

From Evidence to Verdict: An Agent-Based Forensic Framework for AI-Generated Image Detection

Mengfei Liang Yiting Qu Yukun Jiang Michael Backes Yang Zhang*

CISPA Helmholtz Center for Information Security

Abstract

The rapid evolution of AI-generated images poses unprecedented challenges to information integrity and media authenticity. Existing detection approaches suffer from fundamental limitations: traditional classifiers lack interpretability and fail to generalize across evolving generative models, while vision-language models (VLMs), despite their promise, remain constrained to single-shot analysis and pixel-level reasoning. To address these challenges, we introduce AIFo (Agent-based Image Forensics), a novel training-free framework that emulates human forensic investigation through multi-agent collaboration. Unlike conventional methods, our framework employs a set of forensic tools, including reverse image search, metadata extraction, pre-trained classifiers, and VLM analysis, coordinated by specialized LLM-based agents that collect, synthesize, and reason over cross-source evidence. When evidence is conflicting or insufficient, a structured multi-agent debate mechanism allows agents to exchange arguments and reach a reliable conclusion. Furthermore, we enhance the framework with a memory-augmented reasoning module that learns from historical cases to improve future detection accuracy. Our comprehensive evaluation spans 6,000 images across both controlled laboratory settings and challenging real-world scenarios, including images from modern generative platforms and diverse online sources. AIFo achieves 97.05% accuracy, substantially outperforming traditional classifiers and state-of-the-art VLMs. These results demonstrate that agent-based procedural reasoning offers a new paradigm for more robust, interpretable, and adaptable AI-generated image detection.

1 Introduction

In recent years, image generative models such as GLIDE [37], Imagen [45], DALL-E 2 [43], and Stable Diffusion [44], have advanced rapidly. They can synthesize photo-realistic images from natural language in seconds [23, 44, 45]. However, the realism of AI-generated images has raised serious societal concerns. Because AI-generated images can now easily fool human observers, malicious actors are increasingly leveraging them to spread disinformation and impersonate individuals [48]. For example, during the 2024

U.S. presidential election cycle, sophisticated deepfakes (i.e., AI-generated images or videos that convincingly fabricate real people or events) have appeared in campaign ads and on social media, which potentially manipulate public opinions and disrupt the voting behaviors [15]. Beyond elections, AI-generated images also introduce broader risks [48, 50], including misinformation and privacy infringement.

In response to these risks, substantial research has been devoted to the detection of AI-generated images. Current methodologies can be generally classified into two main categories: traditional machine learning classifiers and advanced approaches leveraging large vision language models (VLMs).

Traditional machine learning classifiers typically rely on training convolutional neural networks (CNNs) or transformer-based models to distinguish between real and fake images [17, 19, 53, 54, 59]. Early studies reveal that AI-generated images tend to exhibit shared low-level artifacts, allowing detectors trained on labeled images to identify them [46, 53]. For example, DE-FAKE [46] trains a set of classifiers on AI-generated and real images to learn AI-specific artifacts. However, these works often depend on a limited number of training datasets with fake images produced by only a few specific generative models, which makes them prone to *generalizability* issues: they perform well on seen data but struggle to generalize to unseen images from new generative models [51, 60]. They also often lack *explainability*: most models act as black boxes, producing binary outputs without offering human-interpretable justifications [33, 60].

More recently, **vision language models (VLMs)** have shown promise for more generalizable and explainable detection [28, 33, 57, 60]. Due to the large-scale pre-training, VLMs can be transferred to image detection tasks in a zero-shot or few-shot manner [51, 57, 60], without relying on specialized labeled datasets. Beyond identifying pixel-level artifacts like traditional classifiers, VLMs can also apply semantic and world knowledge in the detection process, e.g., a photo of a flying cat is AI-generated because a flying cat cannot exist in the physical world. In addition, they can provide human-interpretable justifications through prompt engineering for explainable detection [28, 51, 57].

Despite these advancements, both traditional and VLM-based approaches suffer from fundamental limitations com-

*Yang Zhang is the corresponding author.

pared to human forensic experts. First, they rely heavily on pixel-level image features as the main detection evidence. By contrast, human experts can not only employ off-the-shelf classifiers and VLMs for pixel-level analysis, but also actively seek evidence beyond the images themselves, e.g., Exchangeable Image File Format (EXIF) metadata that contains important information about the image or online contextual information to make a comprehensive decision. Second, both approaches treat image detection as a single-shot classification task, using a fixed model or pipeline. Human experts, on the other hand, approach image detection as a dynamic reasoning process: they flexibly employ different tools and iteratively refine their judgments based on the gathered evidence. Finally, human experts can improve over time through accumulated experience, enabling them to adapt to new generative models and real-world scenarios, whereas fixed models remain static unless frequently fine-tuned on new datasets. Motivated by this, we aim to explore a new paradigm that combines the advantages of existing classifiers and VLMs with the capabilities of human experts for AI-generated image detection.

Our Work. Our approach fundamentally differs from conventional methods based on static classifiers or single VLM. Rather than developing another standalone image classifier, we design a cognitive AI agent system that automates the entire human-like forensic workflow, simulating reasoning and decision-making in the detection of AI-generated images. We present AIFo (Agent-based Image Forensics), an LLM-based multi-agent framework that integrates functionalities such as visual recognition, semantic understanding, and provenance analysis. AIFo consists of a forensic Toolbox, an Evidence Gatherer, a Reasoning Agent, and a Debate module. Given a test image, the Evidence Gatherer first applies tools from the Toolbox and aggregates the resulting evidence. The Reasoning Agent then evaluates the quality of the collected evidence, i.e., whether it is sufficient and consistent enough to support a reliable judgment. If so, the Reasoning Agent produces a final decision and explains why the image is classified as AI-generated or real. Otherwise, we introduce a Debate module [21,34–36,49] to handle cases where the evidence is insufficient or conflicting. In this module, two Debate Agents exchange arguments over multiple rounds, each adopting an opposing stance (for or against the claim that the image is AI-generated), while a Judge Agent oversees the debate process and produces the final judgment. Additionally, we demonstrate the potential of our framework to continuously and progressively enhance its detection effectiveness by learning from historical testing data. By incorporating a memory module that stores all testing history, the framework behaves like a human expert, accumulating experience and improving performance over time. Overall, such a cognitive AI agent framework offers a new paradigm for AI-generated image detection.

Main Findings. We evaluate AIFo on a dataset of 6,000 AI-generated and real images, comprising 3,000 samples from five established benchmarks (e.g., Flickr30K [40], GenImage [61], and FakeBench [33]) and 3,000 in-the-wild images collected from six online platforms. AIFo is benchmarked

against a range of baselines methods, including traditional classifiers such as CNNSpot [53], DE-FAKE [46], and state-of-the-art VLMs (e.g., GPT-4.1 [8] and GPT-4o [39]). To ensure a fair comparison, we disable the memory module and prevent our framework from learning during the main evaluation. Even under these conditions, AIFo achieves the best overall performance, reaching 0.9705 accuracy and surpassing GPT-4o (0.9483), GPT-4.1 (0.9416), and other baselines. It also maintains higher robustness (0.9047–0.9690) than GPT-4o (0.8818–0.9462) under three types of perturbations. To demonstrate the effectiveness of the memory module, we conduct a case study on 50 images that are misclassified in the main evaluation. The results show that when the memory module stores similar historical cases, approximately 40% of these errors are successfully corrected.

The main contributions of our work are as follows:

- **Human-Like Procedural Reasoning.** We present the first agentic framework (AIFo) that simulates human-like procedural reasoning for AI-generated image detection. Our framework combines the strengths of conventional classifiers and VLMs, employing them as complementary tools, and incorporates human-like reasoning processes such as integrating multiple sources of evidence, debating conflicting evidence, and improving through accumulated experience. This highlights a transition from static classification toward a dynamic reasoning process in AI-generated image detection.
- **Comprehensive Evaluation.** We contribute a valuable benchmark dataset covering 6000 images from both existing benchmark datasets and in-the-wild images from internet platforms. This dataset serves as the foundation for rigorous evaluation, through which we demonstrate the superior performance, generalizability, and robustness of our AIFo compared to state-of-the-art baselines.
- **Training-Free Core and Cross-Model Generalizability.** Unlike conventional detectors that require large-scale training data and frequent updates, the core of AIFo is training-free and designed to generalize across evolving generative models. It leverages generalizable forensic evidence rather than model-specific features, enabling robust detection of AI-generated images. An optional memory module further enhances performance over time by accumulating historical cases.

2 Threat Model

2.1 Detector’s Goals

Distinguishing Between AI-Generated and Real Images. The detector aims to determine whether a given image is AI-generated or a real image. In this work, AI-generated images include those produced by generative models like text-to-image diffusion models, as well as AI-edited images, such as cartoons generated from real photographs and images replaced with AI-generated backgrounds. Real images refer to those captured by physical cameras or manually created/drawn by human artists. Notably, if a real image has

been edited or post-processed by humans (without the aid of AI models), e.g., by applying filters, adjusting color contrast, or increasing brightness, we still consider it a real image.

Enhanced Interpretability. Unlike conventional binary detectors, our detector aims not only to provide a decision but also to offer human-interpretable justifications. For example, if our detector identifies an image as an AI-generated image, it should also point out evidence that supports this decision, such as unreasonable object placement (semantic-level evidence), distorted lighting in the image (pixel-level evidence), editing history (metadata-level evidence), AI watermarks (source-level evidence), etc. Providing such explanations helps users understand, verify, and trust the detector’s decisions.

Cross-Model Generalizability. Given the rapid evolution of text-to-image generative models, new architectures and generated images are continually emerging, which may degrade the performance of detectors trained on previous generative models. Therefore, the detector should be robust to various model iterations and new architectures, and capable of leveraging generalizable evidence to assess image provenance, rather than relying solely on model-specific features.

Training-Free Design. Furthermore, traditional AI image detection frameworks typically depend on collecting large-scale training datasets of AI-generated images. As generative models evolve, these frameworks require continual updates to their training data, incurring significant maintenance costs. In contrast, the core of our detector is designed to be training-free: it does not require any dedicated training set, and can be directly applied to detect AI-generated images from any model, including those unseen during development.

2.2 Detector’s Capabilities

Our detector is assumed to have access to the following resources and conditions:

- **Image-File Access.** The detector can access the image file itself, including its metadata if available (e.g., camera details, timestamps, and editing history).
- **External Forensic Tools.** The detector is permitted to utilize a suite of external forensic tools to analyze the image and gather supporting evidence. This includes pretrained AI detection models, reverse image search engines, and vision-language models.

