ZOOM-IN TO SORT AI-GENERATED IMAGES OUT

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid growth of AI-generated imagery has blurred the boundary between real and synthetic content, raising critical concerns for digital integrity. Vision-language models (VLMs) offer interpretability through explanations but often fail to detect subtle artifacts in high-quality synthetic images. We propose **ZoomIn**, a two-stage forensic framework that improves both accuracy and interpretability. Mimicking human visual inspection, ZoomIn first scans an image to locate suspicious regions and then performs a focused analysis on these zoomed-in areas to deliver a grounded verdict. To support training, we introduce **MagniFake**, a dataset of 20,000 real and high-quality synthetic images annotated with bounding boxes and forensic explanations, generated through an automated VLM-based pipeline. Our approach achieves 96.39% accuracy with strong generalization across external datasets, and providing human-understandable explanations grounded in visual evidence.

1 INTRODUCTION

The rapid advancement of image generation models (Wang et al., 2025b; Li et al., 2025; Chadebec et al., 2025) has enabled the creation of AI-generated images with unprecedented photorealism, increasingly blurring the boundary between authentic and synthetic content. There is a critical need for platforms to deploy accurate and effective detection methods. However, the current landscape of detection methods is dominated by classification-based approaches. While often effective on specific datasets, these methods typically operate as "black- or gray-boxes", offering little insight into their decision-making process. This lack of explainability is coupled with poor generalizability, as models trained to detect artifacts from one generative architecture often fail when confronted with novel, unseen ones.

The emergence of Vision-Language Models (VLMs) (Fang et al., 2025; Chen et al., 2024; Man et al., 2025; Wu et al., 2025c;b; Zhang et al., 2025) has opened a new frontier, offering a promising path towards semantic-level analysis and greater interpretability. Many recent approaches re-frame the detection task as a visual question answering (VQA) problem (Gao et al., 2025) or an image captioning task (Keita et al., 2025). However, to reach high accuracy, these methods typically rely on external modules such as segmentation networks or additional classification heads (Zhou et al., 2025; Kang et al., 2025), which underutilize the intrinsic knowledge and common-sense reasoning already embedded in VLMs and reduce them to passive feature extractors. More fundamentally, most approaches perform a single global pass over the image: the visual encoder compresses the scene into a limited set of tokens, attention is spread across the entire image, and fine-grained forensic cues (*e.g.*, tiny text artifacts, stitching seams, periodic textures, specular edges) are weakened by downsampling and pooling (Park & Owens, 2024; Liu et al., 2023). Without revisiting localized regions to test hypotheses, small but decisive artifacts are often overlooked, resulting in unreliable judgments on high-quality synthetic images. As demonstrated in Figure 1(a), when the model cannot clearly perceive all details, it may rely on prior knowledge to "guess", producing flawed reasoning and incorrect verdicts.

In this work, we propose a paradigm shift from passively classifying real and AI-generated images to actively reasoning *with* them. Rather than training a model for a single, holistic classification task, our approach emulates the process of a human expert, who first identifies suspicious regions and then "zooms in" for a closer look. We bridged the reasoning and world knowledge of language models with the vision modality, building a system that inspects an image, hypothesizes about potentially synthetic regions, and then re-evaluates those specific regions to make a final, grounded decision.

Figure 1: (a) Without revisiting specific details, VLMs may overlook critical cues and produce false reasoning with incorrect decisions. (b) Our two-stage **ZoomIn** pipeline. The VLM first performs a global scan to query region(s) of interest (Query 1), then analyzes the cropped regions for a detailed, final verdict with grounded explanations (Query 2, "Local Evidence Check").

To support this objective, we construct a dataset that provides cropped regions and grounded explanations for training. As illustrated in Figure 1(b), our model delivers reliable reasoning and correct decisions under the zoom-in paradigm. Our experiments show that the proposed iterative, foveated approach enables a more robust and interpretable analysis. Our major contributions are threefold:

1. **A Paradigm Shift to "Think *with* Images" in Forensics:** We introduce a two-stage framework where a VLM first scans an image to hypothesize regions potentially indicating synthetic origin, and then performs a second, focused analysis on cropped regions to refine the verdict. This process grounds the final decision in explicit visual evidence.
2. **A Novel Data Annotation Pipeline:** Leveraging state-of-the-art VLMs (*e.g.*, GPT-4o) with detection-capable VLMs (*e.g.*, Qwen-2.5-VL), we construct MagniFake, a dataset of 20,000 real and generated images annotated with bounding boxes and fine-grained forensic explanations.
3. **Robust Generalization with Interpretability:** By directly linking the final verdict to visual evidence within identified bounding boxes, our method provides clear, human-intelligible explanations for its high accuracy robust to degradation and out-of-distribution data.

2 Related Works

Detection of AI-generated images. Image forgery detection has evolved from traditional feature engineering to modern deep learning methods. Early detection methods rely on handcrafted features and statistical anomalies (Li & Zhou, 2019; Chen et al., 2021; Frank et al., 2020) or perform frequency-domain analysis for identifying GAN-specific artifacts (Jeong et al., 2022). The advent of deep learning marked a paradigm shift, with CNN-based detectors such as CNNSpot (Wang et al., 2020) demonstrating remarkable generalization capabilities across various GAN architectures when trained on ProGAN-generated (Gao et al., 2019) images. As generative models evolved beyond GANs to include diffusion models (Song et al., 2020; Ho et al., 2020; Le et al., 2025; Ye et al., 2025), detection strategies adapted accordingly (Corvi et al., 2023; Tan et al., 2023; Park & Owens,

