MIRAGE: Towards AI-Generated Image Detection in the Wild

OuCheng Huang[1*], Manxi Lin[1*], Jiexiang Tan[1*], Xiaoxiong Du[1], Yang Qiu[1]
Junjun Zheng[1†], Xiangheng Kong[1†], Yuning Jiang[1], Bo Zheng[1]

[1]Algorithm Tech Team Taobao & Tmall Group of Alibaba

Abstract

The spreading of AI-generated images (AIGI), driven by advances in generative AI, poses a significant threat to information security and public trust. Existing AIGI detectors, while effective against images in clean laboratory settings, fail to generalize to **in-the-wild** scenarios. These real-world images are noisy, varying from "obviously fake" images to realistic ones derived from multiple generative models and further edited for quality control. We address **in-the-wild AIGI detection** in this paper. We introduce **MIRAGE**, a challenging benchmark designed to emulate the complexity of in-the-wild AIGI. **MIRAGE** is constructed from two sources: (1) a large corpus of Internet-sourced AIGI verified by human experts, and (2) a synthesized dataset created through the collaboration between multiple expert generators, closely simulating the realistic AIGI in the wild. Building on this benchmark, we propose **MIRAGE-R1**, a vision-language model with heuristic-to-analytic reasoning, a reflective reasoning mechanism for AIGI detection. **MIRAGE-R1** is trained in two stages: a supervised-fine-tuning cold start, followed by a reinforcement learning stage. By further adopting an inference-time adaptive thinking strategy, **MIRAGE-R1** is able to provide either a quick judgment or a more robust and accurate conclusion, effectively balancing inference speed and performance. Extensive experiments show that our model leads state-of-the-art detectors by **5%** and **10%** on **MIRAGE** and public benchmark, respectively.

Code — https://github.com/och-alibaba/mirage

Introduction

The rapid development of generative AI has made the creation of AI-generated images (AIGI) widely accessible (Rombach et al. 2022; Cai et al. 2025). While this technology benefits creative industry, it also raises serious concerns. On one hand, photorealistic forgeries can lead to problems like copyright issues (Rombach et al. 2022; Ren et al. 2024); on the other hand, the mass spread of obviously-fake AI images can flood digital platforms, harming user experience and authenticity of the platform (Yin and Liu 2024). In response, recent works have developed AIGI detectors (Xu et al. 2024; Huang et al. 2025; Zhou et al. 2025; Gao et al. 2025) that shows impressive performance in separating real photos from even the most convincing synthetic ones.

In long-established computer vision tasks like object detection (Chen et al. 2018; Xiang, Mottaghi, and Savarese 2014) and face recognition (Liu et al. 2015), "in-the-wild" evaluation is standard to measure real-world performance. In this paper, we address the **in-the-wild AIGI detection** task. Inspired by similar tasks in other vision fields, we define this problem as identifying *naturally-occurring* AI-generated images: those that have already circulated online or been altered in a real-world setting.

Benchmark. Most existing benchmarks (Wang et al. 2020; Zhu et al. 2023) confine AIGI to images from certain types of generative models in a controlled environment, without capturing the complexity and messiness of real-life scenarios. In contrast, recent studies (Yan et al. 2024a; Zhang et al. 2025) have emphasized the importance of using "real-world" fake images sourced from online AIGI communities, which are often realistic enough to deceive human experts (Yan et al. 2024a). However, the "in-the-wild" scenario is not limited to these two types of images. It also encompasses a large volume of fake images that, while potentially recognizable by humans, still exhibit a data distribution distinct from that of single-model outputs due to real-world factors like post-processing and quality control.

To address these gaps, we introduce **MIRAGE** — the first dataset focused on in-the-wild AIGI detection. As shown in Fig. 1, besides the images from the controlled environment, i.e., *vanilla generators* without further editing, our dataset also includes two types of positive examples: (1) *Human Curation:* We scrawled and annotated images, typically fake images, from different sources on the Internet. Many of these images clearly show signs of AI generation and are "obviously fake" to most viewers. (2) *AIGI from Composite Pipelines:* We also generate photorealistic AIGI by applying multiple pipelines, involving the collaboration between many models, as well as a range of post-processing steps (such as face restoration). These images serve as more challenging synthetic examples, aiming to capture the variety seen in uncontrolled, real-world settings.

Method. Recent works have revealed two critical shortcomings of existing AIGI detectors: limited generalization and a lack of explainability (Zhou et al. 2025; Gao et al.

*These authors contributed equally.
†Co-corresponding authors.