3 The AIFo Framework

3.1 Design Rationale

Unlike existing approaches that rely solely on pre-trained classifiers or VLMs, our main design principle is to emulate the procedural reasoning processes employed by human forensic experts [22, 38], while still leveraging the strengths of existing approaches. To this end, we leverage LLMs as autonomous agents to develop an agentic framework that incorporates the key capabilities of human forensic experts. First, the framework is designed to employ and coordinate

a suite of forensic tools, such as pre-trained classifiers, metadata extraction, reverse image search, and VLM-based analysis. Second, rather than relying on a single fixed source of evidence, the framework should be able to cross-check evidence from multiple tools before forming the final decision. Next, when the collected evidence indicates conflicting signals, the framework is expected to resolve these conflicts through human-like reasoning activities, such as assessing the reliability of evidence sources and engaging in structured debate among the expert team. The core framework is training-free, building on top of pre-trained classifiers and state-of-the-art VLMs as forensic tools. At the same time, an optional memory module allows the framework to accumulate historical cases and progressively improve performance. Altogether, we aim to build a training-free, interpretable, and robust framework for AI-generated image detection, which not only leverages existing detection approaches but also possesses key capabilities of human forensic experts, e.g., integrating various evidence, evaluating their reliability, and resolving conflicts among them.

3.2 Overview

Our proposed framework, Agent-based Image Forensics (AIFo), is a training-free, LLM-based multi-agent system for AI-generated image detection, designed to mimic the procedural reasoning workflow of human forensic experts. As illustrated in Figure 1, the core of our framework¹ consists of a **Toolbox** $\mathcal{T} = \{T_1, T_2, \dots, T_n\}$ containing n different forensic tools, an **Evidence Gatherer Agent** A_{EG} , a **Reasoning Agent** A_R , two **Debate Agent** A_{D1} and A_{D2} and a **Judge Agent** A_J . Given an input image of unknown authenticity from the set of all images, $I \in \mathcal{V}$, the Evidence Gatherer A_1 first employs tools from the Toolbox. Specifically, the Toolbox includes four main types of forensic tools: (1) **Image Reverse Search** to locate visually similar images and retrieve their provenance information from online sources; (2) **Metadata Extraction** to recover embedded EXIF or generated metadata from the image file; (3) **VLM-Based Analysis** to perform semantic-level reasoning and contextual understanding of the image leveraging VLMs; and (4) **Pre-Trained Classifiers** to apply pre-trained models for binary classification of the image.

Then, the Evidence Gatherer Agent executes each tool $T_i \in \mathcal{T}$ to produce a piece of evidence:

$$e_i = T_i(I), \quad T_i : \mathcal{V} \rightarrow \mathcal{E}_i,$$

where \mathcal{E}_i denotes the evidence space for tool T_i . For example, the metadata extraction tool may return detailed EXIF information embedded in the image file, such as "EXIF:Model": "Canon EOS 5D Mark IV" or "EXIF:LensModel": "EF24-70mm f/2.8L II USM", which can serve as important evidence for authenticity assessment. The complete evidence set is:

$$E = \{e_i \mid T_i \in \mathcal{T}\}.$$

¹We introduce the optional memory module in Section 5.

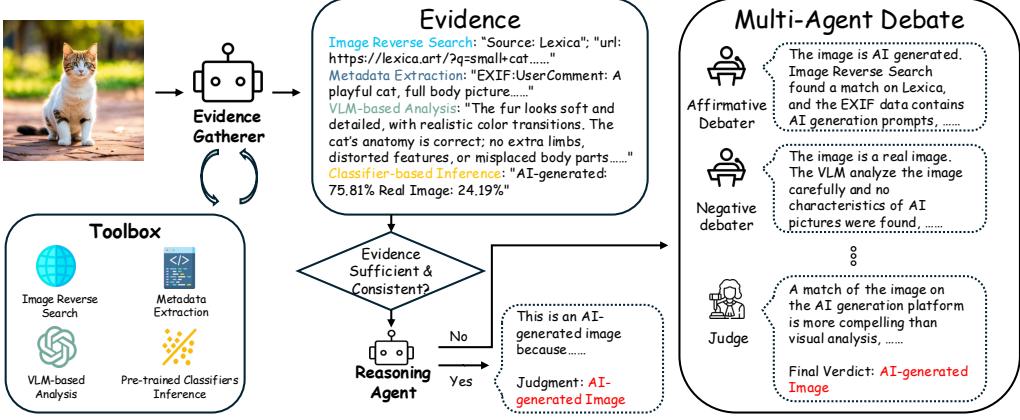


Figure 1: High-level overview of our proposed AIFo.

Next, the Reasoning Agent A_R examines the entire evidence set \mathcal{E} and determines whether the evidence is *sufficient* and *consistent* enough to make a reliable decision. Evidence sufficiency refers to whether the collected evidence covers enough aspects of the input image, e.g., without missing key metadata or reverse search results. Evidence consistency measures the extent to which different pieces of evidence point toward the same conclusion, e.g., when the majority of evidence supports that the image is AI-generated. If the evidence is considered sufficient and consistent enough, the Reasoning Agent generates a final judgment $D \in \mathcal{D}$ and a human-readable explanation $R \in \mathcal{R}$:

$$A_R : \mathcal{E} \rightarrow \mathcal{D} \times \mathcal{R}.$$

Otherwise, the system initiates a multi-round debate process. Two Debate Agents A_{D1} and A_{D2} take opposing stances (pro vs. contra) regarding whether the image is AI-generated or real. They exchange arguments over n rounds, and a Judge Agent A_J observes the debate history H together with the tool-derived evidence \mathcal{E} , and produces the final judgment $D \in \mathcal{D}$ along with an explanation $R \in \mathcal{R}$:

$$A_J : (H, \mathcal{E}) \rightarrow \mathcal{D} \times \mathcal{R}.$$

This hierarchical reasoning-debate framework ensures that decisions are made either directly from sufficient evidence or through a careful debating process that simulates human forensic reasoning when the evidence is inconclusive or conflicting.

3.3 LLM-Based Agents

The LLM-based agents act as the central reasoning and orchestration units of our multi-agent framework, responsible for applying forensic tools, evaluating the reliability of collected evidence, and making decisions based on existing evidence. All agents are instantiated from a LLM (GPT-4.1-2025-04-14), with specialized prompts defining their respective roles and responsibilities. The detailed prompts used for each agent are provided in [Appendix A](#).

The **Evidence Gatherer** has full access to the Toolbox and is responsible for invoking the forensic tools to collect evidence. Its decision-making process is guided by a carefully

engineered prompt that contains three critical components: (1) a clear definition of the agent’s task, which is acting as an AI image forensics expert to determine whether the input image is AI-generated or real; (2) rigorous definitions of “AI-generated image” and “real image” (see [Section 2.1](#)) to ensure consistent interpretation during analysis; (3) a comprehensive description of all available tools, including their capabilities. With these explicit task definitions and resource descriptions, the LLM gains both the contextual understanding and the operational awareness to invoke the tools and collect results.

The **Reasoning Agent** is responsible for conducting an initial assessment of the evidence collected by the evidence gatherer. It first evaluates whether the evidence set is sufficient and consistent enough. This evidence set check is implemented via a dedicated prompt, which instructs the LLM to internally assess the completeness (sufficiency) and consistency of the evidence and return a binary decision (*true / false*). If true, the reasoning agent proceeds to generate a final judgment and an explanation with another prompt that: (1) defines its responsibility to synthesize the collected evidence and produce a final binary judgment (AI-generated or real) with an accompanying explanation; (2) instructs the agent to evaluate the reliability of each evidence source, taking into account the confidence level, potential biases, and the trustworthiness of the generating tool; (3) emphasizes the importance of generating a human-readable, logically structured reasoning chain. If the evidence is insufficient or conflicting, the reasoning agent abstains from producing a final decision and instead triggers the multi-agent debate mechanism.

The multi-agent debate mechanism is designed to resolve ambiguity and enhance the robustness of the final judgment when the evidence is inconclusive or conflicting.

The **Debate Agents** are responsible for engaging in a structured multi-round debate when the evidence is deemed insufficient or inconsistent. The debate process lasts up to n rounds. In the first round, the two debate agents are guided by designed prompts to take opposing stances: one agent provides arguments supporting that the image is AI-generated based on collected evidence, while the other provides arguments supporting that the image is real. In subsequent

rounds, both agents are instructed to refine or strengthen their reasoning based on the arguments made by the opposing side in the previous round. Through this iterative exchange, the debate agents progressively sharpen their analyses and solve potential conflicts in the evidence interpretation.

A **Judge Agent** is responsible for supervising the debate process. At the end of each round, the Judge Agent internally evaluates whether the debate has reached sufficient coverage and clarity to make a final decision. It may decide to terminate the debate early if the arguments are deemed sufficient, or allow the debate to proceed for additional rounds if further clarification is needed. Once the debate concludes, the Judge Agent generates the final judgment and an explanation. Unlike the Reasoning Agent, which relies solely on the collected evidence, the Judge Agent is guided to synthesize both the tool-derived evidence and the debate history. This design ensures that the final decision integrates multiple sources of evidence with structured deliberation, leading to a transparent and defensible forensic conclusion.

3.4 Forensic Tools

The forensic Toolbox is a collection of specialized modules that the Evidence Gatherer can invoke to analyze the input image. The tools can be broadly categorized into the following four classes:

- **Reverse Image Search Tools:** Assess image provenance by querying external online sources.
- **Metadata Extraction Tool:** Parse metadata information from the image file.
- **Pre-Trained Classifiers Tool:** Apply static models to assess authenticity.
- **VLM-Based Reasoning Tool:** Perform semantic-level analysis using VLMs.

First, **reverse image search tools** are designed to identify the provenance and distribution history of input images by searching for exact matches or visually similar content across the Internet. These tools are particularly useful for determining image provenance, e.g., AI-generated images may appear on generative art platforms, while real images are more likely to be found on reputable news or photographic sites. We implement two complementary reverse image search tools. The first tool leverages the Google Cloud Vision API’s Web Detection service [7], which provides programmatic access to Google’s extensive image indexing capabilities. It primarily returns exact matches of the query image by analyzing its visual features and comparing them against Google’s web-scale image database. The API outputs structured data such as webpage titles, URLs, and contextual information where identical or highly similar images are found. However, the returned information can sometimes be sparse or insufficient for reliable judgment. To address this limitation and broaden the search scope, we implement a second approach inspired by Xu et al. [55] that simulates authentic human user behavior through web interface automation. This tool directly interacts with the Google Images search interface using Play-

wright², a modern web automation framework, to perform searches beyond the API’s restricted results. Unlike the first tool, this approach captures not only exact matches but also a wider range of visually similar images, thereby providing richer contextual evidence. It automatically uploads the input image, captures the results section from the returned webpage while filtering out advertisements and noise, and summarizes key information such as image match identification and the provenance of similar images.

Metadata analysis tool focuses on extracting and analyzing technical metadata embedded within image files to identify authenticity markers that distinguish real photographs from AI-generated content. Digital cameras and imaging devices typically embed rich metadata (EXIF data) including camera settings, GPS coordinates, timestamps, and device-specific information that are often absent or inconsistent in synthetically generated images. For example, as shown in Table 1, a genuine photograph may contain entries like EXIF:Make and EXIF:Model indicating the camera manufacturer and model. They may also include realistic optical parameters or position data such as EXIF:LensInfo and Composite:GPSPosition that are indicative of a real-world capture event. In contrast, AI-generated images frequently lack such detailed metadata or contain synthetic traces. For instance, some AI images explicitly embed generation information or even the original text prompt used for synthesis, e.g., EXIF UserComment. These anomalies serve as strong indicators of AI generation.

