arXiv:2509.19841v1 [cs.CV] 24 Sep 2025

ThinkFake: Reasoning in Multimodal Large Language Models for AI-Generated Image Detection

Tai-Ming Huang[1,2], **Wei-Tung Lin**[2,5], **Kai-Lung Hua**[4,5],
Wen-Huang Cheng[1], **Junichi Yamagishi**[3], **Jun-Cheng Chen**[2]
National Taiwan University[1], Academia Sinica[2], National Institute of Informatics[3],
Microsoft[4] National Taiwan University of Science and Technology[5]

Abstract

The increasing realism of AI-generated images has raised serious concerns about misinformation and privacy violations, highlighting the urgent need for accurate and interpretable detection methods. While existing approaches have made progress, most rely on binary classification without explanations or depend heavily on supervised fine-tuning, resulting in limited generalization. In this paper, we propose ThinkFake, a novel reasoning-based and generalizable framework for AI-generated image detection. Our method leverages a Multimodal Large Language Model (MLLM) equipped with a forgery reasoning prompt and is trained using Group Relative Policy Optimization (GRPO) reinforcement learning with carefully designed reward functions. This design enables the model to perform step-by-step reasoning and produce interpretable, structured outputs. We further introduce a structured detection pipeline to enhance reasoning quality and adaptability. Extensive experiments show that ThinkFake outperforms state-of-the-art methods on the GenImage benchmark and demonstrates strong zero-shot generalization on the challenging LOKI benchmark. These results validate our framework's effectiveness and robustness. Code will be released upon acceptance.

1 Introduction

The rapid development of generative models (Ramesh et al. 2021; Dhariwal and Nichol 2021; Li, Hou, and Loy 2024) has led to AI-generated images that are increasingly realistic and difficult to distinguish from real ones. While these advancements benefit creative industries, they also raise serious concerns regarding misuse, including misinformation, fraud, and privacy violations. As techniques such as GANs (Goodfellow et al. 2014), diffusion models (Esser et al. 2024), and flow-based approaches (Rombach et al. 2022; Labs 2024) continue to evolve, the challenge of detecting synthetic content grows more severe. This trend highlights the urgent need and importance of detection methods that are both accurate and interpretable.

Recent studies (Wang et al. 2020; Yan et al. 2025; Tan et al. 2024; Ojha, Li, and Lee 2023; Chang et al. 2023) have shown significant progress in detecting AI-generated images. However, most existing detection models are limited to binary classification and lack meaningful explanations, which makes their outputs difficult to interpret and unreliable in real-world scenarios. To improve interpretability, some recent approaches (Gao et al. 2025; Zhou et al. 2025; Ji et al. 2025; Kang et al. 2025; Wen et al. 2025) leverage Multimodal Large Language Models (MLLMs) due to their strong capabilities in commonsense reasoning and natural language generation. Despite their promise, these methods often rely on large-scale annotated datasets for supervised fine-tuning. This heavy dependence on labeled data leads to memorization rather than genuine reasoning, ultimately resulting in poor generalization. While these advancements are promising, two key challenges remain: **1) lack of interpretability:** most models provide only binary outputs without explaining their decisions, which limits their practical applicability; and **2) lack of reasoning and generalization:** methods that incorporate MLLMs for explanation typically rely on extensive supervision and lack autonomous reasoning, making it difficult to generalize to complex, unseen scenarios. Therefore, developing AI-generated image detectors that are both capable of reasoning and generalization has become increasingly urgent. "Reasoning" in MLLMs represents a major step forward in handling complex analytical tasks. Unlike traditional classifiers that rely on pattern matching or likelihood scores, reasoning-capable MLLMs can produce step-by-step explanations that resemble human cognition. The principle of "SFT memorizes, RL generalizes" has been applied in recent work (Chu et al. 2025), raising an important question: can we leverage this reasoning ability not only to detect AI-generated images, but also to explain decisions through logic, comparison, and context?

To address the above challenges, we propose a reasoning-based and generalizable framework for AI-generated image detection, inspired by recent progress in reasoning tasks (Liu et al. 2025; Huang et al. 2025a) and reinforcement learning (Shao et al. 2024). Our method builds on a reasoning MLLM, guided by carefully designed prompts and reward functions, to enhance both performance and generalization. We further introduce a structured detection pipeline combined with Group Relative Policy Optimization (GRPO) reinforcement learning to strengthen the model's reasoning and adaptability. Through extensive experiments and ablation studies, our approach outperforms existing state-of-the-art methods on the GenImage (Zhu et al. 2024) bench-

---

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

mark. Moreover, on the more challenging and comprehensive LOKI (Ye et al. 2024) benchmark, our method demonstrates strong zero-shot generalization, validating the effectiveness of our framework. Our main contributions are summarized as follows:

- We develop a novel reasoning MLLM framework that not only achieves strong detection performance but also provides interpretable reasoning for AI-generated image detection.
- We design specialized prompts and reward functions that significantly enhance the model's detection capability, performance, and explainability.
- Our method outperforms state-of-the-art approaches on the GenImage benchmark. Extensive experiments further demonstrate the framework's generalization ability and robustness under real-world challenges.

2 Related Work

2.1 AI-Generated Image Detection

The rapid advancement of generative AI models has intensified the challenge of detecting AI-generated images, as recent models including MidJourney (mid 2025), DALLE-3 (Ramesh et al. 2021), Stable Diffusion (Rombach et al. 2021), and Flux (Labs 2024) produce increasingly realistic images. FreDect (Frank et al. 2020) identifies anomalies by analyzing frequency domain characteristics of GAN-generated images. UnivFD (Ojha, Li, and Lee 2023) leverages pre-trained CLIP-ViT features with nearest neighbor and linear probing for cross-domain generalization. DIRE (Wang et al. 2023) exploits reconstruction errors to distinguish diffusion-generated images from real ones, though it suffers from poor generalization and high computational overhead. PatchCraft (Zhong et al. 2023) segments images into patches and applies SRM filters to analyze pixel correlations for detection. AEROBLADE (Ricker, Lukovnikov, and Fischer 2024) introduces a training-free approach using autoencoder reconstruction errors for latent diffusion model detection. AIDE (Yan et al. 2025) employs a dual-stream architecture combining frequency and semantic features. However, these approaches often fail to generalize to out-of-distribution data due to limited training data and the lack of prior knowledge from Multimodal Large Language Models (MLLMs) (Bai et al. 2025; Liu et al. 2023). In addition, their binary predictions lack explanations, limiting their practical applicability.

2.2 Explainability

Explainability in MLLMs is crucial for mitigating risks associated with AI misuse, and recent research has increasingly investigated this domain. Some studies (Huang et al. 2025b; Jia et al. 2024) directly prompt MLLMs with queries such as "explain what the artifacts are." However, research reveals that directly generating textual explanations frequently results in hallucinations or overthinking, leading to inaccurate outcomes or refusal to respond (Jia et al. 2024). Furthermore, MLLMs often fail to comprehensively perceive all relevant features, which constrains their explainability effectiveness. To overcome these limitations, researchers have employed fine-tuning approaches such as LoRA (Hu et al. 2022) or Direct Preference Optimization (DPO) (Rafailov et al. 2023) for MLLMs (Huang et al. 2024; Wen et al. 2025; Gao et al. 2025; Zhou et al. 2025). However, these methods tend to memorize training patterns rather than generalize to real-world scenarios. To address this limitation, we employ GRPO (Guo et al. 2025), which enhances generalization by letting the model learns how to "think", for improved MLLM detection and explainability.

3 Method

3.1 Overview

In this paper, we aim to train a model that effectively distinguishes AI-generated images from real images captured by cameras. Leveraging recent advancements in multimodal large language models (MLLMs) (Chen et al. 2024; Bai et al. 2025), we utilize their capabilities in image understanding and text generation to analyze image authenticity and provide reliable explanations. We propose ThinkFake, a carefully designed framework that employs a reasoning MLLM $\mathcal{F}_\theta$ with a reinforcement learning process to classify and reason about whether an image is real or AI-generated. Given an input image I and a corresponding prompt P, with a simple question *"is this image real or generated by AI?"*, the reasoning MLLM $\mathcal{F}_\theta$ produces both verdict V and reasoning explanation R. This process can be represented by the following equations:

$$\mathrm{V}, \mathcal{R} = f_{post}(\mathcal{F}_\theta(\mathrm{I}, \mathrm{P})), \tag{1}$$

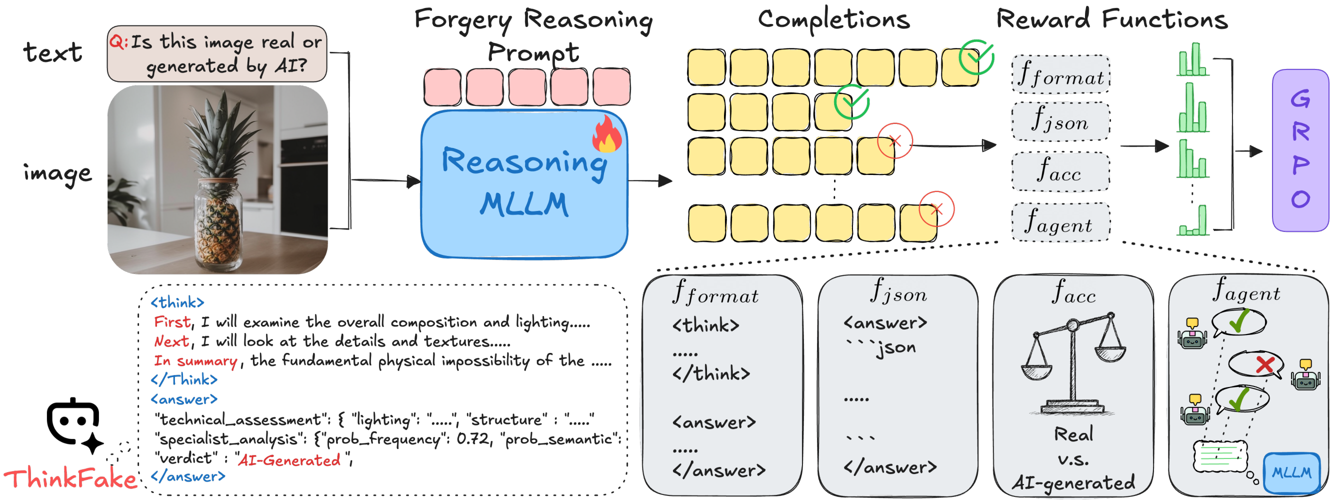
where $f_{\text{post}}(\cdot)$ denotes the post-processing function.

Specifically, we developed a comprehensive detection pipeline and training process for the MLLM. First, we designed a task-specific *Forgery Reasoning Prompt* tailored for the MLLM to address this task. Inspired by (Guo et al. 2025), we then employed reinforcement learning with four carefully designed reward functions to enhance the MLLM's reasoning capabilities for this task, enabling it to perform both detection and explanation, as illustrated in Fig. 1.

