotated per Image

Image Generators are getting better,

one pass is not enough.

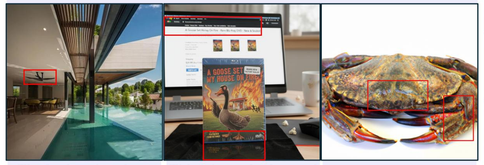
Just like how humans perceive, we

Let models Look Closer.

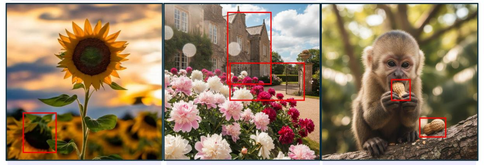


LTE Refined

from *Real* to *Fake*

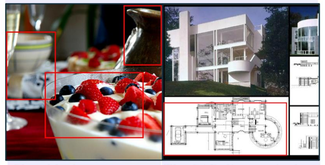


The fan hanging above has a wrong geometry | Title mismatch; the towel should not reflect | Extra claw; missing eye



Missing petals replaced by leaves | The gates are too far away from each other | Peanuts not cracked and poorly-textured

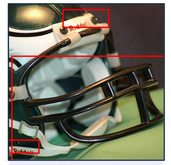
from *Fake* to *Real*



Clear blueberry contour, natural camera blur | The CAD diagram is clean with identifiable text

Precise metallic keyrings; consistent shadows | Symmetrical structure with intricate details

LTE Unchanged



Branding consistent, correct geometry

Natural fog and scenery

▲ Pipeline

LTE In Action ▼

In an era where generative AI booms,
We tackle OoD with explainability,
And took it further via the LTE paradigm.

Initial Verdict is weak, but the refinement is strong.

I. Locate

Which specific region(s) in the image, when examined, can tell the authenticity of the image?

This step typically identifies suspicious regions such as facial features, hands, fine text, and complex textures where generators commonly fail. Consequently, correctly rendering these regions makes the image look more authentic.

II. Examine !

After programmatically cropping the regions, models should now become more evident, effectively polishing their initial verdict.

Local evidence check will utilize the regional close-ups to detect subtle artifacts, or determine that the details are evident that the artwork is indeed realistic.

During training, we used SFT + pGRPO with IoU, BLEU, Accuracy, Formatting as rewards. We only take final verdict's accuracy into account.

METHOD
LOCATE-THEN-EXAMINE

Locate-Then-Examine: Grounded Region Reasoning Improves Detection of AI-Generated Images

Yikun Ji · Yan Hong · Bowen Deng · Jun Lan · Huijia Zhu · Weiqiang Wang · Liqing Zhang · Jianfu Zhang

ANT GROUP