Our metadata extraction tool employs ExifTool³, a comprehensive metadata manipulation library, to perform deep analysis of image files. Rather than extracting all available metadata fields, which can be overwhelming and include irrelevant information, our system implements a selective extraction strategy based on two key filtering mechanisms. First, we maintain a curated list of exact-match key fields KEY_FIELD_EXACT that have been empirically determined to provide strong authenticity signals. These include critical camera parameters such as focal length, aperture settings, ISO values, and camera manufacturer information. Second, we implement prefix-based filtering KEY_FIELD_PREFIXES to capture related metadata families. For example, all fields for camera manufacturer custom information are captured through the “MakerNotes” prefix. This dual-filtering approach ensures comprehensive coverage of relevant metadata while filtering out noise. The complete list of all key fields and prefixes is provided in the Appendix B.

Pre-trained classifiers tool leverages pre-trained models specifically fine-tuned for AI-generated image classification. These models have been trained on diverse datasets containing both authentic and AI-generated images, enabling them to learn discriminative features that distinguish between the two categories. We select the top five most downloaded classification models for AI-generated image detection available on Hugging Face [1]:

- haywoodsloan/ai-image-detector-deploy [3]

²<https://playwright.dev/>

³<https://exiftool.org/>

Table 1: Comparison of metadata signals for distinguishing real and AI-generated images.

Metadata Field	Example Value
Real Image Signal	
EXIF:Make	Canon
EXIF:LensInfo	4.1 123 3.5 6.4
Composite:GPSPosition	28.35 N, 81.59 W
AI Image Signal	
JUMBF:Description	AI Generated Image
EXIF:UserComment	"A full body of a cat..."

- Organika/sdxl-detector [12]
- legekka/AI-Anime-Image-Detector-ViT [4]
- Smogy/SMOGY-Ai-images-detector [13]
- NYUAD-ComNets/NYUAD_AI-generated_images_detector [11]

Our implementation employs an ensemble approach that integrates multiple transformer models to enhance detection robustness and accuracy. The ensemble includes models trained on different synthetic image generation techniques, ensuring comprehensive coverage of various AI generation paradigms including GANs, diffusion models, and other neural synthesis methods. Each transformer model performs independent classification, producing probability distributions over the binary classification space. The final prediction score is obtained through a weighted voting mechanism, where each model’s contribution is scaled by its weight parameter. This weighted ensemble score is computed as:

$$\text{Prediction Score} = \frac{\sum_{i=1}^N w_i \cdot s_i}{\sum_{i=1}^N w_i},$$

where w_i represents the weight of the i -th model θ_i , s_i is the AI confidence score from θ_i , and N is the total number of loaded models. Each model is assigned an equal weight $w_i = 1$ to ensure balanced contribution and avoid over-reliance on any individual model during inference.

VLM-based reasoning tools employs VLMs to perform sophisticated visual analysis. These models leverage their extensive training on diverse image-text pairs to identify subtle visual patterns, artifacts, and inconsistencies that may indicate AI generation. Our implementation utilizes GPT-4.1 [8] to conduct detailed image analysis. A crafted prompt guides the model to focus on specific visual characteristics that are indicative of AI generation versus authentic photography. The complete prompt is provided in Table 15 in the Appendix. When the model identifies an image as AI-generated, it is prompted to provide detailed evidence in the form of specific visual artifacts. These include unnatural textures or patterns that deviate from expected material properties, inconsistent lighting or shadow directions that violate physical principles, anatomical errors in human or animal subjects, unusual distortions in object boundaries, text rendering abnormalities, symmetry issues, and contextual inconsistencies in background elements. For images classified

as real, the model is instructed to explain the supporting visual characteristics, such as realistic anatomical proportions and coherent environmental context. Unlike the other forensic tools, VLMs not only enable binary classification but also provide detailed justifications and fine-grained visual analyses that enhance the interpretability of the final decision. Additionally, the model provides a confidence assessment (high, medium, or low) based on the strength and clarity of the observed evidence. The deterministic configuration (temperature = 0, seed = 42) ensures consistent and reproducible analysis results across multiple runs.

4 Evaluation

4.1 Dataset Construction

To comprehensively evaluate our agentic framework, we consider both controlled and real-world scenarios. We therefore construct our benchmark dataset to cover two distinct settings: *in-the-lab* and *in-the-wild*. The *in-the-lab* setting consists of images collected from well-curated, controlled datasets commonly used in prior research, while the *in-the-wild* setting comprises images sourced from diverse, unconstrained online platforms, reflecting the complexity and unpredictability of real-world data.

Our curated dataset comprises a total of 6000 images, evenly distributed across the two settings, with each containing 1500 AI-generated images and 1500 real images. For the *in-the-lab* setting, real images are sampled from well-established, curated datasets frequently used in image generation and detection research. Specifically, we sample 500 images each from Flickr30k [40], ImageNet [20], and the DIV2K dataset [16]. The AI-generated counterparts are obtained from GenImage [61] and FakeBench [33]. We sample 100 images from each of the 8 generative models included in GenImage, and 70 images from each of the 10 models in FakeBench, resulting in a total of 1500 AI-generated images.

For the *in-the-wild* setting, real images are collected from a diverse set of publicly available online sources, including a subset of photographs sampled from Flickr [6] and Wikimedia Commons [14], as well as images sampled from the Holopix50k dataset [26]. In order to ensure diversity, we select ten keywords: *animal*, *building*, *food*, *indoor*, *landscape*, *person*, *plant*, *snow*, *sport*, *transportation* and *water*. For Flickr and Wikimedia Commons, we search using these keywords and randomly sampled images from the results. For Holopix50k [26], which lacks an explicit label for each image, we employ BLIP [31] model to perform semantic analysis to categorize images according to the same set of keywords, followed by random sampling within each category. For each keyword and each source, we randomly sampled 50 images, resulting in a balanced and diverse collection of real-world content. The corresponding AI-generated images are also sourced from three online generative art platforms: Lexica [9], NightCafe [10], and Civitai [5]. Images from Lexica are primarily generated using the Lexica Aperture series models, while NightCafe and Civitai include images produced by a wide range of text-to-image models such as DALL·E [2, 43], Stable Diffusion [44], SDXL [41], and

Table 2: Dataset construction statistics.

Source Datasets / Platforms	Type	# Images
<i>In-the-lab Setting</i>		
Flickr30k [40]	Real Image	500
ImageNet [20]	Real Image	500
DIV2K [16]	Real Image	500
GenImage [61]	AI Image	800
FakeBench [33]	AI Image	700
<i>In-the-wild Setting</i>		
Holopix50k [26]	Real Image	500
Flickr [6]	Real Image	500
Wikimedia Commons [14]	Real Image	500
Lexica [9]	AI Image	500
NightCafe [10]	AI Image	500
Civitai [5]	AI Image	500

numerous community finetuned variants. This ensures that our AI-generated image collection reflects the variety and complexity of generative models encountered in real-world scenarios. We similarly use the same set of ten predefined keywords to search and randomly sample images from the three AI image platforms, collecting 50 images per keyword from each platform. Appendix D provides a comprehensive overview of the AI models used for generating images in our benchmark’s AI-sourced datasets, which shows our dataset encompasses over 20 generative models, such as the StyleGAN [30] series, the Stable Diffusion [44] series, and the DALL·E [2] family. Table 2 summarizes the detailed statistics of our dataset construction across both settings. This comprehensive benchmark enables rigorous evaluation of our agent framework under both controlled and real-world conditions, ensuring that our results reflect practical deployment scenarios.

4.2 Experimental Setup

Implementation Details. To evaluate the performance of our AIFo framework, we conduct experiments on the constructed benchmark dataset, comparing our agentic approach against a range of baseline methods. Our framework is implemented using LangGraph,⁴ which is specifically designed for building stateful, multi-agent applications with LLMs. We select GPT-4.1 (gpt-4.1-2025-04-14 version) as the backbone for our LLM agents, and we set the temperature to 0 and the seed to 42 to ensure deterministic behavior across runs.

Evaluation Metrics. We formulate the evaluation as a standard binary classification task, where the goal is to assess the framework’s ability to distinguish between AI-generated and real images. Specifically, we define correctly identifying an AI-generated image as a True Positive (TP) and correctly identifying a real image as a True Negative (TN). Based on these definitions, we compute the following standard evaluation metrics: **Accuracy**, **Precision**, **Recall**, and **F1-score**. In addition to evaluating the framework’s overall performance on the entire dataset, we further analyze its behavior separately on the *in-the-lab* and *in-the-wild* subsets to assess its robustness across controlled and real-world scenarios.

⁴<https://langchain-ai.github.io/langgraph/>

Baseline Methods. To comprehensively assess the effectiveness of our proposed multi-agent framework, we compare it against a diverse set of representative baseline methods spanning both conventional and vision-language modeling paradigms. For traditional classifier-based approaches, we include CNNSpot [53], DE-FAKE [46], and PatchCraft [59], with their details discussed in Section 7.1. For VLM baselines, we adopt state-of-the-art multimodal models including GPT-4o and GPT-4.1, which have demonstrated advanced visual reasoning capabilities. During evaluation, each model is provided with the input image along with the prompt *Is this a fake or real image?*, and instructed to return a structured binary classification result.

4.3 Results

Table 3 presents the comprehensive performance comparison of our multi-agent framework against five baseline methods across three evaluation settings. A detailed breakdown of their performance is provided in Table 17 in the Appendix. These results demonstrate the superior effectiveness of our proposed approach in AI-generated image detection under both controlled and real-world conditions.

Traditional Classifier Baselines. The traditional methods show varied performance patterns. CNNSpot [53] exhibits poor overall detection performance, close to random guessing. Its extremely low recall and F1-score indicate that it fails to effectively detect AI-generated images. DE-FAKE [46] achieves moderate performance, with an accuracy of 0.7142 and an F1 score of 0.7374, slightly outperforming PatchCraft [59]. These results highlight the inherent limitations of traditional classification methods in coping with rapidly evolving and continuously updated diverse AI-generated content in complex real-world scenarios.

Vision-Language Model Baselines. Both GPT-4.1 and GPT-4o demonstrate strong performance, with accuracy exceeding 0.94 and F1-scores above 0.93. However, the recall rates for both models are relatively lower compared to their precision, indicating that they still miss a significant portion of AI-generated images. This suggests challenges remain for these models in detecting certain types of AI-generated content. Our agent framework addresses this limitation through the multi-agent architecture, which combines the strengths of multiple specialized tools to achieve more reliable detection.