3.2 Pipeline and Forgery Reasoning Prompt

Pipeline: Inspired by (Gao et al. 2025), detecting AI-generated content requires a multifaceted approach that examines various visual attributes, such as lighting, shadows, textures, and edges. By analyzing these features, it is possible to identify recurring patterns or artifacts commonly associated with AI-generated content, thereby facilitating robust detection. Drawing on recent advancements in multimodal chain-of-thought (CoT) reasoning on Multimodal Large Language Models (MLLMs), we propose a novel chain-of-thought pipeline for AI-generated image detection. This pipeline incorporates hierarchical reasoning steps to enhance the visual reasoning capabilities of MLLMs. First, we decompose the detection pipeline into five steps, as shown in the upper part of Fig. 2:

- **Initial Quick Scan and Intuitive Judgment:** In the first step, an initial impression of the input image is formed, evaluating its overall harmony, content, and composition.



text
Q: Is this image real or generated by AI?
image
Forgery Reasoning Prompt
Reasoning MLLM
Completions
Reward Functions
f_{format}
f_{json}
f_{acc}
f_{agent}
GRPO
ThinkFake
f_{format}
<think>
.....
</think>
<answer>
.....
</answer>
f_{json}
<answer>
```json
.....
```
</answer>
f_{acc}
Real
v.s.
AI-generated
f_{agent}
MLLM

```
<think>
First, I will examine the overall composition and lighting.....
Next, I will look at the details and textures.....
In summary, the fundamental physical impossibility of the .....
</Think>
<answer>
"technical_assessment": { "lighting": ".....", "structure" : "....."
"specialist_analysis": {"prob_frequency": 0.72, "prob_semantic":
"verdict" : "AI-Generated ",
</answer>
```

Figure 1: **Overview of our ThinkFake:** We propose a reasoning MLLM framework capable of deliberate reasoning for detecting AI-generated images. Guided by a forgery reasoning prompt, and trained with carefully designed reward functions and GRPO-based reinforcement learning, the model demonstrates strong generalization and self-reflective explanatory capabilities.

- **Detailed In-Depth Observation and Analysis:** Next, a meticulous examination is conducted, focusing on texture quality, lighting, shadows, reflections, edges, and blending, with each aspect analyzed in detail.
- **Technical Analysis and Source Tracing:** During the technical analysis and source tracing phase, specific attention is given to artifacts commonly found in AI-generated content, such as inconsistencies in structure, surfaces, edges, and contextual elements.
- **Auxiliary Tool Inspection:** Several specialized AI-generated image detection tools, each focusing on different aspects, are available as reference or auxiliary tools to support comprehensive detection and evaluation.
- **Comprehensive Judgment and Conclusion:** In the final stage of the pipeline, evidence and analysis results from all previous steps are synthesized to deliver a final judgment and reasoning, completing the comprehensive detection process.

Forgery Reasoning Prompt: Inspired by recent advancements in structured output for MLLMs (Tam et al. 2024; He et al. 2024), we unify the analytical process described above and provide clear guidance for evaluating each attribute. As illustrated in the lower part of Fig. 2, we introduce the *Forgery Reasoning Prompt* to enhance the reasoning capability of MLLMs, specifically for the task of detecting AI-generated images. Each analytical step is incorporated into the prompt using special tokens, <think> and <answer>, together with a strict JSON format to constrain the model's output. This structured format improves both interpretability and detection performance. In the "Auxiliary Tool Inspection" stage, we define three expert perspectives: a semantic expert that focuses on high-level features, a frequency expert that targets low-level features, and a dual-stream expert that considers both. This design allows the *Forgery Reasoning Prompt* to support comprehensive reasoning and guide the model through a complete detection pipeline, ultimately producing interpretable and structured outputs.

3.3 Reward Functions

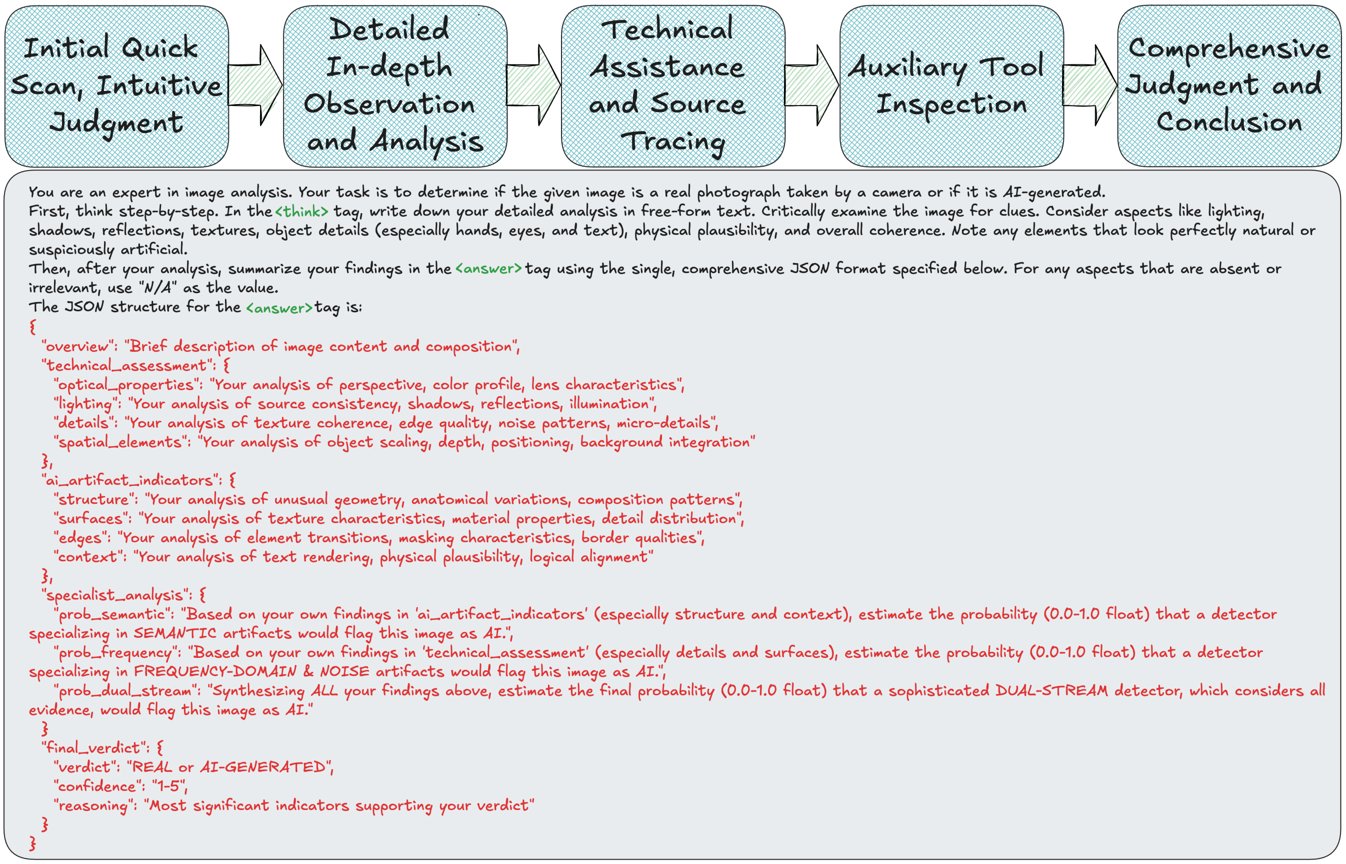
Reward functions guide model optimization in reinforcement learning. We design four binary rewards, each set to 1 if satisfied and 0 otherwise.

Reasoning Format Reward (f_{format})**.** The model's output must adhere to the <think>...</think> and <answer>...</answer> format. This reward constrains the model to output its reasoning process within the <think> tags and provide the final answer within the <answer> tags.

JSON Format Reward (f_{json})**.** To support Forgery Reasoning Prompt, we impose stricter constraints on the model's response format, requiring answers to adhere to a structured JSON format. This constraint improves detection performance, enhances the efficiency of answer extraction, and provides interpretable outputs.

Accuracy Reward (f_{acc})**.** To evaluate the model's performance, we implemented a reward function that measures binary accuracy by comparing the predicted answer to the ground-truth label. Specifically, the model receives a reward of 1 when its prediction matches the ground-truth label, and 0 otherwise. This straightforward yet effective scheme offers a clear and reliable signal for guiding the model toward accurate classification.

Agentic Reward ($f_{agentic}$)**.** The agentic reward corresponds to the "Auxiliary Tool Inspection" step in our pipeline. We incorporated state-of-the-art models as expert agents to enhance the inspection process from multiple perspectives: UnivFD (Ojha, Li, and Lee 2023) for high-level semantic artifact analysis, NPR (Tan et al. 2024) for low-level frequency-based artifact features, and AIDE (Yan et al. 2025) as a dual-stream agent addressing both levels. The


Initial Quick Scan, Intuitive Judgment
Detailed In-depth Observation and Analysis
Technical Assistance and Source Tracing
Auxiliary Tool Inspection
Comprehensive Judgment and Conclusion

```
You are an expert in image analysis. Your task is to determine if the given image is a real photograph taken by a camera or if it is AI-generated.
First, think step-by-step. In the <think> tag, write down your detailed analysis in free-form text. Critically examine the image for clues. Consider aspects like lighting,
shadows, reflections, textures, object details (especially hands, eyes, and text), physical plausibility, and overall coherence. Note any elements that look perfectly natural or
suspiciously artificial.
Then, after your analysis, summarize your findings in the <answer> tag using the single, comprehensive JSON format specified below. For any aspects that are absent or
irrelevant, use "N/A" as the value.
The JSON structure for the <answer> tag is:
{
  "overview": "Brief description of image content and composition",
  "technical_assessment": {
    "optical_properties": "Your analysis of perspective, color profile, lens characteristics",
    "lighting": "Your analysis of source consistency, shadows, reflections, illumination",
    "details": "Your analysis of texture coherence, edge quality, noise patterns, micro-details",
    "spatial_elements": "Your analysis of object scaling, depth, positioning, background integration"
  },
  "ai_artifact_indicators": {
    "structure": "Your analysis of unusual geometry, anatomical variations, composition patterns",
    "surfaces": "Your analysis of texture characteristics, material properties, detail distribution",
    "edges": "Your analysis of element transitions, masking characteristics, border qualities",
    "context": "Your analysis of text rendering, physical plausibility, logical alignment"
  },
  "specialist_analysis": {
    "prob_semantic": "Based on your own findings in 'ai_artifact_indicators' (especially structure and context), estimate the probability (0.0-1.0 float) that a detector
specializing in SEMANTIC artifacts would flag this image as AI.",
    "prob_frequency": "Based on your own findings in 'technical_assessment' (especially details and surfaces), estimate the probability (0.0-1.0 float) that a detector
specializing in FREQUENCY-DOMAIN & NOISE artifacts would flag this image as AI.",
    "prob_dual_stream": "Synthesizing ALL your findings above, estimate the final probability (0.0-1.0 float) that a sophisticated DUAL-STREAM detector, which considers all
evidence, would flag this image as AI."
  }
  "final_verdict": {
    "verdict": "REAL or AI-GENERATED",
    "confidence": "1-5",
    "reasoning": "Most significant indicators supporting your verdict"
  }
}
```

