EXPLAINABLE AI-GENERATED IMAGE DETECTION WITH MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Detecting AI-generated images has become increasingly challenging as modern generative models produce highly realistic content. While existing detectors achieve reasonable classification accuracy, they often suffer from limited interpretability and poor generalization to unseen generators, especially in real-world scenarios. Recently, multimodal large language models (MLLMs) have shown promise for explainable AI-generated image detection; however, their effectiveness is constrained by the lack of large-scale, high-quality explanatory datasets. In this work, we introduce Fake-Reason, a large-scale and high-quality dataset for AI-generated image detection with 80K grounded samples provide fine-grained artifact explanations and precise bounding boxes. Unlike datasets dominated by recent generators, Fake-Reason deliberately spans early GANs and diffusion models as well as modern generators. Based on Fake-Reason, we fine-tune a lightweight open-source MLLM and identify an off-region visual shortcut in standard supervised training: the model can rely on unrelated image regions to infer the global label and then fabricate a plausible explanation for the target region. To alleviate this problem, we introduce region-level attention supervision that aligns the attention of each explanation with its annotated bounding box. Extensive generalization experiments and real-world evaluations demonstrate that our resulting detector, RAFE, achieves state-of-the-art detection performance while providing accurate localization and human-verifiable explanations. We believe that Fake-Reason and the proposed training strategy offer a strong foundation for advancing explainable and robust AI-generated image detection.

1 INTRODUCTION

With the rapid advancement of generative models, AI-generated images have become increasingly photorealistic, blurring the boundary between synthetic and authentic visual content Roca et al. (2025). Although AI image generation has brought significant benefits to various domains such as digital art, entertainment, and content creation, it simultaneously poses substantial risks to information integrity, privacy, and social trust.

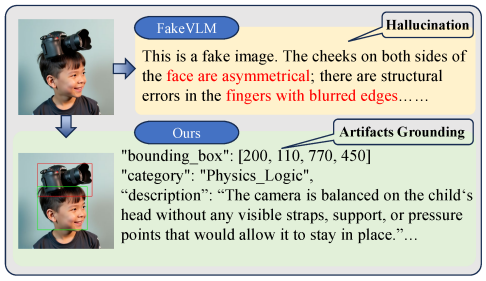


Figure 1: Comparison of the detection results produced by our detector, RAFE, and FakeVLM. RAFE provides more accurate explanations as well as more detailed localization information.

To address this issue, an increasing number of AI-generated image detection methods have been proposed. To date, most existing approaches formulate the task directly as a binary classification problem Wang et al. (2020); Ojha et al. (2023); Tan et al. (2024b); Yan et al. (2025b)Ḍespite achieving promising classification performance, existing approaches suffer from two major limitations. First, the **lack of explanation**. These black-box detectors are unable to provide human-verifiable explanations for their predictions, which significantly undermines their credibility and practical applicability. Second, **insufficient performance**. With the rapid advancement of genera-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tive models, AI-generated images have become increasingly realistic, making it necessary to analyze images from multiple perspectives to reliably distinguish real from fake content. However, formulating the problem as a direct binary classification task often causes models to overfit to low-level feature distributions, resulting in poor generalization to unseen generators, particularly in real-world scenarios.

The rapid advancement of multimodal large language models (MLLMs) has made explainable AI-generated image detection increasingly feasible, leading to the emergence of recent detection methods built upon MLLMs. However, fine-tuning multimodal models for this task requires high-quality explanatory datasets. Existing works, such as FakeClue and Holmes-Set, rely on open-source multimodal large language models to annotate data in order to construct large-scale datasets. Nevertheless, since current open-source MLLMs perform significantly worse than commercial models on AI-generated image detection tasks, the quality of these datasets is difficult to guarantee. Models trained on such datasets are prone to severe hallucinations, as illustrated in Figure 1. Although a few works resort to human annotation to ensure data quality, the substantial labor required for this task severely limits the dataset scale, making it impractical for large-scale model training. As a result, the field still lacks a large-scale, high-quality dataset with reliable and verifiable explanations.

We further identify an overlooked issue in the generator composition of existing MLLM-based training data. Recent datasets and methods, including AnomReason Tan et al. (2026a) and Veritas++ Tan et al. (2026b), emphasize images produced by strong contemporary generators or perception tasks designed around recent content. Such images often contain fewer conspicuous object deformations or violations of physical logic, leaving mainly subtle, low-level statistical traces. We observe that detectors trained predominantly on this distribution can perform well on benchmarks containing similar modern generators yet fail on images from earlier GANs or diffusion models, even when those images exhibit more obvious structural artifacts. This counter-intuitive behavior suggests that the detectors may rely on generator fingerprints rather than learning generalizable cues that distinguish authentic from synthetic objects. We therefore argue that images from earlier generators remain valuable training data rather than obsolete easy examples.

To address these issues, we propose Fake-Reason, which deliberately covers the evolution of image generation. Its core subset consists of 80K images annotated with rich forgery explanations ranging from low-level texture artifacts (e.g., over-smoothing) to high-level semantic inconsistencies (e.g., violations of physical logic). We intentionally include outputs from earlier generators such as ProGAN and Stable Diffusion v1.5 so that the model can learn explicit structural evidence of synthetic objects. Each explanation is associated with a precise bounding box, enabling simultaneous forgery localization and explanation. We further add 10K realistic samples derived from Pico-Banana-400K with only real/fake labels, covering modern generated images for which a reliable region-level explanation may not exist. To further improve annotation quality, unlike prior works that rely on open-source MLLMs for labeling, we leverage state-of-the-art commercial multimodal models for annotation to ensure the reliability and accuracy of the labels.

Based on this dataset, we fine-tune a lightweight open-source multimodal large language model. We observe that standard supervised learning permits an off-region visual shortcut. Because an AI-generated image is often synthetic as a whole, when explaining a particular object the model may attend strongly to unrelated objects or the background, infer the global label from those regions, and fabricate a plausible explanation without learning the actual artifacts inside the target box. To alleviate this issue, we introduce region-level attention supervision. For every annotated description, our objective encourages its output tokens to allocate sufficient visual attention to the corresponding bounding box, using fractional patch–box overlap to provide spatially precise supervision. Through extensive generalization experiments and evaluations in real-world scenarios, the resulting model, RAFE, demonstrates superior detection performance.

In summary, our contributions are threefold:

- **Dataset:** We construct **Fake-Reason**, a high-quality 90K-sample dataset comprising 80K grounded samples with detailed explanations and 10K label-only samples from a modern generator. By deliberately covering both early generators with conspicuous structural artifacts and modern generators with subtle statistical traces, Fake-Reason mitigates generator-era bias and encourages detectors to learn visual authenticity rather than generator-specific fingerprints.

- **Detector:** By fine-tuning a multimodal large language model on Fake-Reason, we obtain a detector, **RAFE**, capable of simultaneous forgery detection, explanation, and localization. Our region-level attention supervision grounds each explanation in its annotated visual evidence and reduces off-region shortcut learning.
- **Performance:** Our method achieves state-of-the-art detection performance across images generated by a wide range of advanced generative models, and demonstrates significantly superior performance in real-world scenarios compared to existing detectors.

2 Related Works

2.1 Generated Image Detection

AI models, due to limitations in their computational and architectural design (e.g., upsampling operations), often leave subtle fingerprints in generated images that are imperceptible to the human eye Dzanic et al. (2020); Durall et al. (2020). This observation has motivated a line of image-based detection methods that focus on low-level features Liu et al. (2022); Nguyen et al. (2025); Cozzolino et al. (2024b). Particularly, NPR Tan et al. (2024b) models discrepancies among neighboring pixels through image resampling operations, PiD Fu et al. (2025) extracts pixel-level residual noise via color space transformations and quantization, and SAFE Li et al. (2024) captures high-frequency components using discrete wavelet transforms. Such high-frequency–oriented features have been shown to be effective for detecting images produced by a wide range of generative models.