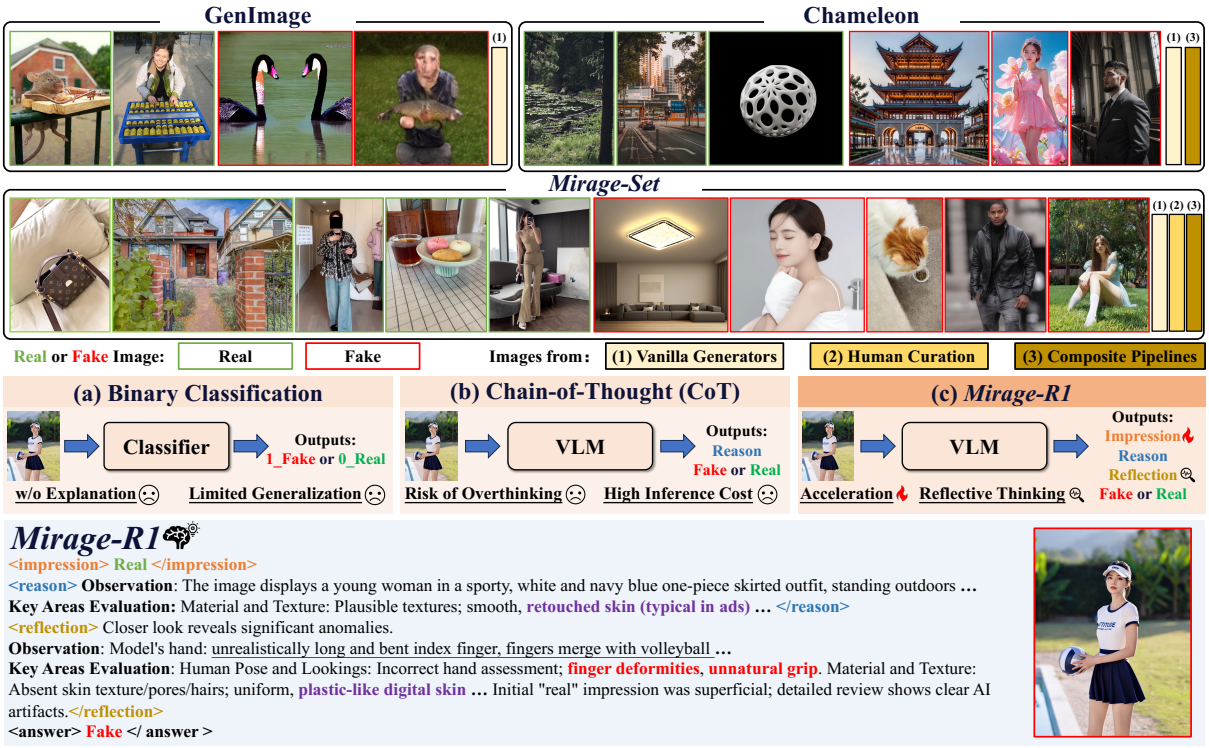

Figure 1: Overall review of our **MIRAGE** and **MIRAGE-R1**.

2025). Prior works (Zhang et al. 2025; Zhou et al. 2025) suggest that Vision-Language Models (VLMs), particularly those employing Chain-of-Thought (CoT) reasoning, hold promise in addressing these challenges due to their strong prior knowledge from large-scale pretraining. In CoT, complex problems are broken down into well-defined steps, such as image captioning and criterion-based judgment, which the VLM completes sequentially. The final decision is made after this explicit reasoning process, and the reasoning process serves as an explanation for the answer. However, CoT-based approaches usually increase inference costs and can be prone to "overthinking": on "easy" AIGI samples with obvious flaws, the model may make unnecessary mistakes during the long reasoning process. This issue is especially prominent in noisy, in-the-wild AIGI data, which often contains such "obviously fake" images.

To address these challenges, we introduce **MIRAGE-R1**, a VLM specifically tailored for robust AIGI detection in the wild, building upon **MIRAGE**. Fig. 1 shows an example of our model reasoning. **MIRAGE-R1** is progressively trained in two stages to think in a way, namely heuristic to analytic reasoning. As a consequence, the model is able to generate both a rapid initial prediction (`<impression>` in the figure) and a more deliberate final answer (`<answer>`), inspecting the first one. As shown in the figure, our model is able to perform reflection (`<reflection>`), that is, correcting the issues in its initial reasoning ((`<reason>`)). In inference, **MIRAGE-R1** adaptively selects its response according to its confidence in the fast answer. This allows **MIRAGE-R1** to efficiently balance accuracy and computational cost in the noisy in-the-wild scenarios.

Our main contributions are as follows:

- We define the task of in-the-wild AIGI detection and introduce **MIRAGE**, a comprehensive benchmark for real-world evaluation.
- We propose **MIRAGE-R1**, a VLM capable of adaptive and reflective reasoning for AIGI detection.
- Extensive experiments demonstrate that **MIRAGE-R1** significantly outperforms existing methods on both the **MIRAGE** benchmark and other public datasets.

Related Works

Due to the page limits, we put a detailed related works section in the appendix.

Benchmarks in AIGI Detection. Existing AIGI detection benchmarks (Wang et al. 2020; Zhu et al. 2023; Hong and Zhang 2024) often fail to model real-world complexity. Even recent benchmarks that simulate real-world AIGI (Yan et al. 2024a; Zhang et al. 2025) suffer from two critical limitations: 1) domain bias, introduced by sourcing real and AI-generated images from disparate online contexts, and 2) a lack of examples from complex pipelines involving fine-tuned models and post-processing, which is common in in-the-wild AIGI.

To bridge these gaps, we introduce **MIRAGE**, a benchmark designed for more rigorous generalization assessment.