Figure 2: Top: Overview of ThinkFake detection pipeline. Bottom: Detail of our proposed *Forgery Reasoning Prompt*.

outputs of these agents serve as ground-truth labels for their respective inspection tasks. Rewards are computed via cross-entropy between these labels and the model's predictions. Leveraging supervision from expert agents enables the model to achieve more robust and comprehensive artifact detection. *Further details are provided in the Appendix. A.*

3.4 Training Process

Data Preparation. We follow the GenImage protocol (Zhu et al. 2024) for dataset construction. Real images are sourced from ImageNet, while fake images are generated using SD v1.4 based on 1k ImageNet labels. To ensure balance, we uniformly sample each class to create a sub-dataset with equal numbers of real and fake images. This sub-dataset is split into two parts: one for cold-start supervised fine-tuning (SFT-set) and one for reinforcement learning (RL-set). **1) SFT-set:** Since the original data contains only binary labels, we design a data annotation pipeline using a commercial MLLM (e.g., Gemini-1.4-pro) to generate reasoning paths and attribute-level answers. The ground-truth label is included in the prompt to reduce hallucinations. We further apply rule-based filtering to remove low-quality responses, resulting in 638 high-quality samples for SFT training. **2) RL-set:** RL training uses only binary labels, so no additional annotation is needed. We use 5,000 samples (2,500 real and 2,500 fake) as the RL training set.

GRPO. We fine-tune the pre-trained Qwen2.5-VL-7B-Instruct model using the proposed reward signals to guide policy learning. For optimization, we adopt the Group Relative Policy Optimization (GRPO) algorithm (Shao et al. 2024; Guo et al. 2025), which eliminates the need for a separate critic network by leveraging group-level reward statistics to estimate baselines, thereby reducing computational overhead.

Given a question q, GRPO draws a set of G responses $o_1, o_2, \ldots, o_G$ from the old policy $\pi_{\theta_{\text{old}}}$ and updates the current policy π_θ by maximizing the following objective:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\Bigg(\min\Bigg(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}A_i,\\ \text{clip}\Bigg(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\epsilon, 1+\epsilon\Bigg)A_i\Bigg)\\ -\beta\,\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Bigg)\Bigg],\end{aligned} \tag{2}$$

where the hyperparameters ϵ and β regulate the clipping threshold and the KL regularization strength, respectively. The KL-divergence term is computed as:

$$\mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{\text{ref}}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \tag{3}$$

Model	**Params.**	Midjourney	SD v1.4	SD v1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	***Mean***
Closed-Source MLLMs										
ChatGPT-4o (2024)	–	56.8	75.3	75.7	65.8	79.3	84.4	**79.7**	**84.6**	75.2
ChatGPT-o3 (2025)	–	62.0	81.6	81.2	68.8	80.5	91.0	79.0	81.4	78.2
Gemini-2.0-flash (Comanici et al. 2025)	–	55.5	80.1	77.8	65.4	66.5	88.2	78.8	79.4	73.9
Open-Source General MLLMs										
Qwen2.5-VL-7B-Instruct (Bai et al. 2025)	7B	54.6	51.9	50.6	52.3	58.0	56.3	54.7	54.2	54.1
Qwen2.5-VL-72B-Instruct (Bai et al. 2025)	72B	54.5	56.0	54.9	55.5	62.3	65.1	61.1	61.6	58.9
InternVL2.5-78B (Chen et al. 2024)	78B	52.7	53.4	53.0	55.0	63.0	59.2	61.7	72.3	58.8
LLaVA-NeXT (Li et al. 2024)	72B	52.0	51.5	49.4	51.5	50.8	55.3	56.8	59.3	53.3
Open-Source Reasoning MLLMs										
LLaVA-CoT-7B (Xu et al. 2024)	7B	51.5	50.8	51.5	51.3	54.8	55.3	53.9	54.8	53.0
Mulberry-7B (Yao et al. 2024)	7B	50.4	50.2	50.5	51.2	63.4	54.5	53.4	56.0	53.7
Our Model										
ThinkFake (Ours)	7B	**92.5**	**93.1**	**95.3**	**73.1**	**87.4**	**93.6**	66.2	70.8	**84.0**

Table 1: Performance comparison in Acc. (%) on selected generative models. For ThinkFake, except for SD v1.4 (used in training), all others are unseen test sets. The best results are highlighted in bold.

and the normalized advantage A_i is derived from the group-wise reward set $r_1, r_2, \ldots, r_G$ as:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\}))}{\text{std}(\{r_1, r_2, \ldots, r_G\}))}. \tag{4}$$

Training ThinkFake. Overall, our ThinkFake approach begins with supervised fine-tuning (SFT) using the SFT-set for cold-start initialization. The high-quality SFT-set data enables the reasoning MLLM to develop initial capabilities for AI-generated image detection while aligning with the desired output format. Subsequently, we employ reinforcement learning with the RL-set using GRPO to enhance the MLLM's generalization ability. At this stage, we utilize four carefully designed reward functions, as described in Sec. 3.3, to guide the model. Two format reward functions constrain the model's reasoning and output structure. The "Accuracy Reward" improves yes/no discrimination, while the "Agentic Reward" enables multi-perspective evaluation for better performance and interpretability. *Full data pipeline and training details are in Appendix B.*

4 Experiments

4.1 Experimental Settings

Implementation Details. We initialize the reasoning MLLM $\mathcal{F}_\theta$ using Qwen2.5-VL-7B-Instruct (Bai et al. 2025). The cold-start phase comprises 58 iterations with a batch size of 40 and a learning rate of $5e{-}6$, implemented via LLaMA-Factory [1]. This is followed by GRPO training for 1,250 iterations with a batch size of 256, a learning rate of $1e{-}6$, and 8 generations, conducted using the VLM-R1 [2] framework. To implement the expert agents—UnivFD (Ojha, Li, and Lee 2023), NPR (Tan et al. 2024), and AIDE (Yan et al. 2025)—we adopt their official open-source code and pre-trained weights. All experiments are conducted on 8 NVIDIA H200 GPUs.

Datasets. As noted in Sec. 3.4, we adhere to the GenImage protocol (Zhu et al. 2024) but employ our constructed SFT-set and RL-set sub-training sets for cold-start SFT and RL training, respectively. Testing follows the GenImage benchmark, with training performed on SD v1.4 and generalization evaluated on unseen models, including SD v1.5, Midjourney, ADM, GLIDE, Wukong, VQDM, and BigGAN.

Evaluation Metrics. Following prior work (Wang et al. 2020; Yan et al. 2025), we report classification accuracy (Acc.) averaged over both real and AI-generated images.

4.2 Can MLLMs Handle AI-generated Image Detection?

We first investigate whether existing state-of-the-art multimodal large language models (MLLMs) possess AI-generated image detection capabilities. We evaluated open-source general MLLMs, including the Qwen2.5-VL series, InternVL-2.5, and LLaVA-Next, as well as open-source reasoning MLLMs such as LLaVA-CoT and Mulberry. Additionally, we assessed closed-source commercial models like ChatGPT-4o, Gemini-2.0-Flash, and the widely recognized reasoning model ChatGPT-o3. The results, presented in Tab. 1, reveal that these large-scale MLLMs, including reasoning models, fail to distinguish real images from AI-generated ones, underscoring the importance and urgency of this task. In contrast, our ThinkFake model demonstrates superior detection capabilities, performing competitively despite having significantly fewer parameters. These results highlight ThinkFake's strong reasoning capabilities, enabling it to generalize to unseen categories and effectively handle complex detection and explanation tasks.

4.3 Comparison with Other SOTA Methods

To assess detection performance, we compare our proposed method with SOTA method (Wang et al. 2020; Yan et al. 2025; Tan et al. 2024; Ojha, Li, and Lee 2023; Chang et al. 2023). For a fair comparison, all baseline detectors are re-trained using our experimental configuration, which adheres to the GenImage protocol and trains solely on the

---

[1] https://github.com/hiyouga/LLaMA-Factory

[2] https://github.com/om-ai-lab/VLM-R1

	Overall	Scene	Animal	Person	Object	Medicine	Doc	Satellite
Human	27.3	24.0	25.8	19.9	26.9	26.1	22.1	39.3
AIDE (Yan et al. 2025)	63.1	-	**89.9**	62.5	**96.5**	53.4	49.7	39.3
VILA1.5-40B	48.8	53.7	39.3	50.0	33.4	52.5	59.9	50.6
InternVL2-40B	49.6	55.7	37.3	59.2	34.8	55.5	64.8	40.8
Qwen2-VL-72B	53.2	55.9	43.4	66.9	38.0	55.9	73.7	38.2
LLaVA-OneVision-72b	46.3	54.7	31.6	53.1	27.8	52.1	**67.9**	36.6
Claude-3.5-Sonnet	53.6	51.6	51.6	55.2	51.4	51.9	59.1	50.9
Gemini-1.5-Pro	43.5	53.7	35.7	51.5	30.3	50.0	47.2	38.1
GPT-4o	63.4	70.1	69.7	**84.4**	70.3	54.3	60.1	45.0
ThinkFake (Ours)	**75.4**	**78.5**	87.2	71.7	89.4	**64.5**	54.2	**82.2**

Table 2: Quantitative results in Acc. (%) on the LOKI (Ye et al. 2024) benchmark. All categories are unseen zero-shot test cases. The best results are highlighted in bold.