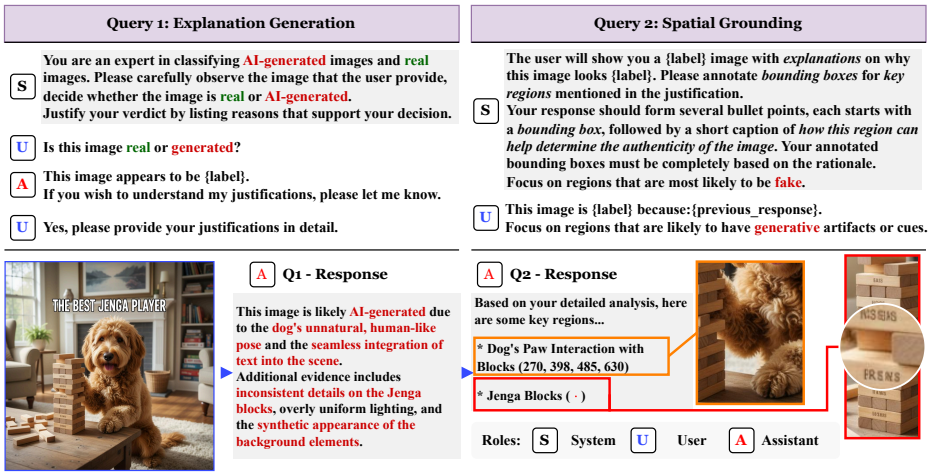


Figure 2: The proposed data annotation pipeline. We ask the forensics expert VLM in Query 1 "Explanation Generation" to identify key reasons that make this image look real or AI-generated, followed by Query 2 "Spatial Grounding", which uses the explanation to extract bounding boxes.

2024; Ricker et al., 2024). Notably, DIRE (Wang et al., 2023) pioneered the use of reconstruction error metrics specifically tailored for diffusion-generated content. NPR (Tan et al., 2023) leveraged frozen CLIP encoders to maintain domain invariance.

Despite these advances, explainability and robust generalization remain major challenges. The emergence of VLMs offers a new frontier by enabling semantic-level analysis and natural language reasoning. Many approaches re-formulate this classification problem to visual question answering (VQA) questions (Chang et al., 2023; Zhang et al., 2024b) or image captioning tasks (Keita et al., 2025). Several forensics datasets (Li et al., 2024; Zhang et al., 2024b; Gao et al., 2025) are curated using VLMs, creating training corpora that combine visual analysis with natural language reasoning. Specifically, Zhou et al. (2025) combines NPR (Tan et al., 2023) with LLM, achieving high detection accuracy with good interpretability. In terms of localization, FakeShield (Xu et al., 2024) introduces the Segment Anything (Kirillov et al., 2023) module to acquire the tampered mask for the manipulated image. LEGION (Kang et al., 2025) postfixes the vision encoder with an MLP to identify the authenticity of the input image. Our work builds on this direction by introducing spatial grounding and iterative refinement, transforming VLMs from passive analyzers into active visual investigators that fully leverage their inherent common-sense reasoning.

Training and Fine-Tuning Reasoning-Capable VLMs. Improving the reasoning abilities of VLMs is essential for tasks demanding sophisticated comprehension (Wu et al., 2025a;c; Yang et al., 2025a). Early approaches focused on transforming images into structured textual representations to facilitate language-based reasoning (Yang et al., 2025b). More recent studies have emphasized cultivating advanced cognitive skills, such as self-verification, self-correction, fostering "slow thinking" capabilities (Wang et al., 2025a), and regulating reasoning depth to mitigate issues like "overthinking" (Xiao et al., 2025). Additionally, efforts have concentrated on developing high-quality multi-modal Chain-of-Thought (CoT) datasets (Huang et al., 2025) to steer the reasoning process effectively. DeepSeek-Math (Shao et al., 2024) provides a solid foundation and methodology for fine-tuning large language models. For reward design, in addition to the outcome reward used in DeepSeek-Math, expert LLMs are also widely used as an online training reward provider (Lambert et al., 2024). Researchers are also experimenting with using IoU (Pan et al., 2025) and BLEU (Chang et al., 2025) metrics as rewards. Building on these advances, we train our VLM with spatially grounded supervision and iterative reasoning objectives, where cropped regions and fine-grained explanations guide the model to link decisions with explicit visual evidence, thereby enhancing both accuracy and interpretability.

3 METHODOLOGY

3.1 SCAN, CHECK AND VERDICT WITH THE "MAGNIFYING GLASS"

Recently, thinking *with* image methods equip VLMs with visual tools to enable structured interaction with images, allowing step-wise reasoning and improved interpretability (Su et al., 2025c; Shen et al., 2025; Fan et al., 2025; Su et al., 2025a; Cheng et al., 2025). Motivated by this idea, we design a detector that explicitly "zooms-in", encouraging the model to take an additional step of focused reasoning. Given an input image I, our goal is to build a VLM that outputs a final verdict $v \in \{\text{real}, \text{generated}\}$ along with interpretable evidence, including bounding boxes and explanations, by mimicking the human forensic process of scanning, localizing suspicious regions, and then re-examining them under a magnifying glass. To achieve this, we propose a two-stage inference pipeline inspired by how humans handle uncertainty: *when uncertain about the whole image, they re-examine and focus more closely on regions that are most likely to be fake in order to increase confidence*. Figure 1 shows the two-stage inference workflow. The setup is detailed below:

Query 1: Global Scan. Given an input image I, we prompt a grounding-capable VLM to perform comprehensive visual analysis. The model generates:

- Initial verdict $v_1 \in \{\text{real}, \text{generated}\}$.
- Suspicious regions $B = \{b_1, ..., b_n\}$ where $b_i = (x_1, y_1, x_2, y_2)$ represents bounding box.
- Preliminary explanation E_1 articulating the reasoning.

Since VLMs may overlook such regions due to resolution or encoding limitations, magnified and focused inspection enables refined predictions and more reliable explanations. This stage leverages the VLM's ability to process global context while identifying locally anomalous regions, encouraging the model to "think *with* images" (Su et al., 2025b). Figures 1 and 2 show examples for real and fake cases, respectively, illustrating the types of regions that warrant closer inspection. We focus on two types: (i) *regions inherently challenging for generative models*, such as human faces or hands (Fig. 1), and fine-grained animal attributes like paws or poses (Fig. 2); and (ii) *image-specific details that are difficult to reproduce*, including logos on a referee's shirt (Fig. 1) or small texts (Fig. 2). By zooming into these regions, the model reduces global uncertainty, strengthens prediction accuracy, and provides more reliable explanations.