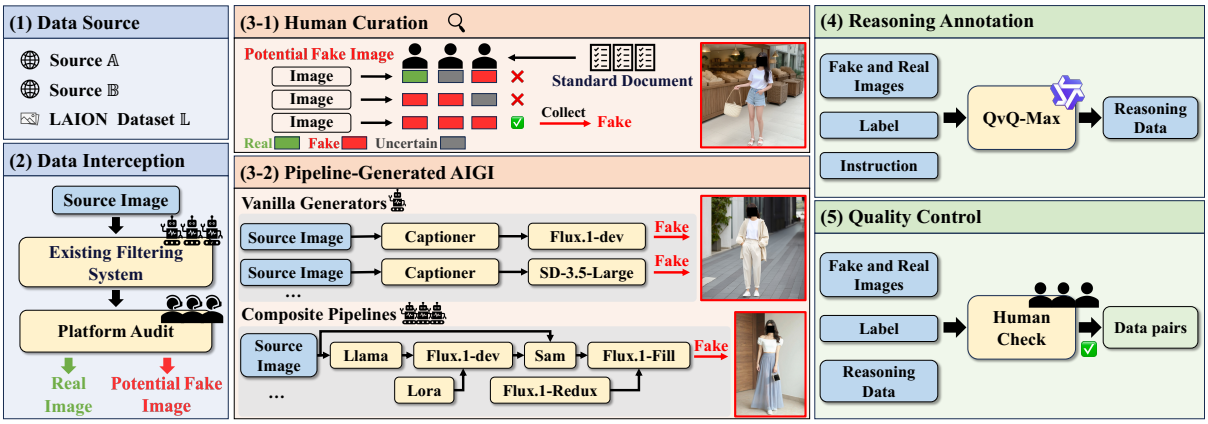

Figure 2: Illustration of the dataset construction of our **MIRAGE**. We have anonymized the faces.

It addresses these issues by including: (a) real and AI-generated images curated from the same online communities to eliminate domain bias, and (b) challenging forgeries created through advanced generation and post-processing pipelines, thus providing a more realistic evaluation.

AIGI Detection. AIGI detection has evolved from early, low-level artifact detectors (Wang et al. 2020; Tan et al. 2024) to modern approaches leveraging large pre-trained models like CLIP (Ojha, Li, and Lee 2023; Yan et al. 2024b,a) for improved generalization. However, these advanced methods suffer from two critical limitations: 1) they are often "black-box" models, lacking *explainability* (Gao et al. 2025; Zhou et al. 2025), and 2) their *generalization* is rarely tested on "in-the-wild" images that undergo unknown post-processing. To address this, recent work has shifted towards VLMs, yet these have introduced their own fundamental trade-off. They are split between fast but potentially unfaithful *heuristic (answer-first)* reasoning (Li et al. 2024; Wen et al. 2025), and robust but slow *analytic (reason-first)* reasoning (Gao et al. 2025; Zhang et al. 2025).

To address these challenges, we introduce **MIRAGE-R1**, a VLM that pioneers adaptive and reflective reasoning. It first provides a fast, heuristic answer, then adaptively decides whether to engage in deeper, reflective thinking by estimating its confidence in the fast answer. This allows **MIRAGE-R1** to achieve generalizable and explainable detection while dynamically allocating computational resources based on task difficulty, offering a unified solution to in-the-wild AIGI detection.

MIRAGE Dataset

Problem Formulation. We define in-the-wild AIGI to be naturally-occurring AI-generated images. Compared to conventional benchmark data, in-the-wild AIGI have unique characteristics: (1) *Mixed Presence with Real Photos:* In-the-wild AIGI appear in the same settings where real photos are expected, such as social media and daily news, instead of AI art communities. (2) *Generation from Pipeline:* These images often come from a pipeline: a combination of diverse fine-tuned generative models and potential further manual modification through, e.g., editing, compositing, or retouching for use in social media, advertisements etc. (3) *Wide Realism Range:* Real-world AIGI can span a wide range in realism, from obviously synthetic or heavily edited to highly convincing forgeries that are difficult for humans to identify.

Dataset	Vanilla Generators	Human Curation	Composite Pipelines
CNN-Detection	✓	✗	✗
GenImage	✓	✗	✗
LOKI	✓	✗	✗
Chameleon	✓	✗	✓
AIGI-Holmes	✓	✗	✗
MIRAGE	✓	✓	✓

Table 1: Data distribution across various benchmarks. Unlike existing benchmarks, our **MIRAGE** dataset is distinguished by its inclusion of AIGI sourced from different provenances. The columns correspond to these provenances: direct generation from vanilla generators without further post-processing, human-curated examples sourced online, and complex composite pipelines.

Based on these features, we build the dataset of **M**ixed **I**mages from **R**eal-world **A**I **Ge**neration (**MIRAGE**). As shown in Tab. 1, our **MIRAGE** has the most complete coverage of in-the-wild AIGI compared to existing benchmarks.

Dataset Construction

The construction of **MIRAGE** is a multi-stage process encompassing data sourcing, expert annotation of real-world AIGI, advanced pipeline-based generation, and rigorous quality control. An overview is illustrated in Fig. 2.

Data Sourcing and Human Curation. We sourced our initial data from three diverse origins: two large-scale proprietary websites, 𝔸 and 𝔹 [1], and the public LAION DATASET 𝕃 (Schuhmann et al. 2022). This collection contains a mix of real and potentially AI-generated images from scenarios such as social media, e-commerce, news, etc. A portion of this data, totaling 15,000 images with preliminary "real"

[1]To preserve anonymity, the sources are not named but represent major online platforms with a mix of user-generated content.