Method	**Reasoning**	**Mean Acc.**
CNNSpot		62.6
AntifakePrompt		76.9
UnivFD		72.3
NPR		59.6
AIDE		81.7
ThinkFake (Ours)	✓	**84.0**

Table 3: **Comparison of State-of-the-Art Methods:** All baselines are re-trained on our training set to ensure a fair comparison, following the GenImage evaluation protocol.

sub-training set. Evaluation is conducted on unseen generators, as described in Sec.4.1. In this experiment, we report the mean image-level accuracy (Mean Acc.) across all test sets, as shown in Tab. 3. Our ThinkFake model achieves strong detection performance. Unlike the other SOTA methods, which are limited to binary classification, our method also generates interpretable reasoning, further demonstrating the effectiveness of our reinforcement learning framework in guiding the model's decision-making process for AI-generated image detection. *A comprehensive evaluation will be presented in the Appendix. C.*

4.4 Generalization in Real-World Scenarios

In Tab. 2, we assess the generalization capability of our ThinkFake model against the LOKI benchmark (Ye et al. 2024). LOKI benchmark is a novel, comprehensive evaluation for MLMMs to detect synthetic data across multimodal, with 18,000 questions across 26 subcategories. We selected the "Image Judgment Task", where MLLMs identify synthetic or real images in categories like Scene, Animal, Person, Object, Medicine, Doc, and Satellite. Following the LOKI protocol, we zero-shot tested our ThinkFake model. As shown in Tab. 2, ThinkFake exhibited strong generalization, performing well across unseen categories, with slight performance drops in Doc and Medicine but still competitive results. Notably, ThinkFake exhibits strong and comprehensive robustness under unseen scenarios. Even for rare or uncommon categories, the model maintains high performance, suggesting that its reasoning and explanatory capabilities can effectively generalize to real-world challenges. These results underscore the effectiveness of our approach. *Additional baseline methods and visual examples will be introduced in the Appendix. D.*

Method	**Cold Start**	**GRPO**	**Mean Acc.**
Baseline			54.1
ThinkFake-Zero		✓	72.4
ThinkFake-SFT	✓		76.7
ThinkFake-R1 (Ours)	✓	✓	**84.0**

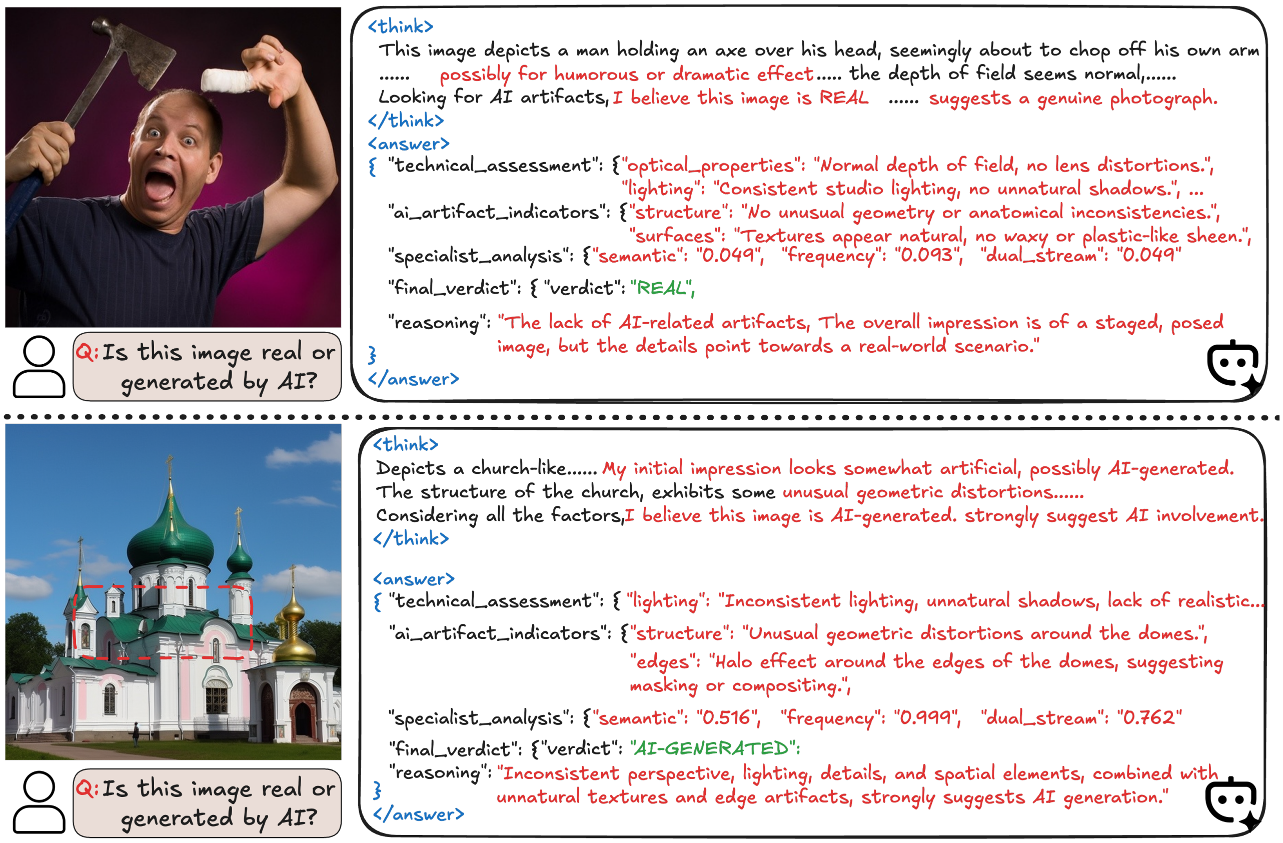
Table 4: **Comparison of Training Strategies:** We compare ThinkFake with various training methods under the GenImage evaluation protocol. The baseline refers to our initial reasoning model, Qwen2.5-VL-7B-Instruct.

f_{format}	f_{json}	f_{acc}	$f_{agentic}$	**Mean Acc.**
		✓		✗
✓		✓		74.6
✓	✓	✓		82.6
✓	✓	✓	✓	**84.0**

Table 5: **Comparison of Reward Functions:** We evaluate the impact of different reward functions on the performance of ThinkFake under the GenImage evaluation protocol. A ✗ symbol indicates training failure.

4.5 Ablation Study

Training Strategy Inspired by DeepSeek-R1-Zero (Guo et al. 2025), we explored training strategies that guide models toward self-reflected reasoning for AI-generated image detection. Specifically, we followed the DeepSeek-R1-Zero pipeline to train a reasoning MLLM using GRPO with our RL-set, as described in Sec. 3.4. This variant, named "ThinkFake-Zero", was trained using only the "Reasoning Format Reward" and "Accuracy Reward" (Sec. 3.3) to



Q: Is this image real or generated by AI?

```
<think>
 This image depicts a man holding an axe over his head, seemingly about to chop off his own arm
 ......    possibly for humorous or dramatic effect..... the depth of field seems normal,......
 Looking for AI artifacts, I believe this image is REAL   ...... suggests a genuine photograph.
</think>
<answer>
{  "technical_assessment": {"optical_properties": "Normal depth of field, no lens distortions.",
                            "lighting": "Consistent studio lighting, no unnatural shadows.", ...
   "ai_artifact_indicators": {"structure": "No unusual geometry or anatomical inconsistencies.",
                              "surfaces": "Textures appear natural, no waxy or plastic-like sheen.",
   "specialist_analysis": {"semantic": "0.049",  "frequency": "0.093",  "dual_stream": "0.049"

   "final_verdict": { "verdict": "REAL",

   "reasoning": "The lack of AI-related artifacts, The overall impression is of a staged, posed
}                image, but the details point towards a real-world scenario."
</answer>
```

Q: Is this image real or generated by AI?

```
<think>
Depicts a church-like...... My initial impression looks somewhat artificial, possibly AI-generated.
The structure of the church, exhibits some unusual geometric distortions......
Considering all the factors, I believe this image is AI-generated. strongly suggest AI involvement.
</think>

<answer>
{ "technical_assessment": { "lighting": "Inconsistent lighting, unnatural shadows, lack of realistic...
  "ai_artifact_indicators": {"structure": "Unusual geometric distortions around the domes.",
                             "edges": "Halo effect around the edges of the domes, suggesting
                             masking or compositing.",
  "specialist_analysis": {"semantic": "0.516",  "frequency": "0.999",  "dual_stream": "0.762"
  "final_verdict": {"verdict": "AI-GENERATED":
  "reasoning": "Inconsistent perspective, lighting, details, and spatial elements, combined with
}               unnatural textures and edge artifacts, strongly suggests AI generation."
</answer>
```

Figure 3: **Qualitative Results:** We present qualitative examples demonstrating ThinkFake's capability in detecting AI-generated images and providing corresponding explanations.

encourage reasoning behavior. We also implemented the DeepSeek-R1 pipeline, which begins with a cold-start initialization followed by GRPO training. Using the SFT-set, we first obtained "ThinkFake-SFT", then applied reinforcement learning to produce the final model, "ThinkFake-R1". As shown in Tab. 3, ThinkFake-Zero improved upon the base model but exhibited limited generalization. This suggests that applying RL alone to enhance reasoning in MLLMs remains challenging. In contrast, ThinkFake-SFT, trained with high-quality supervision, achieved stable output formatting and basic task competency. Subsequent RL fine-tuning led to ThinkFake-R1, which demonstrated strong generalization across diverse scenarios.

Reward Function. In Tab. 5, we evaluate the impact of different reward functions on the framework's generalization across all test set. Our results highlight the critical role of carefully designed rewards. The accuracy reward (f_{acc}) is key to driving convergence during reinforcement learning, while omitting the reasoning format reward (f_{format}) leads to unstructured and unstable outputs. Together, f_{acc} and f_{format} provide a stable foundation for effective training. Adding the JSON structure reward (f_{json}) boosts performance by enforcing outputs aligned with the detection pipeline. Incorporating the agentic reward ($f_{agentic}$), guided by expert agents, further enhances reasoning and enables robust, comprehensive detection.

4.6 Qualitative Results

Fig. 3 showcases the detection and explanation results generated by ThinkFake on both real and AI-generated images. Our method provides detailed thinking processes, diverse perspectives of judgment, and intuitive, human-friendly explanations, clearly demonstrating its effectiveness. *Additional qualitative examples are provided in the Appendix. E.*

5 Conclusions

In this paper, we introduced ThinkFake, a reasoning MLLM framework for detecting AI-generated images. By integrating a forgery reasoning prompt, GRPO reinforcement learning, and carefully designed reward functions, our method achieves both accurate classification and interpretable reasoning. Evaluations under the GenImage protocol show that ThinkFake surpasses existing state-of-the-art detectors. Moreover, the model generalizes effectively to challenging real-world LOKI benchmarks, highlighting its robustness and practical value. These results highlight the potential of reasoning-based methods for trustworthy AI detection. ThinkFake offers a strong foundation for future research in explainable and generalizable media forensics.

References

2025. Midjourney. https://www.midjourney.com/home.

(2024), O. 2024. GPT-4o System Card. arXiv:2410.21276.

(2025), O. 2025. OpenAI o3 and o4-mini System Card.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chang, Y.-M.; Yeh, C.; Chiu, W.-C.; and Yu, N. 2023. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 24185–24198.

Chu, T.; Zhai, Y.; Yang, J.; Tong, S.; Xie, S.; Schuurmans, D.; Le, Q. V.; Levine, S.; and Ma, Y. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *NeurIPS*, 34: 8780–8794.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; Podell, D.; Dockhorn, T.; English, Z.; and Rombach, R. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *ICML*.

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging frequency analysis for deep fake image recognition. In *ICML*, 3247–3258. PMLR.