Query 2: Local Evidence Check. For each identified region b_i, we first extract crops $C_i = \text{Crop}(I, b_i)$, and then provide the VLM with both the original image I and the crop collection $\{C_i\}_{i=1}^n$, enabling comparative analysis between global context and local details. This dual-input mechanism produces:

- Final verdict $v_2 \in \{\text{real}, \text{generated}\}$.
- Refined explanation E_2 grounded in specific visual evidence.

By providing these image crops C_i, we enrich the input with fine-grained visual tokens, allowing the model to correct earlier misjudgments through magnified analysis, much like a forensic expert using a "magnifying glass". This two-stage verification significantly enhances the accuracy and interpretability of the prediction. Enabling such reasoning requires training data with fine-grained annotations and supervision, which motivates the construction of our ***MagniFake*** dataset and the training of ***ZoomIn***.

3.2 MAGNIFAKE DATASET CONSTRUCTION

Our two-stage pipeline requires localization information for both real and AI-generated images. As discussed earlier, such information should highlight (i) regions inherently challenging for generative models and (ii) image-specific regions that are difficult to reproduce. While the zoom-in process mimics the intuition of human forensic experts, VLMs do not naturally possess this capability. To enable this behavior and improve both grounding accuracy and detection performance, we design a targeted training strategy based on (I, E, B, C) tuples, providing supervision for classification and grounding tasks within the pipeline. To acquire grounding-aware training data, we create the ***MagniFake*** dataset by leveraging different VLM experts to generate textual explanations and bounding

boxes. Previous works (Gao et al., 2025; Ji et al., 2025) and our preliminary tests have confirmed that the OpenAI GPT-4o can produce detailed forensic explanations describing why an image is real or AI-generated. Additionally, the Qwen-2.5-VL series of VLMs (Qwen Team, 2025) acan extract spatial regions from these explanations and output bounding boxes, making them well-suited for generating spatially grounded annotations. We address the lack of spatially grounded forensic annotations through an automated pipeline as follows:

1. **Explanation Generation**. For images with known labels, GPT-4o generates forensic explanations focusing on specific visual evidence. The prompt used at this stage is designed to elicit detailed reasoning about the depicted objects, arrangement, perspective, and other relevant aspects of the given images.
2. **Spatial Grounding.** Qwen-2.5-VL extracts bounding boxes from these explanations, creating (I, E, B, C) tuples.

An overview of the pipeline is provided in Figure 2, showing the prompts used in the automated annotation process.

Data Purification. During the spatial grounding phase, the Qwen-2.5-VL model occasionally generates bounding boxes that cover more than 50% of the image, associating them with over-saturated global image imperfections. In certain instances, these bounding boxes may revert to object detection when the primary object is clearly flawed but occupies a limited portion of the image. To ensure only high-quality fake regions with logical captions are included in MagniFake, we further leverage a VLM to filter out redundant or excessively large regions from the dataset. This filtration process evaluates whether the bounding box fully encapsulates the primary object (indicating a regression to object detection) or is disproportionately large, covering over 50% of the image, effectively removing them from the MagniFake dataset.

Image Distribution. MagniFake consists of 10,000 real images and 10,000 AI-generated images. All of which are annotated with explanations and spatial grounding information, among which 99.5% of the images have at least one bounding box, with an average of 3.24 bounding boxes per image after filtering. The real images are sourced equally from ImageNet (Deng et al., 2009) and COCO (Chen et al., 2015), and the AI-generated images are equally sourced from GPT-Image-1 (OpenAI, 2025a) and Gemini 2.5 Flash Image (Google, 2025).

3.3 Training Procedure of ZoomIn

With (I, E, B, C) tuples from ***MagniFake***, we fine-tune VLMs to serve as the ZoomIn detector. Inspired by DeepSeek-Math (Shao et al., 2024), our fine-tuning approach employs a two-phase training paradigm that combines supervised fine-tuning (SFT) with reinforcement learning (RL) implemented through Group Relative Policy Optimization (GRPO).

Supervised Fine-tuning Phase. The training begins with SFT to establish foundational capabilities and ensure stable model behavior. During this phase, all trainable parameters across the model's vision encoder, projection layers, and language modeling components undergo optimization using supervised signals from the dataset. This initial phase serves two critical purposes: (1) teaching the model to generate outputs that conform to our specified structured format, and (2) establishing baseline performance before applying reinforcement learning techniques.

Reinforcement Learning with Enhanced Rewards. Following SFT, we implement RL through two GRPO stages. Our reward design extends beyond traditional classification and localization metrics by incorporating linguistic quality assessment through BLEU scores (Chang et al., 2025). We define reward functions tailored to two distinct query stages, each addressing specific aspects of the model's performance. In the first query, the model generates an initial hypothesis, focusing on format compliance and localization precision. The reward structure is defined as follows:

- **Format Compliance:** The reward for correct output formatting is given by:

$$\mathcal{R}_F = 1 \text{ if output contains valid } \texttt{<verdict>} \text{ tags.} \tag{1}$$