Another line of research leverages pretrained models to extract features and performs classification of AI-generated images in the semantic space. Representative works include UnivFD Ojha et al. (2023) , which observes a clear and easily separable distribution gap between features of real and generated images extracted by CLIP Radford et al. (2021). C2P-CLIP Tan et al. (2025) further injects forgery-related concepts through the text encoder of CLIP. Effort Yan et al. (2025b) proposes to preserve the original knowledge of the model by freezing its principal components via singular value decomposition.

2.2 Multimodal Large Model

The development of Large Language Models (LLMs) Touvron et al. (2023a;b) and Vision-Language Models (VLMs) Liu et al. (2023); Zhang et al. (2024) has significantly enhanced multimodal understanding, enabling seamless integration between visual and textual data. The LLaVA series Liu et al. (2023; 2024a) enhances visual question answering by aligning visual features with textual representations.The Qwen-VL series Wang et al. (2024); Bai et al. (2025) is a family of open-source multimodal large models that integrate visual encoders with large language models, providing strong general-purpose vision–language understanding and reasoning capabilities.

Recent progress in the fusion of multimodal data has notably improved the performance of models in detecting ai-generated images Chang et al. (2023); Huang et al. (2024); Zhou et al. (2025).ForgeryGPT Liu et al. (2024b) combines advanced forensic knowledge with a mask-oriented forgery extractor to enhance both detection and localization, achieving pixel-level identification of manipulated regions. Meanwhile, FakeShield Xu et al. (2025)leverages the capabilities of LLaVA to identify and localize tampered areas while providing interpretable analytical results. FakeVLM Wen et al. (2025) constructs a large-scale explanatory dataset using open-source multimodal large models and then trains a multimodal large model on this dataset for AI-generated image detection. Veritas++ Tan et al. (2026b) strengthens fine-grained visual, semantic, and pixel-level perception.

3 Dataset

Existing explanation-oriented datasets are commonly annotated by open-source MLLMs Wen et al. (2025); Zhou et al. (2025). Because these models remain substantially weaker than commercial systems for AI-generated image detection Li et al. (2025), they are often given the ground-truth label before producing an explanation, which can encourage result-conditioned hallucination. Human annotation improves reliability but is difficult to scale Ye et al. (2025); Li et al. (2025).

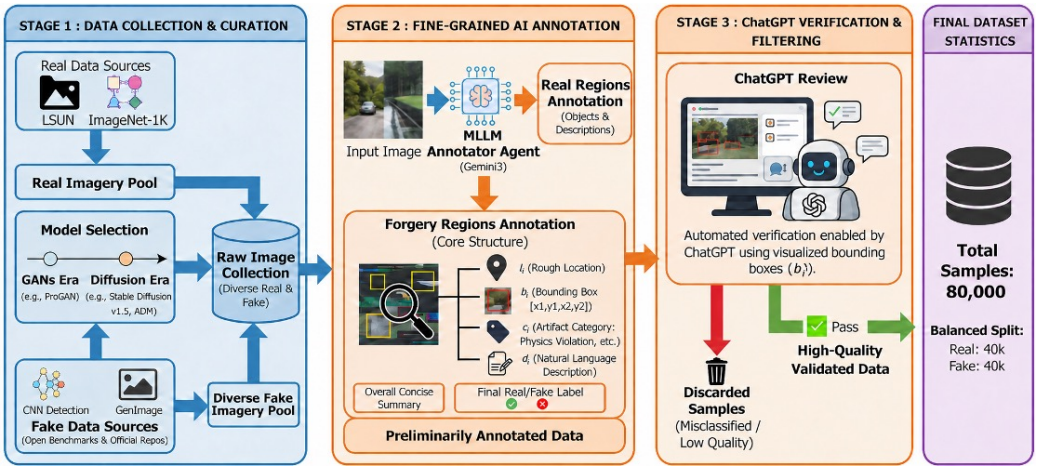


Figure 2: Construction pipeline for the 80K grounded subset of Fake-Reason.

To address these limitations, we construct **Fake-Reason**, a balanced dataset of 90K real and generated images. Its core 80K samples contain structured region-level annotations. Unlike datasets dominated by recent generators, Fake-Reason deliberately covers different generator eras. We retain outputs from earlier models, such as ProGAN Karras et al. (2018) and Stable Diffusion v1.5 Rombach et al. (2022), whose conspicuous structural artifacts provide direct supervision for learning visual authenticity rather than generator-specific fingerprints.

Modern generated images, however, frequently contain no localized object deformation or clear violation of physical logic, making region-level natural-language rationales unreliable. We therefore add 10K samples from Pico-Banana-400K Qian et al. (2026), consisting of 5K original photographs and 5K edited images, and provide only real/fake labels for this subset. These label-only samples expose the model to subtle statistical differences between highly realistic generated and real images without forcing hallucinated explanations.

For the grounded subset, Gemini 3.0 produces initial object-level explanations and bounding boxes for both real and generated images. We then remove misclassified samples and use ChatGPT-5.0 together with manual inspection to filter incorrect or low-quality explanations. The resulting 80K grounded samples support detection, localization, explanation, and region-level attention supervision, whereas the additional 10K samples contribute only to the SFT objective. Full data sources, annotation format, verification procedure, construction pipeline, and comparisons with existing datasets are provided in Appendix A.

4 Method

Current open-source multimodal large language models (MLLMs) remain unreliable at identifying AI-generated content without task-specific adaptation. We therefore fine-tune an MLLM on Fake-Reason using supervised learning. As illustrated in Figure 3, in addition to the standard language-modeling objective, we introduce region-level attention supervision to ground each forgery explanation in the visual region it describes.

4.1 Supervised Fine-Tuning

Given an image I, a prompt q, and an annotated response $\mathbf{y} = (y_1, \ldots, y_T)$, we optimize the standard autoregressive objective over the complete response:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{T}\sum_{t=1}^{T} \log P_\theta(y_t \mid I, q, \mathbf{y}_{<t}). \tag{1}$$

For the 80K grounded samples, the response contains the global real/fake decision as well as multiple region-level descriptions paired with bounding boxes. For the 10K Pico-Banana-400K samples, the

