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

Supplementary Material

1. More Details for the TRACE Dataset

The TRACE dataset comprises 20,000 images (10,000 real and 10,000 AI-generated) annotated with forensic explanations and spatially grounded bounding boxes through an automated pipeline combining GPT-4o and Qwen-2.5-VL. This dataset addresses the critical need for grounding-aware training data in AI-generated image detection, particularly providing spatial localization capabilities alongside textual reasoning.

1.1. Source of Images

To ensure comprehensive coverage across different image categories and generation techniques, we sourced images from established datasets and state-of-the-art generation models.

Real Images. The authentic images are sourced from two widely used computer vision datasets: ImageNet [3] and COCO [2]. These datasets provide diverse natural images spanning various object categories, ensuring broad coverage of real-world visual content. The selection from these datasets guarantees high-quality, authentic photographs that serve as reliable negative examples for training.

AI-Generated Images. Half of the synthetic images are created using the OpenAI GPT-Image-1 model [8]; the other half is generated by Gemini 2.5 Flash Image [5]. Both models represent state-of-the-art text-to-image generation capabilities. This choice ensures that our dataset captures contemporary AI generation artifacts and challenges, providing relevant training examples for current detection scenarios.

1.2. Annotation Process

We developed a completely automated pipeline that leverages the complementary strengths of different VLMs to generate comprehensive annotations without requiring extensive human labeling.

Explanation Generation. For images with known ground truth labels, GPT-4o generates detailed forensic explanations focusing on specific visual evidence that indicates whether an image is real or AI-generated. The prompts are designed to elicit detailed reasoning about depicted objects, spatial arrangements, perspective consistency, lighting patterns, and other forensic indicators.

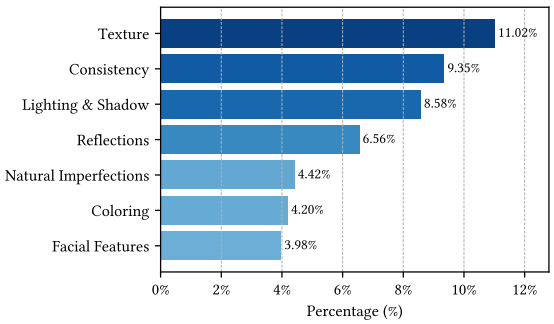

Figure 1. A statistical analysis of keywords in TRACE explanations.

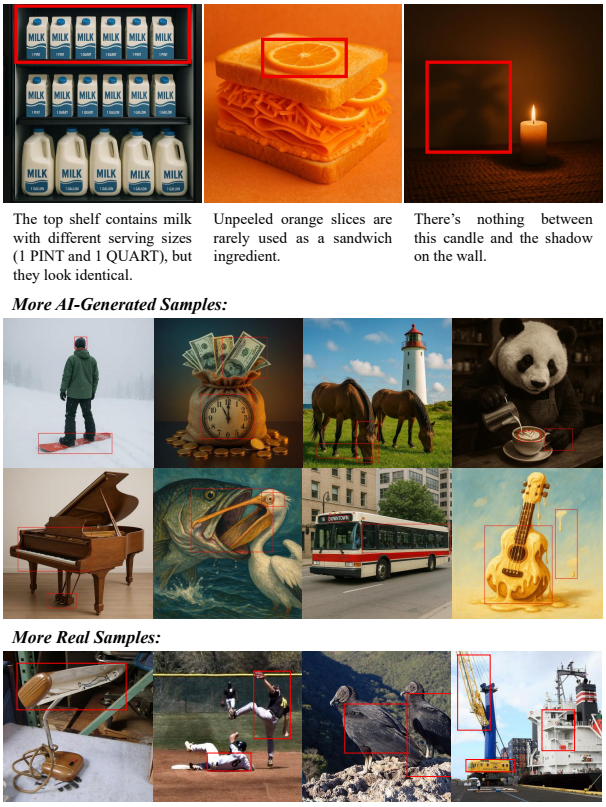

Figure 2. A collection of images from TRACE with rendered bounding boxes. The first row shows three AI-generated images with bounding boxes and corresponding explanations. The second row presents additional AI-generated samples, while the third row illustrates real images, all annotated with bounding boxes.