Gao, Y.; Chang, D.; Yu, B.; Qin, H.; Chen, L.; Liang, K.; and Ma, Z. 2025. FakeReasoning: Towards Generalizable Forgery Detection and Reasoning. *arXiv preprint arXiv:2503.21210*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *NeurIPS*, 27.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

He, J.; Rungta, M.; Koleczek, D.; Sekhon, A.; Wang, F. X.; and Hasan, S. 2024. Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

Huang, C.-P.; Wu, Y.-H.; Chen, M.-H.; Wang, Y.-C. F.; and Yang, F.-E. 2025a. ThinkAct: Vision-Language-Action Reasoning via Reinforced Visual Latent Planning. *arXiv preprint arXiv:2507.16815*.

Huang, Z.; Hu, J.; Li, X.; He, Y.; Zhao, X.; Peng, B.; Wu, B.; Huang, X.; and Cheng, G. 2025b. SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model. In *CVPR*.

Huang, Z.; Xia, B.; Lin, Z.; Mou, Z.; and Yang, W. 2024. FFAA: Multimodal Large Language Model based Explainable Open-World Face Forgery Analysis Assistant. *arXiv preprint arXiv:2408.10072*.

Ji, Y.; Hong, Y.; Zhan, J.; Chen, H.; Zhu, H.; Wang, W.; Zhang, L.; Zhang, J.; et al. 2025. Towards Explainable Fake Image Detection with Multi-Modal Large Language Models. *arXiv preprint arXiv:2504.14245*.

Jia, S.; Lyu, R.; Zhao, K.; Chen, Y.; Yan, Z.; Ju, Y.; Hu, C.; Li, X.; Wu, B.; and Lyu, S. 2024. Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics. *CVPRW*.

Kang, H.; Wen, S.; Wen, Z.; Ye, J.; Li, W.; Feng, P.; Zhou, B.; Wang, B.; Lin, D.; Zhang, L.; and He, C. 2025. LEGION: Learning to Ground and Explain for Synthetic Image Detection. *arXiv preprint arXiv:2503.15264*.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Li, B.; Zhang, K.; Zhang, H.; Guo, D.; Zhang, R.; Li, F.; Zhang, Y.; Liu, Z.; and Li, C. 2024. LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild.

Li, X.; Hou, X.; and Loy, C. C. 2024. When stylegan meets stable diffusion: a w+ adapter for personalized image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2187–2196.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *NeurIPS*, volume 36, 34892–34916.

Liu, Y.; Peng, B.; Zhong, Z.; Yue, Z.; Lu, F.; Yu, B.; and Jia, J. 2025. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*.

Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 24480–24489.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *NeurIPS*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *ICML*, 8821–8831.

Ricker, J.; Lukovnikov, D.; and Fischer, A. 2024. AEROBLADE: Training-Free Detection of Latent Diffusion Images Using Autoencoder Reconstruction Error. In *CVPR*, 9130–9140.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv preprint arXiv:2112.10752*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Tam, Z. R.; Wu, C.-K.; Tsai, Y.-L.; Lin, C.-Y.; Lee, H.-y.; and Chen, Y.-N. 2024. Let me speak freely? a study on the impact of format restrictions on performance of large language models. *arXiv preprint arXiv:2408.02442*.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *CVPR*, 28130–28139.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-generated images are surprisingly easy to spot...for now. In *CVPR*.

Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Hu, H.; Chen, H.; and Li, H. 2023. Dire for diffusion-generated image detection. In *ICCV*, 22445–22455.

Wen, S.; Ye, J.; Feng, P.; Kang, H.; Wen, Z.; Chen, Y.; Wu, J.; Wu, W.; He, C.; and Li, W. 2025. Spot the Fake: Large Multimodal Model-Based Synthetic Image Detection with Artifact Explanation. *arXiv preprint arXiv:2503.14905*.

Xu, G.; Jin, P.; Li, H.; Song, Y.; Sun, L.; and Yuan, L. 2024. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. In *ICCV*.

Yan, S.; Li, O.; Cai, J.; Hao, Y.; Jiang, X.; Hu, Y.; and Xie, W. 2025. A Sanity Check for AI-generated Image Detection. In *ICLR*.

Yao, H.; Huang, J.; Wu, W.; Zhang, J.; Wang, Y.; Liu, S.; Wang, Y.; Song, Y.; Feng, H.; Shen, L.; et al. 2024. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*.

Ye, J.; Zhou, B.; Huang, Z.; Zhang, J.; Bai, T.; Kang, H.; He, J.; Lin, H.; Wang, Z.; Wu, T.; et al. 2024. LOKI: A Comprehensive Synthetic Data Detection Benchmark using Large Multimodal Models. *arXiv preprint arXiv:2410.09732*.

Zhong, N.; Xu, Y.; Li, S.; Qian, Z.; and Zhang, X. 2023. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*.

Zhou, Z.; Luo, Y.; Wu, Y.; Sun, K.; Ji, J.; Yan, K.; Ding, S.; Sun, X.; Wu, Y.; and Ji, R. 2025. AIGI-Holmes: Towards Explainable and Generalizable AI-Generated Image Detection via Multimodal Large Language Models. In *ICCV*.

Zhu, M.; Chen, H.; Yan, Q.; Huang, X.; Lin, G.; Li, W.; Tu, Z.; Hu, H.; Hu, J.; and Wang, Y. 2024. Genimage: A million-scale benchmark for detecting ai-generated image. *NeurIPS*, 36.

A Reward Functions

We carefully design four reward functions as follows:

Reasoning Format Reward (f_{format}). Pseudocode is provided in Algorithm 1.

JSON Format Reward (f_{json}). Pseudocode is provided in Algorithm 2.

Accuracy Reward (f_{acc}). Pseudocode is provided in Algorithm 3.

Agentic Reward ($f_{agentic}$). Pseudocode is provided in Algorithm 4.

B Training Process Details

As described in Section 3.4, we follow the GenImage protocol for data preparation. After obtaining the full GenImage (Zhu et al. 2024) training set, we uniformly sample 6,000 examples, consisting of 3,000 real and 3,000 fake images. This sampled data is then split into two subsets: 1,000 samples for supervised fine-tuning (SFT-set) and 5,000 samples for reinforcement learning (RL-set).

The original dataset provides only binary labels (real/-fake). To construct a high-quality SFT-set, we design a three-stage data preparation pipeline: **1)** We adapt the "Forgery Reasoning Prompt" as described in Sec. 3.2 and inject prior ground-truth information into the prompt to reduce hallucinations (see Fig. 4 for full prompt example). **2)** Using a commercial MLLM (e.g., Gemini-1.4-pro), we generate attribute-level explanations based on the prompts and corresponding images. **3)** A rule-based post-processing step filters out low-quality outputs to ensure consistency and accuracy. Through this process, we obtain 638 high-quality annotated samples, which form the final SFT-set. This carefully designed pipeline serves as a critical foundation for the strong performance of our model.

As described in Section 4.1, the training of our ThinkFake model consists of two stages. The first stage is a cold-start initialization, where we perform supervised fine-tuning using the SFT-set under the LLaMA-Factory [3] framework to align the model's output with the AI-generated image detection task. In the second stage, we adopt the VLM-R1 [4] framework and train the model on the RL-set using the GRPO reinforcement learning algorithm. During this phase, we employ four reward functions—"Reasoning Format Reward", "JSON Format Reward", "Accuracy Reward", and "Agentic Reward"—to guide and optimize the training process.

C Compare with SOTA Methods

We compare our method with several state-of-the-art approaches, including CNNSpot (Wang et al. 2020), Antifake-Prompt (Chang et al. 2023), UnivFD (Ojha, Li, and Lee 2023), NPR (Tan et al. 2024), and AIDE (Yan et al. 2025). For fair comparison, we reproduce their results using their official open-source codes and retrain all models on the same sub-training set constructed under the GenImage (Zhu et al. 2024) protocol, as described in Section 3.4. Evaluation is conducted on a cross test-set to assess generalization. Note that only SD v1.4 overlaps with the training domain, while all other test sets contain unseen categories.

Algorithm 1: Reasoning Format Reward

```
Input: $Completions$: A list of content generated by the model.
Output: $Rewards$: A list of reward values.
Initialize $Rewards \leftarrow []$
Define $Pattern \leftarrow$ <think>...</think> <answer>...</answer>
for all $C$ in $Completions$ do
    if $C$ matches the full $Pattern$ then
        Append 1.0 to $Rewards$
    else
        Append 0.0 to $Rewards$
    end if
end for
return $Rewards$
```

Algorithm 2: JSON Format Reward

```
Input: $Completions$: A list of content generated by the model.
Output: $Rewards$: A list of reward values.
Initialize $Rewards \leftarrow []$
for all $C$ in $Completions$ do
    $R \leftarrow 0.0$
    try
        $ExtractedJSON \leftarrow$ Extract from $C$ within ```json...```
        if $ExtractedJSON$ is not empty then
            Parse $ExtractedJSON$
            $R \leftarrow 1.0$
        end if
    catch JSONParseException
        $R \leftarrow 0.0$
    end try
    Append $R$ to $Rewards$
end for
return $Rewards$
```

The full cross-evaluation results are reported in Table 6, with accuracy (ACC. %) as the evaluation metric. Despite being trained on a limited dataset, our ThinkFake model demonstrates strong and consistent generalization across diverse test domains. This suggests that the model effectively performs reasoning-based detection. Furthermore, unlike other detectors that provide only binary predictions, ThinkFake generates interpretable, structured outputs—highlighting its effectiveness in explainable AI-generated image detection.