AIFo Framework Performance. Compared to all baseline methods, our AIFo framework consistently achieves the best performance across almost all metrics and evaluation settings. In the overall evaluation, AIFo attains an accuracy of 0.9705 and an F1-score of 0.9698, surpassing the strongest baseline GPT-4o by absolute margins of 2.22% in accuracy and 2.40% in F1-score. Moreover, this improvement is observed consistently in both *in-the-lab* and *in-the-wild* data, which span over 20 generative models, demonstrating the strong generalization capability of our framework. Through a training-free paradigm, our method is minimally affected by the evolution and iteration of generative models.

Impact of Multi-Agent Debate Mechanism. We also evaluate the impact of disabling the multi-agent debate mechanism on the performance. Notably, even without the debate

Table 3: Performance comparison of different methods on our benchmark dataset comprising three evaluation subsets: *Overall*, *In-the-Lab*, and *In-the-Wild*. Metrics reported are Accuracy (Acc), Precision (Prec), Recall (Rec), and F1-score (F1). Best results are highlighted in bold and second best results are underlined.

Method	Overall				In-the-Lab				In-the-Wild			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
CNNSpot [53]	0.5277	0.9826	0.0563	0.1066	0.5553	0.9882	0.1120	0.2012	0.5000	0.5000	0.0007	0.0013
PatchCraft [59]	0.6517	0.7423	0.4647	0.5715	0.8123	0.8704	0.7340	0.7964	0.4910	0.4780	0.1953	0.2773
DE-FAKE [46]	0.7142	0.6820	0.8027	0.7374	0.6720	0.6673	0.6860	0.6765	0.7563	0.6933	<u>0.9193</u>	0.7905
GPT-4.1 [8]	0.9416	0.9913	0.8910	0.9385	0.9332	<u>0.9932</u>	0.8723	0.9288	0.9500	0.9895	0.9097	0.9479
GPT-4o [39]	0.9483	0.9920	0.9038	0.9458	0.9537	0.9938	0.9130	0.9517	0.9428	0.9900	0.8947	0.9399
AIFo w/o debate (ours)	<u>0.9635</u>	0.9922	<u>0.9343</u>	0.9624	0.9730	0.9924	<u>0.9533</u>	<u>0.9725</u>	<u>0.9540</u>	<u>0.9921</u>	0.9153	0.9521
AIFo (ours)	0.9705	<u>0.9920</u>	0.9487	0.9698	0.9740	0.9917	0.9560	0.9735	0.9670	0.9923	0.9413	0.9661

mechanism, our AIFo framework still significantly outperforms GPT-4o. With the debate mechanism incorporated, our framework’s accuracy is further improved, particularly in *in-the-wild* setting, where it surpasses GPT-4o by 2.4%. These results demonstrate that the debate mechanism serves as an effective refinement stage, resolving potential reasoning uncertainties caused by conflicting evidence and enhancing the reliability of the final decision.

Laboratory vs. Wild Environment Analysis. A comparison between controlled (*in-the-lab*) and real-world (*in-the-wild*) environments reveals key differences. Traditional classifiers often fail to generalize: for example, PatchCraft’s F1-score drops from 0.7964 in the laboratory to 0.2773 in the wild, and CNNSpot nearly collapses entirely. This highlights the difficulty of transferring models trained on curated datasets to diverse real-world content. In contrast, our agent framework achieves consistently strong results, with F1-scores of 0.9735 in the laboratory and 0.9661 in the wild, across more than 20 generative models. These findings confirm the framework’s ability to generalize beyond controlled datasets and effectively handle the diversity of real-world AI-generated content.

Quantitative and Qualitative Analysis. To better understand where AIFo achieves improvements over GPT-4o, we analyze all samples misclassified by GPT-4o but correctly identified by AIFo. We find that AIFo correctly identifies 136 more AI-generated images compared to GPT-4o out of the total 6,000 samples. Among these corrected cases, approximately 38% are attributed to decisive reverse image search evidence, primarily from *in-the-wild* sources. Another 35% are resolved through metadata signals, which prove particularly effective for images containing detailed camera parameters or generation prompts embedded by community diffusion models. The remaining 27% rely on transformer-based classifier results as supporting evidence. In approximately 30% of all corrected cases, AIFo invokes the debate mechanism to resolve ambiguous or conflicting evidence, ultimately overturning the initial VLM-based assessment.

To further demonstrate the effectiveness of AIFo, we present qualitative examples in Figure 2 that show how the system integrates multiple sources of evidence and, when necessary, employs a debate mechanism to reach accurate conclusions. In the first real-image case, the Evidence Gatherer agent successfully collects evidence from four forensic tools, including image sources, camera parameters, VLM

analysis, and classifier scores. The Reasoning Agent judges the evidence sufficiency and, by synthesizing all sources, correctly concludes that the image is real.

In the second AI-generated case, the realistic appearance misleads the baseline GPT-4o and the VLM tool, but other tools indicated an AI origin, creating conflicting signals. In this case, the Reasoning Agent judges that the evidence is inconsistent for a direct decision and instead triggers the debate mechanism. The pro-debate agent and con-debate agent present arguments supporting their respective stances, and the decisive argument came from the pro side, which highlighted the exact match of the image found on an AI platform. Ultimately, the Judge Agent weighs the conflicting evidence and reaches the correct conclusion that the image is AI-generated.

These cases highlight two key strengths of our framework: its ability to override misleading visual evidence using stronger provenance signals, and its capacity to reconcile conflicting evidence through structured debate. These capabilities are essential for reliable detection in both benchmarks and real-world scenarios.

Tool Reliability and Decision Pattern Analysis. To gain deeper insights into the internal decision-making processes of our multi-agent framework, we conduct a quantitative analysis of the reliability and usage patterns of individual tools, as illustrated in Figure 3. We define a consistent judgment as a case in which the agent’s final decision is in agreement with the evidence provided by a given tool. The reliability rate of a tool is then computed as the ratio of its consistent judgments to the total number of valid (i.e., informative) pieces of evidence it produces. A higher reliability rate indicates that the agent frequently trusts and aligns with that tool’s output, suggesting its potential influence on the overall decision-making process. We also compute the coverage rate of each tool, defined as the proportion of cases in which the tool provides valid evidence relative to the total number of agent decisions.

From Figure 3a and Figure 3b, metadata extraction and the first reverse search tool focused on exact matches achieve the highest reliability rates, meaning their evidence plays a significant role in shaping the agent’s final decisions when available. However, their lower coverage rates show they are only relevant to a subset of images. This is because in our benchmark dataset, only a small fraction of images contain rich EXIF metadata or have provenance information retriev-

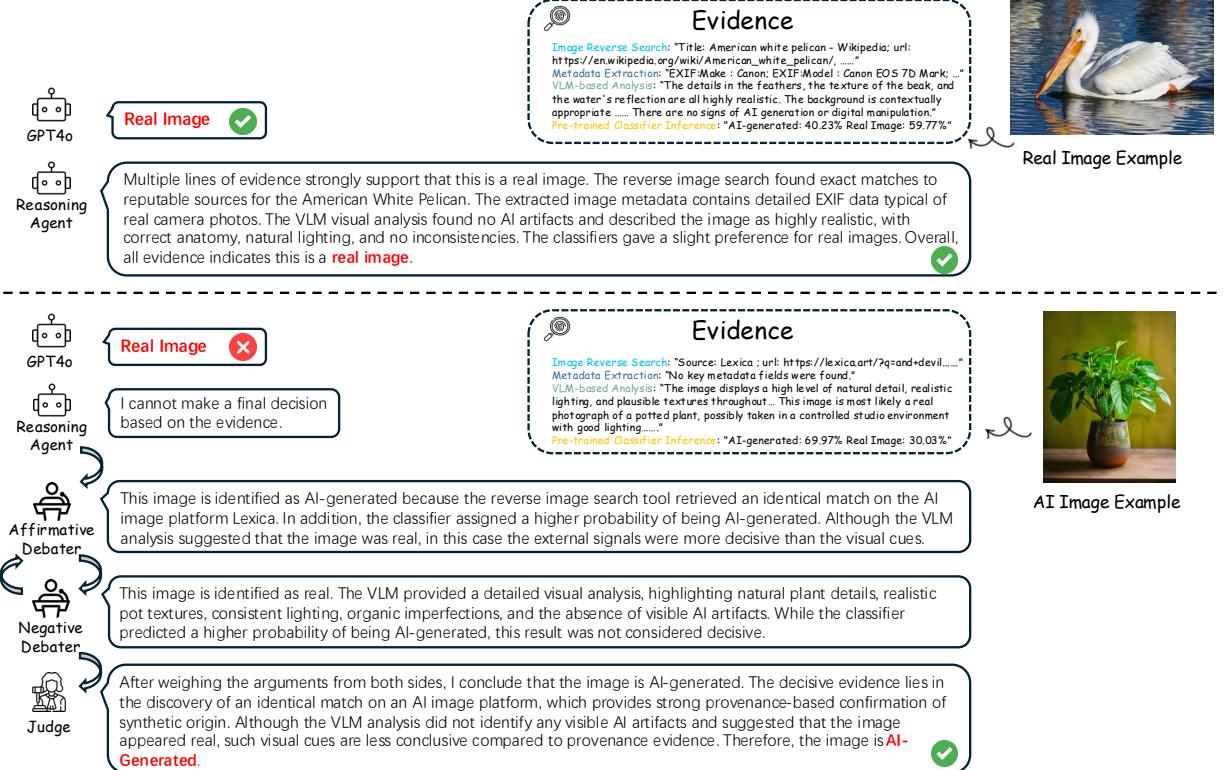


Figure 2: Examples of our agent framework’s decision-making process, demonstrating diverse evidence integration across different image types and sources.

able using the first reverse search tool. These tools thus act as high-precision, low-frequency decision anchors, exerting strong influence when available but not contributing in all cases. In contrast, the VLM analysis tool, pre-trained classifiers, and the second reverse search tool specialized in similar image retrieval achieve nearly full coverage. While the VLM’s visual reasoning remains relatively reliable, the outputs of the second reverse search tool and the classifiers are less dependable. This suggests that the agent will internally assign lower weights to these pieces of evidence. Hence, these tools serve as broad-coverage, moderate-confidence sources of evidence, providing additional context and support for the final decision but not dominating the reasoning process.

Overall, the agent appears to follow a tiered weighting strategy: prioritizing high-reliability tools when available, while leveraging high-coverage tools to maintain decision robustness.

Inference Efficiency and Cost Analysis. While AIFo achieves high detection accuracy, its architecture introduces additional computational cost compared to single model such as GPT-4o. To assess the efficiency of our framework, we measure the end-to-end latency and token usage for processing a single image. Table 4 summarizes the inference latency and token consumption per image for different methods. On average, AIFo requires 40.08 seconds per image, approximately 7.5 times slower than GPT-4o’s 5.31 seconds. The token usage also increases significantly, with AIFo consuming an average of 5230.86 tokens per image, compared to GPT-

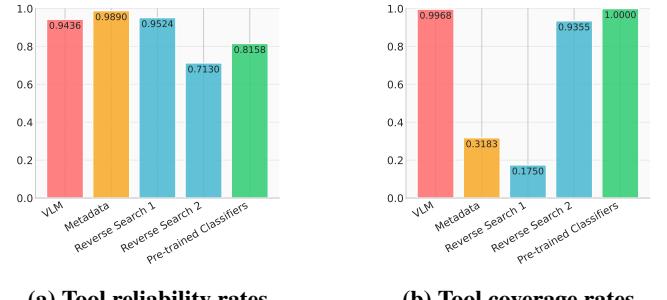


Figure 3: Analysis of individual tool contributions to the agent framework: (a) reliability rates measuring agent trust in each tool’s evidence, and (b) coverage rates showing the proportion of decisions where each tool provides informative evidence.