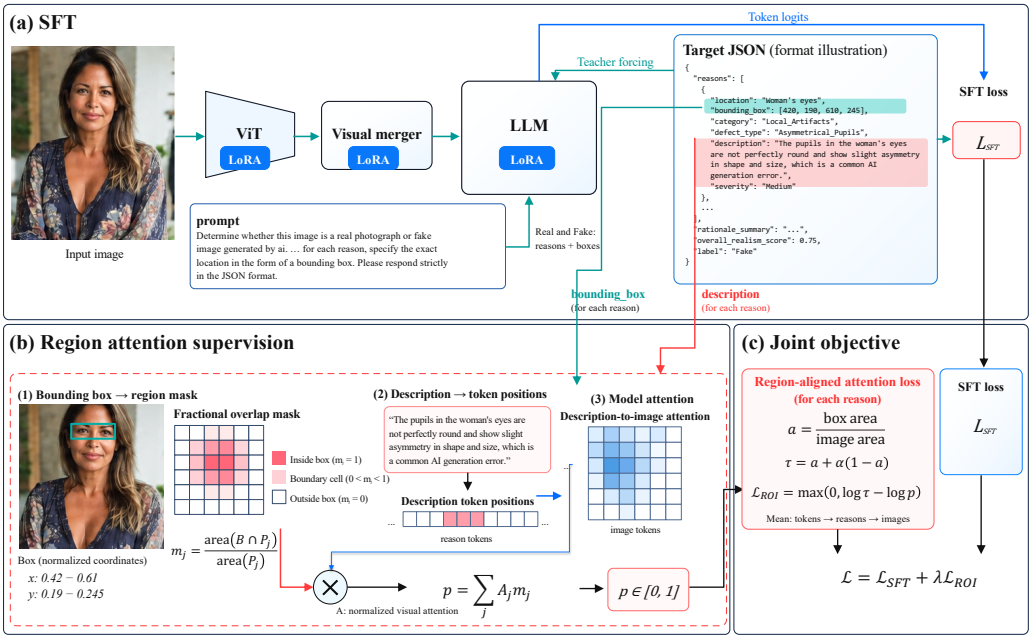


Figure 3: Overview of the proposed training framework. (a) The MLLM is supervised using structured responses containing authenticity labels, region descriptions, and bounding boxes. (b) Region-level attention supervision aligns each description with its corresponding visual evidence using fractional patch–box overlap. (c) The final objective combines the standard SFT loss with the region-aligned attention loss.

response contains only the authenticity label. Both subsets contribute to $\mathcal{L}_{\text{SFT}}$, whereas the region-level objective below is applied only to samples with grounded descriptions.

Risk of Off-Region Visual Shortcuts. Directly applying standard SFT does not guarantee that a region-level explanation is derived from its corresponding region. Because an AI-generated image is typically synthetic as a whole, when generating an explanation for one object, the model may assign excessive attention to other objects or the background. It can then infer the global image label from these unrelated regions and fabricate a plausible explanation for the queried region without learning its actual artifacts. This shortcut weakens the grounding and verifiability of the generated explanations.

4.2 Region-Level Attention Supervision

To prevent the model from explaining a target region using evidence from unrelated image areas, we explicitly supervise the visual attention of each region description. For description token t associated with bounding box B_r, we aggregate its attention $A_{t,j}$ to visual token j over all heads and all Transformer layers of the backend LLM. Let $m_{r,j} \in [0,1]$ denote the fraction of visual token j covered by B_r. The in-region attention ratio is

$$p_{r,t} = \frac{\sum_{j=1}^{HW} A_{t,j} m_{r,j}}{\sum_{j=1}^{HW} A_{t,j}}. \tag{2}$$

To account for region size, we introduce a tunable attention-margin coefficient α and set an area-adaptive target τ_r. We penalize only insufficient in-region attention:

$$\tau_r = a_r + \alpha(1 - a_r), \qquad \ell_{r,t} = \max(0, \log \tau_r - \log p_{r,t}), \tag{3}$$

where a_r is the image-area ratio of B_r. Averaging $\ell_{r,t}$ over description tokens, regions, and images gives $\mathcal{L}_{\text{ROI}}$. The complete response retains the standard SFT objective, with a tunable loss

coefficient λ:

$$\mathcal{L} = \mathcal{L}_{\text{SFT}} + \lambda \mathcal{L}_{\text{ROI}}. \tag{4}$$

This objective encourages each explanation to use its annotated visual evidence without constraining the model's full response. Detailed token mapping, attention aggregation, and loss computation are provided in Appendix B.

	Synthbuster					New Generator					
Method	DALLE3	Firefly	MJv5	SDXL	Avg	Infinity	FLUX	SD3	GPT-Image-2	Nano Banana 2	Avg
FreqNet Tan et al. (2024a)	43.9	47.6	61.45	82.6	58.9	68.65	67.3	67.2	66.9	47.8	63.6
NPR Tan et al. (2024b)	55.6	47.5	86.8	89.0	69.7	92.4	92.8	91.7	**94.4**	65.0	87.3
CNNSpot Wang et al. (2020)	34.4	52.9	59.5	54.9	50.4	61.8	53.9	64.6	44.2	45.0	53.9
UnivFD Ojha et al. (2023)	49.3	65.5	42.8	56.0	53.4	60.6	43.1	51.8	70.8	75.2	60.3
Effort Yan et al. (2025b)	68.8	84.9	66.9	78.5	74.8	85.0	69.6	79.5	88.9	79.1	80.4
BFree Guillaro et al. (2025)	98.4	**99.9**	**99.8**	**100.**	**99.5**	98.5	67.0	**99.6**	48.2	83.3	79.3
FakeVLM Wen et al. (2025)	57.1	22.6	49.1	50.6	44.9	57.1	56.6	55.6	80.8	90.8	68.2
Veritas++ Tan et al. (2026b)	**100.**	66.5	91.6	97.3	88.9	99.9	99.9	98.2	55.8	64.0	83.6
RAFE (Ours)	99.5	96.2	98.1	99.3	94.6	**100.**	**100.**	99.4	90.3	**95.6**	**97.1**

Table 1: Accuracy comparisons (%) across different generators on Synthbuster and New Generator.

	SynthWildX				WildRF				Chamelon
Method	DALLE3	Firefly	MJv5	Avg	Facebook	Reddit	Twitter	Avg	
FreqNet Tan et al. (2024a)	45.3	49.7	45.2	46.7	57.5	64.2	45.8	55.8	53.3
NPR Tan et al. (2024b)	48.8	50.6	48.8	49.4	54.4	77.6	55.1	62.4	51.7
CNNSpot Wang et al. (2020)	53.8	58.6	64.7	59.0	47.5	62.0	59.7	56.4	60.0
UnivFD Ojha et al. (2023)	55.0	69.0	55.9	60.0	61.6	61.2	68.4	63.7	60.6
Effort Yan et al. (2025b)	56.1	77.6	61.9	65.2	65.7	67.8	72.4	68.6	52.6
BFree Guillaro et al. (2025)	95.7	**95.7**	**95.6**	**95.7**	95.4	86.5	96.2	92.7	76.3
FakeVLM Wen et al. (2025)	68.0	67.6	68.1	67.9	59.1	69.2	63.0	63.8	65.6
Veritas++ Tan et al. (2026b)	94.7	84.8	93.7	91.1	95.0	95.9	**97.9**	96.3	**91.3**
RAFE (Ours)	**96.0**	87.3	95.3	92.8	**97.0**	**97.0**	97.5	**97.2**	90.5

Table 2: Accuracy comparisons on SynthWildX, WildRF, and Chameleon datasets.

5 Experiments

5.1 Experiment Setup

Baselines. We compare our method with three categories of detectors, including three approaches based on low-level features: FreqNet Tan et al. (2024a)(AAAI'24) and NPR Tan et al. (2024b)(CVPR'24),CNNSpot Wang et al. (2020)(CVPR'20), three approaches based on pre-trained models: UnivFD Ojha et al. (2023)(CVPR'23), and Effort Yan et al. (2025b)(ICML'25),BFree Guillaro et al. (2025), and two detectors built upon multimodal large language models: FakeVLM Wen et al. (2025)(NIPS'25) and Veritas++ Tan et al. (2026b). For the first two categories of methods that are trained via direct binary classification, we re-train the models on our dataset using their official codebases to ensure a fair comparison (except for BFree, which was trained on their proprietary dataset, we directly use the publicly released model weights.). In contrast, methods based on multimodal large language models are typically trained on their own specifically constructed large-scale datasets; therefore, we directly adopt their officially released pre-trained models for evaluation.