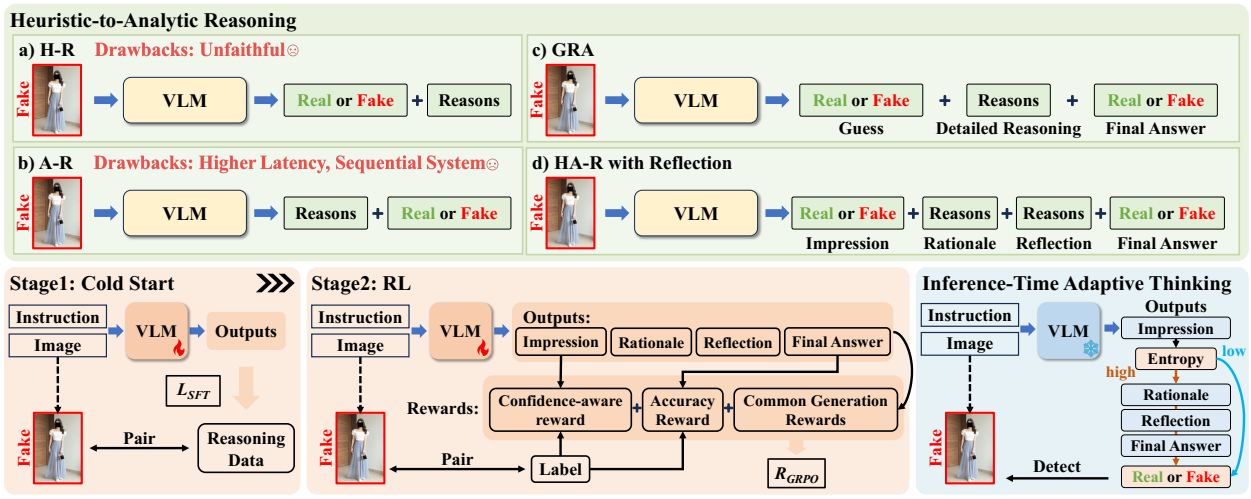

Figure 3: Training framework of our **MIRAGE-R1**.

or "AI-generated" flags from the source platforms, underwent a rigorous expert re-annotation process. We employed 29 trained annotators, with each image independently assessed by three different experts. They could label an image as "Real" "AI-generated", or "Uncertain". To ensure the highest data quality, we applied a strict filtering protocol: (1) Only images where all three annotators and the original platform label reached a unanimous decision were kept; (2) Images unanimously labeled as "Uncertain" were discarded. This process yielded 11,559 high-confidence images (8,465 Real, 3,094 AI-generated), forming our human-curated dataset.

AIGI Generation via Vanilla Generators and Composed Pipelines. Human annotation can only capture existing, identifiable forgeries. To cover emerging threats and complex manipulations, we identified 8 common real-world generation patterns: Text-to-Image (*T2I*), Inpainting/Outpainting (*IP&OP*), Instruction-based Editing (*IE*), Face Swapping (*FW*), Background Change (*CB*), Virtual Try-on (*VTO*), Realistic Model Generation (*RMG*), and Pose-Consistent Model Generation (*PCMG*). Here, *T2I* refers to the AIGI construction from vanilla generators (similar to most existing benchmarks), and the rest 7 patterns are all complicated composite pipelines that simulate the real-world application.

Using ComfyUI [2] and the Python interpreter, we built 64 automated pipelines integrating 53 different models (including text-to-image generators, LoRAs (Hu et al. 2022), ControlNets (Zhang, Rao, and Agrawala 2023), and other image processing models) to realize these generation patterns. For example, our *PCMG* pipeline first uses a VLM to generate a detailed prompt from a real photograph while extracting the subject's pose. Then, a combination of a fine-tuned LoRA, a T2I model, and ControlNet generates a new character in the same pose. As a final step, segmentation models (SAM (Kirillov et al. 2023) and GroundingDINO (Liu et al. 2024)) and inpainting are used to transfer the original clothing onto the generated character, creating a highly convincing forgery. Further details are available in the Appendix.

Reasoning Annotation and Quality Control. To train our VLM, we generated textual reasoning for each image using QvQ-Max (The Qwen Team 2025), providing it with the ground-truth label and an in-context example, similar to (Zhang et al. 2025; Huang et al. 2025). Prompts are given in the appendix.

Finally, both curated and generated images underwent a final quality control step. Five AIGI experts filtered out noisy or low-quality examples, ensuring that the final benchmark contains only challenging and representative samples.

Dataset Partition

The final **MIRAGE** benchmark is partitioned into two primary splits for comprehensive evaluation: In-Distribution (ID) and Out-of-Distribution (OOD).

ID Split. The ID split is designed for training and standard ID evaluation. It combines human-curated and T2I images from sources $\mathbb{A}$ and $\mathbb{L}$, maintaining a 1:1 real-to-fake ratio. This split contains 21,685 images, divided into a training set (20,000 images) and an ID test set (1,685 images).

OOD Split. The OOD split is crafted to rigorously test model generalization against unseen data sources. It contains 12,087 images and is composed of images from the unseen source $\mathbb{B}$ and images generated with the full set of 8 generation patterns from all sources. This split is further divided into multiple test subsets for granular analysis. This includes an OOD human-curated set to test generalization to new real image distributions. Furthermore, each of the 8 generation patterns forms an OOD pipeline-generated test set, composed of the real images from $\mathbb{B}$, and the generated images from the real images from $\mathbb{B}$, across all the 8 generation types. This allows for a fine-grained evaluation of a model's robustness to each manipulation type. Note that the T2I models for this split are different from the ones used in the ID split to keep the generated images OOD.

[2] https://github.com/comfyanonymous/ComfyUI

Methodology

The detection of AIGI in the wild demands a delicate balance of: (1) *Generalization:* As generative models rapidly evolve, detectors must perform reliably on OOD images from unseen models and diverse scenarios. (2) *Explainability:* In sensitive applications like content moderation, providing clear, human-understandable rationales for decisions is crucial for user trust. (3) *Efficiency:* Many practical scenarios, such as real-time media analysis, impose strict low-latency requirements.