4o’s 715.05 tokens. The majority of the additional latency and token usage originates from the multi-round llm invocation and the execution of external tools. Despite the increased computational cost, AIFo achieves a consistent 2-3 % improvement in accuracy and over 4 % gain in recall. Meanwhile, it is important to note that AIFo operates in a training-free manner, eliminating the need for costly model retraining or fine-tuning as generative models evolve. Most importantly, unlike single model reasoning that relies solely on visual analysis, our framework integrates diverse sources of evidence, offering more verifiable interpretability that aligns with human forensic reasoning.

Table 4: Average inference latency and token usage per image.

Method	Avg. Latency	Avg. Token Usage
GPT-4o	5.31s	715.05
AIFo (w/o debate)	25.43s	2728.29
AIFo	40.08s	5230.86

Table 5: Performance comparison of AIFo vs GPT-4o under different image perturbations on the overall dataset. AIFo superior results are highlighted in bold.

GPT-4o				AIFo (Ours)			
Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Blu 0.8818	0.9662	0.7913	0.8701	0.9047	0.9380	0.8667	0.9009
Noi 0.9462	0.9866	0.9047	0.9438	0.9690	0.9879	0.9497	0.9684
Sha 0.9410	0.9926	0.8887	0.9377	0.9670	0.9902	0.9433	0.9662

4.4 Robustness Analysis

To assess the robustness of our multi-agent detection framework under realistic conditions, we evaluate its performance on perturbed versions of the dataset. The goal of this analysis is to determine whether the framework can maintain consistent accuracy when the input images undergo quality degradation. We apply a series of standard image degradation techniques, including: **Blurring**. We apply Gaussian blur with a fixed radius of 2, simulating defocus or motion blur; **Sharpening**. We enhance image sharpness using a sharpening factor of 2.0, which may alter local edge distributions; **Gaussian Noise**. We add Gaussian noise with mean 0 and variance 2 to simulate sensor-level or environmental noise. These image transformations represent common distortions encountered in real-world image acquisition or compression pipelines. For each perturbation type, we measure the detection performance on the perturbed images and compare the results against the best baseline, GPT-4o.

Table 5 presents the comprehensive performance comparison between our AIFo framework and the GPT-4o baseline across three types of image perturbations: Blurring (Blu), Gaussian noise (Noi), and sharpening (Sha). The results demonstrate that our framework maintains superior performance under these robustness testing conditions.

Our AIFo framework consistently outperforms GPT-4o across all perturbation types and most evaluation metrics. Under blurring conditions, both GPT-4o and AIFo experience noticeable performance degradation, as visual analysis tools are inherently sensitive to loss of a great amount of image pixel features. Nevertheless, AIFo still achieves an accuracy of 0.9047, outperforming GPT-4o's 0.8818. In contrast, under noisy and sharpened conditions, AIFo demonstrates remarkable robustness: its accuracy remains at 0.9710 and 0.9670 respectively, with only minimal decline compared to the clean setting, and consistently surpasses GPT-4o. This robustness can be attributed to the fact that, apart from pre-trained classifiers, the majority of tools integrated in our framework are largely insensitive to the perturbations, thereby stabilizing overall performance.

Table 6: Performance of AIFo under two evasive attack scenarios.

Attack Type	Acc	Prec	Rec	F1
None (clean)	0.9705	0.9920	0.9487	0.9698
Reverse-search manipulation	0.8971	0.8690	0.9353	0.9010
Metadata forgery	0.8702	0.8399	0.9147	0.8757

4.5 Evasive Attack Analysis

To evaluate the resilience of AIFo against evasive attacks, we simulate two representative attack scenarios that target the framework's key evidence sources. The first attack employs reverse image search manipulation techniques and the second attack involves metadata forgery. In the first scenario, we manipulate the provenance evidence returned by the reverse image search tool. Specifically, we use GPT-4o to generate counterfactual search results for each image. For real images, we fabricate search results indicating that the image was sourced from an AI generation platform, while for AI-generated images, we create results suggesting the image originated from a reputable photography website. These results are then injected into the evidence set returned to the agent. In the second scenario, we perform metadata forgery by swapping EXIF metadata between AI-generated and real images. Real images are randomly assigned metadata extracted from AI-generated samples, while AI-generated images receive real images' metadata.

As shown in Table 6, AIFo experiences a moderate performance degradation under both evasive attack settings. This degradation is primarily due to the framework's limitation in verifying the authenticity of evidence returned by external tools. Since AIFo is designed to treat tool outputs as trustworthy forensic sources, deliberately falsified information can mislead the agent, resulting in false judgments. To enhance robustness against such attacks, several potential defenses can be considered. Firstly, implementing cross-tool consistency validation can help identify conflicting evidence that may indicate manipulation such as verifying that metadata timestamps and reverse-search provenance sources are mutually coherent. Secondly, trust-weighted evidence aggregation can be introduced, where each tool's output is dynamically weighted based on historical reliability. Finally, incorporating external verification layers such as digital watermark authentication can help validate evidence before feeding it into the reasoning pipeline. These strategies would allow AIFo to better distinguish adversarially manipulated evidence, thereby improving its resilience against real-world evasive attacks.

4.6 Ablation Study

To investigate the individual contributions of different tool categories within our multi-agent framework, we conduct ablation experiments to quantify the impact of each of the four tool types on overall detection performance. To assess the importance of each category, we adopt a leave-one-out strategy: in each ablation run, we disable one tool category while keeping the others unchanged, and evaluate the resulting per-

formance on the full benchmark dataset. Meanwhile, we temporarily disable the debate mechanism to better isolate and focus on the contributions of the tools themselves. By observing the performance drop associated with the removal of each tool category, we can quantify its relative contribution to the framework’s overall effectiveness. This analysis provides insight into which components are most essential for robust and reliable detection, and helps guide future efforts in optimizing or simplifying the system without substantial loss in performance.

[Figure 4](#) presents the performance degradation observed when each tool category is disabled. The results clearly demonstrate that the vision-language model (VLM) tool contributes the most critical evidence to the overall detection pipeline. Removing the VLM tool causes the largest performance degradation across all metrics, with accuracy dropping below 0.85, and recall falling below 0.70. This indicates that direct image analysis from state-of-the-art VLMs remains the main source of evidence for the reasoning agent.

Disabling any of the *metadata extraction*, *reverse image search*, or *pre-trained classifier* modules leads to a slight decrease in both accuracy and F1 score. This indicates that the evidence provided by these tools offers complementary information that enhances the capability of the reasoning agent. Notably, even without these modules, our framework still surpasses the GPT-4o baseline, confirming that these tools contribute complementary signals that strengthen the robustness of the overall reasoning process.

Overall, the ablation study validates the necessity of a diverse Toolbox for reliable detection. These findings highlight that performance gains in our multi-agent framework arise not from a single dominant component, but from the effective integration of cross-source and complementary evidence sources.

4.7 Takeaway

Our evaluation demonstrates the advantages of the proposed multi-agent framework: it consistently achieves the highest detection performance on five benchmark datasets in the lab and in-the-wild images from six online platforms, covering AI-generated content produced by at least 20 models. The framework also shows stronger robustness than the best baseline, GPT-4o, against three common perturbations. We find that our framework can dynamically weigh the confidence of evidence gathered from each tool: when metadata and reverse search tools provide valid evidence, it tends to prioritize them; otherwise, it falls back on VLM analysis and traditional classifiers. The debate module further addresses conflicting evidence and prevents the framework from over-relying on a single source of evidence. Altogether, the framework demonstrates greater effectiveness and robustness compared to single-model approaches for AI-generated image detection.

5 Memory-Augmented Reasoning

5.1 Motivation and Design

Our current designed detectors operate as stateless processes, analyzing each image independently without leveraging historical context or learning from past decisions. However, forensic analysis is inherently iterative and cumulative, where insights from previous cases can inform future judgments [25]. To address this limitation, we introduce a memory-augmented reasoning module that enables the framework to maintain a knowledge base of past cases and decisions. As illustrated in [Figure 5](#), the module captures, indexes, and retrieves relevant historical context to inform current detection decisions. The designed knowledge base systematically collects both successful and failed detection cases, storing image embedding vectors together with their associated classification results, collected evidence, and reasoning processes. For failed cases, it additionally stores a reflective analysis that captures the causes of misclassification. When analyzing a new image, the system performs similarity-based retrieval over this knowledge base to identify relevant historical cases.

We hypothesize that the introduction of similar historical cases enables the agent to learn from experience and potentially adjust the weights of evidence from different forensic tools based on past performance patterns. Therefore, the enhanced system first queries its accumulated experience to identify potentially relevant historical cases rather than approaching each image as an isolated classification problem.

5.2 Implementation Details

Knowledge Base Dataset. To avoid data leakage, we construct the knowledge base image repository using a completely separate set of images from those used in the main benchmark dataset described in [Section 4.1](#). Following the same collection and sampling methodology as the benchmark dataset, we curate 600 new images for the knowledge base, containing 300 AI images and 300 real images.

Building Knowledge Base. Our knowledge base system is built on a hybrid architecture that combines vector similarity search for content retrieval with structured metadata storage for case management and analysis. The implementation leverages CLIP [42] embeddings to capture high-level semantic and visual features that enable effective similarity matching across images. Each image processed by the framework undergoes feature extraction using the CLIP-ViT-B/32 model, generating 512-dimensional dense vector representations. These embeddings serve as the primary indexing mechanism, enabling efficient similarity-based retrieval of historically relevant cases. The knowledge base maintains separate indices for successful and failed detection cases. For successful cases, the system records the complete set of evidence used in decision-making and the full analytical process. For failed cases, the system records all the above information and additionally stores a structured reflection, providing deeper insight into the causes of misclassification. The reflection is generated by GPT-4.1 to examine the complete analytical pathway and identify potential failure points. This

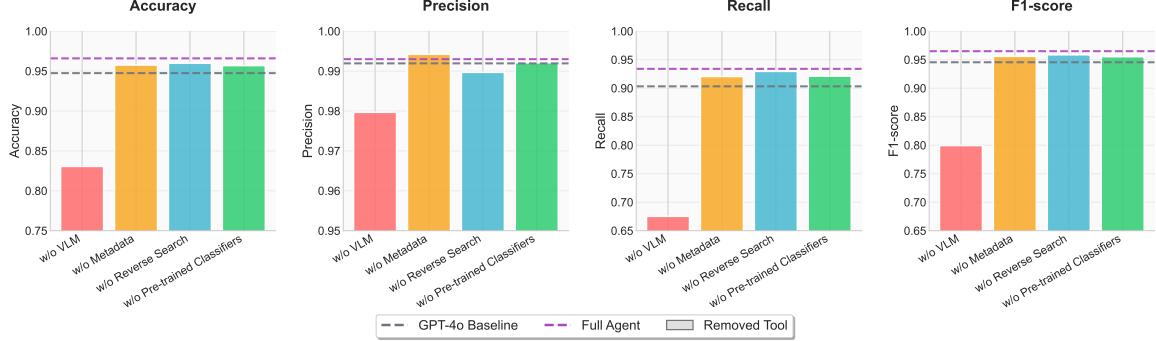


Figure 4: Performance degradation when each tool is disabled from the framework.