Datasets. To comprehensively evaluate the performance of our method, we conduct extensive experiments on a diverse set of large-scale datasets. **Synthbuster** Bammey (2023) consists of 1K real images from RAISE Dang-Nguyen et al. (2015) and synthetic images generated by advanced generative models such as DALLE3 Betker et al. (2023). All images are stored in RAW format to avoid artifacts introduced by JPEG compression. **Chameleon** Yan et al. (2025a) collects a large number of real images and high-quality synthetic images from online platforms, which are carefully refined by professional photographers and AI artists. **New-Generator:** Considering the rapid evolution

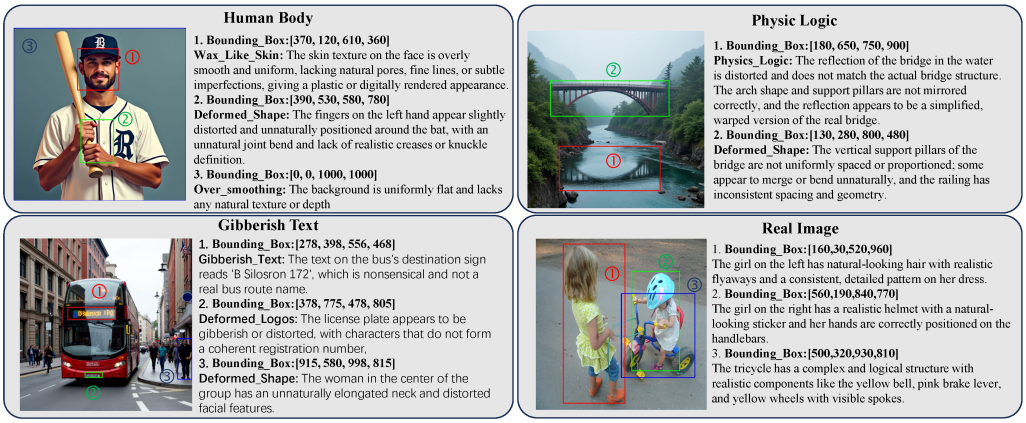


Figure 4: Synthetic and real image detection cases.We explicitly visualize the bounding boxes predicted by the model to facilitate inspection.

of generative models, we further evaluate our method on images produced by several state-of-the-art generators, including FLUX Labs (2024), Infinity Han et al. (2025), GPT-Image Schulman et al. (2022), Stable Diffusion 3 Esser et al. (2024), as well as the latest generators GPT-Image-2 and Nano Banana-2. Details of the construction of this evaluation set are provided in Appendix C. In addition, to assess the performance of our detector under real-world conditions, we include two datasets collected from social media platforms: **WildRF** Cavia et al. (2024) and **SynthWildX** Cozzolino et al. (2024a), which contain both real and AI-generated images.

Implementation Details. We adopt Qwen3-VL-8B-Instruct Bai et al. (2025) as the backbone model and apply LoRA with rank $r = 16$ and scaling factor $\alpha_{\text{LoRA}} = 32$. We set the attention-margin coefficient to $\alpha = 0.2$ and the auxiliary-loss coefficient to $\lambda = 0.05$. The learning rate is set to $1\text{e}{-5}$ with warm-start initialization. All experiments are conducted on four NVIDIA A100 GPUs.

5.2 Generalization Evaluation

To evaluate the generalization ability of different detectors, we benchmark their performance on Synthbuster and New Generator. The quantitative results are reported in Table 1.

It is worth noting that our training data are primarily derived from GenImage, which is also a widely adopted benchmark dataset in prior work. As shown in the results, most detectors exhibit noticeably degraded performance when evaluated on images generated by these more advanced generative models.

A key observation is that, unlike GenImage, the real images in Synthbuster and New Generator are preserved in their original formats without compression. This distribution discrepancy appears to be an important factor contributing to the poor generalization performance of most detectors. Our findings further corroborate previous studies showing that detection models tend to overfit to dataset-specific distribution biases.

Among the compared methods, NPR mitigates the influence of semantic content by explicitly modeling differences between adjacent pixels, thereby achieving relatively stronger detection performance. Although FakeVLM also leverage multimodal large language models fine-tuned for this task, we observe that it similarly overfit to the training distribution, leading to suboptimal generalization. In contrast, our detector consistently outperforms the other methods by a clear margin across all evaluated benchmarks.

5.3 Robustness in the Wild

As AI-generated images are increasingly disseminated across the Internet, evaluating detectors under real-world conditions is essential to ensure their practical applicability. In realistic scenarios, images shared on social media platforms are often subject to various degradations, such as compres-

Table 3: Accuracy comparisons on GenImage.

Method	BigGAN	SD v1.5	DALL-E 2	ADM	VQDM	Midjourney	GLIDE	Avg.
FreqNet Tan et al. (2024a)	57.2	91.7	52.7	78.5	76.1	62.0	76.5	70.7
NPR Tan et al. (2024b)	54.9	95.5	88.2	93.8	89.8	83.6	93.7	85.6
CNNSpot Wang et al. (2020)	56.2	75.0	74.2	85.9	85.9	65.0	88.4	75.8
Effort Yan et al. (2025b)	93.8	98.1	93.6	69.5	89.8	81.6	77.4	86.3
BFree Guillaro et al. (2025)	90.8	**99.4**	83.6	78.8	89.1	94.5	94.5	89.4
UnivFD Ojha et al. (2023)	97.7	88.5	92.8	63.3	96.7	75.0	93.7	86.8
FakeVLM Wen et al. (2025)	98.4	97.8	98.0	**97.7**	**98.0**	**97.3**	97.5	**97.8**
Veritas++ Tan et al. (2026b)	71.1	92.3	96.4	52.0	74.8	90.9	73.7	78.7
RAFE(Ours)	**99.2**	96.9	**99.0**	94.8	95.3	95.5	**99.4**	97.5

sion, resizing, and transmission artifacts or unknown operations from users. To assess robustness under such conditions, we further evaluate all detectors on three in-the-wild AI-generated image detection datasets: WildRF, SynthWildX, and Chameleon. These datasets consist of both real and AI-generated images collected directly from social media platforms, and therefore better reflect the distribution shifts and degradations encountered in real-world deployments. Consequently, they provide a more reliable assessment of detector performance in practical settings.

The quantitative results of all detectors on these benchmarks are summarized in Table 2. As can be observed, the vast majority of detectors perform poorly across all three datasets, with accuracy dropping to around 50%, which is close to random guessing. Notably, NPR, despite achieving approximately 87% detection accuracy in the previous generalization experiments, suffers a dramatic performance degradation in real-world settings, dropping to 51.7%. In contrast, our method leverages the strong semantic understanding capability of multimodal large models and is able to maintain relatively robust performance under real-world conditions.

These results highlight a substantial gap between the data distributions used for validating current detectors, such as GenImage, and those encountered in real-world scenarios. This gap suggests that existing benchmarks may be insufficient to fully characterize detector robustness in practical deployments, underscoring the need for more realistic evaluation protocols. The overall comparison across all datasets is provided in Appendix D.

5.4 Performance on GenImage

Recent MLLM-based detectors increasingly emphasize training on high-quality images and perception data associated with contemporary generators. However, strong performance on such distributions does not necessarily imply that the model has learned generator-agnostic distinctions between authentic and synthetic objects. As shown in Table 3, Veritas++ Tan et al. (2026b) achieves 96.4% accuracy on DALL-E 2 images, but drops to 71.1% on BigGAN, 52.0% on ADM, and 74.8% on VQDM. This degradation is particularly striking because images from these earlier generators often contain more conspicuous structural defects than modern photorealistic generations. These results indicate that recent MLLM-based methods may still rely substantially on generator-specific low-level statistical traces rather than robustly distinguishing real and synthetic object structures. This observation motivates our deliberate inclusion of earlier generators with explicit, semantically interpretable artifacts in Fake-Reason.