To address these challenges, we introduce **MIRAGE-R1**, a VLM architected for in-the-wild AIGI detection. As illustrated in Fig. 3, our core innovation is a novel reasoning framework, Heuristic-to-Analytic Reasoning, which emulates the multi-stage cognitive process of human experts. We operationalize this framework through a two-stage progressive training paradigm. The resulting **MIRAGE-R1** is uniquely capable of delivering a fast initial verdict for efficiency, a detailed rationale for explainability, and a final, more accurate answer derived from reflective thinking for superior generalization.

Heuristic-to-Analytic Reasoning (HA-R)

VLM-based AIGI detectors typically adopt one of two reasoning paradigms. Let I be the input, A be the binary verdict, and R be the textual rationale. The key distinction lies in how they model the joint probability $p(A, R \mid I)$:

- **Heuristic Reasoning (H-R)** first generates an answer, modeling the probability as $p(A, R \mid I) = p(A \mid I) \cdot p(R \mid I, A)$. This approach is fast, as inference can be terminated after A is generated. However, the rationale R risks being a post-hoc rationalization rather than a cause for the verdict.
- **Analytic Reasoning (A-R)** first formulates a rationale, modeling the probability as $p(A, R \mid I) = p(R \mid I) \cdot p(A \mid I, R)$. This promotes causal thinking and often improves generalization, but at the cost of higher latency since the entire rationale R must be generated.

We aim to fuse these paradigms to harness their respective strengths. We develop our HA-R framework through an iterative design process.

Initial Formulation: Guess-Reason-Answer (G-R-A). Our first design was a direct combination of the two paradigms, inspired by human cognition: forming an initial "guess" (G), followed by deliberate reasoning (R) to reach a final answer (A). This sequence, $G \rightarrow R \rightarrow A$, is modeled probabilistically as: $p(G, R, A \mid I) = p(G \mid I) \cdot p(R \mid I, G) \cdot p(A \mid I, G, R)$. While this structure begins to emulate human thought, the connection between G and A remains implicit.

Final Formulation: HA-R with Reflection. To create a more structured and powerful reasoning process, we introduced an explicit self-correction step, inspired by the concept of reflective thinking in models like DeepSeek-R1 (Guo et al. 2025). This led to our final HA-R framework, which follows a four-stage sequence: $A_i \rightarrow R_1 \rightarrow R_2 \rightarrow A_f$. (1) **Impression ($A_i$):** The model generates a fast, heuristic answer. (2) **Reasons (R_1):** It provides an explanation for its initial impression. (3) **Reflection (R_2):** The model critically re-evaluates its own rationale (R_1), explicitly modeling a self-correction mechanism. (4) **Answer (A_f):** Based on all preceding steps, the model outputs its final, more robust analytic answer.

This process is modeled as $p(A_i, R_1, R_2, A_f \mid I) = p(A_i \mid I) \cdot p(R_1 \mid I, A_i) \cdot p(R_2 \mid I, A_i, R_1) \cdot p(A_f \mid I, A_i, R_1, R_2)$. By conditioning the final answer A_f on a dedicated reflection step, we enable the model to achieve higher accuracy and robustness. The HA-R structure inherently provides two verdicts, a Fast Answer (A_i) and a Robust Answer (A_f), laying the groundwork for an adaptive inference strategy.

Progressive Training for Adaptive Reasoning

We instill the HA-R capability into a VLM via a progressive two-stage training paradigm, combining Supervised Fine-Tuning (SFT) and Reinforcement Learning with Verifiable Reward (RLVR) (Guo et al. 2025).

Stage 1: Cold-Start via Supervised Fine-Tuning. We first "cold-start" a pre-trained VLM using parameter-efficient SFT with LoRA (Hu et al. 2022). The objective is to teach the model the HA-R format by maximizing the log-likelihood of the ground-truth sequence:

$$\mathcal{L}_{SFT} = -\log p(A_i, R_1, R_2, A_f \mid I). \quad (1)$$

However, while SFT effectively teaches the desired output structure, it primarily encourages mimicry of annotated reasoning rather than fostering genuine, "free-form" analytical skills, which can damage model generalization.

Stage 2: RLVR for Adaptive and Free-Form Thinking. We use GRPO (Shao et al. 2024) to perform RLVR to encourage the model to generate more diverse and higher-quality reasoning. RLVR's key advantage is its ability to use programmatic, rule-based rewards without requiring a separate value model or human preference data. This allows us to guide the model towards our desired reasoning principles while empowering it to "think freely" beyond the SFT data.

Traditional accuracy reward in RLVR gives an reward of 1 when the answer is correct other wise 0, which encourages a form of brittle overconfidence, where the model is incentivized to make a high-risk guess rather than honestly assessing its own uncertainty. For our reflective reasoning, the model must learn to produce a well-calibrated confidence score in its initial assessment; otherwise, it has no reliable internal signal to trigger deeper, more costly analytic reasoning when it is truly uncertain. We use a "soft" reward function for A_i as follows:

$$\mathcal{R}_{\text{conf}} = 1 - cos(\frac{\pi}{2} p(A_i = c|I)), \quad (2)$$

where c is the ground truth label of the sample. The final reward is a weighted sum of the GRPO component, the confidence reward R_{conf}, and an accuracy reward $\mathcal{R}_{acc}$ for the final answer A_f, a length reward $\mathcal{R}_{len}$ for longer response, a repetition reward $\mathcal{R}_{rep}$ penalizing the repetition, along with a format reward $\mathcal{R}_{fmt}$ for correct formatting. To sum, our RLVR reward is as follows:

$$\mathcal{R}_{GRPO} = \mathcal{R}_{acc} + \mathcal{R}_{conf} + \mathcal{R}_{len} + \mathcal{R}_{rep} + \mathcal{R}_{fmt}. \quad (3)$$