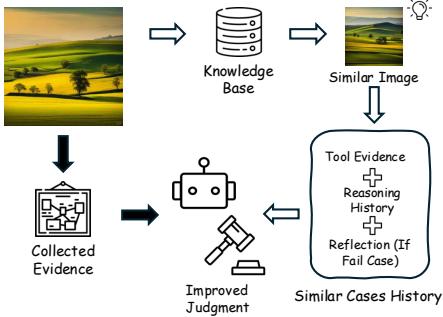


Figure 5: Overview of the memory-augmented reasoning module.

process produces insights covering evidence misinterpretation, tool reliability assessment, and reasoning inconsistencies. These generated reflections create a systematic learning mechanism that helps the system avoid repeating similar analytical errors. Finally, the knowledge base module is integrated seamlessly as one extra forensic tool. During the evidence collection phase, the module is activated.

Inference. Upon receiving a new image, the system computes its CLIP embedding and performs a similarity search over the knowledge base indices to retrieve the top- k most relevant historical cases, with defaults of $k = 1$ and a similarity threshold of 0.85 to ensure only sufficiently close matches are returned. The retrieved cases are then passed to the reasoning agent as additional context, allowing it to leverage past experiences to inform the current analysis.

5.3 Results Analysis

We did not incorporate the memory module in the main evaluation, as it requires ground-truth labels during the experience accumulation phase. Allowing such label-dependent learning at test time would introduce unfair advantages to our framework, compared to baselines that do not access these labels. Instead, to highlight the potential of the memory module, we directly focus on failure cases from the main evaluation. These are the scenarios where the knowledge base and additional experience are most needed.

To more directly assess its effectiveness, we conduct a targeted evaluation focusing on failure cases from the original framework. Specifically, we identify 50 misclassified im-

Table 7: Number of errors before and after incorporating similar case history. FP: real images misclassified as AI images; FN: AI images misclassified as real images.

	In-the-Lab		In-the-Wild		Total
	FP	FN	FP	FN	
Before	2	12	1	35	50
After	1	6	0	22	29

ages from the benchmark set for which the knowledge base contained semantically similar counterparts. We then re-evaluated these images with the knowledge base module enabled, examining whether the additional contextual evidence could help correct the prior misjudgments. Table 7 summarizes the results by different image types and settings, showing the number of errors successfully corrected after incorporating knowledge base evidence. The results demonstrate that over 40% of error cases are successfully corrected when the memory module is employed. Furthermore, the example illustrated in Figure 6 highlights how retrieved similar cases influenced the reasoning process. When presented with similar past cases, the agents internally reweighted the credibility of different forensic tools, reinterpreted conflicting evidence, and ultimately arrived at the correct conclusion. These observations suggest that the memory-augmented reasoning offers substantial benefits in failure recovery and demonstrates strong potential for enhancing adaptive reasoning in forensic scenarios.

6 Discussion and Limitations

Our AIFo framework potentially represents a paradigm shift in AI-generated image detection by emulating human forensic reasoning through multi-agent collaboration. The framework’s training-free nature and cross-model generalizability address key limitations of existing detection methods, offering a more sustainable and adaptable solution for the rapidly evolving landscape of generative AI.

Despite the promising results, our approach has several important limitations that warrant careful consideration.

Scalability and Computational Efficiency. While our multi-agent approach achieves high accuracy, it comes with increased computational overhead compared to single-model solutions. The sequential and parallel execution of multiple

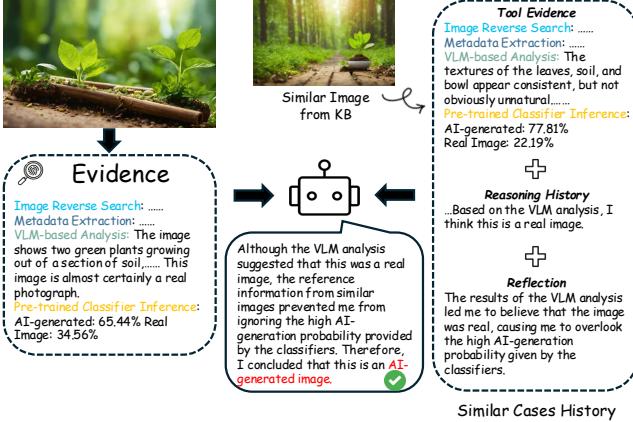


Figure 6: Example of memory-augmented reasoning showing how historical cases influence the agent’s decision-making process and evidence weighting.

forensic tools, combined with LLM-based reasoning, results in higher latency and resource consumption. For large-scale deployment scenarios, optimization strategies such as tool prioritization, caching mechanisms, and selective tool activation based on image characteristics could help balance accuracy and efficiency.

Dependency on External Services. The framework’s effectiveness is inherently tied to the availability and reliability of external services, particularly reverse image search APIs and metadata extraction tools. Changes in API policies, service outages, or modifications to search algorithms could impact the framework’s performance. This dependency creates potential points of failure that are beyond the system’s direct control.

Adversarial Metadata Manipulation. Our framework lies in its reliance on image EXIF metadata as a key source of forensic evidence. However, adversaries could potentially manipulate image metadata to mislead the detection system. For instance, attackers could inject fake EXIF data mimicking legitimate camera parameters into AI-generated images. Such metadata spoofing attacks could compromise the reliability of our metadata extraction tool, which currently shows high reliability rates in our evaluation.

7 Related Work

7.1 Fake Image Detection

Early approaches to AI-generated image detection mainly rely on training image classifiers using machine learning, often using datasets produced by specific generative models. Over time, research has shifted towards more generalizable and robust detection strategies, including leveraging the visual capabilities of large multimodal models (LMMs) such as CLIP [42]. CNNSpot [53] is one of the first works to propose a “universal” detector capable of distinguishing real images from CNN-generated ones without being dependent on a particular network architecture or training dataset. De-Fake [46] extends the detection pipeline by incorporating both the image content and generated prompts using CLIP to train a hy-

brid classifier. DIRE [54] and ZeroFake [47] exploit intrinsic differences between real and fake images revealed during the diffusion model reconstruction process to build detection models with improved generalization. PatchCraft [59] focuses on local texture patches, identifying subtle artifacts left by generative models in fine-grained regions.

Recent works utilize VLMs as fake image detectors, such as AntifakePrompt [18], the method proposed by Jia *et al.* [29], and the approach by Ji *et al.* [28]. AIGI-Holmes [60] proposes a complete framework to train a VLM for explainable and generalizable detection, aiming to produce human-verifiable justifications. The work by Yu *et al.* [57] develops a framework to enhance generalization and explainability by using a knowledge-guided detector and a forgery-aware prompt learner. FakeBench [33] and DFBench [51] introduced a large-scale benchmark to rigorously test the detection performance of LMMs against a wide range of modern generative models.

Despite these progresses, existing methods exhibit inherent limitations. These approaches predominantly rely on internal, pixel-level visual features, while overlooking complementary external sources of evidence that can be critical for reliable forensic analysis. In contrast to standalone detector architectures, our work adopts an agent-based framework that emulates the investigative workflow of human forensic experts. The proposed AI agent leverages a diverse set of external tools to systematically collect, integrate, and reason over different evidence, which allows for a more robust, context-aware, and explainable framework for AI-generated image detection.

7.2 LLM-Based Multi-Agent Frameworks

Recent advances in large language models (LLMs) have catalyzed the development of agentic frameworks that simulate complex human-like workflows by coordinating multiple specialized agents. ReAct [56] introduces a framework that switches between reasoning and acting within language models. Other works such as MetaGPT [24] simulate various roles in a software company and build a multi-agent software development framework. In the security domain, several studies have developed multi-agent frameworks tailored to diverse security-related downstream tasks, including adversarial defense [58], harmful content detection [36], bias identification in generative models [52], fake news verification [32], and comparative evaluation of misinformation detection strategies [27]. However, existing approaches do not specifically target zero-shot detection tasks nor address the unique challenges of AI-generated image detection. To the best of our knowledge, our work is the first to develop a training-free, zero-shot multi-agent framework specifically designed for AI-generated image detection. Combining the strengths of existing image detection methods and LLM agents, we present a novel and practical solution for addressing the evolving challenges in AI-generated image forensics.

8 Conclusion

In this work, we introduce AIFo, a novel multi-agent framework that advances AI-generated image detection by emulating human forensic reasoning. Unlike conventional methods, AIFo integrates diverse forensic tools and uses multi-agent collaboration to synthesize evidence. Our evaluation on a comprehensive benchmark spanning both lab and real-world settings shows that AIFo achieves 97.05% accuracy, substantially outperforming traditional classifiers and state-of-the-art VLMs like GPT-4o.

The key contributions include a training-free, agent-based paradigm that ensures generalizability through procedural reasoning, a finding that multi-source evidence integration and structured debate mechanisms significantly enhance both accuracy and interpretability, and the establishment of a comprehensive benchmark that enables evaluation under real-world conditions. While our framework represents a significant advancement in AI-generated image detection, it also faces some limitations including computational overhead and dependency on external services.

As generative AI technologies continue to evolve rapidly, the need for robust, interpretable, and adaptable detection systems becomes increasingly critical. Our agent-based approach offers a sustainable foundation for addressing these challenges, providing a framework that can generalize alongside advancing generative models while maintaining the transparency and reliability essential for real-world deployment in security-critical applications.

Ethical Considerations

In this study, we adopted a stakeholder-oriented perspective to examine the ethical dimensions of our work. For the research team, the development and validation of our new detection framework contributed to advancing technical expertise and academic reputation. For the general public, the framework offers a practical tool to mitigate the spread of misinformation by improving the detection of AI-generated images, thereby safeguarding individuals from being misled. Companies such as social media platforms and news organizations may also benefit by employing the framework to verify content authenticity and maintain the credibility of their services. Our research is guided by several core ethical principles. First, the principle of beneficence is reflected in our aim to protect society from the harmful consequences of misinformation. Second, respect for persons is ensured by using only publicly available datasets that do not involve personal or sensitive information. Third, the principle of justice informed our effort to design a framework whose outcomes can be applied broadly across different social groups, thereby promoting fair access to reliable information. We think that this research provides substantial value in promoting information authenticity and strengthening public trust. We are therefore confident that the study is ethically sound and makes a meaningful contribution to the ongoing development of AI-generated image detection.