5.5 Qualitative Results

Figure 4 illustrates the detection results of our method on several AI-generated and real images. As shown, our detector is able not only to correctly identify artifact regions in the images (e.g., deformed shapes and distorted text), but also to localize them with high precision. This significantly enhances the credibility and verifiability of our detection results.

Table 4: Ablation study of different backbone models and region-level attention supervision.

Model	**DALLE3**	**Firefly**	**MJv5**	**SDXL**	**Avg**
Qwen3-4B	**100.0**	80.4	85.6	90.2	89.1
Qwen3-8B	99.5	**96.2**	98.1	**99.3**	**98.3**
Qwen3.5-9B	99.1	92.4	**99.0**	98.9	97.4
Qwen3-8B w/o Attn. Sup.	**100.0**	85.5	92.0	93.5	92.8

5.6 Ablation Study

To validate the effectiveness of the proposed method, we further conduct ablation studies on the Synthbuster dataset. We evaluate the effect of backbone architecture and scale using Qwen3-VL-4B-Instruct, Qwen3-VL-8B-Instruct, and Qwen3.5-9B. The detailed results are reported in Table 4. This comparison examines whether the effectiveness of our training framework persists across different model scales and generations. In addition, we investigate the effectiveness of the proposed region-level attention supervision. Removing this objective consistently degrades average performance, with the largest drop occurring on the Firefly subset. This result indicates that explicitly grounding description tokens in their corresponding regions helps the model learn transferable forgery evidence. Additional sensitivity experiments for the attention-margin coefficient α and auxiliary-loss coefficient λ are provided in Appendix B.4.

Attention Visualization. Figure 5 compares the inference-time attention maps of standard SFT and SFT with the proposed attention objective. Without attention supervision, the description tokens distribute substantial attention over unrelated objects and background regions. After adding our objective, the model assigns noticeably more weight to the region corresponding to each description, while reducing diffuse off-region responses. This visualization confirms that the auxiliary loss changes not only the training objective but also the model's visual grounding behavior during inference.



Figure 5: Visualization of description-to-image attention during inference. Green boxes indicate the regions corresponding to the generated descriptions, and brighter colors represent larger attention weights.

6 Conclusion

In this work, we introduce Fake-Reason, a large-scale and high-quality dataset for AI-generated image detection. We highlight that an exclusive focus on modern generators can cause detectors to rely on generator-specific fingerprints while overlooking the conspicuous structural artifacts of earlier generators. Fake-Reason therefore spans multiple generator eras: 80K samples provide grounded forgery explanations, and 10K highly realistic modern-generation samples provide authenticity labels without forcing unreliable region-level rationales.

Built upon an open-source multimodal large language model, we train an AI-generated image detector named RAFE on this dataset and introduce region-level attention supervision to reduce off-region visual shortcut learning. The proposed objective aligns each explanation with its annotated evidence while retaining standard supervision over the complete response. Extensive generalization experiments, together with evaluations in real-world scenarios, demonstrate the superior detection performance of RAFE, as well as its accurate localization and explanation capabilities.

We hope that the proposed dataset and training methodology will facilitate the development of explainable AI-generated image detection methods, thereby effectively countering increasingly realistic image generators.

Ethics Statement

This work aims to advance explainable AI-generated image detection by developing both a large-scale, high-quality dataset (Fake-Reason) and a multimodal detector (RAFE) capable of simultane-

ous detection, localization, and human-verifiable explanation. The primary positive societal impact is to enable more trustworthy detection of manipulated or AI-generated images, which can help mitigate the spread of misinformation and protect visual content integrity. Additionally, by providing precise explanations and localization, our approach facilitates transparency and accountability in AI-driven image analysis.

However, as with any generative-media or detection technology, there are potential risks. The dataset and models could be misused for surveillance, privacy intrusion, or adversarial adaptation by content-generation systems. We therefore encourage responsible use of the released resources and careful consideration of privacy, security, and downstream deployment contexts.

Reproducibility Statement

The key information required for reproduction is provided throughout the paper: dataset construction and verification are detailed in the Dataset section and Appendix A, the complete attention-supervision objective is specified in the Method section and Appendix B, and the backbone, LoRA configuration, optimization hyperparameters, and evaluation settings are reported in the Experiments section. Upon formal publication of this work, we will publicly release the complete training and evaluation code, model configurations, processed annotations, and trained model weights.

AI Use Statement

Generative AI systems are integral to the research methodology: Gemini 3 is used to produce initial structured annotations, which are subsequently verified and filtered as described in the paper. During manuscript preparation, generative AI tools were also used to assist with surveying related work and polishing the language of individual sentences.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, and Keqin Chen. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2023.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024.

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.

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*, pp. 24185–24198, 2024.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4356–4366, 2024a.

Davide Cozzolino, Giovanni Poggi, Matthias Nießner, and Luisa Verdoliva. Zero-shot detection of ai-generated images. In *European Conference on Computer Vision*, pp. 54–72. Springer, 2024b.

Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM multimedia systems conference*, pp. 219–224, 2015.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7890–7899, 2020.

Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. *Advances in neural information processing systems*, 33:3022–3032, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Xinghe Fu, Zhiyuan Yan, Zheng Yang, Taiping Yao, Yandan Zhao, Shouhong Ding, and Xi Li. Pid: Generalized AI-generated images detection with pixelwise decomposition residuals. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=gye2zYytx6.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10696–10706, 2022.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18685–18694, 2025.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15733–15744, 2025.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. *arXiv preprint arXiv:2412.04292*, 2024.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. *arXiv preprint arXiv:2408.06741*, 2024.

Yixuan Li, Xuelin Liu, Xiaoyang Wang, Bu Sung Lee, Shiqi Wang, Anderson Rocha, and Weisi Lin. Fakebench: Probing explainable fake image detection via large multimodal models. *IEEE Transactions on Information Forensics and Security*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European Conference on Computer Vision*, pp. 95–110. Springer, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024a.

Jiawei Liu, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. *arXiv preprint arXiv:2410.10238*, 2024b.

Tai D Nguyen, Aref Azizpour, and Matthew C Stamm. Forensic self-descriptions are all you need for zero-shot detection, open-set source attribution, and clustering of ai-generated images. *arXiv preprint arXiv:2503.21003*, 2025.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pp. 16784–16804. PMLR, 2022.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-banana-400k: A large-scale dataset for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37226–37235, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Thomas Roca, Anthony Cintron Roman, Jehú Torres Vega, Marcelo Duarte, Pengce Wang, Kevin White, Amit Misra, and Juan Lavista Ferres. How good are humans at detecting ai-generated images? learnings from an experiment. *arXiv preprint arXiv:2507.18640*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, et al. Chatgpt: Optimizing language models for dialogue. *OpenAI blog*, 2(4), 2022.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5052–5060, 2024a.

Chuangchuang Tan, 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024b.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025.

Chuangchuang Tan, Xiang Ming, Jinglu Wang, Renshuai Tao, Bin Li, Yunchao Wei, Yao Zhao, and Yan Lu. Semantic visual anomaly detection and reasoning in ai-generated images. In *International Conference on Learning Representations*, 2026a.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Zijian Yu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas++: Value-aware on-policy distillation for perception-enhanced aigi detection, 2026b.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, wenjun wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=xLFYd1owiP.