Method	Venues	ID			OOD-C			T2I	IP&OP	IE	FS	CB	VTO	RMG	PCMG	*Mean*
		ACC.	P.	R.	ACC.	P.	R.	ACC.	ACC.	ACC.	ACC.	ACC.	ACC.	ACC.	ACC.	ACC.
Qwen-VL-Max	preprint 2025	66.21	61.35	95.88	75.35	71.08	86.51	76.71	58.17	62.33	64.61	60.64	66.08	74.12	67.01	67.12
QvQ-Max	preprint 2025	47.66	51.02	2.86	53.21	94.23	8.06	24.40	39.04	46.34	73.00	68.15	60.10	21.87	52.33	48.61
Gemini2.5-Pro-0617	preprint 2025	71.45	90.67	50.74	**82.30**	98.29	66.01	66.48	50.57	56.93	77.51	**74.60**	63.88	61.89	66.95	67.26
CNNSpot	CVPR 2020	77.57	94.85	60.61	57.29	89.92	17.60	69.08	42.14	57.64	71.89	66.33	60.29	35.80	52.76	59.08
UnivFD	CVPR 2023	68.26	74.00	61.07	55.45	75.89	17.60	72.31	52.37	**66.31**	73.86	70.53	65.20	63.07	64.16	65.15
NPR	CVPR 2024	73.21	91.86	53.80	50.29	60.00	5.43	79.81	36.61	39.96	71.12	64.00	55.69	27.18	51.93	54.63
AIDE	ICLR 2025	89.98	94.29	86.15	62.95	88.26	64.34	84.66	**44.22**	56.72	73.24	67.80	67.80	59.90	44.47	59.41
Effort	ICML 2025	91.23	97.05	85.93	57.04	96.00	15.79	86.40	43.71	45.42	**77.68**	71.10	63.92	44.40	59.67	64.06
Qwen-VL-2.5-3B	preprint 2025	96.44	98.02	95.14	65.92	96.68	33.50	68.24	39.80	45.26	75.31	69.98	62.51	56.24	63.43	64.31
Qwen-VL-2.5-7B	preprint 2025	96.67	97.59	96.04	69.07	95.72	40.39	86.15	40.05	45.26	74.45	68.85	62.22	52.47	59.66	65.48
LLaMA3.2-Vision-11B	preprint 2024	96.08	98.69	93.78	67.66	98.23	36.45	82.79	40.05	49.08	74.33	68.28	59.69	54.57	64.81	65.73
MIRAGE-R1	-	**97.26**	96.45	98.41	78.46	96.30	59.87	**88.82**	43.96	55.63	77.10	71.12	**67.94**	**75.79**	**72.03**	**72.81**

Table 2: The comparision of model performance on our proposed **MIRAGE** benchmark cross various distributions: In-distribution test set (*ID*), human-curated OOD set (*OOD-C*), and 8 subsets from different image generation types. We report the **MIRAGE-R1** performance with inference-time adaptive thinking in the table. All the models below "Gemini2.5-Pro-0617" were trained/finetuned on our **MIRAGE**.

Inference-Time Adaptive Thinking

We perform adaptive thinking at inference time. Given A_i, we quantify its confidence by its probability entropy:

$$H(A_i) = -\sum_{c\in\{real,fake\}} p(A_i = c \mid I)\log_2 p(A_i = c \mid I). \tag{4}$$

Here, $H(A_i)$ values between 0 and 1, and the high value represents a high uncertainty, i.e., a low confidence. Using a empirical threshold τ, the model dynamically chooses its reasoning path. If confidence is high ($h(A_i) \leq \tau$), it terminates early and returns the fast answer A_i. If confidence is low (($h(A_i) > \tau$)), it proceeds with the full HA-R process to generate the more robust answer A_f. This mechanism allows **MIRAGE-R1** to efficiently allocate computational resources, performing deep reasoning only when the task is sufficiently challenging.

Experiments

Experiment Settings

Method	Training Set	Chameleon		
		ALL	FAKE	REAL
CNNSpot*	GenImage	68.89	9.86	99.25
UnivFD*	GenImage	60.42	**85.52**	41.56
NPR*	GenImage	57.81	1.68	**100.00**
AIDE*	GenImage	65.77	26.80	95.06
CNNSpot	**MIRAGE**	64.21	38.08	83.85
UnivFD	**MIRAGE**	50.29	63.19	40.60
NPR	**MIRAGE**	62.95	15.49	98.62
AIDE	**MIRAGE**	60.69	45.32	72.25
Effort	**MIRAGE**	64.36	17.07	99.91
MIRAGE-R1	**MIRAGE**	**78.31**	50.29	98.37

Table 3: Model performance on the Chameleon benchmark. The numbers in the table refer to the accuracy on different subsets of the benchmark. The result of methods with * is taken from (Yan et al. 2024a).