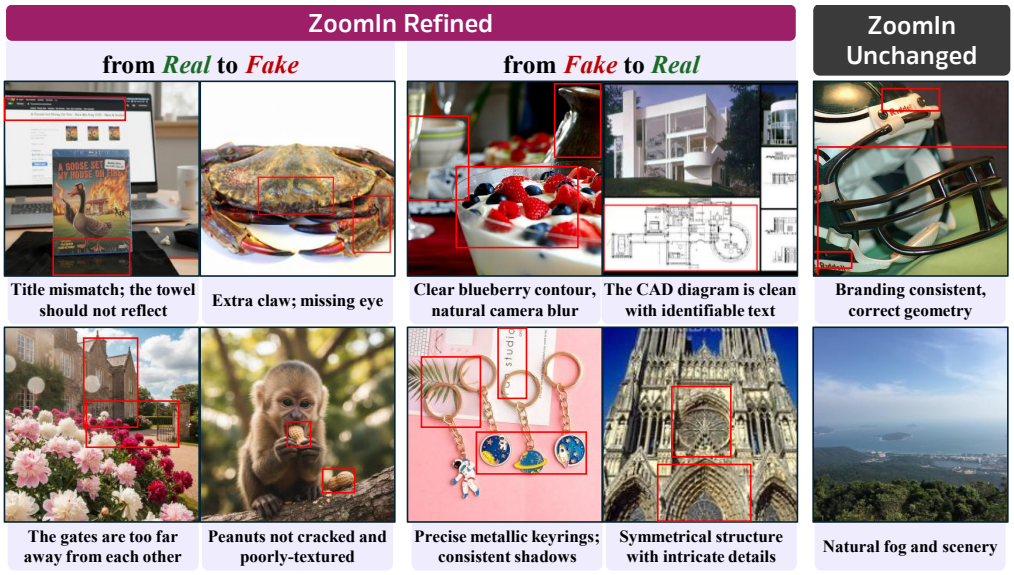


Figure 3: Examples from the test set of MagniFake, captions are summarized from the Query 2 response, generated by ZoomIn-32B.

- **Localization Precision:** Grounding accuracy is measured by IoU, rewarding precise spatial alignment. With b_i denoting model-predicted boxes and $\hat{b}_j$ annotated boxes, we compute:

$$\mathcal{R}_{\text{IoU}} = \frac{1}{|B|} \sum_i \max_j \text{IoU}(b_i, \hat{b}_j). \tag{2}$$

This reward structure ensures the model prioritizes accurate spatial localization and adherence to the expected output format during the hypothesis generation phase.

In the second query, the model refines its hypothesis, emphasizing correct verdict prediction and high-quality explanation generation. The reward structure comprises:

- **Classification Accuracy:** The binary reward for correct verdict prediction is defined as:

$$\mathcal{R}_C = \mathbb{1}[v_2 = y]. \tag{3}$$

- **Explanation Quality:** To encourage contextually appropriate explanations, we compute BLEU scores between generated explanations and reference texts:

$$\mathcal{R}_{\text{BLEU}} = \text{BLEU}_2(E', E_{\text{ref}}). \tag{4}$$

where E' is the explanation text from the model's output, and E_{ref} represents the reference explanation from our annotated dataset.

This reward structure drives the model to produce accurate verdicts and coherent, high-quality explanations during the refinement phase. Section 4.3 shows the model performance when BLEU reward is ablated from the GRPO procedure, or when a correct verdict counts as a reward in the first query.

4 Experiments

4.1 Setup

We use Qwen-2.5-VL (7B- and 32B-Instruct variants) (Qwen Team, 2025) trained on 8x NVIDIA A100 GPUs. The learning rates are set to 2×10^{-5} (SFT) and 10^{-5} (GRPO). For ablation, we evaluate single-turn variants of Qwen-2.5-VL-Instruct: the untrained base model (Base), a version using only explanations (E-), explanations with grounding but without refinement (E+G-), and zoom-in variants trained with SFT only, without GRPO. For baselines, we compare our models against traditional classification methods, including Community Forensics (Park & Owens, 2024), Antifake

Table 1: Detection accuracy and reasoning quality metrics on the MagniFake test set. E- stands for explanation-only, and E+G- stands for explanation and grounding-only.

Method	Acc.	I-Acc.	C-Acc.	C-Cases (%)	BLEU-1	BLEU-2	ROUGE-L	IoU
ZoomIn-7B	0.942	0.866	0.915	9.2	0.314	0.209	0.291	0.316
ZoomIn-32B	**0.972**	0.873	0.956	10.9	0.346	0.211	0.327	0.359
E-7B (one-turn)	0.855	-	-	-	0.280	0.146	0.301	-
E-32B (one-turn)	0.869	-	-	-	0.294	0.153	0.315	-
E+G-7B (one-turn)	0.906	-	-	-	0.275	0.136	0.269	0.245
E+G-32B (one-turn)	0.914	-	-	-	0.282	0.149	0.295	0.254
Base-7B	0.553	-	-	-	0.110	0.032	0.073	-
Base-32B	0.587	-	-	-	0.102	0.043	0.079	-
SFT-7B	0.719	0.724	0.447	4.8	0.156	0.042	0.129	0.094
SFT-32B	0.715	0.706	0.584	5.3	0.159	0.076	0.130	0.105
No BLEU Reward-32B	0.929	0.868	0.871	8.2	0.187	0.095	0.153	0.296
Verdict Only-32B (one-turn)	0.897	0.897	-	-	-	-	-	-
Dual Verdict Reward-32B	0.944	0.948	0.473	7.4	0.276	0.164	0.252	0.260
Random Cropping-32B	0.842	0.873	0.421	19.6	0.105	0.043	0.109	-
Largest 4 Bboxes-32B	0.934	0.873	0.938	7.0	0.303	0.182	0.158	-
Largest 3 Bboxes-32B	0.924	0.873	0.919	6.1	0.299	0.183	0.147	-
FakeShield (Xu et al., 2024)	0.801	-	-	-	0.097	0.056	0.067	0.096
LEGION (Kang et al., 2025)	0.654	-	-	-	0.102	0.058	0.054	0.061

Prompt (Chang et al., 2023), DIRE (Wang et al., 2023), CNNSpot (Wang et al., 2020) and NPR (Tan et al., 2023). For fair comparison, all baselines are retrained on MagniFake's training split. Due to their specific design, FakeShield (Xu et al., 2024) and LEGION (Kang et al., 2025) cannot be directly trained or fine-tuned on MagniFake because of incompatible dataset formats. Instead, we adopt the released pre-trained weights from FakeShield, and train LEGION on SynthScars (Kang et al., 2025) to reproduce their results.

4.2 Experimental Results

MagniFake Results. In Table 1, we report MagniFake results for all VLM-based methods, including all ZoomIn variants and the baselines LEGION (Kang et al., 2025) and FakeShield (Xu et al., 2024). Accuracy is reported on the test split of the MagniFake dataset. The results show that our method's accuracy surpasses all existing detection methods, even with the smaller 7B variant. Notably, compared to the original, untrained VLM baselines, our full pipeline with zoom-in reasoning yields an accuracy improvement of over 30%. Furthermore, when compared to single-turn variants, the zoom-in mechanism consistently contributes an additional 3.6% accuracy gain for the 32B and 5.8% for the 7B models, validating our core hypothesis that progressive visual reasoning enhances detection, reaffirming the paradigm shift to reason and think *with* images.