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. In *The Thirteenth International Conference on Learning Representations*, 2025.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=ODRHZrkOQM.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. In *Forty-second International Conference on Machine Learning*, 2025b.

Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, Zhizheng Wu, Yiping Chen, Dahua Lin, Conghui He, and Weijia Li. LOKI: A comprehensive synthetic data detection benchmark using large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=z8sxoCYgmd.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in neural information processing systems*, 37:71737–71767, 2024.

Zirui Zhang, Yinbo Yu, Donghai Guan, Chunwei Tian, Daoqiang Zhang, and Qi Zhu. A benchmark dataset for mllm-generated image detection: Gpt image2 & nano banana2. *arXiv preprint arXiv:2608.01258*, 2026.

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. In *Proceedings of the IEEE international conference on computer vision*, 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 Detailed Construction of Fake-Reason

This section provides the complete data collection, annotation, and verification procedure for Fake-Reason.

A.1 Motivation and Annotation Quality

Most previous explanation-oriented datasets Wen et al. (2025); Zhou et al. (2025) use multiple open-source MLLMs to generate annotations. However, prior studies Li et al. (2025) show a substantial performance gap between open-source multimodal models Wang et al. (2024); Chen et al. (2024) and commercial models Schulman et al. (2022); Team et al. (2023) for AI-generated image detection. Because open-source models cannot reliably determine image authenticity independently, previous pipelines commonly reveal the ground-truth label and ask the model to explain that predetermined outcome. Such result-conditioned annotation can introduce hallucinated evidence. Other datasets employ human annotators Ye et al. (2025); Li et al. (2025), but the associated labor cost limits their scale.

We instead use Gemini 3.0 to produce initial structured annotations and subsequently apply model-assisted and manual verification. This design combines scalable annotation with explicit quality control.

A.2 Data Collection and Generator Diversity

Existing MLLM-oriented datasets increasingly favor strong contemporary generators. Although their images are highly realistic, they often lack the conspicuous object deformation, broken geometry, or physical inconsistency observed in earlier generations, leaving mainly subtle statistical artifacts. Training exclusively on this distribution may encourage detectors to memorize generator-specific fingerprints rather than learn whether depicted objects and interactions are visually authentic.

We therefore curate fake images from eight representative models spanning GAN and diffusion-based generation. ProGAN Karras et al. (2018) and Stable Diffusion v1.5 Rombach et al. (2022) provide abundant and interpretable structural artifacts. We additionally include BigGAN Brock et al. (2018), VQDM Gu et al. (2022), ADM Dhariwal & Nichol (2021), GLIDE Nichol et al. (2022), DALL-E 2 Betker et al. (2023), and StyleGAN2 Karras et al. (2020). The samples are collected from CNNDetection Wang et al. (2020), GenImage Zhu et al. (2023), or the generators' official repositories. Corresponding real images are drawn from LSUN Yu et al. (2015) and ImageNet-1K Russakovsky et al. (2015).

A.3 Structured Region Annotation

For every image in the grounded subset, the annotator identifies several salient visual regions. Each region is represented by a tuple (l_i, b_i, c_i, d_i), where l_i denotes its semantic location, b_i is an optional bounding box, c_i specifies the evidence category (e.g., a physics/logic violation or local visual artifact), and d_i provides a natural-language description. The response additionally contains a concise image-level summary and final real/fake label.

Real images receive the same form of supervision: the model identifies genuine objects, provides bounding boxes, and briefly describes the visual evidence supporting authenticity. This symmetric design prevents explanations from becoming a fake-only output pattern. The complete annotation prompt is provided in Appendix E.

We manually remove misclassified images. For correctly classified samples, ChatGPT-5.0 cross-validates each image against the explanations produced by Gemini 3.0, after which incorrect and low-quality annotations are removed through manual inspection. This process yields 80K balanced samples with structured region-level annotations.

Dataset	Annotator	MLLM	Grounding	Real Explain	Number
LOKI Ye et al. (2025)	Human	–	✗	✗	3k
FakeBench Li et al. (2025)	Human	–	✗	✗	6k
Holmes-Set Zhou et al. (2025)	MLLM	Open-Source	✗	✓	65k
FakeClue Wen et al. (2025)	MLLM	Open-Source	✗	✓	100k
Fake-Reason	MLLM + Human	Gemini3.0	✓ (80k)	✓ (80k)	90k

Table 5: Comparison of Fake-Reason with other AI-generated image explanation datasets.

A.4 Label-Only Modern-Generator Subset

Recent generators often produce images without localized object corruption or clear semantic contradictions. Forcing region-level explanations for such images would create arbitrary or hallucinated rationales. We therefore collect a complementary 10K samples from Pico-Banana-400K Qian et al. (2026): 5K original real photographs and 5K images edited by Nano-Banana. These samples retain only binary authenticity labels. They contribute to standard SFT but not to region-level explanation or attention supervision. Together with the grounded subset, Fake-Reason contains 90K balanced real and fake samples.

A.5 Comparison with Existing Datasets

Table 5 compares Fake-Reason with existing explanation-oriented datasets. LOKI Ye et al. (2025) and FakeBench Li et al. (2025) provide human annotations but contain considerably fewer samples. AIGI-Holmes Zhou et al. (2025) and FakeClue Wen et al. (2025) scale annotation using open-source MLLMs such as Qwen2-VL Wang et al. (2024) and InternVL2 Chen et al. (2024). In contrast, Fake-Reason combines commercial-model annotation with verification, covers multiple generator eras, and supplies verifiable spatial grounding for 80K samples. Representative annotations are shown in Appendix F.

B Details of Region-Level Attention Supervision

This section provides the complete construction of the region-level attention objective. For every region description, we identify its description-token set $\mathcal{T}_r$ and corresponding bounding box B_r.

B.1 Mapping Visual Tokens to Image Regions

We derive the visual grid from the processed `image_grid_thw`. Dividing its spatial dimensions by the vision encoder's spatial-merging factor gives the $H \times W$ visual tokens passed to the LLM. Using normalized image coordinates, the visual token at row i and column j corresponds to

$$P_{ij} = \left[\frac{j}{W}, \frac{i}{H}, \frac{j+1}{W}, \frac{i+1}{H}\right]. \tag{5}$$

Here, a visual patch refers to the image region represented by one spatially merged LLM visual token.

We use fractional rather than binary box membership. The coverage weight of patch P_{ij} for bounding box B_r is

$$m_{r,ij} = \frac{\mathrm{Area}(B_r \cap P_{ij})}{\mathrm{Area}(P_{ij})}. \tag{6}$$

Consequently, a patch fully inside the box has weight one, a patch outside the box has weight zero, and a partially intersecting patch receives its overlap ratio. For example, if a 32×32 patch overlaps a box by 16×24 pixels, its weight is $16 \times 24/(32 \times 32) = 0.375$.