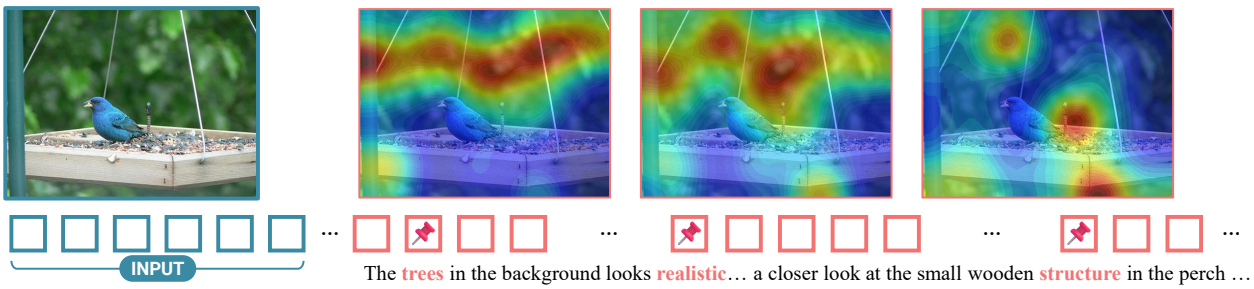

Figure 3. A visualization of the attention mechanisms of the VLM in Query 1.

GPT-4o's strong reasoning capabilities and knowledge of image forensics make it well-suited for generating high-quality explanatory text that identifies key visual cues.

Spatial Grounding. Qwen-2.5-VL extracts bounding boxes from the GPT-4o generated explanations, creating spatially grounded annotations in the form of (I, y, E, B) tuples, where I represents the image, $y \in \{\text{real}, \text{generated}\}$ the ground-truth label, E the explanation, and B the bounding boxes. Image crops C are deterministically derived from (I, B) at training time. The Qwen-2.5-VL model demonstrates strong capabilities in extracting spatial regions based on textual descriptions, making it suitable for converting explanatory text into precise spatial coordinates.

Quality Control and Filtering. During the spatial grounding phase, we observed that Qwen-2.5-VL occasionally generates bounding boxes encompassing over 50% of the image area, often associated with global image characteristics such as over-saturation mentioned in the explanations. In some instances, the model reverts to object detection behavior when the primary flawed object occupies only a small portion of the image. To address these issues, we implement a filtering mechanism that leverages Qwen-2.5-VL to evaluate and remove bounding boxes that either fully encapsulate the primary object (indicating object detection regression) or cover an excessive portion of the image area. This filtration process ensures that the final annotations focus on specific forensic regions rather than global characteristics or entire objects.

Word Frequency Analysis. Figure 1 displays the most frequently occurring keywords in the annotations. Texture and object consistency emerge as the primary concerns, followed by unnatural lighting, shadows, and reflections, indicating that these are the features most commonly leveraged by the model for detecting synthetic content.

1.3. More Samples from TRACE

Figure 2 presents additional examples from the TRACE dataset, demonstrating the diversity of forensic indicators captured by our automated annotation pipeline. The first row shows three AI-generated images with a brief summary of the explanation. We can see that the explanations cover both fine-grained and general semantic reasoning of why this image should be considered real or AI-generated. TRACE features both real and AI-generated images. Four real image samples are provided at the bottom row. This dataset covers a wide range of reasons, fostering explainability for fine-tuned models, and also demonstrating the complexity and variability inherent in synthetic imagery.

1.4. Ethical Considerations

All AI-generated images in the dataset are created specifically for this research and do not depict real individuals. The real images sourced from ImageNet and COCO are used in accordance with their respective licensing terms and ethical guidelines.

1.5. Known Limitations

Automated Annotation Bias. While our automated pipeline reduces human annotation costs, it may inherit biases from the underlying VLMs used for annotation. The quality of explanations and spatial grounding depends on the capabilities and training data of GPT-4o and Qwen-2.5-VL, potentially limiting the coverage of subtle or novel forensic indicators.

Language Limitation. All explanations are generated in English, limiting the applicability of the dataset for multilingual forensic applications. Translation of the nuanced forensic explanations to other languages would require careful validation to maintain technical accuracy.

2. VLM Attention Visualization

In our multi-stage reasoning pipeline, we aim to enable the model to actively "look" for suspicious or diagnostically relevant regions within images, thereby facilitating deeper,

more focused analysis in the second step. A natural question arises: can the VLM truly identify image regions that are semantically aligned with the generated text and relevant to the real/fake detection task?

To verify whether our VLM, Qwen-2.5-VL-32B-Instruct, is indeed attending to specific, meaningful patches of the input image, rather than relying solely on global context or textual priors, we conducted a visualization study using gradient-based attention mapping on a representative sample from Query 1. Specifically, we generated LLaVA-CAM [13] heatmaps to highlight the regions of the image that most strongly influence the model's output predictions. As shown in Figure 3, there is a clear and compelling correspondence between the highlighted areas in the heatmap and the content of the model's generated textual explanation, confirming that the model can localize and focus on relevant regions, which lays a solid foundation for the LTE process. Moreover, we observe that attention often centers around keywords (*e.g.*, "realistic") in the textual explanation, reinforcing the connection between visual grounding and real/fake decision-critical semantics. This confirms that the model can meaningfully propose LTE regions that support reliable second-stage analysis.