Baseline. We evaluated 5 specialized state-of-the-art (SOTA) AIGI detection models, including CNN-Spot (Wang et al. 2020), UnivFD (Ojha, Li, and Lee 2023), NPR (Tan et al. 2024), AIDE (Yan et al. 2024a), and Effort (Yan et al. 2024b). Besides, we selected 3 state-of-the-art (SOTA) closed-source models for zero-shot evaluation: Qwen-VL-Max (Bai et al. 2023), QvQ-Max (The Qwen Team 2025), and Gemini2.5-Pro (Comanici et al. 2025). We also fine-tuned 3 open-source models, Qwen-VL-2.5-3B (Bai et al. 2025), Qwen-VL-2.5-7B (Bai et al. 2025), and LLaMA3.2-Vision-11B (Grattafiori et al. 2024), as foundational models to evaluate the effectiveness of our approach.

Method	ID			OOD-C		
	ACC.	P.	R.	ACC.	P.	R.
w/o Reasoning	96.67	**97.59**	96.04	69.07	95.72	40.39
H-R	96.12	95.53	97.16	72.36	94.23	48.36
A-R	93.90	93.39	95.10	73.13	89.29	53.90
G-R-A	96.54	96.49	96.93	75.25	**96.99**	52.87
HA-R-A_i	97.26	97.18	97.62	76.63	96.33	56.09
HA-R-A_f	97.14	96.34	98.30	78.13	96.01	59.38
HA-R	**97.26**	96.45	**98.41**	**78.46**	96.30	**59.87**

Table 4: Ablations on the reasoning format. The model with the best performance is in **bold**. Here, HA-R-A_i, HA-R-A_f, HA-R refer to our model prediction from the "impression", "final answer", and inference-time adaptive thinking.

Dataset. To comprehensively evaluate the capability of existing methods in ID and OOD scenarios, we conducted experiments according to the setting in the dataset section. Additionally, we used a well-known public benchmark for real-world AIGI detection, Chameleon (Yan et al. 2024a), for cross-benchmark evaluation.

Implementation Details. We selected hyperparameters on a sit-alone development set sampled from the training set. We employed a pre-trained Qwen2.5-VL-7B (Bai et al. 2025) as the backbone of **MIRAGE-R1**. The LoRA rank was set to 256 in the SFT stage for 4000 steps on 64 NVIDIA H20 GPUs, and to 8 in the RLVR stage for 50 steps on 32 NVIDIA H20 GPUs. In both stages, the batch size was set to 8. We empirically selected $\tau = 0.96$, where 10% samples

go to "deep thinking" in our development set. In the cold-start stage, we constructed the training set to be with 70% A_i the wrong answer and 30% the correct answer, so as to teach the model both to correct or enhance A_i in A_f. Other implementation details and hyperparameters are included in the appendix.

Metrics. Following previous works (Yan et al. 2024a,b; Wang et al. 2020), we report the classification accuracy (ACC.), precision (P.), and recall (R.) in our experiments, where ACC. is our primary evaluation metric.

Results

On benchmark MIRAGE. We evaluated **MIRAGE-R1** against SOTA specialized detectors and VLMs on our **MIRAGE** benchmark, with results summarized in Tab. 2. The findings underscore the effectiveness of our approach. Overall, **MIRAGE-R1** not only achieves the highest accuracy on the ID test set but also establishes a new SOTA on the OOD splits, surpassing the next-best competitor, Gemini-2.5 Pro, by a significant margin of over 5%. This robust generalization extends to a fine-grained level: when analyzed across the eight distinct generative attack patterns, **MIRAGE-R1** secures the top rank in four categories and places in the top three for seven out of eight, proving its consistency against a wide spectrum of sophisticated manipulation techniques. Notably, its performance on the T2I subset, where it leads by 2%, specifically evaluates generalization to unseen vanilla generative models — a common setting in prior work. The comprehensive generalization across diverse in-the-wild scenarios validates the advantages of our adaptive and reflective reasoning framework.

On benchmark Chameleon. We conducted a rigorous cross-benchmark evaluation on the public Chameleon dataset (Yan et al. 2024a) to further assess the OOD generalization of **MIRAGE-R1**. The results, presented in Tab. 3, demonstrate the generalization of our approach. Our model surpasses all other methods by a significant margin, establishing a new state-of-the-art on this challenging benchmark. This is a critical finding, as Chameleon's images are sourced from online AIGI communities and are known to be difficult even for human experts to identify. Remarkably, without any exposure to Chameleon's data during training, **MIRAGE-R1** achieves a balanced performance, correctly identifying half of the sophisticated forgeries while maintaining a near-perfect accuracy (98.37%) on real images. This showcases its ability to learn generalizable forgery cues rather than overfitting to artifacts from a specific training distribution. Moreover, this experiment highlights an equally important contribution of our work: the **MIRAGE** dataset itself. We observed that when baseline models are trained on **MIRAGE** training set instead of a conventional dataset like GenImage, their accuracy on Chameleon's FAKE subset consistently improves. This provides strong empirical evidence that our **MIRAGE** benchmark, with its diverse and challenging pipeline-generated examples, better prepares models for the complexities of in-the-wild AIGI detection.

Ablation studies. We conducted a comprehensive ablation study, with results presented in Tab. 4. The findings confirm our hypotheses. First, any form of reasoning (H-R, A-R) significantly enhances OOD generalization compared to a non-reasoning baseline. Second, introducing an explicit reflection step in our HA-R-A_f model yields a substantial OOD accuracy gain over our initial G-R-A design (78.13% vs. 75.25%), demonstrating the critical role of self-correction. Most importantly, our final adaptive HA-R model achieves the best overall performance (78.46% OOD Acc.), even surpassing the strategy of always using the final answer. This indicates that by dynamically choosing when to engage in deep, reflective thinking, our model strikes an optimal balance between accuracy and efficiency, preventing "overthinking" and "wrong impression".