OoD Results. In addition to our own dataset, we also evaluate our models on out-of-distribution (OoD) datasets, including GenImage (Zhu et al., 2023), MMFR-Dataset (Gao et al., 2025) and SynthScars (Kang et al., 2025). For GenImage, we uniformly sample 10,000 images from the dataset and ensure that there is no image overlap with MagniFake as we also source some of the real images from ImageNet. Note that LEGION is trained on SynthScars, so its results on this dataset are not considered OoD. The scores are reported in Table 2. Our ZoomIn-32B model has an outstanding OoD performance, surpassing most baseline models and the 7B variant.

Reasoning Quality. The generated reasoning is assessed for similarity to the MagniFake dataset using BLEU-1, BLEU-2, and ROUGE-L (Lin, 2004) metrics. The grounding capability is evaluated by the IoU, which represents the grounding accuracy of our models. Table 1 shows that the explanation quality improves for all model variants when introducing the zoom-in mechanism, consistently improving the responses' alignment with the ground truth.

Discussions on refinement corrections. Figure 4 demonstrates the correlation of accuracy and the number of zoom-in regions. We found that after training, the ZoomIn-32B and 7B models exhibit different tendencies in Query 1. In Query 1, the ZoomIn-32B model outputs an average of 3.58 bounding boxes per image, while ZoomIn-7B gives 1.95. The 32B model tends to ask for two to four close-ups per image, summing up to 83.15% among all input samples; meanwhile, the

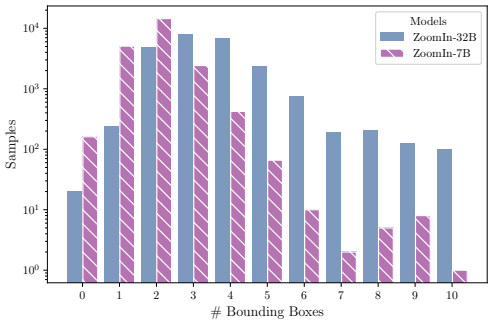


Figure 4: The number of bounding boxes in Query 1 for ZoomIn-32B/7B models.

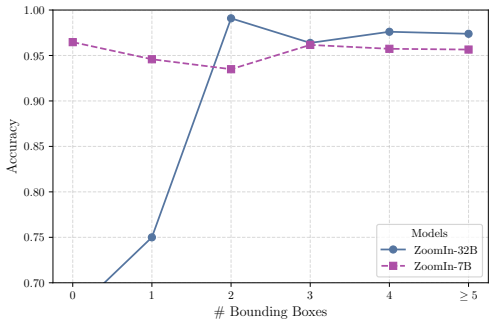


Figure 5: The relation of accuracy with regard to the number of detected bounding boxes.

Table 2: Accuracy (%) of ZoomIn and other comparing methods on MagniFake, GenImage (Zhu et al., 2023), MMFR-Dataset (Gao et al., 2025) and SynthScars (Kang et al., 2025).

Datasets	GenImage	MMFR	SynthScars	MagniFake	Average
ZoomIn-32B	0.915	**0.893**	0.852	**0.972**	**0.908**
ZoomIn-7B	0.878	0.892	0.826	0.942	0.885
CommunityForensics (Park & Owens, 2024)	0.854	0.838	0.760	0.876	0.832
Antifake Prompt (Chang et al., 2023)	**0.916**	0.862	0.819	0.927	0.881
DIRE (Wang et al., 2023)	0.872	0.889	0.815	0.922	0.874
CNNSpot (Wang et al., 2020)	0.853	0.846	0.807	0.855	0.840
NPR (Tan et al., 2023)	0.859	0.816	0.823	0.880	0.844
LEGION (Kang et al., 2025)	0.230	0.193	**0.861**	0.654	0.485
FakeShield (Xu et al., 2024)	0.864	0.710	0.765	0.801	0.785

7B model provides one or two bounding boxes for 86.34% of the cases. To further investigate the model performance with regard to the number of bounding boxes, Figure 5 shows that as the number of bounding boxes grows, the accuracy of ZoomIn-7B slightly lowers, while ZoomIn-32B has its accuracy spike at two bounding boxes, and performs consistently better than 7B on cases with more than two bounding boxes. This phenomenon aligns with the fact that most training samples have an average of 3.24 bounding boxes.

4.3 Ablation Studies

We conducted several ablation studies on the 32B model variant to validate our design choices. Table 1 shows the experimental results. We report three metrics: the initial accuracy of the first turn (I-Acc.), the proportion of images forwarded to the second turn (C-Cases), and the corrected accuracy after refinement in the second turn (C-Acc.).

Contribution of the Zoom-In Mechanism. For ZoomIn-7B and 32B models, 9.2% and 10.9% of the inputs will have a different verdict in Query 1 and Query 2 (i.e. $v_1 \neq v_2$). Among these corrected cases, the accuracy is 91.5% and 95.6%, respectively, proving that models can refine their initial verdict with accurate grounding prior.

Impact of BLEU Rewards. Removing linguistic quality rewards R_B ("No BLEU Reward") lowers the BLEU-n and ROUGE-L metrics as expected. In addition to these reasoning quality metrics, we also note a 4.3% accuracy drop and a 6.3% IoU drop. This demonstrates that encouraging the model to generate coherent explanations improves not just text quality, but also its underlying forensic reasoning capabilities.