3. More Experimental Details

Since real and AI-generated images are not of the same resolution or aspect ratio, we performed center-cropping and resizing to ensure all input images have a resolution of 512×512 during training.

We use ms-swift to fine-tune VLMs. The batch size is set to 1. During the GRPO stage, the number of generations is set to 2. For LTE-32B, the full training pipeline took 42.6 hours on 8x NVIDIA A100 GPUs. We found that at least 600 GB of VRAM is required to perform GRPO. For LTE-7B, the training took 35.3 hours on 4x NVIDIA A100 GPUs. The training process is generally stable. A few loss spikes are observed during the first 1,000 steps of training, but the model quickly converges after that and recovers from the spike.

Details of Baseline Methods A range of methodologies has been proposed for detecting synthetic content, each grounded in distinct theoretical assumptions and detection paradigms.

CNNSpot [11] hypothesizes that CNN-based generative models leave consistent, detectable artifacts and achieve cross-generator generalization through data augmentation. We trained CNNSpot from scratch on the training set of TRACE. The training settings are the same as described in the original work.

Community Forensics [9] adopts a data-centric approach, positing that detection performance scales with the diversity and quantity of training generators, and introduces a large-scale dataset comprising thousands of generators to train robust classifiers.

DIRE [12] takes a process-centric perspective, exploiting the asymmetric reconstruction behavior of diffusion models: real and generated images exhibit differing error patterns when reverse-denoised, forming a discriminative signal known as the DIRE map.

Antifake Prompt [1] leverages VLMs and reformulates detection as a visual question-answering task, employing parameter-efficient soft prompt tuning on a frozen VLM to enable generalization.

NPR [10] utilizes neighboring pixel relationships to identify AI-generated images with good accuracy and generalizability, as CNN-based generative methods exhibit patterns in neighboring pixels.

Collectively, these methods represent diverse strategies from artifact analysis to semantic reasoning, advancing the state of synthetic content detection. During evaluation, all models are trained on the training set of TRACE with the same setup as the original work.

4. Analysis of Bounding Boxes on OoD Datasets

We collected the responses from LTE models when evaluating on OoD datasets, GenImage [14], MMFR-Dataset [4] and SynthScars [6]. On these OoD datasets, Figure 4 and 5 display the relation of bounding boxes with regard to model performance, and the number of detected bounding boxes for each LTE model variant (7B and 32B). The trend of Figure 4 highly resembles *Figure 5* in the main paper, while Figure 5 is slightly different than *Figure 6*, where 32B performs better than 7B for cases with more regions selected. This proves that knowledge from the TRACE dataset can be adapted beyond the dataset, and 32B generalizes better than 7B on OoD datasets.

5. Robustness Against Degradations

We evaluate the robustness on TRACE dataset under four common degradations: JPEG compression at 80% and 30% quality, random cropping, and $0.5\times$ downsampling (Table 1). All methods exhibit performance drops relative to clean images, with the extent varying across perturbations and models.

LTE attains the highest accuracy in every setting and the best IoU among methods that produce localization, with modest declines across degradations (IoU: 0.355 at JPEG 80%, 0.347 at JPEG 30%, 0.346 under downsampling, and 0.306 with random cropping). Random cropping is most detrimental to localization quality, while heavy JPEG compression tends to reduce classification accuracy the most for several baselines. Among the baselines, DIRE

Table 1. Performance on TRACE with degradation, including JPEG compression artifacts, random cropping and image down-sampling.

Degradation	Metric	**LTE**	FakeShield	LEGION	ComFor.	AfPr.	DIRE	CNNSpot	NPR
JPEG Compression (80% Quality)	Acc.	**0.970**	0.781	0.518	0.832	0.873	0.913	0.849	0.869
	IoU	**0.355**	0.089	0.067	-	-	-	-	-
JPEG Compression (30% Quality)	Acc.	**0.964**	0.768	0.505	0.791	0.852	0.896	0.837	0.835
	IoU	**0.347**	0.086	0.066	-	-	-	-	-
Random Cropping	Acc.	**0.965**	0.756	0.513	0.835	0.877	0.909	0.848	0.866
	IoU	**0.306**	0.061	0.063	-	-	-	-	-
Downsampling (0.5x)	Acc.	**0.969**	0.759	0.514	0.890	0.886	0.912	0.851	0.874
	IoU	**0.346**	0.075	0.070	-	-	-	-	-