B.2 Attention Aggregation

Let $\omega_{t,j}^{\ell,h}$ denote the attention from description token t to visual token j at backend LLM layer ℓ and attention head h. Let N_L and N_h denote the numbers of backend LLM layers and attention heads, respectively. Within each layer, we first average over heads and normalize the resulting attention over visual tokens. We then average the normalized maps across all backend LLM layers:

$$A_{t,j} = \frac{1}{N_L} \sum_{\ell=1}^{N_L} \frac{\frac{1}{N_h} \sum_{h=1}^{N_h} \omega_{t,j}^{\ell,h}}{\sum_{k=1}^{HW} \frac{1}{N_h} \sum_{h=1}^{N_h} \omega_{t,k}^{\ell,h}}. \tag{7}$$

Using the soft coverage weights, the fraction of visual attention assigned to region B_r is

$$p_{r,t} = \frac{\sum_{j=1}^{HW} A_{t,j} m_{r,j}}{\sum_{j=1}^{HW} A_{t,j}}. \tag{8}$$

B.3 Area-Adaptive Objective

A fixed target would disproportionately penalize small boxes. Let a_r denote the fraction of the complete image covered by B_r, and let $\alpha \in [0, 1]$ be a tunable attention-margin coefficient. We define

$$\tau_r = a_r + \alpha(1 - a_r), \tag{9}$$

which assigns the region its area-proportional attention plus an α fraction of the remaining attention mass. Only attention below this target is penalized:

$$\ell_{r,t} = \max\left(0, \log \tau_r - \log p_{r,t}\right). \tag{10}$$

Let $\mathcal{G}$ denote the grounded samples in a batch and $N_g = |\mathcal{G}|$. The loss is averaged successively over the tokens of each description, the K_n region descriptions of each grounded image, and the grounded images:

$$\mathcal{L}_{\text{ROI}} = \frac{1}{N_g} \sum_{n \in \mathcal{G}} \frac{1}{K_n} \sum_{r=1}^{K_n} \frac{1}{|\mathcal{T}_{n,r}|} \sum_{t \in \mathcal{T}_{n,r}} \ell_{n,r,t}. \tag{11}$$

The complete response is still optimized with the ordinary SFT loss. Using a tunable coefficient $\lambda \geq 0$ for the auxiliary objective, the final loss is

$$\mathcal{L} = \mathcal{L}_{\text{SFT}} + \lambda \mathcal{L}_{\text{ROI}}. \tag{12}$$

B.4 Sensitivity to Attention-Supervision Strength

We further study the sensitivity of the attention-supervision objective on Synthbuster. We first fix the auxiliary-loss coefficient to $\lambda = 0.05$ and vary the attention-margin coefficient over $\alpha \in \{0.1, 0.2, 0.5\}$. We then fix $\alpha = 0.2$ and evaluate two additional loss weights, $\lambda = 0.01$ and $\lambda = 0.1$. The evaluated configurations are summarized in Table 6. Among these five configurations, $\alpha = 0.2$ and $\lambda = 0.05$ achieve the best average detection performance.

This result indicates that the strength of region-level attention supervision must be balanced. A small α or λ provides insufficient pressure to suppress off-region shortcuts, whereas an excessively large value imposes an overly restrictive attention target or allows the auxiliary objective to dominate standard supervised learning. In particular, increasing α to 0.5 or λ to 0.1 degrades performance, suggesting that forcing too much attention into the annotated box may suppress useful contextual evidence. We therefore use $\alpha = 0.2$ and $\lambda = 0.05$ in all main experiments.

C Construction of the New-Generator Evaluation Set

We construct the New-Generator evaluation set from two complementary sources. First, we randomly select 500 images from the COCO test split Lin et al. (2014) as real samples. We use the corresponding COCO captions as prompts for the open-source FLUX, Infinity, and Stable Diffusion 3 models, thereby obtaining synthetic images whose semantic content is aligned with that of the real set.

Table 6: Sensitivity of region-level attention supervision on Synthbuster. Detection accuracy (%) is reported for each generator and averaged in the last column.

α	λ	DALL-E 3	Firefly	MJv5	SDXL	Avg.
0.1	0.05	99.5	93.7	97.9	**99.4**	97.6
0.2	0.05	99.5	**96.2**	98.1	99.3	**98.3**
0.5	0.05	94.2	84.5	90.4	97.6	91.7
0.2	0.01	**99.9**	90.0	**98.8**	99.3	97.0
0.2	0.10	98.4	87.2	93.7	99.1	94.6

Table 7: Comparison of different detection methods across all datasets.

Method	SynthBuster	NewGen	SynthWildX	WildRF	Chameleon	GenImage	Avg
FreqNet Tan et al. (2024a)	58.9	63.6	46.7	55.8	53.3	70.7	58.2
NPR Tan et al. (2024b)	69.7	87.3	49.4	62.4	51.7	85.6	67.7
CNNSpot Wang et al. (2020)	50.4	53.9	59.0	56.4	60.0	75.8	59.3
UnivFD Ojha et al. (2023)	53.4	60.3	60.0	63.7	60.6	86.8	64.1
Effort Yan et al. (2025b)	74.8	80.4	65.2	68.6	52.6	86.3	71.3
BFree Guillaro et al. (2025)	**99.5**	79.3	**95.7**	92.7	76.3	89.4	88.8
FakeVLM Wen et al. (2025)	44.9	68.2	67.9	63.8	65.6	**97.8**	68.0
Veritas++ Tan et al. (2026b)	88.9	83.6	91.1	96.3	**91.3**	78.7	88.3
RAFE (Ours)	94.6	**97.1**	92.8	**97.2**	90.5	97.5	**95.0**

Second, we draw additional samples from the benchmark proposed for detecting images generated by GPT-Image-2 and Nano Banana-2 Zhang et al. (2026). We include its real-image subset and the texture-image subsets generated by GPT-Image-2 and Nano Banana-2. Together, these sources cover both recent open-source generators and state-of-the-art proprietary image-generation systems.

D Overall Results

For completeness, Table 7 summarizes the average detection accuracy of all compared methods across the generalization and in-the-wild evaluation datasets.

E Prompts Provided to Gemini 3



Prompt Specification: Real vs. AI-Generated Image Classification & Localization

Determine whether this image is a real photograph or fake image generated by ai. If it is ai_generated, please provide the specific reasons, and for each reason, specify the exact location in the form of a bounding box. If it is real, you also need to provide some reasons, and for each reason, specify the exact location in the form of a bounding box. Please respond strictly in the JSON format.

Defect Taxonomy (Reference)

Use these Taxonomy for the "artifacts" section.

1. Local_Artifacts:

- Common Defects: e.g., Wrong_Texture, Deformed_Shape, Asymmetrical_Pupils, Extra_Fingers, Fused_Joints, Wax_Like_Skin, Blending_Hair, Gibberish/Alien_Text, Deformed_Logos, Nonsensical_Signs

2. Physics_Logic:

- Common Defects: e.g., Missing_Tools, Floating_Objects, Impossible_Architecture.

3. Global_Artifacts:

- Common Defects: e.g., Over_smoothing, Hyper_Idealized, Painterly_Background_Textures.

Output Schema (JSON)

You must output ONLY a JSON object based on this structure. Do not include markdown backticks or conversational text.

e.g.,

For ai_generated,

```
{
  "reasons": [
    {
      "location": "Description of location (e.g., 'Man's right hand','Upper left quadrant')",
      "bounding_box": [y_min,x_min, y_max, x_max],
      "category": "Category from Taxonomy",
      "defect_type": "Specific short name (e.g., Missing_Object, Gibberish_Text)",
      "description": "Detailed explanation of what is wrong. Be specific about location and nature of the error.",
      "severity": "Low" | "Medium" | "High",  //strictly follow this format.
    }
  ],
  "label": "Fake"
}
```

for real,

```
{
  "reasons": [
    {
      "location": "Description of location (e.g., 'Man's right hand','Upper left quadrant')",
      "bounding_box": [y_min,x_min, y_max, x_max],
      "category": "Real",
      "description": "Detailed explanation of what is real. Be specific about location.",
    }
  ],
  "rationale_summary": "A brief 1-sentence summary of the main reason for the verdict.",
  "overall_realism_score": Float (0.0 to 1.0, where 1.0 is perfectly real),
  "label": "Real"
}
```

Constraints & Guidelines

1. Be Skeptical: Look for subtle errors (pupil shape, finger joints, background texture) even if the image looks good at first glance.