Stage-Specific Rewards. We experimented with applying the classification accuracy reward ($\mathcal{R}_C$) during Query 1, in addition to Query 2. The result in row "Dual Verdict Reward" shows a slight performance degradation. The model became overly cautious, often failing to propose suspicious regions unless it was already highly confident in its initial verdict, thus undermining the purpose of the second-stage refinement. Specifically, with this reward setting, the zoom-in process will no longer benefit the performance, and the IoU was also lower than the original reward setup. This con-

firms that focusing Stage 1 rewards on localization is the optimal strategy in exploiting the VLM's intrinsic and learned capabilities in image forgery detection. We also attempt to use the verdict as the only reward and remove the entire zoom-in process ("Verdict Only"). This results in slightly lower accuracy when compared to the E+G group and is more likely to overfit on the MagniFake dataset.

Random Cropping. Our two-stage inference pipeline is dependent on both the original image and cropped images ("Random Cropping"). Therefore, we further ablate the bounding box selection stage and use random cropping instead. While the model remains as ZoomIn-32B, the bounding boxes given by the first query are no longer passed on to the second query. Evaluation shows that when zooming in on random regions, the accuracy can drop to 84.2%, 13.0% lower than the initial verdict. BLEU-n and ROUGE-L metrics also drop to a level similar to the base model. This indicates that without intelligent region proposal, the zoom-in mechanism is ineffective and the explanation quality is low.

Limiting Bounding Box Count. Since Figure 5 indicates that more bounding boxes may negatively impact model performance, we analyzed the impact of limiting the maximum number of proposed regions in Query 1 by selectively using the largest n bounding boxes from the response. As shown in Table 1, using the largest three or four bounding boxes can cause reduced overall performance, with a 1.8% and 3.7% drop in correction accuracy, respectively.

4.4 Qualitative Analysis

Figure 3 illustrates representative examples of successful outcomes, while Figure 6 presents a specific "corrected case". In this instance, ZoomIn-32B determines that the highly realistic and detailed rendering of window regions effectively refutes its initial assessment, thereby confirming that the image is likely a genuine photograph rather than an AI-generated picture. Certain images with small text that are not clearly identifiable at first sight, minor anatomical issues (*e.g.,* extra finger), or texture problems are highly likely to be corrected during the refinement query.

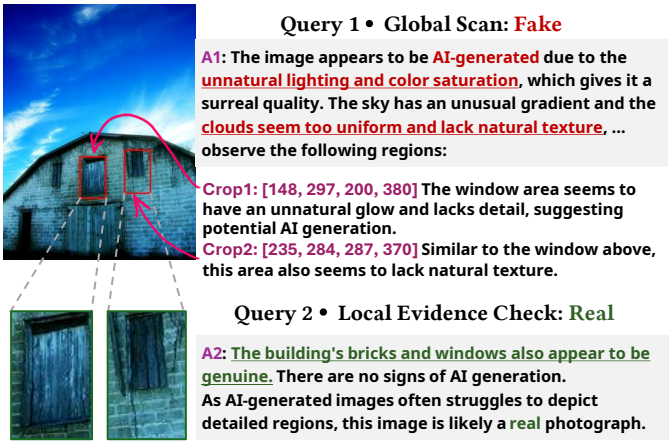


Figure 6: An example of where ZoomIn-32B corrects its initial mistake upon zooming-in, finalizing with the correct conclusion that this image is real.

We observe the following common explanation patterns across successful predictions:

Lighting inconsistencies (12.4%): Unnatural reflections, impossible shadows.
Anatomical anomalies (8.6%): Distorted fingers, impossible joint positions.
Texture artifacts & blurry text (8.1%): Unnatural skin textures, clothing patterns.
Perspective errors (4.7%): Impossible spatial relationships, such as overlapping object parts.

Failure cases often involve highly realistic AI-generated images, especially human or animal faces, where visual artifacts are subtle or imperceptible, even under magnification. Conversely, some real images with too complex textures or patterns are occasionally misidentified as synthetic.

5 Conclusion

We present the MagniFake dataset and the zoom-in mechanism in image forensics, mimicking human visual examination patterns. Our method achieves both high accuracy (97.2%) and interpretability through spatially grounded explanations. The success of this two-stage approach suggests that complex problems may benefit from decomposition into interpretable sub-tasks rather than end-to-end black-box solutions.

REFERENCES

Clement Chadebec, Onur Tasar, Eyal Benaroche, and Benjamin Aubin. Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Yapei Chang, Yekyung Kim, Michael Krumdick, Amir Zadeh, Chuan Li, Chris Tanner, and Mohit Iyyer. Bleuberi: Bleu is a surprisingly effective reward for instruction following. *arXiv preprint arXiv:2505.11080*, 2025.

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.

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. URL https://api.semanticscholar.org/CorpusID:2210455.

Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14185–14193, October 2021.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, et al. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *arXiv preprint arXiv:2505.15510*, 2025.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095167.

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*, pp. 248–255, 2009. URL https://api.semanticscholar.org/CorpusID:57246310.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*, 2025.

Wenlong Fang, Qiaofeng Wu, Jing Chen, and Yun Xue. guided mllm reasoning: Enhancing mllm with knowledge and visual notes for visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Hongchang Gao, Jian Pei, and Heng Huang. Progan: Network embedding via proximity generative adversarial network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 1308–1316, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330866. URL https://doi.org/10.1145/3292500.3330866.

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.

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.

Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. URL https://api.semanticscholar.org/CorpusID:219955663.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi. Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1060–1068, 2022.

Yikun Ji, Yan Hong, Jiahui Zhan, Haoxing Chen, jun lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang. Towards explainable fake image detection with multi-modal large language models, 2025. URL https://arxiv.org/abs/2504.14245.

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.

Mamadou Keita, Wassim Hamidouche, Hessen Bougueffa Eutamene, Abdelmalik Taleb-Ahmed, David Camacho, and Abdenour Hadid. Bi-lora: A vision-language approach for synthetic image detection. *Expert Systems*, 42(2):e13829, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

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.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Duong H Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. One diffusion to generate them all. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Jia Li, Lijie Hu, Jingfeng Zhang, Tianhang Zheng, Hua Zhang, and Di Wang. Fair text-to-image diffusion via fair mapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Yixuan Li, Xuelin Liu, Xiaoyang Wang, Shiqi Wang, and Weisi Lin. Fakebench: Uncover the achilles' heels of fake images with large multimodal models. *ArXiv*, abs/2404.13306, 2024. URL https://api.semanticscholar.org/CorpusID:269293612.