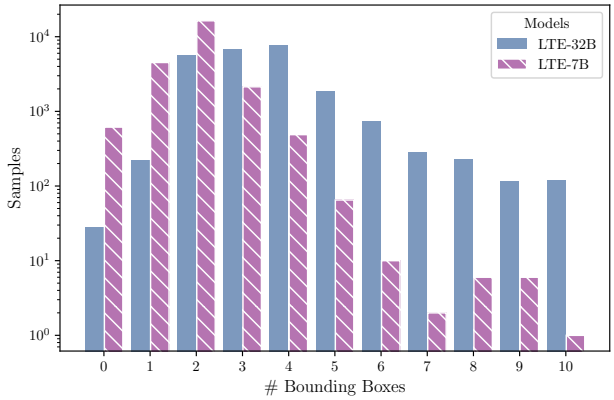

Figure 4. Number of samples grouped by the bounding boxes on OoD datasets.

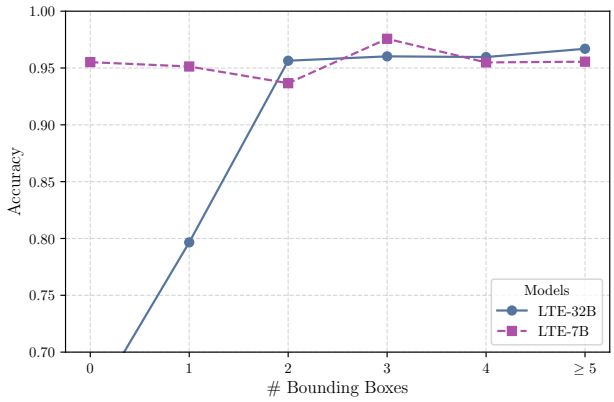

Figure 5. The relation of accuracy with regard to the number of detected bounding boxes on OoD datasets.

consistently produces the best results, followed by AntifakePrompt and CommunityForensics, whereas all VLM-based methods show good robustness against JPEG compression. Downsampling by 50% affects the performance across the models nonuniformly.

These results indicate that, despite relatively strong detection accuracy, most current detectors remain sensitive to common real-world degradations, while VLM-based methods have a lower rate, suggesting opportunities for improvement via degradation-aware training, stronger invariances, and reduced reliance on dataset-specific biases.

Table 2. Mean and standard deviation of inference time per image.

Method	Seconds / Image
LTE-32B	**24.0±3.0**
E-32B (one-turn)	10.9±2.2
LTE-7B	**8.94±1.21**
Base-7B (one-turn)	4.38±0.71
LEGION	11.3±2.61
FakeShield	60.4±5.33
NPR	0.105±0.02
AntifakePrompt	0.182±0.05

6. Computational Efficiency

Table 2 lists the end-to-end inference time for our trained VLMs. We use vllm [7] to accelerate the inference process. The deployment of LTE-32B takes 4x NVIDIA A100-40G GPUs connected with PCI-E. LTE-7B, however, is deployed on one NVIDIA A100-40G GPU. While our two-stage approach increases inference time from traditional classification methods, this overhead is justified by accuracy improvements and interpretability gains. For high-stakes forensic applications, the trade-off favors thoroughness over speed.

7. Limitations & Future Works

The generalization capability of our method has not yet been thoroughly evaluated. In future work, we aim to conduct more comprehensive assessments using a broader range of datasets and image sources to better understand the model's detection accuracy across diverse scenarios.

While our current approach equips the model with a cropping tool to facilitate image-based reasoning, this rep-

resents only a preliminary step toward enabling true visual thinking. Despite its effectiveness, there remains significant potential for further exploration in this direction, particularly in developing more sophisticated interactive mechanisms that empower the model to dynamically analyze and reason over visual content.

References

[1] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419*, 2023. 3

[2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325, 2015. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[4] Yueying Gao, Dongliang Chang, Bingyao Yu, Haotian Qin, Lei Chen, Kongming Liang, and Zhanyu Ma. Fakereasoning: Towards generalizable forgery detection and reasoning. *arXiv preprint arXiv:2503.21210*, 2025. 3

[5] Google. Introducing gemini 2.5 flash image, our state-of-the-art image model. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, 2025. 1

[6] 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. *arXiv preprint arXiv:2503.15264*, 2025. 3

[7] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 4

[8] OpenAI. Introducing 4o image generation, 2025. 1

[9] Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors, 2024. 3

[10] Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection, 2023. 3

[11] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 3

[12] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. 3

[13] Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Enhancing explainability in multimodal large language models. *arXiv e-prints*, pages arXiv–2406, 2024. 3

[14] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36:77771–77782, 2023. 3