References

- [1] <https://huggingface.co/>. 5
- [2] <https://openai.com/dall-e-3>. 6, 7
- [3] AI Image Detector. <https://huggingface.co/haywoodsloan/ai-image-detector-deploy>. 5
- [4] Anime Image Detector. <https://huggingface.co/legekka/AI-Anime-Image-Detector-ViT>. 6
- [5] Civitai. <https://civitai.com>. 6, 7, 20
- [6] Flickr. <https://flickr.com>. 6, 7
- [7] Google Cloud Vision. <https://cloud.google.com/vision?hl=en>. 5
- [8] GPT-4.1. <https://openai.com/index/gpt-4-1/>. 2, 6, 8, 20
- [9] Lexica. <https://lexica.art/>. 6, 7, 20
- [10] NightCafe. <https://creator.nightcafe.studio/>. 6, 7, 20
- [11] NYUAD AI Image Detector. https://huggingface.co/NYUAD-ComNets/NYUAD_AI-generated_images_detector. 6
- [12] SDXL-Detector. <https://huggingface.co/Organika/sdxl-detector>. 6
- [13] SMOGY AI Image Detector. <https://huggingface.co/Smogy/SMOGY-Ai-images-detector>. 6
- [14] Wikimedia Commons. <https://commons.wikimedia.org>. 6, 7
- [15] As Social Media Guardrails Fade and AI Deepfakes Go Mainstream, Experts Warn of Impact on Elections. <https://apnews.com/article/election-2024-misinformation-ai-social-media-trump-6119ee6f498db10603b3664e9ad3e87e>, 2023. 1
- [16] Eirikur Agustsson and Radu Timofte. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1122–1131. IEEE, 2017. 6, 7
- [17] Xiuli Bi, Bo Liu, Fan Yang, Bin Xiao, Weisheng Li, Gao Huang, and Pamela C. Cosman. Detecting Generated Images by Real Images Only. *CoRR abs/2311.00962*, 2023. 1
- [18] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. AntifakePrompt: Prompt-Tuned Vision-Language Models are Fake Image Detectors. *CoRR abs/2310.17419*, 2023. 13
- [19] Davide Cozzolino, Giovanni Poggi, Matthias Nießner, and Luisa Verdoliva. Zero-Shot Detection of AI-Generated Images. In *European Conference on Computer Vision (ECCV)*, pages 54–72. Springer, 2024. 1
- [20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009. 6, 7
- [21] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving Factuality and Reasoning in Language Models through Multi-Agent Debate. In *International Conference on Machine Learning (ICML)*. JMLR, 2024. 2
- [22] Teppo Felin and Matthias Holweg. Theory is All You Need: AI, Human Cognition, and Causal Reasoning. *Strategy Science*, 2024. 3

- [23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2020. 1
- [24] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *International Conference on Learning Representations (ICLR)*. JMLR, 2024. 13
- [25] Md. Tanzib Hosain, Salman Rahman, Md. Kishor Morol, and Md. Rizwan Parvez. Xolver: Multi-Agent Reasoning with Holistic Experience Learning Just Like an Olympiad Team. *CoRR abs/2506.14234*, 2025. 11
- [26] Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. Holopix50k: A Large-Scale In-the-wild Stereo Image Dataset. *CoRR abs/2003.11172*, 2020. 6, 7
- [27] Tianyi Huang, Jingyuan Yi, Peiyang Yu, and Xiaochuan Xu. Unmasking Digital Falsehoods: A Comparative Analysis of LLM-Based Misinformation Detection Strategies. *CoRR abs/2503.00724*, 2025. 13
- [28] Yikun Ji, Yan Hong, Jiahui Zhan, Haoxing Chen, Jun Lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang. Towards Explainable Fake Image Detection with Multi-Modal Large Language Models. *CoRR abs/2504.14245*, 2025. 1, 13
- [29] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4324–4333. IEEE, 2024. 13
- [30] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116. IEEE, 2020. 7
- [31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *CoRR abs/2201.12086*, 2022. 6
- [32] Xinyi Li, Yongfeng Zhang, and Edward C. Malthouse. Large Language Model Agent for Fake News Detection. *CoRR abs/2405.01593*, 2024. 13
- [33] Yixuan Li, Xuelin Liu, Xiaoyang Wang, Bu Sung Lee, Shiqi Wang, Anderson Rocha, and Weisi Lin. FakeBench: Probing Explainable Fake Image Detection via Large Multimodal Models. *CoRR abs/2404.13306*, 2024. 1, 2, 6, 7, 13, 20
- [34] Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving Multi-Agent Debate with Sparse Communication Topology. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7281–7294. ACL, 2024. 2
- [35] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 17889–17904. ACL, 2024. 2
- [36] Ziyan Liu, Chunxiao Fan, Haoran Lou, Yuexin Wu, and Kaiwei Deng. MIND: A Multi-agent Framework for Zero-shot Harmful Meme Detection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 923–947. ACL, 2025. 2, 13
- [37] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *CoRR abs/2112.10741*, 2021. 1
- [38] Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, and Ming Li. Large Language Models and Cognitive Science: A Comprehensive Review of Similarities, Differences, and Challenges. *CoRR abs/2409.02387*, 2024. 3
- [39] OpenAI. GPT-4o. <https://openai.com/index/hello-gpt-4o/>. 2, 8, 20
- [40] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649. IEEE, 2015. 2, 6, 7
- [41] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *CoRR abs/2307.01952*, 2023. 6
- [42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 11, 13
- [43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR abs/2204.06125*, 2022. 1, 6
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695. IEEE, 2022. 1, 6, 7
- [45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasedipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *CoRR abs/2205.11487*, 2022. 1
- [46] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Diffusion Models. *CoRR abs/2210.06998*, 2022. 1, 2, 7, 8, 13, 20
- [47] Zeyang Sha, Yicong Tan, Mingjie Li, Michael Backes, and Yang Zhang. ZeroFake: Zero-Shot Detection of Fake Images Generated and Edited by Text-to-Image Generation Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 4852–4866. ACM, 2024. 13
- [48] Mohamed R. Shoaib, Zefan Wang, Milad Taleby Ahvanooy, and Jun Zhao. Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models. In *International Conference on Computer and Applications (ICCA)*, pages 1–7. IEEE, 2023. 1

- [49] Andries P. Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs. In *International Conference on Machine Learning (ICML)*. JMLR, 2024. [2](#)
- [50] Luisa Verdoliva. Media Forensics and DeepFakes: An Overview. *Journal of Selected Topics in Signal Processing*, 2020. [1](#)
- [51] Jiarui Wang, Huiyu Duan, Junlong Wang, Ziheng Jia, Woo Yi Yang, Xiaorong Zhu, Yu Zhao, Jiaying Qian, Yuke Xing, Guangtao Zhai, and Xiongkuo Min. DFBench: Benchmarking Deepfake Image Detection Capability of Large Multimodal Models. *CoRR abs/2506.03007*, 2025. [1, 13](#)
- [52] Qichao Wang, Tian Bian, Yian Yin, Tingyang Xu, Hong Cheng, Helen M. Meng, Zibin Zheng, Liang Chen, and Bingzhe Wu. Language Agents for Detecting Implicit Stereotypes in Text-to-image Models at Scale. *CoRR abs/2310.11778*, 2023. [13](#)
- [53] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8692–8701. IEEE, 2020. [1, 2, 7, 8, 13, 20](#)
- [54] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for Diffusion-Generated Image Detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 22388–22398. IEEE, 2023. [1, 13](#)
- [55] Jialiang Xu, Michael Moor, and Jure Leskovec. Reverse Image Retrieval Cues Parametric Memory in Multimodal LLMs. *CoRR abs/2405.18740*, 2024. [5](#)
- [56] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*. ICLR, 2023. [13](#)
- [57] Peipeng Yu, Jianwei Fei, Hui Gao, Xuan Feng, Zhihua Xia, and Chip-Hong Chang. Unlocking the Capabilities of Vision-Language Models for Generalizable and Explainable Deepfake Detection. *CoRR abs/2503.14853*, 2025. [1, 13](#)
- [58] Yifan Zeng, Yiran Wu, Xiao Zhang, Huazheng Wang, and Qingyun Wu. AutoDefense: Multi-Agent LLM Defense against Jailbreak Attacks. *CoRR abs/2403.04783*, 2024. [13](#)
- [59] Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. PatchCraft: Exploring Texture Patch for Efficient AI-generated Image Detection. *CoRR abs/2311.12397*, 2024. [1, 7, 8, 13, 20](#)
- [60] Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. AIGI-Holmes: Towards Explainable and Generalizable AI-Generated Image Detection via Multimodal Large Language Models. *CoRR abs/2507.02664*, 2025. [1, 13](#)
- [61] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2023. [2, 6, 7, 20](#)

A Prompts Used in Our Framework

This appendix documents the exact prompts used in our experiments to facilitate reproducibility. We group prompts according to their roles in the multi-agent framework.

A.1 Evidence Gatherer Agent Prompt

The Evidence Gatherer Agent collects cross-source forensic signals. [Table 8](#) is the prompt template used to instruct the LLM:

You are an AI Image Forensics Expert. Your task is to determine whether the input image is AI-generated or real using the available forensic tools.

- A real image refers to images created by humans, including photographs captured by cameras, photos that have been edited with software such as Photoshop, or human artistic creations such as hand-drawn sketches and paintings.
- An AI-generated image refers to images that are fully or partially generated by AI models.

Available Tools:

- reverse_search: Perform a reverse image search to find exact matches or similar appearances online.
- extract_image_metadata: Inspect technical EXIF metadata for authenticity cues.
- vlm_analysis: Obtain expert-level visual analysis of the image content.
- pre-trained_classifiers: Apply dedicated AI-generated image detection models.

Your role is to systematically invoke these tools as needed and collect evidence that will later be assessed to determine the authenticity of the input image.

Table 8: Prompt template for the Evidence Gatherer Agent.

A.2 Reasoning Agent Prompt

The Reasoning Agent first assesses whether the evidence is sufficient and consistent to support a decision. If so, it synthesizes all sources to produce a final binary judgment with an explanation that evaluates source reliability. [Table 9](#) and [Table 10](#) are the prompt templates used to instruct the LLM:

A.3 Debate Agents Prompt

The Debate Agents engage in a structured debate to resolve conflicts and ambiguities in the evidence. [Table 11](#) is the prompt template used to instruct the Pro-Agent LLM:

[Table 12](#) is the prompt template used to instruct the Con-Agent LLM:

A.4 Judge Agent Prompt

The Judge Agent is tasked with overseeing the debate process and synthesizing the final decision based on both the tool-derived evidence and the debate history. The Judge also

You are an AI Image Forensics Expert. Your task is to determine if the following evidence collected from multiple tools is sufficient and consistent enough to make a final judgment.

{tool_results}

Answer 'True' if the evidence is both sufficient and consistent enough to confidently reach a final decision and 'False' if the evidence is incomplete, ambiguous, or contains major conflicts that require further debate and analysis.