2. Severity Logic:

- **High**: Artifacts severely disrupt the image's structure or semantics, making it obviously unrealistic at a glance.
- **Medium**: Artifacts are clearly visible and noticeably reduce local realism, and an average viewer can detect the abnormalities.
- **Low**: Artifacts are present but subtle, becoming noticeable only upon close inspection and having limited impact on overall realism.

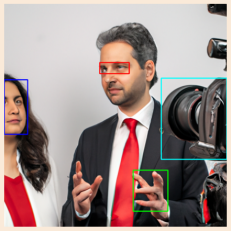
3. JSON Only: Your output must be parsable JSON.

Figure 6: Prompts provided to Gemini 3 for constructing Fake-Reason.

F EXAMPLES FROM FAKE-REASON




```
{
  "reasons": [
    {  "location": 'Top surface of the desk' ,
       "bounding_box": [267, 181, 403, 711],
       "category": 'Local Artifacts',
       "defect_type": 'deformed_shape',
       "description": 'The books on the left side of the desk
       appear '
       'fused and lack distinct edges, while the '
       'decorative object on the right side has a '
       'mangled, unrecognizable shape.',
       "severity": "High"                                    }
    {  "location": 'Upper section of the desk body' ,
       "bounding_box": [390, 222, 491, 918],
       "category": 'Physics_Logic',
       "defect_type": 'impossible_architecture',
       "description": 'The top row of drawers/panels on the
       desk is '
       'structurally nonsensical, with uneven spacing, '
       'mismatched decorative patterns, and knobs that '
       'don't align with functional drawer units.',
       "severity": "High"                                    }
    {  "location": 'Bottom right drawer' ,
       "bounding_box": [642, 696, 718, 882],
       "category": 'Local Artifacts',
       "defect_type": 'deformed_shape',
       "description":  'The two knobs on the large bottom-
       right drawer '
       'are not horizontally aligned, which is a '
       'structural inconsistency for high-end '
       'furniture.',
       "severity": "Medium"                                  }
    {  "location": 'Background picture frame and foreground
rug' ,
       "bounding_box": [0, 252, 189, 586],
       "category": 'Global_Artifacts',
       "defect_type": 'painterly_background_textures',
       "description": 'The contents of the picture frame and
       the '
       'pattern on the rug exhibit a blurry, '
       'impressionistic quality rather than realistic '
       'photographic detail.',
       "severity": "Medium"                                  }],
  "overall_realism_score": 0.3,
  "rationale_summary": 'The image exhibits several classic
AI artifacts, '
  'including fused and deformed objects on the desk, '
  'inconsistent drawer alignment, and painterly textures '
  'in the background and foreground.'
  "label": "Fake"
}
```


```
{
  "reasons": [
    {  "location": 'Man's eyes' ,
       "bounding_box": [263, 429, 313, 561],
       "category": 'Local Artifacts',
       "defect_type": 'Asymmetrical_Pupils',
       "description": 'The man's eyes are severely malformed;
       the '
       'left eye (viewer's right) has a distorted iris '
       'shape and the right eye (viewer's left) '
       'appears misplaced and lacks proper anatomical '
       'structure.',
       "severity": "High"                                    }
    {  "location": 'Man's left hand' ,
       "bounding_box": [746, 579, 930, 735],
       "category": 'Local Artifacts',
       "defect_type": 'Fused_Joints',
       "description": 'The man's left hand (viewer's right) is a '
       'jumbled mass of distorted fingers and fused '
       'skin, lacking any realistic skeletal '
       'structure.',
       "severity": "High"                                    }
    {  "location": 'Woman's face' ,
       "bounding_box": [340, 0, 585, 105],
       "category": 'Local Artifacts',
       "defect_type": 'deformed_shape',
       "description": 'The woman's face is heavily distorted, '
       'particularly around the eyes and nose, with a '
       'lack of symmetry and realistic detail.',
       "severity": "High"                                    }
    {  "location": 'Camera equipment on the right' ,
       "bounding_box": [333, 705, 697, 1000],
       "category": 'Local Artifacts',
       "defect_type": 'Deformed_Shape'
       "description":  'The camera lens and body exhibit
       nonsensical, '
       'warped geometry and textures that do not '
       'correspond to any real-world broadcast '
       'equipment.',
       "severity": "Medium"                                  }],
  "overall_realism_score": 0.1,
  "rationale_summary":  'The image exhibits significant
anatomical deformities '
  'in the hands and eyes, along with nonsensical camera '
  'geometry and irrational graphical artifacts on the '
  'clothing.'
  "label": "Fake"
}
```


Figure 7: Examples from Fake-Reason (part 1).




```
{
  "reasons": [
    {
      "location": 'Award ribbons in the foreground' ,
      "bounding_box": [538, 234, 989, 715],
      "category": 'Real',
      "description": 'The text on the award ribbons is perfectly '
      'legible, spelling out 'Delaware County Kennel ' 'Club, Inc.' and 'Agility Trial', which is ' 'highly specific and consistent with real-world ' 'objects.',
    }
    {
      "location": 'The dog's coat and head' ,
      "bounding_box": [15, 6, 732, 798],
      "category": 'Real',
      "description": "The red digital date stamp '2007 4 7' is a "
      'characteristic feature of consumer digital ' 'cameras from that era and appears integrated ' "correctly with the image's overall resolution " 'and noise profile.',
    }
  ],
  "overall_realism_score": 0.98,
  "rationale_summary": 'The image exhibits high-fidelity details, legible and '
  'contextually accurate text on the awards, and ' 'naturalistic textures consistent with an authentic ' 'photograph from a digital camera.'
  "label": "Real"
}
```





```
{
  "reasons": [
    { "location": 'Upper side of the bus' ,
      "bounding_box": [426, 621, 543, 940],
      "category": 'Local Artifacts',
      "defect_type": 'Gibberish_Text',
      "description": "The text on the side of the bus 'TROUUS' "
      'followed by a distorted character is ' 'nonsensical and typical of AI text generation ' 'failures.',
      "severity": "High" }
    { "location": 'Front wheel area' ,
      "bounding_box": [685, 283, 859, 431],
      "category": 'Local Artifacts',
      "defect_type": 'Deformed_Shape',
      "description": 'The front wheel hubcap is asymmetrical and '
      'lacks the consistent mechanical geometry of a ' 'real wheel.',
      "severity": "Medium" }
    { "location": 'Side and roof section of the bus' ,
      "bounding_box": [253, 388, 558, 981],
      "category": 'Physics_Logic',
      "defect_type": 'Impossible_Architecture',
      "description": 'The structural lines of the bus, particularly' 'the white diagonal divider and the window ' 'frames, blend into each other or end abruptly ' "in a way that doesn't follow physical " 'construction logic.',
      "severity": "Medium" }
    { "location": 'Main deck windows' ,
      "bounding_box": [486, 264, 637, 659],
      "category": 'Local Artifacts',
      "defect_type": 'Nonsensical_Content'
      "description": 'The silhouettes visible through the passenger '
      'windows are amorphous and do not clearly ' 'represent people, seats, or any recognizable ' 'interior bus components.',
      "severity": "Medium" }],
  "overall_realism_score": 0.25,
  "rationale_summary": 'The image exhibits several classic AI artifacts, '
'including nonsensical text, distorted mechanical ' 'components like the wheels, and inconsistent structural ' 'lines on the vehicle.'
  "label": "Fake"
}
```


Figure 8: Examples from Fake-Reason (part 2).