Method	Original	JPEG Compression				Gaussian Blur	
		QF=50	QF=70	QF=90	QF=95	σ =1.0	σ =2.0
CNNSpot	77.57	77.13	77.85	78.45	77.97	75.10	73.19
UnivFD	68.26	69.07	68.00	67.94	68.06	69.13	67.94
NPR	73.21	58.69	66.33	71.52	71.82	71.88	62.39
AIDE	89.98	87.04	85.97	89.13	89.97	89.85	87.94
Effort	91.23	84.54	88.30	88.60	88.78	80.12	62.57
MIRAGE-R1	**97.26**	**95.43**	**96.14**	**96.08**	**95.19**	**91.33**	**88.30**

Table 5: Robustness of Classification Accuracy on different levels of JPEG compression and Gaussian blurring attack. The accuracy is calculated over the ID subset of the **MIRAGE** Benchmark.

Robustness evaluation. The robustness of an AIGI detector against real-world corruptions is critical for its practical application, as perturbations like compression and noise can erase forgery artifacts. We therefore evaluated **MIRAGE-R1** and baselines on our ID test set subjected to standard JPEG compression and Gaussian noise, following (Zhou et al. 2025). The results in Tab. 5 highlight our model's robustness. Under JPEG compression, **MIRAGE-R1** maintains consistently high accuracy while baseline performance drops significantly. It also outperforms competitors against Gaussian noise. Crucially, **MIRAGE-R1** achieves this robustness without being trained on corresponding data augmentations (e.g., Gaussian noise), which were commonly used to develop the baseline models.

Conclusion

In this work, we take a step towards addressing the challenge of in-the-wild AIGI detection by introducing **MIRAGE**, a benchmark designed to better reflect real-world complexity, and proposing **MIRAGE-R1**, a VLM with a novel adaptive reasoning framework. Our experiments show that **MIRAGE-R1** achieves strong performance across our benchmark and public datasets, with notable improvements in both generalization and robustness. We believe this work offers a useful step toward developing more practical and reliable AIGI detection systems for our increasingly complex digital world.

Acknowledgements

This work is supported by Taobao&Tmall Group of Alibaba through Alibaba Research Intern Program and Alibaba Group Innovative Research Program.

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

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*.

Cai, Q.; Chen, J.; Chen, Y.; Li, Y.; Long, F.; Pan, Y.; Qiu, Z.; Zhang, Y.; Gao, F.; Xu, P.; et al. 2025. HiDream-I1: A High-Efficient Image Generative Foundation Model with Sparse Diffusion Transformer. *arXiv preprint arXiv:2505.22705*.

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3339–3348.

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*.

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*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; and Yang, A. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

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*.

Hong, Y.; and Zhang, J. 2024. Wildfake: A large-scale challenging dataset for ai-generated images detection. *arXiv preprint arXiv:2402.11843*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, Z.; Li, T.; Li, X.; Wen, H.; He, Y.; Zhang, J.; Fei, H.; Yang, X.; Huang, X.; Peng, B.; et al. 2025. So-Fake: Benchmarking and Explaining Social Media Image Forgery Detection. *arXiv preprint arXiv:2505.18660*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Li, Y.; Liu, X.; Wang, X.; Lee, B. S.; Wang, S.; Rocha, A.; and Lin, W. 2024. Fakebench: Probing explainable fake image detection via large multimodal models. *arXiv preprint arXiv:2404.13306*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 38–55. Springer.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24480–24489.

Ren, J.; Xu, H.; He, P.; Cui, Y.; Zeng, S.; Zhang, J.; Wen, H.; Ding, J.; Huang, P.; Lyu, L.; et al. 2024. Copyright protection in generative ai: A technical perspective. *arXiv preprint arXiv:2402.02333*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402.

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*.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28130–28139.

The Qwen Team. 2025. QvQ-Max-Preview. https://qwenlm.github.io/blog/qvq-max-preview/. Blog Post.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8695–8704.

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*.

Xiang, Y.; Mottaghi, R.; and Savarese, S. 2014. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, 75–82. IEEE.

Xu, Z.; Zhang, X.; Li, R.; Tang, Z.; Huang, Q.; and Zhang, J. 2024. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761*.

Yan, S.; Li, O.; Cai, J.; Hao, Y.; Jiang, X.; Hu, Y.; and Xie, W. 2024a. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*.

Yan, Z.; Wang, J.; Jin, P.; Zhang, K.-Y.; Liu, C.; Chen, S.; Yao, T.; Ding, S.; Wu, B.; and Yuan, L. 2024b. Orthogonal Subspace Decomposition for Generalizable AI-Generated Image Detection. *arXiv preprint arXiv:2411.15633*.

Yin, R.; and Liu, X. 2024. Enabling media production with AIGC and its ethical considerations. In *2024 IEEE 10th International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, 100–105. IEEE.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, W.; Jiang, C.; Zhang, Z.; Si, C.; Yu, F.; and Peng, W. 2025. IVY-FAKE: A Unified Explainable Framework and Benchmark for Image and Video AIGC Detection. *arXiv preprint arXiv:2506.00979*.

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. *arXiv preprint arXiv:2507.02664*.

Zhu, M.; Chen, H.; Yan, Q.; Huang, X.; Lin, G.; Li, W.; Tu, Z.; Hu, H.; Hu, J.; and Wang, Y. 2023. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. arXiv:2306.08571.