Yuanman Li and Jiantao Zhou. Fast and effective image copy-move forgery detection via hierarchical feature point matching. *IEEE Transactions on Information Forensics and Security*, 14(5): 1307–1322, 2019. doi: 10.1109/TIFS.2018.2876837.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

OpenAI. Gpt-image-1. *https://openai.com/index/image-generation-api/*, 2025a.

OpenAI. Introducing 4o image generation, Mar 2025b. URL https://openai.com/index/introducing-4o-image-generation/.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors, 2024. URL https://arxiv.org/abs/2411.04125.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9130–9140, 2024. URL https://api.semanticscholar.org/CorpusID:267335007.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. URL https://api.semanticscholar.org/CorpusID:267412607.

Chuming Shen, Wei Wei, Xiaoye Qu, and Yu Cheng. Satori-r1: Incentivizing multimodal reasoning with spatial grounding and verifiable rewards. *arXiv preprint arXiv:2505.19094*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020. URL https://api.semanticscholar.org/CorpusID:222140788.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025a.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025b.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025c.

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.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025a.

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*, pp. 8695–8704, 2020.

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*, pp. 22445–22455, 2023.

Zhendong Wang, Jianmin Bao, Shuyang Gu, Dong Chen, Wengang Zhou, and Houqiang Li. Design-diffusion: High-quality text-to-design image generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025b.

Mengyang Wu, Yuzhi Zhao, Jialun Cao, Mingjie Xu, Zhongming Jiang, Xuehui Wang, Qinbin Li, Guangneng Hu, Shengchao Qin, and Chi-Wing Fu. Icm-assistant: Instruction-tuning multimodal large language models for rule-based explainable image content moderation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025a.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. *arXiv preprint arXiv:2503.12799*, 2025b.

Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. Combating multimodal llm hallucination via bottom-up holistic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025c.

Wenyi Xiao, Leilei Gan, Weilong Dai, Wanggui He, Ziwei Huang, Haoyuan Li, Fangxun Shu, Zhelun Yu, Peng Zhang, Hao Jiang, et al. Fast-slow thinking for large vision-language model reasoning. *arXiv preprint arXiv:2504.18458*, 2025.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761*, 2024.

Fan Yang, Ru Zhen, Jianing Wang, Yanhao Zhang, Haoxiang Chen, Haonan Lu, Sicheng Zhao, and Guiguang Ding. Heie: Mllm-based hierarchical explainable aigc image implausibility evaluator. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025a.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025b.

Zilyu Ye, Zhiyang Chen, Tiancheng Li, Zemin Huang, Weijian Luo, and Guo-Jun Qi. Schedule on the fly: Diffusion time prediction for faster and better image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, et al. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

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*, pp. arXiv–2406, 2024a.

Zicheng Zhang, Haoning Wu, Chunyi Li, Yingjie Zhou, Wei Sun, Xiongkuo Min, Zijian Chen, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. A-bench: Are lmms masters at evaluating ai-generated images? *arXiv preprint arXiv:2406.03070*, 2024b.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. *arXiv preprint arXiv:2507.02664*, 2025.

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.

A Dataset Appendix for MagniFake

The MagniFake dataset comprises 10,000 images (5,000 real and 5,000 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.

A.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 (Deng et al., 2009) and COCO (Chen et al., 2015). 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. All synthetic images are created using the OpenAI GPT-Image-1 model (OpenAI, 2025b), representing 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.

A.2 Annotation Process

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

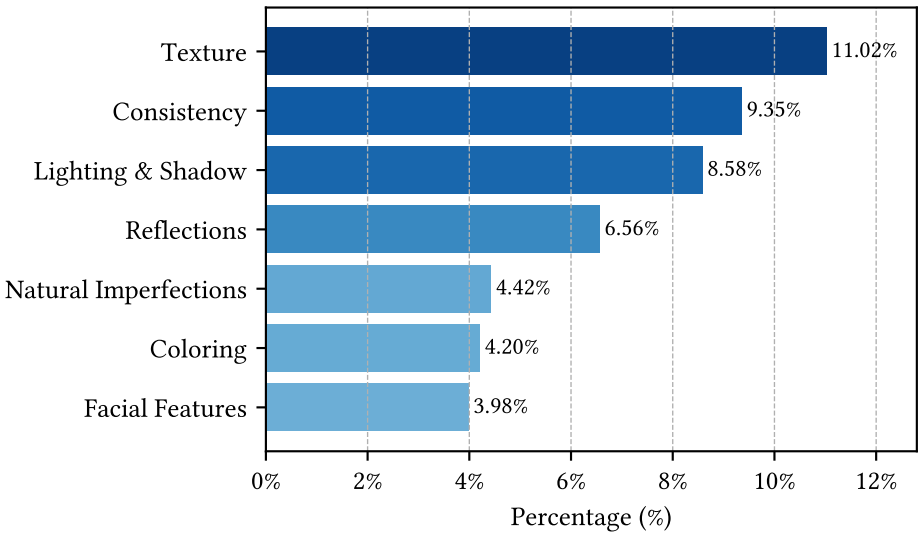
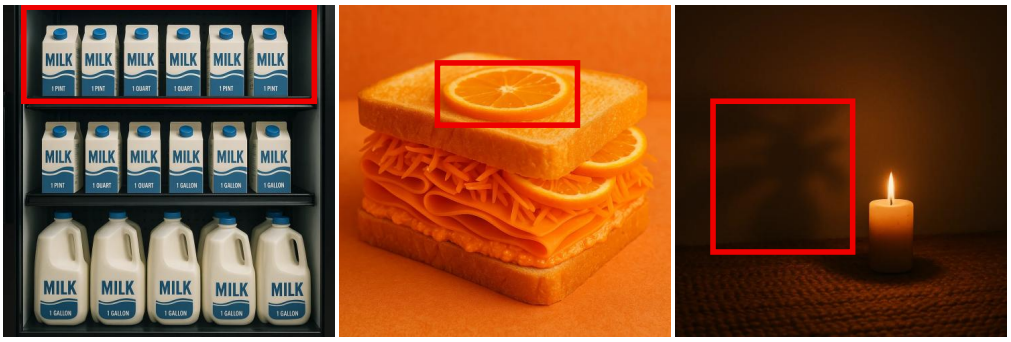


Figure 7: A statistical analysis of keywords in MagniFake explanations.