D LOKI Benchmark

Generalization has long been one of the key challenges in AI-generated image detection, limiting the practical applicability of previous methods. The Image modality of the LOKI (Ye et al. 2024) benchmark offers a valuable setting

---

[3] https://github.com/hiyouga/LLaMA-Factory

[4] https://github.com/om-ai-lab/VLM-R1

Algorithm 3: Accuracy Reward

```
Input: $Completions$: A list of content generated by the model. $Solutions$: The ground truth label for each completion.
Output: $Rewards$: A list of calculated reward values.
Initialize $Rewards \leftarrow []$
for all $(C, S)$ in zip($Completions$, $Solutions$) do
    $R \leftarrow -1.0$
    try
        $ExtractedJSON \leftarrow$ Extract JSON from $C$
        $ParsedAnswer \leftarrow$ Parse $ExtractedJSON$
        $Verdict \leftarrow ParsedAnswer.final\_verdict.verdict$
        if ($S = 0$ and $Verdict =$ “REAL”) or ($S = 1$ and $Verdict =$ “AI-GENERATED”) then
            $R \leftarrow 1.0$
        else
            $R \leftarrow 0.0$
        end if
    catch JSONParseException
        $R \leftarrow -1.0$
    end try
    Append $R$ to $Rewards$
end for
return $Rewards$
```

Algorithm 4: Agentic Reward

```
Input: $Completions$: A list of content generated by the model. $Solutions$: The ground truth label for each completion. $AgenticSolutions$: The ground truth probability distribution data.
Output: $Rewards$: A list of calculated reward values.
Initialize $Rewards \leftarrow []$
$ProbKeys \leftarrow [prob_{semantic}, prob_{frequency}, prob_{dual}]$.
for all $(C, S, A)$ in zip($Completions$, $Solutions$, $AgenticSolutions$) do
    $R \leftarrow -1.0$
    try
        $ExtractedJSON \leftarrow$ Extract JSON from $C$
        $ParsedAnswer \leftarrow$ Parse $ExtractedJSON$
        $Verdict \leftarrow ParsedAnswer.final\_verdict.verdict$
        if ($S = 0$ and $Verdict =$ “REAL”) or ($S = 1$ and $Verdict =$ ‘AI-GENERATED”) then
            Initialize $Losses \leftarrow []$
            $PredProbs \leftarrow ParsedAnswer.specialist\_analysis$
            for all $key$ in $ProbKeys$ do
                $p_{\text{pred}} \leftarrow PredProbs[key]$
                $p_{\text{gt}} \leftarrow A[key]$
                $\mathcal{L}_{\text{BCE}} \leftarrow$ BinaryCrossEntropy$(p_{\text{pred}}, p_{\text{gt}})$
                Append $\mathcal{L}_{\text{BCE}}$ to $Losses$
            end for
            $\mathcal{L}_{\text{avg}} \leftarrow$ Mean$(Losses)$
            $R \leftarrow e^{-\mathcal{L}_{\text{avg}}}$
        else
            $R \leftarrow 0.0$
        end if
    catch JSONParseException
        $R \leftarrow -1.0$
    end try
    Append $R$ to $Rewards$
end for
return $Rewards$
```

for evaluating generalization in real-world scenarios. This evaluation dataset contains over 2,200 images from seven real-world subcategories (e.g., Scene, Animal, Person, Object, Medicine, Document, Satellite), collected through existing datasets, internet sources, and newly synthesized data. Image generation methods include FLUX, Midjourney, Stable Diffusion, and ten additional techniques to ensure high quality and diversity. The task requires the model to determine whether each image is real or synthetic, presenting a challenging evaluation of generalization to complex and diverse scenarios.

Following the LOKI protocol and its comprehensive evaluation results, we conduct a zero-shot evaluation of our ThinkFake method. The results are presented in Table D. Our method demonstrates unprecedented detection performance and robustness across diverse categories. While performance slightly decreases in rare and domain-specific categories such as “Medicine” and “Document”, the overall results highlight the effectiveness of our framework. We attribute this success to the combination of our reasoning-based design and the generalization capability enabled by reinforcement learning, which allows the model to reason effectively even in unseen scenarios. Notably, although “Satellite” is also a rare category, our ThinkFake model achieves outstanding performance. As shown in Fig. 5, the model is able to think accurately and provide interpretable answers. This success can be attributed not only to the strong knowledge base of the underlying base MLLM, but more importantly, to our carefully designed reasoning pipeline. This pipeline guides the model to analyze, differentiate, and summarize before making a final decision with explanation—demonstrating practical applicability in real-world scenarios.

E Qualitative Results

We provide additional qualitative examples in Fig.6 and Fig.7, showcasing the complete outputs generated by ThinkFake. These outputs include the full step-by-step reasoning process, along with clearly structured answers. ThinkFake demonstrates not only accurate detection but also user-friendly and interpretable explanations.

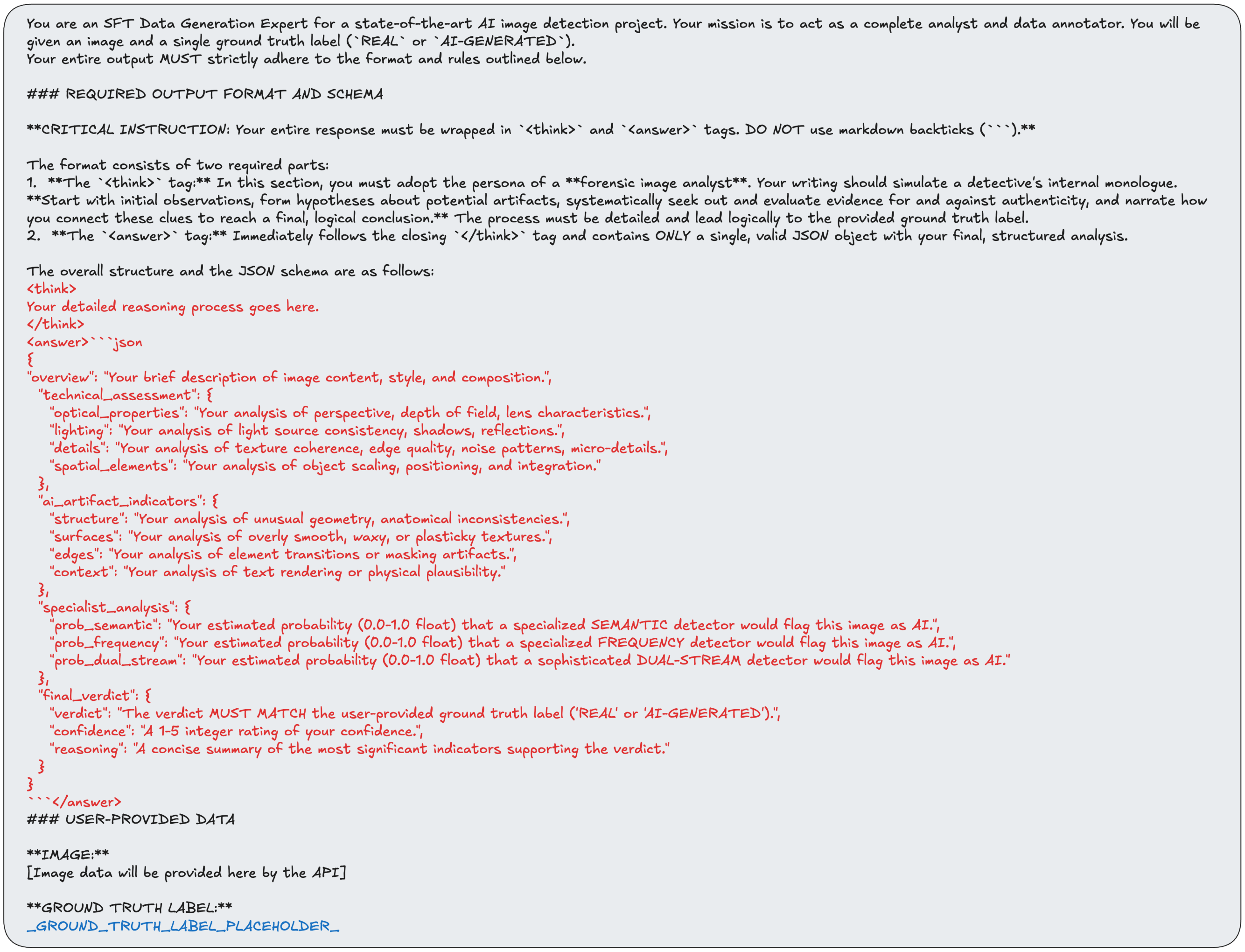
F Limitation and Future Work

Limitation. Our proposed ThinkFake framework establishes a complete detection pipeline and leverages reinforcement learning with carefully designed reward functions for training. Unlike previous methods (Gao et al. 2025; Zhou et al. 2025), which rely heavily on large-scale supervised fine-tuning (SFT) data and techniques such as prompt optimization or in-context learning to simulate reasoning behavior, ThinkFake achieves more generalized and effective results using only a small amount of SFT data and coarse-labeled RL data. Through a well-designed pipeline, ThinkFake enables the model to perform self-reasoning and structured detection.

However, since reinforcement learning is applied directly to the base MLLM, we freeze the visual encoder and fully fine-tune the LLM component to reduce computational cost. Despite this, a key limitation of ThinkFake is its requirement for substantial computational resources during training. Additionally, because the model relies on reasoning to produce decisions, real-time detection remains challenging during inference.

Future Work. In this paper, ThinkFake has demonstrated strong detection performance, interpretability, and robust generalization to real-world challenges. However, for end users, visual cues are often more intuitive than textual explanations. Future work could integrate segmentation models to map ThinkFake's reasoning outcomes directly onto the image, highlighting relevant regions. Combining visual annotations with textual reasoning may lead to more human-friendly and understandable explanations.

As AI-generated media continues to grow, so do the associated risks. We believe that detection technologies must evolve alongside generation techniques to ensure a safer and more trustworthy digital future.



~~~
You are an SFT Data Generation Expert for a state-of-the-art AI image detection project. Your mission is to act as a complete analyst and data annotator. You will be
given an image and a single ground truth label (`REAL` or `AI-GENERATED`).
Your entire output MUST strictly adhere to the format and rules outlined below.

### REQUIRED OUTPUT FORMAT AND SCHEMA

**CRITICAL INSTRUCTION: Your entire response must be wrapped in `<think>` and `<answer>` tags. DO NOT use markdown backticks (```).**

The format consists of two required parts:
1.  **The `<think>` tag:** In this section, you must adopt the persona of a **forensic image analyst**. Your writing should simulate a detective's internal monologue.
**Start with initial observations, form hypotheses about potential artifacts, systematically seek out and evaluate evidence for and against authenticity, and narrate how
you connect these clues to reach a final, logical conclusion.** The process must be detailed and lead logically to the provided ground truth label.
2.  **The `<answer>` tag:** Immediately follows the closing `</think>` tag and contains ONLY a single, valid JSON object with your final, structured analysis.

The overall structure and the JSON schema are as follows:
<think>
Your detailed reasoning process goes here.
</think>
<answer>```json
{
"overview": "Your brief description of image content, style, and composition.",
  "technical_assessment": {
    "optical_properties": "Your analysis of perspective, depth of field, lens characteristics.",
    "lighting": "Your analysis of light source consistency, shadows, reflections.",
    "details": "Your analysis of texture coherence, edge quality, noise patterns, micro-details.",
    "spatial_elements": "Your analysis of object scaling, positioning, and integration."
  },
  "ai_artifact_indicators": {
    "structure": "Your analysis of unusual geometry, anatomical inconsistencies.",
    "surfaces": "Your analysis of overly smooth, waxy, or plasticky textures.",
    "edges": "Your analysis of element transitions or masking artifacts.",
    "context": "Your analysis of text rendering or physical plausibility."
  },
  "specialist_analysis": {
    "prob_semantic": "Your estimated probability (0.0-1.0 float) that a specialized SEMANTIC detector would flag this image as AI.",
    "prob_frequency": "Your estimated probability (0.0-1.0 float) that a specialized FREQUENCY detector would flag this image as AI.",
    "prob_dual_stream": "Your estimated probability (0.0-1.0 float) that a sophisticated DUAL-STREAM detector would flag this image as AI."
  },
  "final_verdict": {
    "verdict": "The verdict MUST MATCH the user-provided ground truth label ('REAL' or 'AI-GENERATED').",
    "confidence": "A 1-5 integer rating of your confidence.",
    "reasoning": "A concise summary of the most significant indicators supporting the verdict."
  }
}
```</answer>
### USER-PROVIDED DATA