Table 9: First prompt template for the Reasoning Agent.

You are an AI Image Forensics Expert. Your task is to determine whether the image is ai-generated or a real image.

- A real image refers to images created by humans, including photographs captured by cameras, photos that have been edited with software such as Photoshop, or human artistic creations such as hand-drawn sketches and paintings.
- An AI-generated image refers to images that are fully or partially generated by AI models.

Please make a final judgment based on the following evidence collected from multiple tools:

{tool_results}

Critically evaluate each evidence source and its reliability.

Required output format:

1. is_ai_generated: boolean (True if AI-generated, False if real image)
2. analysis_details: A detailed analysis explaining your decision

Table 10: Second prompt template for the Reasoning Agent.

evaluates the sufficiency of each debate round and can decide to terminate the debate early if the arguments are deemed sufficient. [Table 13](#) and [Table 14](#) are the prompt templates used to instruct the LLM:

A.5 VLM Analysis Tool Prompt

The VLM Analysis Tool utilizes vision-language models to conduct in-depth visual analysis of images. The prompt used to guide the VLM Analysis Tool is detailed in [Table 15](#), ensuring the model focuses on key visual characteristics and provides comprehensive evidence to support its classification.

You are an AI Image Forensics Expert. Your goal is to correctly classify an image as either AI-generated or real.

Your analysis must be based on the evidence provided in the tool results below.

Tool Results:

{tool_results}

#First Round Only:

You are arguing in favor of the image being AI-generated.

Scrutinize the tool results for any artifacts, inconsistencies, or patterns typical of AI generation. Present your findings as a concise, bullet-pointed list. Focus on the strongest pieces of evidence that support your assigned perspective.

#Subsequent Rounds Only:

Review the other expert's points from the previous round and re-evaluate your own position.

- Acknowledge any valid points they made.
- Re-examine the tool results to see if their perspective reveals something you missed.
- Refine or strengthen your analysis based on this new information. Your updated analysis should be more nuanced.

You are arguing in favor of the image being AI-generated.

The other expert's (arguing for "Real") points:

{negative_history}

Provide your updated, refined analysis as a concise bullet-pointed list.

Table 11: Prompt template for the Pro-Agent.

B Metadata Analysis Tool Key Fields

[Table 16](#) provides the exact key fields and prefixes used in the metadata analysis tool to identify authenticity markers in images.

C Detailed Accuracy performance

[Table 17](#) provides a detailed breakdown of the accuracy performance of various methods across different image sources, including both in-the-lab and in-the-wild scenarios.

D AI Model Sources

[Table 18](#) provides an overview of the AI models used for generating images in our benchmark's AI-sourced datasets.

You are an AI Image Forensics Expert. Your goal is to correctly classify an image as either AI-generated or real.

Your analysis must be based on the evidence provided in the tool results below.

Tool Results:
{tool_results}

First Round Only:

You are arguing in favor of the image being authentic (real).

Look for signs of naturalness, photographic properties, and details that are hard for AI to replicate, based on the tool results.

Subsequent Rounds Only:

Review the other expert's points from the previous round and re-evaluate your own position.

- Acknowledge any valid points they made.
- Re-examine the tool results to see if their perspective reveals something you missed.
- Refine or strengthen your analysis based on this new information. Your updated analysis should be more nuanced.

You are arguing in favor of the image being authentic (real).

The other expert's (arguing for "AI-generated") points:

{positive_history}

Provide your updated, refined analysis as a concise bullet-pointed list.

As an impartial judge, review the debate history so far. Your task is NOT to make the final decision, but to determine if the debate is sufficient to support a final decision.

Arguments for 'AI-generated':
{positive_args}

Arguments for 'Authentic Image':
{negative_args}

Your Decision Criteria:

1. If one side's evidence is strong and the other's is weak or has been effectively countered, the information is likely sufficient.
2. If both sides have presented compelling but conflicting evidence that has not yet been reconciled, more analysis is needed.
3. If the discussion become repetitive, further rounds are unlikely to be productive.

Based on these criteria, decide if you have enough information to make a high-confidence final judgment. Answer 'True' if sufficient, 'False' if more debate and analysis would be helpful.

Table 12: Prompt template for the Con-Agent.

Table 13: First prompt template for the Judge Agent.

You are an AI Image Forensics Judge. Your role is to synthesize all available information and deliver a definitive, well-reasoned judgment on whether the image is AI-generated or real.

- A real image refers to images created by humans, including photographs captured by cameras, photos that have been edited with software such as Photoshop, or human artistic creations such as hand-drawn sketches and paintings.
- An AI-generated image refers to images that are fully or partially generated by AI models.

Raw Evidence from tools: tool_results

Arguments for 'AI-generated':
{positive_args}

Arguments for 'Authentic Image':
{negative_args}

Your analysis must be a comprehensive synthesis. Follow these steps in your reasoning:

1. Weigh the Evidence: Identify the most compelling piece of evidence from EACH side.
2. Resolve the Core Conflict: Directly address the central disagreement.
3. State Your Final Conclusion: Based on your analysis, provide a clear final verdict.

Required output format:

1. is_ai_generated: boolean (True if AI-generated, False if real image)
2. analysis_details: A detailed analysis explaining your decision

Format the response as a structured object.

As a professional AI image detector, please analyze this image carefully:

1. Determine if this is an AI-generated image or a real image.
 - Real images include images that are created by humans, including photographs captured by cameras, photos that have been edited with software such as Photoshop, or human artistic creations such as hand-drawn sketches and paintings.
 - AI-generated images include images that are fully or partially generated by AI models.
2. If you determine it's an AI-generated image, please specifically identify and list the visual artifacts or characteristics that indicate AI generation, such as:
 - Unnatural textures or patterns
 - Inconsistent lighting or shadows
 - Anatomical errors in humans or animals
 - Unusual distortions or blending of elements
 - Text or writing abnormalities
 - Symmetry issues or repeating patterns
 - Unusual backgrounds or contextual inconsistencies
3. If you determine it's a real image, explain what characteristics support this conclusion.
4. Provide your final classification with confidence level (high, medium, or low).

Table 15: Prompt template for the VLM Analysis Tool.

Table 14: Second prompt template for the Judge Agent.

Table 16: Metadata fields and prefixes considered in the analysis tool.

Category	Field / Prefix	Description
<i>Exact Key Fields</i>		
XMP:CreatorTool	Creator tool	Software used to generate or edit the image.
EXIF:Software	Software tag	Image editing or generation software information.
EXIF:UserComment	User comment	Arbitrary comments added to the image metadata.
File:Comment	File comment	Comments embedded directly in the file container.
XMP:Description	Description	Textual description of the image.
XMP:Title	Title	Title field embedded in XMP metadata.
XMP:Rights	Rights	Usage rights or copyright information.
XMP:Source	Source	Original source reference of the image.
EXIF:Make	Camera make	Manufacturer of the recording equipment.
EXIF:Model	Camera model	Camera model used for the photo.
EXIF:LensModel	Lens model	Lens information recorded by the camera.
EXIF:LensInfo	Lens info	Technical specifications of the lens.
EXIF:LensSerialNumber	Lens serial number	Unique identifier of the lens.
EXIF:ExposureTime	Exposure time	Shutter exposure duration.
EXIF:FNumber	F-number	Aperture size of the lens.
EXIF:ISO	ISO	Sensitivity setting of the camera.
EXIF:FocalLength	Focal length	Lens focal length value.
EXIF:SerialNumber	Camera serial number	Unique identifier of the camera.
EXIF:GPSLatitude	GPS latitude	Geographic latitude of capture.
EXIF:GPSLongitude	GPS longitude	Geographic longitude of capture.
EXIF:GPSTimeStamp	GPS timestamp	Time recorded by GPS.
EXIF:DateTimeOriginal	Original datetime	Original capture time of the image.
EXIF:CreateDate	Creation date	File creation date.
Composite:GPSPosition	GPS position	Combined GPS coordinates.
Composite:Aperture	Aperture	Derived aperture value.
Composite:ShutterSpeed	Shutter speed	Derived shutter speed.
Composite:LensID	Lens ID	Identifier for the lens model.
ICC_Profile:ProfileDescription	ICC profile description	Description of the color profile.
ICC_Profile:ProfileCopyright	ICC profile copyright	Copyright information for the ICC profile.
IPTC:DocumentNotes	Document notes	Notes in IPTC metadata.
IPTC:ApplicationRecordVersion	Record version	Version of the IPTC application record.
<i>Key Field Prefixes</i>		
MakerNotes:	Camera-specific notes	Manufacturer-specific EXIF metadata.
JUMBFI	JUMBFI metadata	Metadata block for embedding auxiliary information.
MPF:	Multi-picture format	Metadata for multi-frame images.

Table 17: Detailed Accuracy performance of different methods on each image sources.

Method	In-the-Lab					In-the-Wild					
	Real Images			AI Images		Real Images			AI Images		
	Flickr30k	ImageNet	DIV2k	GenImage	FakeBench	Holopix50k	Flickr	W. Commons	Lexica	Nightcafe	Civitai
CNNSpot [53]	0.9980	0.9980	1.0000	0.0475	0.1857	1.0000	1.0000	0.9980	0.0000	0.0020	0.0000
PatchCraft [59]	0.9900	0.9140	0.7680	0.8525	0.5986	0.8600	0.9040	0.5960	0.0140	0.2360	0.3360
DE-FAKE [46]	0.8980	0.7280	0.3480	0.7225	0.6443	0.7180	0.5700	0.4920	0.9840	0.9380	0.8360
GPT-4.1 [8]	1.0000	0.9920	0.9880	0.8350	0.9086	0.9980	0.9860	0.9920	0.7840	0.9900	0.9500
GPT-4o [39]	1.0000	0.9860	0.9960	0.8850	0.9471	0.9980	0.9820	0.9940	0.7520	0.9900	0.9420
AIFo (ours)	1.0000	0.9840	0.9920	0.9475	0.9657	0.9960	0.9880	0.9940	0.8420	0.9880	0.9940

Table 18: Overview of the AI models and platforms used for generating images in our benchmark’s AI-sourced datasets. The table is categorized by the *In-the-Lab* and *In-the-Wild* settings.

Dataset Source	AI Models and Platforms Used for Generation
<i>In-the-Lab AI Image Sources</i>	
GenImage [61]	BigGAN, GLIDE, VQDM, ADM, Midjourney, Wukong, and Stable Diffusion (v1.4, v1.5).
FakeBench [33]	ProGAN, StyleGANs, CogView2, FuseDream, VQDM, GLIDE, Midjourney, Stable Diffusion, DALL-E 2, and DALL-E 3.
<i>In-the-Wild AI Image Sources</i>	
Lexica [9]	Lexica Aperture Series (v3.5, v4, v5, Max).
Nightcafe [10]	DALL-E 2, DALL-E 3, Stable Diffusion, and various other community fine-tuned models.
Civitai [5]	A vast collection of community fine-tuned models, predominantly based on Stable Diffusion (including SDXL variants) series.