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. 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.



The top shelf contains milk with different serving sizes (1 PINT and 1 QUART), but they look identical.

Unpeeled orange slices are rarely used as a sandwich ingredient.

There's nothing between this candle and the shadow on the wall.

More AI-Generated Samples:

More Real Samples:

Figure 8: A collection of images from MagniFake 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.

Spatial Grounding. Qwen-2.5-VL extracts bounding boxes from the GPT-4o generated explanations, creating spatially grounded annotations in the form of (I, E, B, C) tuples, where I represents the image, E the explanation, B the bounding boxes, and C the cropped regions. 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.

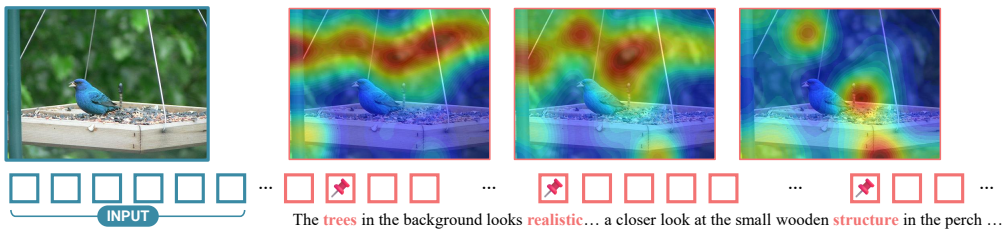


Figure 9: A visualization of the attention mechanisms of the VLM in Query 1.

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 7 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.

A.3 More Samples from MagniFake

Figure 8 presents additional examples from the MagniFake 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. MagniFake 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.

A.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.

A.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.

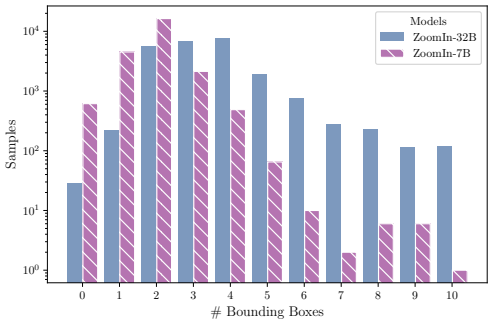


Figure 10: Number of samples grouped by the bounding boxes on OoD datasets.

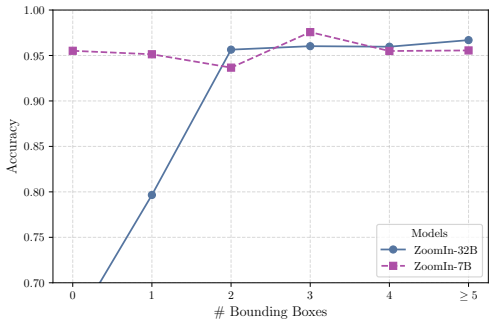


Figure 11: The relation of accuracy with regard to the number of detected bounding boxes on OoD datasets.

B 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 (Zhang et al., 2024a) heatmaps to highlight the regions of the image that most strongly influence the model's output predictions. As shown in Figure 9, 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 zoom-in 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 zoom-in regions that support reliable second-stage analysis.

C 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. During the GRPO stage, the number of generations is set to 2. For ZoomIn-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 ZoomIn-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 (Wang et al., 2020) 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 MagniFake. The training settings are the same as described in the original work.

Table 3: Performance on MagniFake with degradation, including JPEG compression artifacts, random cropping and image down-sampling.

Degradation	Metric	**ZoomIn**	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	-	-	-	-	-

Community Forensics (Park & Owens, 2024) 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 (Wang et al., 2023) 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 (Chang et al., 2023) 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 (Tan et al., 2023) 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 MagniFake with the same setup as the original work.

D Analysis of Bounding Boxes on OoD Datasets

We collected the responses from ZoomIn models when evaluating on OoD datasets, GenImage (Zhu et al., 2023), MMFR-Dataset (Gao et al., 2025) and SynthScars (Kang et al., 2025). On these OoD datasets, Figure 10 and 11 display the relation of bounding boxes with regard to model performance, and the number of detected bounding boxes for each ZoomIn model variant (7B and 32B). The trend of Figure 10 highly resembles *Figure 5* in the main paper, while Figure 11 is slightly different than *Figure 6*, where 32B performs better than 7B for cases with more regions selected. This proves that knowledge from the MagniFake dataset can be adapted beyond the dataset, and 32B generalizes better than 7B on OoD datasets.

E Robustness Against Degradations

We evaluate the robustness on MagniFake under four common degradations: JPEG compression at 80% and 30% quality, random cropping, and 0.5$\times$ downsampling (Table 3). All methods exhibit performance drops relative to clean images, with the extent varying across perturbations and models.

ZoomIn 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 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 robustness, 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.

F Computational Efficiency

Table 4 lists the end-to-end inference time for our trained VLMs. We use vllm (Kwon et al., 2023) to accelerate the inference process. The deployment of ZoomIn-32B takes 4x NVIDIA A100-40G GPUs connected with PCI-E. ZoomIn-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.

Table 4: Mean and standard deviation of inference time per image.

Method	Seconds / Image
ZoomIn-32B	**24.0±3.0**
E-32B (one-turn)	10.9±2.2
ZoomIn-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

G Limitations & Future Work

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 represents 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.

H The Use of Large Language Models

For this submission, we used large language models solely as writing assistants to improve grammar and fluency. They were not used for research ideation, methodological design, or generation of experimental results. The research itself studies vision-language models as part of the proposed forensic framework, including their use in dataset annotation, which is integral to the contribution of this work. All ideas, experiments, and analyses were conceived, designed, and verified by the authors, who take full responsibility for the content of the paper.