**IMAGE:**
[Image data will be provided here by the API]

**GROUND TRUTH LABEL:**
_GROUND_TRUTH_LABEL_PLACEHOLDER_
~~~

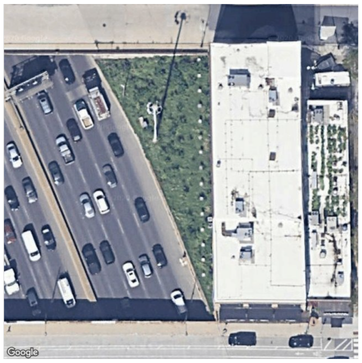
Figure 4: **SFT-set Preparation Prompt:** We adapt the "Forgery Reasoning Prompt" as the base and design a ground-truth injection mechanism to guide the MLLM in generating high-quality reasoning without hallucination.

Method	Midjourney	SD v1.4	SD v1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	*Mean*
CNNSpot	49.9	84.9	85.1	50.6	53.2	70.0	51.9	55.2	62.6
AntifakePrompt	77.1	81.6	81.4	72.3	82.1	80.4	69.8	70.4	76.9
UnivFD	76.6	80.5	80.6	58.5	78.9	73.9	53.1	**76.7**	72.3
NPR	61.1	61.3	59.2	55.9	61.4	60.0	58.2	60.1	59.6
AIDE	73.6	**98.7**	**99.3**	64.7	81.6	**95.6**	**73.6**	67.0	81.7
ThinkFake (Ours)	**92.5**	93.1	95.3	**73.1**	**87.4**	93.6	66.2	70.8	**84.0**

Table 6: **Comparison of State-of-the-Art Methods:** All baselines are re-trained on our training set to ensure a fair comparison, following the GenImage evaluation protocol. Performance comparison in Acc. (%)

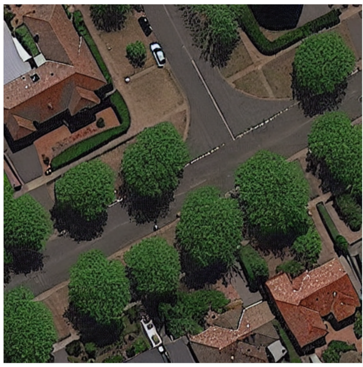
	Overall	**Scene**	**Animal**	**Person**	**Object**	**Medicine**	**Doc**	**Satellite**
Random Choice	18.0	21.6	18.3	18.6	26.0	22.2	22.1	22.1
Human	27.3	24.0	25.8	19.9	26.9	26.1	22.1	39.3
AIDE (Yan et al. 2025)	63.1	-	**89.9**	62.5	**96.5**	53.4	49.7	39.3
MiniCPM-V-2.6	44.8	52.0	34.4	53.1	31.5	53.8	51.5	38.3
Phi-3.5-Vision	52.5	50.8	41.7	71.5	34.1	57.3	54.3	60.5
LLaVA-OneVision-7B	49.8	59.2	41.9	58.1	37.3	52.3	53.0	50.1
InternLM-XComposer2.5	46.4	52.7	40.0	56.7	32.5	56.1	49.8	38.2
mPLUG-Owl3-7B	45.9	52.1	37.3	52.9	31.4	55.3	53.8	38.1
Qwen2-VL-7B	47.8	54.7	38.9	57.9	30.3	56.0	59.6	36.9
LongVA-7B	46.2	57.6	37.4	52.5	34.1	54.4	49.8	39.7
Mantis-8B	54.6	54.9	52.2	54.8	53.5	53.1	51.9	63.3
Idefics2-8B	45.0	51.8	35.3	52.3	29.2	52.3	53.9	40.6
InternVL2-8B	49.7	58.8	39.4	54.4	37.8	53.9	60.2	44.2
Llama3-LongVILA-8B	49.8	49.8	50.5	50.6	47.2	50.0	49.9	50.0
VILA1.5-13B	49.3	52.0	38.6	54.2	31.0	50.1	56.6	62.4
InternVL2-26B	44.3	51.6	35.4	50.8	28.2	51.3	54.4	37.6
VILA1.5-40B	48.8	53.7	39.3	50.0	33.4	52.5	59.9	50.6
InternVL2-40B	49.6	55.7	37.3	59.2	34.8	55.5	64.8	40.8
Qwen2-VL-72B	53.2	55.9	43.4	66.9	38.0	55.9	**73.7**	38.2
LLaVA-OneVision-72b	46.3	54.7	31.6	53.1	27.8	52.1	67.9	36.6
Claude-3.5-Sonnet	53.6	51.6	51.6	55.2	51.4	51.9	59.1	50.9
Gemini-1.5-Pro	43.5	53.7	35.7	51.5	30.3	50.0	47.2	38.1
GPT-4o	63.4	70.1	69.7	**84.4**	70.3	54.3	60.1	45.0
ThinkFake (Ours)	**75.4**	**78.5**	87.2	71.7	89.4	**64.5**	54.2	**82.2**

Table 7: Accuracy (%) of different models on judgment questions from the Image modality of the LOKI benchmark. All categories are unseen zero-shot test cases. Best results are in bold, and second-best are underlined.



Q: Is this image real or generated by AI?

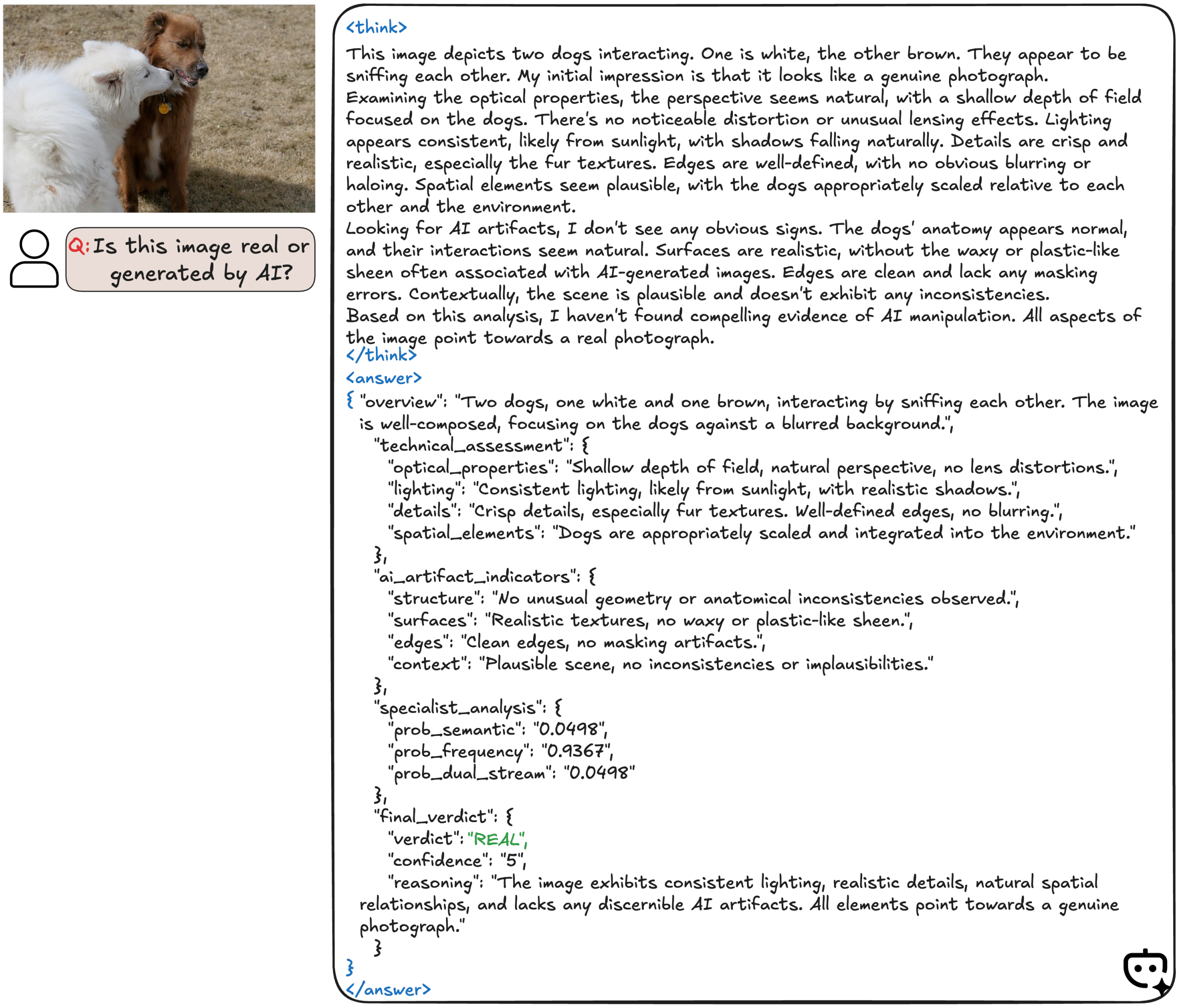
```
<think>
This image appears to depict a highway interchange with a large building and some greenery. My
initial impression is that it's a real photograph.
**Optical Properties:** The perspective seems natural, typical of satellite imagery.
......
**Final Verdict:** Based on my analysis, I conclude that this image is REAL. The image
exhibits all the hallmarks of a genuine satellite photograph, including natural lighting, realistic
textures, and consistent spatial relationships.
</think>
<answer>
{"overview": "Aerial view of a highway interchange with a large building and greenery.",
 "technical_assessment": {
     "optical_properties": "Natural perspective and depth of field, consistent with real-world
     aerial photography.",
     "lighting": "Even and consistent lighting, plausible shadows.",
     "details": "Realistic textures, well-defined edges, no discernible noise.",
     "spatial_elements": "Natural scaling and positioning of objects." },
 "ai_artifact_indicators": {
     "structure": "No unusual geometry or inconsistencies.",
     "surfaces": "No overly smooth or plasticky textures.",
     "edges": "Clean and natural transitions between elements.",
     "context": "Plausible scene context." },
 "specialist_analysis": { "prob_semantic": 0.05, "prob_frequency": 0.05, "prob_dual_stream": 0.05 },
 "final_verdict": {
     "verdict": "REAL",
     "confidence": 5,
     "reasoning": "The image exhibits all the characteristics of a real satellite photograph,
     including natural lighting, realistic textures, and consistent spatial relationships." }
}
</answer>
```



Q: Is this image real or generated by AI?

```
<think>
This aerial view of a residential area appears initially plausible.
However, closer inspection reveals several inconsistencies.
......
Considering these observations, I suspect this image is AI-generated. The overly perfect
trees, the uniformity of the houses and landscaping, and the inconsistencies in lighting and
shadows all point towards artificial origins. The lack of subtle imperfections and the almost
"too perfect" appearance of the scene are strong indicators of AI involvement.
</think>
<answer>
{ "overview": "Aerial view of a residential neighborhood with houses, trees, and streets. The
  image has a somewhat artificial, staged quality.",
  "technical_assessment": {
       "optical_properties": "Perspective seems slightly distorted, depth of field is shallow,
       and the overall image lacks natural lens characteristics.",
       "lighting": "Lighting appears inconsistent, shadows are unconvincing, and there's a
       lack of natural light gradation.",
       "details": "Textures are overly smooth, particularly on trees and hedges. Edge
       quality is too sharp, lacking natural organic variation.",
       "spatial_elements": "Scaling of elements is slightly off, and the interaction between
       trees and the environment appears unnatural." },
  "ai_artifact_indicators": {
       "structure": "Trees exhibit an unnatural uniformity in shape and size. The overall
       composition feels artificially planned.",
       "surfaces": "Textures appear overly smooth and plasticky, lacking the complexity of
       real-world materials.",
       "edges": "Edges of trees and hedges are too sharp and lack natural irregularities.",
       "context": "The overall context feels staged and lacks the randomness of a real
       neighborhood."},
  "specialist_analysis": {"prob_semantic":"0.95","prob_frequency":"0.85","prob_dual_stream":"0.98"},
  "final_verdict": {
       "verdict": "AI-GENERATED",
       "confidence": "4",
       "reasoning": "The overly perfect trees, uniform landscaping, inconsistent lighting, and
       unnatural spatial relationships strongly suggest AI generation."}
}
</answer>
```

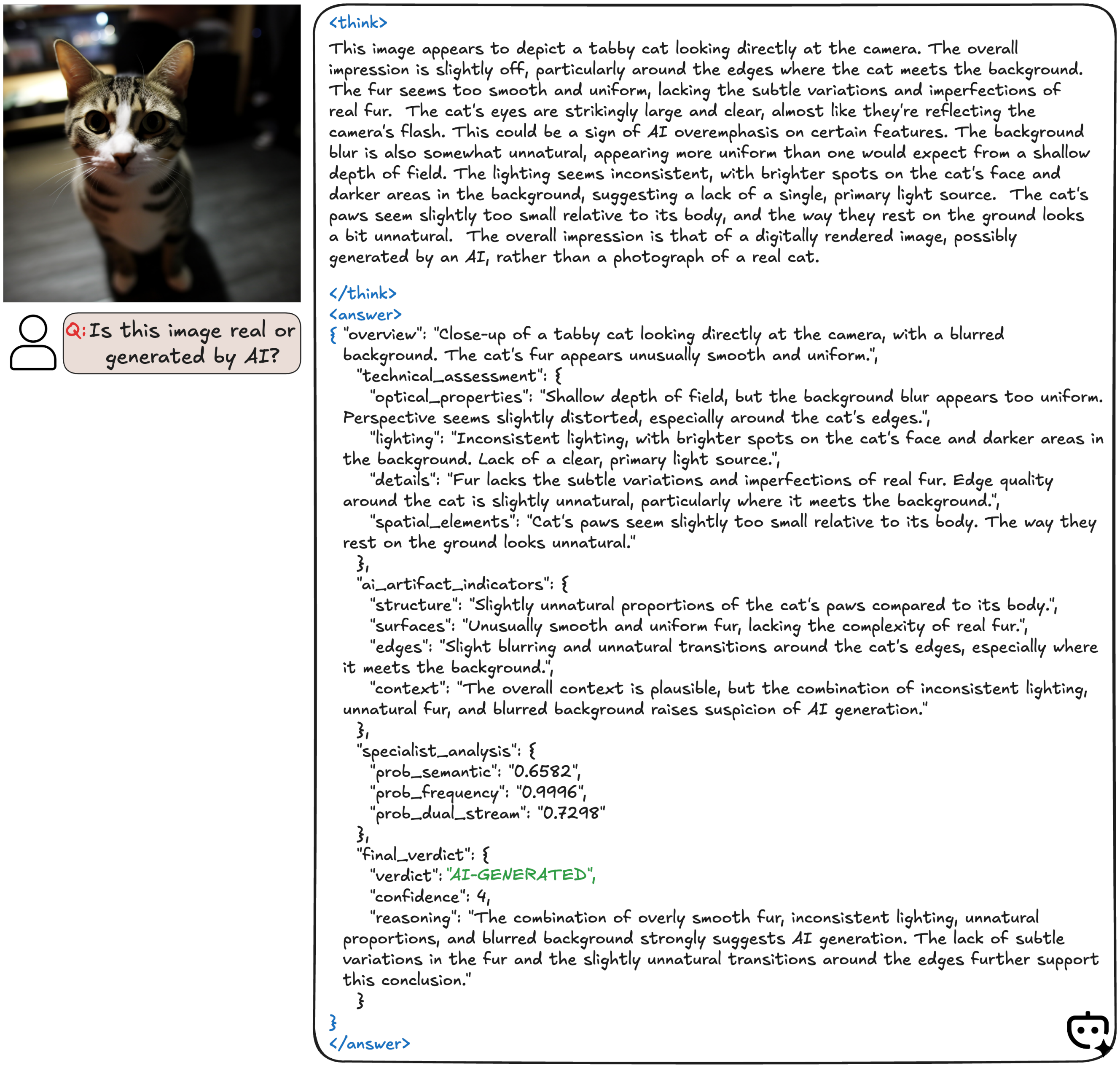
Figure 5: **Qualitative Results:** We present qualitative examples of ThinkFake on the "Satellite" category from the LOKI benchmark. Despite the rarity of such images, ThinkFake consistently provides high-quality and robust predictions along with clear explanations.



Q: Is this image real or generated by AI?

```
<think>
This image depicts two dogs interacting. One is white, the other brown. They appear to be
sniffing each other. My initial impression is that it looks like a genuine photograph.
Examining the optical properties, the perspective seems natural, with a shallow depth of field
focused on the dogs. There's no noticeable distortion or unusual lensing effects. Lighting
appears consistent, likely from sunlight, with shadows falling naturally. Details are crisp and
realistic, especially the fur textures. Edges are well-defined, with no obvious blurring or
haloing. Spatial elements seem plausible, with the dogs appropriately scaled relative to each
other and the environment.
Looking for AI artifacts, I don't see any obvious signs. The dogs' anatomy appears normal,
and their interactions seem natural. Surfaces are realistic, without the waxy or plastic-like
sheen often associated with AI-generated images. Edges are clean and lack any masking
errors. Contextually, the scene is plausible and doesn't exhibit any inconsistencies.
Based on this analysis, I haven't found compelling evidence of AI manipulation. All aspects of
the image point towards a real photograph.
</think>
<answer>
{ "overview": "Two dogs, one white and one brown, interacting by sniffing each other. The image
  is well-composed, focusing on the dogs against a blurred background.",
    "technical_assessment": {
      "optical_properties": "Shallow depth of field, natural perspective, no lens distortions.",
      "lighting": "Consistent lighting, likely from sunlight, with realistic shadows.",
      "details": "Crisp details, especially fur textures. Well-defined edges, no blurring.",
      "spatial_elements": "Dogs are appropriately scaled and integrated into the environment."
    },
    "ai_artifact_indicators": {
      "structure": "No unusual geometry or anatomical inconsistencies observed.",
      "surfaces": "Realistic textures, no waxy or plastic-like sheen.",
      "edges": "Clean edges, no masking artifacts.",
      "context": "Plausible scene, no inconsistencies or implausibilities."
    },
    "specialist_analysis": {
      "prob_semantic": "0.0498",
      "prob_frequency": "0.9367",
      "prob_dual_stream": "0.0498"
    },
    "final_verdict": {
      "verdict": "REAL",
      "confidence": "5",
      "reasoning": "The image exhibits consistent lighting, realistic details, natural spatial
  relationships, and lacks any discernible AI artifacts. All elements point towards a genuine
  photograph."
    }
}
</answer>
```

Figure 6: **Qualitative Result:** Detection and explanation of real image by ThinkFake.



Q:Is this image real or generated by AI?

```
<think>
This image appears to depict a tabby cat looking directly at the camera. The overall
impression is slightly off, particularly around the edges where the cat meets the background.
The fur seems too smooth and uniform, lacking the subtle variations and imperfections of
real fur.  The cat's eyes are strikingly large and clear, almost like they're reflecting the
camera's flash. This could be a sign of AI overemphasis on certain features. The background
blur is also somewhat unnatural, appearing more uniform than one would expect from a shallow
depth of field. The lighting seems inconsistent, with brighter spots on the cat's face and
darker areas in the background, suggesting a lack of a single, primary light source.  The cat's
paws seem slightly too small relative to its body, and the way they rest on the ground looks
a bit unnatural.  The overall impression is that of a digitally rendered image, possibly
generated by an AI, rather than a photograph of a real cat.

</think>
<answer>
{ "overview": "Close-up of a tabby cat looking directly at the camera, with a blurred
  background. The cat's fur appears unusually smooth and uniform.",
    "technical_assessment": {
      "optical_properties": "Shallow depth of field, but the background blur appears too uniform.
  Perspective seems slightly distorted, especially around the cat's edges.",
      "lighting": "Inconsistent lighting, with brighter spots on the cat's face and darker areas in
  the background. Lack of a clear, primary light source.",
      "details": "Fur lacks the subtle variations and imperfections of real fur. Edge quality
  around the cat is slightly unnatural, particularly where it meets the background.",
      "spatial_elements": "Cat's paws seem slightly too small relative to its body. The way they
  rest on the ground looks unnatural."
    },
    "ai_artifact_indicators": {
      "structure": "Slightly unnatural proportions of the cat's paws compared to its body.",
      "surfaces": "Unusually smooth and uniform fur, lacking the complexity of real fur.",
      "edges": "Slight blurring and unnatural transitions around the cat's edges, especially where
  it meets the background.",
      "context": "The overall context is plausible, but the combination of inconsistent lighting,
  unnatural fur, and blurred background raises suspicion of AI generation."
    },
    "specialist_analysis": {
      "prob_semantic": "0.6582",
      "prob_frequency": "0.9996",
      "prob_dual_stream": "0.7298"
    },
    "final_verdict": {
      "verdict": "AI-GENERATED",
      "confidence": 4,
      "reasoning": "The combination of overly smooth fur, inconsistent lighting, unnatural
  proportions, and blurred background strongly suggests AI generation. The lack of subtle
  variations in the fur and the slightly unnatural transitions around the edges further support
  this conclusion."
    }
}
</answer>
```

Figure 7: **Qualitative Result:** Detection and explanation of AI-generated image by ThinkFake.